2.5-Dimensional VLSI System Integration

Yangdong Deng, *Student Member, IEEE,* and Wojciech P. Maly, *Fellow, IEEE*

***Abstract*—The excessive interconnection delay and fast increasing development cost, as well as complexity of the single-chip integration of different technologies, are likely to become the major stumbling blocks for the success of monolithic system-on-chips. To address the above problems, this paper investigates a new VLSI integration paradigm, the so-called 2.5-dimensional (2.5-D) integration scheme. Using this scheme, a VLSI system is implemented as a three-dimensional stacking of monolithic chips. A cost analysis framework was developed to justify the 2.5-D integration scheme from an economic point of view. Enabling technologies for the new integration scheme are also reviewed.**

***Index Terms*—Cost, monolithic integrated circuits, VLSI, 2.5-D integration.**

I. Background

FUNCTIONALITY increase has been and will continue to be the major driving force for the semiconductor industry. As a matter of fact, the spectacular success of IC industry in the past 30 years depends on the ability to continuously shrink the feature size of IC fabrication process and at the same time pack more devices on a single silicon die. However, when the main-stream fabrication technology is now moving to the 90-nm node, the feasibility of the monolithic integration paradigm is severely stumbled by the following factors.

Interconnection "Crisis": Historically, functionality to be integrated in a single chip at every technology generation always exceeds the capacity provided by pure scaling. To accommodate the extra transistors, chip size has always been increasing since the invention of the first integrated circuit [1]. The problem is that, interconnection length, especially worst-case interconnection length, has to increase as long as IC chip size is expanding. Starting from the 0.25 μm generation, the interconnection delay of long on-chip wires has become the dominant part determining system performance [2]–[4]. As a result, the timing of global signals has become a critical concern. Unfortunately, interconnection delay is very hard to predict before the circuit is actually laid out. As a result, current synthesis-based VLSI design methodology often has difficulty to achieve timing closure.

Fabrication Cost: The extreme complexity of today's semiconductor process leads to a skyrocketing of the fabrication facility cost [5]. It has been reported that the cost of a single mask set and corresponding probe will soon reach $1 million [6], [7]. Meanwhile, modern system-on-chips (SoCs), especially those for wireless applications, typically integrates heterogeneous components. These components are originally targeted for different fabrication technologies. This further complicates the merged process and raises fabrication cost. For example, in a RF-CMOS process, the price of a finished wafer is higher than that of pure CMOS by at least 15% [8]. Meanwhile, for RF circuits, it is difficult to achieve further performance improvement and cost reduction by using a scaled technology. For instance, many analog transistors and passive components have to occupy a relatively constant die area to meet performance requirements no matter in which technology generation they are fabricated [8].

Memory Gap: Memory bandwidth has already become the limiting factor impeding the performance of general-purpose microprocessors and multimedia appliances, as well as other data-intensive applications. It has been reported that the processor performance has been improving by 35% annually from 1980 to 1986 and by 55% annually thereafter [9], [10]. In the same period, the access latency of DRAM has been improving by only 7% per year [9]. Now this problem is mainly addressed by introducing cache hierarchy and integrating memories with the logic on the same chip. For most current processors, at least 50% of the die area is occupied by cache memories [11]. Also, a PDA-type phone could use as much as 128-Mb flash and 128-Mb DRAM [12]. Embedded memory requires a merged memory/logic process, which is more expensive and leads to inferior memory devices [13], [14]. Moreover, the long interconnects of the memory buses can also become a bottleneck when a large amount of memory is integrated.

It must be indicated that the above problems are inherent to the monolithic integration. Therefore, the key question must be raised: How to build modern systems that avoid the shortcomings of monolithic SoC, while maintaining momentum in the increase of the functionality?

II. Nonmonolithic Integration Schemes

There have been many nonmonolithic integration solutions proposed in the past to address the problems inherent to the monolithic integration scheme. As a result, for a large IC system that is difficult and/or costly to fabricate with a single set of reticles, the following integration schemes can be considered: Multiple-Reticle Wafer derived from Wafer Scale Integration [15], MultiChip Module (MCM) [16], and three-dimensional (3-D) integration [17]. In this section, we first review these nonmonolithic VLSI integration schemes and then propose a new integration scheme by extending the "smart substrate MCM" [18] concept with a 3-D stacking approach.

A. Multiple-Reticle Wafer

This approach is a revised version of wafer scale integration [15]. Under such a context, a system is partitioned into multiple dies and these dies are built on the same wafer. Since faulty

Manuscript received April 7, 2004; revised December 20, 2004. This work was supported in part by the Giga-Scale Research Center.

The authors are with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: yangdon@ece.cmu.edu; maly@ece.cmu.edu).

Digital Object Identifier 10.1109/TVLSI.2005.848814

TABLE I
WAFER BONDING BASED 3-D INTEGRATION TECHNOLOGIES

Research Group	Stacking Style	Bonding Interface[1]	Inter-chip Contact		
			Alignment Accuracy (μm)	Footprint (μm^2)	Height (μm)
IBM [34~36]	Face-to-back	Adhesive	N/A	Variable	25~60
Tohoku Univ. [29, 30]	Face-to-face or face-to-back	Adhesive	± 1	3 × 3	~30
NEU [31]	Face-to-face	Adhesive	~ ± 3	~ 10	<5
MIT [32]	Face-to-back	Cu-Cu interface	± 3	3~5 × 3~5	<10
RPI/UAlbany [33]	Face-to-face or face-to-back	Low-k dielectric glue	± 1 ~ 2	3 × 3	~30

[1] By bonding interface we indicate the manner in which two chips are attached.

die(s) may be produced during the manufacturing process, redundant dies are introduced to guarantee correct functioning of the system. However, the major reason hindering this approach is the prohibitive fabrication cost. As a matter of fact, a general VLSI application without a regular system architecture will require multiple sets of masks, which can be extremely expensive unless the cost can be amortized by a really large volume. From the performance point of view, this multiple-die scheme could hardly outperform its monolithic equivalent, especially because global interconnects have to be longer due to the introduction of redundant dies and extra bypass logic.

B. Multiple Chip Module (MCM)

MCM system assembly [16] has been extensively considered as an alternative packaging solution for VLSI systems. Under such a context, bare dies are mounted on a common substrate, which can be simply a miniaturized PCB (MCM-L), a piece of glass ceramic (MCM-C), or alternately deposited layers of high-density thin-film metal and low dielectric materials (MCM-D). Commonly used techniques to bond the chips and the substrate include wire bonding, Tape Automated Bonding (TAB), and flip-chip or Controlled-Collapse Chip Connections (C4). The problem here is that wafer probing testing is very difficult for at-speed tests and thus bare dies usually can only be partially tested. Due to the imperfect testing, faulty dies can be introduced into a MCM system and lead to poor yield of the whole module. One way to address the problem is through the so-called "known good dies" that have a very high probability to correctly function. However, the techniques to guarantee the high probability tend to be very expensive.

C. 3-D Integration

Research and industrial efforts along this direction can be dated back to as early as 1980s [19], [20] and has been followed by many recent developments [21]. Fabrication technologies for the 3-D integration can be classified into two categories: silicon growth and wafer bonding.

