Two-Dimensional and Three-Dimensional Integration of Heterogeneous Electronic Systems Under Cost, Performance, and Technological Constraints

Roshan Weerasekera, *Member, IEEE*, Dinesh Pamunuwa, *Member, IEEE*,
Li-Rong Zheng, *Member, IEEE*, and Hannu Tenhunen, *Member, IEEE*

***Abstract*—Present day market demand for high-performance high-density portable hand-held applications has shifted the focus from 2-D planar system-on-a-chip-type single-chip solutions to alternatives such as tiled silicon and single-level embedded modules as well as 3-D die stacks. Among the various choices, finding an optimal solution for system implementation deals usually with cost, performance, power, thermal, and technological tradeoff analyses at the system conceptual level. It has been estimated that decisions made in the first 20% of the design cycle influence up to 80% of the final product cost. In this paper, we discuss realistic metrics appropriate for performance and cost tradeoff analyses both at the system conceptual level in the early stages of the design cycle and in the implementation phase, for verification. In order to validate the proposed metrics and methodology, two ubiquitous electronic systems are analyzed under various implementation schemes and the performance tradeoffs discussed. This case study is used to highlight the importance of a cost and performance tradeoff analysis early in the design flow.**

***Index Terms*—Die stacking, performance and cost tradeoffs, power consumption, system-in-package (SiP), system-on-chip (SoC), system-on-package (SoP), thermal analysis, wafer-level integration (WLI), 3-D integration.**

I. Introduction

HIGH-PERFORMANCE electronic processor systems in portable applications need to satisfy increasingly stringent requirements on energy efficiency under ever more severe performance, cost, weight, and technological restrictions. The solutions explored by the semiconductor industry to meet these challenges are migrating toward 3-D integration options. A major driver behind this trend is the plethora of implementation problems facing gigascale 2-D integration, ranging from technological to architectural. From a fabrication point of view, integrating disparate technologies such as sensors, MEMS structures, and other heterogeneous elements demanded by many applications on a single die is more challenging than connecting separate dies by external interconnections. The 2-D architecture also results in numerous bottlenecks due to area and routing congestion, such as the memory bottleneck in multimedia systems-on-a-chip (SoCs) [1]. Recent developments in fabrication technology have resulted in 3-D integration being a potentially viable option for gigascale integration [2], [3]. Major potential benefits of vertical integration include improved form factor and the reduction in the total length of wiring required for a given system configuration. The wire length reduction alone is reported to reduce the interconnect energy and propagation delay by up to 51% and 54%, respectively, at the 45-nm technology node in [4]. However, the potential gain in performance is a strong function of the die area, as we show in this paper. The reduced parasitics for interconnects can significantly simplify the circuit and power distribution network design for high-performance applications. In mixed-signal systems, noise-sensitive analog/RF circuitry is prone to failure due to interference from their digital counterpart through the base silicon substrate. Three-dimensional integration aids in the solution for noise isolation as it separates the analog/RF and digital circuits into different substrates, with the metal or the dielectric bonding layer used in wafer-bonding technology providing an effective guard ring [5]. The final footprint of the packaged system can also be smaller for a 3-D implementation.

One of the main obstacles to 3-D integration is poor thermal conductivity and heat dissipation and the resultant temperature rise due to the high power density [6]. Another is the relatively poor yield and correspondingly high cost due to the possibility of one faulty die causing an entire stack to be faulty. Balanced against this is the possibility of improving yield by reducing the area of individual dies. A well-known method to transfer heat out of the die is to use thermal vias, which however further increases the routing congestion [7], [8]. Nevertheless, as we show in this paper, careful thermal-via placement in high-performance systems has the potential to effectively control the temperature in 3-D ICs. Some alternative methods that have been proposed, such as integrated microchannel cooling [9], [10] may also be a viable option. Moreover, we show that even though the increased temperature reduces the highest operating frequency, the overall system performance can still be comparatively better than in a 2-D implementation.

Manuscript received June 3, 2008; revised August 14, 2008, October 28, 2008, and January 6, 2009. Current version published July 17, 2009. This work was supported in part by the Swedish International Development Cooperation Agency (SIDA) under the auspices of the research capacity building project at the Department of Electrical and Electronic Engineering, University of Peradeniya, Sri Lanka, and in part by European Union research funding under Grant FP7-ICT-215030 (ELITE) of the 7th framework programme. This paper was recommended by Associate Editor V. Narayanan.

R. Weerasekera and D. Pamunuwa are with the Center for Microsystems Engineering, Lancaster University, LA1 4YR Lancaster, U.K. (e-mail: r.weerasekera@lancaster.ac.uk; d.pamunuwa@lancaster.ac.uk).

L.-R. Zheng and H. Tenhunen are with the School of Information and Communication Technologies, Royal Institute of Technology, 164 40 Stockholm, Sweden (e-mail: lirong@kth.se; hannu@kth.se).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCAD.2009.2021734

However, even as designers are presented with an extra spatial dimension, the complexity of the layout and the architectural tradeoffs also increase. To get a true improvement in performance, an accurate analysis using detailed models at different hierarchical levels is crucial. Even though several previous works have addressed this issue [11], [12], [13], they mostly concentrate on isolated model development, or target some specific type of system. In this paper, we present a cohesive analysis of the technological, cost, and performance tradeoffs for digital and also crucially mixed-mode systems, outlining the choices available at different points in the design and their ramifications. To this end, we collate existing models for area, yield, cost, thermal profile, and performance metrics from the literature, and modify them as necessary in order to facilitate their use for analysis of various 2-D and 3-D integration options using modern technologies.

The main contribution of this paper is in developing a generic methodology for performance and cost estimations of 3-D systems that can be modified for different applications, as well as providing 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 3-D integration. We build on our previous work in [14] and include the testing cost in the cost models. We also crucially model the connectivity between temperature and performance.

The rest of this paper is organized as follows: In Section II, we present a short overview of 3-D integration technologies, while in Section III, all models used are discussed in detail, including comparisons between models where a choice exists, and justification and validation for the choice where relevant. This section also presents our estimation and tradeoff analysis methodology. Section IV presents a case study where the models and methodology proposed in this paper are applied to two different applications and cost and performance metrics derived for various 2-D and 3-D integration options, highlighting the importance of a tradeoff analysis early in the design flow. We end with a discussion and our conclusions.

II. 3-D Integration Technologies

Terms that have been coined for packaging technologies in the literature are sometimes interpreted in different ways depending on the context of usage. Hence, we define a few terms at the outset to avoid ambiguity. The term System-On-Package (SoP) is used to refer to a 2-D multichip module (MCM) arrangement where packaged chips are situated on a single substrate or across a board. A 3-D stacked arrangement of chips or dies is referred to as a System-In-Package (SiP). An electronic system that is laid out on a single die in 2-D is referred to as an SoC.

Three-dimensional integration techniques can be categorized into two major approaches [15]: folding and stacking. In folding, a planar assembly with a flexible substrate is folded into several layers in order to form a compact shape. In this approach, the interconnect length is longer than in the stacked approach described below, but a very small size can be achieved.

Stacking can be carried out at the chip level with either chip-to-chip, package-on-package, or MCM-to-MCM bonding using epoxy or glues and wire bonding, as shown in Fig. 1(a). These techniques present the opportunity to stack Known-Good-Dies (KGDs) in layers [16], improving system yield. As an alternative to chip stacking, 3-D integration can be performed at the wafer level. Different blocks can be processed on separate wafers, and interconnected vertically using through-hole vias (THV) or through-Si vias (TSV) to form global communication links [Fig. 1(b)]. This effectively reduces the latency and power drawbacks inherent to global communication in SoCs by reducing the average interconnect length and providing thermal dissipation channels. 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) [17]. Some other possibilities not considered here include capacitive [18], [19] or inductive [20] links for wireless communication between chips [21].

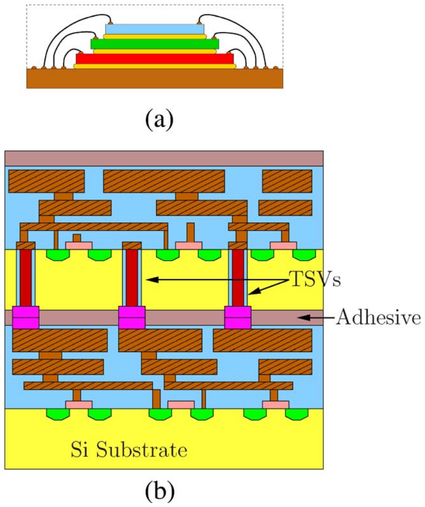

Fig. 1. Three-dimensional integration options. (a) System-in-package (die stacking using wire bonding). (b) WLI with vertical interconnects.

