Metal Layer Sharing: A Routing Optimization Technique for Monolithic 3D ICs

Sai Pentapati and Sung Kyu Lim, *Senior Member, IEEE*

***Abstract*— As Moore's law with traditional process node scaling is slowing down, other techniques are required for the advancement of process nodes. In this work, we focus on one such alternative: 3-D physical design of integrated circuits (ICs). While many recent studies have shown the benefits of 3-D IC design on timing and power consumption of circuits, routing in 3-D is solely done with the automatic commercial routers and has not been well studied. In this article, we discuss the various routing scenarios that arise from cell partitioning and the metal layer stack in 3-D. Unlike a 2-D IC, the metal layer configuration in 3-D depends on the orientation in which the dies are bonded together. Due to this, depending on the configuration, cells in one tier tend to use routing layers from the other tier. This is referred to as metal layer (or) routing sharing. This depends on the metal layer stack and the cell partitioning in 3-D, as well as the via pitch used for 3-D connections. By analyzing metal layer sharing in detail, we see that it can help reduce metal layer costs in 3-D as well as improve the power consumption and, in some cases, the maximum achievable performance of the circuits. Overall, the 3-D metal layer cost can decrease by 9% along with an improved power delay product of up to 7.5% just from the routing sharing in monolithic 3-D ICs.**

***Index Terms*— 3-D integrated circuits (3-D ICs), interconnect analysis, metal layer sharing, monolithic 3-D.**

I. Introduction

SINCE its advent, the advancement of Moore's law has mostly been possible due to the improvements to transistor technology with the help of semiconductor physics and the improvements to the manufacturing abilities that make such transistors realizable. The physical scaling that accompanies and drives such improvements cannot be extended endlessly and several new methods are proposed to go "Beyond Moore's law" [1] by the IRDS. The 3-D manufacturing technology is one such method proposed to continue the trend of creating circuits that are ever more powerful. This represents a shift in focus from device physics to the physical design of integrated circuits (ICs) where the 3-D design is applied. Depending on the type of 3-D integration, the change to the manufacturing flow is either at packaging or fabrication stage.

Manuscript received 15 February 2022; revised 9 May 2022; accepted 26 May 2022. Date of publication 13 June 2022; date of current version 1 September 2022. This work was supported by the DARPA ERI 3DSOC Program under Award HR001118C0096 and in part by the Semiconductor Research Corporation under GRC Task 2929. *(Corresponding author: Sai Pentapati.)*

The authors are with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: sai.pentapati@gatech.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TVLSI.2022.3181185.

Digital Object Identifier 10.1109/TVLSI.2022.3181185

The 3-D integration has already entered commercial production up to a certain extent with the advent of consumer-grade electronics using 2.5-D ICs, where various chiplets are placed and connected through an interposer. Memory packaging has been another area, which already utilizes 3-D packaging using through-silicon vias (TSVs) or micro-bump-based bonding. Intel [2] and AMD [3] recently also announced 3-D processors with micro-bumping and hybrid-bonding designs showing the growing trend of 3-D ICs with ASIC design.

While both Intel and AMD's implementation of the 3-D ICs follow a coarse partitioning with relatively few number of 3-D connections, various types of partitioning are possible with 3-D based on the 3-D via pitch. The maximum 3-D connections and the 3-D pitch depend on the bonding types: monolithic 3-D ICs, hybrid bonded 3-D IC, and micro-bump 3-D IC [4]. With micro-bumping, two known good dies are bonded together using large micro-bumps 10–100 μm. The large pitch and high parasitics of the micro-bumps limit their use in high-performance designs requiring large 3-D bandwidth.

Hybrid bonding designs have 3-D pitch values around 1–10 μm, which increases the allowed 3-D bandwidth by around 100$\times$ compared to micro-bumps. The fine pitch for hybrid bonding at wafer level is still hard to accomplish with current technology processes and not many foundries offer this at present. However, more research is being done into enabling sub-micrometer pitch for hybrid bonding [5], which can potentially bring this to consumer electronics in the near future.

Finally, monolithic 3-D IC design is the most advanced type of fabrication for 3-D ICs. In this, the different tiers of the 3-D IC are sequentially fabricated on top of each other. The resulting lithographic alignment of vias can achieve a pitch of $\sim$0.1 μm and can roughly scale with the interconnect technology size. The process of fabricating dies on top of each other is extremely challenging with many limitations related to the thermal budget and the materials that could be used in the fabrication. In recent years, CEA-LETI showcased a significant breakthrough in low-temperature fabrication of the devices that shows potential for monolithic 3-D IC to lead the "more than Moore" era of computing [6]. The resulting nanometer-scale 3-D via pitch can unlock a variety of 3-D partitioning and routing scenarios.

While different bonding techniques determine the number of 3-D connections and the partitioning type, the orientation of the 3-D ICs also alters the overall 3-D routing. The two different stacking orientations are: face-to-face (F2F) and face-to-back (F2B), whose cross sections are shown in Fig. 1. The micro-bump-based routing in Fig. 1 shows the result

1063-8210 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on December 11,2022 at 00:39:06 UTC from IEEE Xplore. Restrictions apply.

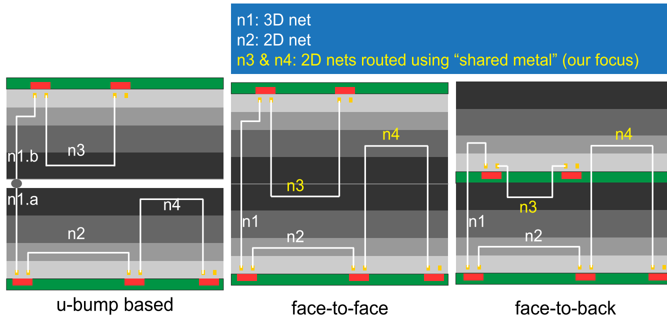

Fig. 1. Routing layer sharing in F2F and F2B 3-D ICs. Green portion represents the active FEOL layers, gray represents the dielectric and various routing layers. The darker shade corresponds to higher thickness, pitch, and lower parasitic values of metal layers.

of routing the two dies independently. The routing within each tier would be similar to a 2-D IC routing, and only the 3-D nets will have a different kind of routing, which would pass through a micro-bump as shown. With a complete 3-D routing, as depicted in the case of F2B and F2F designs, the nets can be routed freely across the entire metal stack resulting in nets such as $n3$ and $n4$ that borrow routing tracks from different tiers to be fully routed. The main focus of this article is to explore the usefulness of such metal layer sharing.

A. Literature Review and Motivation

Routing in 3-D ICs has been a focus in recent literature such as [7]–[9], but they focus on reducing the routability issues in 3-D and are not applicable for our work. We use state-of-the-art 3-D flows such as [10] and [11] that effectively use commercial electronic design automation (EDA) tools and are not limited by congestion in 3-D.

Yan *et al.* [7] addressed the routability in 3-D but are limited to micro-bump-based 3-D ICs and routability in the redistribution layer (RDL). While this is very useful for micro-bumping 3-D IC, which is commercially most relevant, the results cannot be transferred to 3-D designs with high bump density where the integration is done at a wafer level and do not require RDLs. In such cases, compared to commercial router, any manual bump-assignment would lack with respect to the quality of routing under many design constraints, such as timing, power, and congestion. This is why most of the monolithic 3-D and hybrid bonding 3-D design flows use a commercial router for monolithic intertier via (MIV) placement [10]–[13].

Panth *et al.* [8] proposed two different partitioning models for 3-D ICs: one based on placement and another based on routing. The placement-driven partitioning is useful to achieve a fewer number of 3-D connections but can cause routability issues due to clusters of cells. On the other hand, the routing-driven placement is well-suited for decreasing congestion at the cost of increased connection density. While these were useful in the older design flows for 3-D ICs, Pentapati *et al.* [11] proposed a routing and timing-driven placement legalization removing the need for the approximate routing estimation(s) and instead utilized the commercial router's congestion modeling.

In [9], the routability in transistor-level 3-D is shown to be poor due to its high cell density and is targeted to improve the congestion using changes to the 3-D cell structure. At gate-level 3-D, we do not see such an issue since the cells are arranged in a 3-D fashion, and the pin layers are distributed across both the layers unlike in transistor-level 3-D IC design. As such, improving the routability further is not necessarily useful in gate-level 3-D ICs and the current work is mostly focused on different types of routing in monolithic and hybrid bonding 3-D ICs under different orientations.

For the first time, we analyze a type of routing that is specific to 3-D ICs—metal layer sharing. We investigate the effect of different 3-D arrangements on metal layer sharing and how it can be leveraged to efficiently use the metal layers in the 3-D stack. We show that, with metal layer sharing, an entire metal layer can be dropped from the routing stack without negatively affecting the maximum performance or power efficiency of the design. We also see how this phenomenon creates better 3-D designs with up to 8% higher power efficiency. Finally, we analyze the routing of the 3-D ICs with respect to congestion and design rule violations (DRVs) due to the metal layer sharing.

