Power, Performance, Area and Cost Analysis of Memory-on-Logic Face-to-Face Bonded 3D Processor Designs

Anthony Agnesina[1], Moritz Brunion[2], Jinwoo Kim[1], Alberto Garcia-Ortiz[2],
Dragomir Milojevic[3], Francky Catthoor[3], Manu Perumkunnil[3] and Sung Kyu Lim[1]
[1]School of ECE, Georgia Institute of Technology, Atlanta, USA
[2]ITEM.ids, University of Bremen, Germany
[3]IMEC, Leuven, Belgium
agnesina@gatech.edu

***Abstract*—In this paper, we present a power, performance, area and cost (PPAC) analysis for large-scale 3D processor designs based on wafer-to-wafer bonding. From the evaluation of our cost model, we investigate a typically disregarded opportunity in 3D that is area savings due to buffer savings and better routability, offering unexpected cost savings. We explore the viability of this factor with the feedback of a state-of-the-art 3D 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 3D implementation offers 7% silicon area savings, resulting in 53.5% footprint reduction. We also obtain a 40% power-performance-cost improvement compared with 2D counterparts.**

I. Introduction

The semiconductor industry is innovating new techniques to prolong Moore's Law as CMOS transistor geometries approach physical limits resorting to novel devices or extreme ultra-violet (EUV) lithography to reduce masks count. But it often discounts the benefit of 3D vertical integration believing integration cost exceeds its economic benefit. In this paper, we propose a high-level study of a generic cost model to capture the dominant trade-offs in 2D and 3D. Our analyses reveal that the cost is very sensitive to the newly explored 3D area savings, giving a novel perspective to potentially alleviate traditional economic barriers of 3D vertical integration.

We focus on the wafer-to-wafer (W2W) hybrid bonding method that stacks two separately pre-fabricated wafers, aligns them, bonds them face-to-face (F2F) and then dices them to create multiple single two-tier dies interconnected by high-density metallic interconnections. Utilizing 3D integration, heterogeneous devices and technologies (memory, logic, RF, analog, sensors, etc.) can be optimized for cost and performance for each individual layer. We claim the following contributions of this paper:

- We develop a high-level cost model to capture the conceptual trade-offs among area, cost, and performance in 2D and 3D, independent of the precise knowledge of foundry parameters.
- We develop an effective cost analysis and explore the significance of the defined parameters, revealing an unexpectedly large impact of the often disregarded 3D area savings factor γ.
- Our experiments show that our proposed cost optimization axis of area savings is viable. Due to the better exploitation in 3D of the F2F-bonded back-end-of-line (BEOL) stack, we obtain improved power, performance, as well as *footprint and silicon* area results. Such area gains lead to a significant cost saving. This PPA plus cost saving make 3D ICs more attractive than previous analysis has indicated.

Our industry representative benchmark designs of a large-scale manycore processor done at GDS-level and simulated with sign-off simulations convincingly highlight the performance as well as area and cost savings opportunities of 3D ICs.

978-1-6654-3922-0/21/$31.00 ©2021 IEEE

II. Cost Model Overview

We build a highly parametrized and generic cost model for 3D IC integration. It handles W2W hybrid bonding and 3D memory-on-logic integration.

Previous work on cost modeling for 3D IC [1], [2], [3] do not consider the silicon area savings opportunities, assuming a fixed footprint area reduction of 50% in 3D compared to 2D. In sharp contrast, we introduce the 3D area savings as an independent parameter γ.

A. Parameters

Our cost model integrates two types of parameters 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 real full-chip GDS designs obtained using a physical design flow.

B. Assumptions

To set the foundry constants of Table I, we use information from ITRS reports, previous literature [4], [5], as well as inside expertise validated by silicon measurements and industry feedback. We propose the three cases presented in Table II as illustrative examples of a sub-28nm foundry-grade technology node. A BEOL configuration called $4M_x2M_y1Mz$ corresponds to a stack of 4 expensive and 2 less expensive metal layers. The remaining layers in the stack are of the cheaper M_z type. The cost of W2W 3D integration includes wafer alignment, bonding, Si thinning and via realization. In

TABLE I
COST MODEL PARAMETERS.

