Cost Analysis and Cost-Driven IP Reuse Methodology for SoC design Based on 2.5D/3D Integration

Invited Paper

Dylan Stow, Itir Akgun, Russell Barnes, Peng Gu, Yuan Xie
University of California, Santa Barbara
1155 Phelps Hall
Santa Barbara, California
{dstow,yuanxie}@ece.ucsb.edu

ABSTRACT

Due to the increasing fabrication and design complexity with new process nodes, the cost per transistor trend originally identified in Moore's Law is slowing when using traditional integration methods. However, emerging die-level integration technologies may be viable alternatives that can scale the number of transistors per integrated device while reducing the cost per transistor through yield improvements across multiple smaller dies. Additionally, the escalating overheads of non-recurring engineering costs like masks and verification can be curtailed through die integration-enabled reuse of intellectual property across heterogeneous process technologies. In this paper, we present an analytical cost model for 3D and interposer-based 2.5D die integration and employ it to demonstrate the potential cost reductions across semiconductor markets. We also propose a methodology and platform for IP reuse based on these integration technologies and explore the available reductions in overall product cost through reduction in non-recurring engineering effort.

1. INTRODUCTION

The semiconductor industry has seen tremendous growth because of the economic and performance scaling of increasing integration complexity, with each new process node delivering more transistors per silicon area with better performance and lower cost per transistor. By following the cadence of this scaling trend, formally recorded by Moore [10], the industry has been able to scale integrated circuits from several transistors to several billion, enabling the integration of floating point units, cache memories, multiple cores, graphics processing units, and power management units. In the maturing System-on-Chip era, a single die may additionally include modems, digital signal processors, heterogeneous cores, and application-specific accelerators to provide further system efficiency and market differentiation.

After several decades, however, these industry-defining scaling trends may be reaching their conclusions. Power scaling has slowed significantly, leading to the current "dark silicon" era of power-efficiency driven design. While the number of transistors per die continues to increase with smaller process nodes, many foundries have already failed to achieve the targeted area scaling per transistor and the cadence of new process technologies is expected to slow down for future nodes. Perhaps most importantly, Moore's famous observation on the cost per transistor may no longer hold, as the fabrication of sub-20nm FinFET transistors is sufficiently complex to require difficult and expensive fabrication technologies that translate to yield challenges and additional wafer cost. Thus, circuit designers and computer architects can no longer depend on the free availability of additional transistors and integration opportunity with each new process node. Additionally, non-recurring engineering costs have also increased quickly during the last several process nodes due to fabrication complexity issues, such as complex layout design rules and multiple masks per layer, and due to the system complexity challenges of verifying billions of logic gates.

Despite these major setbacks, alternate integration technologies may be able to function in tandem with traditional process node scaling to provide cost reductions and more transistors per circuit. Through die-level integration technologies, multiple circuit dies can be connected electrically and physically to produce a larger integrated circuit. These technologies include 3D die stacking with connections through either face-face micro-bumps or face-back Through-Silicon Vias (TSVs), or through interposer-based 2.5D integration that uses a large passive routed die to provide interconnect between dies and the substrate. By partitioning a monolithic SoC across multiple small die, yield per die can be greatly improved and metal layer count can be lowered, reducing the total IC cost if integration overheads are sufficiently low. Die-level integration also allows for new integration strategies, such as heterogeneous process integration with different process technologies between dies, that can be used to reduce costs or optimize performance. These technologies can also act as a platform for the reuse of hard intellectual property, allowing for the reconfiguration of SoCs through the integration of different die combinations while amortizing non-recurring overheads of design, verification, and masks.[1]

ICCAD '16,
© 2016 ACM. ISBN 978-1-4503-4466-1/16/11...$15.00
DOI: http://dx.doi.org/10.1145/2966986.2980095

[1] This project was supported by the DARPA DAHI program through ONR award N00014-16-1-2087. We thank Dr. Daniel Green for his help and guidance on this project.