In the silicon growth approach, a new device layer of silicon is formed on top of an existing substrate by one of the following methods: 1) polysilicon deposition (which can be used to build thin-film transistor) [22]; 2) epitaxy growth [23]; and 3) amorphous silicon deposition and crystallization [20], [24], [25]. The inter-chip contacts can be constructed by extending via-formation techniques, e.g., etching through the new layer of silicon.

An important concern of this approach is that the formation of each new device layer of silicon must be compatible with metal interconnects underneath. For Cu wires, the processing temperature must be under 450 °C so that the Cu diffusion effect doesn't occur. Repeated exposure to high temperature also tends to impair the quality of transistors in the lower layer wafers.

Under the context of wafer bonding (e.g., [26]–[41]), wafers can be built with traditional processes. An upper layer wafer is first grinded from back to a thickness of around 10 micrometers and then bonded on the top of the lower layer wafer. To construct the inter-chip interconnection, one solution is to etch through-wafer vias all the way from the top level wafer to the uppermost metal layer on the lower layer of wafer. An alternative solution involves three steps: initially building the inter-chip interconnects as embedded contacts, then exposing them by grind the wafer from the back to a proper thickness, and finally bonding them with the bump built on the top of the lower layer wafer. Alignment accuracy determines the lower limit of the size of inter-chip interconnects. In the recently reported processes listed in Table I, alignment accuracy is within the range of ± 3 μm. Thus, the footprint of inter-chip contact could be as small as ~ 10 μm^2.

Since it allows wafers built with different processes to be assembled, the wafer bonding technology is more flexible than the silicon growth approach. In addition, the stacking technology does not involve high-temperature processing. On the other hand, the wafer thinning step still could lower the fabrication yield or even lead to the failure of wafer manufacture.

However, the key hindering factor for the 3-D integration scheme is the accumulative yield loss. If one die on a certain layer of wafer is faulty, then the 3-D VLSI system containing this die has to be defective after the assembling process (either silicon growth or wafer bonding). In other words, since it is not possible to selectively replace a faulty die, the accumulative yield lose is inevitable as long as the assembling process performed at the wafer level.

D. 2.5-D Integration

In the work reported in this paper, we consider a new approach, 2.5-dimensional (2.5-D) integration, which is a revision of the concept of "smart substrate MCM" proposed in the past [18], [42]. To implement a system (given e.g., in the form of a system-level or RTL description) using the 2.5-D system integration scheme, one partitions the system into a number of

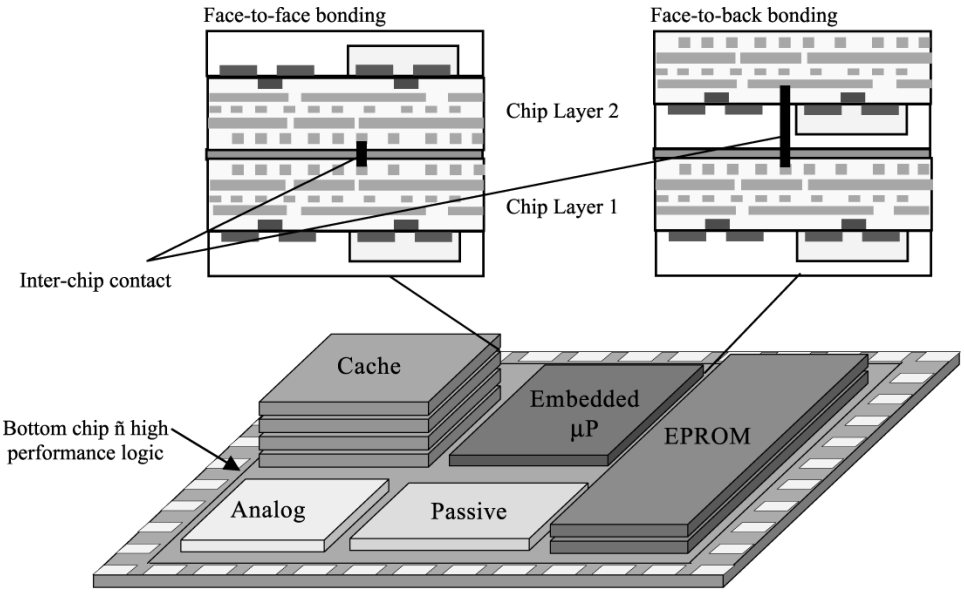

Fig. 1. Potential 2.5-D system.

clusters, each containing components that are going to be fabricated in a specific technology. Logic synthesis and layout design for every cluster of components are performed such that each cluster can be fabricated as an unpackaged die, optimized for performance and/or cost. Finally these chips can be stacked together in the manner, for instance, illustrated in Fig. 1. In this particular implementation the inter-die communication and power distribution might be accomplished through "vertical" interconnects between stacked dies. We refer to the vertical interconnect as "inter-chip contact."

To overcome the "imperfect testing" problem associated with the MCM approach, the 2.5-D integration scheme is enhanced with an **incremental, hierarchical testing and assembling** methodology proposed in [42]. The bottom level chip can be designed and fabricated with relatively sufficient testing support (e.g., partially packaged), while upper layer chips can be designed with adequate self-testing logic and isolation capability. This way the bottom level chip can be properly tested during the assembling process and then used as a "chassis" to test upper layer chip(s). As a result, a die can be tested as soon as it is "plugged" into a partially assembled 2.5-D system using available hardware/software components. In addition, if the stacking process can be designed to support **rework** to some extent, it is possible to replace a faulty die with a new one without damaging the whole system.

Intuitively, many advantages can be expected through the adoption of the 2.5-D integration scheme, including the following.

- Smaller system footprint. By removing intermediate packaging levels, the 2.5-D integration scheme allows a system to be constructed with a much smaller volume and weight. This advantage has significant implications on the portable appliances, which are extremely sensitive to volume and weight.
- Reduced interconnection length. Generally speaking, the underlying graph (or hyper-graph) corresponding to a VLSI circuit is not necessarily planar and so embedding it into a monolithic surface tends to result in an overhead in the interconnection length. On the other hand, 2.5-D/3-D integration enables designers or CAD tools to find a more efficient packing of circuits according to their inherent topology. This way a systematic reduction in the on-chip wirelength can be expected, which can be translated into performance gain and/or power saving. For instance, significant increased memory bandwidth could be achieved by integrating memory dies on top of CPU dies and exploiting very wide memory buses.
- Decoupling between functionality increase and technology selection. Using the new paradigm, a VLSI system can be properly divided into multiple chips so that the fabrication cost and system performance can be optimized. For instance, to build a wireless application using 2.5-D integration concept, the RF transceiver circuit can be manufactured with a SiGe-BiCMOS process with excellent RF performance, especially optimized for sensitivity and low power consumption. Meanwhile, high-performance digital signal processing circuits can be fabricated with a high-speed CMOS process, while other logic circuit for user-application built with a high-V_T, low-power CMOS process. Finally, high-density, low-power DRAMs fabricated with a dedicated DRAM process can be stacked on the top of the logic chips.
- New reuse opportunity. The 2.5-D integration scheme enables reusing verified components at die level. IP cores can be delivered as pre-fabricated and fully characterized dies with standard interfaces. As a result, systems for different applications can be realized as different combinations of standard IP dies. From a design perspective, integrating these IP-dies requires much less effort than the IP-core integration does. This new paradigm of mask reuse promises that VLSI systems with significantly reduced system cost could be designed and implemented in significantly reduced time.