This article is organized as follows.

1) In Section II, we discuss the different effects of routing in 2-D and 3-D ICs of different configurations.
2) In Section III, we present our setup and methodology to analyze the effect of metal layer sharing. In Section IV, we quantitatively discuss metal layer sharing in various scenarios of 3-D ICs with place and route (PnR) simulations of a processor design.
3) In Section V, we discuss several aspects of metal layer sharing in the monolithic 3-D IC design of three commercial processors.
4) Finally, Section VI concludes this article.

II. Characteristics of Routing

Here, we discuss and analyze the general routing characteristics of 2-D ICs and the various routing scenarios of 3-D ICs to understand the impact of the metal stack arrangement on 3-D IC routing.

A. 2-D IC Routing Characteristics

In a traditional 2-D IC, there is a single layer of front end of the line (FEOL) followed by multiple layers of the back end of the line (BEOL) for routing the cells together. Within cell routing is mostly limited to the metal layer closest to the FEOL (M1), while the next layer, M2, is only utilized in cells with more complex internal connectivity, such as flip-flops. Due to the close proximity to standard cells, M2 is used effectively to route short wires rather than longer wires. This is because long wires block a contiguous portion of the available routing tracks that can block M2 tracks over several cells. All the nets have some routing on M2 (sometimes M1 is used based on the cell placement) for pin access making this a highly utilized layer.

As we go further in the metal layer stack from M3 to M6, we see that the number of nets on the layer decreases, while

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on December 11,2022 at 00:39:06 UTC from IEEE Xplore. Restrictions apply.

TABLE I

INTERCELL ROUTING LAYER USAGE ANALYZED BASED ON OUR OPENPITON 2-D IC IMPLEMENTATION. A WIRE SEGMENT IS A SINGLE CONTINUOUS PIECE OF METAL ROUTED IN A STRAIGHT LINE

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

the average length of uninterrupted wire segment routed on the layer increases. These two phenomena are closely related to each other. As more nets are routed on a layer, it is effective to use it for shorter wire segments while letting longer portions of the nets route on higher layers. In addition, as we go up the stack, most of the nets would have sufficient routing tracks available to complete routing without needing to route further up the stack, leaving only a few long nets as we reach M6. This wire distribution behavior of 2-D ICs is shown in Table I.

The metal layers are also engineered in a way to consider the wirelength trends and to reduce the wire load. One such technique is to gradually increase the metal layer width and the pitch as we move higher up the stack. In conjunction with the metal layer design, the surrounding dielectric medium and the thickness are also increased with the pitch to reduce the parasitics of the wire. The smaller width of the lower metal layers accommodates more nets closer to the routing stack. Also, as the metal layer at the top does not route many nets, it can accommodate a larger pitch. This reduces the wire resistance, and the increased spacing helps to limit any coupling parasitics for the long wires to keep the wire delay in check.

B. 3-D IC Routing Characteristics

Fig. 1 show the example cross sections of the two 3-D IC orientations along with their BEOL. The BEOL of each tier within the 3-D IC is colored to reflect the fact that the parasitics, average wire segment length decreases monotonically along the stack. By joining two different BEOL stacks in 3-D IC, we see that the overall routing stack can change significantly not only from 2-D but also between the two 3-D orientations. To understand different routing scenarios that are created due to this stacking, we split the nets into different categories based on connectivity and routing.

1) *3-D Nets:* Nets connecting cells from more than one tier.
2) *2-D Nets:* Nets connecting cells located in a single tier.
    a) *No Sharing:* 2-D nets that are routed on the BEOL of its own tier. Referred to as "default 2-D nets."
    b) *With Sharing:* 2-D nets that borrow tracks from metal layers of other tier for their routing

1) F2F Bonded 3-D IC: In the F2F orientation, the two tiers are attached at the metal layer face. Assuming a single-tier BEOL stack to be from metals 1 through x (*M1–Mx*), the 3-D BEOL stack would be as follows: *M1_bottom–Mx_bottom–Mx_top–M1_top*. In this configuration, the metal layer sharing is limited as the two FEOL layers are separated by the BEOLs of the two tiers.

Consider the 2-D net $n2$ from Fig. 1 (F2F case) that only connects the cells from a single tier (here, bottom FEOL). Specifically, this is a type of net that does not use metal layer sharing as the routing is limited to its own BEOL (*M1_bottom–Mx_bottom*). As such, the routing characteristics of this net are not different than a normal net in 2-D IC with only a single FEOL and a BEOL with monotonic decrease in parasitics per unit length.

From the pin connectivity, nets $n3$ and $n4$ are also classified as 2-D nets, as they connect to pins that are in a single tier (top for $n3$ and bottom for $n4$). However, these nets use routing tracks from metal layers that are not from their own BEOL. In our example (Fig. 1), this is shown by the use of *Mx_top* for the bottom-tier 2-D net $n3$ and routing on metal layer *Mx_bottom* for the top-tier 2-D net $n4$. As we have discussed previously, the number of nets that require higher metal layers for routing decreases gradually (from $M1$ to Mx), and thus, metal layer sharing can be very limited in F2F designs.

Finally, the nets of type $n1$ are examples of the 3-D nets as they connect cells from different tiers. In order to achieve full connectivity, these nets must be routed across all the layers in the 3-D metal stack (BEOL of bottom layer + BEOL of the top layer). This adds, what is referred to as a "3-D overhead," excess routing that needs to be done for 3-D nets due to taller BEOL stack and the placement of pins on either end of the stack.

2) F2B Bonded 3-D IC: The F2B stacking creates a 3-D metal layer stack that is the most different from any of those discussed above. By connecting the top layer of bottom tier (*Mx_bottom*) to the back side of the top-tier FEOL, the 3-D stack becomes *M1_bottom–Mx_bottom–M1_top–Mx_top*. This places the bottom tier *Mx*, which usually has the least routing, easily accessible to the cell pins of the top tier. This encourages metal layer sharing in the top-tier 2-D nets such as $n3$ from Fig. 1 (F2B case). Since the 3-D vias compete with standard cells for silicon area, there will additional detour to find legal locations for the 3-D vias going into bottom-tier BEOL.

Moreover, sharing on the bottom-tier nets such as net $n4$ is further restricted as they are placed further away from the *Mx_top* compared to the same type of net under F2F stacking. The high track utilization of metals above the top-tier FEOL and the added dependence between 3-D vias and the top-tier cell placement create additional restriction for metal layer sharing of $n4$-type nets.

3) 3-D Bonding and via Pitch: Apart from the orientation of the tiers, the bonding type also plays an important role in determining the feasibility of metal layer sharing, while a higher pitch discourages metal layer sharing, due to fewer available connections that can be made. The parasitics of the bonding structures also play an important part. For example, with micro-bump bonding, the bumps have a high parasitic value and can significantly add delays for the paths. Thus, micro-bump bonding is left out of consideration when talking about metal layer sharing. With sequential fabrication and hybrid bonding, neither the pitch size nor the parasitics become

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on December 11,2022 at 00:39:06 UTC from IEEE Xplore. Restrictions apply.

very important in the overall path delays and can be aggressively used for metal layer sharing.

III. Experimental Setup

A. 3-D PnR and Controlling the Metal Layer Sharing

For 3-D PnR, we use state-of-the-art tool flows Macro-3D [10] for logic-on-memory partitioning and Pin-3D [11] for logic-on-logic partitioning along with Innovus PnR tool (v20.15). This allows us to implement a wide range of partitioning and 3-D bonding types to analyze the routing. Macro-3D flow is well suited for designing memory-on-logic 3-D IC designs using hybrid bonding or monolithic integration, while Pin-3D flow is a more generalized flow that works well with any type of partitioning and supports both the 3-D bonding types that are possible with Macro-3D.

Neither Pin-3D nor Macro-3D offers any differentiation between net types (2-D or 3-D) during the routing stage, leaving the routing fully driven by the router. In order to analyze the metal layer sharing separately, we need to control the signal routing, which is done using custom scripts in our work. In PnR flow, early global routing of the signals starts as early as the placement stage where the trial routing is done to improve the placement quality. Detail routing is first done at the clock tree synthesis stage, where the clock network is routed before any other nets are routed to have the best possible clock tree design. Followed by this, the entire design is routed based on the results from the global routing stage.

In order to control the metal layer sharing, we identify the nets that connect to different tiers (3-D nets) and the nets limited to single tier (2-D nets). The 2-D nets are then restricted to be routed in their respective metal layers while letting the 3-D nets to be routed on the entire 3-D metal stack. In addition, the clock nets are handled separately to be routed on only the desired routing layers to limit the metal layer sharing. Clock tree requires special care as the clock optimization engine does not honor the routing rules initially set for the whole design. By controlling the metal layer sharing, we can effectively isolate and study its effects on the full chip PPA of 3-D IC designs.