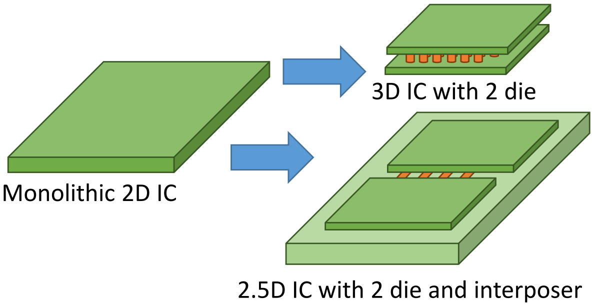

Figure 1: Die-level integration through TSV-based 3D and interposer-based 2.5D technologies.

In this paper, we present a high-level, analytical cost model for 3D and 2.5D integrated circuits and apply the model to current and future process technologies to demonstrate the feasible cost savings that are achievable through die-level integration. We also propose techniques for IP reuse that take advantage of these emerging integration technologies and provide analysis of achievable cost reductions across markets.

2. METHODOLOGY FOR COST ANALYSIS

In order to accurately study the relative costs of traditional 2D integrated circuits versus 3D or 2.5D die-level integrated circuits, an analytical cost model was developed to determine the approximate manufacturing cost per integrated circuit. The methodology includes flexible estimation of die area, metal layer count, and die yield to determine cost per individual die. The model extends to include 3D processing and bonding overhead, additional TSV area, 2.5D interposer cost, and packaging cost to estimate the final manufacturing cost of the packaged integrated circuit.

2.1 Die Cost Methodology

The cost of an individual die, whether the monolithic die in a traditional 2D circuit or one of several dies in a 3D or 2.5D circuit, can be estimated from a few parameters, allowing for a wide range of design costs to be studied and compared. The choice of process technology has a major impact on the die cost as it determines the cost per wafer, number of metal layers and cost per additional layer, the average area per transistor and gate, and the defect density. Once a process technology has been selected, the expected area can be calculated from the number of transistors or gates, with previous work assuming an average of four transistors per gate. In this paper, the area is estimated with the equation $A_{die} = N_g * \beta\lambda^2$ where N_g is the number of gates, λ is the feature size, and β is an empirical scaling term such that $\beta\lambda^2$ is the average area per gate. A value for β can be determined from a survey of previous market designs, with values ranging from 450 million for dense graphics processors, 700 million for consumer CPUs, and up to 850 million for some SoCs. As some sub-28nm foundry process nodes have scaled less than expected [11][12], the λ value should be adjusted to the true effective feature size.

With the process technology, area, and number of gates per die, it is then possible to estimate the required number of metal layers. Rent's rule, an observation on the number of nets crossing a boundary between groups of gates [8], can be extended to estimate the average metal length given the number of gates and routing efficiency [3], expressed as

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

where p is the Rent's exponent value that expresses the amount of route optimization [1]. The number of metal layers can then be approximated from the average wire length with the equation

$$N_{metal} = \frac{f.o.N_g\bar{L}\omega}{\eta A_{die}} \qquad (2)$$

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 minus the overheads of vias and power and clock tracks. The formula employed in this model assumes a uniform metal pitch across layers to maintain flexibility between processes, but layer-based assignment with variable wire pitch and utilization can be employed if known and specified [4][6].

Metal layer estimation is an important step in the cost model because the maximum number of metal layers impacts the cost per wafer. By partitioning a single large die into multiple smaller die, the number of metal layers per die can be reduced, helping to reduce the die cost. As shown in Table 1, multiple metal layers can be removed from a design with sufficient die partitioning.

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

After the determination of process technology, die area, and required metal layer count, the manufacturing cost of the die can be determined by calculating the number of die per wafer and the resulting die yield. First, the cost per wafer can be defined as $C_{wafer} = C_{process} + N_{metal} * C_{metal}$, where C_{wafer} is the total wafer cost, $C_{process}$ is the base wafer cost, and C_{metal} is an additional cost per metal layer. From a given wafer, the number of rectangular die per circular wafer is approximated by

$$N_{die} = \frac{\pi \times (\phi_{wafer}/2)^2}{A_{die}} - \frac{\pi \times \phi_{wafer}}{\sqrt{2 \times A_{die}}} \qquad (3)$$

where N_{die} is the number of dies per wafer, ϕ_{wafer} is the wafer diameter, and A_{die} is the die area. The process technology, and its maturity, also determines the defect density, or number of failure-causing defects per area. With the defect density D_0, the yield per die is

$$Y_{die} = Y_{wafer} \times \left(1 + \frac{A_{die}D_0}{\alpha}\right)^{-\alpha} \qquad (4)$$