In this analysis, we concentrate on stacking methodologies and compare between 3D-SiP, 3D-D2W, and 3D-W2W technologies, the most commonly practiced approaches in 3-D integration.

III. Performance and Cost Estimation Models

Previous works that addressed cost and performance tradeoffs include [11] and [12], 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 [13] describe a yield and cost model for 3-D stacked chips with particular emphasis on the effect on yield of the number of THV.

We have previously discussed issues around the design choices of SoC and SoP for mixed signal circuits [11]. The overall cost estimation process that we propose here is shown in Fig. 2. The first task is to find chip/module area, as the cost and performance is predicated on the 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,

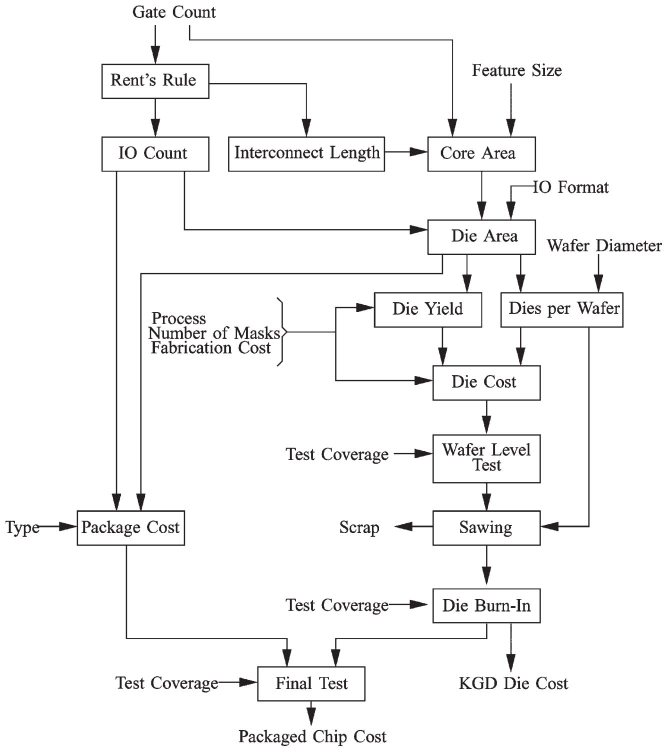

Fig. 2. Overall cost modeling flow for a chip.

using gate information and technology scaling. This procedure takes the gate count and technology information as inputs, and estimates a core area which in turn yields the die area. The die area, together with input-output (I/O) information, is used to estimate the die cost as well as the package cost, leading to the overall chip cost after the addition of other relevant costs such as testing, dicing, and burn-in costs.

However, the area of an analog chip depends not only on the number of transistors and their sizes (in practice, minimum-sized transistors are not used in analog circuits) but also on the circuit architecture. For example, in a voltage-controlled oscillator, the area of the on-chip inductor may be hundreds of times larger than that of a transistor. In an analog-to-digital or digital-to-analog converter, on-chip resistors and capacitors may occupy a large fraction of the total area. Full custom design experiences are necessary to estimate the size of an analog chip. The models for core area described below are for digital implementations; it is assumed that area information for analog blocks is available. However, all models following on from the core area are valid for mixed-mode systems. Instances where variations with respect to pure digital systems occur are identified and supported by appropriate models.

A. Chip/Module Area Models

In this section, die, chip, and module area models are presented, based largely on the SUSPENS model proposed in [22], modified by revisions proposed in the literature over the years, and some additional refinements proposed in this paper.

1) Die/Chip Area Model: The area occupied by the transistors and their interconnects is termed the core area (A_{core}) of the chip. This area can either be interconnect-capacity limited or transistor-area limited depending on the logic type and the available resources such as number of metal layers. For example, memories usually have a very regular structure and do not require many interconnection levels, resulting in a very high packing density. However, digital logic circuits are less regular and require more connectivity, resulting in the area being either interconnect limited or gate area limited. The core area of a chip is estimated from

$$A_{\mathrm{core}} = \max\left\{N_{\mathrm{g}}d_{\mathrm{g}}^2, N_{\mathrm{g}}A_{\mathrm{g}}\right\} \tag{1}$$

where N_{g} is the average number of logic gates, A_{g} the average logic gate area, and d_{g} the gate dimension. For static CMOS, the average logic gate is considered to be a two-input NAND gate with a fan-out of three identical NAND gates; for dynamic logic, [23] proposes a two-input NOR with fan-out of an inverter as the representative gate. The reasons for these choices are that the NAND gate is one of the commonest gates in random logic and is widely used in density metrics; NOR gates are widely used in high-speed dynamic logic.

The estimation of A_{g} in (1) can be carried out based on the height and width of the cell layout, determined by the contacted metal pitches of the local metal layers. As per [23], the size of a two-input NAND gate for a standard drive strength is 4 metal pitches across by 16 metal pitches, i.e., $4p_{\mathrm{wL}} \times 16p_{\mathrm{wL}}$.

The gate dimension is defined in [22] as

$$d_{\mathrm{g}} = \frac{f_{\mathrm{g}}\overline{R_{\mathrm{m}}}p_{\mathrm{w}}}{e_{\mathrm{w}}n_{\mathrm{w}}} \tag{2}$$

where f_{g} refers to the gate fan-out, p_{w} to interconnection pitch, n_{w} to the number of interconnection layers, e_{w} to the utilization efficiency of interconnects, and $\overline{R_{\mathrm{m}}}$ to the average interconnect length. This model was further validated and used in [24], and also used in [25]. However, in modern technologies, the number of available wiring levels is much higher, and the variation in wire pitch between the lowest and highest levels is significant; for example, the pitch of a global wire is typically several times that of local wires. Additionally, the higher the number of levels, the greater the congestion introduced by the presence of vias and studs needed for the interconnection of adjacent wiring levels. Therefore, (2) requires a refinement in order to be used with multilevel interconnect structures. One proposal, in [24], is to use an average value for p_{w}, while another, in [26], is to estimate $p_{\mathrm{w}}/n_{\mathrm{w}}$ considering only local and global wires. In addition, due to unequal usage of power and clock lines on the different metal layers and via blockage, the wiring efficiency for signal wires varies from one level to another. In this paper, the change in wiring pitch and different wiring efficiency factors for each layer, as well as the effects of via blockage are considered by modifying (2) to get

$$d_{\mathrm{g}} = \frac{1+f_{\mathrm{g}}}{2}\,\frac{\overline{R_{\mathrm{m}}}}{\sum\limits_{i=1}^{n_{\mathrm{w}}}\frac{e_{\mathrm{w},i}k_{\mathrm{p},i}}{p_{\mathrm{w},i}}} \tag{3}$$

where $k_{\mathrm{p},i}$ and $e_{\mathrm{w},i}$ are the wiring utilization factor (modeling the effect of via blockage) and wiring efficiency factor (modeling the routing efficiency), respectively, for the ith layer. Such an approach is suggested in [23] and [27]. The term $(1+f_{\mathrm{g}})/2$

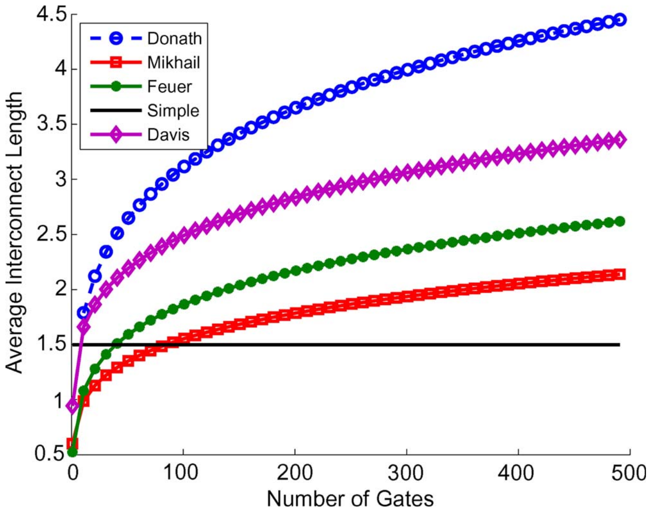

Fig. 3. Average wire length prediction models.

