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

Anthony Agnesina¹, Moritz Brunion², Jinwoo Kim¹, Alberto Garcia-Ortiz², Dragomir Milojovic³, Francky Catthoor³, Manu Perumkunnil³ and Sung Kyu Lim¹

¹School of ECE, Georgia Institute of Technology, Atlanta, USA
²ITEM.ids, University of Bremen, Germany
³IMEC, 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 $\gamma$.

• 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.

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 $\gamma$.

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 $4M_x 2M_y 1M_z$ 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

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>wafer diameter</td>
<td>foundry related parameters</td>
</tr>
<tr>
<td>α_FEOL, Logic</td>
<td>cost of logic wafer</td>
</tr>
<tr>
<td>α_MOL, Logic</td>
<td>MOL cost of logic wafer</td>
</tr>
<tr>
<td>D_logic</td>
<td>logic die defect density</td>
</tr>
<tr>
<td>D_logic clustering parameter</td>
<td>logic die clustering parameter</td>
</tr>
<tr>
<td>α_FEOL, Memory</td>
<td>cost of memory wafer</td>
</tr>
<tr>
<td>α_MOL, Memory</td>
<td>MOL cost of memory wafer</td>
</tr>
<tr>
<td>D_Memory</td>
<td>memory die defect density</td>
</tr>
<tr>
<td>D_Memory clustering parameter</td>
<td>memory die clustering parameter</td>
</tr>
<tr>
<td>y_B</td>
<td>3D integration yield</td>
</tr>
<tr>
<td>c_B</td>
<td>3D integration cost</td>
</tr>
<tr>
<td>c_M (L)</td>
<td>cost of metal layer L in BEOL of configuration C</td>
</tr>
</tbody>
</table>

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

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:

\[
r_D(a, \gamma) = \frac{C_{D_{2D}}(\gamma a)}{C_{D_{2D}}(a)} = \frac{C_{W_{2D}}}{k_d g_{2D}(\gamma a)} \cdot \frac{k_d g_{2D}(\gamma a)}{C_{W_{2D}}} = \frac{C_{W_{2D}}}{C_{W_{2D}}} \cdot \frac{dpw(a)}{dpw(\gamma a)} \cdot \frac{y_{2D}(a)}{y_{2D}(\gamma a)} \tag{1}
\]

where \( k_d, dpw, \gamma \) 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 \( \gamma \) 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 \( \gamma \). 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 \( \gamma \).

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:

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

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:

\[
y_{\text{die}}(a) = \left( 1 + \frac{D_a}{\alpha} \right)^{-\alpha} \tag{4}
\]
where $D$ is the defect density and $\alpha$ the clustering parameter of the faults. Parameter $\alpha$ 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:

$$y_{\text{stack}}(a) = \prod_{i=1}^{\#\text{dies}} y_{\text{die},i}(a)$$  \hspace{1cm} (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:

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

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

$$C_{W_{2D}}(N, C) = c_{FEOL} + c_{MOL} + \sum_{i=1}^{N} c_{BEOL}(N, C)$$  \hspace{1cm} (7)

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

$$C_{W_{3D}} = c_B + C_{W_{2D}}(N_1, C_1) + C_{W_{2D}}(N_2, C_2)$$  \hspace{1cm} (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.

---

**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 $\gamma$ below 0.5 has a very positive impact on the 3D yield.

---

**A. Die-Per-Wafer Ratio**

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

$$r_{DPW}(a, \gamma) = \frac{dpw(a)}{dpw(\gamma a)} \approx 2^\sqrt{\frac{w_d}{\gamma a}} - 1$$  \hspace{1cm} (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 $\gamma$ bounds the DPW ratio.

