Cost and Thermal Analysis of High-Performance 2.5D and 3D Integrated Circuit Design Space

Dylan Stow, Itir Akgun, Russell Barnes, Peng Gu, Yuan Xie
Electrical and Computer Engineering
University of California, Santa Barbara
Email: {dstow,yuanxie}@ece.ucsb.edu

***Abstract*—3D integration is a promising technology to continue the trend of Moore's law. However, higher density from die stacking introduces thermal challenges that require more expensive packaging and cooling solutions. An alternative integration technology is interposer-based 2.5D design, which has fewer thermal issues but adds extra interposer cost. Designers must be aware of the system-level cost benefits of these choices early in the design process. This paper presents a cost analysis model with wafer costs, 3D bonding costs, and thermal modeling for the optimization of package and cooling costs. The cost model is used to explore the design space of integrated circuits to determine cost-driven enabling points of 2.5D and 3D integration under consideration of design size and power density. Our results suggest that proper use of die-integration technologies can realize substantial cost savings over traditional 2D design, even with the inclusion of packaging and cooling costs. When thermal properties are considered, interposer-based 2.5D integration is predicted to be more cost effective than TSV-based 3D integration, especially when power density is high.**

I. Introduction

Three-dimensional integrated circuit design (3D IC) is a promising technology that offers performance, power, and footprint improvements by vertically integrating multiple dies [1]. With 3D die-stacking, IC designs are partitioned across layers for performance and fabrication cost improvements. Partitioning can reduce interconnect distance between units, reducing interconnect delay and power. Smaller die sizes improve production yields and allow for higher transistor counts, extending Moore's law even as transistor scaling becomes increasingly difficult.

Despite the opportunities and benefits of 3D IC design, new challenges are introduced. High transistor density of vertical stacking leads to elevated power density and higher die temperatures, requiring more expensive packaging and cooling solutions. Through-silicon vias (TSV) also impose area overhead and require floorplan constraints to ensure connections between layers.

Consequently, even though 3D IC design and architecture have been explored for more than a decade [1], [2], interposer-based 2.5D integration was recently explored as an alternative technology to TSV-based 3D integration. Interposer-based 2.5D integration provides the benefit of close die integration with fewer design and thermal requirements. It also decouples the design of CPU/GPU and the design of the memory stack, reducing the design complexity with great flexibility. As a result, industry has adopted such 2.5D approaches in commercial products, such as Xilinx's FPGA [3] and the AMD Fury X GPU [4].

When a design strategy (either 2D or 2.5D or 3D design) has to be made, all benefits ultimately have to be justified with cost evaluation. Consequently, system-level cost analysis at early design stages is imperative to decide on whether 2.5D or 3-D integration should be adopted.

Realizing the importance of cost analysis, cost models for 3D integration have previously been proposed [5] to estimate fabrication costs. However, prior work was done a half-decade ago, considered technology nodes up to 45nm and a maximum of 200M gates, and only compared TSV-based 3D integration. With advances in fabrication technology and multi-billion transistors per chip, as well as the emergence of interposer-based products, it warrants to revisit the cost analysis methodology with the inclusion of 2.5D technology and with adjustments to area and cost estimation methodology to properly perform cost analysis in modern processes. In addition, previous models have provided insight into cost-driven design decisions, but have not included a flexible thermal model for measuring packaging and cooling costs across the range of possible IC designs.

In this paper, a cost model is presented for 2D, 2.5D, and 3D ICs to determine the optimal design choice for minimizing silicon fabrication cost. This model includes estimation and calculation for area, metal layer count, yield, and die cost. A thermal model is included to determine packaging and cooling costs. We then use this model to explore and characterize the design space for the emerging integration options. We also present the best choice between different 2D, 2.5D, and 3D partitioning schemes across the range of high-performance power densities and gate counts.

II. Cost Analysis Methodology

In this section, we present the proposed cost analysis methodology, including the die cost model, fabrication cost exploration, and package/cooling cost model.

A. Die Cost Estimation Model

Widespread adoption of 2.5D and 3D IC technology requires cost effectiveness to justify the risks of new design and production methodologies. Therefore, it is critical that a product's system cost is estimated early to demonstrate the benefits of new packaging arrangements. The total production cost of an integrated circuit is influenced by the costs of the silicon die, the packaging, and the required cooling solution. While each component should be studied at the onset of the design process, the silicon die generally has the greatest impact on system cost and constraints. Silicon die cost is dependent on the die area and on manufacturing details, including process technology and metal layer count. In the following sections, we detail methodologies for determining the required area and metal layer count given a design size and process technology. These models are then applied towards the calculations of the silicon yields and costs of 2D, 2.5D, and 3D circuits.

