Routing Layer Sharing: A New Opportunity for Routing Optimization in Monolithic 3D ICs

Sai Pentapati
sai.pentapati@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Sung Kyu Lim
limsk@ece.gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT

A 3D Integrated Circuit consists of two or more dies bonded to each other in the vertical direction. This allows for a high transistor density without a need for shrinking the underlying transistor dimensions. While it has been shown to improve design power, performance, and area (PPA) due to the stacked Front End Of the Line (FEOL) layers, the Back End Of the Line (BEOL) structure of the stacked IC also allows for novel routing scenarios. With the split dies in 3D, nets would need to connect cells from different tiers, across many vertical layers and multiple FEOLs. More importantly, nets connecting cells in a single tier could still use metal layers from the BEOL of other tiers to complete routing. This is referred to as routing / metal layer sharing. While such sharing creates additional 3D connections, it can also be utilized to improve several aspects of the design such as cost, routing congestion, and performance. In this paper, we analyze the nets with metal layer sharing in 3D and provide ways to control the number of 3D connections. We show that the configuration of the 3D BEOL stack helps with metal layer cost reduction with up to 1-2 fewer layers needed to complete routing without a noticeable timing impact. Sharing also allows for a better distribution of wirelength in the BEOL stack that can achieve significant reduction in metal layer congestion of top most layer by up to a 50% reduction of its track usage. Finally, we also see performance benefits of up to 16% with the help of metal layer sharing in 3D IC design.

CCS CONCEPTS

• **Hardware → 3D integrated circuits**; **Wire routing**.

KEYWORDS

Monolithic 3D IC; Interconnect Analysis; Metal Layer Sharing

ACM Reference Format:
Sai Pentapati and Sung Kyu Lim. 2022. Routing Layer Sharing: A New Opportunity for Routing Optimization in Monolithic 3D ICs. In *Proceedings of the 2022 International Symposium on Physical Design (ISPD'22), March 27–30, 2022, VirtualEvent, Canada.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3505170.3506729

ISPD'22, March 27–30, 2022, Virtual Event, Canada

ACM ISBN 978-1-4503-9210-5/22/03…$15.00
https://doi.org/10.1145/3505170.3506729

1 INTRODUCTION

The traditional approach to technology scaling has been useful to increase transistor PPA and density, creating ever powerful consumer chips. On the other hand, 3D stacking of dies adds a new dimension to connect and place the standard cells [1]. This improves the overall chip density and the power, performance, and area (PPA) of the chip. In these chips, the stacking technology influences the maximum number of connections possible and the types of partitioning supported. Micro bump bonding, Hybrid bonding, Monolithic 3D (M3D) IC are the three main categories of the bonding structures [2].

Micro bump bonding uses large bumps to bond two independently fabricated dies on top of each other. The pitch of the bumps is usually in the order of 10 µm to help with the alignment. The large pitch doesn't allow for a high 3D connection density and limits the partitioning options possible. Hybrid bonding is similar to micro-bump as the two dies are fabricated independently but are bonded using bumps with a much finer pitch of 1 µm due to the better alignment accuracy and the smaller size of the bond pads. This allows for a wide range of partitioning options. By independently fabricating the two dies, it is possible to increase overall yield in 3D using known good dies of half footprint of 2D design.

On the extreme end of the 3D bonding techniques, is the sequential fabrication of the two tiers – a Monolithic 3D (M3D) IC. This removes the need for die alignment and allows for 3D pitch in the order of 0.1 µm. Each style of 3D IC provides different benefits in terms of cost, PPA, and partitioning types. In our discussions, we limit our analysis to two tiers of 3D ICs.

The orientation of stacked dies is another important part of the 3D IC design, and it can differ as shown in Fig. 1: Face-to-Face, Face-to-Back. In general, a 2D IC has the devices on the Front End Of Line (FEOL), and metal layers in the Back End of Line (BEOL). In a Face-to-Face 3D IC, the two dies are attached at the BEOL boundary of two tiers with the FEOLs facing each other at the either ends of the full die stack. As the top tier devices are inverted, a sequentially fabricated M3D IC does not allow for Face-To-Face stacking. Only Face-To-Back orientation is allowed for M3D, where the top-side of the bottom layer BEOL is attached to the back of the top tier FEOL (shown in green in Fig. 1). Here, the 3D via connecting the metal layers of the two tiers blocks the silicon area in the top tier. Therefore, it is important to use smaller sized vias to minimize this blocked area. The 3D IC stack allows for an interaction between the BEOL and FEOL of different tiers creating a new paradigm for placement and routing in 3D ICs. While several studies were conducted to develop better algorithms for placement [3–5] and partitioning [6, 7] in 3D, routing has been largely left unexplored.

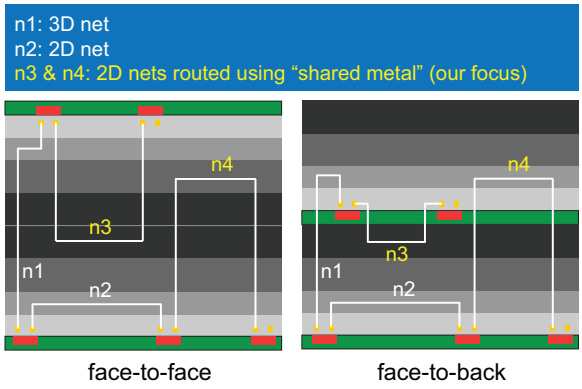


Figure 1: Routing layer sharing in face-to-face and face-to-back 3D ICs. Green portion represent the active FEOL layers, Gray represents the dielectric and various routing layers. The darker shade corresponding to higher thickness, pitch, and lower parasitic values of metal layers

Table 1: Inter-cell routing layer usage in OpenPiton 2D IC used as a reference. Wire segment is a single piece of metal routed only along a single direction.

