Power, Performance, Area, and Cost Analysis of Face-to-Face-Bonded 3-D ICs

Anthony Agnesina, Moritz Brunion, Jinwoo Kim, Alberto Garcia-Ortiz, *Senior Member, IEEE*, Dragomir Milojevic, Francky Catthoor, *Fellow, IEEE*, Gioele Mirabelli, Manu Komalan, *Member, IEEE*, and Sung Kyu Lim, *Fellow, IEEE*

***Abstract*— This article presents a power, performance, area, and cost (PPAC) analysis for large-scale 3-D processor designs based on face-to-face wafer-to-wafer (W2W) and collective die-to-wafer (Co-D2W) bonding technologies. From evaluating our cost model on a comprehensive portfolio of technology nodes, we investigate a typically disregarded opportunity in 3-D through diverse computer-age statistical methods: area savings due to buffer savings and better routability, offering unforeseen, and considerable cost savings. We explore the viability of this factor with the feedback of a state-of-the-art 3-D memory-on-logic implementation flow. We show how this affects the PPAC of full-chip GDS implementations of a large-scale manycore processor design. Experiments show that our memory-on-logic 3-D implementation offers 7% silicon area savings, resulting in a 53.5% footprint reduction. We also obtained a 40% power-performance-cost improvement compared with the 2-D counterparts.**

***Index Terms*— 3-D integration, collective die-to-wafer (Co-D2W) bonding, cost, statistical analysis, wafer-to-wafer (W2W) bonding.**

NOMENCLATURE

A. W2W and Co-D2W

Symbol	Definition
$r_.$	3-D versus 2-D ratio of (die cost/wafer cost/DPW/yield)
F	decomposition of die-cost ratio into $(r_W, r_{\text{DPW}}, r_Y)$
C_d	total cost of a single die (= a stack for 3-D)
kgd	# known good dies
kgs	# known good stacks
dpw	# dies per wafer
y	yield
C_W	wafer cost

B. Co-D2W Only

$t_{r/nr}$	wafer type: reconstituted or not reconstituted
$k_{r/nr}$	# fabricated wafers per wafer type
n_r	# dies for wafer reconstitution

Manuscript received 16 December 2022; revised 19 March 2023; accepted 26 March 2023. Date of publication 5 April 2023; date of current version 4 May 2023. Recommended for publication by Associate Editor S. Lin upon evaluation of reviewers' comments. *(Corresponding author: Anthony Agnesina.)*

Anthony Agnesina, Jinwoo Kim, and Sung Kyu Lim are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: agnesina@gatech.edu; jinwookim@gatech.edu; limsk@ece.gatech.edu).

Moritz Brunion and Alberto Garcia-Ortiz are with the Institute of Electrodynamics and Microelectronics, University of Bremen, 28359 Bremen, Germany (e-mail: moritz.brunion@imec.be; agarcia@item.uni-bremen.de).

Dragomir Milojevic, Francky Catthoor, Gioele Mirabelli, and Manu Komalan are with IMEC, 3001 Leuven, Belgium (e-mail: dragomir.milojevic.ext@imec.be; francky.catthoor@imec.be; gioele.mirabelli@imec.be; manu.perumkunnil@imec.be).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCPMT.2023.3264520.

Digital Object Identifier 10.1109/TCPMT.2023.3264520

I. INTRODUCTION

THE semiconductor industry innovates new techniques to prolong Moore's Law as CMOS transistor geometries approach physical limits resorting to novel devices or extreme ultraviolet (EUV) lithography to reduce mask count. However, it often discounts the benefit of 3-D vertical integration, believing that integration cost exceeds its economic benefit. This article proposes a high-level study of a generic cost model to capture the dominant trade-offs in 2-D and 3-D. Our analyses reveal that the cost is very susceptible to the newly explored 3-D area savings, potentially giving a novel perspective to alleviate traditional economic barriers to 3-D vertical integration.

We focus on wafer-to-wafer (W2W) and collective die-to-wafer (Co-D2W) hybrid-bonding technologies. Both stack two pre-fabricated wafers, align them, bond them face-to-face (F2F), and dice them to create multiple single two-tier dies interconnected by high-density metallic interconnections. However, compared with the W2W stacking, the Co-D2W stacking includes a pre-stacking stage where one of the two wafers is diced, tested, and reconstituted before the shared W2W bonding step. With 3-D integration, heterogeneous devices, and technologies (e.g., memory, logic, RF, analog, and sensors) can be optimized for cost and performance for each layer.

In this article, we propose a new analysis of the 3-D cost for F2F designs to reveal spots in the 3-D design space where cost savings emerge. We claim the contributions of this article, an extension of [1], are as follows.

1) We propose a novel, generic, and rigorous derivation of the 3-D cost in the W2W and Co-D2W stackings.
2) We develop a high-level cost model to capture the conceptual trade-offs among area, cost, and performance in

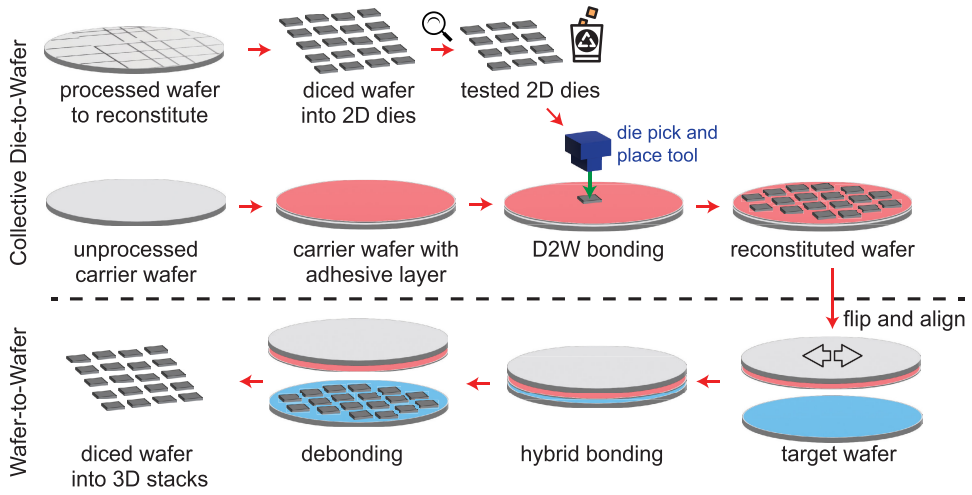


Fig. 1. Manufacturing flows for collective die-to-wafer (Co-D2W) and wafer-to-wafer (W2W) bonding processes. The W2W and Co-D2W share the last part of the process flow. Some standard steps are not shown (e.g., cleaning, plasma activation).

2-D and 3-D, independent of the technology nodes, F2F stacking methods, and precise knowledge of foundry parameters.

3) We develop a practical cost analysis and explore the significance of the defined parameters, revealing an unexpectedly significant impact of the often disregarded 3-D area savings factor γ.
4) We propose various modern statistical methods to study the cost and its complex relationship with the design and foundry parameters, including a novel explanatory approach based on machine learning to provide simple guidance to a designer and improve the cost of the 3-D design.
5) Thanks to the use of diverse cost parameters from industrial knowledge of technology nodes ranging from time-honored 28 nm to futuristic 3 nm, the conclusions of our analysis have broad relevance and applicability.
6) Our in-house memory-on-logic physical design flow experiments on an industry-representative benchmark of a large-scale manycore processor show that our proposed cost optimization axis of 3-D inherent area savings is viable. Due to the better exploitation in 3-D of the F2F-bonded back-end-of-line (BEOL) stack, we obtain improved power, performance, footprint, and silicon area results. Such area gains lead to significant cost savings. This PPA plus cost savings make 3-D-integrated circuits (ICs) more attractive than previous analysis has indicated.

The rest of this article is organized as follows. Section II presents the cost modeling's components, assumptions, and derivation. Section III mathematically analyzes the analytical cost formulations, while Section IV explains the statistical analysis. Section V demonstrates the experimental results. Finally, Section VI concludes this article.

II. Modeling

This section will review the F2F bonding processes considered in our study, explain the principal components and assumptions for the cost modeling, and derive the mathematical formulations of the die cost.

A. Bonding Processes

We consider the two mature F2F W2W and Co-D2W stacking methods. They differ mainly by introducing a pre-stacking step of die dicing, testing, and selection in the Co-D2W case, affecting yield and wafer cost [2]. The schematic of the two manufacturing flows is shown in Fig. 1.

1) Wafer-to-Wafer: The W2W hybrid bonding method is a renowned and established method for building 3-D ICs, where (heterogeneous) wafers are directly stacked and electrically connected at high throughput [3]. However, dies can only be tested after bonding, potentially causing yield degradation. Moreover, W2W requires matching die sizes, which prevents some cost-effective applications. For example, in a memory/logic 3-D partitioning scheme, the silicon area requirements of logic and memory are unlikely to match precisely. Co-designing the dies with independent footprint constraints would allow optimizing the implemented functionalities at the lowest silicon area cost.

2) Collective Die-to-Wafer: The Co-D2W technology is a more practical option that does not require matching die sizes but has a limited volume of production [4]. This technique relies on the collective die transfer by reconstituted carrier: after a die pick and place step that repopulates a wafer (diced and tested dies from a fabricated wafer are individually picked, aligned, and placed on the wafer, secured by a polymer), collective transfer bonding of the dies is done using standard W2W bonding equipment to transfer all the dies in a single process step. The collective nature is faster than the more traditional sequential die bonding and avoids the several bonding temperature cycles on the landing wafer [5]. Moreover, the known good die selection by testing dies pre-stack improves the yield after the 3-D assembly process. However, sub-micrometer accuracy issues for die alignment and additional cost of carrier preparation and purchase/use of the pick and place equipment exist.

TABLE I
COST MODELING PARAMETERS FOR 2-D AND 3-D W2W AND COLLECTIVE DIE-TO-WAFER (CO-D2W) STACKS

Parameter	Description
Foundry	
tn	technology node $\in$ Table II
stacking	hybrid bonding method $\in \{\mathrm{W2W}, \mathrm{D2W}\}$
w_d	wafer diameter
y_B	3D wafer bonding yield
$c_{B,\mathrm{W2W}}$	W2W integration cost
$y_{A,\mathrm{D2W}}$	D2W die alignment yield degradation
$c_{B,\mathrm{D2W}}$	D2W integration cost
f_T	testing fault coverage
c_T	testing cost of silicon per mm^2
D	defect density
θ	clustering defect parameter
Implementation	
a	2D die area (ref.)
γ (=$\gamma_{\mathrm{bot,D2W}}$)	2D vs. 3D bottom-die-area ratio
$\gamma_{\mathrm{top,D2W}}$	2D vs. 3D top-die-area ratio
N_{bot} (=N_{2D})	# BEOL metal layers of bottom 2D/3D die
N_{top}	# BEOL metal layers of top 3D die

TABLE II
TECHNOLOGY NODES CONSIDERED. THE COST VALUES OF THE FEOL, MOL, AND BEOL ARE SCALED TO THE N28 FEOL COST. CPP DENOTES THE CONTACTED POLY PITCH, AND MxP IS THE MINIMUM METAL PITCH

Technology (industry name)	Production Year	CPP (nm)	MxP (nm)
N28	2011	117	90
N20	2014	90	64
N16/14	2015	90	64
N10	2017	66	44
N7	2018	54	40
N5	2020	42	32
N3	2022+	42	21
N3-NSH	?	45	21

B. Parameters

Our cost model integrates two types of parameters, as presented in Table I. Foundry-related parameters are relative to a given technology node from a manufacturing foundry. The other type is implementation-dependent, relative to full-chip GDS designs obtained using a physical design flow.