is a correction to take into account the fact that a logic gate fans out to several gates. The term $k_{\mathrm{p},i}$ is the fraction of metal layer i occupied by wires, while e_{w_i} can be expressed as the product of three factors as

$$e_{\mathrm{w},i} = e^{i}_{\mathrm{rout}}\left(1 - b^{i}_{\mathrm{PGC}}\right)\left(1 - b^{i}_{\mathrm{via}}\right) \tag{4}$$

where e^{i}_{rout} is the efficiency of the routing tool for the ith layer (approximately constant over all layers), b^{i}_{PGC} the blockage due to power/ground/clock nets, and b^{i}_{via} the blockage due to vias [27]. Sai-Halasz [28] estimates that in the case of two layers of identical wire pitch, the top layer blocks between 12%–15% of the wiring capacity of the lower layer, and further recommends that the blocking percentage between levels of varying pitch be scaled proportionally with pitch. Hence, b^{i}_{via} on the first wiring layer can be between 10% and 50%, and progressively smaller on higher metal levels [29]. By contrast, b_{PGC} for the topmost level is around 30%–40%, and less than 3% for the lower levels [23]. It is possible, however, to assume a constant e_{w} for all wiring layers by a process of averaging the different values in a first-order approximation.

The simplest method to estimate the average wire length $\overline{R_{\mathrm{m}}}$ is to assume that exactly half the wires in a module traverse the length of a gate pitch F_{p} and the remainder the length of two gate pitches $2F_{\mathrm{p}}$ [30]. Under this assumption, the average wire length $\overline{R_{\mathrm{m}}}$ in gate pitches is equal to 1.5. However, more sophisticated theoretical and empirical treatments for wire length prediction based on Rent's rule have been proposed in the literature, such as Donath's [31], Feuer's [32], Mikhail's [33], and Davis's [34] models. The basic premise in these approaches is the recursive application of Rent's rule (see next paragraph for more details on Rent's rule) to smaller and smaller logic blocks, and the calculation of their external communication requirements. For the tradeoff analysis, we use Donath's model, since it gives an upper bound (Fig. 3) on average wire length, when the average interconnection length is given by

$$\overline{R_{\mathrm{m}}} = \frac{2}{9}\,\frac{1-4^{(p-1)}}{1-N_{\mathrm{g}}^{(p-1)}}\left(7\frac{N_{\mathrm{g}}^{(p-0.5)}-1}{4^{(p-0.5)}-1} - \frac{1-N_{\mathrm{g}}^{(p-1.5)}}{1-4^{(p-1.5)}}\right) \tag{5}$$

for $p \neq 0.5$, and

$$\overline{R_{\mathrm{m}}} = \frac{2}{9}\,\frac{1-4^{p-1}}{1-Ng^{p-1}}\left(7\log_4 Ng - \frac{1-Ng^{p-1.5}}{1-4^{p-1.5}}\right) \tag{6}$$

for $p = 0.5$. Here, p is Rent's exponent.[1]

When packaging the core, the I/O pads providing connectivity 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 pad pitch. Then, the die area is given by

$$A_{\mathrm{die}} = \max\left\{(\sqrt{A_{\mathrm{core}}} + 2p_{\mathrm{p}})^2, \left(\frac{N_{\mathrm{p}}p_{\mathrm{p}}}{4} + 2p_{\mathrm{p}}\right)^2\right\} \tag{7}$$

where p_{p} is the peripheral in-line pad pitch and N_{p} is the total number of I/O pads. When the I/O pads are area array distributed, formulas for the die area is given in [35].

Once the die area is known, the packaged chip area can be estimated according to the following equation:

$$A_{\mathrm{pkged\_chip}} = (\sqrt{A_{\mathrm{die}}} + 2L_{\mathrm{bnd}})^2 \tag{8}$$

where L_{bnd} is the bond wire length.

N_{p} in (7) is calculated from the well-known Rent's rule, the empirical equation that estimates the growth in the number of signal pins on a circuit as a function of the logic components in it. It usually takes the form

$$N_{\mathrm{p}} = K \cdot N_{\mathrm{g}}^{\rho} \tag{9}$$

where ρ is Rent's exponent, K is Rent's coefficient, and N_{g} is the number of logic gates on the chip or logic partition. Rent's rule in this form is valid only for homogeneous systems, not for more complex systems where modules with different architectures are integrated to form an SoC. A form of Rent's rule described in [36], which argues that the same power-law expression holds with modified K and ρ parameters, is given by

$$K_{\mathrm{eq}} = \sqrt[N_{\mathrm{g\_eq}}]{\left(\prod_{i=1}^{n} K_i^{N_{\mathrm{g\_}i}}\right)}$$
$$\rho_{\mathrm{eq}} = \frac{\sum_{i=1}^{n} \rho_i N_{\mathrm{g\_}i}}{N_{\mathrm{g\_eq}}} \tag{10}$$

where K_i and ρ_i are the usual Rent's rule parameters, $N_{\mathrm{g\_}i}$ is the number of gates in block i, and $N_{\mathrm{g\_eq}} = \sum_{i=1}^{n} N_{\mathrm{g\_}i}$.

The chip area models defined in (1) through (7) are compared against actual data from two microprocessors, the Alpha 21164 and Intel Pentium, reported in [23], which gives transistor counts as well as area breakdowns for cache memory, CPU logic and I/O pad ring (Table I). The gate-area-limited and interconnect-limited areas reveal that for both processors, the total area is interconnect limited according to the models, and that the final predicted value is within 9% of the actual value. In general, the greater the amount of data available in a particular

[1] The Rent exponent for wire length estimation is approximately 0.1 higher than that used for estimating pin count [23].

TABLE I
VALIDATION OF AREA MODELS FOR TWO MICROPROCESSORS

Parameter	Alpha 21164	Pentium
Die Area (mm^2)	299	163
Memory Area (mm^2)	102	44
CPU Logic Area (mm^2)	180	111.8
I/O Pad Area (mm^2)	17	7.2
Transistor Count (M)	9.3	3.1
Memory Transistors (M)	6.71	0.971
Technology	0.5 μm CMOS	0.6 μm BiCMOS
n_w	4	4
Wiring Pitch Values ($p_{w,i}$) (μm)	1.125,1.125,3,3	1.4,1.7,1.7,3.5
Rent's Constant p	0.35	0.35
A_g (μm^2)	81	125
e_w	0.26	0.32
N_g (M)	0.648	0.532
f_g	3	3
Gate-Area-Limited Area (mm^2)	52	89
Interconnect-Limited Area (mm^2)	172	108

technology for the accurate empirical estimation of relevant constants, the higher the accuracy of the models for future predictions.

2) Module Area Model: MCM technology can be used as a possible implementation of SoP. The MCM substrate area A_{sub} can easily be estimated by the method outlined by Bakoglu [22]. If the chip pitch F_{p} is known, the SoP area is

$$A_{\text{sub}} = N_{\text{c}} F_{\text{p}}^2 \tag{11}$$

where N_{c} is the number of chips. It is understood that if only one chip carrier is available on the module, the footprint cannot be smaller than the chip carrier's size, and will be limited by the interconnection capacity of the module. In Bakoglu's method, the interconnect-capacity-limited substrate area is predicated on the average interconnect length at the module level, $\overline{R_{\text{M}}}$, calculated using the same approach as in the chip level estimations. In (6), the number of gates N_{g} is replaced with the number of chips N_{c}. Furthermore, the Rent's exponent for modules is different from that for chips. Hence, F_{p} can be limited by either the die-size or the chip carrier-related size. Therefore, the chip pitch is given by the limiting constraint [22]

$$F_{\text{p}} = \text{MAX}\left\{\frac{F_{\text{o}}}{F_{\text{o}}+1}\frac{\overline{R_{\text{M}}}N_{\text{mcm}}P_{\text{w\_mcm}}}{N_{\text{c}}e_{\text{w}}n_{\text{w}}}, D_{\text{c}}, P_{\text{c}}\right\} \tag{12}$$

where N_{c} is the chip count, F_{o} the average fan-out of a chip's I/O (typically 1.5), N_{mcm} the total number of chip I/Os and the I/Os to and from the MCM, and n_{w} and $P_{\text{w\_mcm}}$ the number and pitch of module wiring levels, respectively, D_{c} the dimension of the chip and P_{c} the dimension of the chip carrier.