Metal Layer	Number of nets	Avg. wire segment (µm)
M1	1400	0.77
M2	212900	0.44
M3	181100	1.27
M4	79200	1.96
M5	34800	4.87
M6	15000	10.90

In this paper, we analyze the different signal routing scenarios in 3D. Specifically, we—for the first time—report and analyze an interesting routing phenomenon in 3D called routing/metal layer sharing (Fig. 1). We analyze sharing under different 3D IC design scenarios with three general purpose processors as our test-benches. Compared to a default 3D BEOL stack, we show BEOL stack can be optimized to complete routing with 2 fewer metal layers. Simultaneously, enabling sharing with the shorter stack is useful to negate any loss of performance. Metal layer sharing is also useful in decreasing the routing congestion in higher metal layers. We observe up to 50% reduction in wirelength/track-usage of the top-most layer with sharing enabled. We also quantify the impact of metal layer sharing on the PPA, as we see up to a 16% performance drop in designs without sharing enabled. While the default metal layer sharing leads to significantly large number of 3D connections and routing overhead, we present a way to control the metal layer sharing without losing its timing benefit.

2 CHARACTERISTICS OF ROUTING

2.1 Routing in 2D ICs

In a traditional 2D IC, the BEOL consists of multiple metal layers on top of each other in a dielectric medium. The layers are generally monotonic in terms of their default pitch, unit RC parasitics of the layers. The short nets connecting cells close to each other are generally routed on the bottom layers that have the most routing tracks possible. Longer nets are preferably routed on higher metal layers that have a larger pitch and smaller RCs for better timing.

Table 1 shows the number of nets per layer and the average length of wire routed per metal layer for OpenPiton processor design in 2D. Note that each net consists of multiple wires and on various layers. This trend shows that bottom layers have more nets but with short wires, and higher layers have fewer nets with long wires. The majority of M1 is used for routing within the standard cells and does not have enough tracks for signal routing of the full chip design. Therefore, the BEOL of a 2D IC usually start with thin high resistance metal layers near the FEOL and the parasitics monotonically decrease higher up in the BEOL stack.

2.2 Routing in 3D ICs

The 3D BEOL stack does not have a monotonic progression of metal layers (Fig. 1) like 2D and can affect the routing quality. In a 3D IC, the underlying principles of routing remains unchanged, but the presence of multiple FEOLs and BEOLs creates novel routing scenarios. In this work, we separate the 3D routing into:

1. 3D nets: These connect cells from different tiers
2. 2D nets† without routing layer sharing
3. 2D nets† with routing layer sharing

†2D nets are the nets that connect to cells within a single tier. 3D nets and 2D nets with metal layer sharing use the 3D interface layer and so their routing behavior depends on the 3D stacks.

2.2.1 Metal Sharing in Face-to-Face Bonding. By bonding the top most metal layers of the two tiers in 3D IC, spacing between the two FEOLs become significant. The 3D net ($n1$ in Fig. 1) connecting the cells from two tiers must pass through the entire 3D routing stack adding "3D overhead". The shared 2D nets have to go through the BEOL stack of its tier to access metal layers of different tier. This decreases some of the benefits of metal layer sharing to improve timing and congestion. For example, the shared 2D net $n3$ goes through the BEOL stack of its own tier and only then can it borrow the metal routing resources of another tier. This doesn't add much benefit to such shared 2D nets except when there is significant routing congestion in one of the tiers. In the scenario depicted in Fig. 1 $n3$, $n4$ are shared 2D nets of different tiers. As the 3D BEOL stack in face-to-face bonding is symmetrical, the nets of type $n3$, $n4$ behave in a similar fashion.

2.2.2 Metal Sharing in Face-to-Back Bonding. In this stacking style, the top tier is placed back down on to the top face of the bottom tier. As discussed earlier in Section 1, 3D vias in this configuration block some of the active area in top tier FEOL creating placement obstacles. In addition, as the FEOL layers are closer than the Face To Face stacking, 3D nets have a smaller overhead to connect to cells from the two tiers. Unlike the default 2D net such as $n2$, the shared 2D nets ($n3$, $n4$) now have different routing behaviors as the 3D stack is no longer symmetrical.

Net $n3$ connecting the top FEOL cells has an easier access to the bottom BEOL's top-most metal layers. Therefore, by using the closer BEOL of bottom tier rather than its own BEOL. This frees up top tier BEOL for more timing critical or congested nets. The same cannot be said of the shared 2D net $n4$, as it has to go through the entire 3D BEOL, to access the higher metal layers of top tier.

Routing similar to net $n4$ is therefore not very useful to reduce overall design congestion. The net $n4$ in Face-To-Back is also worse than its Face-To-Face counterpart as now the net has to go through the FEOL layer. So, $n4$ in Face-To-Back will have even higher detour to find a free space in the top FEOL not blocked by cells.

3 EXPERIMENTAL SETUP

While 3D nets are a necessary part of routing completion in a 3D IC, the shared 2D nets are not mandatory and behave differently than the default 2D nets. In order to quantify their impact on overall design, we first analyze the different 3D scenarios and the characteristics of shared 2D nets compared to other nets. We then analyze different partition types that have a very asymmetric partitioning and metal layer usage between the two tiers, and the routing sharing in these cases. Macro-3D [8] is used for Logic on Memory partitioning, and Pin-3D [9] for Logic on Logic partitioning.