Previous works on cost modeling for 3-D IC [6], [7], [8] do not consider the silicon area savings opportunities, assuming a fixed footprint area reduction of 50% in 3-D compared to 2-D. In sharp contrast, we introduce the 3-D area savings as an independent parameter γ. We introduce an area savings factor per die to model the absence of a requirement in matching die sizes for the Co-D2W case.

C. Assumptions

We use information from ITRS reports, previous literature [9], [10], and inside expertise validated by silicon measurements and industry feedback to set the foundry constants in Tables I and II. These values, including cost model components for the BEOL layers and 3-D integration technologies, have been used to construct, verify, and calibrate internal IC cost models.

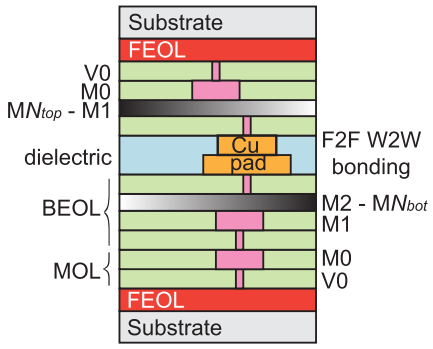


Fig. 2. Conceptual view (not to scale) of the 3-D stack for our wafer-to-wafer (W2W) and collective die-to-wafer (Co-D2W) face-to-face (F2F) wafer cost analysis.

When sensible, we abstract the costs or parameters of complex processes into single values (e.g., for stacking or testing), as it is unpractical and ungeneralizable to model manufacturing effects that depend heavily on foundry characteristics.

1) Technology Node: Table II shows the eight technology nodes considered, ranging from planar N28 to N3 FinFet and N3 Nanosheet [11]. The raw values of the cost components used in our analyses are normalized to the front-end-of-line (FEOL) cost of the N28 process. For every node, individual metal layer costs and stack configurations are defined. In contrast, the cost associated with the 3-D processing is constant when maintaining F2F pitch and wafer-handling assumptions, assuming the CMOS technology is independent of the 3-D bonding technology.

2) Metal Stack: Fig. 2 presents our exemplar stack, which includes a middle-of-line (MOL) to interconnect standard cells locally. Note that the MOL is not used in N28. We set for simplicity and without loss of generality that the number of metal layers used for routing in the BEOL of the bottom die is the same as the 2-D reference. Only N_{top} will differ in the BEOL comparison of 3-D versus 2-D.

3) Integration: The cost of 3-D integration includes wafer alignment, bonding, Si thinning, and via realization for both W2W and Co-D2W. The 3-D integration cost in Co-D2W also includes the extra wafer-processing steps, such as wafer dicing, carrier application, and wafer reconstitution. Capturing the difference in throughput between the two manufacturing processes is challenging. It depends on the number of processing, bonding, and testing machines available. As such, it is abstracted in the Co-D2W integration cost.

4) Yield: Since the grid of copper pads on the wafer is manufactured independently of the number of 3-D structures required to interconnect the design electrically, we abstract the nonperfect wafer-to-wafer assembly yield into a single factor y_B. Note that the alignment procedure in Co-D2W is very complex. On top of a global key for aligning wafers, it also requires a local alignment key for every die in the pick and place. This can cause a yield degradation abstracted as yield factor $y_{A,\mathrm{D2W}}$.

5) Testing: The testing cost for the Co-D2W depends on the automated test equipment (ATE) characteristics and testing time. It is abstracted into a cost per square millimeter c_T,

assuming the cost of testing the active dies pre-stacking is proportional to the die area. We do not consider the post-stacking testing cost, assumed to be a constant offset per design for all the integration scenarios. This is pessimistic for the Co-D2W stacking method, where one of the dies is already tested for wafer reconstitution. Finally, the introduction of additional test circuitry required for good die identification in Co-D2W affects the die area of the 3-D dies, which is an implementation-related input parameter and thus can inherently be considered during the die cost calculation.

6) Packaging: We do not consider the cost of packaging in our study, because the 2-D and 3-D packaging methods exhibit varied possibilities, which are very challenging to model. Nonetheless, the area and routing overhead due to the insertion of through-silicon vias (TSVs) for IOs is captured in the die area of the 3-D implementations through the intermediary of γ, and the additional routing resources required are encapsulated in the number of BEOL layers N.

D. Die-Cost Ratio

We build a parameterized and generic analytical cost model for 3-D ICs that handles both W2W and Co-D2W hybrid bonding processes. The cost model is written in Python [12], while its analyses are done in R [13] and Python.

For convenience, we write $f(x_1, x_2, \ldots, x_n)$ as $f(x_k)$ to highlight the dependence of function f on x_k. This way, we define $C_d(a) = C_d(a, \ldots)$ as the cost of a single die of area a, where the dependence to the other parameters of Table I is abstracted into an ellipsis $(\ldots)$. We focus on the high-level die-cost ratio between 2-D and 3-D to derive accurate cost optimization techniques

$$r_d(a,\gamma) \doteq \frac{C_{d_{3-D}}(\gamma a)}{C_{d_{2-D}}(a)} \overset{\text{some way}}{\sim} F\left\{\frac{C_{W_{3-D}}}{\text{kgd}_{3-D}(\gamma a)}, \frac{\text{kgd}_{2-D}(a)}{C_{W_{2-D}}}\right\} = F\left\{\frac{C_{W_{3-D}}}{C_{W_{2-D}}}, \frac{\text{dpw}(a)}{\text{dpw}(\gamma a)}, \frac{\text{y}_{2-D}(a)}{\text{y}_{3-D}(\gamma a)}\right\} \quad (1)$$

where kgd, dpw, y, and C_W denote the number of known good dies, the number of die-per-wafer (DPW), the yield, and the wafer cost, respectively. The nomenclature summarizes the symbols we will be using in the equations in addition to the main parameters in Table I. The implementation-dependent parameter γ models the shrinking of the die area in 3-D. We rewrite the formula to highlight the die cost trade-offs

$$\boxed{r_d(a,\gamma) = F\{r_W(a,\gamma), r_{\text{DPW}}(a,\gamma), r_Y(a,\gamma)\}} \quad (2)$$

where F is a functional $\mathcal{R}^3 \to \mathcal{R}$ of the three ratios: the wafer-cost ratio, the die-per-wafer ratio, and the yield ratio. The ratio r_W embodies the cost difference in manufacturing a 2-D die versus a 3-D die. The ratio r_{DPW} describes how many more 3-D dies we can obtain using 3-D integration and reducing the footprint a with the factor γ. Finally, the ratio r_Y compares the 3-D yield with the 2-D yield. We will show that F is the product map for W2W, while it is more complex for Co-D2W as a composition of product and linear mappings. Moreover, the dependency of r_W on (a, γ) will be shown to hold only in the Co-D2W case.

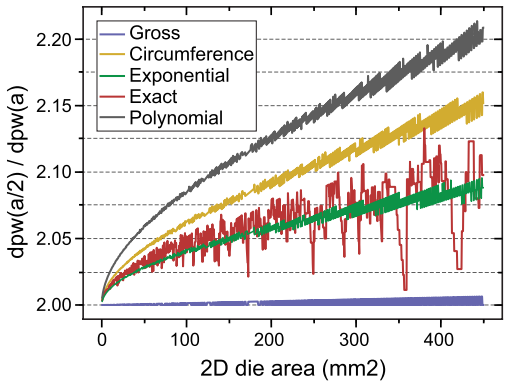


Fig. 3. Inverse die-per-wafer (DPW) ratio $r_{\text{DPW}}(a, 0.5)^{-1}$ for different analytical expressions and $w_d = 300$ mm. We observe that the Exponential model is the most accurate versus the Exact solution.

E. Components

We present classical approximations for each component of the cost model, such as die-per-wafer, yield, and wafer cost. In addition, we highlight their dependence on the area a, which is linearly modified by γ in the 3-D stacking.

1) Die-Per-Wafer: We select the most accurate analytical formula for the gross number of die-per-wafer. Fig. 3 compares various second-order approximations with the exact brute force solution presented in [14]. It shows that the exponential formula due to Ferris-Prabhu [15] is the most accurate

$$\text{dpw}(a) = \left\lfloor \left(\frac{\pi w_d^2}{4a}\right) e^{-2\sqrt{a}/w_d} \right\rfloor. \quad (3)$$

Notice that a larger die area a reduces the number of dies printed per wafer, and edge effects decrease the effective utilization of the total wafer area.

2) Yield: For 2-D integration, the total compound yield corresponds to the yield of a die. In contrast, the yield for 3-D W2W and Co-D2W ICs is more complicated and will be derived mathematically later. We use the negative binomial yield model [16] suggested in ITRS reports that considers the clustering of manufacturing defects rather than their independent occurrences, that is,

$$\Pr\{\#\text{defects on chip} = k\} = \frac{\Gamma(\theta + k)}{k!\Gamma(\theta)} \frac{(Da/\theta)^k}{(1 + Da/\theta)^{\theta+k}} \quad (4)$$

where $\Gamma(x)$ is the gamma function, D is the defect density obtained from the average number of defects per unit area, that is, the average number of defects on a chip of area a is $\mu = E\{\#\text{defects}\} = Da$, and θ is the clustering parameter of the faults, that is, $\text{Var}[\#\text{defects}] = \mu + \mu^2/\theta$. Parameter θ depends on the technology and design (e.g., mask steps), and small values indicate increased clustering. In practice, we use $\theta = 2$ to model slight nonuniformity of fault distribution. It is straightforward to extend the current framework to other more precise models, as those proprietary ones available inside the foundry companies. Finally, the die yield is obtained as

$$\text{y}(a) = \Pr\{\#\text{defects on chip} = 0\} = \left(1 + \frac{Da}{\theta}\right)^{-\theta}. \quad (5)$$

Thus, a larger active die area a reduces the processing yield exponentially.