parameters	description
foundry related parameters	
w_d	wafer diameter
$c_{\text{FEOL, Logic}}$	FEOL cost of logic wafer
$c_{\text{MOL, Logic}}$	MOL cost of logic wafer
D_{Logic}	logic die defect density
α_{Logic}	logic die clustering parameter
$c_{\text{FEOL, Memory}}$	FEOL cost of memory wafer
$c_{\text{MOL, Memory}}$	MOL cost of memory wafer
D_{Memory}	memory die defect density
α_{Memory}	memory die clustering parameter
y_B	3D integration yield
c_B	3D integration cost
$c_{M_i}(\mathcal{C})$	cost of metal layer i in BEOL of configuration $\mathcal{C}$
implementation related parameters	
a	2D die area
γ	2D vs. 3D die area ratio: $a_{3D} = \gamma a$
$\mathcal{C}$	BEOL configuration/aggressivity
N	number of BEOL metal layers

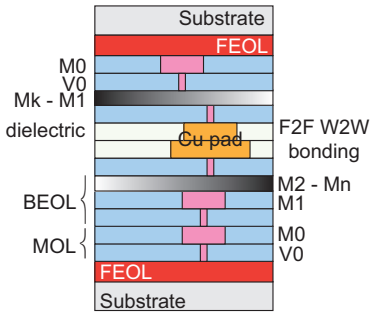

Fig. 1. Conceptual view (not to scale) of the 3D stack for our wafer-on-wafer face-to-face wafer cost analysis.

Figure 1 we present our exemplar stack which includes a middle-of-line (MOL) used for local interconnect of standard cells. A value of 2 for the cluster parameter of the yield models slight non-uniformity of fault distribution. The effective memory defect density is specified to be better than the logic defect density as it is assumed that the memory logic? implements ECC or redundancy. Moreover, the memory front-end-of-line (FEOL) is cheaper as some FEOL processing steps in memory wafers can be omitted.

C. Analytical Model

For convenience, we will write $f(x_1, x_2, \ldots, x_n)$ as $f(x_k)$ to highlight the dependence of 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 $(\ldots)$. To derive accurate cost optimization techniques, we focus on the die cost ratio between 2D and 3D:

$$\begin{aligned} r_D(a,\gamma) = \frac{C_{D_{3D}}(\gamma a)}{C_{D_{2D}}(a)} &= \frac{C_{W_{3D}}}{kgd_{3D}(\gamma a)} \cdot \frac{kgd_{2D}(a)}{C_{W_{2D}}} \\ &= \frac{C_{W_{3D}}}{C_{W_{2D}}} \cdot \frac{dpw(a)}{dpw(\gamma a)} \cdot \frac{y_{2D}(a)}{y_{3D}(\gamma a)} \end{aligned} \tag{1}$$

TABLE II
COST MODEL PARAMETERS IN OUR THREE ILLUSTRATIVE SCENARIOS. BEOL CONFIGURATION IS ASSUMED AS FOLLOWS: HIGH = $4M_x2M_y1M_z$, MEDIUM = $3M_x2M_y1M_z$, LOW = $2M_x3M_y1M_z$.

parameters	case A	case B	case C
$w_d\ (mm)$	300	300	300
$c_{\text{FEOL, Logic}}$	1	1	1
$c_{\text{MOL, Logic}}$	0.18	0.18	0.18
$D_{\text{Logic}}\ (mm^{-2})$	0.001	0.0009	0.005
α_{Logic}	2	2	2
$c_{\text{FEOL, Memory}}$	0.85	0.85	0.85
$c_{\text{MOL, Memory}}$	0.18	0.18	0.18
$D_{\text{Memory}}\ (mm^{-2})$	0.0009	0.0008	0.004
α_{Memory}	2	2	2
y_{B}	0.98	0.99	0.95
c_{B}	0.26	0.20	0.30
$c_{\text{expensive layer Mx}}$	0.19	0.19	0.19
$c_{\text{medium layer My}}$	0.11	0.09	0.11
$c_{\text{cheap layer Mz}}$	0.07	0.05	0.07
$\text{BEOL}_{\text{Logic}}$	medium	high	high
$\text{BEOL}_{\text{Memory}}$	medium	medium	high

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 implementation-dependent parameter γ models the shrinking of the die area in 3D. The equation can be rewritten to highlight the die cost trade-offs:

$$r_D(a,\gamma) = r_W \cdot r_{DPW}(a,\gamma) \cdot r_Y(a,\gamma) \tag{2}$$

where r_W, the wafer cost ratio, is independent of the die areas. The ratio r_{DPW} describes how many more 3D dies we can obtain by using 3D integration and reducing footprint a with factor γ. The ratio r_Y compares the 3D yield with the 2D yield.