1) Area Estimation: Die area influences dies per wafer and die yield and is a critical parameter for silicon cost. The die area can be estimated from the number of gates in the design and the selected process node with the equation:

$$A_{die} = N_g * \beta\lambda^2 \quad (1)$$

978-1-4673-9039-2/16 $31.00 © 2016 IEEE
DOI 10.1109/ISVLSI.2016.133

where N_g is the number of gates, λ is the feature size, and β is an empirical scaling term where $\beta\lambda^2$ is the average area per gate. Previous work used data from commericiall designs to estimate a single value of β [5]. However, a survey of modern designs reveals considerable variation between markets, resulting in incorrect area estimation of up to 3x the actual area when using prior scaling values. Additionally, the formula assumes ideal scaling between each process generation. In reality, the gate pitch and minimum bit cell sizes have both scaled slower in process technologies since 28nm [6][7]. We propose that λ be adjusted to the effective feature size, as calculated from actual gate pitch and bit cell scaling rather than the advertised feature size. Table I presents scaling coefficients for several markets, illustrating the variability in average gate sizing, with data surveyed from 90nm to 14nm processes. Average power densities for different markets are also included.

Design Type	Scaling Coefficient β (M)	Power Density (W/mm^2)
CPU (desktop)	720	0.45
CPU (mobile)	610	0.24
CPU (server)	670	0.44
GPU (desktop)	440	0.47
GPU (mobile)	450	0.40
GPU (server)	440	0.33
Desktop SoC	840	0.27
Mobile SoC	710	0.19

Table I: Gate sizing coefficients and average power densities for commercial products from 90nm to 14nm.

2) Metal Layer Estimation: The metal layer count is also an important parameter for die cost, as additional layers require extra fabrication steps and resources. The number of required layers depends on the interconnect distance that must be routed in the design and therefore depends on design complexity. The range of available layers may be limited by the foundry, but estimation methodology can predict the required number of metal layers in new designs. First, Rent's rule can be used to estimate the average wire length [8]:

$$\bar{L} = \frac{2}{9}\left(\frac{1-4^{p-1}}{1-N_g^{p-1}}\right)\left(\frac{7N_g^{p-0.5}-1}{4^{p-0.5}-1} - \frac{1-N_g^{p-1.5}}{1-4^{p-1.5}}\right) \quad (2)$$

where p is Rent's exponent value that expresses the route complexity. The number of metal layers can then be approximated from the average wire length with the equation:

$$N_{metal} = \frac{f.o.N_g\bar{L}\omega}{\eta A_{die}} \quad (3)$$

where N_{metal} is the number of required metal layers, $f.o.$ is the average fanout, ω is the wire pitch, and η is the average interconnect utilization rate with consideration of percent metalization and overheads of vias and power and clock tracks. This formula assumes a uniform metal pitch across metal layers, but if specific metal stack wire dimensions are known, layer-based assignment with variable wire pitch and utilization can be employed [5][9].

Metal layer estimation is also useful for anticipating the reduction in meta layerl count from die partitioning in a 2.5D and 3D design. As shown above in Table II, partitioning a design into multiple dies can reduce the number of required layers per die, thus decreasing the wafer cost.

Area(mm^2)	Gate Count	1 die	2 dies	3 dies	4 dies
5	21	7	7	6	6
10	41	8	7	7	7
25	103	9	8	8	7
50	207	9	9	8	8
100	413	10	9	9	9
250	1033	11	10	10	9
500	2065	12	11	11	10

Table II: Estimated metal layer counts of single and partitioned die with 14nm process, $\beta = 650M$, and $\eta = 0.3$

3) 2D Yield and Cost Model: The cost of an individual silicon die prior to any additional steps for 3D integration can be estimated from the process technology, area, and metal layer count. To determine the die cost, the first step is to model the cost per wafer, which depends upon the process details, wafer diameter, and foundry vendor. This process technology choice is often determined early in the design process and is an input to this model. Within each process, the price per wafer can vary by the number of required metal layers, as extra metal layers require additional processing steps. The cost per wafer can be calculated from:

$$C_{wafer} = C_{process} + N_{metal} * C_{metal} \quad (4)$$

where C_{wafer} is the total wafer cost, $C_{process}$ is the base cost per wafer, and C_{metal} is the additional cost per metal layer. The values employed in our model were calculated using a cost model from industry [10].