TABLE III
OUTCOMES IN THE PRE-STACKING TEST OF THE DICED DIES, WHERE (+) AND (−) DENOTE THE DEFECTIVE AND FUNCTIONAL DIES, RESPECTIVELY. THE NUMBER OF DIES AFTER TEST n_r = TNEG + FNEG, f_T = TPOS / POS, y_r = NEG / (POS + NEG), k_r = POS + NEG

		Test Result		
		(+)	(−)	Total
Actual	(+)	TPos	**FNeg**	Pos
	(−)	FPos (= 0)	**TNeg**	Neg

3) Wafer Cost: The 2-D wafer cost is the sum of the individual costs of each layer

$$C_{W_{2-D}}(\mathrm{tn}, N) = c_{\mathrm{Substrate}}(\mathrm{tn}) + c_{\mathrm{FEOL}}(\mathrm{tn}) + c_{\mathrm{MOL}}(\mathrm{tn}) + c_{\mathrm{BEOL}}(\mathrm{tn}, N) \tag{6}$$

where $c_{\mathrm{BEOL}}(\mathrm{tn}, N) = \sum_{i=1}^{N} c_{M_i}(\mathrm{tn})$ and c_{M_i} is the cost of metal layer i in the BEOL whose configuration and individual metal layer values depend on the technology node tn. For the 3-D case, we add the integration cost to the wafer costs

$$C_{W_{3-D}}(\mathrm{tn}, N_1, N_2) = c_B + C_{W_{2-D}}(\mathrm{tn}, N_1) + C_{W_{2-D}}(\mathrm{tn}, N_2) \tag{7}$$

where c_B is, as highlighted, independent of the CMOS node chosen, but dependent on the minimum 3-D structure pitch, the 3-D structure density, and the stacking method.

4) 2-D Die Cost: The total die cost in 2-D is simply

$$\boxed{C_{d_{2-D}} = \frac{C_{W_{2-D}}}{\mathrm{kgd}_{2-D}(a)} = \frac{C_{W_{2-D}}}{\mathrm{y}(a)\,\mathrm{dpw}(a)}.} \tag{8}$$

F. 3-D Die Cost

We present a derivation of the 3-D stack "die" cost considering a two-tier stack of two dies. This can be generalized to stacks with more than two tiers by considering each wafer interface independently and aggregating the costs based on the bonding assumptions.

We first derive the cost for Co-D2W, where stacks can contain heterogeneous dies of different sizes. For Co-D2W, a wafer can be of reconstituted type or not ($t_{r/nr}$). For each type, we abbreviate $\mathrm{dpw}(\gamma a)|\,\mathrm{y}(\gamma a)$ by $\mathrm{dpw}\,|\,\mathrm{y}$ and let k denote the number of wafers. After testing and selecting the active dies passing the pre-stacking test, the number of available dies from wafers t_r for wafer reconstitution and stacking is

$$n_r = k_r\,\mathrm{dpw}_r\left[1 + f_T\left(\mathrm{y}_r - 1\right)\right] \tag{9}$$

where f_T is the test coverage. This equation can be proven from Table III describing the test true/false positive/negative rates, assuming the tests have a probability of false alarm of zero. This assumption aligns with traditional testing capabilities, where a die test can only fail by not recognizing an existing fault, but it cannot identify a fault when there is none.

The need to bond two wafers together where a die from wafer-type t_r is stacked with a die from t_{nr} yields the condition

$$\left\lfloor \frac{n_r}{\mathrm{dpw}_{nr}} \right\rfloor = k_{nr} \tag{10}$$

where the integer part can be neglected as a first approximation for high-volume manufacturing. The number of good stacks obtained after bonding is

$$\begin{aligned}\mathrm{kgs} &= \#\text{ stacks fabricated} \times \Pr\{\text{die } t_{nr} \text{ good}\} \times \Pr\{\text{die } t_r \text{ good}\} \times \Pr\{\text{stacking good}\} \\ &= k_{nr}\,\mathrm{dpw}_{nr} \times \mathrm{y}_{nr} \times \frac{\mathrm{y}_r}{1 + f_T\left(\mathrm{y}_r - 1\right)} \times y_A y_B\end{aligned} \tag{11}$$

since the yield of the reconstituted wafer is by definition $k_r\,\mathrm{dpw}_r\,\mathrm{y}_r\,/n_r$, and y_A is the yield loss due to the misalignment of the dies on the reconstituted wafers. Moreover, the total cost spent is

$$C_t = \underbrace{k_{nr} C_{W_{nr}}}_{\text{non-reconst. wafers}} + \underbrace{k_r C_{W_r}}_{\text{reconst. wafers}} + \underbrace{k_{nr} c_B}_{\text{3-D manufacturing}} + \underbrace{k_r\,\mathrm{dpw}_r\,c_T \gamma_r a}_{\text{testing dies for reconstitution}} \tag{12}$$

where c_B is the 3-D integration cost for two wafers and c_T is the cost of testing per mm^2, which is multiplied by the total area of the diced dies. Equations (9)–(12) give the 3-D die stack cost

$$\boxed{C_{d_{3-D,D2W}} = \frac{C_t}{\mathrm{kgs}} = \frac{1 + f_T(\mathrm{y}_r - 1)}{y_A y_B\,\mathrm{y}_{nr}\,\mathrm{y}_r} \cdot \left\{ \frac{C_{W_{nr}} + c_B}{\mathrm{dpw}_{nr}} + \frac{C_{W_r} + \mathrm{dpw}_r\,c_T \gamma_r a}{\mathrm{dpw}_r[1 + f_T(\mathrm{y}_r - 1)]} \right\}} \tag{13}$$

which is fully generic for Co-D2W. There is a clear trade-off of c_T versus f_T, as advanced tests increase the duration of the testing procedure and thus the testing cost, but will result in more reliable dies. The 3-D Co-D2W compound stack yield is then

$$y_{3-D} = \frac{\#\text{ good stacks}}{\#\text{ stacks fabricated}} = \frac{\mathrm{kgs}}{k_{nr}\,\mathrm{dpw}_{nr}} = \frac{y_A y_B\,\mathrm{y}_{nr}\,\mathrm{y}_r}{1 + f_T\left(\mathrm{y}_r - 1\right)} \tag{14}$$

which is improved when the testing is more accurate. This formula embodies the die yield cost loss from the faulty dies escaping the test, weighted by the testing accuracy.

In practice, we choose to reconstitute the wafer that yields the lower 3-D cost by evaluating (13) for both top and bottom dies. We will show that an approximate rule is to always dice the wafer where the dies are the smallest.

The W2W formula can be directly derived from the Co-D2W formulation, a reassuring fact and proof of the generality of the previous Co-D2W demonstration. Here, there is no wafer reconstitution, no testing of the dies, and dies have matching areas, thus $c_T \leftrightarrow 0$, $f_T \leftrightarrow 0$, $y_A \leftrightarrow 1$, $\gamma_{nr} = \gamma_r \leftrightarrow \gamma$, and the formula simplifies as

$$\boxed{C_{d_{3-D,W2W}} = \frac{C_{W_{\mathrm{bot}}} + C_{W_{\mathrm{top}}} + c_B}{y_B \mathrm{y}(\gamma a)^2\,\mathrm{dpw}(\gamma a)}} \tag{15}$$

which results in a formal 3-D W2W die stack yield of $y_B \mathrm{y}^2$. The square dependency is due to stacking faulty dies with good dies from the two wafers. This compound yield is much worse than in (14) due to the die yield exponential trend, as long as

$f_T > (1 - y_A)/(1 - y)$, a condition most likely verified in reasonable scenarios.

III. Analytical Analysis

This section establishes cost optimization priorities robust to changes in the foundry parameters, relying on the mathematical analysis of the analytical formulas. We focus on the die cost ratio behavior relative to the deeply linked parameter γ and area a. The discussions of W2W and Co-D2W are separated due to the differences in their 3-D cost formulations.

A. Wafer-to-Wafer

For W2W, (15) gives a die cost ratio in the conceptual decomposition of (2) where F is the product

$$\frac{\mathrm{dpw}(a)}{\mathrm{dpw}(\gamma a)} \frac{\mathrm{y}(a)}{y_B \mathrm{y}(\gamma a)^2} \frac{C_{W_{\mathrm{bot}}} + C_{W_{\mathrm{top}}} + c_B}{C_{W_{2-D}}} \tag{16}$$

which simplifies its study to the independent study of each ratio.

1) Die-Per-Wafer Ratio: The floor function in (3) can be neglected to obtain

$$r_{\mathrm{DPW}}(a, \gamma) = \frac{\mathrm{dpw}(a)}{\mathrm{dpw}(\gamma a)} \approx \gamma e^{-2\sqrt{a}(1-\sqrt{\gamma})/w_d} \tag{17}$$

from which three direct conclusions can be made as follows.

1) Increasing the wafer diameter w_d, a typical trend with the advancements of fabs, increases the DPW ratio.
2) A larger die area reduces the DPW ratio.
3) The value of γ bounds the DPW ratio, and its reduction favors 3-D.

To gain insight into the driving forces of r_{DPW}, we compute the growth rates for γ and a

$$\frac{\left|\frac{\partial r_{\mathrm{DPW}}(a,\gamma)}{\partial \gamma}\right| \Delta\gamma}{\left|\frac{\partial r_{\mathrm{DPW}}(a,\gamma)}{\partial a}\right| \Delta a} = \left(\frac{\Delta\gamma}{\gamma}\right)\left(\frac{a}{\Delta a}\right)\frac{\sqrt{\gamma}}{1-\sqrt{\gamma}}\left(1 + \frac{w_d}{\sqrt{\gamma a}}\right) \gg 1. \tag{18}$$

As a result, γ has a higher effect on the DPW ratio than the die area a alone.

2) Yield Ratio: The formal stack yield is the product of the yield of the top and bottom dies, which are assumed identical. Thus

$$r_Y(a, \gamma) = \frac{\mathrm{y}(a)}{y_B \mathrm{y}(\gamma a)^2} \propto \frac{\left(1 + \frac{Da}{\theta}\right)^{-\theta}}{\left(1 + \frac{D\gamma a}{\theta}\right)^{-2\theta}} \tag{19}$$

from which we see that

$$\frac{\partial}{\partial \gamma} r_Y(a, \gamma) \geq 0 \tag{20}$$

always holds, revealing a monotonic relationship between γ and the yield ratio, where decreasing γ always favors 3-D.

Next, the implication of the area on the yield ratio is discussed. We have

$$\mathrm{sgn}\left(\frac{\partial}{\partial a} r_Y(a, \gamma)\right) = \mathrm{sgn}(aD\gamma + \theta(2\gamma - 1)). \tag{21}$$

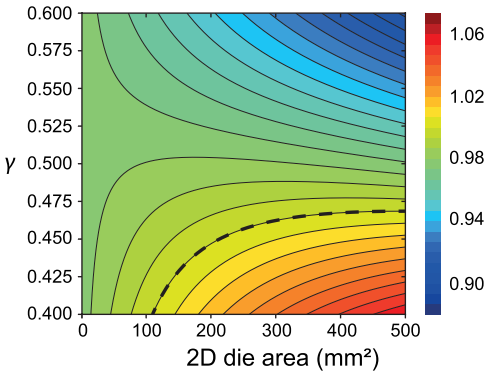


Fig. 4. Inverse yield ratio $r_Y(a, \gamma)^{-1}$ in the wafer-to-wafer (W2W) case: hotter colors indicate improved 3-D yield versus 2-D. The dashed contour line corresponds to a ratio of 1, under which the yield is favorable to 3-D. A γ below 0.5 positively impacts the 3-D yield.