3.1 Benchmark and Cell Library

Three different CPUs are used to study the impact of routing layer sharing: Openpiton [10], and two *large-scale commercial industry cores* referred to as Industry-A, Industry-B to hide sensitive information.[1] All of these circuits are general purpose CPUs. The OpenPiton design is a single core processor with 256 kB of L3 cache. Industry-A is a dual core processor with 512 kB of shared L2 cache, and Industry-B design is a single core processor with 1 MB L2 cache.

In our experiments, we use a commercial 28 nm PDK/library. The 3D via or the Monolithic Inter-tier Via (MIV) pitch is fixed to 0.1 µm to all Face-To-Back designs corresponding to Monolithic 3D IC process. 1.0 µm pitch is used for Face-To-Face designs inline with the hybrid bonding process. The default/baseline BEOL stack is 6 metal layers per tier. All the Place and Route was done with Innovus v20.15 with in-house scripts to support 3D design.

3.2 Controlling the 3D Routing

One of the issues we tackle is controlling the number of 3D connections in the 3D design. By default, the commercial EDA tool uses an extremely high number of 3D connections. In order to restrict this, we add individual net attributes that constrain the routing engine. Longer nets are selectively allowed to use the memory tier layers, while all other nets are discouraged to use the memory tier. This is done by setting the preferred routing layer range of the 2D nets along with the effort level. With this added limitation, the tool stops using a large number of MIVs. By varying the threshold for long nets, we can achieve different ranges of MIV/3D via count. MIVs are also added to the final set of nets added in the post-route optimization. Since these nets are added in the last stage of the design process, they won't have the net attributes that restrict routing. The criteria for selection can also be changed from net length to timing criticality or other design metrics.

4 RESULTS AND ANALYSIS

4.1 Which Die Bonding Styles Benefit More?

In this section, metal layer sharing is compared between the two bonding styles using the OpenPiton design. The RTL is partitioned in Logic-On-Memory fashion and the final layouts are shown in Fig. 3 (a), (b). The L3 data cache and tag memories are assigned to the bottom (memory) tier of the 3D IC.

[1]We are not able to reveal the names due to NDA.

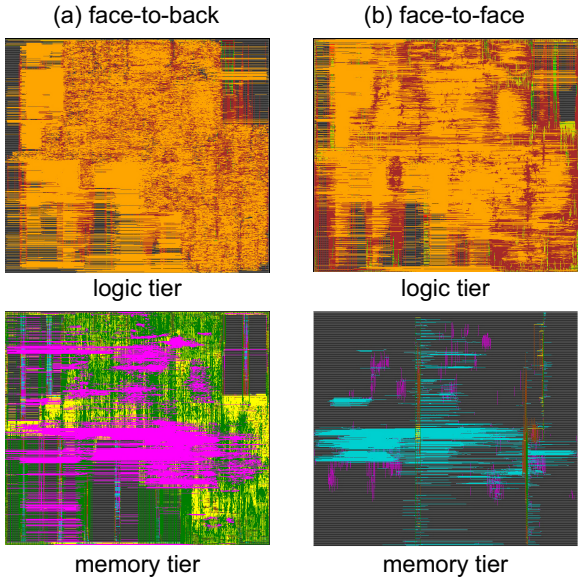


Figure 2: Comparing the routing in two different bonding styles of Logic-On-Memory 3D ICs. (a) Face-To-Back, (b) Face-To-Face. The logic tier BEOL layouts are on the top, and memory tier BEOL layouts the bottom. Each color corresponds to a routing layer.

Table 2: Metal layer sharing in different 3D bonding styles using OpenPiton RTL. # MIVs on 2D nets shows the amount of metal layer sharing.

	Units	Face-To-Back	Face-To-Face
Frequency	MHz	1400	1400
Chip Area	mm^2	0.638	0.638
# MIVs	–	120,351	3,112
# MIVs on 2D nets	–	119,317	2293
# MIVs on 3D nets	–	1034	819
Wirelength	m	6.36	5.81
Worst Neg Slack	ns	-0.384	-0.438
Effective Frequency	MHz	910.5	867.8
Total Neg Slack	ns	-864.5	-540.2
Total Power	mW	414.6	411.2

Table 2 shows the various design metrics for the two bonding styles implemented using the Macro-3D flow [8]. The number of MIVs in each scenario is further split into MIVs on 2D and 3D nets. The number of MIVs on 3D nets depends on the partitioning and are very similar in the two scenarios. The small difference comes from 3D nets that route using multiple MIVs. In Face-To-Back, a lot of the 2D nets use metal layer sharing as seen from the $\sim 120k$ MIVs in this scenario. In Face-To-Face, only $\sim 2.2k$ MIVs are used for metal layer sharing due to the 3D BEOL stack difference and the higher contact pitch of the 3D vias.

This difference is also seen in the routing layouts of logic and memory tiers in Fig. 2. Memory tier of the Face-To-Back style has a much higher metal layer usage. The memory tier routing is also sparser compared to logic tier as only a few nets are connected to these macros. In Face-To-Back bonding style, metal layer sharing is encouraged and so the memory tier layers are well utilized.

The increased metal layer sharing in Face-To-Back causes the higher total wirelength due to the '3D overhead' (more in Section 4.4). This increased wirelength is also the reason for the worse effective frequency. The limited sharing in Face-To-Face bonding coupled with the higher '3D overhead' makes metal layer sharing not as effective in the Face-To-Face orientation.