However, this approach assumes that the components to be arranged in an MCM substrate are homogeneous, which is usually not true for mixed-signal systems. It is understood that this restriction is critical in two respects [30]: 1) in the derivation of the wiring capacity limited footprint and 2) in the determination of the module size. This limitation can be overcome by recomputing an effective chip count and corresponding average interconnect length for each component as follows:

$$\text{Effective } Nc_i = \frac{NIO_{\text{mcm}}}{NIO_{\text{chip\_}i}} \tag{13}$$

where NIO_{mcm} is the total number of I/O connections in the whole module, and $NIO_{\text{chip\_}i}$ is the number of I/O connections that the ith component requires. The following summation can be used to find the total SoP (MCM) area [30]:

$$A_{\text{SOP}} = \sum_{i=1}^{N_{\text{c}}} F_{p_i}^2. \tag{14}$$

B. Yield and Cost Analysis

The yield of a bare silicon die, Y_{d}, depends on electrical defects created on each mask layer in the fabrication process and the total area of the chip. In [37] a yield function for the bare silicon die, i.e., the fraction of dies that is estimated to be fault free before wafer-level testing, is proposed

$$Y_{\text{d}} = \frac{1}{(1+SD_0A)^{\frac{N}{S}}} \tag{15}$$

where D_0 is the average electrical defect density, S the shape factor of (what is assumed to be) the Gamma distribution of electrical defect density, N the number of mask layers, and A the chip area. System yield is a function of the yield of individual components and the yield of the integration methodology used. This is basically the multiplication of the yields of all dies, substrate fabrication processes, and bonding processes. Thus, overall yield can be uneconomically low for complex systems, unless KGD methods are used.

The chip yield, Y_{chip}, is defined as the die yield after wafer-level testing, i.e., the fraction of dies that is estimated to be fault free after wafer-level testing. This is estimated from the fault coverage, which is defined as the fraction of defects that are identified in the test, and the actual yield of the die on the wafer. When the fault coverage level is denoted by Fc, [30] shows the chip yield is given by

$$Y_{\text{chip}} = Y_{\text{d}}^{(1-Fc)}. \tag{16}$$

After dicing, known defective dies are scrapped and the rest sent on for burn-in. The fraction of dies that are available for burn-in and test is known as the pass fraction PF defined as the fraction of dies estimated to be fault free, after the dies that were detected to be faulty in the wafer-level test are discarded. Pass fraction is given by

$$PF = Y_{\text{d}}^{Fc}. \tag{17}$$

However, the higher the fault coverage, the higher the cost; the extra testing time results in extra cost including labor, and equipment usage costs. Assuming that a higher fault coverage level requires significantly increased testing time, the following exponential model correlating test coverage level with testing time t_{test} is proposed in [38]

$$F_{\text{c}} = 1 - e^{-kt_{\text{test}}}. \tag{18}$$

Here, k is an empirical constant that defines the steepness of the exponential function. It is assumed that 60 s is enough to achieve 99.99% coverage, in order to estimate the value of k [38]. Thus, k is calculated to be 0.1. In addition, it is assumed that wafer-level testing achieves 80% fault-coverage, and testing after burn-in achieves a fault coverage level of 99%. Then, the testing cost can be assumed to be linearly propotional to the testing time [38]

$$C_{\text{test}} = C_t t_{\text{test}}. \tag{19}$$

Now, the cumulative cost per die or MCM at the end of each process step is computed as follows:

$$C_{1,i} = \frac{C_{1,i-1} + C_i}{PF} \tag{20}$$

where $C_{1,i-1}$ is the accumulated cost of all steps up to but not including the present step, C_i is the cost of the present step, and PF is the percentage of the die or MCMs that passes the current step. The bare-die cost is a function of the raw materials, process cost, and mask cost of a wafer, which are specific to a given technology, as well as the number of dies per wafer and the die yield

$$C_1 = \frac{C_{\text{wafer}}(raw, process, mask)}{N_{\text{die}} Y_{\text{die}}}. \tag{21}$$

The package cost is calculated using a price versus pin count assumption as in [39]. For a peripheral I/O single-chip plastic package, the cost in U.S. dollars is

$$C_{\text{pkg}} = 0.01 e^{1.16 \log(NIO) - 2.09}. \tag{22}$$

C. Analytical Die Thermal Model for 2-D and 3-D Integration

Thermal integrity is a critical issue in all high-performance chips because system reliability is strongly dependent on the temperature. For vertically stacked chips, due to the higher power density in the stacked arrangement, it is difficult to remove the excess heat from chips or dies that have more than 1° of separation from the heat sink. The increased heat causes further leakage, which, in turn, increases the temperature, an undesirable cycle which can cause catastrophic breakdown. In the following analysis, the contribution to the chip temperature from interconnect joule heating is disregarded.

Assuming the heat dissipates through the Silicon substrate, the average die temperature can be usually described using a 1-D heat equation when the die size is much larger than its thickness (t) [41]

$$T_{\text{die}} = T_{\text{ambient}} + \left(\frac{t}{kA}\right) P_{\text{chip}} \tag{23}$$

where T_{ambient} is the ambient temperature, P_{chip} is the chip power dissipation, A is the chip area, and k is the thermal conductivity of the material. The factor t/kA in (23) is known as the effective thermal resistance (R) of the substrate layer and the package.

If the same assumption is made that the die size is much larger than its thickness, the maximum temperature in a 3-D IC occurs at the highest device layer. Then, as described in [41], the average die temperature of a 3-D IC with m layers is

$$T_{\text{3D}} = T_{\text{ambient}} + \sum_{i=1}^{m} R_{(i-1),i} \sum_{j=i}^{m} P_j \tag{24}$$

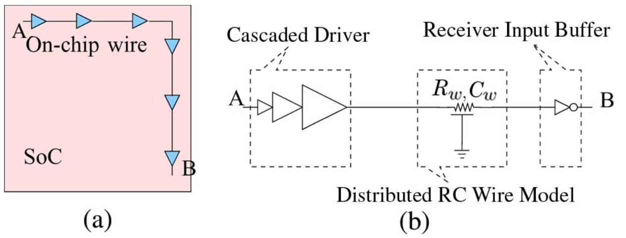

Fig. 4. Signal propagation for worst-case latency in a SoC. (a) Signal propagation link. (b) Interconnect model.

where $R_{(i-1),i}$ is the effective thermal resistance between the i^{th} and $(i-1)$th layer including the glue layer where applicable, and P_j is the power dissipation in the kth active layer.

Recently, a 2-D thermal model called Hotspot [42] that takes into account lateral as well as vertical heat dissipation in 2-D SoCs was proposed. This model requires information of at least the block-level architecture and power density of each block, and heat resistances and capacitances of the resulting 3-D heat flow network of all layers up to and including the heat sink. When such information is available, the 1-D thermal model in the proposed estimation flow can be replaced by a more accurate model. In the absence of such information, as in most *a priori* estimations, a 1-D model with an average junction temperature across the die appears to be reasonably accurate, and has been used in many prior works [41], [43], [44].

With the increasing power density of nanoscale chips, die temperatures are expected to rise substantially. The thermal problem is further aggravated by the fact that leakage power is exponentially dependent on temperature. Hence, rising temperatures lead to larger leakage power dissipation and vice versa in a positive feedback relationship. As was mentioned in Section I, one effective way to alleviate the excessive heat generated in 3-D ICs is to incorporate dummy thermal vias (T-vias); the thermal conductivity of a die layer is significantly improved by the existence of thermal vias. When k_{thv} and k_{layer} are the thermal conductivity of a thermal via and the layer, respectively, and m is the fraction of area occupied by the thermal vias to the total area, the effective thermal conductivity of the layer is [15]

$$k_{\text{eff}} = m k_{\text{thv}} + (1-m) k_{\text{layer}}. \tag{25}$$

To estimate the thermal resistance, the effective thermal conductivity coefficient for each pair of layers, for example, die and glue, should be found.

D. Interconnect Performance Models

In the performance estimations, the latency for the longest possible link is the characteristic metric used for comparison. For example, in a planar system, the latency between two diagonal corners is considered, while in a 3-D system, the delay from a corner on the bottom chip to a diagonally opposite corner on the top chip is considered. Please refer Figs. 4–6 for the schematics.

1) On-Chip Wire Delay: Typically, on-chip global signal wires are highly resistive while the inductance is negligible. Hence, signal transmission obeys the diffusion equation.