If $\gamma > 0.5$, then this is > 0 always, and the area therefore worsens the yield ratio for 3-D. If $\gamma < 0.5$, there is a regime where the yield ratio decreases when a increases

$$a \leq a_{\mathrm{Cutoff}} = \frac{\theta(1 - 2\gamma)}{D\gamma} \tag{22}$$

until it reaches a minimum, after which it increases onward. Therefore, if γ is small enough, an area increase will always be positive for 3-D. We summarize the fundamental yield trade-offs with the graph shown in Fig. 4.

B. Collective Die-to-Wafer

The die-cost ratio in the Co-D2W can be written concisely as

$$\mathcal{A}\left\{\frac{\left[1 + f_T(\mathrm{y}(\gamma_r a) - 1)\right]}{\mathrm{y}(\gamma_r a)} \frac{\mathrm{y}(a)}{\mathrm{y}(\gamma_{nr} a)}\right\}\left[\frac{\mathrm{dpw}(a)}{\mathrm{dpw}(\gamma_{nr} a)}\right] \left\{\mathcal{B} + \left[\frac{\mathrm{dpw}(\gamma_{nr} a)}{\mathrm{dpw}(\gamma_r a)}\right]\frac{C_{W_r} + \mathrm{dpw}(\gamma_r a)\gamma_r a c_T}{\left[1 + f_T(\mathrm{y}(\gamma_r a) - 1)\right]}\right\} \tag{23}$$

where $\mathcal{A} = 1/(C_{W_{2-D}} y_A y_B)$, and $\mathcal{B} = C_{W_{nr}} + c_B$. Here, F is much more complicated than for W2W. It can, however, still be decomposed as a set of linear sums and multiplications of the three defined ratios.

1) Die-Per-Wafer Ratio: The DPW intervenes thrice in (23). First is the competition of 2-D versus 3-D, second is that of nonreconstituted versus reconstituted die, and third is for the reconstituted die only. Considering only the DPW, r_d is small when

$$\begin{cases} \dfrac{\mathrm{dpw}(a)}{\mathrm{dpw}(\gamma_{nr} a)} \approx \gamma_{nr} e^{-2\sqrt{a}/w_d(1-\sqrt{\gamma_{nr}})} \\ \dfrac{\mathrm{dpw}(\gamma_{nr} a)}{\mathrm{dpw}(\gamma_r a)} \approx \dfrac{\gamma_r}{\gamma_{nr}} e^{-2\sqrt{a}/w_d(\sqrt{\gamma_{nr}}-\sqrt{\gamma_r})} \\ \mathrm{dpw}(\gamma_r a)\gamma_r a \approx \left(\pi w_d^2\right) e^{-2\sqrt{\gamma_r a}/w_d} \end{cases} \tag{24}$$

are all small, which is satisfied when γ_{nr} is small, $\gamma_r < \gamma_{nr}$ (i.e., the reconstituted die is the smallest), and γ_r is small. Moreover, a larger area a is better for 3-D for all cases when $\gamma_r < \gamma_{nr}$.

2) Yield Ratio: The yield appears twice in (23). Considering only the yield, r_d is small when

$$\begin{cases} \dfrac{\mathrm{y}(a)}{\mathrm{y}(\gamma_{nr}a)} = \dfrac{\left(1+\frac{Da}{\theta}\right)^{-\theta}}{\left(1+\frac{D\gamma_{nr}a}{\theta}\right)^{-\theta}} \\ \dfrac{1+f_T(\mathrm{y}(\gamma_r a)-1)}{\mathrm{y}(\gamma_r a)} \text{ and } \dfrac{1}{1+f_T(\mathrm{y}(\gamma_r a)-1)} \end{cases} \tag{25}$$

are all small, which is satisfied when γ_{nr} and γ_r are small (one can check derivatives with respect to γ_{nr} and γ_r are positive). Also, a larger a diminishes the first ratio. Still, it increases the last two ratios in a trend opposite to the DPW. In contrast with W2W, this case makes it too difficult to estimate the combined effect of a analytically; hence, selecting optimal parameter settings in a simple analytical way is also infeasible. This is one of the motivations for our next section.

C. Wafer-Cost Ratio

A simple cost improvement is obtained by reducing the total metal layer count of the stack. Assuming a similar metal pitch, a similar floorplan aspect ratio, and a γ of 0.5, the available length of the routing tracks of one metal layer in 2-D is equal to that of two layers in 3-D. Therefore, the metal layer count can be adjusted to the routing resource requirements of a design more accurately in 3-D, resulting in a more optimized cost. This granularity also enables a second gain from the more diverse BEOL configurations available for the top or bottom dies. Finally, a third gain comes from the empirical wire length reduction in 3-D, estimated grossly as a maximum half-perimeter wire length reduction of $1/\sqrt{\gamma}$. Still, more complex and accurate models are available in [7].

In the W2W stacking, the wafer-cost ratio r_W is independent of the die area and γ, which is not the case in the Co-D2W due to the testing cost. For a W2W balanced stack, we get

$$r_W = \frac{C_{W_{\text{bot}}} + C_{W_{\text{top}}} + c_B}{C_{W_{2-D}}} = 2 + \frac{c_B}{C_{W_{2-D}}} \geq 2 \tag{26}$$

which is minimized for taller stacks and very aggressive BEOL configurations. This statement also stands for Co-D2W, but deriving it is more challenging. While this reduces the wafer cost ratio, this effect is not a good optimization. In the imbalanced case, 3-D tiers with fewer metal layers than the 2-D counterpart yield additional cost-saving opportunities.

IV. Statistical Analysis

The analytical analysis highlighted general trends of the die-cost ratio based on γ and the area. This section studies the die-cost ratio instead through the statistical lens. Because many parameters intervene in its nonlinear form [see Table I and (13), (15)], traditional analytical or geometrical analyses of r_d are challenging, especially when considering the variability in the parameters due to the variety of foundries and uncertainty of measurements.

We first present how to produce good query samples of the die-cost ratio, $\{(\boldsymbol{x}_i, r_d(\boldsymbol{x}_i))\}$, where $\boldsymbol{x}$ is the vector of parameters of dimension p, with $p = 9$ for W2W and $p = 13$ for Co-D2W. The considered domain of r_d is presented in Table IV, where parameter ranges have been carefully designed to cover a broad but realistic catalog of scenarios. These samples are then used for a sensitivity analysis to understand which parameter variations are fundamental in the 2-D versus 3-D cost trade-off. We also train an interpretable machine-learning model with these samples, enabling a human-readable high-level evaluation that can be directly fed into an exploration or optimization engine. Finally, we study the evolution of the die-cost ratio with the technology node and propose strategies to improve the cost of 3-D ICs.

TABLE IV
Parameters' Variation Ranges for Our Statistical Analysis of the Die-Cost Ratio r_d. Values Are Set to Cover a Wide Range of Scenarios (Transistor, Design, and 3-D Pitch)

Parameter	Variation Range
technology	$\{\mathrm{N28}, \ldots, \mathrm{N3}\}$
stacking	$\{\mathrm{W2W}, \mathrm{D2W}\}$
w_d (mm)	[100, 400]
a (mm^2)	[10, 500]
γ, $\gamma_{\text{top,D2W}}$	[0.45, 0.55]
N_{bot}, N_{top}	[10, 20]
y_{B}	[0.90, 1.00]
$c_{\text{B,W2W}}$	[0.20, 0.30]
$y_{\text{A,D2W}}$	[0.98, 1.00]
$c_{\text{B,D2W}}$	[0.20, 0.40]
f_{T}	[0.95, 1.00]
c_{T} (mm^{-2})(e^{-6})	[6.15, 13.5]
D (mm^{-2})	[0.0005, 0.005]
θ	[1, 10]

A. Sobol's Sequence

We use Sobol's low-discrepancy sequence to sample the die-cost ratio. For large sample sizes and functions with unknown topologies, Sobol's sequences exhibit much better characteristics (fast computation, sampling uniformity, and sampling efficiency) than other popular methods, such as the Latin Hypercube Sampling [17]. Sobol's sequence simulates the uniform distribution on the p-dimensional cube. We generate it based on digital nets, which can be constructed efficiently through matrix multiplications [18]. The random uniformity of the sequence has two crucial implications for our practical applications.

1) The point sets are spread evenly and avoid clusters and gaps, increasing the sensitivity analyses' convergence rate.
2) The point sets cover the entire parameter space evenly, which helps the machine-learning model generalize to unseen inputs in any subset of the domain through local properties [19].

Sobol's sequence produces independent and identically distributed parameters. However, actual foundry and design parameters are usually correlated; testing cost and fault coverage are positively correlated; yield-related parameters (e.g., defect density, bonding yield, and alignment yield) correlate by originating from the same fab. However, finding such intrinsic correlations is only attainable with an extensive database of foundry parameters and design implementations. Because such

TABLE V

LOCAL MORRIS SENSITIVITY ANALYSIS OF THE DIE-COST RATIO r_d FOR W2W AND CO-D2W. THE ABSOLUTE MEAN $\mu^\star$ AND VARIANCE σ OF THE ELEMENTARY EFFECTS ARE FIRST AVERAGED OVER ALL TECHNOLOGIES, THEN SORTED AND REPORTED AS NORMALIZED %. (a) W2W. (b) Co-D2W

(a)

p	$\mu^\star$	σ
γ	41.3	26.2
N_{bot}	11.1	4.5
N_{top}	10.5	3.9
a	10.3	25.0
D	9.6	22.4
y_B	8.1	2.9
w_d	4.6	5.7
θ	3.8	9.2
$c_{\text{B,W2W}}$	0.8	0.3

(b)

p	$\mu^\star$	σ
$\gamma_{\text{top,D2W}}$	16.1	10.1
γ	16.1	10.1
a	13.4	20.4
w_d	11.5	8.2
N_{bot}	9.9	5.6
N_{top}	8.5	4.4
D	7.9	16.5
y_B	7.0	3.1
θ	3.9	14.1
c_T	2.6	5.1
$y_{A,\text{D2W}}$	1.3	0.6
$c_{\text{B,D2W}}$	1.1	0.8
f_T	0.5	1.1

TABLE VI

GLOBAL SOBOL SENSITIVITY ANALYSIS OF THE DIE-COST RATIO r_d FOR W2W AND CO-D2W. THE CLOSED τ^2 AND TOTAL ρ^2 SENSITIVITY INDICES ARE FIRST AVERAGED OVER ALL TECHNOLOGIES, THEN SORTED AND REPORTED AS NORMALIZED %. (a) W2W. (b) Co-D2W

(a)

p	τ^2	ρ^2
γ	60.3	53.5
N_{bot}	12.6	10.3
N_{top}	11.1	9.1
y_B	5.9	4.8
D	4.4	9.0
w_d	2.4	2.4
a	1.7	7.9
θ	1.6	2.9
$c_{\text{B,W2W}}$	0.05	0.04

(b)

p	τ^2	ρ^2
$\gamma_{\text{top,D2W}}$	22.2	20.9
γ	22.0	20.8
a	16.4	17.4
w_d	11.3	11.0
N_{bot}	9.4	8.9
N_{top}	6.8	6.3
D	5.2	6.8
y_B	4.1	3.8
θ	1.6	2.9
c_T	0.7	0.8
$y_{A,\text{D2W}}$	0.2	0.2
$c_{\text{B,D2W}}$	0.08	0.01
f_T	0.03	0.03

a database is currently unavailable, we conduct our study on a purely abstract view of the model, where the effects of the potential dependencies of input parameters are not captured separately. Instead, we focus on understanding the high-level mapping of inputs to the model's output, where variables are seen as independent factors.