4.2 Which Tier Partitioning Benefit More?

Fig. 2(b) with the asymmetrical logic-on-memory partitioning of the RTL leads to an uncongested bottom tier that encourages metal layer sharing by the 2D nets connecting standard cells in top tier. To quantify this effect, two different partitioning options are considered as shown in Fig. 3. The top row corresponds to the Logic-On-Memory partitioning of OpenPiton discussed in Section 4.1. On the bottom are the layouts for a more symmetrical Logic-on-Logic partitioning. Due to the additional complexities in such partitioning, it is implemented using the Pin-3D flow [9]. Only Face-To-Back bonding styles are considered and BEOL is comprised of 6 metal layers per tier.

Table 3 compares the two partitioning implementations of Face-To-Back OpenPiton at 1400 MHz. The Logic-On-Logic design has a smaller footprint due to the flexibility of the partitioning. In Logic-On-Memory design with Macro-3D [8], the memory tier is only made of macros, which makes for an inefficient use of footprint if the macros don't fit well together. In the Logic-On-Logic design, the standard cells can fill in the white-spaces efficiently allowing for a good usage of the silicon area. Note that Pin-3D cannot be used in partitions with a very asymmetric memory placement. Pin-3D flow builds up on Compact-2D [11] which breaks down with a highly asymmetric partitioning.

The total number of MIVs are similar in both cases at $\sim 100k$, but the breakdown into MIVs on 2D and 3D nets shows that Logic-On-Memory partitioning have significantly more MIVs used for shared 2D nets. In Logic-On-Logic, the standard cell partitioning leads to a significantly large cut-size and therefore has a lot of 3D nets in the design. The presence of standard cells on both tiers also increases the local usage of metal layers within BEOL of each tier. Even though the total routed wirelength in Logic-On-Logic case is significantly smaller, the free tracks are not easily available for metal layer sharing due to the symmetrical layout. The more uncongested tracks of M5, M6 are located above the memories which are stacked on top of each other in this configuration. This makes these tracks hard to access by the logic cells. As almost all the nets are connected to the logic cells, most of the routing is made above them. Further, the # of MIVs on nets that borrow routing tracks from bottom tier (like net *n*3 in Fig. 1), and the # of MIVs that help borrow routing tracks from top tier (like net n4 in Fig. 1) shows almost all the metal layer sharing is of type *n*3.

Most (> 90%) of the routing in bottom BEOL M4–M6 of Logic-on-Memory partitioning is occupied by the shared 2D nets. This decreases as we go farther away from the top/logic tier FEOL. A similar trend is seen with Logic-On-Logic albeit with a much smaller percentage of wirelength (<30%) used for shared net routing. The routing results of memory tier M6 M5 in Fig. 4 depict the metal layer sharing in the layouts.

Table 3: Metal layer sharing in 3D partitioning options: Logic+Memory, Logic+Logic. #MIVs on 2D nets shows the abundance of metal sharing in the designs.

	Units	Logic+Mem	Logic+Logic
Frequency	MHz	1400	1400
Chip Area	mm^2	0.638	0.603
# MIVs	–	121,714	104,606
# MIVs on 2D nets	–	119,317	17,575
# MIVs on 3D nets	–	1034	87,031
# MIVs on clk nets	–	1363	13,278
Borrow from bottom	–	119,317	17,421
Borrow from top	–	0	154
Wirelength	m	6.36	4.66
Shared Wirelength	%	25.1	6.4
Worst Negative Slack	ns	-0.384	-0.403
Effective Frequency	MHz	910.5	895.0
Total Negative Slack	nHz	-864.5	-631.6
Total Power	mW	414.6	378.3
% WL of shared nets in the memory tier			
M6	%	97.4	29.7
M5	%	95.5	18.1
M4	%	91.2	2.6
M3	%	36.0	0.1
M2	%	2.3	0.0
M1	%	0.0	0.0

Table 4: Design metrics of the three RTLs considered in our work. The designs are implemented in a Face-To-Back 3D fashion with 6 metal layers per tier. These are the baseline designs for further comparisons

	Units	Piton	Ind-A	Ind-B
Frequency	MHz	1400	1500	1375
Chip Area	mm^2	0.638	1.109	1.893
# Metal Layers	–	12	12	12
# MIVs	–	120351	247158	441365
# MIVs on 2D nets	–	119317	243373	439953
# MIVs on 3D nets	–	1034	3785	1412
# MIVs on clk nets	–	1363	28576	44859
Wirelength	m	6.36	12.25	19.03
Worst Neg Slack	ns	-0.384	-0.296	-0.323
Effective Frequency	MHz	910.5	1038.8	952.1
Total Neg Slack	ns	-864.5	-2838.2	-2028.3
Total Power	mW	414.6	862.0	979.1

4.2.1 Baseline Experiments. Table 4 shows the baseline PPA and MIV/3D via related metrics for the three RTLs considered (openpiton, Instry-A,B). These are useful to compare against the PPA under various scenarios (like changing metal layer stack, suppressed metal layer sharing, etc.).