III. Yield Analysis of Different VLSI Integration Approaches

As discussed in the previous section, there exist multiple schemes to implement a given VLSI system into silicon. In this section, we compare various integration strategies using a unified cost analysis framework.

For a given semiconductor technology, the fabrication cost of a VLSI system can be measured by its total consumed (i.e., fabricated) silicon area, S_A, which is calculated as

$$S_A = \frac{A}{Y} \tag{1}$$

where A is the actual silicon area, or the working silicon area, used by the VLSI silicon implementation and Y is the fabrication yield.

The yield, Y, is given by the formula [43]

$$Y = \prod_{i=1}^{m} Y_i = \prod_{i=1}^{m} \prod_{j=1}^{n_i} Y_{ij} \tag{2}$$

where m is the number of components or dies in the system, n_i is the number of mask layers of device layer i or die i, Y_i is the yield of the ith component, and Y_{ij} is the yield of the ith component of the system due to the defects in the jth layer of the IC layout structure. There exist a number of approaches to compute Y_{ij} [43]. In this paper, we use a simple one [44]

$$Y_{ij} = e^{-A_{ij} \cdot D} \tag{3}$$

where A_{ij} is the area of ith component on the jth layer and D is defect density. It is generally assumed that layout structures on different layers have the same area. In other words, A_{ij} can be replaced as A_i (for $1 \le j \le n$) and we have

$$Y_i = e^{-n \cdot A_i \cdot D}. \tag{4}$$

As a result, (2) can be simplified as

$$Y_i = \prod_{i=1}^{m} e^{-n \cdot A_i \cdot D}. \tag{5}$$

In our computation, we consider implementing a common application using different integration schemes. It is assumed that this application needs a die size of 4 cm^2 when manufactured as a monolithic chip. For all the integration schemes, the underlying manufacturing process is a 0.13-mm, 6-metal CMOS process with a wafer diameter of 300 mm. It is assumed that the wafer has a fabrication cost of $2500 [45]. In this process there are 19 layout layers: n-well, active region, n-select, p-select, thin oxide, polysilicon, oxide, 6 metal layers, and 6 inter-metal isolation layers. It should be noted that, in the nonmonolithic integrations approaches, certain device layers or dies (e.g., memory circuits) may require a smaller number of mask layers. For a general analysis here we assume an identical number of mask layers for all device layers or dies, i.e., $n_i = n$. On the other hand, although our derivation is simplified, a similar cost trend should be observed if different n_i values are set for different device layers or dies as long as the system is partitioned in the same manner. The defect density, D, is assumed to have a value of 0.025 particles/cm^2. The values of major parameters are shown in Table II. In the rest of this section we will discuss the cost implications of different VLSI integration strategies under the framework given by (1) to (5).

TABLE II
Values for the Major Parameters of Our Cost Model

Symbol	Parameter	Value
D	Defect density	0.025 particles/cm^2
A	Silicon area when implemented as a single chip	4 cm^2
A_i	Silicon area of one component (die)	A/m
n_i	# Layers of layout structure	19
S_w	Wafer area	706.86 cm^2
C_w	Wafer cost	$2,500
C_t	Testing cost per second	0.12 $/s
k	Steepness of fault coverage with regard to time	0.116495
$C_{Carrier}$	KGD testing carrier cost	$2.4
$F_{c\text{-}MCM}$	Fault coverage level of MCM	0.999
Y_a	3-D assembling yield	0.95
Y_R	Multi-reticle reconfiguration yield	0.99
O_M	Multi-reticle cost overhead	0.125 per extra die

A. Monolithic SoC

Given the parameter values listed in Table II, the single-chip implementation ($m = 1$) of the system has a yield of only 15.0%, making it very cost-inefficient. According to (1), the total silicon area needed to fabricate a single working chip, S_A, is 26.7 cm^2.

B. Multiple-Reticle Wafer (MRW)

With this approach, a VLSI system is partitioned into m parts and then each part is fabricated as a separated die on the same wafer. For this discussion, the m parts are assumed to have identical die areas for a general analysis. To improve defect tolerance, r identical dies are built for every component. After one wafer is fabricated, faulty dies will be bypassed through reconfiguring the interconnection among different dies. Here we simply assume the whole system will correctly function as long as at least one die of every component is free of faults. As a result, the yield is given by the formula

$$\begin{aligned} Y_{\mathrm{MRW}} &= \prod_{i=1}^{m} [1 - (1 - Y_i)^r] \cdot Y_R \\ &= \prod_{i=1}^{m} \left[1 - \left(1 - e^{-\frac{A}{m} D}\right)^r\right] \cdot Y_R \\ &= \left[1 - \left(1 - e^{-\frac{A}{m} D}\right)^r\right]^m \cdot Y_R \end{aligned} \tag{6}$$

where Y_R represents the yield of the reconfiguration process. It bears mentioning that the fabrication time of a MRW wafer is considerably longer due to the usage of multiple reticles [15]. We assumed that every die introduces a modest cost overhead of 12.5%, which is actually lower than the typical values for most

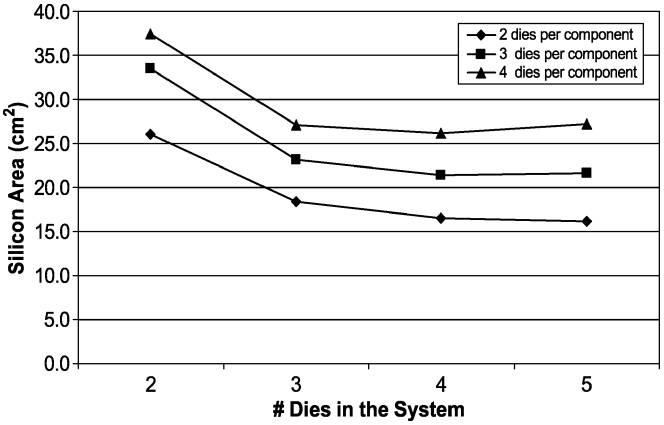

Fig. 2. Total consumed silicon area of multiple-reticle wafer.