B. Sensitivity Analyses

We present local and global sensitivity analyses to study the die-cost ratio. While differing in how they look for and establish significant individual predictors, both methods display the unchallenged impact of γ on the die-cost ratio, irrespective of the chosen technology node and stacking process.

1) Local: We use the Method of Morris [20] to calculate the elementary effects of each parameter one at a time in a discretized way, providing a pointwise view of partial derivatives. The Sobol sampling helps build an accurate finite distribution of elementary effects for each parameter. From there, we define $\mu^\star$ as the absolute mean of the distribution and σ as its standard deviation.

The results averaged over all technologies are shown in Table V, where larger $\mu^\star$ values imply a general local influence on the resulting die-cost ratio r_d. This highlights the extreme sensitivity of r_d to the γ parameters explicitly for all technology nodes and stacking methods. Notice in Table V(b) that the total effect of area savings is

$$\left(\mu^\star_\gamma + \mu^\star_{\gamma_{\text{top,D2W}}} = 32.2, \sqrt{\sigma^2_\gamma + \sigma^2_{\gamma_{\text{top,D2W}}}} = 10.1\right) \tag{27}$$

surpassing the effects of other parameters. The large σ values for a and D show that r_d is highly affected by these parameters in some regions of the parameter space. In contrast, the local sensitivity to f_T and c_B is negligibly small. Moreover, the commonly acknowledged cost optimization option of metal layer reduction is shown to have a high impact on the overall cost in W2W but significantly less than γ.

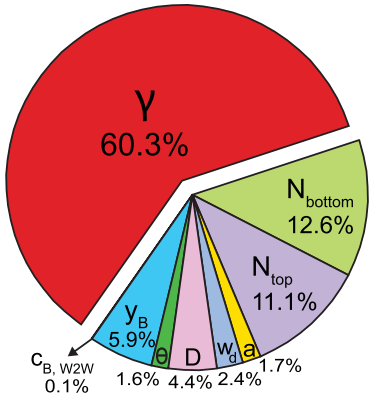


Fig. 5. Pie chart of the closed sensitivity index contribution for global sensitivity analysis of each cost parameter on the die-cost ratio r_d for wafer-to-wafer (W2W). The values are averaged over the technologies of Table II.

2) Global: We use the analysis of variance (ANOVA) [21] for the global analysis, decomposing r_d into a sum of independent factors

$$r_d(\boldsymbol{x}) \equiv \sum_{u \subseteq \{1,\ldots,p\}} r_d(\boldsymbol{x}_u), \quad \text{where} \quad \boldsymbol{x}_u = \left(x_j\right)_{j \in u}. \tag{28}$$

The quasi-Monte Carlo method based on the Sobol's sequence is used to estimate the factors through integral methods [22] and obtain $\sigma^2_u = \text{Var}[r_d(\boldsymbol{x}_u)]$. From there are defined the two Sobol's sensitivity indices

$$\tau^2_u = \sum_{v \subseteq u} \sigma^2_v \quad \text{(closed)} \quad \text{and} \quad \rho^2_u = \sum_{v : v \cap u \neq \varnothing} \sigma^2_v \quad \text{(total)} \tag{29}$$

respectively quantifying the proportion of variance of r_d attributable to subsets of u, and quantifying the proportion of variance of r_d attributable to subsets intersecting u.

Table VI presents the global sensitivity indices averaged over all technologies. The pie chart in Fig. 5 summarizes the individual contribution of each parameter to the variance of r_d in the W2W stacking. These results reinforce the previous

observation that γ is the most influential parameter, impacting the die-cost ratio more than all the other parameters combined. Interestingly, the effect of γ is more important in the W2W case, which can be attributed to the square die yield with an exponential dependency of γa.

C. Rules Extraction

While sensitivity analyses carry valuable insights on the die-cost ratio, they fail to guide the designer in improving the 3-D design, i.e., which parameter to modify first, in which direction, and by how much. To answer such questions, we build an interpretable, stable, and accurate machine-learning regression of the die-cost ratio. We use the SIRUS model (Stable and Interpretable Rule Set) [23], which simplifies random forests with high predictivity power toward a rule-based linear structure with high interpretability attributes. This rule-based model is intuitive enough to help designers identify the fundamental parameters in their 3-D design and give recommendations to improve it cost-wise.

The SIRUS algorithm first grows a random forest following Breiman's original algorithm [24] but restricts the depth and number of node splits. Then, the large collection of forest decisions/paths, aka rules, are pruned, simplified, and merged to conserve the forest's accuracy and stability. The remaining rules are then aggregated with ridge regression so that the model output is an estimate of the die-cost ratio

$$\hat{r}_d(\boldsymbol{x}) = \widehat{E}\{r_d|\boldsymbol{x}\} = \sum_{r\in\mathcal{S}} g_r \text{rule}_r(\boldsymbol{x}) \tag{30}$$

where $g_r > 0$, and the sum is taken over the set of all rules $\mathcal{S}$.

We define three metrics to judge the quality of the trained machine-learning model. First, the error is computed as the unexplained variance

$$\widehat{\text{err}} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \quad \text{where } \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{31}$$

and $\hat{y}_i = \hat{r}_d(\boldsymbol{x}_i)$ is the model prediction. Next, the Sorensen–Dice coefficient is used to estimate the stability of the trained model. Given two rule sets obtained from two runs of the SIRUS algorithm fit on independent samples, R_0 and R_1

$$\widehat{\text{stab}} = \frac{2|R_0 \cap R_1|}{|R_0| + |R_1|}. \tag{32}$$

Thus, stability is the average proportion of rules shared by two SIRUS models fit on two distinct parts of the data. Finally, we use the number of rules as the simplicity metric, indirectly set by β, the selection threshold on the frequency of appearance of a rule in the forest; a rule r is selected when $\sum_{t\in\text{trees}} \mathbf{1}[r \in t]/\#\text{trees} \geq \beta$.

We perform cross-validation of the algorithm on the training data drawn without replacement from Sobol's sequence (2.5M samples), yielding the curves of Fig. 6. To realize the Pareto optimum of our three desired metrics, we select β_0 computed from the two curves as

$$\beta_0 = \underset{\beta>0}{\text{argmin}} \left(\widehat{\text{stab}}_\beta - 0.9\right)^2 + \widehat{\text{err}}_\beta^2. \tag{33}$$

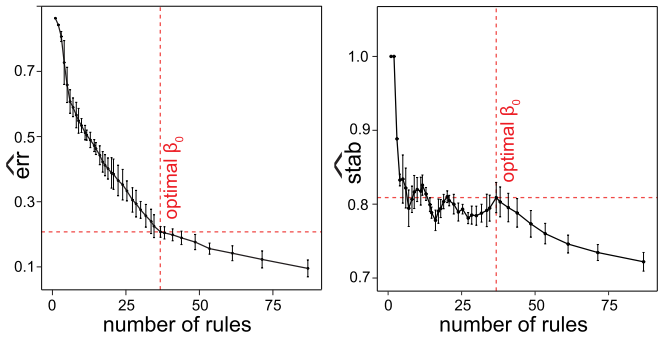


Fig. 6. Pareto front β_0 of unexplained variance ($\widehat{\text{err}}$) versus stability ($\widehat{\text{stab}}$) versus the number of rules, estimated via 10-fold cross-validation.

TABLE VII

EXTRACTED RULES TO PREDICT THE DIE-COST RATIO r_d. TO GENERATE A PREDICTION FOR A GIVEN PARAMETER SET $\boldsymbol{x}$, FOR EACH RULE, THE CORRESPONDING $\hat{r}_d$ IS RETRIEVED DEPENDING ON WHETHER $\boldsymbol{x}$ SATISFIES THE RULE CONDITIONS. THEN ALL RULE OUTPUTS FOR $\boldsymbol{x}$ ARE MULTIPLIED BY THEIR ASSOCIATED WEIGHT AND SUMMED

Weight	Rule (**if** condition **then** $\hat{r}_d =$ **else** $\hat{r}_d =$)
0.29	**if** $\gamma \geq 0.5$ & stack is W2W **then** 1.2 **else** 1
0.28	**if** $a \geq 160$ & stack is Co-D2W **then** 0.99 **else** 1.1
0.25	**if** $\gamma < 0.48$ **then** 0.96 **else** 1.1
0.21	**if** $\gamma > 0.5$ & $N_{\text{bot}} \geq 15$ **then** 1.2 **else** 1
0.19	**if** $y_B \geq 0.94$ & $N_{\text{top}} < 16$ **then** 1 **else** 1.1
0.16	**if** $\gamma_{\text{top,D2W}} \leq 0.5$ & stack is Co-D2W **then** 0.97 **else** 1.2
...	...

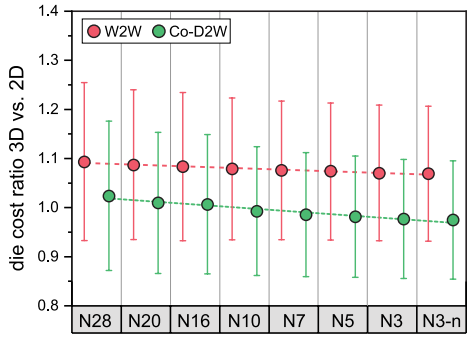


Fig. 7. Evolution of the mean and variation of the die-cost ratio r_d per transistor technology and 3-D face-to-face (F2F) stacking method, obtained from the uniform sampling of parameters defined in the ranges in Table IV.

Using β_0, the model is trained and then tested on separate data (500 K samples). The test metrics of $\widehat{\text{err}} = 0.14$, $\widehat{\text{stab}} = 0.95$, and # rules = 32 highlight a satisfactory trade-off of accuracy, stability, and simplicity. We present in Table VII the top few rules. As each tree in the original forest has a maximum depth of two, rules have at most two clauses. Examining the predominant rules indicates the importance of the (two) γ factor(s) and the preferred use of the Co-D2W stacking. Furthermore, the rules frequently include the metal layer count, confirming the trends observed in Tables V and VI.

D. Technology Impact

We plot in Fig. 7 the statistics gathered per technology and stacking method from the uniform sampling of r_d. We observe that as follows.