D. Cost Model Components

We present classical approximations for each component of the cost model such as die-per-wafer, yield, wafer cost, and we highlight their dependence on γ.

1) Die-Per-Wafer: To derive accurate and interpretable guidance from the model, we select the most accurate analytical formula for the gross number of die-per-wafer. Figure 2 compares various second-order approximations with the exact brute force solution presented in [6]. It shows that the exponential formula due to Ferris-Prabhu [7] is most accurate:

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

2) Yield: We present the different components for the wafer yield calculation. For 2D integration, it corresponds to the yield of a die, while the yield modeling for 3D W2W ICs is more complicated, as presented here.

Die Yield We use the negative binomial yield model [8] suggested in ITRS reports that considers the clustering of manufacturing defects rather than their independent occurrences:

$$y_{die}(a) = \left(1 + \frac{Da}{\alpha} \right)^{-\alpha} \tag{4}$$

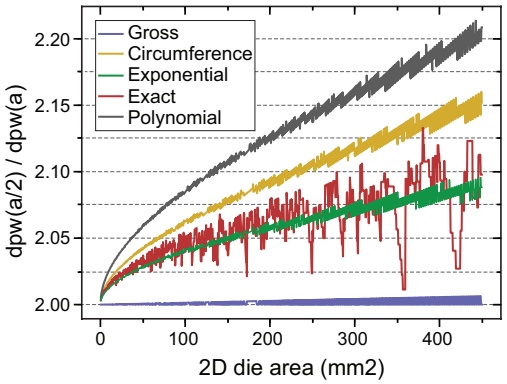

Fig. 2. Inverse die-per-wafer ratio $r_{DPW}(a, 0.5)^{-1}$ for different analytical approximations and $w_d = 300$mm. We observe the Exponential model is the most accurate vs. the Exact solution.

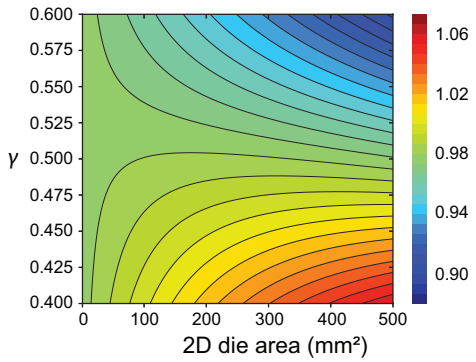

Fig. 3. Inverse yield ratio $r_Y(a, \gamma)^{-1}$: hotter colors indicate improved 3D yield vs. 2D. The parameters used are from Case A, but the observed trends are irrespective of the case chosen. It is clear that a γ below 0.5 has a very positive impact on the 3D yield.

where D is the defect density and α the clustering parameter of the faults. Parameter α depends on the technology and the design itself (e.g., mask steps) and small values indicate increased clustering. It is straightforward to extend the current framework to other more precise models, as those proprietary ones available inside the foundry companies.

Stack Yield In W2W integration, each die can usually be tested only after bonding. Therefore, some bad dies are forced to bond on the good dies, resulting in a stack yield of:

$$y_{stack}(a) = \prod_{i=1}^{\#dies} y_{die,i}(a) \tag{5}$$

Assembly Yield Since the grid of copper pads on the wafer is produced independently of the number of vias required to electrically interconnect the design, we abstract the yield of assembly into a single factor y_B.

Total Yield The total yield ratio for a two-tier stack is:

$$r_Y(a, \gamma) = \frac{y_{2D}(a)}{y_{3D}(\gamma a)} = \frac{y_{die,2D}(a)}{y_B \cdot y_{die,1}(\gamma a) \cdot y_{die,2}(\gamma a)} \tag{6}$$

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

$$C_{W_{2D}}(N, \mathcal{C}) = c_{FEOL} + c_{MOL} + \underbrace{\sum_{i=1}^{N} c_{M_i}(\mathcal{C})}_{c_{BEOL}(N,\mathcal{C})} \tag{7}$$

For the 3D case, we add the integration cost to the wafer costs:

$$C_{W_{3D}} = c_B + C_{W_{2D}}(N_1, \mathcal{C}_1) + C_{W_{2D}}(N_2, \mathcal{C}_2) \tag{8}$$

III. Key Metrics for 3D IC Cost Analysis

Relying on mathematical analysis of precise analytical formulas, our intention is to establish cost optimization priorities that are robust to changes of the foundry parameters.