Each fabricated wafer contains a finite number of silicon dies within its area, as determined by the wafer diameter and die area. The number of dies per wafer is calculated by:

$$N_{die} = \frac{\pi \times (\phi_{wafer}/2)^2}{A_{die}} - \frac{\pi \times \phi_{wafer}}{\sqrt{2 \times A_{die}}} \quad (5)$$

where N_{die} is the number of dies per wafer, ϕ_{wafer} is the wafer diameter, and A_{die} is the die area.

For a given wafer, only a percentage of dies will properly yield after fabrication. Assuming a negative binomial yield model [11], the die yield is calculated from:

$$Y_{die} = Y_{wafer} \times \left(1 + \frac{A_{die}D_0}{\alpha}\right)^{-\alpha} \quad (6)$$

where Y_{die} is the die yield and D_0 is the defect density, which is determined by the process. For our model, we select $\alpha = 3$ and therefore use the Dingwall yield model [12]. Note that as area increases, yield rapidly decreases, encouraging an SoC design of multiple small dies over a monolithic die.

If C_{test} is the die testing cost, the final die cost C_{die} is:

$$C_{die} = (\frac{C_{wafer}}{N_{die}} + C_{test})/Y_{die} \quad (7)$$

4) 3D Yield and Cost Model: Based on Equations (6) and (7), for complex circuits with high transistor count and large area, a monolithic 2D die will have a lower yield and will thus be expensive to produce. Partitioning a large design into multiple small dies will improve the yield per die, as shown earlier in Equation 6, and can reduce the overall silicon cost, but these dies must be tightly integrated to maintain performance. 3D IC includes several techniques for creating vertical interconnects between stacked die. The most mature technique is Through-Silicon Via (TSV) technology, which provides high-bandwidth connections

between layers with latency that matches on-die global routes [13]. Although silicon yields improve from partitioning, 3D integration also introduces manufacturing overheads that may add to silicon costs, including wafer thinning, via production, die bonding, and extra TSV area overheads.

For this model, dies in a 3D stack are assumed to use face-to-back arrangement, which allows for stacking beyond 2 layers at the expense of additional via area. TSVs, with manufactured diameter below 1 μm [13], introduce area overheads to the die that must be considered for accurate cost modeling. The number of TSVs between two layers will depend on partitioning decisions and circuit organization. For many circuit designs, the number of TSVs will be set by global interconnect buses between layers. From the known TSV count X_{TSV}, the total adjusted silicon area for a die in the 3D stack A_{3D} is:

$$A_{3D} = A_{die} + X_{TSV} A_{TSV} \tag{8}$$

where A_{die} is the total original silicon area and A_{TSV} is the area per TSV, including keep-out boundary. As TSVs block routable area, only A_{die} is available for metal interconnect.

In an ASIC design, the via count between two layers can be estimated from the respective gate counts (N_1 and N_2) and Rent's Rule coefficients (k_1, k_2, p_1, and p_2):

$$\begin{aligned} X_{TSV} =& \alpha k_{1,2}(N_1+N_2)(1-(N_1+N_2)^{p_{1,2}-1}) - \\ & \alpha k_1 N_1 (1 - N_1^{p_1-1}) - \alpha k_2 N_2 (1 - N_2^{p_2-1}) \end{aligned} \tag{9}$$

where $k_{1,2}$ and $p_{1,2}$ are equivalent Rent coefficients [14].

To calculate the cost of a 3D die stack, the individual die costs are first calculated. For each die in the stack, the cost is increased by extra TSV area overhead and additional process costs for wafer thinning and TSV processing. This model assumes die-to-wafer stacking and known good die testing (KGD) before bonding, which have been shown to reduce net cost when die yields are low [15]. It also assumes no testing between bonding steps, which has also been shown to be cost effective when bond yields are high [16].

The net cost of the 3D stack C_{3D} is calculated from:

$$C_{3D} = \frac{\sum_{i=1}^{n} (\frac{C_i}{y_i}) + (n-1)C_{bond}}{Y_{bond}^{n-1}} \tag{10}$$

where Y_{bond} is the bond yield, n is the number of die, C_i and y_i are the silicon cost and yield of a given die, and C_{bond} is the cost of alignment and bonding between die. Note that die layers that require TSVs will have higher process costs $C_{process}$ during the wafer cost C_{wafer} calculation.

5) 2.5D Yield and Cost Model: 2.5D packaging may be used to reduce system cost in the same manner as 3D integration: through the partitioning of large die into multiple small dies to improve yield and to reduce the required metal layer count. These savings will be reduced by the overhead of interposer silicon cost, but interposer costs are significantly less than die costs for comparable areas due to the lack of active transistors and small number of routing layers. Figure 1 compares the costs of a 65 nm interposer process and a 65nm CMOS logic process with 7 metal layers to illustrate the relative price difference.