TABLE VIII
3-D IC COST OPTIMIZATION STRATEGIES PROPOSED AND VALIDATED IN OUR EXPERIMENTS

Strategy	Rationale
footprint reduction beyond 50%	3D IC enables less buffering and better routability to provide total silicon area reduction
reduction of metal layer count	3D IC offers wire length savings plus sharing of metal layers across the tiers

1) The Co-D2W method appears more advantageous cost-wise for 3-D stacking.
2) The cost competition of 3-D versus 2-D benefits from advancing transistor scaling in both W2W and Co-D2W, with a faster pace for Co-D2W.

However, considering the significant variations, there is no guarantee that technology scaling will inherently render the cost competition favorable to 3-D. This information is crucial to emphasize the importance of the *design* part in the overall VLSI flow for cost optimization, i.e., in the proper choice of the 3-D stacking type and especially the 3-D partitioning strategy (which can affect γ immensely), as well as 3-D CAD physical design methodology. Some research is underway to address both issues [25], [26].

E. Strategies

From examining the sensitivity analyses and extracted rules, we propose two simple but effective optimization strategies for a designer in 3-D, summarized in Table VIII. The neglected silicon area reduction in 3-D offers substantial potential cost gains.

1) *1st strategy:* Most 3-D implementations exhibit reduced wire length and associated buffer count reduction, theoretically translating into a footprint and silicon area reduction. However, fully exploiting this potential may require expanding the exploration space by the less restrictive logic/logic partitioning scheme, where logic and memory macros can be assigned to both dies in the stack. Thus, the detailed exploration of this axis may require new CAD flows to enable a higher degree of global optimization capability [26].
2) *2nd strategy:* The routing resources available during the implementation are shared across both dies. As a result, the metal layer count can be adjusted with finer granularity in 3-D, leading to more efficient resource utilization and lower manufacturing costs. Note, however, that reducing the number of metal layers can significantly impact the power integrity of the chip due to fewer available routing resources for the power delivery network [27].

This work does not explore the implications of heterogeneous integration—different technology nodes used for different wafers—on the cost, though this approach can effectively promote 3-D IC in the industry. However, our cost model methodology can be readily applied to study such heterogeneity.

V. EXPERIMENTAL RESULTS

The previous analysis highlighted the importance of γ for cost optimization. However, its value can only be explored with feedback from actual physical design implementations. This section integrates and validates our cost model with sign-off quality GDS and PPA data from 3-D IC physical layouts. We provide a representative case study for modern multicore SoCs to study the viability of the γ reduction in a 3-D memory-on-logic design flow. We show that the footprint of our 3-D designs can be reduced to achieve γ values small enough to make 3-D more attractive cost-wise while retaining superior PPA results.

A. Benchmark

We choose OpenPiton+Ariane [28] as our benchmark design, an open-source silicon-proven manycore RISC-V processor. It includes small-cache tiles, including 8 KiB of L1 instruction cache, 16 KiB of L1 data cache, 16 KiB of L2 cache, and 256 KiB of L3 cache per tile. In addition, it features a representative memory hierarchy structure with a significant, coherent, and distributed last-level cache and network-on-chip (NoC) routers inside each tile to enable the scalability of the design.

B. Technology Settings

We use a commercial 28- nm PDK technology for our implementations. The F2F via size, pitch, resistance, and capacitance are 0.5 μm $\times$ 0.5 μm, 1.0 μm, 0.5 Ω, and 1 fF, respectively [3]. The SRAM memory blocks require four BEOL layers for internal routing.

For both W2W and Co-D2W, the CAD flow must ensure that the achieved density/pitch of 3-D bumps follow the capabilities of the manufacturing process. The effective 3-D pitch can be affected by restricting the sharing of metal resources between BEOLs by appropriately setting each net's preferred top and bottom routing layers. For Co-D2W specifically, as it allows different-sized dies to be stacked, the CAD flow must additionally ensure that the 3-D interconnections are located only in the overlapping region of the two dies.

We only consider the W2W stacking method because our implementations require a high 3-D interconnectivity density. First, the memory-on-logic partitioning scheme at the tile level requires small pitches well below 10 μm. Indeed, the memories placed in the top tier result in 3,359 authentic 3-D connections, which must be fanned out from the memory pins on the left side of the dense floorplan. Second, tiles must be placed tightly at the top level of the implementation to achieve a good γ, which tremendously increases the local F2F density. Achieving these required pitches is only possible in the W2W that can attain pad size and pitch as small as 0.35-μm pad on 0.7-μm pitch [29]. At the same time, Co-D2W lacks high-yield and high-throughput equipment for pad size and pitch below 5 and 10 μm, respectively [29].

C. Single-Tile PPAC Optimization

Single-tile designs are implemented first with the same target frequency of 500 MHz at the slow corner. The tile

TABLE IX
W2W COST MODEL PARAMETERS IN OUR THREE ILLUSTRATIVE SCENARIOS FOR POWER, PERFORMANCE, AREA, AND COST (PPAC) COMPARISON

Parameters	Case A	Case B	Case C
w_d (mm)	300	300	300
technology	N28	N28	N28
stacking	W2W	W2W	W2W
y_B	0.98	0.99	0.95
$c_{B,W2W}$	0.26	0.20	0.30
D (mm^{-2})	0.001	0.0009	0.005
θ	2	2	2

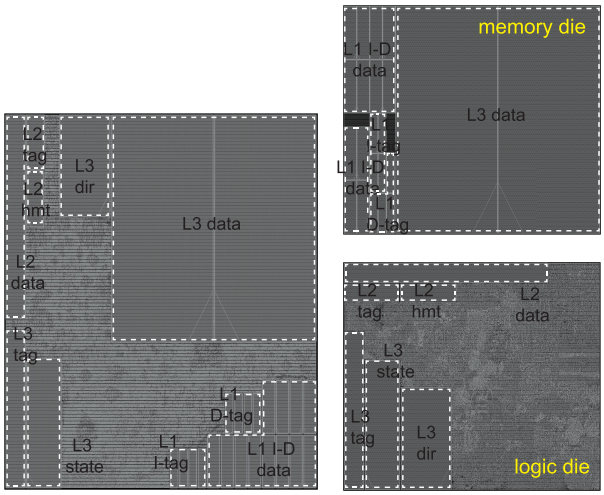


Fig. 8. Floorplans of the OpenPiton single-tile designs using a commercial 28 nm technology: 2-D (1.0 × 1.2 mm) versus 3-D (0.82 × 0.73 mm).

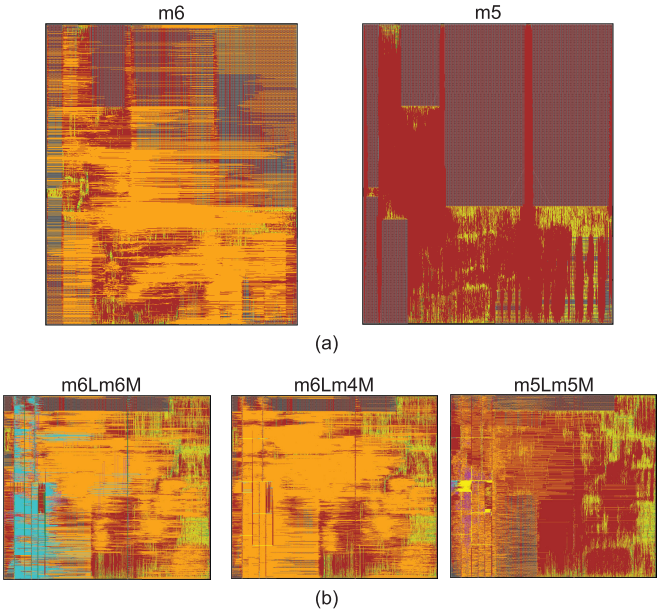


Fig. 9. Layouts (scaled equally) of the different GDS implementations for the single-tile OpenPiton+Ariane [28] design using a commercial 28- nm technology: 2-D (1.0 × 1.2 mm) versus 3-D (0.82 × 0.73 mm). (a) 2-D OpenPiton tiles. (b) 3-D OpenPiton tiles.

floorplans shown in Fig. 8 are obtained after refining aspect ratios and macro placements. We use our in-house physical design flow named Macro-3-D [30], a state-of-the-art physical design flow for memory-on-logic for the P&R of the 3-D tile. The memory tier contains memory macros only, while standard cells and potentially memory macros are placed on another tier. Finally, the memory tier macros are "projected" onto the logic tier, allowing a commercial 2-D EDA engine to perform P&R.

We study the cost optimization axes on the 3-D memory-on-logic implementation of a single tile to estimate the benefits of memory-on-logic 3-D integration in terms of PPAC. We use the *power-performance-cost* metric to compare the designs

$$\frac{\text{Frequency}}{\text{Die Cost} \times \text{Power}} \tag{34}$$

where the die cost is computed with the parameters shown in Table IX of three illustrative examples of a foundry.

1) 2-D Baseline: We first optimize the 2-D tile with six metal layers as a reference. The 2-D area is optimized with a high logic density of ∼80%. In 2-D, the cost model analysis prompts a reduction of the BEOL layer count by removing the topmost metal layers. Table X shows that the 2-D m5 version exhibits an exponential rise in the number of design rule check (DRC) violations post-route. Therefore, we choose 2-D m6 as our baseline.

2) 3-D Exploration: For 3-D, a γ under 50% is scarcely feasible due to the simplicity and small area of the tile and the requirement for matching die sizes in the W2W stacking, which limits area-optimized memory and logic partitioning. Thus, we proceed to the second axis of BEOL reduction and follow the same approach for the 3-D design as for 2-D, starting with a balanced stack mirror of 2-D (m6Lm6M). Finally, we perform P&R using different BEOL configurations, namely (6,6), (6,4), and (5,5).

3) Observations: We report our PPA results for the single-tile design in Table X and show the corresponding layouts in Fig. 9. Our main takeaways are as follows.

1) *Performance:* For the 2-D design, our attempts at meeting timing by floor planning and target density tuning were unsuccessful. On the other hand, most 3-D implementations met timing effortlessly, thanks to improved routability. However, the poor quality of the m5Lm5M implementation with an explosion of DRCs stems from the large obstructions from the F2F bumps on both dies top metal layers, heavily hindering routability between stacked memory macros.
2) *BEOL Reduction:* Reducing the metal layer count of the memory die from six to four is possible without PPA degradation in 3-D. At the same time, the 2-D implementation suffers unacceptable degradation when removing its top metal layer. Routability in 3-D is improved by the displacement of significant macro blockages from the logic tier to the memory tier, allowing memory wires to connect to standard cells without excessive detours. The other noticeable effect shown in Fig. 10 is the general net wire length reduction, allowing for BEOL savings.