where Y_{die} is the die yield and in this model $\alpha = 3$ [2]. Of important concern is the negative exponential trend of yield with area, as shown in Figure 2, that causes the yield of large die to rapidly diminish. By having more small dies on or across wafers, each defect will cause a lower percentage of the wafer to fail and yield per die will be higher. Therefore, this trend suggests the possibility of reducing manufacturing cost through die-level integration by achieving higher die yields. Finally, the cost per die can be calculated with

$$C_{die} = (\frac{C_{wafer}}{N_{die}} + C_{test})/Y_{die} \qquad (5)$$

where C_{test} is the cost to verify each die, which is dependent on the circuit complexity. An industry cost projection tool was used to calculate the specific process costs for each technology [5].

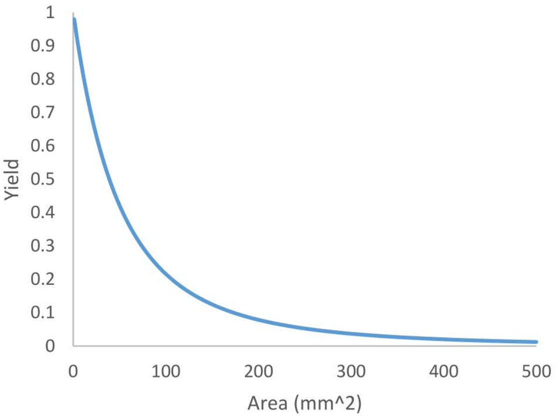

Figure 2: Yield versus Die Area with $D_0 = 0.3$ defects/cm^2

2.2 Die-Integration Cost Overheads

The manufacturing cost of a traditional 2D integrated circuit is the cost of the single silicon die. 3D and 2.5D integrated circuits, however, have additional fabrication overheads, and therefore additional wafer costs, that need to be accounted for in the analysis model. For 3D circuits with face-back integration and TSVs, it is necessary to first thin the dies to achieve the proper aspect ratio and to then fabricate the TSVs. The bonding process between dies also introduces extra process costs and potential yield decrease from improper alignment. Additionally, for many current integration techniques, TSVs will block off active device area, so a given partitioned die will need slightly more area for the X number of TSVs: $A_{3D} = A_{die} + X_{TSV}A_{TSV}$. The number of TSVs can be calculated by again using Rent's rule with the equation

$$\begin{aligned} X_{TSV} =& \alpha k_{1,2}(N_1 + N_2)(1 - (N_1 + N_2)^{p_{1,2}-1}) - \\ & \alpha k_1 N_1(1 - N_1^{p_1-1}) - \alpha k_2 N_2(1 - N_2^{p_2-1}) \end{aligned} \qquad (6)$$

where N_1 and N_2 are the gate counts in each layer, k_1, k_2, p_1, and p_2 are the Rent's Rule coefficients per layer, and $k_{1,2}$ and $p_{1,2}$ are equivalent Rent coefficients between layers [15]. Overall, the cost of a 3D integrated circuit can be estimated as

$$C_{3D} = \frac{\sum_{i=1}^{n}(\frac{C_i + C_O}{y_i}) + (n-1)C_{bond}}{Y_{bond}^{n-1}} \qquad (7)$$

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, C_O is the additional overhead cost per die for 3D processing, and C_{bond} is the cost of alignment and bonding between die.

Interposer-based 2.5D circuits must similarly include the overheads of bonding dies, but thinning and TSV creation are not necessary if the dies are not 3D stacks, since they can be connected with micro-bumps and copper pillars directly from their top medal layer to the interposer. Of course, these designs must also pay for the cost of the interposers, which are currently produced using mature BEOL technologies from standard foundry processes. The large interposer suffers from the same yield versus area trend as a CMOS die, but the passive nature of the interposer without active devices makes the yield and wafer costs of the interposer much better. Figure 3 shows the relative costs of a 65nm CMOS die and 65nm interposer at several different sizes, demonstrating the much lower interposer cost per area. Overall, the cost of the full 2.5D integrated circuit can be calculated with:

$$C_{2.5D} = \frac{\frac{C_{int}}{y_{int}} + \sum_{i=1}^{n}(\frac{C_i}{y_i} + C_{bond_i})}{Y_{bond}^{n-1}} \qquad (8)$$

where C_{int} and y_{int} are the interposer silicon cost and yield, calculated just as an active silicon die but with lower wafer cost and higher yields.