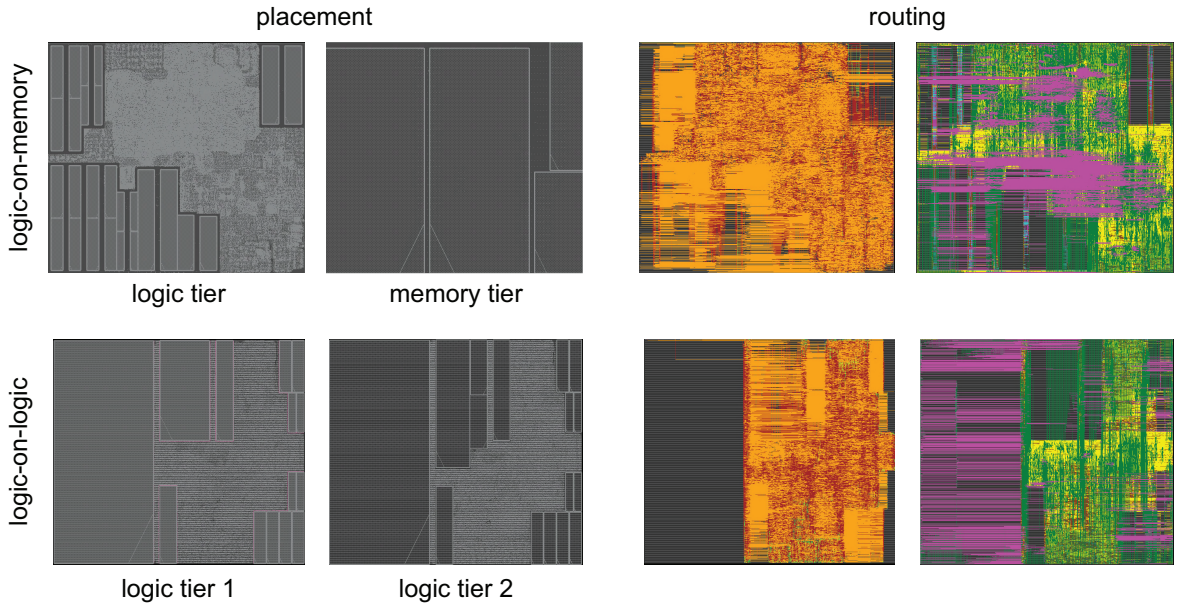


Figure 3: Comparing tier partitioning impact on routing in OpenPiton. The placement and routing layouts in the two tiers are provided for the two styles of partitioning. Memory tier and Logic tier 2 are the bottom FEOL in their corresponding designs.

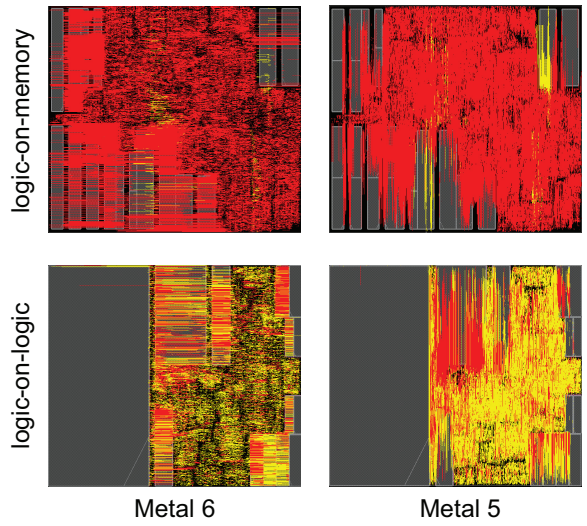


Figure 4: Routing in shared metal layers of 3D OpenPiton design with Face-To-Back bonding style. We show M5 and M6 of the memory tier and logic tier 2. Red are routing with metal sharing, and yellow is everything else.

4.3 Can We Save Metal Layers?

4.3.1 Removing Metals from the Logic Tier. Table 5 shows the PPA of the design with the top most metal layer removed. With one fewer metal layer to route, the routing layer sharing becomes more important and is reflected in the increased MIVs compared to sharing enabled design with all 12 metal layers. Design metrics such as wirelength, total power, and effective frequency change only by a small amount within 1-2% of the baseline. Total negative slack on the other hand degrades due to the lack of the top-most metal layer M6 in logic tier. In general this layer is used for most timing critical nets in logic tier and helps with overall TNS.

Table 5: Impact of metal layers on the overall design metrics. Only the percentage delta w.r.t to the baseline in Table 4 are reported. Negative implies the current design is worse compared to baseline. Frequency and footprint of the following are same as the corresponding design baselines.

	OpenPiton	Ind-A	Ind-B
Removed logic tier's M6 (top-most) metal layer			
# Metal Layers	11	11	11
# MIVs	119599	252016	449047
# MIVs on 2D nets	118611	248143	447617
# MIVs on 3D nets	988	3873	1430
# MIVs on clk nets	1289	28711	44509
% Wirelength	-0.8	-1.4	0.3
% Effective Frequency	0.8	-1.8	-0.7
% Total Neg Slack	-34.5	6.5	-21.8
% Total Power	-1.8	-1.3	-1.0
Removed memory tier's M5, M6 metal layers			
# Metal Layers	10	10	10
# MIVs	11378	14380	3591
# MIVs on 2D nets	10436	10690	2023
# MIVs on 3D nets	942	3690	1568
# MIVs on clk nets	1624	1213	252
% Wirelength	8.8	6.7	3.4
% Effective Frequency	-2.2	1.2	-1.2
% Total Neg Slack	21.5	24.5	-4.6
% Total Power	-1.8	0.6	-1.9

Table 6: Impact of metal layer sharing on overall PPA in OpenPiton, Industry-A, and Industry-B. The percentage values are w.r.t to the baseline in Table 4. Negative implies the current design is worse compared to baseline. Frequency and footprint same as baseline.