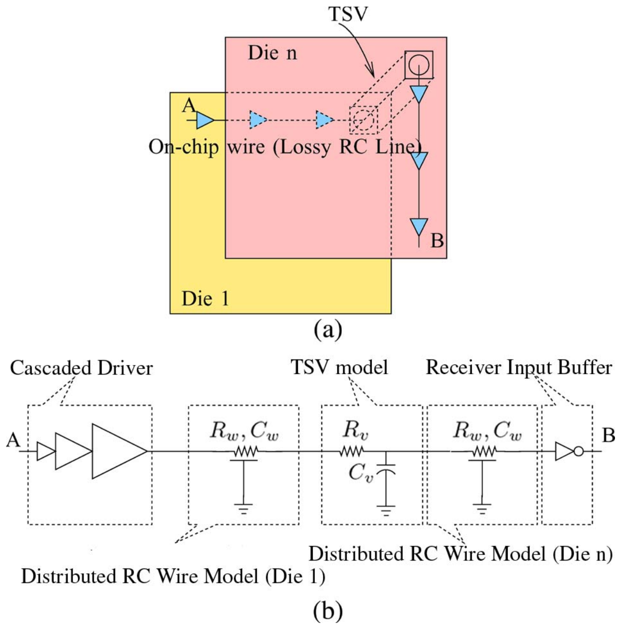

Fig. 5. Signal propagation for worst-case latency in 3-D wafer-level chip stack. (a) Signal propagation link. (b) Interconnect model.

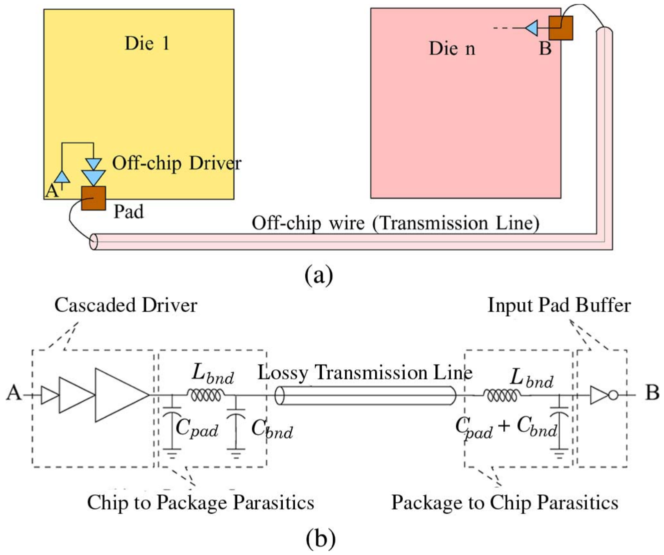

Fig. 6. Signal propagation for worst-case latency in 2D-SoP and 3D-SiP type arrangements. (a) Signal propagation link. (b) Interconnect model.

The appropriate model therefore, is a distributed resistance–capacitance (RC) line [22], [46]. A very good approximation to the delay over an RC dominated wire with capacitive load C_{L} connected at the far-end is given by the first-order Elmore approximation [22]

$$t_{\mathrm{rc\_d}} = 0.693\left\{R_{\mathrm{d}}(C_{\mathrm{d}} + c_{\mathrm{w}}L + C_{\mathrm{L}}) + r_{\mathrm{w}}LC_{\mathrm{L}}\right\} + 0.377r_{\mathrm{w}}c_{\mathrm{w}}L^2 \tag{26}$$

where R_{d} is the driving inverter's equivalent output impedance and C_{d} the self-loading drain diffusion capacitance, while c_{w} and r_{w} are the per-unit-length capacitance and resistance of the interconnect and L its length.

The wire capacitance incorporates the coupling capacitance to adjacent wires, if necessary with an appropriate switching factor to allow for worst-case coupling [47], resulting in a combined total equivalent capacitance. It can be seen that the delay increases exponentially with length when the wire parasitics dominate. The most common method of reducing this delay over long interconnects is to insert repeaters at appropriate positions, which makes the wire delay linear with length. However, repeater insertion is effective only when wire time constant $(r_{\mathrm{w}}c_{\mathrm{w}}L^2)$ is at least seven times the time constant of a repeater $(R_{\mathrm{d}}(C_{\mathrm{d}} + C_{\mathrm{g}}))$ [22]. The 50% delay for a repeater inserted on-chip wire of length L is

$$t_{\mathrm{rc}} = k\left\{0.69\frac{R_{\mathrm{d}}}{H}\left[H(C_{\mathrm{d}} + C_{\mathrm{g}}) + \frac{c_{\mathrm{w}}L}{k}\right] + 0.69\frac{r_{\mathrm{w}}L}{k}HC_{\mathrm{g}} + 0.377\frac{r_{\mathrm{w}}L}{k}\frac{C_{\mathrm{w}}L}{k}\right\} + 0.69\frac{R_{\mathrm{d}}}{H}(C_{\mathrm{d}} + C_{\mathrm{L}}) \tag{27}$$

where H and k are the delay optimal repeater sizes and numbers given by $H = \sqrt{R_{\mathrm{d}}(C_{\mathrm{L}} + cwL)/r_{\mathrm{w}}LC_{\mathrm{g}}}$ and $k = L\sqrt{0.4r_{\mathrm{w}}c_{\mathrm{w}}/0.69r_{\mathrm{w}}LC_{\mathrm{g}}}$, respectively.

Finally, the total propagation delay of the on-chip global wire, as shown in Fig. 4(b), is the sum of the cascaded buffer delay (t_{drv}) at the near-end and the repeater-inserted delay of the RC wire

$$t_{\mathrm{intra}} = t_{\mathrm{drv}} + t_{\mathrm{rc}}. \tag{28}$$

2) Off-Chip Wire Delay: Interchip wires on a typical package substrate are characterized by conductors with low resistivity and a relatively large cross section in a low-loss dielectric, making losses due to shunt conductance negligible. Hence, signal transmission exhibits transmission line behavior. In a lossy transmission line, both RC and LC delays coexist. For LC dominated wires, the signal propagation delay is equal to its time of flight

$$t_{LC} = t_{\mathrm{tof}} = L\sqrt{l_{\mathrm{w}}c_{\mathrm{w}}}. \tag{29}$$

If a wire is a very resistive transmission line, the following empirical formula for adding the time-of-flight (t_{tof}) delay and conventional RC delay $(t_{\mathrm{rc\_tl}})$ was found in [28] to accurately predict the total wire delay

$$t_{\mathrm{RLC}} = \left(t_{\mathrm{tof}}^{1.6} + t_{\mathrm{rc\_tl}}^{1.6}\right)^{\frac{1}{1.6.}} \tag{30}$$

For the interchip communication link shown in Fig. 6(b), the following expressions can be derived:

$$t_{\mathrm{rc\_tl}} = 0.693\left[Z_0(Cd + C_{\mathrm{pad}} + C_{\mathrm{bnd}} + 0.5C_{\mathrm{L}}) + \frac{L_{\mathrm{bnd}}}{Z_0} + r_{\mathrm{w}}L(C_{\mathrm{pad}} + C_{\mathrm{bnd}} + C_{\mathrm{L}})\right] + 0.4r_{\mathrm{w}}c_{\mathrm{w}}L^2. \tag{31}$$

Finally, the total delay for the interchip communication link is the summation of the cascaded driver delay (t_{drv}) and the RLC-wire delay (t_{RLC})

$$t_{\mathrm{inter}} = t_{\mathrm{drv}} + t_{\mathrm{RLC}}. \tag{32}$$

3) Thermal Effect on Interconnect Performance: The driver resistance R_{d} and wire resistance r_{w} both increase with temperature. Usually, R_{d} is expressed in terms of the saturation current of the device when the gate voltage is equal to the supply voltage

$$R_{\mathrm{d}}(T) = \frac{V_{\mathrm{dd}}}{K v_{\mathrm{sat}}(T) W \left(V_{\mathrm{dd}} - V_{\mathrm{th}}(T)\right)^{\alpha}} \tag{33}$$

where K is a constant that is specific to a given technology, T is the temperature in kelvin, V_{th} is the threshold voltage at temperate T, and v_{sat} is the saturation velocity. As validated in [48], when V_{dd} is sufficiently larger than V_{th}, the change in V_{th} with temperature is relatively insignificant. However, as V_{dd} is scaled down, V_{th} has a comparable and counter effect to the change in v_{sat}. Therefore, for a 65-nm CMOS technology the driver resistance can be taken as a constant with increasing temperature [49].