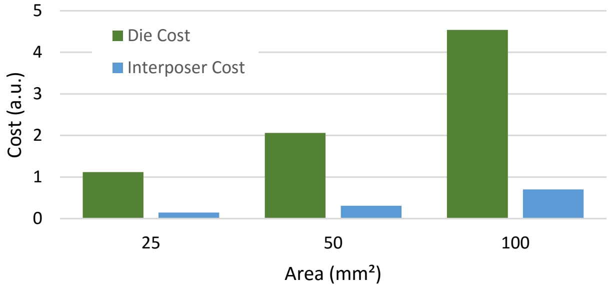

Figure 3: Costs of 65nm interposer and CMOS die vs. area

3. DIE COST ANALYSIS

Using the cost estimation methodology described in the previous section, we can compare the relative manufacturing costs of a given design across 2D, 2.5D, and 3D integration configurations. Using the assumed values stated in Table 2, Figure 4 shows the cost impact of partitioning a single 2D design into multiple equally-sized dies, either 2 or 4 dies, using either interposer-enabled 2.5D integration or TSV-based 3D integration. As evident in the figure, smaller designs with less area remain most cost effective as a single-die, due to the overheads of additional processing, bonding, and testing with die-level integration, as well as the TSV area for 3D and interposer cost for 2.5D. However, these overheads become less significant as the design size increases and the cost becomes dominated by the decreasing yield.

Relative breakdowns of the cost contributions at two design sizes can be seen in Figures 5 and 6. The points of equal cost between 2D and different die-level integration configurations are shown across process technologies in Table 3 and across bond yield in Table 4.