A. Die-Per-Wafer Ratio

The floor function of Eq. 3 can be neglected to obtain:

$$r_{DPW}(a, \gamma) = \frac{dpw(a)}{dpw(\gamma a)} \approx \gamma e^{-2\sqrt{a}/w_d(1-\sqrt{\gamma})} \tag{9}$$

Three direct conclusions can be made from this formula:

- Increasing w_d, a typical trend with the advancements of fabs, reduces the DPW ratio.
- A larger die area reduces the DPW ratio.
- The value of γ bounds the DPW ratio.

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

$$\left| \frac{\frac{\partial r_{DPW}(a,\gamma)}{\partial \gamma} \Delta\gamma}{\frac{\partial r_{DPW}(a,\gamma)}{\partial a} \Delta a} \right| = \left(\frac{\Delta\gamma}{\gamma}\right)\left(\frac{a}{\Delta a}\right)\frac{\sqrt{\gamma}}{1-\sqrt{\gamma}}\left(1+\frac{w_d}{\sqrt{\gamma a}}\right) \gg 1 \tag{10}$$

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

B. Yield Ratio

Eq. 6 can be rewritten as:

$$r_Y(a, \gamma)^{-1} = y_B \left(\frac{1 + \frac{D_1 \gamma a}{\alpha_1}}{1 + \frac{D_1 a}{\alpha_1}}\right)^{-\alpha_1} \left(1 + \frac{D_2 \gamma a}{\alpha_2}\right)^{-\alpha_2} \tag{11}$$

We see that

$$\frac{\partial\, r_Y(a, \gamma)^{-1}}{\partial\, \gamma} \leq 0 \tag{12}$$

always holds, revealing a monotonic relationship between γ and the yield ratio, where decreasing γ always favors 3D. Next, the implications of logic-on-logic and memory-on-logic designs on the yield ratio are discussed.

1) Logic-on-Logic: For $(D_1, \alpha_1) = (D_2, \alpha_2) = (D, \alpha)$ we can show that:

$$\frac{\partial\, r_Y(a, \gamma)^{-1}}{\partial\, a} \geq 0 \iff a \leq a_{Cutoff} = \frac{\alpha(1 - 2\gamma)}{D\gamma} \tag{13}$$

Therefore, if $\gamma \geq 0.5$, the yield worsens for 3D when the die area increases, and the resulting maximum inverse yield ratio is y_B. If $\gamma < 0.5$, the yield ratio is increased with the area until it reaches its maximum for a 2D footprint of a_{Cutoff}.

2) Memory-on-Logic: In this case, the equation is more complex to analyze analytically. However, we can summarize the fundamental trade-offs with the graph shown in Figure 3. We show mathematically that the following trends hold regardless of the absolute values of the foundry parameters.

- If $\gamma \leq \frac{D_1}{D_1+D_2}$, the inverse ratio yield improves with the die area until it reaches a peak where it starts decreasing onwards. The imbalance in the upper bound on γ introduced by the heterogeneity in the defect densities of memory-on-logic shows that if the memory tier exhibits a smaller defect density (D_2), the requirement for the 3D die area increase is more relaxed.
- If $\gamma > \frac{D_1}{D_1+D_2}$, the inverse yield ratio rises with the area.

C. Wafer Cost Ratio

An obvious cost improvement is obtained with the reduction of the total metal layer count of the stack. This axis is readily available in 3D, as the routing resources of two metal layers in 3D correspond to about one 2D metal layer. The metal layer count can therefore be adjusted to the routing resource requirements of a design more accurately in 3D, resulting in more optimized cost. This granularity also enables a second gain from the more diverse BEOL configurations available for top or bottom die.

For the logic-on-logic balanced stack, we get:

$$r_W = 2 + \frac{c_B}{C_{W_{2D}}(N, \mathcal{C}_{\text{Logic}})} \geq 2 \tag{14}$$

which is minimized for taller stacks and very aggressive BEOL configurations. While this reduces the wafer cost ratio, this effect is not a reasonable optimization. Instead, the memory-on-logic balanced stack is a more promising approach:

$$r_W = 1 + \frac{c_B}{C_{W_{2D}}(N, \mathcal{C}_{\text{Logic}})} + \frac{C_{W_{2D}}(N, \mathcal{C}_{\text{Memory}})}{C_{W_{2D}}(N, \mathcal{C}_{\text{Logic}})} \tag{15}$$

as it introduces a difference in wafer costs due to different FEOL processing and BEOL configurations.

