Early Selection of System Implementation Choice among SoC, SoP and 3-D Integration

Roshan Weerasekera, Li-Rong Zheng
ECS/ICT/KTH,
ELECTRUM 229,
164 40 Kista, Sweden.
Email: {roshan,lirong}@imit.kth.se

Dinesh Pamunuwa
Centre for Microsystems Engineering
Lancaster University
Lancaster LA1 4YR, UK.
Email: d.pamunuwa@lancaster.ac.uk

Hannu Tenhunen
ECS/ICT/KTH,
ELECTRUM 229,
164 40 Kista, Sweden.
Email: hannu@imit.kth.se

***Abstract*— Recently there is a tendency for shifting the planar SoC single-chip solutions to different alternative options as tiled silicon and single-level embedded modules as well as 3-D integration, and the designers confronted with several system design options. To get a true improvement in performance, a very careful analysis using detailed models at different hierarchical levels is crucial. In this work, we present a cohesive analysis of the technological, cost and performance trade-offs for implementing digital and mixed-mode systems considering the choices between 2-D and 3-D integration and their ramifications.**

I. Introduction

As consumer demand for products that keep getting smaller, lighter and offer more functionality and performance for less power continues unabated, experimental electronic system implementation technologies are migrating towards 3-D solutions [1]. However, even as designers are presented with an extra spatial dimension, the complexity of the layout and the architectural trade-offs also increase. To get a true improvement in performance, a very careful analysis using detailed models at different hierarchical levels is crucial. Even though several previous works have addressed this issue [2][3][4], they mostly concentrate on isolated model development, or target some specific type of system. In this work, we collate existing models from the literature, and modify them and also derive new models as necessary. The main contribution of this paper is in developing a generic methodology for performance and cost estimations of 3D systems that can be modified for different applications, and a comprehensive set of estimation models as building blocks. We also use this methodology to provide detailed estimates for two applications that showcase the potential benefits of 3D integration.

Previous works that addressed cost and performance trade-offs include [2] and [3], where Liu et. al. discuss the mapping from 2-D to 3-D under the constraints of performance, cost and temperature. However, they omit many 3-D technological details. The authors of [4] describe a yield and cost model for 3-D stacked chips with particular emphasis on how the yield is affected by the number of through-hold vias.

3-D integration techniques can be basically categorized into two major schemes: Folding and Stacking. In folding, a planar assembly with flexible substrate is folded into several layers in order to form a very compact shape. In this approach the interconnect length is longer than in the stacked approach described below, but a very compact size can be achieved. Stacking can be done at the chip level with either chip-to-chip (C2C), Package-on-Package (PoP) or MCM-to-MCM bonding using epoxy or glues and creating electrical connections by wire-bonding techniques. As an alternative to chip stacking, 3-D integration can be performed at the wafer-level too. Different blocks can be processed on separate wafers, and they can be interconnected vertically using through-hole vias (THV) or through-Si vias (TSV) to form global communication links. Wafer-Level integration (WLI) can be performed in two ways; entire wafers can be bonded together before dicing (an approach herein after termed 3D-W2W) or KGDs are bonded on top of a host wafer containing other KGD sites termed (3D-D2W) [5].

In this analysis, we concentrate on stacking methodologies and compare between 3D-SiP, 3D-D2W and 3D-W2W technologies.

The rest of the paper is organized as follows; first, we present our methodology for cost and performance estimation, including all models; and then in Section III, we discuss the cost and performance issues for two different applications in detail. We end with a discussion and our conclusions.

II. Cost and Performance Estimation Models

A. Yield and Cost Analysis

The yield of a bare silicon die, Y_{die}, depends on electrical defects on each mask layer in the fabrication process and the total area of the chip. As given by [6], a yield function for a bare silicon die is:

$$Y_{die} = \frac{1}{(1 + SD_0A)^{\frac{N}{S}}} \quad (1)$$

where D_0 is the average electrical defect density, S is the shape factor of (what is assumed to be) the Gamma

978-1-4244-1593-9/07/$25.00 ©2007 IEEE

distribution of electrical defect density, N is the number of mask layers, and A is the chip area. If not provided by the IP vendor the area of a digital module implemented in some target technology can be estimated in a straightforward manner, using gate information and technology scaling. However, the area of an analog chip depends not only on the number of transistors and their sizes (in practice, minimum size transistors are not used in analog circuits), but also the circuit architecture.