Table 2: 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

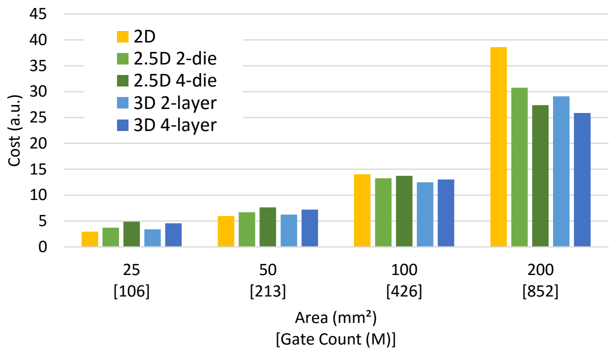

Figure 4: 14nm cost vs. gate count for 2D, 2.5D, and 3D

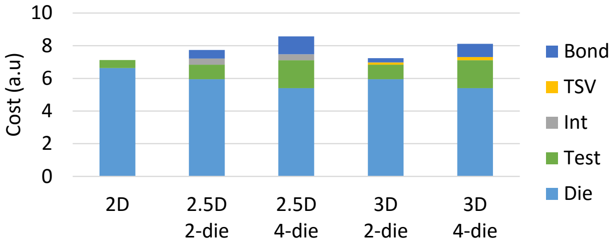

Figure 5: Cost contributions in 14nm with 250M gates

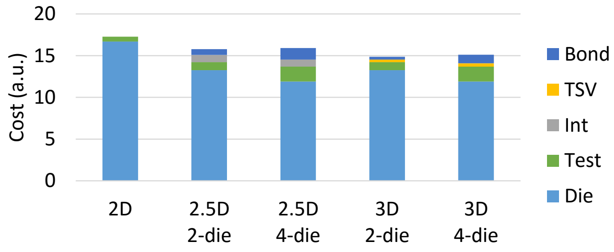

Figure 6: Cost contributions in 14nm with 500M gates

Table 3: 2.5D/3D cost enabling points vs. process

		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

4. NON-RECURRING ENGINEERING COST

The cost model and analysis presented in the previous sections cover the manufacturing cost of an integrated circuit, which can be examined without concern for specific product volume quantity. The total cost of an integrated circuit must also include the contribution of non-recurring engineering (NRE) costs that are paid once and then amortized across the number of produced circuits. For sufficiently large volumes, these non-recurring costs are distributed across all devices for a small impact on cost per device. Unfortunately, increases in design complexity with smaller process nodes and higher levels of integration are resulting in growing NRE costs. This includes physical complexity from smaller feature size, requiring multiple masks per layer and more difficult design rules, as well as system complexity from more transistors per integrated circuit, resulting in more difficult design and verification. Industry analysts have observed a rise in the design cost of a standard SoC by 2.7x between 28nm and 14nm designs, and anticipate a further increase to 9x, over $270 million, from 28nm to 7nm [9]. Similarly, projections from ARM suggest that because of non-recurring costs, average sized designs will be prohibitively expensive with volumes less than 10 million units and even at high volumes there will be an increase in cost per constant area [14]. To demonstrate the relative impact of NRE costs, Figure 7 uses the constant area die cost projections from [14] to break down the total die cost between silicon manufacturing cost and amortized NRE overhead for a sample 14nm chip at different product volumes. Similar but less extreme trends are seen for older process technologies with reduced process complexity, while future technology nodes will require even larger unit volumes to amortize the expensive non-recurring costs.

Table 4: 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

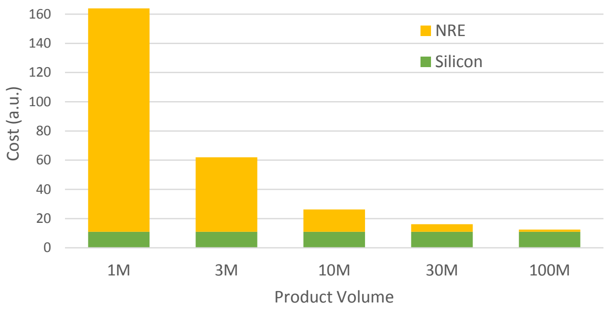

Figure 7: Total die cost breakdown between silicon fabrication and NRE overhead for sample 14nm chip

As demonstrated in the previous sections, die-level integration may be a cost-driven design method towards the reduction of manufacturing costs and the improvement of integrated circuit yields, especially for larger designs like SoCs. However, non-recurring engineering costs may still dominate the overall product cost, especially at lower volumes, if a design is only partitioned across multiple dies. To further reduce costs and combat the trend of increasing NRE, die-level integration may also be employed as a platform for the reuse of intellectual property at the die level. The dies of non-critical logic can be used across SoC designs to further amortize the initial non-recurring costs of mask and design and to greatly reduce the verification effort. Additionally, heterogeneous process technologies can be employed, so any critical logic can move to the highest performing node while reusable logic may remain in an older process technology with mature yields.

5. FLEXIBLE INTERCONNECT ARCHITECTURE FOR IP-REUSE

In traditional SOC integration, various IP blocks are connected with standardized bus (such as AMBA bus) or customized point-to-point interconnects. Such interconnect architecture is typically fixed for one particular design, and does not provide flexibility to be reused for future chip integration, because the interconnect fabrics and IP blocks are all fabricated on the same 2D chip. Such interconnect fabric is normally built for *providing best average case performance* across a generic set of applications. The reason being, no fixed on-chip network design efficiently supports all different types of communication requirements. Therefore it lacks flexibility and is unable to adapt to dynamically changing and special communication requirements at runtime. In addition, in 2D as well as 3D designs the interconnect fabric is tightly coupled with computing (core) and storage (cache) components. This pushes chip designers to adopt low complexity, structured and regular interconnect design to *limit pre-fabrication and post-fabrication verification efforts and costs.*

We propose to leverage 2.5D/3D integration technology to design flexible interconnect topology that can ensure existing or future IP block can be easily swapped, so that each device technology's unique performance characteristic can be preserved while enabling fast and compatible integration.

The reuse of intellectual property at the die-integration level may be realized through both 3D and 2.5D technologies. Reuse with TSV-based 3D integration would normally be limited by the predetermined placement and connectivity of the TSVs. In our work, the key concept is to adopt the *Network-on-Chip* concept to replace the traditional bus interconnect structure, and decouple the interconnect fabric from the IP blocks, and implement the interconnect architecture in a separate silicon layer called **interconnect service layer (ISL)** [13], either on the interposer (for 2.5D) or as an independent layer in 3D stacking. Such decoupling can provide reduced manufacture cost and offer more reliable and flexible interconnect layer compared to its traditional 2D counterparts. The decoupled ISL can contain multiple on-chip networks such as mesh, ring, hierarchical bus topologies, etc. With ISL the constraints on the on-chip network router area and link bandwidth can be relaxed, and it can also support different manufacture volume for each die in 3D to reduce the overall cost. For example, the proposed ISL (either as 2.5D interposer or as a separate layer in true 3D stacking) can be manufactured with much larger volume than the IP blocks, then it can be integrated with various IP blocks on various design, such as with different number of CPU cores and various analog/mixed signal IP blocks.