In the imbalanced case, both 3D tiers having less metal layers than the 2D counterpart yields additional cost saving opportunities.

D. Sensitivity Analysis of the Die Cost Ratio

It is difficult to study and interpret a multivariate function such as r_D that depends on all the parameters presented in Table I. Therefore, we perform a sensitivity analysis to highlight the primary driving parameters of the cost and bring out trends irrespective of the imprecision on the parameters.

In our sensitivity analysis, the input parameters of the model are seen as variables, which are varied to estimate their contributions to the output of the model $r_D(a, \gamma, \ldots)$.

We use a variance-based analysis based on the Sobol sampling and variance estimation [9]. Input values are sampled according to a quasi–Monte Carlo low-discrepancy sequence. In this paradigm, the die cost ratio variance $\mathbb{V}(r_D)$ is decomposed into parts attributable to our cost parameters. The metric of choice is the sensitivity index:

$$S_p = \frac{V_p}{\mathbb{V}(r_D)} \quad \forall p \in \text{Table I} \tag{16}$$

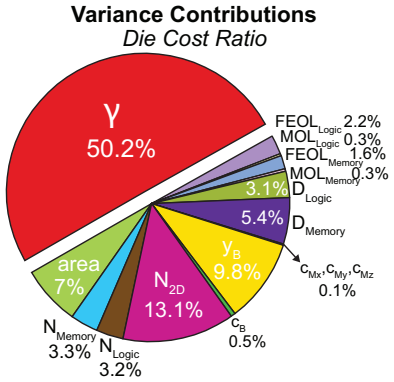

Fig. 4. Pie chart of the contribution of each cost parameter on the die cost ratio r_D. The values are obtained in our global variance-based sensitivity analysis and averaged over the three cases of Table II.

where $\sum_p S_p = 1$ and $V_p = \mathbb{V}_p\left(\mathbb{E}_{X_{i\neq p}}(r_D \mid X_i)\right)$. V_p measures the effect of varying p alone, but averaged over variations in other input parameters.

1) Local Sensitivity Analysis: In order to estimate the local effects on the cost of the different parameters, we first analyze the sensitivity of each parameter "one-at-a-time". We vary one input variable locally around its nominal value in the range $p_0 \pm \Delta p$ while keeping others at their nominal values (see Table II). For each parameter p, we find the Δp to match $S_p = S_{c_{\text{FEOL, Logic}}} \pm 3\%$ when $\Delta c_{\text{FEOL, Logic}} = 5\% \cdot c_{\text{FEOL, Logic}}$.

We record the resulting Δp's in Table III. Smaller values for $\Delta p/p_0$ imply a significant local influence on the resulting cost r_D. This specifically highlights the extreme sensitivity of r_D to the parameters γ and y_B, in that their very small variations have large effects on the cost. The sensitivity to a, $c_{\text{exp. Mx}}$, $c_{\text{med. My}}$ and $c_{\text{cheap Mz}}$ is in contrast very small.

The commonly discussed cost optimization option of metal layer reduction is shown to have a high impact on the overall cost, but significantly less than γ. As a simple exercise, take $a = 150mm^2$ in 2D (die area of an Intel Core i7-8700 processor) and case B parameters. Starting from a balanced 3D stack of $(8, 8)$ metal layers, a reduction of 2 metal layers on the memory tier $(8, 6)$ corresponds to a γ reduction from 0.500 to 0.491, so only a $1.43mm^2$ area reduction needed in the 3D implementation.

2) Global Sensitivity Analysis: To bring out the effects of the parameters on a large spectrum of values, we perform a global sensitivity analysis. In contrast with the previous approach, this offers an overall view on the influence of parameters on the cost as opposed to a local view of partial derivatives. The variation ranges for each parameter are set as is described in Table IV.

The pie chart in Figure 4 presents the individual contribution of each parameter to the variance of r_D. This reinforces the previous observation that γ is the most influential parameter, with an overall impact on the die cost ratio superior to that of all the other parameters combined.