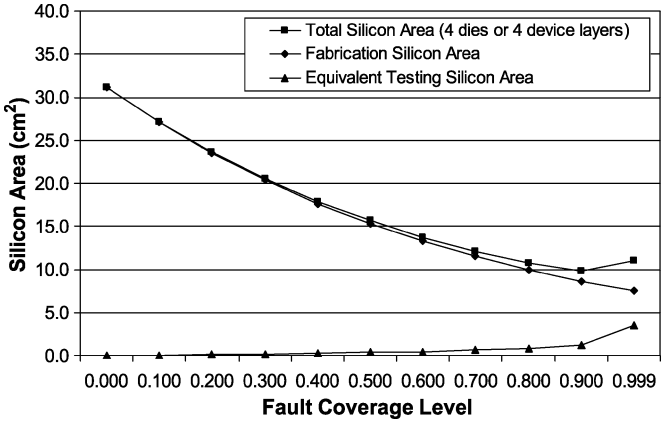

Fig. 3. Silicon area of the 2.5-D implementation with 4 device layers of chips.

real design cases [15]. This way the silicon area of the MRW approach is given by the formula

$$S_{A-\mathrm{MRW}} = \frac{A \cdot (I + O_{\mathrm{M}} \cdot m) \cdot r}{Y_{\mathrm{MRW}}} \tag{7}$$

where O_{M} is the cost overhead.

Using (6) and (7), the total consumed silicon area, S_A, under different values of r are shown in Fig. 2. Compared with the monolithic integration approach, the redundancy can be quite effective to reduce system cost. Meanwhile, it is the most cost efficient to install one extra copy for each die.

C. 2.5-D System Integration

Under the 2.5-D integration context, a VLSI system is partitioned into m parts and then each part is fabricated as a separated die on different wafers. Finally these dies are assembled on a common substrate. Again we assume that every die has the same area as A/m.

As a result, the cumulative yield of one single die, $Y_{i-2.5\mathrm{D}}$, can be computed as the product of three components: 1) Y_i, which is the yield loss due to its own fabrication process; 2) Y_{Others}, the yield loss due to the assembling of other dies; and 3) Y_a, yield loss due to the final 3-D stacking process.

$$Y_{i-2.5\mathrm{D}} = Y_i \cdot Y_{\mathrm{Others}} \cdot Y_a. \tag{8}$$

Y_i can be straightforwardly determined by (5) and Y_a can be assumed to have a constant value of 0.95. The computation of Y_{Others} depends on the fault coverage level of the dies (designated as F_c) in a 2.5-D system [44]

$$Y_{\mathrm{Others}} = \left(Y_i^{1-F_C}\right)^{m-1} \tag{9}$$

where the term $m-1$ in the exponent is to take into account the yield loss due to the assembling of all other components.

On the other hand, the high fault coverage level comes with a price: extra testing time implies extra cost. Generally, a modest fault coverage level, e.g., 80%, can be achieved in relatively short testing time. However, a higher fault coverage level requires significantly increased testing time. Based on this observation, we propose to use an exponential model to correlate the test coverage level with the testing time t

$$F_C = 1 - e^{-\mathrm{k} \cdot t} \tag{10}$$

where k is a constant defining the steepness of exponential function, which can be derived by assuming 60 seconds is long enough to achieve a 99.9% fault coverage and 10% of total time is enough for 80% fault coverage. The testing cost, C_{test}, can be assumed to be linearly proportional to the testing time:

$$C_{\mathrm{Test}} = C_t \cdot t. \tag{11}$$

In this work, we use a reasonable value of \$0.12 for C_t [46]. Meanwhile, since we already know the wafer cost and wafer size, the testing cost can be translated into silicon area by the formula

$$S_{A-\mathrm{Test}} = \frac{C_{\mathrm{Test}}}{C_w} \cdot S_w = \frac{C_t \cdot t}{C_w} \cdot S_w. \tag{12}$$

From the above analysis, it can be seen that the silicon area of each die in a 2.5-D system consists of two components: fabrication silicon area and equivalent testing silicon area.

$$S_{Ai-2.5\mathrm{D}} = S_{Ai0} + S_{A-\mathrm{Test}} = \frac{\frac{A}{m}}{Y_{i-2.5\mathrm{D}}} + S_{A-\mathrm{Test}}. \tag{13}$$

The silicon area of a 2.5-D system is the summation over all its components

$$S_{A-2.5\mathrm{D}} = \sum_{i=I}^{m} S_{Ai-2.5\mathrm{D}} = \frac{A}{Y_{i-2.5\mathrm{D}}} + m \cdot S_{A-\mathrm{Test}}. \tag{14}$$

The drawing in Fig. 3 shows the trend of silicon areas when the input VLSI design is partitioned into four layers and implemented as four separate dies. Clearly, fabrication cost decrease very rapidly with the increasing fault coverage level of the components. However, it seems unnecessary to have fault coverage very close to 100% due to the excessive testing cost. For the application under discussion, a 90% fault coverage level is cost optimal. Fig. 4 demonstrates silicon area values of the 2.5-D integration with different numbers of components/dies. These results lead to the following observations.

1) For dies with very incomplete or no testing at all, it is more cost efficient to implement the system in smaller number of dies.
2) For dies with reasonable (>40%) fault coverage level, 4 or 5 layers of dies, each with a die area of around 1 cm^2, would be the cost optimum choice.

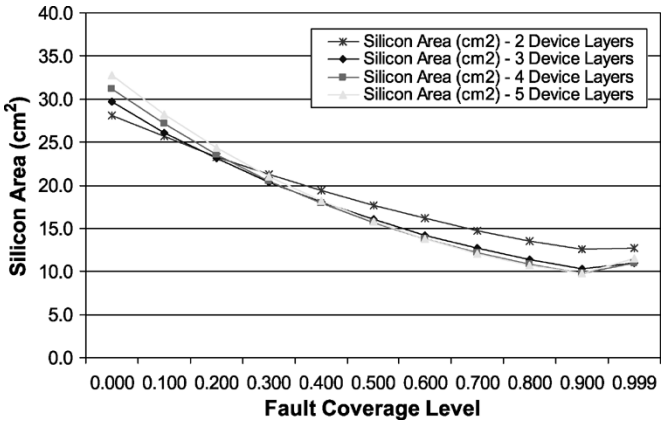

Fig. 4. Silicon area of the 2.5-D implementation.

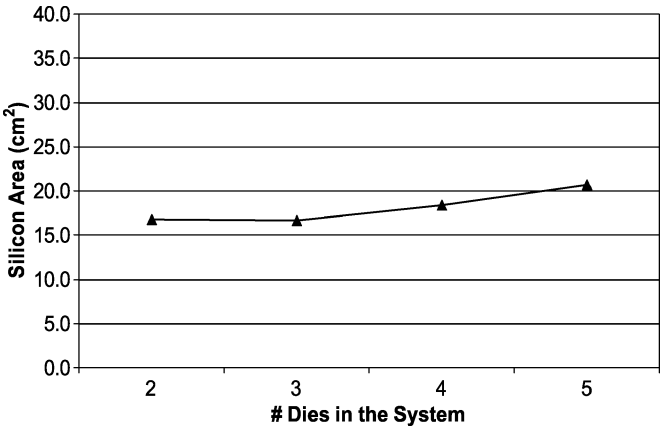

Fig. 5. Silicon area of the MCM implementation.