The cost of the 2.5D silicon stack can be calculated as:

$$C_{2.5D} = \frac{\frac{C_{int}}{y_{int}} + \sum_{i=1}^{n} (\frac{C_i}{y_i} + C_{bond_i})}{Y_{bond}^{n-1}} \tag{11}$$

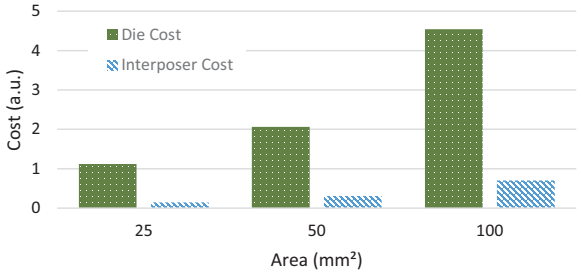

Figure 1: Costs of 65nm interposer and CMOS die vs. area

where $C_{2.5D}$ is the net silicon cost of the 2.5D stack, Y_{bond} is the bond yield between a die or die stack and the interposer, n is the number of die/stacks being bonded to the interposer, C_{int} and y_{int} are the interposer silicon cost and yield, C_i and y_i are silicon cost and yield of a given die or stack, and C_{bond_i} is the bond cost for a given die or stack, which can vary with required accuracy and other manufacturing considerations. For our model, we again assume known good die testing before bonding to the interposer and no die testing between bonding steps. Required interposer area can be approximated as the sum of the footprint areas of the bonded die and die stacks. This model does not consider die bonding on both sides of the interposer, which reduces required interposer area but requires vertical interconnect spacers to attach the interposer to the substrate and introduces cost and thermal complexity outside of the scope of this model.

B. Fabrication Cost Exploration

With cost estimation methodology for 2D, 2.5D, and 3D integrated circuits, the silicon fabrication costs at different design sizes can be compared to determine the enabling points of the different die integration technologies. Unless otherwise noted, the parameters outlined in Table III are assumed for cost estimation.

Table III: Assumed values for design exploration.

Feature Size (λ)	14 nm	Y_{wafer}	98%
Area Scaling (β)	650M	Y_{bond}	99%
Rent's Coefficient (k)	4.0	D_{TSV}	1 μm
Rent's Exponent (p)	0.6	$D_{\mu bump}$	25 μm
Metal Utilization (η)	30%	Interposer Feature Size	65 nm
Gate Pitch	4.5 × λ	Average Fan-out ($f.o.$)	4
Wire Pitch	3.6 × λ	Defect Density (D_0)	0.2-0.3

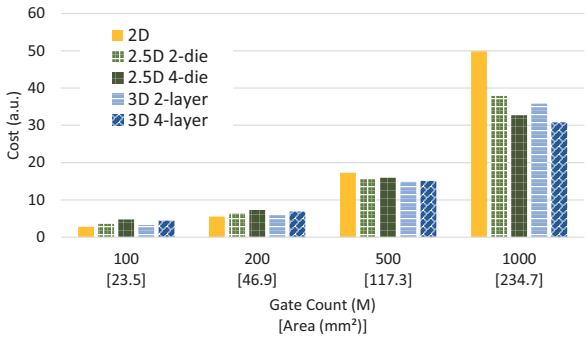

Figure 2: 14nm cost vs. gate count for 2D, 2.5D, and 3D

Figure 2 shows the relative fabrication costs for designs of various gate counts using 2D fabrication, interposer-based 2.5D

fabrication with the design partitioned into either 2 or 4 smaller dies, and TSV-based 3D integration with the design partitioned into either 2 or 4 layers. Fabrication cost for the 2.5D circuits includes costs and yields for the interposer and bonding steps. Costs of the 3D circuits includes process overhead for the addition of TSVs and extra thinning, TSV area overhead, and bonding costs and yields.

As gate count increases, both 2.5D and 3D circuits become more cost effective than single-die designs because of the area-dependent yield trend described in Equation (6). For the same number of die partitions, 3D fabrication is lower cost than 2.5D fabrication because of the interposer silicon overhead, which, although much cheaper than the active silicon, also exhibits reduced yield at large die area. Table V shows, for multiple bond yields, the number of gates at which 2.5D and 3D integration become cheaper to fabricate than single-die designs. The enabling points are also dependent upon the process technology, as shown in Table IV. Figures 3 and 4 show the relative cost contributions of different fabrication factors, including die cost, testing cost, interposer cost, TSV overhead, and bonding cost, at two different design sizes. The relatively high cost enabling points, in terms of gate count and area, of 2.5D and 3D integration confines their cost effective use to high-performance IC markets with greater design complexity.