TABLE III
SAMPLING RANGES FOR THE DIFFERENT PARAMETERS TO ACHIEVE SIMILAR LOCAL SENSITIVITY ON THE r_D COST.

parameter	$c_{\text{FEOL, Logic}}$	$c_{\text{MOL, Logic}}$	$c_{\text{FEOL, Memory}}$	$c_{\text{MOL, Memory}}$	D_{Logic}	D_{Memory}	γ	a
nominal p_0	1	0.18	0.85	0.18	$0.001mm^{-2}$	$0.0009mm^{-2}$	0.50	$100mm^2$
Δp	0.050	0.0483	0.0528	0.0523	$0.000260mm^{-2}$	$0.000240mm^{-2}$	0.0052	$75.62mm^2$
$\Delta p/p_0$	5%	26.8%	6.2%	29.1%	26%	26.7%	1%	75.6%

parameter	$c_{\text{expensive layer Mx}}$	$c_{\text{medium layer My}}$	$c_{\text{cheap layer Mz}}$	y_B	c_B	N_{2D}	N_{Logic}	N_{Memory}
nominal p_0	0.19	0.11	0.07	0.98	0.26	8	8	8
Δp	0.6219	1.4838	1.0842	0.012133	0.05551	0.36061	0.78766	0.78766
$\Delta p/p_0$	327.3%	1348.9%	1548.9%	1.2%	21.35%	4.5%	9.8%	9.8%

TABLE IV
VARIATIONS RANGES IN OUR GLOBAL SENSITIVITY ANALYSIS. VALUES OF p_0 ARE SET FOR EACH PARAMETER ACCORDING TO THE THREE CASES OF TABLE II.

parameter	variation range
γ	$[0.46, 0.54]$
a	$[0.1mm^2, 500mm^2]$
$c_{\text{FEOL, Logic}}$	$p_0 \pm 10\%\ p_0$
$c_{\text{MOL, Logic}}$	$p_0 \pm 20\%\ p_0$
D_{Logic}	$p_0 \pm 20\%\ p_0$
$c_{\text{FEOL, Memory}}$	$p_0 \pm 10\%\ p_0$
$c_{\text{MOL, Memory}}$	$p_0 \pm 20\%\ p_0$
D_{Memory}	$p_0 \pm 20\%\ p_0$
y_B	$[0.90, 0.99]$
c_B	$p_0 \pm 20\%\ p_0$
$c_{\text{expensive layer Mx}}$	$p_0 \pm 30\%\ p_0$
$c_{\text{medium layer My}}$	$p_0 \pm 30\%\ p_0$
$c_{\text{cheap layer Mz}}$	$p_0 \pm 30\%\ p_0$
N_{2D}	$[6, 12]$
N_{Logic}	$[6, 12]$
N_{Memory}	$[6, 12]$

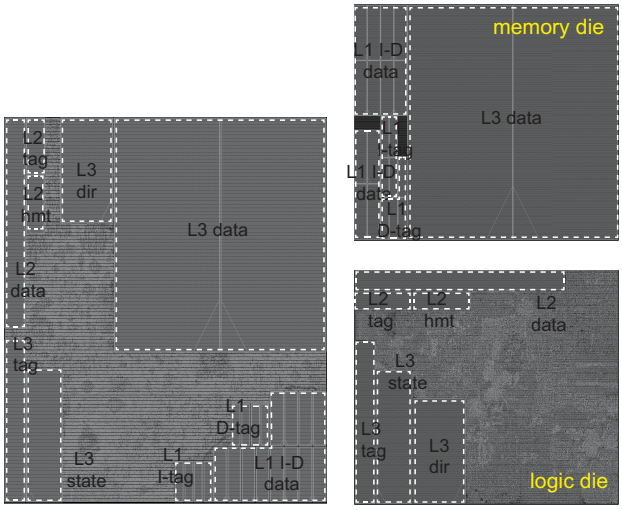

Fig. 5. Floorplans of the OpenPiton single-tile designs using a commercial 28nm technology: 2D (1.2×1.0mm) vs. 3D (0.73×0.82mm).