D. 3-D Integration

Under the 3-D integration paradigm, the input VLSI system is built into m device layers (not necessarily having an equal area). Due to the nature of wafer-level processing, it is not possible to support rework even if the fault coverage testing techniques introduced for the 2.5-D approach could be applied. With no redundancy or rework-ability enhancement structure installed, the yield of the 3-D implementation is the cumulative yield over all layers:

$$Y_{3D} = Y_i \cdot \prod_{i=1}^{m-1} (Y_i \cdot Y_a) = Y_i^m \cdot Y_a^{m-1} \tag{15}$$

where Y_a is the yield loss due to the 3-D assembling process. The factor Y_a^{m-1} is to take into account the fact that integration of m layers of chips requires $m-1$ silicon growth or wafer bonding procedures. This way the silicon area of the 3-D integration is given by

$$S_{A-3D} = \frac{A}{Y_{3D}} = \frac{A}{Y_i^m \cdot Y_a^{m-1}}. \tag{16}$$

Observing (16), the shortcoming of 3-D integration is clear: the yield loss has to be accumulated, which means the 3-D integration scheme is inherently more costly than the monolithic scheme. As result, when the 3-D stacking process has an assembling yield of 95%, the total consumed silicon area, S_A, of the 3-D implementation is 28.1 cm^2.

E. Multi-Chip Module

From the perspective of silicon area, the MCM scheme is similar to the 2.5-D scheme expect the following factors:

1) Typically, MCM assembling usually requires a fixed fault coverage level of 99.9% or even higher to ensure the correct functioning of the whole system. Accordingly, we assume a fixed $F_c = 0.999$ in our computation to $S_{A-\text{Test}}$.
2) Due to the difficulty of test access, MCM requires very expensive die carriers to achieve high fault coverage. The carrier cost as well as the extra test preparation time is modeled as having a cost corresponding to a testing time of 60 seconds [47]. As a result, the silicon area of MCM can be computed as

$$\begin{aligned} S_{A-\text{MCM}} &= S_{Ai0} + S_{A-\text{Test}} \\ &= \frac{\frac{A}{m}}{Y_{i-\text{MCM}}} + S_{A-\text{Test}} + S_{A-\text{Carrier}} \end{aligned} \tag{17}$$

where $Y_{i-\text{MCM}}$ can be derived using (8).

Based on the above analysis, the silicon areas of the MCM implementation for different numbers of dies can be computed and shown in Fig. 5.

F. A Summary of the Integration Schemes

In Fig. 6, we show the silicon area values (normalized to the silicon area of monolithic SoC implementation) for the optimum configuration of each VLSI integration schemes discussed above. Obviously, the 2.5-D integration is justified to have a cost advantage over all other schemes as long as proper test coverage can be achieved.

IV. Enabling Technologies

In the previous section, we have demonstrated the advantage of the 2.5-D integration strategy from a yield perspective. We believe that the success of a new VLSI integration scheme depends on the synergy of three key enabling technologies: fabrication technology, testing methodology, and design technology. In this section, we review the status of these technologies and identify the problems that have to be addressed in the future.

A. Fabrication Technology

Under the 2.5-D integration context, a complete VLSI system is an assembly of fabricated, unpackaged dies [18]. Here the key problem is how to vertically bond chips and construct inter-chip contacts in a yield efficient way. One solution can be developed on the basis of the wafer bonding technology (e.g., [26]–[41]). However, the wafer bonding technology poses very high accuracy requirements (better than 1 micrometer [48]) for the alignment equipments. State-of-the-art aligners could only guarantee a precision of $\sim$3 micrometers [49]. In addition, current high-precision aligners were mainly developed for the MEMS industry and typically need double-sided processing of wafers,

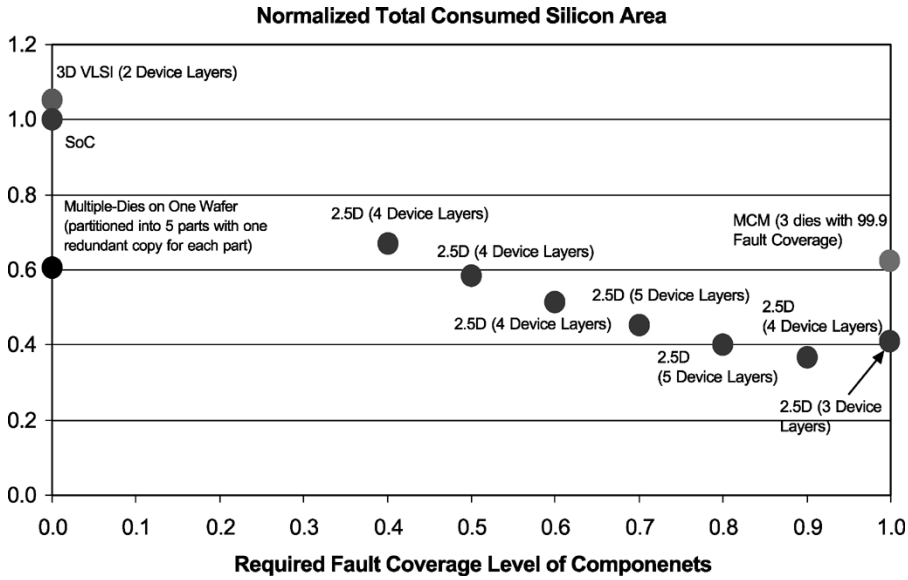

Fig. 6. Silicon area for different integration schemes.

which is not applicable for the purpose of wafer bonding. Recently, aligners designed for wafer bonding have been reported (e.g., [49], [50]), but are still in R&D or early commercial stages.

Accordingly, we envision a bonding technology utilizing a passive, high-precision self-alignment mechanical latch, e.g., a laterally compliant cantilever with a contact clamp. Compliant contacts are desirable for their tolerance to mismatches such as the coefficients of thermal expansion (CTE) of the chips. The alignment process can be organized into multiple stages with increasing accuracies. The refined alignment stage can potentially have very high precision since the alignment features are fabricated in the same process step as the top metallization layer in CMOS technology. Due to the self-assembling process, a key feature of this technology is that it allows "**re-work**." In other words, when a faulty die is identified, it can be removed from the 2.5-D system and replaced by another one. When enhanced with an incremental testing solution, this assembling technology could ultimately overcome the problem of cumulative yield loss.

B. Testing Methodology and Fault Tolerance Technique

To achieve cost-efficient fabrication, a 2.5-D system has to be assembled and tested in an incremental and hierarchical manner. Here the capability of separating different components and selectively testing a single one is critical to guarantee a 2.5-D integrated system to be re-workable. With such a testing methodology, during the assembling process, a die/chip could be selectively tested as soon as it is plugged into a 2.5-D system. This way a faulty die can be identified and replaced immediately and the yield loss would not accumulate during the assembling process.

Such a methodology actually has been developed when the idea of 2.5-D integration was first proposed [18], [42]. Using this testing methodology, every die in the system is isolated with the remaining dies of the system by a dedicated boundary-scan chain and can be selectively powered. These features make it possible to separately and incrementally test every die in either a fully or a partially assembled system.

Meanwhile, testing an embedded IP core is quite similar to testing a die in a 2.5-D system from the testing point of view. Consequently, recently developed testing methods for core-based designs [51] provide another set of testing solutions for the 2.5-D system:

- core isolation techniques such as partial boundary scan chain [52] and test wrapper techniques (e.g., [53]);
- test data propagation to and from a specific core by set other cores into a transparent mode [54];
- reuse of system resource like system bus [55] for test purposes;
- utilization of in-system microprocessor to perform self-testing [56];
- testing solution specifically designed for embedded memories [57], [58].