The service layer is an additional die that provides a network-on-chip (NoC) with interconnect and routers to connect the TSVs between the integrated die above and below it, allowing for non-adjacent units within and across dies to communicate efficiently. An example ISL system is shown in 8a. Because of the relative uniformity and sparseness of the service layer, it can be produced at high yield and volume to minimize the manufacturing cost. A catalog of several ISL designs of varying sizes could be used to address different product markets, but to minimize the number of mask sets each ISL design can allow for physical slicing of the full die to match the area of a given product, reaching a larger product volume and amortizing design and mask costs.

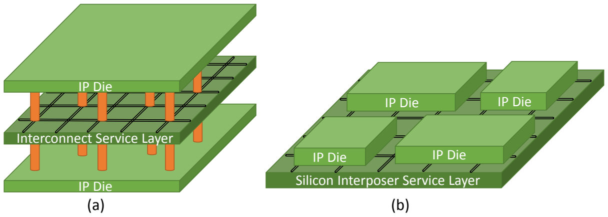

Figure 8: (a) Interconnect Service Layer (ISL) for TSV-based 3D (b) Silicon Interconnect Service Layer (SISL) for active interposer 2.5D

Similarly, interposers can be used as a platform to provide connection between different die-based IPs. For example, Figure 9 shows the 2.5D interposer design with various IP cores sitting on top of an interposer. The ISL interconnect network will be implemented on the interposer layer to provide flexible and reusable interconnect architecture for IP cores which are sitting on top of the interposer.

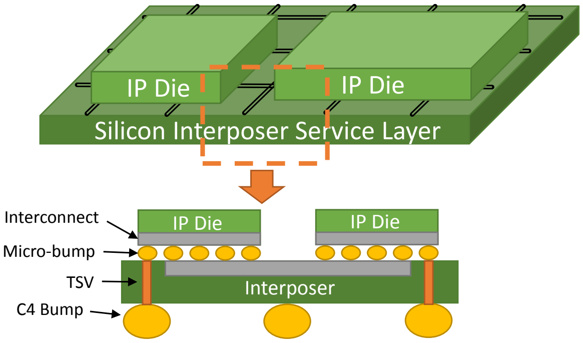

Figure 9: Physical connections of 2.5D interposer with ISL on-chip network

Note that since we propose to use an on-chip network with router design for the ISL, there will be active devices such as logic gates on the ISL. For an interposer-based design, depending on the interposer type, one can implement the on-chip network router on the active interposer, or one can put only the metal routing on the passive interposer but put the active devices for routers in the IP cores (thus reducing flexibility): (1) With current passive interposer technologies, which only have passive interconnect in order to minimize process cost and maximize yield, flexible reconfiguration will require for each IP die to dedicate area for active router hardware. (2) Alternatively, active interposers could be used to provide a disaggregated interconnect network with active routing hardware.

Such Silicon Interposer Service Layer (SISL) provides a base platform with a regular network structure and dedicated routers that can be employed to connect any IP dies that are integrated onto the interposer. Dies are bonded

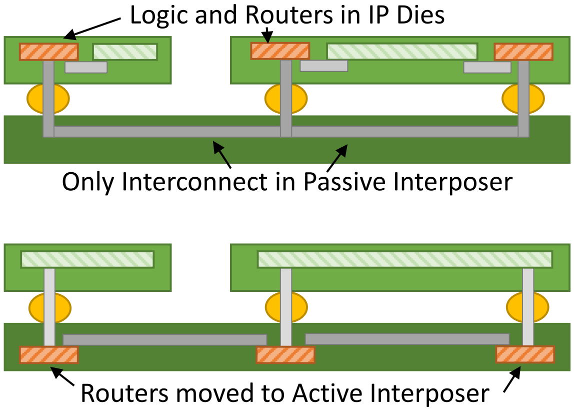
Logic and Routers in IP Dies
Only Interconnect in Passive Interposer
Routers moved to Active Interposer