Wire resistance R_{w} increases linearly with temperature due to the change in the effective metal resistivity in relation to the barrier layer. In order to characterize the dependence of wire resistance with temperature, a linear relationship given by

$$R_{\mathrm{w}}(T) = \frac{\rho(T_0) l}{tw} \left[1 + t_{\mathrm{cr\_bulk}}(T - T_0)\right] \tag{34}$$

can be used [50]. In (34), $R(T)$ is the wire resistance at any given temperature T, $\rho(T_0)$ is wire resistivity at the reference temperature T_0, w and h are wire width and height, and $t_{\mathrm{cr\_bulk}}$ is the temperature coefficient of resistance of the bulk material, a good approximation for which is $t_{\mathrm{cr\_bulk}} = 0.0039\ ^{\circ}\mathrm{C}^{-1}$ [50].

IV. Tradeoff Analysis for SoC, SoP, and 3-D Implementations

The models and methodology proposed in this paper are demonstrated in a case study comprising a comparison of two mixed-signal systems, a wireless sensor, and a 3G mobile terminal. The *wireless sensor* contains a 2-Mb DRAM, an application-specified integrated circuit (ASIC), and Microprocessor with gate counts of 500k and 300k, respectively, and an Analog/RF block occupying an area of 2 mm^2. It also contains a MEMS sensor with an area of 1 mm^2. The *3G mobile terminal* has a similar architecture, but with a larger memory (DRAM) of 128 Mb, and a CMOS image sensor (IS) with a pixel size of 1.75 μm × 1.75 μm, and resolution of 8 $Megapixels$ instead of the MEMS sensor [51]. Furthermore, in the analysis, we consider the ASIC and Microprocessor together as a single logic block, treating our target system as comprising only four megacells: analog/RF, logic, memory, and a MEMS sensor or CMOS IS. For all integration schemes, the underlying manufacturing process is a 65-nm 11-metal CMOS process with a wafer diameter of 300 mm, a lower level wire pitch of 152 nm and a global level pitch of 290 nm [2]. We also assume a peripheral in-line pad arrangement and wire bond packaging. All other key parameters are listed in Table II.

For the different blocks in the systems under consideration, the core areas have been estimated using (1). In the case of DRAM, the core area is simply $N_{\mathrm{g}} A_{\mathrm{g}}$ as memories are cell-area limited due to their densely packed structure. The estimated values for all blocks, in both cases, are shown in Tables IV and V.

TABLE II
Representative Values for a 65-nm technology and Summary of Notation for Major Parameters Used in the Analysis

Notation	Parameter	Value
Do	Defect Density per m^2	250
S	Shape Factor	0.6
N_{dram}	DRAM Mask Layers	13
N_{logic}	Logic mask Layers	18
N_{RF}	CMOS RF Mask Layers	12
N_{MEMS}	MEMS Process Mask Layers	6
N_{CIS}	CMOS Image Sensor Process Mask Layers	10
D_{wafer}	Wafer Diameter	300 mm
C_{lgc}	Process cost per mask layer (logic)	700 $
C_{mixed}	Process cost per mask layer (mixed-signal)	1000 $
C_{mcm}	MCM-D cost per unit area per layer	1000 $
C_{asmb}	Cost of assembly per pin	0.01 $
C_{sub}	Cost of substrate	300 $
C_{3Dvia}	Cost of making a through hole via in WLP	0.01 $
C_{rewrk}	Cost of Rework	3 $
C_{SOI}	Cost of SOI substrate	2000 $
C_{wfr_tst}	Wafer Test Cost	0.1 $
FC_{wfr}	Wafer Test Coverage	80%
C_{burnin}	Die Burn-In and test Cost	0.2 $
FC_{die}	Die Test coverage	99%
C_{mod_tst}	Module/Chip test cost	0.3 $
FC_{mod}	Module/Chip test coverage	95%
Y_{MCMsub}	Yield of MCM substrate production	0.98
Y_{asmb}	Yield of assembly	0.97
Y_{3Dsub}	Yield of Wafer-Level 3D stacking	0.98
α, β, γ	Area merging factors	2,1,1
K_p	Rent's Coeff. (ASIC,DRAM)	2, 4
	Rent's Coeff. (μP, module)	7, 1.4
ρ	Rent's Exp (ASIC,DRAM)	0.35, 0.12
	Rent's Exp (μP, module)	0.4, 0.63
p_g	Global Metal Pitch	290 nm
p_i	Intermediate Wire Pitch	195 nm
p_l	Local Wire Pitch	152 nm
n_w	Number of interconnection layers (on-chip)	11
e_{rout}	Efficiency of routing tool	0.4
f_g	fanout of gates	3
P_p	Peripheral in-line pad pitch	60 μm
A_g	Gate Area	1μm
$A_{dramcell}$	DRAM Cell Area[56], [2]	0.05μm
n_{w_mcm}	Number of interconnection layers (MCM-D)	8
P_{w_mcm}	Interconnect pitch (MCM-D)	20 μm
l_{bw}	Length of bondwire	1 mm
L_{bw}	Inductance of bondwire	2 nH
C_{bw}	Capacitance of bondwire	0.3 pF
R_d	Min. sized Buffer Output Resistance	20.8 $k\Omega$
C_g	Min. sized Buffer Input Capacitance	0.14 fF
C_d	Min. sized Buffer Output Capacitance	0.22 fF
R_v	Resistance of through-hole via[55]	0.35 Ω
C_v	Capacitance of through-hole via[55]	5 fF
C_{pad}	Capacitance of the bond pad	2 pF
t_{layer}	Total Thickness of a Die	20 μm
t_{glue}	Thickness of the glue layer in 3-D stack	2 μm
t_{Cu}	Thickness of Cu metalization layers per die	12 μm
k_{Cu}	Thermal Conductivity of Cu	385 $\frac{W}{mK}$
k_{ILD}	Thermal Conductivity of Dielectric	0.19 $\frac{W}{mK}$
k_{glue}	Thermal Conductivity of Glue layer	0.25 $\frac{W}{mK}$
k_{Si}	Thermal Conductivity of Si	148 $\frac{W}{mK}$
k_{pkg}	Thermal Conductivity of Package Material	0.35 $\frac{W}{mK}$
k_{board}	Thermal Conductivity of PCB	20 $\frac{W}{mK}$

Based on the manufacturers data, the power density for the constituent submodules in our case studies can be estimated. The power density for a DRAM is estimated to be 0.02 W/mm^2 [52], and for a logic block, 0.12 W/mm^2 [53]. A CMOS IS has an average power density of 0.016 W/mm^2. The power dissipation of the MEMS sensor is assumed to be 50 mW, while for the Analog/RF block, it is assumed to be 500 mW. Since these particular mobile applications do not allow the use

TABLE III
ON-CHIP AND OFF-CHIP WIRE PARAMETERS [45]

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 IV
AREAS OF DIFFERENT BLOCKS IN MOBILE TERMINAL

Mobile Terminal			
	A_{int} in mm^2	A_{dev} in mm^2	A_{core} in mm^2
ASIC (500k gates)	0.750	0.528	0.750
μP (300k gates)	0.644	0.317	0.644
$128Mb$ DRAM	-	5.44	5.44
Image Sensor	-	24.5	24.5
Analog/RF	-	-	2

TABLE V
AREAS OF DIFFERENT BLOCKS IN WIRELESS SENSOR

Wireless Sensor			
	A_{int} in mm^2	A_{dev} in mm^2	A_{core} in mm^2
ASIC (500k gates)	0.750	0.528	0.750
μP (300k gates)	0.644	0.317	0.644
$2Mb$ DRAM	-	0.084	0.084
Sensor	-	-	1
Analog/RF	-	-	2

of a heat sink, thermal vias are used to increase the thermal conductivity of the package-board interface. For the stacked arrangement, the highest power dissipating block is placed closest to the board while the other blocks are stacked according to descending value of power dissipation.

In contemporary IC design, a major design consideration is to maintain operating temperature at a level which is not detrimental to the desired performance, reliability, and durability. In most ICs, circuits are designed for a worst-case temperature of 125 °C [44]. However, DRAM data retention depends heavily on operating temperature, and should usually be maintained below approximately 85 °C. In this analysis, we assume that the temperature of the outside of the package is maintained at 35 °C without any loss of generality. The methodology allows for the viability of any operating temperature to be investigated. Wire parasitics used for delay estimation in different implementation scenarios, as discussed in Section III-D, are given in Table III.