B. Benchmarks and Technology Setup

1) Benchmarks Used: In order to analyze the metal layer sharing, three different commercially available CPUs are considered. These are widely used in consumer electronics and their names and layouts are not revealed to protect their IP as per our NDA. In the following, these CPUs are referred to as Industry-A, Industry-B, and Industry-C. Industry-A design is a dual-core processor with 512 kB of shared L2 cache and 32 kB each of L1 instruction and data caches. Industry-B is a larger single-core processor with 1 MB of L2 and 32 kB of L1 instruction and data. Industry-C is the last commercial circuit considered with 512 and 16 kB of L1 instruction and data caches. Finally, an open-source RISC-V processor (OpenPiton) is also considered to freely discuss and show the layouts wherever applicable.

TABLE II
Metal Layer Sharing in Different 3-D Orientations Using OpenPiton RTL. #MIVs on 2-D Nets Show the Amount of Metal Layer Sharing

	Units	F2B	F2F
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

2) Technology Process: The 3-D process design kit (PDK) is heavily based on a commercial 28-nm PDK in this work. For the monolithic integration case, the 3-D via (also referred to as MIV) is assumed to have a pitch of 0.14 μm, which is close to the pitch of an *Mx* via in 2-D. In the hybrid bonded 3-D IC, a larger pitch of 1.00 μm is considered to accommodate for alignment accuracy during bonding.

IV. Metal Layer Sharing Scenarios

In this section, we first empirically analyze the difference in metal layer sharing based on different 3-D types.

1) *Section IV-A:* Different orientations of 3-D ICs—F2F versus F2B.
2) *Section IV-B:* Different partitioning options—logic-on-logic versus logic-on-memory.
3) *Section IV-C:* Varying via pitch—0.1 versus 1.0μm 3-D vias.

The choice of the metal layer count is based on the logic-on-logic partitioned 3-D (Section IV-B), which requires significant routing on metal 6 of both top and bottom tiers.

Why Analyze Metal Layer Sharing?

While 3-D has multiple routing scenarios as discussed in Section II-B, the 2-D nets without sharing are the same as any net in a traditional 2-D IC design. The 3-D nets, while specific to 3-D and are interesting in their own right, are unavoidable and require to be routed to achieve full connectivity of the cells. The 2-D net routing with sharing, on the other hand, is specific to 3-D and can be controlled manually or using commercial tools. Therefore, it is important to understand the characteristics of these nets and their usefulness in the overall physical design of 3-D ICs.

A. Metal Layer Sharing With Different 3-D Bonding Orientations

As mentioned in Section II-B and shown in Fig. 1, the F2F and F2B orientations of the 3-D IC can show significantly different routing characteristics. This is analyzed and quantified in Table II using the OpenPiton RTL. The partitioning is left untouched between the two configurations (L3 data caches and related tag blocks on the memory die, and everything else

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on December 11,2022 at 00:39:06 UTC from IEEE Xplore. Restrictions apply.

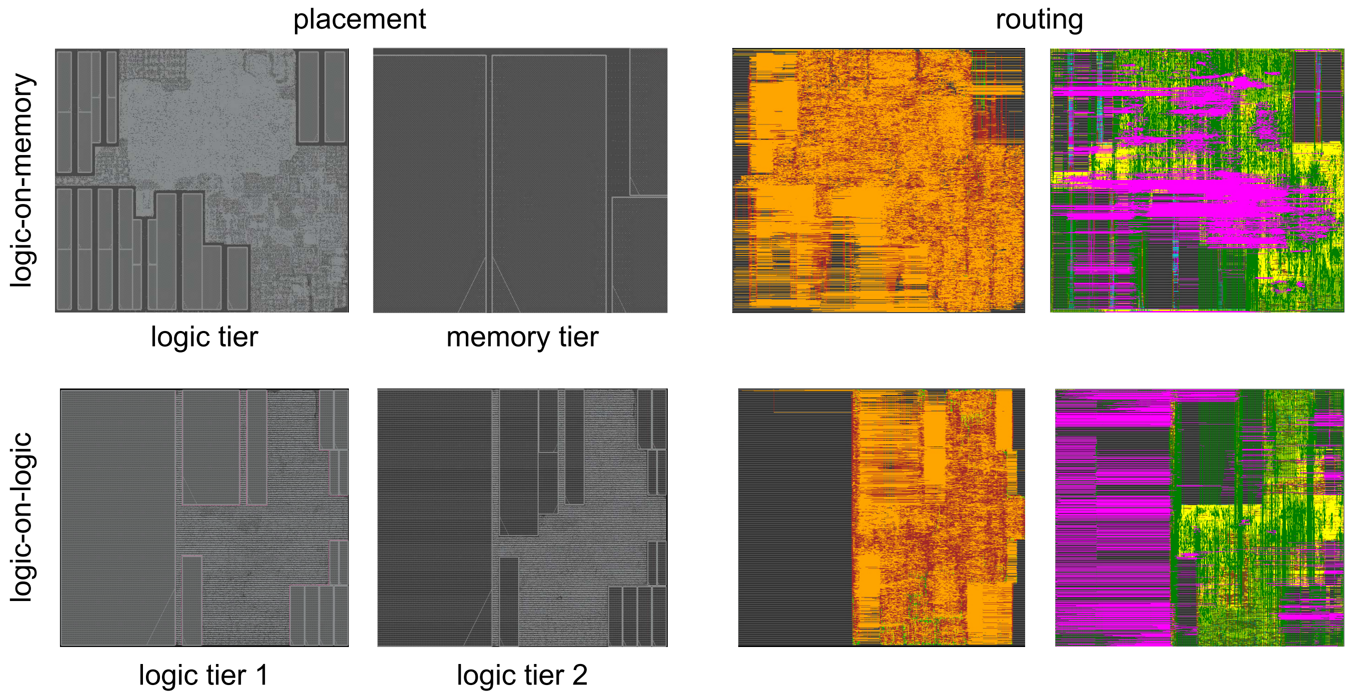

Fig. 2. Comparing tier partitioning impact on routing in OpenPiton. The placement and routing layouts in the two tiers are provided for the two styles of partitioning. Memory tier and logic tier 2 are the bottom FEOL in their corresponding designs.

on the logic die). Macro-3D [10] is used for the 3-D PnR of both these orientations as they have the memory-on-logic partitioning. The logic-on-memory layouts in Fig. 2 show the partitioning of the two tiers.

Comparing the design metrics from Table II, we first observe a few things: the two designs are implemented at the same target frequency and with the same chip area to have a direct comparison between the two in terms of routing and metal layer sharing. The major difference between the two comes in terms of the number of MIVs in the design. The F2B design makes abundant use of 3-D vias: 120 351 vias compared to just 3112 in the F2F case. A further classification of the MIVs based on the type of nets they are located on shows a clearer picture.

On the 3-D nets—which are governed by the partitioning—the number of MIVs is similar between the two with F2B having a slightly larger count by 215 MIVs. The 3-D nets are logically the same between the two orientations, and this increased via count comes from the difference in routing the nets between the two designs. Due to the sudden change of the routing environment between the layers *Mx_bottom* and *M1_top* at the interface of the 3-D connection in F2B, the routing behavior becomes more chaotic. The high routing blockage density of the *M1_top* compared to a lower blockage and routing densities of the *Mx_bottom* layer, the 3-D nets cross the 3-D interface multiple times, thereby accruing multiple MIVs per net. This is referred to as snaking and is a commonly observed phenomenon for 3-D ICs (especially with automatically routed F2B 3-D ICs).

Most of the additional MIVs in the F2B designs are on the 2-D nets. As discussed in Section II-B2, the ease of access to the bottom-tier BEOL from the top-tier nets creates this situation. Moreover, since a 2-D net connects cells within a single tier, in order to borrow routing tracks from metal layers of the bottom BEOL, the nets cross the 3-D interface twice to access the bottom BEOL and then back again to connect the sink pins. Therefore, 2-D nets undergoing metal layer sharing have at least two MIVs per net.

Finally, we analyze the macro-level design properties of the two designs. The wirelength of the F2B version is larger due to the added "snaking" and the additional routing required for metal layer sharing. While this affects the total power slightly by $\approx$3 mW, it is more beneficial for the worst timing path delay. Two factors are at play here that are dependent on the routing stack and the routing behavior. With the two FEOLs separated by both bottom and top BEOL stacks in the F2F configuration (Fig. 1), the 3-D nets would have longer wirelength. Thus, paths through the 3-D nets are vulnerable to delay increase in this configuration. Second, with more metal layer sharing of the F2B configuration, the metal layers are well utilized with a reduced routing usage on the top-tier BEOL stack. This decreases congestion in the design and allows for fewer detours in routing critical paths on the top tier. The routing layouts in Fig. 3 show a noticeable difference in the memory tier routing, which is left underutilized by the F2B option. Consequently, we see that in the logic tier, the F2F routing has more long wires on the top-most *Mx* layer (orange colored in Fig. 3) compared to the F2B option.