The core area (A_{core}) occupying the transistors and their interconnects can either be interconnect-capacity limited or transistor-area limited.

$$A_{core} = \max\left\{N_g d_g^2, N_g A_g\right\}, \quad (2)$$

where N_g is the number of total number of gates, A_g is the gate area (A_g), and d_g is the gate dimension. The gate dimension is defined as $d_g = \frac{f_g R_m P_w}{e_w n_w}$, where R_m is the average interconnect length, which can be determined from Donath's model [7]. When it comes to packaging the core, the number of I/Os to be connected to the outside must be arranged around the periphery and may require a larger perimeter than dictated by the core area in order to facilitate their placement according to the minimum peripheral pitch. Then, the die area is:

$$A_{die} = \max\left\{(\sqrt{A_{core}} + 2P_p)^2, \left(\frac{N_p P_p}{4} + 2P_p\right)^2\right\} \quad (3)$$

where P_p is the peripheral in-line pad pitch and N_p is the total number of IO pads.

When N_{die} is the number of dies that can be processed from a wafer, the bare-die cost is estimated as follows:

$$C_1 = \frac{C_{wafer}(raw, process, mask)}{N_{die} Y_{die} A_{die}} \quad (4)$$

The yield after each testing process depends on the fault coverage level (F_c) of the testing process, and is $Y_d^{(1-Fc)}$. The cumulative cost per die at the end of each process step as follows [8]:

$$C_{1,i} = \frac{C_{1,i-1} + C_i}{Y_d^{F_c}} \quad (5)$$

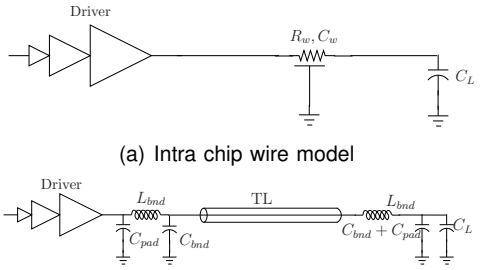
where $C_{1,i-1}$ is the accumulated cost of all the steps up to but not including the present step and C_i, is the cost of the present step.

The package type is assumed to be a peripheral I/O single chip plastic package and its cost is calculated using a price vs pin count assumption as in [9].

B. Interconnect Performance Models

1) On-Chip Wire Delay: Typically the delay over a global on-chip wire is RC dominated. The delay with a capacitive load,C_L, connected at the far-end constitutes the driver delay and the distributed wire delay:

$$t_{rc} = 0.693\left\{R_d(C_d + c_w L + C_L) + r_w L C_L\right\} + 0.377 r_w c_w L^2, \quad (6)$$



(a) Intra chip wire model

(b) Inter chip wire model

Fig. 1. Delay models for Intra and Inter Chip Interconnections

Parameter		On-Chip	Off-Chip
Physical	$W(nm)$	290	15
	$T(nm)$	319	5
	$H(nm)$	290	25
	$S(nm)$	145	50
	k_{ILD}	2.5	3.5
Electrical	$r_w(\Omega/mm)$	237	0.02
	$c_w(fF/mm)$	137	83
	$l_w(nH/mm)$	0.13	0.41
	$Z_0(\Omega)$	31	70

TABLE I
ON-CHIP AND OFF-CHIP WIRE PARAMETERS

where C_d is the driver drain-diffusion capacitance. Therefore the propagation delay on the on-chip wire, as shown in Figure 1(a), is the sum of cascaded buffer delay (t_{drv}) and the Elmore delay of the RC wire: $t_{intra} = t_{drv} + t_{rc}$.