Table IV: 2.5D/3D cost enabling points across processes

		Gate Count (M)	2D Area (mm^2)
16 nm	2.5D	262	75.1
	3D	177	50.7
28 nm	2.5D	231	117.7
	3D	133	67.8
40 nm	2.5D	107	111.3
	3D	87	90.5

Table V: 14nm enabling points, in gate count, vs. bond yield

Bond Yield (%)	2.5D 2-die	2.5D 3-die	2.5D 4-die	3D 2-layer	3D 3-layer	3D 4-layer
0.99	325 M	361 M	376 M	262 M	270 M	326 M
0.95	481 M	536 M	615 M	288 M	394 M	487 M
0.90	747 M	770 M	923 M	383 M	555 M	666 M

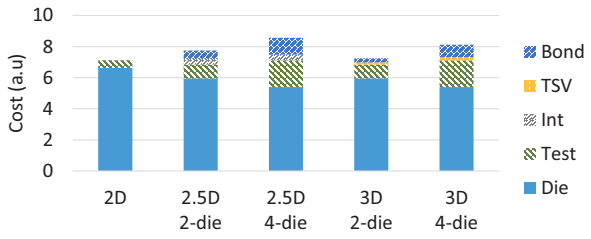

Figure 3: Cost contributions in 14nm with 250M gates

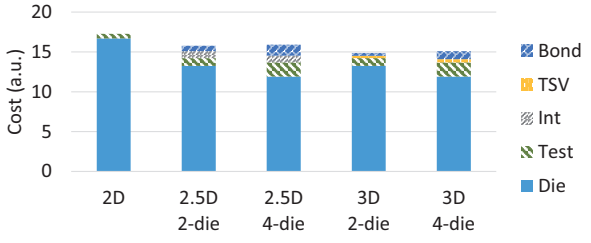

Figure 4: Cost contributions in 14nm with 500M gates

C. Package and Cooling Cost Model

Although the integrated circuit silicon receives the most design focus, a cost-driven design must also consider the system-level costs of packaging and cooling. Rising peak die temperatures compromise circuit robustness and disrupt timing requirements, and thus require more efficient and expensive packaging and cooling solutions. These problems are exacerbated by higher power density and thermal insulation from vertical integration in 3D designs. Therefore, a flexible and accurate thermal model is necessary to select between 2D, 2.5D, and 3D integration.

1) 2D Thermal Model: The one-dimensional heat equation for a 2D die is given by:

$$T_{2D} = T_{ambient} + (\Theta_{JC} + \Theta_{CS} + \Theta_{SA}) \times P + \Theta_{Si} \times P \quad (12)$$

where $T_{ambient}$ is the ambient temperature in $(^{\circ}C)$, Θ_{JC} is the junction-to-case thermal resistance, Θ_{CS} is the thermal resistance of the interface compound between the case and heat sink, Θ_{SA} is the thermal resistance between the heat sink and ambient with units in $(^{\circ}C/W)$, P is the power dissipation, and Θ_{Si} is the thermal resistance of the silicon layer, where the die is integrated face down. For face-up wire bond packaging, the thermal resistance of the silicon layer is replaced with Θ_{CuILD}, the thermal resistance of the metal layers with 50% metalization. For this section, it is assumed that the dies are integrated face down unless noted.

Previous studies [17][18] assume heat removal only from the bottom surface of the die via the package and board. However, greater than 90% of heat in high-performance designs may be transferred out of the heat sink [19]. Moreover, package junction-to-ambient thermal resistance Θ_{JA} values from industry are inadequate for this power range. Therefore, the thermal model assumes an external heat sink and disregards heat removal from the bottom surface. The heat escape path is vertical, with active cooling placed on the top of the chip. Power is generated between the Si substrate and metal layers. Thermal resistances Θ_{JC}, Θ_{CS}, and Θ_{SA} contribute to junction-to-ambient temperature, while Θ_{Si} contributes to the junction temperature. Equation (12) includes these resistances in series to make up the effective thermal resistance.

The choice of package and heat sink is vital for cooling a high-power die, as both contribute significantly to the maximum average temperature of a chip. We integrate the thermal model into our cost model in order to estimate package and cooling costs. Assuming that the chip can reach the allowed maximum temperature T_{max}, we can find the most cost-effective package and heat sink combination that satisfies this constraint.