TABLE X

SINGLE-TILE 2-D VERSUS 3-D F2F W2W FOR DIFFERENT BACK-END-OF-LINE (BEOL) CONFIGURATIONS. COMPARISON DONE AT TARGET FREQUENCY OF 500 MHz. COST METRICS ARE COMPUTED BASED ON THE THREE REPRESENTATIVE CASES IN TABLE IX

Single-Tile	2D m6	2D m5	3D m6Lm6M	3D m6Lm4M	3D m5Lm5M
Full-chip PPA Metrics					
metals used	6	5	6 (logic) 6 (memory)	6 (logic) 4 (memory)	5 (logic) 5 (memory)
footprint (W × H) (mm)	1.0×1.2	1.0×1.2	0.82×0.73	0.82×0.73	0.82×0.73
silicon area (mm^2)	1.20	1.20	0.60	0.60	0.60
total WL (m)	6.91	7.23	5.96	6.09	6.94
std cell area (μm^2)	306.4K	319.7K	305.1K	303.6K	312.1K
gate density (%)	80.4	81.9	77.3	76.9	78.6
worst negative slack (ns)	-0.435	-2.35	-0.041	-0.145	-0.590
total negative slack (ns)	-12.93	-87.60	-2.705	-0.704	-388.55
effective freq (MHz)	410.6	229.6	489.9	466.1	386.1
total power (mW)	115.6	128.2	112.7	112.8	124.6
power-delay product (pJ)	282	558	230	242	323
# F2F bumps	-	-	4,173	3,928	36,547
avg. F2F pitch (µm)	-	-	12.0	12.3	4.1
# 2D nets routed in 3D	-	-	206	15	2,420
buffer count	26,652	31,102	26,600	26,129	29,280
# routing DRCs	3	> 100K	20	312	> 100K
max routing util (%)	40.81 (M4)	46.32 (M4)	33.75 (M3)	34.61 (M3)	51.65 (M4)
W2W Cost Metrics for Cases [A, B, C]					
die-cost ratio r_d	1	0.96	[1.07, 1.05, 1.11]	[0.99, 0.97, 1.03]	[0.99, 0.97, 1.03]
power-performance-cost	1	0.53	[1.14, 1.17, 1.10]	[1.18, 1.20, 1.13]	[0.88, 0.90, 0.85]

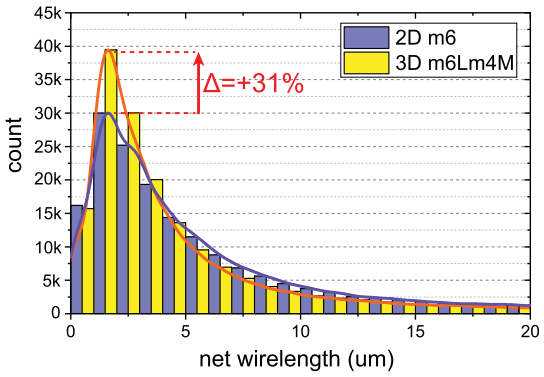


Fig. 10. Wire length distribution of nets in the single-tile design. We observe a distribution shift toward shorter wires for 3-D.

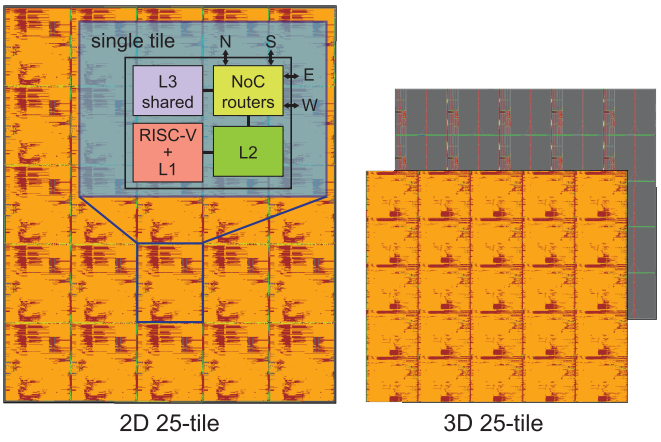


Fig. 11. GDS Layouts (scaled equally) of the 25-tile OpenPiton+Ariane [28] design using a commercial 28-nm technology: 2-D (5.3 × 6.3 mm) versus 3-D (4.2 × 3.7 mm).

D. Viability of the γ Axis

We decide on a large design of 25 tiles connected in a 5 × 5 mesh topology to highlight area savings opportunities in 3-D.

First, we select the PPAC-optimized 3-D single-tile implementation with six BEOL logic layers and four BEOL memory layers. Next, we synthesize with *Synopsys Design Compiler* the 25-tile netlist with the extracted single-tile .db file. We then manually place the single-tile black boxes and route the inter-tile wires using the space between the tiles, using *Cadence Innovus*. Finally, the final design is assembled, and PPA metrics are collected with *Cadence Tempus* after RC extraction of the entire chip.

To examine the potential area reduction, we reduce the tile spacing until timing degrades at a slow corner ($-\text{WNS} > 5\% \cdot T_{\text{period}}$) in 2-D and 3-D. The PPA results are summarized in Table XI, and the implementation layouts and single-tile architecture are shown in Fig. 11.

We see that 3-D considerably reduces the buffer count, which reduces the total silicon area and allows a γ smaller than 50%, down to 46.5%. The significant reduction of the buffers' total silicon area is due to the wire length reduction from the denser packing of the inter-tile channels which reduces RC delays. This effect can be observed in designs with dominating global interconnects, but its extent is design-dependent. The histogram in Fig. 12 shows a more distributed routing utilization of the complete 3-D vertical stack, which alleviates the impact of the smaller footprint. In contrast, shrinking the 2-D floorplan to match $\gamma = 0.5$ significantly increases the DRCs and degrades both frequency and power.

While the actual raw values depend on the foundry parameters, the results showcase the potential PPAC improvements in 3-D when improving γ. The clear die-cost ratio improvement

TABLE XI

25-TILE 2-D VERSUS 3-D F2F W2W. DESIGNS ARE SQUEEZED UNTIL WNS < 0 AND THE NUMBER OF ROUTING DRC VIOLATIONS IS MANAGEABLE. THE TARGET FREQUENCY OF EACH DESIGN IS SET TO THE EFFECTIVE FREQUENCY OF THE CORRESPONDING SINGLE-TILE IMPLEMENTATION. COST METRICS ARE COMPUTED BASED ON THE THREE REPRESENTATIVE CASES IN TABLE IX

25-Tile	2D m6 (Reference)	2D m6 (Squeezed)	3D m6Lm4M (PPAC Opt.)
Full-chip PPA Metrics			
metals used	6	6	6 (logic) 4 (memory)
footprint (W × H) (mm)	5.30×6.30	5.10×6.10	4.20×3.70
silicon area (mm^2)	33.39	31.11 (↓7%)	15.54 (γ=0.465)
total WL (m)	179.1	178.7	156.4
target freq (MHz)	410.6	410.6	466.1
worst negative slack (ns)	-0.20	-0.31	0.030
total negative slack (ns)	-0.90	-2.27	0
effective freq (MHz)	379.5	364.2 (↓4%)	466.1 (↑22%)
total power (W)	3.04	3.09	2.94
power-delay product (pJ)	8011	8484	6308
Inter-Tile PPA Metrics			
std cell area (μm^2)	110K	104K	97K
gate density (%)	4.5	10.6	18.3
total WL (m)	6.25	5.89	4.17
# F2F bumps	-	-	16,490
avg. F2F pitch (µm)	-	-	6.0
# 2D nets routed in 3D	-	-	3,281
power (W)	0.151	0.197	0.117
buffer count	19,845	12,815	5,680
power-delay product (pJ)	398	541 (↑36%)	253 (↓36%)
# routing DRCs	0	4413	11
W2W Cost Metrics for Cases [A, B, C]			
die-cost ratio r_d	1	[0.93, 0.93, 0.92]	[0.91, 0.89, 0.95]
power-performance-cost	1	[1.02, 1.02, 1.03]	[1.40, 1.43, 1.34]

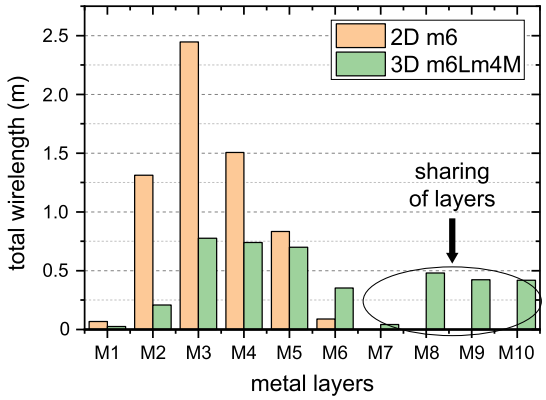


Fig. 12. 2-D versus 3-D total wire length per metal layer for the 25-tile designs. The values only consider the inter-tiles global wires and highlight the sharing of metal resources between both dies.

in 3-D from the single-tile design to the 25-tile design can be explained by the combined positive effects of the area increase on r_{DPW} and r_Y. Both designs' relatively small die areas land below the lower knee-point of the squared yield loss. Indeed, from (22) and worst case C of Table IX, one verifies that $a < a_{\text{Cutoff}} \simeq 60$ mm^2 is satisfied, a regime where an area increase is favorable.

VI. CONCLUSION

We propose a cost analysis of 3-D ICs that considers the additional area savings opportunities in 3-D. We show that these area savings significantly impact the cost and widen the spectrum of designs that could benefit from 3-D cost-wise. The variety in the stacking methods of W2W and Co-D2W, CMOS technologies from N28 to N3, along with the various modern statistical methodologies employed in this work, ensure the robustness and generality of our findings. Finally, we validate the viability of the area savings factor with full-chip implementations based on wafer-to-wafer (W2W) face-to-face (F2F) hybrid bonding. For a sizeable 25-core processor design, despite 3-D integration, we observe potential cost gains with clear PPA benefits compared to the 2-D implementation with 33.4 mm^2 die size, thanks to silicon area reduction.

REFERENCES

[1] A. Agnesina et al., "Power, performance, area and cost analysis of memory-on-logic face-to-face bonded 3D processor designs," in *Proc. IEEE/ACM Int. Symp. Low Power Electron. Design (ISLPED)*, Jul. 2021, pp. 1–6.

[2] D. Velenis et al., "Cost components for 3D system integration," in *Proc. 5th Electron. Syst.-Integr. Technol. Conf. (ESTC)*, Sep. 2014, pp. 1–5.

[3] E. Beyne et al., "Scalable, sub 2 μm pitch, Cu/SiCN to Cu/SiCN hybrid wafer-to-wafer bonding technology," in *IEDM Tech. Dig.*, Dec. 2017, pp. 4–32.

[4] T. Uhrmann, J. Burggraf, and M. Eibelhuber, "Heterogeneous integration by collective die-to-wafer bonding," in *Proc. Int. Wafer Level Packag. Conf. (IWLPC)*, Oct. 2018, pp. 1–7.

[5] G. Lecarpentier et al., "Die-to-wafer bonding of thin dies using a 2-step approach; high accuracy placement, then gang bonding," *Additional Papers Presentations*, vol. 2010, no. DPC, pp. 001254–001281, Jan. 2010.