The essence of the above techniques is to enable separate access to each embedded IP core in a system while trying to reuse existing functionality as much as possible. Accordingly, these testing solutions can be straightforwardly adapted to the purpose of testing 2.5-D systems.

Another approach to overcome the difficulty of test access in a 2.5-D system is to exploit fault tolerance techniques. For a 3-D stacked system, redundant components can be extensively deployed at different granularity levels to compensate the difficulty of testing access. Historically, the fine-grained techniques, such as employing redundant rows or columns of cells in array-styled circuits, have been very successful [43]. On the other hand, coarse-grained techniques (e.g., replicating chip-level functional blocks) have not become popular due to the fact that global failure will impact all modules (including redundant modules) simultaneously [43]. For instance, a short between power and ground in a functional block will lead to the failure of the whole system no matter how many redundant blocks are installed. However, coarse-grained techniques may prove to be very useful in a 2.5-D system since different layers can be fully decoupled and less likely to be affected by a global failure.

TABLE III
DESIGN VARIABLES INVOLVED IN DESIGNING A 2.5-D SYSTEM

Design Variable	Options
Number of Layers	1, 2, 3, …
Process	Standard CMOS, RF-CMOS, SiGe, embedded memory
Inter-Chip Contact Density	With a certain range
Cell Library	High-performance, low-leakage
Design Technology	Digital, analog, optical, MEMS
IP Reuse	Hard/firm/soft IP cores, die-Level IPs, bus standard, embedded operation system
Memory Organization	Memory hierarchy, number of banks per block, cache associativity, memory/cache bus width/line size, memory interface protocol
Communication Protocols	Bus, on-chip networks, asynchronous, global asynchronous local synchronous (GALS)
Heat Dissipation Feature	Heat sink, unconnected inter-chip contact, MEMS micro-pipe

*Monolithic SoC can be considered as a special case of 2.5-D system.

C. Design Technology

With the maturation of 3-D stacking technologies, an essential issue is to develop corresponding CAD tools and design flows. Besides traditional design variables, a design framework for 2.5-D integrated systems has to take into account the following factors.

1) Inter-Chip Contacts: The inter-chip contacts constitute a new level of the interconnection hierarchy. Thus, a key point is how to utilize this physical resource in a systematic manner. In fact, if we assume the inter-chip contacts can be placed everywhere on the chip surface and have an area pitch of $5 \times 5\ \mu\text{m}^2$, then up to 4 million of them will be available on a chip with a die area of $1 \times 1\ \text{cm}^2$. This large amount of communication resource suggests that 2.5-D integration is no longer a packaging option but a design opportunity. Consequently, full potential of 2.5-D system can only be achieved in a strategic design framework.

2) Design Complexity: The formidable complexity of 2.5-D system design is due to both the huge volume of silicon real estate available and the large solution space involved. On one hand, if we stack 4 $1 \times 1\ \text{cm}^2$ chips vertically with the bottom two layers for logic (including SRAM) and top two layers for DRAM, then there would be 5 M gates and 16 MB DRAM available. In addition, the stacking scheme allows virtually any combinations of VLSI technologies to be integrated into a system. As a result, a formidable number of design variables need to be resolved during design process. A sample of these variables is shown in Table III. Among these factors, the memory design is of special importance since the 2.5-D scheme will allow main memory to be directly integrated with the computing resource.

3) Nanometer-Scale Effects: The sub-micrometer effects already have a profound impact on VLSI design. In the 2.5-D system, these issues will be even exaggerated. One crucial problem is the heat dissipation in the upper layer chips that do not directly attached to a heat sink. Excessive heat may stress the system and lead to poor mean time to failure (MTTF). In addition, transistor working under high temperature also has a larger leakage current. Electromagnetic noise will also be an essential concern for 2.5-D systems. For example, the reliability of DRAM and mixed-signal circuitry can be severely affected by the noise generated in the upper and/or lower layer chips.

Considering the above factors, current VLSI design framework should be enhanced with the following components.

- Automatic design exploration. An automatic exploration engine is of key importance to help designers conquer enormous complexity of 2.5-D system integration. Important design variables such as number of layers of chips, technology selection for each device layer of chip, density of inter-chip contact, system partitioning, etc. should be determined through searching the complex solution space in an automatic manner. We envision the exploration engine could extract performance, manufacturability and power consumption estimations from a physical prototype constructed by fast RTL synthesis and coarse-grained placement.
- 2.5-D layout synthesis tools. 2.5-D-aware floorplanning, placement and routing tools can be constructed by extending existing algorithms [60], [61]. These 2.5-D-aware tools could determine the geometric features of a designed system in a stacked space and determine how to assign inter-chip contacts into design hierarchy according to the internal structure of a designed system. An important issue is to consider thermal issues like hot-spot removal and heat dissipation feature placement.
- Memory architecture navigation. Increasing percentage of silicon will be devoted to memory in future VLSI systems. A 2.5-D integrated system could incorporate multiple types of memory blocks (e.g., DRAM, Flash, EEPROM, etc.) with different I/O protocols (SDRAM, DDR, RDRAM) and internal organizations. As a result, we envision memory architecture navigation algorithms to find optimum memory configuration.
- 2.5-D aware analysis tools. Existing parameter extraction, noise analysis, and thermal profiling tools will need to be enhanced to take into account the effects of stacking multiple layers. For instance, thermal and electromagnetic filed solvers several orders faster than existing tools will be indispensable to provide on-line analysis.
- Design for test (DFT) techniques. DFT techniques are essential to enable an incremental and hierarchical testing methodology. Existing testing solutions (e.g., [42], [51]–[58]) could be adapted to meet the requirements of the 2.5-D integration scheme.

V. CONCLUSION

The excessive interconnection delay and cost, as well as the complexity of integrating different technologies on a single chip, are likely to become the major stumbling blocks for the success of monolithic SoCs. To overcome these barriers, the work reported in this paper investigates a 3-dimensional, die stacking based system design strategy, which is designated as the 2.5-D system integration scheme.

We proposed a cost analysis framework to compare various VLSI integration paradigms including monolithic SoC, multiple-reticle wafer, MCM, 3-D integration and the newly pro-

posed 2.5-D scheme. When enhanced with appropriate design and test methodologies, the 2.5-D integration scheme promises a significant cost advantage over other schemes. We then surveyed the fabrication, testing and design technologies that are necessary to enable the new integration paradigm.

REFERENCES

[1] J. D. Plummer, M. D. Deal, and P. B. Griffin, *Silicon VLSI Technology*. Englewood Cliffs, NJ: Prentice-Hall, 2000, ch. 1.

[2] D. Matzke, "Will physical scalability sabotage performance gains?," *IEEE Computer*, vol. 30, no. 9, pp. 37–39, Sep. 1997.

[3] M. Bohr, "Interconnect scaling—the real limiter to high performance ULSI," in *Proc. Int. Electron Device Meeting*, 1995, pp. 241–244.