	OpenPiton	Ind-A	Ind-B
# MIVs	1005	3732	1796
# MIVs on 2D signal nets	0	0	0
# MIVs on 3D signal nets	1000	3607	1600
# MIVs on clk nets	5	125	196
% Wirelength	9.4	4.1	1.1
% Effective Frequency	-2.2	-16.0	-9.0
% Total Neg Slack	27.3	39.1	1.9
% Total Power	0.8	-1.5	-5.1

4.3.2 Removing Metals from the Memory Tier. In Logic-On-Memory partitioning, the memory tier is only comprised of macros and the routing demand (of the inter-block routing of memory tier) is very close to zero. In all our logic-on-memory designs, the # of nets connecting to and from the memory blocks is $< 1\%$ of all the nets. Routing is also present inside the memory blocks and is restricted to metal layers 1–4 by design. Therefore, metal layers 5–6 in the memory tier do not provide a significant benefit to complete routing. And by removing one or more metal layers in the 3D stack, we can save metal layer cost by sacrificing some of the benefits from metal layer sharing. As all the memory macros in the designs, have routing in layers M1–M4, these layers cannot be considered for removal.

Table 5 shows the PPA impact with the metal layers M5, M6 in memory tier removed. With a significant reduction in available number of tracks in the memory tier, the amount of routing layer sharing is severely limited. The number of MIVs on the 3D nets decrease from $\approx 10^5$ in Table 4 with 6 metal layers in the memory tier to $\approx 10^3 - 10^5$ depending on the design. Here, we see that using metal layer sharing (albeit much limited), we were able to remove 2 metal layers with negligible impact to critical paths, total power. The reduced metal layer sharing actually lead to a positive benefit to the wirelength and the total slack due to the reduction in routing detours using 3D vias. In the next section, we try to quantify the exact benefit of the metal layer sharing.

Overall, by partitioning the large macros to the memory die in 3D, we were able to independently control the metal layers M5, M6 in this tier without impacting the overall PPA. In a 2D footprint, the maximum number of layers required is constrained by the logic cell placement and the congestion in the region. This leads to very low metal layer routing in the regions above large memory macros, and a higher routing density over the standard cells in 2D. Finally, we were able to see that removing the memory tier layers actually has a positive impact on the PPA as well as the cost.

4.4 How Do We Control Sharing?

In this section, we isolate and analyze the effect of metal layer sharing by completely blocking shared routing in the memory tier. The MIV breakdown in Table 6 shows that all the MIVs are either on 3D nets or clock nets, effectively restricting the metal layer sharing. The implementations without metal layer sharing have a better

Table 7: Analysis of the designs under limited metal layer sharing. The percentage values are w.r.t to the baseline in Table 4. Negative implies the current design is worse compared to baseline. Frequency and footprint same as baseline.

	Piton	Ind-A	Ind-B
# MIVs Before Control	120351	247158	441365
# MIVs After Control	2407	128417	40981
# MIVs on 2D signal nets	713	124683	41
# MIVs on 3D signal nets	1019	3734	1393
# MIVs on clk nets	675	24553	39547
% Wirelength	9.4	2.3	0.8
% Effective Frequency	0.1	6.8	0.5
% Total Neg Slack	26.3	45.6	1.1
% Total Power	1.5	-0.3	-4.0

overall routed wirelength (by $\approx 1 - 9\%$ based on the design). In full layer sharing, MIVs required for accessing borrowed layers can only be placed where there are no top-tier FEOL cells. So these nets have detours increasing the total wirelength. As seen in previous cases as in Section 4.3, lower wirelength results in better overall Total Slack even in sharing restricted case. But blocking this sharing, negatively impact the critical timing path with a $2 - 16\%$ reduction in the effective frequency compared to full sharing.

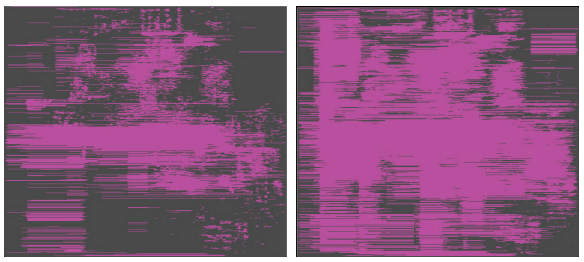
Overall, we see that fully blocking signal sharing can significantly affect the effective frequency of the design. But as we have seen in Table 5, having a very limited sharing can give us most of the effective frequency benefits of the full sharing in Table 4. In order to explore the cause of these benefits and to control the high MIV counts in Table 6 and Table 4, we restrict the metal layer sharing.

By limiting the routing layer ranges of the 2D nets as using the method in Section 3.2, the overall number of MIVs have been greatly reduced from the complete routing as seen in Table 7. With this method, the MIVs can still be added on the clock nets, nets added at the post-route optimization stage, or long nets $> 500\,\mu m$ in length. In Table 7, Industry-A design has the least reduction in MIV count ($\sim 0.5\times$) after the routing control. This is still a very high MIV count, most of which ($108k$ out of $125k$) are added in the final stages of the design. Out of the $100k$ nets added in the post-route optimization of Industry-A, $38k$ nets have metal layer sharing and use $108k$ total MIVs.

The reduction in MIV count for OpenPiton, Industry-B do not negatively impact the overall PPA by an appreciable margin. This tells us that most of the timing benefit is due to a few nets in the design that undergo 3D routing, and most likely the clock design with metal layer sharing. The lower MIV count hasn't negatively impacted the effective frequency of the designs. Note that the effective frequency improved by 6% for Industry-A with MIV count $0.5\times$ the baseline. This shows us that the default routing behavior of Innovus inserts more than the necessary amount of 2D vias for best timing, and MIV control or the limited metal stack in 3D as in Table 5 is a better way to perform routing in 3D.

4.5 What About Routing Congestion?

Table 8 shows the amount of routing per metal layer in the OpenPiton design in the two extreme cases of metal sharing: Case 1.