Figure 10: SISL implementations with passive and active interposer technologies

to the interposer with micro-bumps at the top metal layer, so physical redesign of existing IP onto the SISL platform may be small and could be achieved by altering only the top metal masks. Active interposers are essentially large traditional CMOS die, and as such do not benefit from the reduced wafer cost of passive interposer processes, but the sparse nature of the active die, with only a percentage of active area, means that yields may be very high despite the large area [7]. An example Silicon Interposer Service Layer topology is shown in Figure 8b. For both the 3D ISL and 2.5D SISL platforms, the service layer may contain multiple superimposed heterogeneous networks to provide better communication flexibility and to allow for unnecessary networks to be gated for power. An SISL example is shown in Figure 11, where each node represents a logical connection, realized physically by multiple micro-bumps, between the interposer network and the bonded IP die.

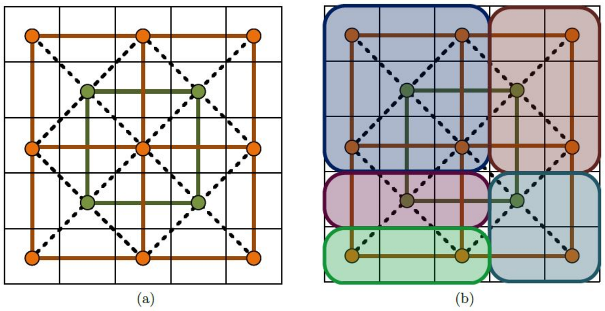

Figure 11: Example active network for SISL implementation: (a) 2x2 and 3x3 superimposed mesh topologies, (b) example overlay of IP blocks

6. CONCLUSION

As modern process complexity increases, the effective cost per transistor reduction is slowing and the cost per constant area is increasing, especially when considering the growing overhead of non-recurring costs like masks, design, and verification. While traditional integration may be reaching cost limits, this paper demonstrates that die-level integration offers the opportunity to reduce costs through the effective partitioning of designs into multiple die in both TSV-based 3D and interposer-based 2.5D integrated circuits. Additionally, die-level integration offers new possibilities for heterogeneous process integration and the reuse of intellectual property, allowing for additional cost benefits and the reduction in non-recurring costs through larger effective volume and reduced redesign effort by taking advantage of flexible interconnect platforms.

7. REFERENCES

[1] P. Christie and D. Stroobandt. The interpretation and application of rent's rule. *IEEE Transactions on VLSI Systems*, 8(6):639–648, Dec 2000.

[2] A. Dingwall. High-yield-processed bipolar LSI arrays. In *IEDM*, volume 14, pages 82–82, 1968.

[3] W. Donath. Placement and average interconnection lengths of computer logic. *IEEE Transactions on Circuits and Systems*, 26(4):272–277, Apr 1979.

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

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

[6] A. Kahng, S. Mantik, and D. Stroobandt. Toward accurate models of achievable routing. *TCAD*, 20(5):648–659, May 2001.

[7] A. Kannan, N. E. Jerger, and G. H. Loh. Exploiting interposer technologies to disintegrate and reintegrate multicore processors. *IEEE Micro*, 36(3):84–93, May 2016.

[8] B. Landman and R. L. Russo. On a pin versus block relationship for partitions of logic graphs. *IEEE Transactions on Computers*, C-20(12):1469–1479, Dec 1971.

[9] M. Lapedus. *10nm Versus 7nm*. Semiconductor Engineering, April 2016. http://semiengineering.com/10nm-versus-7nm/.

[10] G. E. Moore. Cramming more components onto integrated circuits. *Proceedings of the IEEE*, 86(1):82–85, Jan 1998.

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

[12] 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*, pages 3.1.1–3.1.4, Dec 2014.

[13] X. Wu, G. Sun, X. Dong, R. Das, Y. Xie, C. Das, and J. Li. Cost-driven 3d integration with interconnect layers. In *Design Automation Conference (DAC), 2010 47th ACM/IEEE*, pages 150–155, June 2010.

[14] G. Yeric. Moore's law at 50: Are we planning for retirement? In *IEDM*, pages 1.1.1–1.1.8, Dec 2015.

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