B. Metal Layer Sharing With Different 3-D Partitioning

In this section, we turn our attention to the partitioning of the RTL and its impact on routing in 3-D. The two choices considered are the logic-on-memory (same as the partitioning

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on December 11,2022 at 00:39:06 UTC from IEEE Xplore. Restrictions apply.

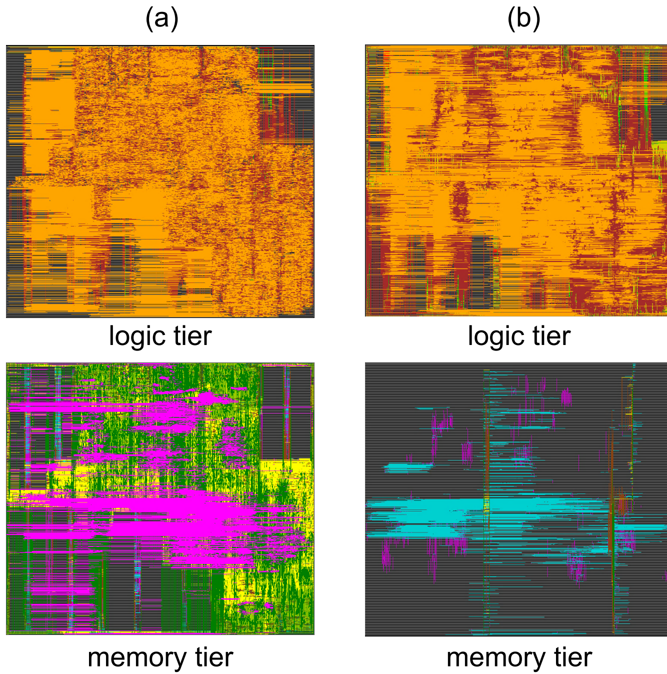

Fig. 3. Routing comparison between two bonding styles of logic-on-memory 3-D ICs. (a) F2B. (b) F2F. The logic tier BEOL layouts are on the top, and memory tier BEOL layouts are on the bottom. Each color corresponds to a routing layer.

discussed in Section IV-A) and the logic-on-logic partitioning. The corresponding layouts are shown in Fig. 2 for both placement and routing. Unlike logic-on-memory partitioning, where the memory tier is limited to preplaced macro blocks, the logic-on-logic style has both logic and memory blocks on both tiers and is implemented using Pin-3-D flow [11]. Pin-3D flow cannot properly optimize the logic-on-memory design due to the largely asymmetrical partitioning, and Macro-3D cannot be applied to designs with logic block on both tiers. Because of this, the two different flows are used to handle the two partitioning types.

F2B orientation is used for both 3-D partitioning options to encourage metal layer sharing. The PPA comparisons and other design metrics are shown in Table III. We first see that the logic-on-memory style has a slightly larger footprint than the logic-on-logic option. This is because the memory tier can only fit macros, which do not fit well together leaving unused white space in the logic-on-memory option. This can be seen in Fig 2 where the memory tier has some unused white space. In terms of the MIV count, both options have a similar MIV count at around 100k and 120k. However, the origin of these MIVs is very different from each other, with most of the MIVs in the logic-on-memory option coming from 2-D nets, but they are majority from 3-D nets in the logic-on-logic option. We have already analyzed the MIV distribution for logic-on-memory case in Section IV-A and are not further mentioned here. In logic-on-logic, the coarse gate-level min-cut partitioning creates a very large cut size that in turn creates a high number of 3-D nets and MIVs. More than 85% of the total MIVs, in this case, are on the 3-D net, which is a stark contrast to the distribution in the logic-on-memory option. Out of the 17 000 MIVs on the 2-D nets, we see that virtually, all of them are used for borrowing tracks from the bottom tier. This further supports our discussion of the metal stack and routing discussion in Section II-B2. The uneven metal layer sharing across the two dies is also seen in the logic-on-memory partitioning.

As a whole, the total wirelength is significantly smaller in the logic-on-logic designs due to the symmetrical nature of the partitioning and the high 3-D net count of this option. As the number of 3-D nets increases, more nets can have shorter wirelengths due to the 3-D placement as well as the net detours. This leads to much smaller total power consumption due to the decreased wire load. Out of the total routed wirelength, more than 25% is on borrowed metal layer tracks showing the abundance of metal layer sharing for the logic-on-memory option. This percentage is only at 6% with the logic-on-logic option. The main reason for the decreased sharing, in this case, is due to the large number of 3-D nets along with the symmetrical partitioning in both tiers.

With logic-on-memory partitioning, the memory die has a lot of unused routing tracks on M5 and M6 that are not used by the memory blocks for intracell routing. This allows for more of the logic tier (top tier) nets to flow onto the bottom tier. In contrast, the symmetrical nature of the logic-on-logic case (memory blocks are placed on top of each other, and the sea of logic cells of both tiers is also largely located in the same region of the tier). This makes the unused routing tracks on top of the memory macros harder to be utilized by the other nets in the design. At the same time, the routing tracks on top of the sea of logic region are heavily utilized with not enough free space for metal layer sharing. This is visualized in Fig. 4, where the routing of layers M5 and M6 in the bottom tier is shown for both partitioning options. Note that the bottom tiers M5 and M6 have the highest metal layer sharing wirelength among all the layers for both options as seen in the "Shared Wirelength" block of Table III. The wires in Fig. 4 are colored based on their type of routing. The wires that belong to nets on the top die (routed on borrowed tracks from bottom/memory tier) are shown in red, and the other nets (default 2-D and 3-D net routing) are shown in yellow. The %age of wirelength in the memory tier layers (bottom tier for logic-on-logic case) used for the purpose of metal layer sharing is also calculated in Table III showing more than 95% shared wirelength on layers M5 and M6 of the memory tier. Metal layers M4–M1 in the bottom BEOL of logic-on-memory option are mostly occupied by the intracell routing of memory macros and for intradie routing in logic-on-logic.

C. Impact of Pitch on the Metal Layer Sharing

Finally, we measure the impact of pitch on the 3-D routing and metal layer sharing using two the logic-on-memory partitioning in the F2B orientation. Two pitch values (0.14 and 1.0 μm) are used for this comparison and the results are tabulated in Table IV. We quickly see a drastic reduction in the number of MIVs used. This is due to the larger pitch, which increases the area occupied by each MIV by 100×. The shared wirelength is also reduced by ≈45% due to this. As seen with different orientations in Section IV-A, the reduced sharing decreases the wirelength but increases the worst slack.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on December 11,2022 at 00:39:06 UTC from IEEE Xplore. Restrictions apply.

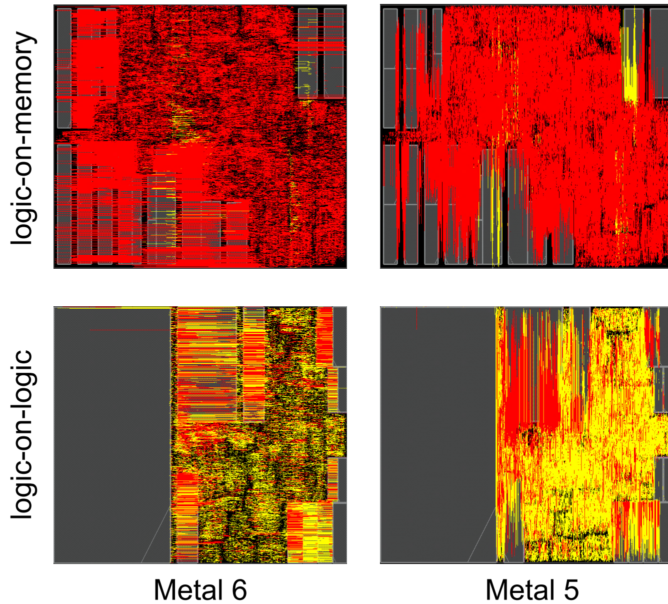

Fig. 4. Routing in shared metal layers of 3-D OpenPiton design with F2B bonding style. We show M5 and M6 of the memory tier and logic tier 2. Red is routing with metal sharing, and yellow is everything else.