2) Off-Chip Wire Delay: For the inter-chip communication link shown in Figure 1(b) the following delay expression can be derived[10]:

$$t_{RLC} = (t_{tof}^{1.6} + t_{rc}^{1.6})^{\frac{1}{1.6}} \quad (7)$$

where,

$$t_{LC} = t_{tof} = L\sqrt{l_w c_w} \quad (8)$$

and

$$t_{rc} = 0.693\left[Z_0(C_d + C_{pad} + C_{bnd} + 0.5C_L) + \frac{L_{bnd}}{Z_0} + r_w L(C_{pad} + C_{bnd} + C_L)\right] + 0.4 r_w c_w L^2, \quad (9)$$

where C_{pad} is capacitance of the pad, and C_{bnd} and L_{bnd} are the capacitance and the inductance of the bond wire.

Finally, the total delay for the inter-chip communication link is the summation of cascaded driver delay (t_{drv}) and the RLC-wire delay (t_{RLC}): $t_{inter} = t_{drv} + t_{RLC}$.

III. TRADEOFF ANALYSIS FOR THE CASE STUDIES

To make the comparison, we begin by selecting two mixed-signal systems. first system is a *Wireless Sensor*, which contain a 2Mb DRAM, and an ASIC and Microprocessor with gate count of 500k and 300k respectively. It also contains an Analog/RF block occupying an area of $2\ mm^2$. Finally , it contains a MEMS sensor with an area of $1mm^2$. The second system is a *3G mobile terminal*. We consider a similar architecture as the first

978-1-4244-1593-9/07/$25.00 ©2007 IEEE

one but with a larger memory of $128\ Mb$ DRAM, and a CMOS image sensor with a pixel size of $1.75\ \mu m \times 1.75\ \mu m$, and resolution of 8 $Megapixel$ [11]. Further, in the analysis, we consider the ASIC and Microprocessor together as a single logic block. For all the integration schemes, the underlying manufacturing process is a 65 nm, 11-metal, CMOS process with a wafer diameter of 300 mm and a lower-level wire pitch of 136 nm. We also assume peripheral in-line pad arrangement and wire bond packaging. The worst-case delay for 2-D systems is estimated diagonally from chip edge to chip edge, while it is estimated from one edge of the bottom chip to the opposite side edge of the top most chip for 3-D systems.

Based on the manufacturers data, the power density for the constituent sub-modules in our case studies can be estimated. The power density for a DRAM is estimated to be $0.02W/mm^2$ [12], and for a logic block, $0.12W/mm^2$[13]. A CMOS Image sensor has an average power density of $0.016W/mm^2$. The power dissipation of the MEMS sensor is assumed to be $50mW$, and that for the Analog/RF block, $500mW$.

For the stacked arrangement, we assume that the logic block is close to the heat sink and other blocks are in the following order: DRAM, Analog/RF block, and MEMS/CMOS Image sensor.

A. Monolithic SoC

The integration mixed signal systems in a single die is a merging of several technologies, such as logic, memory, analog/RF, and this results in increased process complexity and a area change. For example merging logic circuits with memory results in a lower circuit density and hence a larger circuit area, than their logic-only or memory-only counter parts. The area of a single chip implementation is estimated as stipulated in [2]. The total cost for an SoC implementation is given in (38).

B. 2D-SoP

In the 2D-SoP implementation, we assume that four chips (DRAM, RF, Logic and MEMS/Image Sensor) are assembled as a multi chip module (MCM). Hence, the cost of implementing the MCM includes the total cost for each chip including the testing, the assembly cost, the substrate cost, the rework cost, and finally the MCM test cost and packaging cost. The SoP can provide some reworking capability whereas SoC and wafer-level 3-D integration do not. If one rework cycle is assumed for SoP, the yield in assembly is improved from Y_a to $(2 - Y_a)Y_a$. Then the cost for SoP is given by (39) and the overall yield is:

$$Y_{SoP} = Y_{rf}^{(1-Fc)} Y_{dram}^{(1-Fc)} Y_{logic}^{(1-Fc)} Y_{other}^{(1-Fc)} Y_a \qquad (42)$$

C. 3D-SiP

A 3D-SiP implementation is similar to the SoP package integration, except that the SiP implementation integrate dies on top of each other vertically. The cost formula is the same, but the MCM substrate area is reduced, compared to the 2D-SoP implementation.