2) 3D Thermal Model: In order to estimate the temperature increase due to stacking multiple active layers, the 2D thermal model is expanded to include power generation at each layer and thermal resistances between stacked dies. The maximum average die temperature is observed at the layer farthest away from the heat sink. The one-dimensional heat equation of a 3D stacked die with n active layers, is therefore given by [17]:

$$T_{3D} = T_{ambient} + \sum_{i=1}^{n} (\Theta_D (\sum_{j=i}^{n} P_j)) \quad (13)$$

where Θ_D is the thermal resistance between the $(i-1)$ and i^{th} layers, calculated as:

$$\Theta_D = \begin{cases} \Theta_{JC} + \Theta_{CS} + \Theta_{SA} + \Theta_{Si}, & \text{if } i = 1. \\ \Theta_{Si} + \Theta_{glue} + \Theta_{CuILD}, & \text{if } i \neq 1. \end{cases} \quad (14)$$

Figure 5 illustrates the 3D thermal model. The thermal resistance between 3D-stacked dies takes into account the resistances of silicon, glue, and metal layers. According to Equation (14), the die temperature of lower layers are also affected by the power dissipation and the effective thermal resistance of the layers above. Compared to conventional dies, 3D integration results in higher die temperatures and requires better packaging and cooling to maintain the same maximum allowed die temperature.

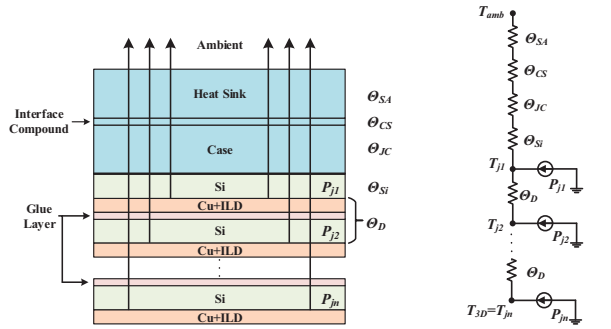

Figure 5: 3D thermal model representation. (a) Schematic of the 3D IC thermals; (b) Effective thermal resistance of 3D IC.

3) 2.5D Thermal Model: In order to estimate the maximum temperature of a 2.5D die-on-silicon interposer, the 3D thermal model is expanded to consider multiple die stacks. Figure 6 describes the 2.5D thermal model in which separate stacks share the same junction-to-ambient thermal pathway but have different stack temperatures in *parallel*. The stack with the maximum die temperature determines the upper bound for cooling and package cost of 2.5D integration.

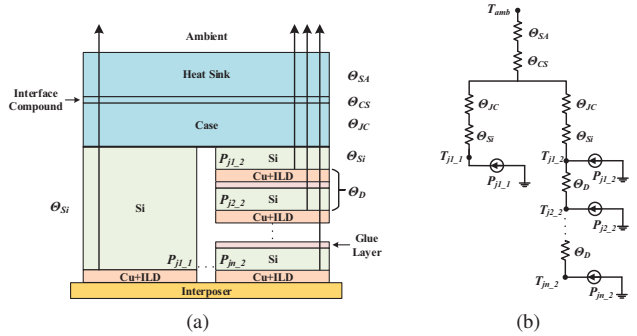

Figure 6: 2.5D thermal model representation. (a) Schematic of the 2.5D IC thermals; (b) Effective thermal resistance of 2.5D IC.

4) Cooling Cost Estimation: The choices of packaging and cooling are major contributors to the final system cost of high-performance integrated circuits. As shown in Equation (13), the package and heat sink thermal resistances are modeled in series, and thus both resistances must be sufficiently low for proper operation of the chip. To determine heat sink cost and effectiveness, commercial hardware was surveyed across a design range that included passive heat sinks, heat sinks with fans, complex heat pipe coolers, and high-end liquid coolers [20] [21]. A continuous cost-versus-thermal resistance curve was extracted from the cooling solution data, shown in Figure 7, to estimate a heat sink cost given a required thermal resistance. Note the steep cost increase as thermal resistance approaches $\Theta_{SA} = 0.07\ ^\circ C/W$ and the lack of commercial solutions beyond this point, suggesting a limit to the currently available heat sink capability.