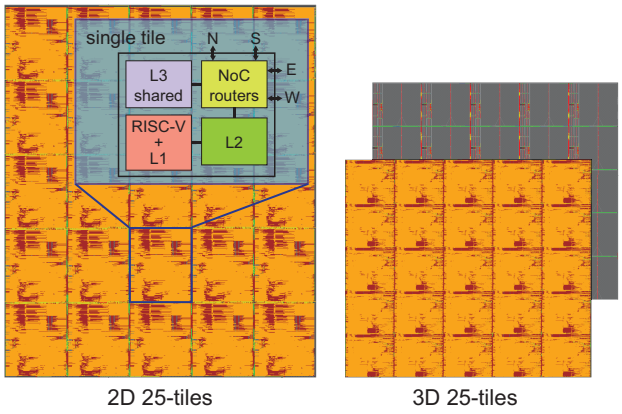

Fig. 6. GDS Layouts (scaled equally) of the 25-tile OpenPiton+Ariane [10] design using a commercial 28nm technology: 2D (5.3×6.3mm) vs. 3D (4.2×3.7mm).

IV. EXPLORATION OF γ WITH FEEDBACK FROM 3D FLOW

The previous analysis brought out the importance of the value of γ for cost optimization. This value can only be explored with the feedback from actual physical design implementations. In this section, we provide a representative case study for modern multi-core SoCs, to study the viability of the γ reduction in a 3D memory-on-logic design flow. We show that the footprint of our 3D designs can be reduced to achieve γ values small enough to make 3D more attractive costwise, while retaining superior PPA results.

A. Benchmark Design

We choose OpenPiton+Ariane [10], an open source silicon-proven manycore RISC-V processor as our benchmark design. It includes small-cache tiles, including 8kB of L1 instruction cache, 16kB of L1 data cache, 16kB of L2 cache, and 256kB of L3 cache per tile. It features a representative memory hierarchy structure with a large, coherent and distributed last-level cache as well as NoC routers inside each tile to enable scalability of the design.

B. Technology Settings

We use a commercial 28nm PDK technology for our implementations. The F2F via size, pitch, resistance and capacitance are $0.5um \times 0.5um$, $1.0um$, 0.5Ω and $1fF$, respectively [11]. The SRAM memory blocks require 4 BEOL layers for internal routing. In our 3D TECH LEF, the facing metal layers have orthogonal routing directions.

C. Viability of the γ axis

To highlight area savings opportunities, we decide on a large design of 25 tiles connected in a 5×5 mesh topology.

Single-tile designs are implemented first with an iso-performance target of 500 MHz at the slow corner. The tile

TABLE V
25-TILE 2D VS. 3D F2F. DESIGNS ARE SQUEEZED UNTIL WNS < 0. COST METRICS ARE COMPUTED BASED ON THE THREE CASES OF TABLE II.

	2D m6	3D m6Lm4M
full-chip metrics		
Metals Used	6	6 (logic) 4 (memory)
Footprint (W × H)	5.30 × 6.30	4.20 × 3.70
Footprint Area (mm^2)	33.39	15.54 ($\gamma = 0.465$)
Total WL (m)	179.1	156.4
Target Freq. (MHz)	410.6	466.1
Worst Negative Slack (ns)	−0.20	0.030
Total Negative Slack (ns)	−0.90	0
Effective Freq. (MHz)	379.5	466.1 (↑22%)
Total Power (W)	3.04	2.94
Power-delay Product (pJ)	8011	6308
inter-tile metrics only		
Std. Cell Area (um^2)	$110K$	$97K$
Total WL (m)	6.25	4.17
F2F bumps	0	16,490
# of 2D Nets routed in 3D	0	3,281
Power (W)	0.151	0.117
Buffer Count	19,845	5,680
Power-delay Product (pJ)	398	253 (↓36%)
# of Routing DRV	0	11
cost metrics for cases [A, B, C]		
Dies-Per-Wafer Ratio r_{DPW}	1	0.46
Yield Ratio r_Y	1	[1.02, 1.01, 1.03]
Wafer Cost Ratio r_W	1	[1.97, 1.89, 1.97]
Die Cost Ratio r_D	1	[0.92, 0.88, 0.93]
Power Performance Cost	1	[1.38, 1.45, 1.37]

floorplans are shown in Figure 5. We use Macro-3D [12], a state-of-the-art physical design flow for memory-on-logic for the P&R of the 3D tile. The memory tier contains memory macros only while standard cells are placed on another tier. The memory macros are projected on the logic tier, allowing for a commercial 2D EDA engine to perform P&R. We select a PPAC optimized 3D single-tile implementation with 6 BEOL logic layers and 4 BEOL memory layers.