A. Monolithic SoC

The integration of mixed-signal systems in a single die requires a merging of several technologies such as logic, memory, and analog/RF, which results in increased process complexity and area. 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. For example, in a United Microelectronics Corporation 0.18-μm technology a 6T-SRAM cell has a footprint of about 4 μm^2 in a pure CMOS implementation, but is 5.6 μm^2 when merged with a logic process. In this case, the cell area increases by a factor of 1.4 as the result of merging processes [11]. The increased process steps for merging different technologies are mentioned in Table VI. If modules P, Q, and R are merged into a single chip, the integrated areas of the composite systems comprising two and three modules, respectively, are shown in [11] to be

$$A_{P\cup Q} = \alpha A_P + \beta A_Q \tag{35}$$

$$A_{P\cup Q\cup R} = \alpha A_P + \beta A_Q + \gamma A_R. \tag{36}$$

The total number of mask layers after merging is

$$N_{P\cup Q} = N_P + N_Q - N_{P\cap Q} \tag{37}$$

$$\begin{aligned} N_{P\cup Q\cup R} =& N_P + N_Q + N_R - N_{P\cap Q} - N_{P\cap R} - N_{Q\cap R} \\ &+ N_{P\cap Q\cap R}. \end{aligned} \tag{38}$$

The total cost for an SoC implementation is given in (39). Note that we assumed a MEMS-CMOS combined process for SoC implementation of the first system, the wireless sensor node.

Multiplying the total power dissipation by the series combination of the substrate and package thermal resistances, we can estimate the average chip temperature

$$\begin{aligned} &C_{\text{SoC}} \\ &= \left[\left(\frac{C_{\text{wafer}}}{Y_{\text{SoC}}N_{\text{die}}} + C_{\text{wafer\_test}}\right)\frac{1}{PF_{\text{w}}} + C_{\text{burn\_in}}\right]\frac{1}{PF_{\text{b}}} + C_{\text{pkg}} \end{aligned} \tag{39}$$

$$\begin{aligned} &C_{\text{SoP}} \\ &= \left\{\frac{\sum_{i=1}^{m} C_{\text{kgd}_i} + \frac{C_{\text{substrate}}}{Y_s} + C_{\text{assembly}} + C_{\text{rework}}}{Y_a} + C_{\text{test}}\right\} \\ &\quad\times \frac{1}{PF_{\text{SoP}}} + C_{\text{pkg}} \end{aligned} \tag{40}$$

$$\begin{aligned} &C_{\text{3D\_SiP}} \\ &= \left\{\frac{\sum_{i=1}^{m} C_{\text{kgd}_i} + \frac{C_{\text{substrate}}}{Y_s} + C_{\text{assembly}} + C_{\text{rework}}}{Y_a} + C_{\text{test}}\right\} \\ &\quad\times \frac{1}{PF_{\text{3D\_SiP}}} + C_{\text{pkg}} \end{aligned} \tag{41}$$

$$\begin{aligned} &C_{\text{3D\_W2W}} \\ &= \left\{\frac{\sum_{i=1}^{m} C_{\text{die}_i} + C_{\text{bonding}}}{Y_{a\_\text{3D\_W2W}}} + C_{\text{test}}\right\}\frac{1}{PF_{\text{W2W}}} + C_{\text{pkg}} \end{aligned} \tag{42}$$

$$\begin{aligned} &C_{\text{3D\_D2W}} \\ &= \left\{\frac{\sum_{i=1}^{m} C_{\text{kgd}_i} + C_{\text{bonding}}}{Y_{a\_\text{3D\_D2W}}} + C_{\text{test}}\right\}\frac{1}{PF_{\text{D2W}}} + C_{\text{pkg}}. \end{aligned} \tag{43}$$

TABLE VI
ADDED PROCESS COMPLEXITY (NUMBER OF MASK LEVEL) FOR SoC TECHNOLOGIES, BASED ON CMOS LOGIC [40]

Added Process	Logic	SRAM	Flash	DRAM	CMOS RF	FPGA	MEMS	FRAM	Chem. Sensors	Electro Optical
Logic	0									
SRAM	1-2	0								
Flash	4	3-4	0							
DRAM	4-5	3-4	7-9	0						
CMOS RF	3-5	5-9	6-9	6-10	0					
FPGA	2	2-4	4-6	3-7	5-7	0				
MEMS	2-10	3-12	6-14	6-15	5-15	4-12	0			
FRAM	4-5	3-4	7-9	2-3	7-10	6-7	9-15	0		
Chem. Sensors	2-6	3-7	6-10	6-11	5-11	4-8	4-16	6-11	0	
Electro Optical	5-8	6-9	9-12	9-13	8-12	7-10	7-18	9-13	7-14	0

B. 2D-SoP

In the 2D-SoP implementation, we assume that several chips such as DRAM, RF, Logic, and MEMS/IS are assembled as an MCM. Hence, the cost of implementing the MCM includes the total cost for each chip including testing cost, assembly cost, substrate cost, rework cost, and finally the MCM test and packaging costs.

The SoP can provide some reworking capability whereas SoC and wafer-level 3-D integration do not. If a single 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 (40), and the overall yield for assembling m number of chips, as described in [38], is

$$Y_{\text{SoP}} = Y_a \prod_{i=1}^{m} Y_{\text{chip}_i}^{(1-Fc)} \tag{44}$$

where Y_{chip_i} is the yield for ith chip.

The overall temperature is found by estimating the effective chip thermal resistance from $R_{\text{eff\_SoP}} = \sum_{i=1}^{n}(t_i/k_iA_i)$ and then multiplying the total power dissipation of all chips by the series combination of thermal resistances $R_{\text{eff\_SoP}}$, $R_{\text{pkg}}(\text{Package})$, and $R_{\text{subs}}(\text{substrate})$. The tradeoff analysis flow used for SoP is described in Fig. 7.

C. 3D-SiP

A 3D-SiP implementation is similar to the SoP package integration, except that the SiP implementation integrates dies vertically. The cost formula is the same, but the MCM substrate area is reduced, compared to the 2D-SoP implementation. The thermal profile is also found in a similar manner, using (24).

D. 3D-WLI

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

$$Y_{\text{3D}} = Y_{\text{2D}} \prod_{i=1}^{m-1} Y_{\text{2D}_i} Y_a \tag{45}$$

where Y_{2D} is the fabrication yield of the 2-D process, and Y_a is the yield loss due to the 3-D assembly process. The Y_a^{m-1} term in the equation takes into account the fact that integration of m layers of chips requires $m - 1$ silicon growth or wafer bonding procedures. In the case of D2W stacking, die yield after testing should be considered, so that KGDs are used. Hence, the overall yield figures for implementing our target system in 3D-W2W and 3D-D2W methods as described in [38] and [54] are as follows:

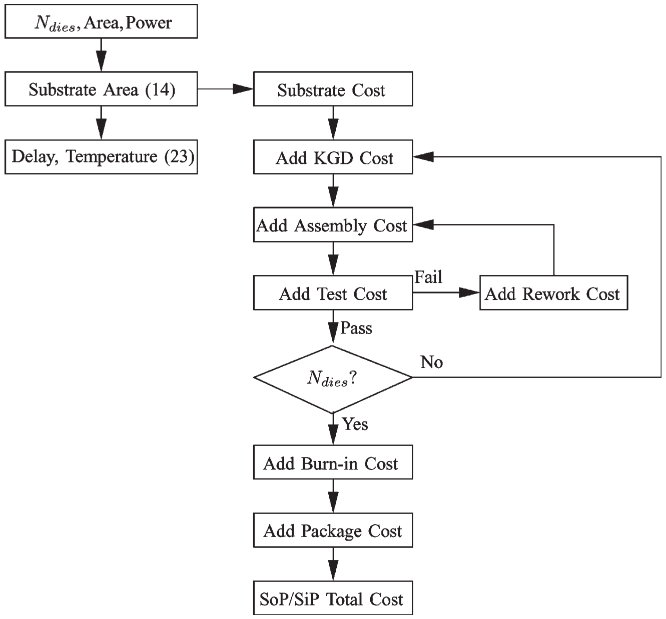

Fig. 7. SoP/SiP integration tradeoff analysis flow.

$$Y_{\text{3D\_w2w}} = Y_{\text{2D}} \prod_{i=1}^{m-1} Y_{\text{2D}_i} Y_a \tag{46}$$

$$Y_{\text{3D\_d2w}} = Y_{\text{2D}}^{(1-F_c)} \prod_{i=1}^{m-1} Y_{\text{2D}_i}^{(1-F_c)} Y_a. \tag{47}$$

The total cost for 3-D WLI is given in (42) and (43). Due to limitations in wafer-level processing, there is no possibility of reworking. In the case of D2W integration methodology, wafer-level test and burn-in costs for each die as well as the cost of testing the final module have been considered. However, in W2W technology, there is no die burn-in process to contribute to the cost.