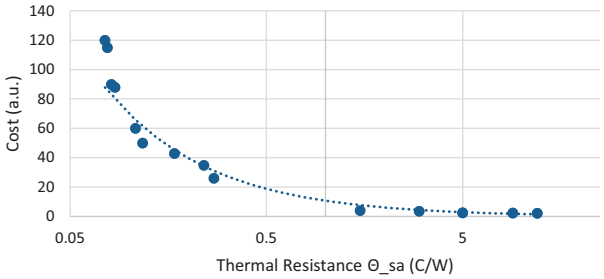

Figure 7: Cooling cost versus efficiency

5) Package Cost Model: Package cost is dependent on multiple factors, but the package technology type has the greatest influence on the thermal resistance and overall cost. A package type can be selected to meet thermal resistance requirements, which then determines the scaling of other package cost contributors. During system optimization, the model selects the most cost efficient package/heat sink combination to meet the required thermal resistance. Cost-efficient package types included are pBGA, fcBGA, and cBGA, with thermal resistances of 0.44, 0.20, and 0.03 $^\circ C/W$ [22].

Other cost parameters include package area, pin count, substrate layer count, and production volume. Package area is determined by chip area, pin count and substrate layer count are determined by electrical requirements, and package volume depends on the target market. The package cost $C_{package}$ can be calculated by:

$$C_{package} = \mu_V(\mu_L N_L)[C_{base} + \mu_A A + \mu_p * N_p] \quad (15)$$

where μ_V is a market volume scalar, N_l and μ_L are the substrate layer count and scaling, C_{base} is the base package cost for the selected type, A and μ_A are the chip area and scaling, and N_p and μ_p are the pin count and scaling.

III. Cost-Driven Design Space Exploration

Although our results show that the silicon cost of 3D integration is consistently less than that of 2.5D integration because of the interposer overhead, the introduction of thermal-dependent packaging and cooling costs results in a new cost-driven design space. Figure 8 shows the system costs across a range of gate counts and power densities for a 14nm process, pin count of 1150, max junction temperature of $100^\circ C$, and ambient temperature of $30^\circ C$. The patterned values in green reflect the best design choice at a given design size and power density. The value outlined in red is too hot to cool with conventional thermal management solutions.

For 14nm designs smaller than $100mm^2$, 2D design is the most cost effective because of minimal fabrication overheads and efficient cooling. 3D stacking is the most cost efficient only when power densities are at or below $0.4W/mm^2$. For reference, average mobile microprocessors have a power density of $0.2W/mm^2$

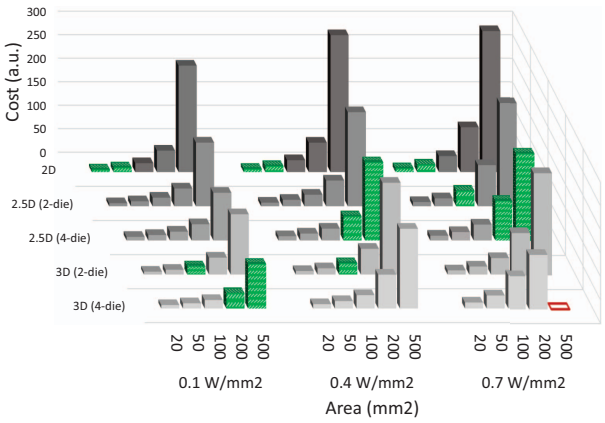

Figure 8: Full 14nm system costs of 2D, 2.5D, and 3D designs at different power densities

and desktop CPU and GPU parts have power densities from $0.3 - 1.0W/mm^2$. At all other power densities and gate counts, 2.5D integration is more cost-efficient than 2D and 3D integration because of the balance of yield improvements from die partitioning and reasonable thermal management. For closer inspection, Figure 9 shows the relative cost breakdown between chip fabrication and package/cooling costs at $0.4W/mm^2$ and $200mm^2$.

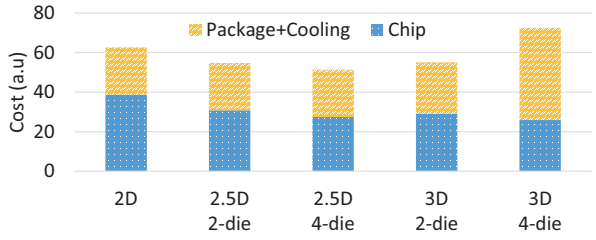

Figure 9: 14nm Cost Breakdown at $0.4W/mm^2$ and $200mm^2$

IV. Conclusion

Cost analysis at an early design stage is a key task towards determing the design strategy of using either monolithic 2D design, interposer-based 2.5D design, or TSV-based 3D integration. This work presents a system-level cost model to compare the silicon fabrication, packaging, and cooling costs between 2D, 2.5D, and 3D systems. The model can be integrated into design flows to enable cost-driven design space exploration.