D. 3D-WLP

The yield of each 3D-implementation method is the cummulative yield over all the layers (m) and given by [14]:

$$Y_{3D} = Y_{2D} \prod_{i=1}^{m-1} Y_{2D_i} Y_a \qquad (43)$$

where Y_{2D} is the yield of 2D process (fabrication yield), and Y_a is the yield loss due to the 3D-assembling process. In the case of D2W stacking, die yield after the KGD testing should be considered. So, the overall yield for implementing our target system in 3D-W2W and 3D-D2W methods are as follows [14]:

$$Y_{3D\_w2w} = Y_{dram} Y_{logic} Y_{rf} Y_{other} Y_a^3 \qquad (44)$$

$$Y_{3D\_D2w} = Y_{dram}^{(1-Fc)} Y_{logic}^{(1-Fc)} Y_{rf}^{(1-Fc)} Y_{other}^{(1-Fc)} Y_a^3 \qquad (45)$$

The total cost for 3-D Wafer-Level integration are given in equations (40) and (41). Due to the limitations in the wafer level processing, there is no possibility of reworking.

IV. Discussion

Results for our case studies are shown in Table II. It is quite obvious that 3-D integration provides very compact designs compared to its 2-D planar counterpart. Except for the 3D-SiP method, 3D-WLI has lower interconnect delays over the 2-D implementations. 3D-SiP and 2D-SoP implementations are more or less equal in implementation cost, but 3D-SiP has a lower interconnect delay. Where the wireless sensor node is concerned, the SoC solution is the better choice, while wafer-level 3D integration provides lower area and higher performance. A SoC solution is the best option for such low memory applications because it is less expensive. However, though quite expensive, for high performance systems, 3D-WLI is the best choice.

The scenario is different when it comes to a mobile terminal. In this case, the overall chip area is 4.25 times larger than that of the wireless sensor node. 3D-WLI technologies outperform SiP implementation technologies, due to the very long RC wires. Also, single chip solution has a very low yield. All the other implementations methods show a lower cost than the single chip implementation. 3D-SiP seems to be the best design choice for low-cost, and high performance in this case.

The case studies show that when the system size becomes very small the thermal resistance becomes high the temperature rise in the top-most chip is unbearable. Hence, extra cooling solutions such as thermal-vias occupying some area, or very thin layers have to be used in the system implementation.

978-1-4244-1593-9/07/$25.00 ©2007 IEEE

$$C_{SoC} = \left[\left(\frac{C_{wafer}}{Y_{SoC}N_{die}} + C_{wafer\_test}\right)\frac{1}{PF_w} + C_{burn\_in}\right]\frac{1}{PF_b} + C_{pkg} \tag{38}$$

$$C_{SoP} = \left\{\frac{\sum_{i=1}^{m} C_{kgd_i} + \frac{C_{substrate}}{Y_s} + C_{assembly} + C_{rework}}{Y_a} + C_{test}\right\}\frac{1}{PF_{SoP}} + C_{pkg} \tag{39}$$

$$C_{3D\_W2W} = \left\{\frac{\sum_{i=1}^{m} C_{die_i} + C_{bonding}}{Y_{a\_3D\_W2W}} + C_{test}\right\}\frac{1}{PF_{3D\_W2W}} + C_{pkg} \tag{40}$$

$$C_{3D\_D2W} = \left\{\frac{\sum_{i=1}^{m} C_{kgd_i} + C_{bonding}}{Y_{a\_3D\_D2W}} + C_{test}\right\}\frac{1}{PF_{D2W}} + C_{pkg} \tag{41}$$