TABLE III

METAL LAYER SHARING IN 3-D PARTITIONING OPTIONS: LOGIC+MEMORY AND LOGIC+LOGIC. #MIVS ON 2-D NETS SHOW 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	–	120,351	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
Shared Wirelength			
Top Layer M1 – M6	μm	0	10,567
M6	μm	690,461	111,529
M5	μm	888,242	138,754
M4	μm	21,197	11,366
M3	μm	14,807	694
M2	μm	47	59
M1	μm	0	7

TABLE IV

METAL LAYER SHARING WITH F2B ORIENTED LOGIC+MEMORY PARTITIONING AT DIFFERENT PITCH VALUES/BONDING TYPES

	Units	Monolithic 3D 0.10 μm Pitch	TSV-Based 1.00 μm Pitch
Frequency	MHz	1400	1400
Chip Area	mm^2	0.638	0.638
# MIVs	–	120,351	33,194
# MIVs on 2D nets	–	119,317	32,264
# MIVs on 3D nets	–	1034	930
Borrow from bottom	–	119,317	32,264
Borrow from top	–	0	0
MIV Occupied Area	mm^2	0.001	0.033
Ideal MIV Occupancy	%	<1	29.6
Wirelength	m	6.36	6.06
Cell Area	mm^2	0.512	0.514
White space	mm^2	0.136	0.134
# Routing violations	–	9353	128,162
Worst Negative Slack	ns	-0.384	-0.410
Effective Frequency	MHz	910.5	889.5
Total Negative Slack	nHz	-864.5	-934.0
Total Power	mW	414.6	408.7
Shared Wirelength			
Top Layer M1 – M6	m	0	0
Bottom Layer M1–M6	m	1.86	1.01

Another key difference that undermines the usage of 1.0-μm bumps for metal layer sharing is the number of violations in the design. The larger 1.0-μm MIV pitch is significantly large ($\approx 10\times$) than the pitch of metal layers and vias around it. This creates routing violations, as shown in Table IV. Similarly, the height of the cells with which the MIVs compete for area in the F2B orientation is 1.2 μm making it comparable to the larger MIV pitch. This severely limits the number of MIVs that can be placed in the design. Ideal MIV occupancy roughly calculates the number of MIVs that can fit in the free area, and the 1.00-μm pitch already utilizes 30% of the available space. In comparison, the smaller MIV pitch has <1% utilization even with its larger MIV count.

Due to the abovementioned reasons, we only use an aggressive pitch of 0.1 μm. This is in-line with the 3-D pitch values used in other works that consider monolithic 3-D integration.

Takeaway: Of all the configurations considered, metal layer sharing is most prevalent in the memory on logic partitioned 3-D IC due to the absence of standard cells in the bottom die and the sparse track usage by the intracell routing of the bottom die. F2B orientation is also crucial for enabling sharing of the tracks between the tiers. Finally, the F2B orientation also necessitates a monolithic integration of the 3-D IC with its fine pitch value rather than hybrid bonding. By identifying the best configuration to analyze the track sharing between BEOLs, we move on to the main portion of the analysis, which investigates the different pros and cons of track sharing in Section V.

V. RESULTS

All the implementations are chosen using a frequency sweep between 1000 and 1500 MHz with the highest effective frequency selected as our candidate for comparison. As power delivery network (PDN) is a crucial aspect of the physical design, all the designs in this section are included with a similar PDN for comparability. The three designs used across

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on December 11,2022 at 00:39:06 UTC from IEEE Xplore. Restrictions apply.

all the below results are the commercial system-on-chip (SoC): Industry-A, Industry-B, and Industry-C.

A. Baseline Experiments

1) 3-D Metal Layer Stack: Before starting the comparisons and analysis of metal layer sharing, we first set baseline designs that would be helpful in all the later discussions. To do so, we need to first find the number of metal layers required to satisfactorily route the logic-on-memory design with the PDN. The preliminary analysis for various designs in Section IV uses the metal layer structure that is limited by the logic-on-logic partitioning, which requires more metal layers due to the distribution and placement of memory macro across the two tiers. This is excessive for logic-on-memory partitioning as evident from the sparse routing in Fig. 3.

Memory macros with the 28-nm technology use up to four metal layers (M1–M4) for internal routing. Thus, at least four metal layers are required in the memory BEOL. However, the F2B nature of the monolithic 3-D ICs creates additional routing issues (especially for PDN) based on the partitioning of the macros. Fig. 5 shows an example of this issue. For the macros or macro pins under the large placement blockages in the top logic tier (here, another memory macro), the signal routing needs to be routed in a small channel toward area not blocked by logic tier placement. This significantly impacts the timing closure capabilities of the 3-D IC. More importantly, power delivery becomes increasingly challenging with just four metal layers in the memory BEOL. In regions where memory-tier macros are not blocked by logic-tier macros in the xy plane, power delivery can be done using MIVs from the top tier directly onto the power rails within the memory macros without the need for additional distribution layer in the memory BEOL. However, with memory macros on logic tier, the portion of or the entire macros that are covered by them in the memory tier cannot be supplied with power and ground supply. Because of this, we use up to five metal layers for the memory BEOL as our baseline.

Table V shows the percentage of available tracks in the metal layers used by the intercell routing in the design. The available track length per each layer is calculated by subtracting away all the routing blockages due to memory macros, logic cells, and other sources of obstruction to routing that are present in the design. The % tracks removed due to blockages are also shown in this table. In the memory tier, we see ≈99% possible routing area blocked by the memory macros on layers M1–M4. Out of the remaining nonblocked tracks, metals M2 and M4 have 15% or more usage, while the usage in M1, M3, and even M5 is limited to under ≤2%. This mainly comes from the orientation of the macros in the memory tier (and the routing channels formed by these macros) and the preferred orientation for routing in these metal layers. In a vertical channel such as the one formed on the memory tier of example Fig. 5, the metal layers with vertical preferred directions have more usage as there can be longer uninterrupted wires in such areas. In the direction orthogonal to the routing channel (vertical direction in this case), the available routing tracks are much shorter making these routing channels underutilized.

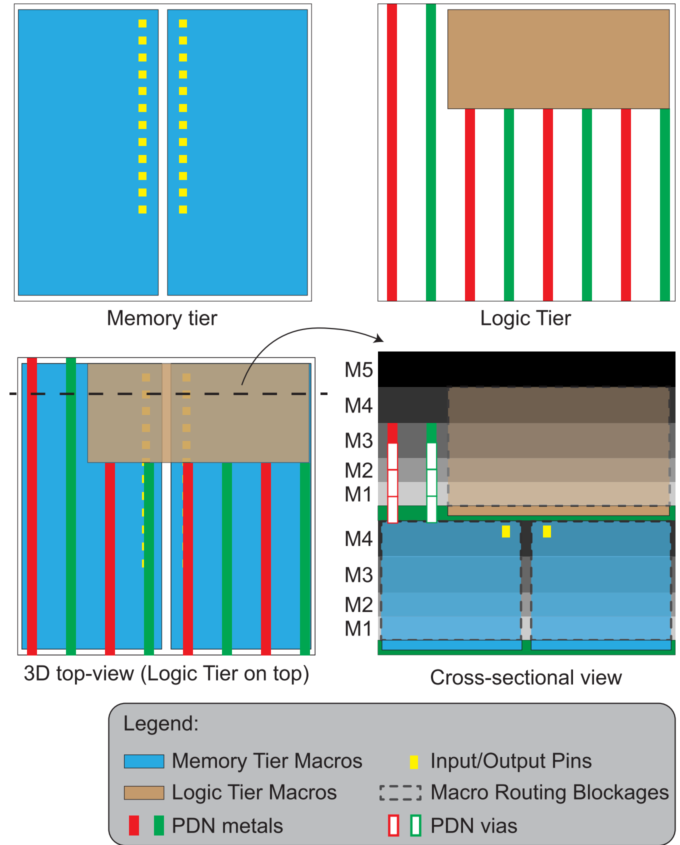

Fig. 5. Partitioning scenario showing the obstructions caused by memory macros with just four layers in the bottom BEOL. The cross-sectional view is shown at the cutline of the 3-D view.

While the results in Table V are particular to Industry-A circuit, the overall signal usage trends are similar across the other circuits in consideration too. Unlike signal routing, PDN is fully controlled by the inputs (layers used for PDN routing, width, and spacing) and is kept constant across all the designs. This generates a PDN with similar metal usage across all the circuits.

2) Results: With the baseline setup established, we perform the PnR of the three Industry circuits (referred to as Ind-A, Ind-B, and Ind-C). As mentioned at the start of this section, the design with the best effective frequency is selected to represent the baseline for each circuit. The number of power and ground (PG) MIVs depends on the PDN pitch used in the top- and bottom-tier metal layers, as well as the overall design footprint. Between the baseline designs, the PDN pitch is identical across all the layers and the varying PG MIV count is strictly attributed to the difference in the footprints of the three circuits. Of the total signal MIVs, most of them are on the 3-D nets, which is the result of the MIV control explained in Section III-A. There exist a few MIVs on the 2-D nets, which is the result of the postroute optimization stage of 3-D ICs. The routing constraints to control the MIVs on 2-D nets are only applied to the nets present in the design during the MIV control. To solve this problem, we reevaluate the constraints at each stage of 3-D IC design. However, after routing, constraints are only added to nets present during routing. Also, the nets added to the design by the automated

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on December 11,2022 at 00:39:06 UTC from IEEE Xplore. Restrictions apply.