(a) Logic M6 with metal layer sharing

(b) Logic M6 without metal layer sharing

Figure 5: Routing on Logic tier's Metal Layer M6 under the two extreme cases of metal layer sharing

Table 8: Routing Analysis of Openpiton with and without the layer sharing. All units are in mm**.**

	Memory Tier		Logic Tier	
Layer	sharing	w/o sharing	sharing	w/o sharing
M1	0.0	0.0	117.5	2.4
M2	3.2	1.0	625.7	548.4
M3	42.3	14.8	1416.1	1298.2
M4	33.6	2.2	1550.8	1274.4
M5	1031.8	55.8	658.9	1516.0
M6	925.7	27.2	0.0	1018.0

Complete sharing with one metal layer removed in logic tier corresponding to Table 5, Case 2: No sharing allowed as in Table 6. The main differences between the two routing behaviors are in the memory tier, and the metal M1, M5, M6 of the logic tier. Fig. 5 shows the usage of top most metal layer M6 in logic tier with metal layer sharing turned on, and so removing it has a minimal impact

In order to accommodate for the frequent access of the bottom BEOL, the routing in the top/logic tier's local metal layers (M1, M2) is increased. As the MIVs need to be placed in a legal location away from the standard cells, the metal layer M1 would have to accommodate for the added detours on shared nets to pass through the top FEOL. Because of this increased routing in local layers, non critical timing nets are most suitable to use the bottom BEOL. The reduction in the routing on global metals of the top BEOL is exploited to save metal layer cost of the die. This increased routing on logic M1 under sharing is still a very small amount compared to overall design routing.

Under metal layer sharing, we see that M5, M6 of the memory tier share a significant load of the total routing. After removing the M6 layer which has preferred direction along horizontal layer, the routing load on the next horizontal layer M4 increases. After this increase, M4 has now similar amount of routing as the M5 logic in the no sharing case. M5 logic still has only 0.4× the routing/track usage of the no sharing case.

4.6 How Are Individual Nets Affected?

Histograms in Fig. 6 show how average behavior of nets vary among the three net categories: All the nets in the design, shared 2D nets, 3D Nets in the baseline designs. These comparisons can provide insight into the routing and timing behavior of these net categories.

The 'Avg. Bounding Box' metric is the half-perimeter bounding box length of the net. This is mainly dependent on the partitioning and placement. This metric is reasonably independent of physical routing problems such as net congestion, 3D overhead, and routing detour of shared nets in finding legal MIV locations. The first histogram (Fig. 6) shows this trend for our net categories. Shared 2D nets have a longer bounding box than an average net in the design implying that the longer nets are most likely to borrow layers from the other tier to reduce congestion of its own tier. This is one of the reasons for choosing the wirelength as a threshold for allowing MIVs in Section. 3.2.

3D nets have an even longer bounding box as they are mostly connected to memories. The macros are very large in size and pins are spread along the edge. This makes it harder for a cells within a certain hierarchy to connect to the memory pins, and the cells would end up farther from the macro pins. Therefore, the logic on memory partitioning option is a main reason for such large bounding box.

For the next figure in Fig. 6, we look at the detour of the net routing from ideal bounding box length. The detour is defined as the % difference between routed wirelength and the ideal shortest net. To illustrate this, Fig. 7 shows two examples of net routing and the detour calculation. The routing is determined by a lot of factors such as congestion in the design, routing blockages, fanout structure of nets, connected cell strengths, timing criticality of the net, etc. For instance, it is sometimes beneficial to use a longer routing to reduce load at some critical cells, and in other cases it is better to take optimal detours to clear tracks for other critical nets. In the case of 3D routing, the detours also occur due to the additional interaction between the MIVs and the standard cells of top tier FEOL.

A large positive value in the detour % shows that net goes back and forth during routing. The detour % is significantly lower for 3D nets due to the large bounding box values. Shared 2D nets, with almost a 2× larger bounding box length compared to an average net still has higher detours. As a shared 2D net need to have at least 2 MIVs (one to access bottom routing layers through the top FEOL, one to come back to its own BEOL containing sink cell), the overhead in finding MIVs is higher than the 3D nets which only require 1 MIV per net (going from one tier to the other). In the three designs here, shared 2D nets have an average of 2.8 – 3.2 MIVs per net while a 3D net has 1.1 – 1.3 MIVs on average. On average, the shared 2D nets have 2.5× more MIVs per net than a 3D net in the design which causes the increased detours.

If these routing detours were not properly considered during timing optimization, the 2D shared nets would be expected to have worse timing than the design average. But the timing histograms of Fig. 6 show that both design average and the shared 2D nets have a similar average negative timing slack. This implies that, on average, the shared 2D nets do not create additional timing bottlenecks. 3D nets again show a worse average timing slack due to the type of partitioning. In smaller designs like OpenPiton and Industry-A, the large memory nets could contribute to a significant delay. But in larger designs, many other long nets exists due to the sheer size of the floorplan and the impact of memory nets wouldn't be as much. The type of memory would also be an important factor here as slack is path attribute and is not specific to a single net.