[6] B. W. Ku, P. Debacker, D. Milojevic, P. Raghavan, and S. K. Lim, "How much cost reduction justifies the adoption of monolithic 3D ICs at 7 nm node?" in *Proc. 35th Int. Conf. Comput.-Aided Design*, Nov. 2016, pp. 1–7.

[7] X. Dong and Y. Xie, "System-level cost analysis and design exploration for three-dimensional integrated circuits (3D ICs)," in *Proc. Asia South Pacific Design Autom. Conf.*, Jan. 2009, pp. 234–241.

[8] Y. Chen, D. Niu, Y. Xie, and K. Chakrabarty, "Cost-effective integration of three-dimensional (3D) ICs emphasizing testing cost analysis," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design (ICCAD)*, Nov. 2010, pp. 471–476.

[9] A. Mallik et al., "The economic impact of EUV lithography on critical process modules," in *Proc. SPIE*, vol. 9048, 2014, Art. no. 90481R.

[10] A. Mallik, J. Ryckaert, A. Mercha, D. Verkest, K. Ronse, and A. Thean, "Maintaining Moore's law: Enabling cost-friendly dimensional scaling," in *Proc. SPIE*, vol. 9422, 2015, Art. no. 94221N.

[11] S. M. Y. Sherazi et al., "Standard-cell design architecture options below 5 nm node: The ultimate scaling of FinFET and nanosheet," in *Proc. SPIE*, Mar. 2019, Art. no. 1096202.

[12] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA, USA: CreateSpace, 2009.

[13] R Core Team, "R: A language and environment for statistical computing," R Found. Stat. Comput., Vienna, Austria, 2023. [Online]. Available: https://www.R-project.org

[14] D. K. D. Vries, "Investigation of gross die per wafer formulas," *IEEE Trans. Semicond. Manuf.*, vol. 18, no. 1, pp. 136–139, Feb. 2005.

[15] A. V. Ferris-Prabhu, "An algebraic expression to count the number of chips on a wafer," *IEEE Circuits Devices Mag.*, vol. 5, no. 1, pp. 37–39, Jan. 1989.

[16] I. Koren and Z. Koren, "Defect tolerance in VLSI circuits: Techniques and yield analysis," *Proc. IEEE*, vol. 86, no. 9, pp. 1819–1838, Sep. 1998.

[17] S. Kucherenko, D. Albrecht, and A. Saltelli, "Exploring multi-dimensional spaces: A comparison of Latin hypercube and quasi Monte Carlo sampling techniques," 2015, *arXiv:1505.02350*.

[18] A. B. Owen, "Quasi-Monte Carlo sampling," in *Monte Carlo Ray Tracing: Siggraph 2003 Course 44*, H. W. Jensen, Ed. Los Angeles, CA, USA: SIGGRAPH, 2003, pp. 69–88.

[19] S. Mallat, "Sciences des données," *L'annuaire du Collège de France. Cours et travaux*, vol. 118, no. 118, pp. 31–42, 2020.

[20] M. D. Morris, "Factorial sampling plans for preliminary computational experiments," *Technometrics*, vol. 33, no. 2, pp. 161–174, 1991.

[21] A. Gelman, "Analysis of variance—Why it is more important than ever," *Ann. Statist.*, vol. 33, no. 1, pp. 1–53.Feb. 2005.

[22] A. B. Owen, *Monte Carlo Theory, Methods and Examples*. 2013. [Online]. Available: https://artowen.su.domains/mc/

[23] C. Bénard, G. Biau, S. Veiga, and E. Scornet, "Interpretable random forests via rule extraction," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 937–945.

[24] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[25] M. Cavalcante et al., "MemPool-3D: Boosting performance and efficiency of shared-L1 memory many-core clusters with 3D integration," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, Mar. 2022, pp. 394–399.

[26] A. Agnesina et al., "Hier-3D: A hierarchical physical design methodology for face-to-face-bonded 3D ICs," in *Proc. ACM/IEEE Int. Symp. Low Power Electron. Design*, Aug. 2022, pp. 1–6.

[27] E.-P. Li, *Electrical Modeling and Design for 3D System Integration: 3D Integrated Circuits and Packaging, Signal Integrity, Power Integrity and EMC*. Wiley, 2012. [Online]. Available: https://ieeexplore.ieee.org/book/6183551

[28] J. Balkind et al., "OpenPiton+ Ariane: The first open-source, SMP Linux-booting RISC-V system scaling from one to many cores," in *Proc. Workshop Comput. Archit. Res. RISC-V (CARRV)*, 2019, pp. 1–6.

[29] G. Van der Plas and E. Beyne, "Design and technology solutions for 3D integrated high performance systems," in *Proc. Symp. VLSI Circuits*, Jun. 2021, pp. 1–2.

[30] 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.

Anthony Agnesina received the diplome d'Ingénieur degree from CentraleSupélec, Gif-sur-Yvette, France, in 2016, and the M.S. and Ph.D. degrees in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2017 and 2022, respectively.

He joined NVIDIA Research, Austin, TX, USA, in 2022. His research interests include 3-D integrated circuits, applied machine learning to EDA, and algorithms for computer-aided design of VLSI circuits.

Dr. Agnesina received the 2022 ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED) Best Paper Award.

Moritz Brunion received the bachelor's and master's degrees in electrical and computer engineering from the University of Bremen, Bremen, Germany, in 2019 and 2022, respectively.

He is currently a Researcher with IMEC, Leuven, Belgium. His current research interests focus on technology-driven interconnect architecture design.

Jinwoo Kim received the B.S. and M.S. degrees in electrical engineering and computer science from Seoul National University, Seoul, South Korea, in 2011 and 2013, respectively, and the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2022.

He was an Analog Circuit Designer with Samsung Electronics, Hwaseong, South Korea, from 2013 to 2017. He was also a Technical Intern with Synopsys Inc., Mountain View, CA, USA, in 2021. He is currently a SoC Design Engineer with Intel Corp., Santa Clara, CA. His research interests include the system and technology co-optimization for 3-D IC designs.

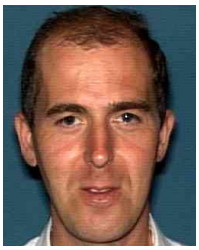
Alberto Garcia-Ortiz (Senior Member, IEEE) received the diploma degree in telecommunication systems from the Polytechnic University of Valencia, Valencia, Spain, in 1998, and the Ph.D. degree (summa cum laude) from the Department of Electrical Engineering and Information Technology, Institute of Microelectronic Systems, Darmstadt University of Technology, Darmstadt, Germany, in 2003.

He was with Newlogic, Austria, Europe. From 2003 to 2005, he worked as a Senior Hardware Design Engineer with IBM Deutschland Development and Research, Böblingen, Germany. Then, he joined the startup AnaFocus, Spain, where he was responsible for the design and integration of AnaFocus' Next Generation Vision Systems-on-Chip. He is currently a Full Professor for the chair of integrated digital systems with the University of Bremen, Bremen, Germany. His research interests include low-power design and estimation, communication-centric design, 3-D integration, and hardware accelerators for AI.

Dr. Garcia-Ortiz received the Outstanding Dissertation Award in 2004 from the European Design and Automation Association. In 2005, he received from IBM an Innovation Award for contributions to leakage estimation. Two patents are issued with that work. He serves as an Editor for JOLPE and is a reviewer for several conferences, journals, and European projects.

Dragomir Milojevic received the M.S. and Ph.D. degrees in electrical engineering from the Université Libre de Bruxelles (ULB), Brussels, Belgium, in 1994 and 2004, respectively.

He holds the position of a Professor with Digital Electronics and Systems Design, ULB. In 2004, he joined IMEC, Leuven, Belgium, working on multiprocessor and network-on-chip (NoC) architectures for low-power multimedia systems. Since 2008, he has been working on enablement of 3-D stacked ICs, and system and design technology co-optimization of advanced technology nodes, and design methodologies for technology aware 3-D ICs.



Francky Catthoor (Fellow, IEEE) received the Ph.D. degree in EE from Katholieke University Leuven, Leuven, Belgium, in 1987.

From 1987 to 2000, he has headed several research domains in the area of synthesis techniques and architectural methodologies. Since 2000, he has been strongly involved in other activities with IMEC including co-exploration of application, computer architecture and deep submicrometer technology aspects, biomedical systems and IoT sensor nodes, and photo-voltaic modules combined with renewable energy systems, with IMEC, Leuven, Belgium. He is also a part-time Full Professor with the EE Department, KULeuven, Leuven.

Dr. Catthoor is currently an IMEC Senior Fellow. He has been an Associate Editor for several IEEE and ACM journals.

Gioele Mirabelli received the B.Eng. and M.Eng. degrees in electronic engineering from the University of Calabria, Calabria, Italy, in 2013 and 2015, respectively, and the Ph.D. degree from Tyndall National Institute, University College Cork, Cork, Ireland, in 2020.

He is working on 2-D-Semiconductors for future electronics. In 2020, he joined the Physical Design Research Group, IMEC, Leuven, Belgium, where he is working on design technology co-optimization of advanced technology nodes.

Manu Komalan (Member, IEEE) received the integrated master's degree in nanotechnology from Amity University, Noida, India, in 2011, the Ph.D. degree in electrical engineering from Katholieke Universiteit Leuven, Leuven, Belgium, and the Ph.D. degree in computer science from the Universidad Complutense de Madrid, Madrid, Spain, in 2017.

He then joined IMEC, Leuven, as a Memory System Architecture Researcher. He worked as the Program Manager with the Memory INSITE Program, IMEC, from 2019 to 2021 on activities involving exploration, analysis, and optimization of NVMs across the different layers of abstraction. He is currently an Architect and Manager with the Compute Systems Architecture Unit, IMEC, wherein, his focus is systems design for the AI and HPC domain.

Sung Kyu Lim (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees from the Computer Science Department, University of California at Los Angeles (UCLA), Los Angeles, CA, USA, in 1994, 1997, and 2000, respectively.

He joined the School in 2001. He is currently a Professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. He has published more than 400 articles on 2.5-D and 3-D ICs. He is the author of *Practical Problems in VLSI Physical Design Automation* (Springer, 2008) and *Design for High Performance, Low Power, and Reliable 3-D Integrated Circuits* (Springer, 2013). His research interests focus on the architecture, design, test, and electronic design automation (EDA) solutions for 2.5-D and 3-D ICs. His research interests are featured as research highlight in the communication of the ACM in January, 2014.

Dr. Lim received the National Science Foundation Faculty Early Career Development (CAREER) Award in 2006. He received the ACM SIGDA Distinguished Service Award in 2008. He received the Best Paper Award from the IEEE TRANSACTIONS ON ELECTROMAGNETIC COMPATIBILITY and the IEEE TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS in 2022. He received the Best Paper Award from several conferences in EDA, including ACM/IEEE International Symposium on Low Power Electronics and Design in 2022. He joined the Defense Advanced Research Projects Agency (DARPA) in 2022 as a Program Manager in the Microsystems Technology Office (MTO).