Acknowledgment

This project is supported by DARPA DAHI program with ONR award N00014-16-1-2087. We thank Dr. Daniel Green for his insight and guidance on this project.

References

[1] J. Zhao, Q. Zou, and Y. Xie, "Overview of 3D architecture design opportunities and techniques," *IEEE Design & Test*, vol. PP, no. 99, 2015.

[2] Y. Xie, J. Cong, and S. Sapatnekar, *Three-dimensional IC: Design, CAD, and Architecture*. Springer, 2009.

[3] K. Saban, *Xilinx Stacked Silicon Interconnect Technology Delivers Breakthrough FPGA Capacity, Bandwidth, and Power Efficiency*. Xilinx white paper, 2012.

[4] B. Black, "Die-stacking is happening: AMD Fury X GPU," in *3D ASIP*, Dec 2015.

[5] X. Dong and Y. Xie, "System-level cost analysis and design exploration for three-dimensional integrated circuits (3D ICs)," in *ASP-DAC*, Jan 2009, pp. 234–241.

[6] T. Song, W. Rim, J. Jung *et al.*, "A 14nm FinFET 128Mb 6T SRAM with Vmin-enhancement techniques for low-power applications," in *ISSCC*, Feb 2014, pp. 232–233.

[7] S.-Y. Wu, C. Lin, M. Chiang *et al.*, "An enhanced 16nm CMOS technology featuring 2nd generation FinFET transistors and advanced Cu/low-k interconnect for low power and high performance applications," in *IEDM*, Dec 2014, pp. 3.1.1–3.1.4.

[8] W. Donath, "Placement and average interconnection lengths of computer logic," *IEEE Transactions on Circuits and Systems*, vol. 26, no. 4, pp. 272–277, Apr 1979.

[9] A. Kahng, S. Mantik, and D. Stroobandt, "Toward accurate models of achievable routing," *TCAD*, vol. 20, no. 5, pp. 648–659, May 2001.

[10] *IC Cost and Price Model, Revision 1506*, IC Knowledge LLC, 2015.

[11] J. Cunningham, "The use and evaluation of yield models in integrated circuit manufacturing," *IEEE Transactions on Semiconductor Manufacturing*, vol. 3, no. 2, pp. 60–71, May 1990.

[12] A. Dingwall, "High-yield-processed bipolar LSI arrays," in *IEDM*, vol. 14, 1968, pp. 82–82.

[13] J. P. Gambino, S. A. Adderly, and J. U. Knickerbocker, "An Overview of Through-silicon-via Technology and Manufacturing Challenges," *Microelectronic Engineering*, vol. 135, pp. 73–106, Mar. 2015.

[14] P. Zarkesh-Ha, J. Davis, W. Loh *et al.*, "On a pin versus gate relationship for heterogeneous systems: Heterogeneous Rent's rule," in *CICC*. IEEE, 1998, pp. 93–96.

[15] Y. Chen, D. Niu, Y. Xie, and K. Chakrabarty, "Cost-effective integration of three-dimensional (3D) ICs emphasizing testing cost analysis," in *ICCAD*, Nov 2010, pp. 471–476.

[16] M. Taouil, S. Hamdioui, E. Marinissen, and S. Bhawmik, "Using 3D-COSTAR for 2.5D test cost optimization," in *3DIC*, Oct 2013, pp. 1–8.

[17] S. Im and K. Banerjee, "Full chip thermal analysis of planar (2-D) and vertically integrated (3-D) high performance ICs," in *IEDM*, Dec 2000, pp. 727–730.

[18] G. Loi, B. Agrawal, N. Srivastava *et al.*, "A thermally-aware performance analysis of vertically integrated (3-D) processor-memory hierarchy," in *DAC*, 2006, pp. 991–996.

[19] J. Galloway, S. Bhopte, and C. Nelson, "Characterizing junction-to-case thermal resistance and its impact on end-use applications," in *ITherm*, May 2012, pp. 1342–1347.

[20] *Digikey*, 2015, www.digikey.com.

[21] *Closed Loop AIO Liquid Coolers*, Anandtech, 2014, http://www.anandtech.com/show/7738/closed-loop-aio-liquid-coolers.

[22] *Challenges in Measuring Theta jc for High Thermal Performance Packages*, Electronics Cooling, 2014, http://www.electronics-cooling.com/2014/05/challenges-measuring-theta-jc-high-thermal-performance-packages/.