[4] R. Ho, K. Mai, and M. Horowitz, "The future of wires," *Proc. IEEE*, vol. 89, no. 4, pp. 490–504, Apr. 2001.

[5] W. Maly, "IC design in high-cost nanometer-technologies era," in *Proc. Design Automation Conf.*, 2001, pp. 9–14.

[6] R. Bryant, K.-T. Cheng, A. B. Kahng, K. Keutzer, W. Maly, R. Newton, L. Pileggi, J. Rabaey, and A. Sangiovanni-Vincentelli, "Limitations and challenges of computer-aided design technology for CMOS VLSI," *Proc. IEEE*, vol. 89, no. 3, pp. 341–365, Mar. 2001.

[7] A. Sangiovanni-Vincentelli and G. Martin, "Platform-based design and software design methodology for embedded systems," *IEEE Des. Test Comput.*, vol. 18, no. 6, pp. 23–33, Nov.–Dec. 2001.

[8] A. Matsuzawa, "RF-SoC—expectations and required conditions," *IEEE Trans. Microwave Theory Tech.*, vol. 50, no. 1, pp. 245–253, Jan. 2002.

[9] J. Hennessy and D. Patterson, *Computer Architecture: A Quantitative Approach*, 3rd ed. San Francisco, CA: Morgan Kaufmann, 2002, ch. 5.

[10] D. Burger and J. Goodman, "Memory bandwidth limitations of future microprocessors," in *Proc. 23rd Int. Symp. Computer Architecture*, May 1996, pp. 78–89.

[11] S. Rusu *et al.*, "A 1.5-GHz 130-nm Itanium 2 processor with 6-MB on-die L3 cache," *IEEE J. Solid-State Circuits*, vol. 38, no. 11, pp. 1887–1895, Nov. 2003.

[12] G. Purvis. Demands on 3G Memory Set to Soar. Wirelessweb. [Online]. Available: http://wireless.iop.org/articles/feature/2/6/2/1

[13] D. Keitel-Schulz and N. Wehn, "Embedded DRAM development: technology, physical design, and application issues," *IEEE Des. Test Comput.*, no. 3, pp. 7–15, May/Jun. 2001.

[14] N. Wehn and S. Hein, "Embedded DRAM architectural trade-offs," in *Proc. Design, Automation and Test Europe*, 1998, pp. 704–708.

[15] W. Maly, "Prospects for WSI: a manufacturing perspective," *IEEE Computer*, vol. 25, no. 4, pp. 58–65, Apr. 1992.

[16] N. A. Sherwani, Q. Yu, and S. Badida, *Introduction to Multichip Modules*. New York: Wiley-Interscience, Dec. 1995, ch. 1.

[17] S. Kawamura, N. Sasaki, T. Iwai, M. Nakano, and M. Takagi, "Three-dimensional CMOS IC's fabricated by using beam recrystallization," *IEEE Electron Device Lett.*, vol. EDL-4, no. 10, pp. 366–368, Oct. 1983.

[18] W. Maly *et al.*, "Smart-substrate multichip-module systems," *IEEE Des. Test Comput.*, vol. 11, no. 2, pp. 64–73, Summer 1994.

[19] M. Nakano, "3-D SOI/CMOS," in *Proc. Int. Electronic Device Meeting*, 1984, pp. 792–795.

[20] S. Kawamura, N. Sasaki, T. Iwai, M. Nakano, and M. Takagi, "Three-dimensional CMOS IC's fabricated by using beam recrystallization," *IEEE Electron Device Lett.*, vol. EDL-4, no. 10, pp. 366–368, Oct. 1983.

[21] K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat, "3-D ICs: a novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration," *Proc. IEEE*, vol. 89, no. 5, pp. 602–633, May 2001.

[22] V. Subramanian, M. Toita, N. R. Ibrahim, S. J. Souri, and K. C. Saraswat, "Low-leakage Germanium-seeded laterally-crystallized single-grain 100 nm TFT's for vertical integration applications," *IEEE Electron Device Lett.*, vol. 20, no. 7, pp. 341–343, Jul. 1999.

[23] G. W. Neudeck, S. Pae, J. P. Denton, and T. Su, "Multiple layers of silicon-on-insulator for nanostructure devices," *J. Vac. Sci. Technol. B*, vol. 17, no. 3, pp. 994–998, 1999.

[24] A. Heya, A. Masuda, and H. Matsumura, "Low-temperature crystallization of morphous silicon using atomic hydrogen generated by catalytic reaction on heated tungsten," *Appl. Phys. Lett.*, vol. 74, no. 15, pp. 2143–2145, 1999.

[25] T. H. Lee. (2002, Jan.) A Vertical Leap for Microchips. Scientific American. [Online]. Available: http://www.sciam.com/article.cfm?articleID=000BD05C-D352-1C6A-84A9809EC588EF21&sc=I100322

[26] M. Chan, "The potential and realization of multilayers three-dimensional integrated circuit," in *Proc. Int. Solid State and Integrated Circuit Technology Conf.*, 2001, pp. 40–45.

[27] S. A. Kuhn, M. B. Kleiner, P. Ramm, and W. Weber, "Performance modeling of the interconnect structure of a three-dimensional integrated RISC processor/cache system," *IEEE Trans. Compon., Packag., Manufact. Technol., Part B: Adv. Packag.*, vol. 19, no. 11, pp. 719–727, Nov. 1996.

[28] ——, "Performance improvement of the memory hierarchy of RISC-systems by application of 3-D technology," *IEEE Trans. Compon., Packag., Manufact. Technol., Part B: Adv. Packag.*, vol. 19, no. 11, pp. 709–718, Nov. 1996.

[29] K. W. Lee, T. Nakamura, T. Ono, Y. Yamada, T. Mizukusa, H. Hashmoto, K. T. Park, H. Kurino, and M. Koyanagi, "Three-dimensional shared memory fabricated using wafer stacking technology," in *Proc. Int. Electron Device Meeting*, 2000, pp. 165–168.

[30] M. Koyanagi, H. Kurino, K. W. Lee, K. Sakuma, N. Miyakawa, and H. Itani, "Future system-on-silicon LSI chips," *IEEE Micro*, vol. 18, no. 4, pp. 17–22, Jul.–Aug. 1998.

[31] S. Strickland, E. Ergin, D. R. Kaeli, and P. Zavracky, "VLSI design in the 3rd dimension," *Integration: The VLSI Journal*, vol. 25/1, pp. 1–16, Sep. 1998.

[32] S. M. Alam, D. E. Troxel, and C. V. Thompson, "A comprehensive layout methodology and layout-specific circuit analyzes for three-dimensional integrated circuits," in *Proc. Int. Symp. Quality Electronic Design*, 2002, pp. 246–251.

[33] J.-Q. Lu *et al.*, "A wafer-scale 3-D IC technology platform using dielectric bonding glues and copper damascene patterned inter-wafer interconnects," in *Proc. Int. Interconnect Technology Conf.*, 2002, pp. 78–80.

[34] H. B. Pogge, "The next chip challenge: effective methods for viable mixed technology SoCs," in *Proc. Design Automation Conf.*, 2002, pp. 84–87.