TABLE V

METAL LAYER USAGE OF SIGNAL AND POWER NETWORKS IN THE BASELINE 3-D METAL STACK WITHOUT SHARING AND THE REDUCED METAL STACK WITH SHARING. USAGE IS CALCULATED AS % OF AVAILABLE TRACKS USED FOR ROUTING. BLOCKED TRACKS IS THE % TRACKS BLOCKED COMPARED TO TOTAL POSSIBLE TRACKS IN THE FOOTPRINT. INDUSTRY-A DESIGN IS USED FOR THE FOLLOWING CALCULATIONS

Metal Layer	Baseline Metal Usage			Usage with reduced metal stack		
	Signal Usage	PDN Usage	Blocked Tracks	Signal Usage	PDN Usage	Shared Usage
Memory Tier						
M1	0.0	0.0	98.8	0.0	0.0	0.0
M2	15.0	0.0	98.8	13.2	0.0	4.9
M3	2.0	2.0	98.8	3.6	2.4	2.2
M4	17.9	1.0	98.8	19.0	1.0	14.3
M5	1.3	9.4	0.0	13.2	9.4	12.3
Logic Tier						
M1	0.0	0.0	25.5	2.0	0.0	0.0
M2	20.4	21.5	20.1	21.9	21.6	0.0
M3	32.2	26.6	19.5	34.2	26.5	0.0
M4	25.8	18.2	19.5	29.3	18.5	0.0
M5	24.0	17.2	0.0	21.9	21.7	0.0
M6	14.8	16.1	0.0	unused		

timing optimization of the PnR tools delete some nets and add others. This last-stage optimization is the reason for the MIVs on the 2-D nets in the design. In all the designs considered, the # MIVs on 2-D nets are insignificant compared to the total number of 2-D nets. As such, there should not be any measurable impact on the PPA due to these nets, acting as a good baseline to measure the impact of metal layer sharing.

B. Metal Layer Sharing and Cost Saving

1) Overall Track Usage: From the left half of Table V, related to baseline design, we have seen that although metal M5 of the memory tier is necessary for power delivery, it is severely underused for signal routing when metal layer sharing is not allowed. With no intracell routing in this metal, as evident by the lack of any routing blockages, it is helpful to use this layer for metal layer sharing. Furthermore, the top-most layer of the logic tier (M6) is also removed as the memory tier's M5 is additionally used for routing the logic tier nets. With this 5+5 metal stack, we can see a one metal layer reduction compared to the baseline. Although it is feasible to simply drop a metal layer in this case, the overall impact on the PPA should be verified.

The right half of Table V shows the track usages for all the metal layers with the reduced metal stack. Since the macro placement is left untouched, the routing blockages are the same and are not repeated. In addition, a new column for the borrowed track usage is added, which signifies the metal layer sharing. In the memory tier, we see that the track usage of the signals is fairly similar with a 2% increase for layers M2–M4. M5 of memory tier, which is the least utilized and blocked, shows the largest difference as its tracks are borrowed by the wires of the logic tier. Specifically, 12.3% of the tracks in memory tier M5 are used by logic tier. This is the additional wirelength rerouted from M6 of top tier (which, otherwise, has 14.8% track usage when metal sharing is not allowed).

In the logic tier, the track usage of M5 decreases as many long wires are preferentially routed on the memory M5. The track usage of the layers M1–M4 see a slight increase as they are rerouting the nets to the memory tier, in the absence of logic M6. Finally, in M1 of logic tier, the usage without metal layer sharing was 0%, which increases to 2% with sharing. In general, because of the large prevalence of intracell routing in this tier, the router is discouraged from any additional routing in the baseline case. The only intercell routing is for pin access for cells that do not have pin shapes on higher metal layers and the routing that is associated with pin access. With metal layer sharing, M1 is needed for all the nets that borrow tracks from memory M5. 58 073 out of a total 561 669 nets in the design are shared 2-D nets that are routed through logic M1 layer to access the memory M5. Even with additional routing, only 2% of the unobstructed tracks in M1 are utilized.

2) Routing Summary: The routing violations from Table VII gives us a better picture of the routing quality than the track usage. Note that only ten iterations are used for fixing the DRVs for all the designs considered. While using more DRV fixing iterations could have reduced the violations, we wanted to compare the type of violations that could be possible. With most of the designs having under 2000 violations over the entire footprint, the DRVs are not a bottleneck. On the memory M5, even with only a relatively lower track usage (12.3%) with metal layer sharing, we see a high DRV count. This can be mainly attributed to the routing layer setup and routing blockages in 3-D, along with the placement of MIVs.

To understand the reason behind these violations on memory M5, Fig. 6 shows a zoomed-in routing in metal M5 of the memory tier (layer with most metal layer sharing). From this, we see a few interesting aspects of the borrowed track routing using five metal layers in each tier. First, we see the MIVs placed in the spaces left behind by the standard cells of the logic tier (shown in gray). When using metal layer sharing, the nets of top tier are routed via an MIV to the bottom (memory) tier and up to the top (logic) tier via another MIV. This presents two challenges. First, since vias can only be located in certain places between the standard cells, we see more detours. Second, as the bottom tier has only five layers, of which four layers are heavily (≈99%) blocked by the intracell routing of macros, the routing is not always optimal. Once the wires are routed through MIV onto the memory tier, the nets that have

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on December 11,2022 at 00:39:06 UTC from IEEE Xplore. Restrictions apply.

TABLE VI

DESIGN METRICS OF THE THREE RTLS CONSIDERED IN OUR WORK. THE DESIGNS ARE IMPLEMENTED IN AN F2B 3-D FASHION. THESE ARE THE BASELINE DESIGNS THAT DO NOT HAVE METAL LAYER SHARING AND THE REDUCED METAL STACK DESIGN WITH METAL LAYER SHARING

		Baseline			Reduced Metal Stack		
	Units	Ind-A	Ind-B	Ind-C	Ind-A	Ind-B	Ind-C
Target Frequency	MHz	1500	1375	1375	1375	1250	1250
Chip Area	mm^2	1.109	1.893	1.051	1.109	1.893	1.051
# Metal Layers	–	11	11	11	10	10	10
# PG MIVs	–	3620	5872	3222	1248	2040	1110
# Signal MIVs	–	4157	1703	2282	159045	312420	190302
# MIVs on 2D nets	–	83	94	78	154765	318074	187868
# MIVs on 3D nets	–	4074	1602	2204	4280	1546	2434
# MIVs on clk nets	–	91	94	96	4221	6823	6720
# Routing Violations	–	1038	1201	1468	1247	327	880
# 2D Nets	–	571650	790350	454312	547745	745072	427412
# 3D Nets	–	3118	1311	1359	3162	1307	1337
# Clock Nets	–	4866	6175	4280	4746	5761	4000
# Shared 2D Nets	–	–	–	–	58073	106795	65235
# MIVs per Shared 2D Net	–	–	–	–	2.66	2.98	2.88
# MIVs per 3D Net	–	–	–	–	1.35	1.18	1.82
Wirelength	m	12.214	19.438	10.541	12.244	19.253	10.436
Worst Neg Slack	ns	-0.352	-0.334	-0.286	-0.231	-0.309	-0.210
Total Neg Slack	ns	-1648	-2287	-913	-1439	-1465	-668
Effective Frequency	MHz	981.7	942.3	986.9	1043.5	901.8	990.1
Effective Power	mW	592.3	730.5	444.9	596.5	648.0	412.6
Eff. PDP (Power×Delay)	pJ	603.3	775.2	450.8	571.6	718.6	416.7

to be routed in both horizontal and vertical directions only have a single layer to do so.

This suboptimality can be seen from the long jogs in the M5 routing. Table VII also shows this from the % routing in the nonpreferred direction values. Values for M1–M4 of memory tier can be safely ignored as they only have limited routing within the channels. M5 of memory tier has 3.5% of its wirelength routed in the nonpreferred direction when metal sharing is turned on, this is relatively high compared to M3–M5 of logic tier under the same implementation. On metals M1 and M2 of logic tier, the jog % is inherently larger due to the proximity to standard cells and the short average wirelength. With metal layer sharing, M1 of logic tier is mostly used for routing to MIVs and has a high % of jogs. Even with the added complexity of the routing, we are assured by the fact that only 112 violations remain in the logic M1 with metal sharing (lower than the baseline).

From the average wire segment lengths, we first see that the averages in the logic tier are always slightly higher in the metal sharing version. This comes from the added routing on these layers to perform metal layer sharing. On M5 metal of memory tier, the average wire segment length becomes shorter with metal layer sharing. This is mainly because of the type of routing. In the baseline, only the long nets that need access to top tier for 3-D nets are routed on this layer. With metal sharing, a lot of logic tier nets access this layer as well the increased jogs add to the number of wire segments (an uninterrupted portion of wire in the horizontal or vertical direction).