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

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

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

**B. Yield Ratio**

Eq. 6 can be rewritten as:

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

We see that

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

always holds, revealing a monotonic relationship between $\gamma$ and the yield ratio, where decreasing $\gamma$ 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:

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

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 $\gamma$ 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:

$$r_W = 2 + \frac{c_B}{C_{W2D}(N, C_{Logic})} \geq 2$$

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:

$$r_W = 1 + \frac{c_B}{C_{W2D}(N, C_{Logic})} + \frac{C_{W2D}(N, C_{Memory})}{C_{W2D}(N, C_{Logic})}$$

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 $V(r_D)$ is decomposed into parts attributable to our cost parameters. The metric of choice is the sensitivity index:

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

where $\sum_p S_p = 1$ and $V_p = V_p \{E_{x_p}(r_D \mid X_i)\}$. $V_p$ 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 $\Delta p$ to match $S_p = S_{FEOL, Logic} \pm 3\%$ when $\Delta \gamma_{FEOL, Logic} = 5\% \cdot c_{FEOL, Logic}$.

We record the resulting $\Delta p$’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 $\gamma$ and $y_B$, in that their very small variations have large effects on the cost. The sensitivity to $a$, $c_{exp. Mx}$, $c_{med. My}$ and $c_{cheap Mz}$ 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 $\gamma$. 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 $\gamma$ 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 $\gamma$ 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 \( \gamma \) COST.

<table>
<thead>
<tr>
<th>parameter</th>
<th>( \gamma ) values</th>
</tr>
</thead>
<tbody>
<tr>
<td>nominal ( p_0 )</td>
<td>( [0.46, 1.14] )</td>
</tr>
<tr>
<td>( \Delta p )</td>
<td>( p_0 \pm 10% p_0 )</td>
</tr>
<tr>
<td>( \Delta p / p_0 )</td>
<td>( p_0 \pm 20% p_0 )</td>
</tr>
</tbody>
</table>

IV. EXPLORATION OF \( \gamma \) WITH FEEDBACK FROM 3D FLOW

The previous analysis brought out the importance of the value of \( \gamma \) 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 \( \gamma \) reduction in a 3D memory-on-logic design flow. We show that the footprint of our 3D designs can be reduced to achieve \( \gamma \) values small enough to make 3D more attractive cost-wise, 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 0.5\( \mu \text{m} \times 0.5\mu \text{m} \), 1.0\( \mu \text{m} \), 0.5\( \Omega \) and 1\( fF \), 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 \( \gamma \) axis

To highlight area savings opportunities, we decide on a large design of 25 tiles connected in a 5\( \times \)5 mesh topology. Single-tile designs are implemented first with an iso-performance target of 500 MHz at the slow corner.

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

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

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.

<table>
<thead>
<tr>
<th>parameter</th>
<th>variation range</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_0 )</td>
<td>( [0.1\mu \text{m}^2, 500\mu \text{m}^2] )</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>( p_0 \pm 10% p_0 )</td>
</tr>
<tr>
<td>( \delta )</td>
<td>( p_0 \pm 20% p_0 )</td>
</tr>
<tr>
<td>( \delta / p_0 )</td>
<td>( p_0 \pm 20% p_0 )</td>
</tr>
<tr>
<td>( \beta )</td>
<td>( [0.90, 0.99] )</td>
</tr>
<tr>
<td>( \epsilon _\text{Mx} )</td>
<td>( p_0 \pm 30% p_0 )</td>
</tr>
<tr>
<td>( \epsilon _\text{My} )</td>
<td>( p_0 \pm 30% p_0 )</td>
</tr>
<tr>
<td>( \epsilon _\text{Mz} )</td>
<td>( p_0 \pm 30% p_0 )</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>( p_0 \pm 20% p_0 )</td>
</tr>
<tr>
<td>( \delta )</td>
<td>( p_0 \pm 20% p_0 )</td>
</tr>
<tr>
<td>( \delta / p_0 )</td>
<td>( p_0 \pm 20% p_0 )</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>( p_0 \pm 20% p_0 )</td>
</tr>
</tbody>
</table>
The absolute values are dependent on foundry parameters, the results clearly showcase the potential PPAC improvements in 3D when improving the \( \gamma \) 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 PPAC benefits with respect to the 2D implementation with 33.4\( \text{mm}^2 \) die size, thanks to silicon area reduction.

References