[35] K. W. Guarini *et al.*, "Electrical integrity of state-of-the-art 0.13 μm SOI CMOS devices and circuits transferred for three-dimensional (3-D) integrated circuit (IC) fabrication," in *Proc. Int. Electron Device Meeting*, 2002, pp. 943–945.

[36] M. Ieong *et al.*, "Three dimensional CMOS devices and integrated circuits," in *Proc. Custom Integrated Circuits Conf.*, 2003, pp. 207–213.

[37] B. Burari, "Bridging the gap between the digital and real worlds: the expanding role of analog interface technologies," in *IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2003, pp. 30–35.

[38] J. Mayega, "3-D direct vertical interconnect microprocessors test vehicle," in *Proc. Great Lakes Symp. VLSI*, 2003, pp. 141–146.

[39] T. Mimura *et al.*, "System module: a new chip-to-chip module technology," in *Proc. Custom Integrated Circuits Conf.*, 1997, pp. 439–442.

[40] 3-D Technology. Matrix Semiconductor. [Online]. Available: http://www.matrixsemi.com/index.shtml

[41] 3D Integration. Ziptronix Inc. [Online]. Available: http://www.ziptronix.com/

[42] A. Gattiker and W. Maly, "Smart substrate MCMs," *J. Electronic Testing Theory Applicat. (JETTA)*, vol. 10, pp. 39–53, 1997.

[43] W. Maly, "Computer-aided design for VLSI circuit manufacturability," *Proc. IEEE*, vol. 78, no. 2, pp. 356–390, Feb. 1990.

[44] ——, "Feasibility of large area integrated circuits," in *Wafer Scale Integration*, E. E. Swartzlander Jr., Ed. Boston, MA: Kluwer, 1989.

[45] (2003) Unprobed Wafer Cost. IC Knowledge. [Online]. Available: http://www.icknowledge.com

[46] J. Khare, H. T. Heineken, and M. d'Abreu, "Cost trade-offs in system on chip designs," in *Proc. 13th Conf. VLSI Design*, 2000, pp. 178–184.

[47] D. Ammann *et al.*, "CostAS—KGD process cost modeling," in *Proc. IEEE Innovative Systems in Silicon Conf.*, Oct. 1996, pp. 313–322.

[48] R. Islam, C. Rrubaker, P. Lindner, and C. Schaefer, "Wafer level packaging and 3-D interconnect for IC technology," in *Proc. IEEE/SEMI Advanced Semiconductor Manufacturing Conf.*, 2002, pp. 212–217.

[49] P. Lindner, V. Dragoi, T. Glinsner, C. Schaefer, and R. Islam, "3D interconnect through aligned wafer level bonding," in *Proc. IEEE Electronic Components and Technology Conf.*, 2002, pp. 1439–1443.

[50] T. Suga, M. M. R. Howlader, and T. Itoh, "A new wafer-bonder of ultrahigh prision using surface activated bonding (SAB) concept," in *Proc. IEEE Electronic Components and Technology Conf.*, 2001, pp. 1013–1018.

[51] Y. Zorian, E. J. Marinissen, and S. Dey, "Testing embedded-core based system chips," *IEEE Computer*, vol. 32, no. 6, pp. 52–60, Jun. 1999.

[52] N. Touba and B. Pouya, "Using partial isolation rings to test core-based designs," *IEEE Des. Test Comput.*, vol. 14, no. 4, pp. 52–59, Oct./Dec. 1997.

[53] S. Koranne, "Design of reconfigurable access wrappers for embedded core based SOC test," in *Proc. Int. Symp. Quality Electronic Design*, 2002, pp. 106–111.

[54] I. Ghosh, N. Jha, and S. Dey, "A low overhead design for testability and test generation techniques for core-based systems," in *Proc Int. Testing Conf.*, 1997, pp. 50–59.

[55] M. Nicolaidis, "IP for embedded robustness," in *Proc. Design, Automation and Test Europe*, 2002, pp. 240–241.
[56] R. Rajsuman, "Testing a system-on-a-chip with embedded microprocessor," in *Proc. Int. Testing Conf.*, 1999, pp. 499–508.
[57] R. McConnell and R. Rajsuman, "Test and repair of large embedded DRAM's 1," in *Proc. Int. Testing Conf.*, 2001, pp. 163–172.
[58] ——, "Test and repair of large embedded DRAM's 2," in *Proc Int. Testing Conf.*, 2001, pp. 173–181.
[59] Y. Deng and W. Maly, "A feasibility study of 2.5-D system integration," in *Proc. Custom Integrated Circuits Conf.*, 2003, pp. 667–670.
[60] ——, "Interconnect characteristics of 2.5-D system integration scheme," in *Proc. Int. Symp. Physical Design*, 2001, pp. 341–345.
[61] ——, "Physical design of the 2.5-D stacked system," in *Proc. Int. Conf. Computer Design*, 2003, pp. 211–217.

Yangdong (Steven) Deng (S'98) received the M.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 1998. He is currently working toward the Ph.D. degree in electrical and computer engineering at Carnegie Mellon University, Pittsburgh, PA.

Since 1998, he has been an ECE Research Fellow at Carnegie Mellon University. His research focuses on defining a new 3-D integration paradigm (2.5-D integration) for VLSI circuits and developing corresponding CAD algorithms. His other research interests include physical design algorithms and layout-driven system-level design for system-on-chips.

Wojciech P. Maly (M'81–SM'86–F'91) received the M.Sc. degree in electronic engineering from the Technical University of Warsaw, Poland, in 1970, and the Ph.D. degree from the Institute of Applied Cybernetics, Polish Academy of Sciences, Warsaw, in 1975.

From 1970 to 1973, he was with the Institute of Applied Cybernetics. In 1973, he was with the Technical University of Warsaw and became an Assistant Professor in 1975. From 1979 to 1981, he was a Visiting Assistant Professor of Electrical and Computer Engineering at Carnegie Mellon University, Pittsburgh, PA. Since September 1983, he has been with Carnegie Mellon University, where he is currently a Whitaker Professor of Electrical and Computer Engineering. He has authored, coauthored, or edited a number of books, journals, and conferences papers as well as patents, which have attempted to promote integration of design, test, and manufacturing. His research interests have focused on the interfaces between VLSI design, testing, and manufacturing with the stress on the stochastic nature of phenomena relating these three VLSI domains.

Dr. Maly has received various awards, including honors for his Ph.D. dissertation, Ministry of Higher Education of Poland Research Award, Carnegie Mellon's Benjamin Richard Teare Teaching Award, AT&T Foundation Award for Excellence in Instructing of Engineering Students, Fellowship from Deutsche Forschungsgemeinschaft, SRC 1992 Technical Excellence Award, the 1990 International Test Conference Best Paper Award, the 1994 ESREF Best Paper Award, the 1994 IEEE TRANSACTIONS ON SEMICONDUCTOR MANUFACTURING Best Paper Award, 1995 Best Paper Award from the European Design and Test Conference, and the Eta Kappa Nu CMU Sigma Chapter Excellence in Teaching Award.