3) PPA Results: Finally, we compare the maximum performance of the three industry circuits with the reduced metal layer stack and metal sharing that are presented in Table VI. The target frequencies correspond to the design, which can

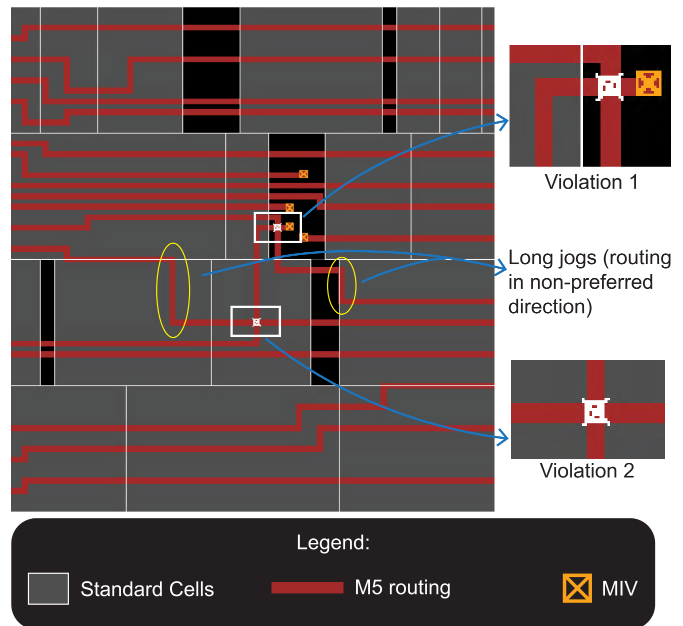

Fig. 6. Zoomed-in shot of M5 routing in the metal layer sharing design. We can see the routing jogs and shorts in this layer.

achieve the highest effective frequency similar to the baseline setup. We see a large number of MIVs used for routing signal nets compared to baseline. As the partitioning is left unchanged, the MIVs on 3-D nets are very similar between the two cases. The MIVs on clock nets also increase significantly with nearly 1 MIV per clock net on average in each circuit. We also see that, on average, the shared 2-D nets as a whole use 2.66–2.98 MIVs per net depending on the design. The 3-D nets, on the other hand, use much fewer MIVs per net between 1.18 and 1.82. This is in-line with our expectations, as the

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on December 11,2022 at 00:39:06 UTC from IEEE Xplore. Restrictions apply.

TABLE VII

ROUTING SUMMARY OF THE INDUSTRY-A DESIGN WITH DIFFERENT METAL LAYER SHARING OPTIONS. THE TWO COLUMNS CORRESPOND TO THE INDUSTRY-A COLUMNS IN TABLE VI

Metal Layer	Without Metal Sharing (Baseline)	With Metal Sharing
# Design Rule Violations		
Memory Tier		
M4	3	3
M5	8	1097
Logic Tier		
M1	144	112
M2	808	20
M3	71	13
M4	0	0
M5	1	0
M6	0	unused
Total Count	1038	1248
Average Wire Segment (in μm)		
Memory Tier		
M4	6.60	2.92
M5	9.57	3.41
Logic Tier		
M1	2.28	0.44
M2	0.54	0.56
M3	1.73	1.89
M4	3.43	4.47
M5	8.59	10.83
M6	20.00	unused
% Routing in Non-Preferred Direction (or) % Jog wirelengths		
Memory Tier		
M4	0.2	0.6
M5	0.7	3.5
Logic Tier		
M1	1.2	12.2
M2	6.0	5.2
M3	0.4	0.4
M4	0.3	0.2
M5	0.0	0.0
M6	0.0	unused

2-D nets with metal sharing need to pass the 3-D interface layer twice or more to be fully routed. Overall, there is only a negligible impact on wirelength with metal sharing compared to the overall design routing. Note that this is with one fewer layer in the logic tier.

When we compare the timing across the two, we see that the total negative slack at the target frequency is worse with sharing. This is due to the worsening of the nets that need to be shared. Although the most critical paths are not severely impacted as seen from the effective frequency, the noncritical paths are all slightly worsened leading to an overall total slack degradation. For the most critical path, except for the Industry-B design, both A and C circuits were able to reach a slightly better or similar frequency with one layer dropped. More importantly, the power benefit is significant for both Industry RTLs B and C, and the overall PDP is better (5%–7%) with metal layer sharing. This shows the usefulness of metal layer sharing in efficiently utilizing the memory tier's M5 and simultaneously dropping the logic tier's M6 without negatively affecting the full-chip performance.

TABLE VIII

TIMING ANALYSIS OF THE CRITICAL PATHS AND CLOCK TREE RESULTS OF INDUSTRY-A DESIGN

Metric	Units	Baseline	Metal Sharing
Effective Period	ns	1.019	0.959
Worst Slack	ns	0.0	0.0
Top-50 Register to Output Path Averages			
Path Delay	ns	0.585	0.644
Setup	ns	0.0	0.0
Skew	ns	0.417	0.271
Slack	ns	0.017	0.044
Path Length	μm	386	397
Logic Depth	–	3.8	3.4
Capture Latency	ns	0.4	0.4
Launch Latency	ns	0.817	0.671
Top-50 Register to Register Paths			
Path Delay	ns	0.801	0.885
Skew	ns	0.016	-0.040
Setup	ns	0.088	0.084
Slack	ns	0.112	0.031
Path Length	μm	458	510
Logic Depth	–	19.7	18.5
Capture Latency	ns	0.654	0.606
Launch Latency	ns	0.700	0.565
Clock Tree Results			
Maximum Latency	ns	1.060	0.952
Minimum Latency	ns	0.440	0.382
Average Latency	ns	0.680	0.746
Std. Deviation Latency	ns	0.066	0.064
Maximum Skew	ns	0.420	0.450

C. Full-Chip Timing Improvements

1) Timing Improvement: In order to find the root of the PPA improvement, we focus our attention first to the Industry-A design. This shows a relatively large improvement in terms of maximum frequency, which is solely responsible for the PDP improvement. To analyze, we look at few of the worst critical paths to analyze their behavior. These results are tabulated in Table VIII for the Industry-A design, which shows the highest performance benefits.

The results presented here correspond to the Industry-A designs reported in Table VI. The values are reported after the clock period is changed to the max achievable period of the respective implementations. This is why the worst slack is 0 ps for both designs. Slack is derived as

$$\text{Path Slack} = \text{Clock Period} - \text{Path Delay} - \text{Setup} - \text{Skew}$$

We analyze the top-50 register-to-output and register-to-register paths in the design. By averaging over a larger number of nets than focusing on the single most critical path, we can isolate some of the eccentricities of a single path. As the two different path groups considered have very different characteristics, we separate them for analysis.

The register-to-output paths only have a portion of path in the chip as can be seen by the very small logic depth of around 3–4. When comparing the overall paths, we see that while the average path delay is marginally better in the baseline design, the main cause of worst period is the clock skew and the clock tree synthesis. The capturing clock

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on December 11,2022 at 00:39:06 UTC from IEEE Xplore. Restrictions apply.

TABLE IX
ENERGY CONSUMPTION PER UNIT CLOCK PERIOD AT MAXIMUM FREQUENCIES FOR INDUSTRY-B DESIGN

Metric	Units	Baseline	Metal Sharing	Delta%
Effective Period	ns	1.061	1.108	4.43
Total Energy	pJ	775.2	718.6	-7.30
Internal Energy	pJ	391.7	360.3	-7.94
Switching Energy	pJ	382.3	357.2	-6.57
Leakage Energy	pJ	1.214	1.124	-7.41
Standard Cell Area	mm^2	1.023	0.931	-8.99
Wirelength	m	19.44	19.25	-0.98
Input Pin Cap	pF	2150	1909	-11.2
Wire Cap	pF	3084	2940	-4.67
Ground Cap	pF	2097	1988	-5.20
Coupling Cap	pF	996.8	951.6	-4.53

latency is the same between the two as it is an estimate based on the clock tree synthesis outside the current chip. The launch latency is significantly different between the two, as the baseline design has a worse latency on these critical paths. The results from the clock tree synthesis further demonstrate this difference. The baseline design has worse overall clock latency, which creates the bottlenecks on the critical paths. While the average latency is worsened with baseline, we see that the tool was still able to do a good clock tree design as the standard deviation of the latency is similar among the two designs.

On the register-to-register paths, we see that the baseline has a better path delay as well as better slack. However, as the bottleneck paths in this design are the register-to-output paths, the slack on these paths is not helpful for the maximum frequency. In the metal sharing design, the average slack of this path group (0.031 ns) is similar to the average slack of the register-to-output path group (0.044 ns). This shows that the two path groups are much more closely distributed in terms of timing in the metal sharing design. Moreover, in the baseline, we see that the paths are shorter in length but have more cells, which is a sign of over-optimization as long nets are broken down into shorter nets by adding buffers. This is required for the cells to meet timing in the bottleneck paths, which is more critical for baseline design.