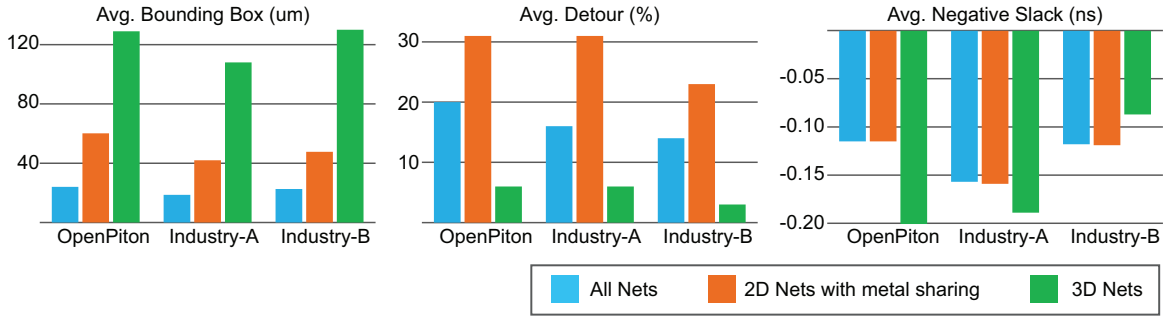


Figure 6: Histograms showing various timing and routing characteristics for shared 2D nets, 3D nets in the three CPU designs considered. The title of each is the Y-axis quantity and unit. Detour is a simple estimate from the bounding box and actual routed wirelength

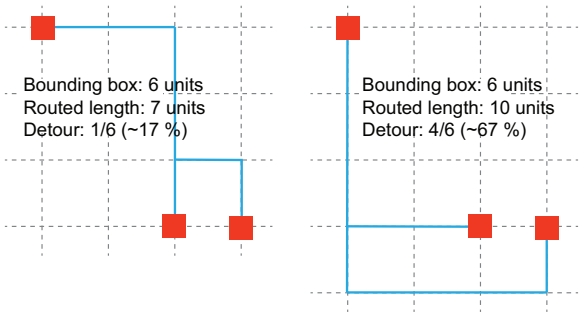


Figure 7: Two routing scenarios with the detour % calculations. Detours can happen when routing tracks are blocked, or timing-driven routing opts for a longer path to reduce overall delays of the cells on the net.

If the memory delay is a bottleneck in the path timing, the nets connecting memories (predominantly 3D nets) would also suffer from higher timing slack.

4.7 Takeaway

By analyzing the different 3D design scenarios, we see that Face-To-Back has the most metal layer sharing. While a symmetric partition can lead to large cut-size and lot of 3D connections, the metal layer sharing can be quite limited as seen in Logic-On-Logic partitioning. A highly asymmetric partitioning with low congested metal layers closer to the FEOL encourages the metal layer sharing significantly. Introducing metal layer sharing counter-intuitively increases the overall wirelength due to the non-overlap constraints between MIVs and standard cells of top tier FEOL in Face-To-Back bonding.

By analyzing the shared nets, we see that long nets are likely to borrow metal layers and have higher than average detour. The timing driven routing optimally route nets so that the extra detour doesn't negatively impact overall timing. We also saw that critical nets are not likely to be using metal layer sharing. Rather, the metal layer sharing reduces the congestion in the BEOL which is exploited by the timing critical cells for better routing. We also saw that the asymmetric partitioning helps with cost saving in the BEOL. Finally, we were able to show that while metal layer sharing is important, most of the PPA benefits can be obtained from just the clock nets and a few 2D nets undergoing with routing sharing.

5 CONCLUSION

In summary, we have analyzed 3D routing and the ways it differs from a traditional 2D routing. In the three circuits considered, disabling the metal layer sharing decreased the maximum frequency by up to 16% which makes it an important part of 3D design. Finally, the reduction of congestion in top tier FEOL's top most routing layers allows for us to drop 2 metal layers without significant cost to power, performance especially for large designs. Further work is required to maximize the impact of sharing by fine-tuning the routing algorithms for the 3D BEOL stack.

ACKNOWLEDGMENTS

This research is funded by the DARPA ERI 3DSOC Program under Award HR001118C0096, the Semiconductor Research Corporation under GRC Task 2929, and the National Research Foundation of Korea under NRF-2020M3F3A2A02082445.

REFERENCES

[1] International Roadmap For Devices and Systems, 2018.
[2] Eric Beyne. The 3-d interconnect technology landscape. IEEE Design and Test, 2016.
[3] Jingwei Lu, Hao Zhuang, Ilgweon Kang, Pengwen Chen, and Chung-Kuan Cheng. Eplace-3d: Electrostatics based placement for 3d-ics. In Proceedings of the 2016 on International Symposium on Physical Design, 2016.
[4] M. Hsu, V. Balabanov, and Y. Chang. Tsv-aware analytical placement for 3-d ic designs based on a novel weighted-average wirelength model. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2013.
[5] B. Goplen and S. Sapatnekar. Efficient thermal placement of standard cells in 3d ics using a force directed approach. In International Conference on Computer Aided Design, 2003.
[6] S. Panth, K. Samadi, Y. Du, and S. K. Lim. Placement-Driven Partitioning for Congestion Mitigation in Monolithic 3D IC Designs. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2015.
[7] Yi-Chen Lu et al. Tp-gnn: a graph neural network framework for tier partitioning in monolithic 3d ics. In 2020 57th ACM/IEEE Design Automation Conference (DAC), 2020.
[8] Lennart Bamberg et al. Macro-3D: A Physical Design Methodology for Face-to-Face-Stacked Heterogeneous 3D ICs. DATE, 2020.
[9] Pentapati et al. Pin-3D: A Physical Synthesis and Post-Layout Optimization Flow for Heterogeneous Monolithic 3D ICs. In 2020 IEEE/ACM International Conference On Computer Aided Design (ICCAD), 2020.
[10] Jonathan Balkind et al. Openpiton: An open source manycore research framework. In Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems, 2016.
[11] B. W. Ku, K. Chang, and S. K. Lim. Compact-2d: A physical design methodology to build two-tier gate-level 3d ics. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2019.