We synthetize with *Synopsys Design Compiler* the 25-tile netlist with the extracted single-tile .db file. We then place the single-tile blackboxes manually and route the inter-tile wires using the space in between the tiles, using *Cadence Innovus*. The final design is assembled and PPA metrics are collected with *Cadence Tempus* after RC extraction of the full-chip.

To examine the potential area reduction, we reduce the tile spacing until timing degrades at a slow corner ($-WNS > 5\% \cdot T_{period}$) in both 2D and 3D. The PPA results are summarized in Table V and the layouts of our implementations along with the single-tile architecture are shown in Figure 6.

We see that 3D reduces considerably the buffer count which reduces the total silicon area and allows a γ smaller than 50%, down to 46.5%.This effect can be observed in designs with dominating global interconnects, but its extent is highly design dependent. The histogram in Figure 7 shows a more distributed routing utilization of the complete 3D vertical stack, which alleviates the impact of the smaller footprint.

D. Cost Comparison

We compare in Table V the designs using the Power Performance Cost = Frequency / (Die Cost × Power). While

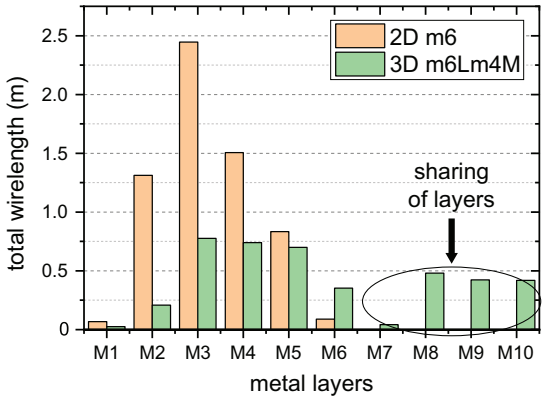

Fig. 7. 2D vs. 3D total wirelength per metal layer for the 25-tile designs. The values only consider the inter-tiles global wires, and clearly highlight the sharing of metal resources between both dies.

the absolute values are dependent on foundry parameters, the results clearly showcase the potential PPAC improvements in 3D when improving the γ parameter.

V. CONCLUSION

We propose a cost analysis of 3D ICs that considers for the first time the additional area savings opportunities in 3D. We show that these area savings have tremendous impact on the cost and widen the spectrum of designs that could potentially benefit from 3D in terms of cost. We validate the viability of this factor with full-chip implementations based on wafer-to-wafer face-to-face hybrid bonding. For a large 25-core processor design, despite 3D integration, we observe potential cost gains with clear PPA benefits with respect to the 2D implementation with $33.4mm^2$ die size, thanks to silicon area reduction.

REFERENCES

[1] Bon Woong Ku et al. How much cost reduction justifies the adoption of monolithic 3D ICs at 7nm node? In 2016 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 2016.

[2] X. Dong and Y. Xie. System-level cost analysis and design exploration for three-dimensional integrated circuits (3D ICs). In 2009 Asia and South Pacific Design Automation Conference, 2009.

[3] Y. Chen et al. Cost-effective integration of three-dimensional (3D) ICs emphasizing testing cost analysis. In 2010 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), 2010.

[4] Arindam Mallik et al. The economic impact of EUV lithography on critical process modules. In EUV Lithography V. SPIE, 2014.

[5] Arindam Mallik et al. Maintaining Moore's law: enabling cost-friendly dimensional scaling. In EUV Lithography VI. SPIE, 2015.

[6] D. K. de Vries. Investigation of gross die per wafer formulas. IEEE Transactions on Semiconductor Manufacturing, 2005.

[7] A. V. Ferris-Prabhu. An algebraic expression to count the number of chips on a wafer. IEEE Circuits and Devices Magazine, 1989.

[8] I. Koren and Z. Koren. Defect tolerance in VLSI circuits: techniques and yield analysis. Proceedings of the IEEE, 1998.

[9] Andrea Saltelli et al. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. Computer Physics Communications, 2010.

[10] Jonathan Balkind et al. OpenPiton+Ariane: The First Open-Source, SMP Linux-booting RISC-V System Scaling From One to Many Cores. 2019.

[11] E. Beyne et al. Scalable, sub 2um pitch, Cu/SiCN to Cu/SiCN hybrid wafer-to-wafer bonding technology. IEDM, 2017.

[12] L. Bamberg et al. Macro-3D: A Physical Design Methodology for Face-to-Face-Stacked Heterogeneous 3D ICs. In DATE, 2020.