2) Power Improvement: Two of the three designs discussed in Table VI have improved PDP due to the power improvement resulting from the metal layer sharing. We analyze the two Industry-B implementations that show the highest power delta. Since power is a function of frequency, we report the power efficiency or the energy consumption per unit clock period (same as the PDP) in order to normalize the power and compare the overall trends. Splitting the total energy into internal, switching, and leakage in Table IX, we see that each component is reduced almost the same amount as the total energy.

Internal energy is the energy consumed within the cells and has a high dependence on the total cell area. Note that we only used a single threshold voltage type (lowest V_{th} available) to keep the technology setup simple. The 9% reduction in cell area is the main reason for the internal energy reduction. The reduced cell area is a combination of better power and timing of the metal sharing designs. These designs reached their maximum frequency at a lower target period than the baseline in each case. Even with the lenient target, the maximum frequency reached is higher or comparable as seen from the reduced metal stack design in Table VI. The combination of these two effects has resulted in a higher power efficiency as discussed throughout our work.

Furthermore, the switching energy, which depends on both the routing and the cell sizing, also sees a significant reduction with metal sharing. The energy consumed due to a net switching is $1/2\alpha CV^2$, where α is the activity factor, C is the load capacitance (= wire cap + input pin cap), and V is the voltage of the signal on the net. With metal layer sharing, the only difference would be in the switching load (C) and some negligible differences in activity factor based on the design optimization. We see that the input pin cap reduces by 11% with metal sharing from the reduced cell area. The wire capacitance also reduces by $\approx$4.5% even though the wirelength reduces by $<$1%. Factors, such as wirelength distribution and fan-out, can also affect the ground capacitance of wire. More importantly, the unit capacitance per layer differs across the different layers of the 3-D metal stack. For example, in the current 3-D baseline metal stack, logic tier M5 has a capacitance of 0.155 fF/μm, logic tier M6 has 0.113 fF/μm, and memory tier M5 has 0.112 fF/μm. Thus, the difference in routing across the metal layers also leads to the difference in the wire ground capacitance. The cross-coupling capacitance between the nets is also reduced by 4.53% showing the improvement of routing by allowing metal layer sharing.

Leakage energy, such as internal energy, is mainly dependent on the standard cells as well, and the 7.4% reduction stems from the cell reduction. The overall power delay product improvement for the Industry-C design also follows the trend presented for Industry-B.

D. How Different Nets Help With Metal Layer Sharing?

As we have seen until now, metal layer sharing can lead to various benefits in 3-D designs. In Fig. 7, we look at power and performance metrics as we change the amount of sharing. In the max-frequency analysis, we see that each circuit behaves very differently from each other as the worst slack is very dependent on the design. Therefore, in the case of Industry-A design, the maximum achievable frequency increases when the clock tree is first allowed to use metal layer sharing and then another rise in frequency once more when the overall nets are allowed to have metal layer sharing. Industry-C design is more or less constant in terms of the maximum frequency. Looking at the total power, the trend becomes more stable as it becomes on the overall nets rather than just a subset of the nets/critical nets. The power trends shown here are normalized with respect to the type-1 implementation in each design. We see that when metal layer sharing is restricted, the power consumption increases across the three designs. This shows the usefulness of metal layer sharing as it can reduce congestion in tiers. Thus, overall, we see that the power benefit comes from the complete metal layer sharing, but performance

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on December 11,2022 at 00:39:06 UTC from IEEE Xplore. Restrictions apply.

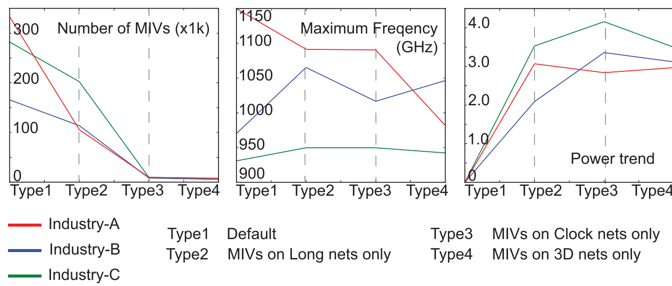

Fig. 7. Max-frequency and iso-performance power impact on the three Industry circuits as we change the metal layer sharing. The # MIVs plot shows how different settings impact the MIV count.

saving can be tuned better when certain nets are targeted based on the design.

Takeaway: Overall, we see that metal layer sharing can allow for efficient usage of metals. By using 12 metal layers for logic-on-logic 3-D ICs, we were able to reduce the 3-D stack to 11 metals with the help of logic-on-memory partitioning. Finally, the metal layer sharing was helpful to effectively use the memory layers and create further cost savings. This also came with an added PDP benefit, which was presented in Table VI. Depending on the type of design, the improvements to clock network or the overall design, in general, were key to generating the PDP benefit with metal sharing.

VI. Conclusion

In this work, we have first analyzed the routing in various 3-D IC types and found that metal layer sharing in 3-D ICs is a novel phenomenon, which can be controlled or enhanced in the designs based on user input. This layer sharing is only meaningful for the monolithic 3-D ICs due to the fine pitch and F2B nature of sequential fabrication. While metal sharing uses a large amount of MIVs, we see that they do not cause any routing issues in the design. Rather, they are helpful in effectively using the metal layers with a high-quality routing using one fewer BEOL layer resulting in cost savings for 3-D ICs. Even with the dropped metal layer, the timing and/or power consumption of the design improved with a power delay product improvement of 5%–7% across the three commercial processor designs.

References

[1] (2018). *International Roadmap for Devices and Systems*. [Online]. Available: https://irds.ieee.org/images/files/pdf/2018/2018IRDS_ES.pdf

[2] D. B. Ingerly *et al.*, “Foveros: 3D integration and the use of face-to-face chip stacking for logic devices,” in *IEDM Tech. Dig.*, Dec. 2019, p. 19.

[3] (2018). *AMD 3D V-Cache Ryzen*. [Online]. Available: https://www.anandtech.com/show/16725

[4] E. Beyne, “The 3-D interconnect technology landscape,” *IEEE Des. Test*, vol. 33, no. 3, pp. 8–20, Mar. 2016.

[5] Y. H. Chen *et al.*, “Ultra high density SoIC with sub-micron bond pitch,” in *Proc. IEEE 70th Electron. Compon. Technol. Conf. (ECTC)*, Jun. 2020, pp. 576–581.

[6] L. Brunet *et al.*, “Record performance of 500 °C low-temperature nMOSFETs for 3D sequential integration using a smart cut layer transfer module,” in *Proc. Symp. VLSI Technol.*, Jun. 2021, pp. 1–2.

[7] J.-T. Yan, Y.-J. Tseng, and C.-H. Yen, “Efficient micro-bump assignment for RDL routing in 3DICs,” in *Proc. 21st IEEE Int. Conf. Electron., Circuits Syst. (ICECS)*, Dec. 2014, pp. 195–198.

[8] S. Panth, K. Samadi, Y. Du, and S. K. Lim, “Placement-driven partitioning for congestion mitigation in monolithic 3D IC designs,” *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 34, no. 4, pp. 540–553, Apr. 2015.

[9] J. Shi, M. Li, S. Khasanvis, M. Rahman, and C. A. Moritz, “Routability in 3D ic design: Monolithic 3D vs. skybridge 3D CMOS,” in *Proc. IEEE/ACM Int. Symp. Nanosc. Architectures (NANOARCH)*, Jul. 2016, pp. 145–150.

[10] L. Bamberg, A. Garcia-Ortiz, L. Zhu, S. Pentapati, D. E. Shim, and S. K. Lim, “Macro-3D: A physical design methodology for face-to-face-stacked heterogeneous 3D ICs,” in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2020, pp. 37–42.

[11] S. Pentapati *et al.*, “Pin-3D: A physical synthesis and post-layout optimization flow for heterogeneous monolithic 3D ICs,” in *Proc. IEEE/ACM Int. Conf. Comput. Aided Design (ICCAD)*, Nov. 2020, pp. 1–9.

[12] S. Panth, K. Samadi, Y. Du, and S. K. Lim, “Shrunk-2-D: A physical design methodology to build commercial-quality monolithic 3-D ICs,” *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 36, no. 10, pp. 1716–1724, Oct. 2017.

[13] B. W. Ku, K. Chang, and S. K. Lim, “Compact-2D: A physical design methodology to build two-tier gate-level 3-D ICs,” *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 39, no. 6, pp. 1151–1164, Jun. 2019.

Authorized licensed use limited to: Georgia Institute of Technology. Downloaded on December 11,2022 at 00:39:06 UTC from IEEE Xplore. Restrictions apply.