Case	Wireless Sensor Node					3G Mobile Terminal				
Parameter	Single Chip	2D-SoP	3D-SiP	3D-W2W	3D-D2W	Single Chip	2D-SoP	3D-SiP	3D-W2W	3D-D2W
Normalized Area	1.00	3.92	0.78	0.71	0.71	1.00	1.94	0.75	0.71	0.71
$\text{Yield}_{overall}$	0.95	0.98	0.98	0.92	0.94	0.56	0.98	0.98	0.71	0.94
Normalized Cost	1.00	4.11	4.04	1.14	2.96	1.00	0.40	0.40	0.38	0.33
Delay (ps)	127.37	176.36	148.33	83.9	83.9	317.88	205.37	168.34	259.63	259.63
$\Delta T(^oC)$	39.16	12.39	52.8	312.74	312.74	26.38	14.67	36.9	73.96	73.96

TABLE II

RESULTS OF COST PERFORMANCE ANALYSIS FOR CASE-STUDIES. NOTE THAT $\Delta T = T_{top\_layer} - T_{ambient}$.

V. CONCLUSION

In this paper, we developed a detailed yield and quantitative cost models, and a quantitative performance metric for 3-D integration. Further, we derived simple yet useful thermal models for 2-D and 3-D integrated circuits. The overall methodology is suitable for early analysis in system explorations for future nanoscale electronic systems. Through some example contemporary mixed signal systems we demonstrate the methodology outlined for different implementations and conclude that the implementation strategy must be carefully selected depending on the circuit complexity, as else the move to 3-D may have a detrimental effect. Design choice early in the design cycle will have a significant impact throughout the design and production lifecycles , and the models and methodology presented in this article can be an important aid in this choice.

REFERENCES

[1] The International Technology Roadmap for Semiconductors(ITRS), 2005. [Online]. Available: http://www.itrs.net

[2] M. Shen, L.-R. Zheng, and H. Tenhunen, "Cost and performance analysis for mixed-signal system implementation: System-on-chip or system-on-package," *Electronics Packaging Manufacturing, IEEE Journal of*, vol. 25, no. 4, pp. 262–272, October 2002.

[3] C. Liu, J.-H. Chen, R. Manohar, and S. Tiwari, "Mapping system-on-chip designs from 2-d to 3-d ics," in *Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on*, 2005, pp. 2939–2942 Vol. 3.

[4] P. Mercier, S. Singh, K. Iniewski, B. Moore, and P. O'Shea, "Yield and cost modeling for 3d chip stack technologies," in *Conference 2006, IEEE Custom Integrated Circuits*, September 2006, pp. 357–360.

[5] T. Fukushima, Y. Yamada, H. Kikuchi, and M. Koyanagi, "New three-dimensional integration technology using chip-to-wafer bonding to achieve ultimate super-chip integration," *Japanese Journal of Applied Physics*, vol. 45, no. 4B, pp. 3030–3035, 2006.

[6] A. George, J. Krusius, and R. Granitz, "Packaging alternatives to large silicon chips: tiled silicon on mcm and pwb substrates," *Components, Packaging, and Manufacturing Technology, Part B: Advanced Packaging, IEEE Transactions on [see also Components, Hybrids, and Manufacturing Technology, IEEE Transactions on]*, vol. 19, no. 4, pp. 699–708, 1996.

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

[8] P. A. Sandborn and H. Moreno, *Conceptual Design of Multichip Modules and Systems*. Kluwer Academic Publishers, 1994.

[9] D. Ragan, P. Sandborn, and P. Stoaks, "A detailed cost model for concurrent use with hardware/software co-design," in *Design Automation Conference, 2002. Proceedings of IEEE/ACM*, 2002, pp. 269–274.

[10] G. Sai-Halasz, "Performance trends in high-end processors," *Proceedings of the IEEE*, vol. 83, no. 1, pp. 20–36, 1995.

[11] Micron CMOS Image Sensor Part Catalog, March 2007. [Online]. Available: http://www.micron.com

[12] Micron 128MB SDRAM Part Catalog, 2007. [Online]. Available: http://www.micron.com

[13] ARM Cortex-A8 Processor Product Brief, March 2007. [Online]. Available: http://www.arm.com

[14] Y. Deng and W. P. Maly, "2.5-dimensional vlsi system integration," *very large scale integration (VLSI) systems , IEEE Transactions of*, vol. 13, no. 6, pp. 668–677, June 2005.

978-1-4244-1593-9/07/$25.00 ©2007 IEEE