It was assumed that standard test equipment can be used for testing of 3-D chips. If specialized equipment is to be used, their depreciation contribution to the cost has to be considered.

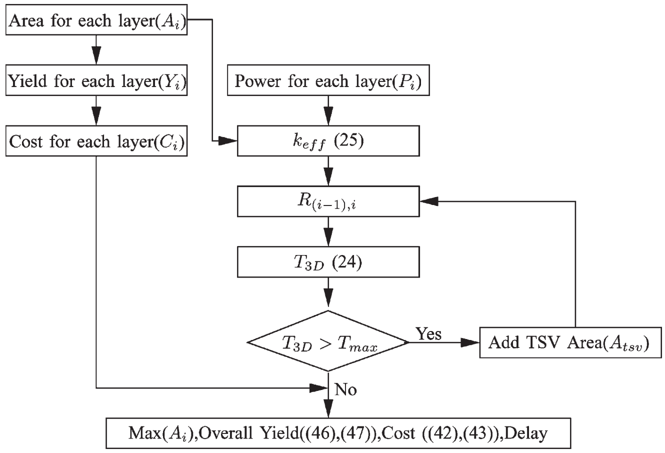
Area for each layer(A_i)
Yield for each layer(Y_i)
Power for each layer(P_i)
Cost for each layer(C_i)
k_{eff} (25)
$R_{(i-1),i}$
T_{3D} (24)
$T_{3D} > T_{max}$
Yes
Add TSV Area(A_{tsv})
No
Max(A_i),Overall Yield((46),(47)),Cost ((42),(43)),Delay

Fig. 8. Three-dimensional IC tradeoff analysis flow.

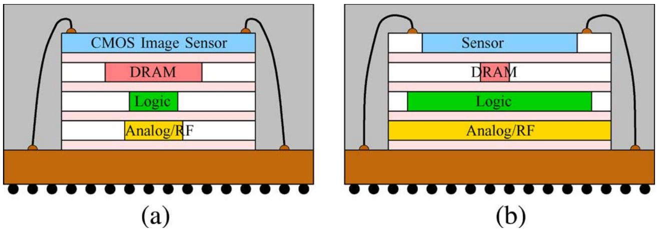

Fig. 9. Three-dimensional arrangement for mobile terminal and wireless sensor. Die sizes are normalized to the largest die in each stack, and are approximately to scale. (a) Mobile terminal. (b) Wireless sensor.

In a W2W integration methodology, all dies must be of the same size in order to alleviate manufacturing difficulties, particularly the precise alignment of wafers to make the vertical interconnections and facilitate dicing, whereas for D2W integration the dies can be of different sizes. The thermal profile is calculated using (24), and when the topmost layer's temperature exceeds the allowable limit, T-via insertion is carried out. The cross-sectional area of a state-of-the-art TSV is on the order of a few μm^2 [55], and inclusion of T-vias result in an appreciable area increase, and consequently, yield reduction. Thus, the chip manufacturing cost increases.

Shown in Fig. 8 is the tradeoff analysis flow used to estimate parameters for comparison. Also, Fig. 9 shows sizes of each die normalized to the largest die in the stack, and their relative position in the vertical direction.

V. Discussion

The results of the case studies are shown in Tables VII and VIII. The following implementation options have been considered in the tradeoff analyses: a single-chip planar SoC, and two-chip and four-chip arrangements of the different implementation options of 2D-SoP, 3D-SiP, 3D-W2W, and 3D-D2W integration. In the two-chip arrangement, Logic and DRAM blocks have been merged to form one chip while the other two blocks have been merged to form the second chip. In the four-chip arrangement, each individual block constitutes a chip. Each case is discussed in detail in the remainder of this section.

For the mobile terminal, 3-D integration provides the most compact design compared to 2-D planar techniques. Irrespective of whether two- or four-layer stacking is carried out, the difference in the final footprint is approximately 5%, due to the area dominance of the IS. Since the area of the mobile terminal is relatively large, the yield of the SoC implementation is rather low, while all other implementations result in higher yields. As can be expected, 3D-W2W integration inherently results in a lower yield in comparison with other 3-D implementations, since untested dies are stacked together. In spite of this though, the final cost of 3D-W2W is the lowest among all options. This is due to the lower test cost in comparison to 3D-D2W, and lower assembly cost in comparison to 2D-SoP and 3D-SiP. The low yield of the SoC solution means that it is relatively expensive when compared to all of the other implementations. A 3D-SiP implementation is slightly more expensive than a 2D-SoP implementation due to the higher assembly cost for a 3-D stack. Overall, the two-chip arrangement is a clear winner due to the lower assembly cost and higher yield in integrating two rather than four chips by stacking.

Interestingly, it appears that an SoC solution is the best choice for the wireless sensor node (Table VIII), because all other implementations show a lower performance, higher cost, and for the most part, a larger area. A 3D-SiP implementation leads to the most compact system, although costing the most. The reason for the comparatively large area in 3D-W2W and 3D-D2W implementations is that the relatively small individual blocks result in a higher power density, and require the addition of a high number of T-vias for thermal management. The total area increases as a consequence, and the cost and worst-case delay increase accordingly. In this case, T-vias occupy about 66% of the total area in the four-chip stack, and 49% in the two-chip stack. The wireless sensor node system is also quite small in comparison to the mobile terminal, and hence has a comparatively higher yield in all implementation choices. For all these reasons, an SoC solution may be the best option for such low area applications.

A comparatively elevated temperature can be seen in the block which is closest to the substrate (T_{top}) for 3-D implementations. As mentioned, this is the result of the increased power density caused by the relatively small area available for dissipation as opposed to the SoC implementation. This is the reason for the higher temperature, for example, in the four-chip arrangement as opposed to the two-chip arrangement in 3D-SiP technology. It should be borne in mind that the accuracy of the 1-D heat model for the particular implementation should be verified, and be replaced with a more accurate model wherever necessary.

One result that might seem counterintuitive is that 3D-WLI technologies result in a higher worst-case delay in some cases, in spite of the reduction in the average interconnect length. For example, the delay in 3D-WLI technologies is significantly higher than that in a 3D-SiP implementation for both case studies, and even than in a 2D-SoP implementation for the mobile terminal. The reason for the increased wire delay in 3D-WLI is due to the use of *package-intermediate-interconnects* [57], [58] in 2D-SoP and 3D-SiP implementations. For global signal transmission, three types of interconnects can be identified in general. These are on-chip wires on a top metal layer, off-chip printed circuit-board-type traces, and TSVs. The off-chip

TABLE VII
RESULTS OF COST AND PERFORMANCE ANALYSIS FOR MOBILE TERMINAL

	Single Chip (SoC)	Two Chips (Logic+DRAM,Analog/RF+IS)				Four Chips (Logic,DRAM,Analog/RF,IS)			
		2D-SoP	3D-SiP	3D-W2W	3D-D2W	2D-SoP	3D-SiP	3D-W2W	3D-D2W
Norm. Area	1.00	1.79	0.79	0.76	0.76	2.20	0.75	0.71	0.71
Yield$_{overall}$	0.56	0.98	0.98	0.88	0.98	0.98	0.98	0.84	0.94
Norm. Cost	1.00	0.54	0.66	0.39	0.47	0.71	0.74	0.54	0.76
Delay (ps)	311	203	171	277	277	213	170	271	271
$T_{top}(^oC)$	58	48	63	92	92	46	80	100	100

TABLE VIII
RESULTS OF COST AND PERFORMANCE ANALYSIS FOR WIRELESS SENSOR NODE. FOR 3D-W2W AND 3D-D2W INTEGRATION, THERMAL-VIAS HAVE TO BE INSERTED IN ORDER TO LIMIT THE TEMPERATURE INSIDE THE TOPMOST CHIP

	Single Chip (SoC)	Two Chips (Logic+DRAM,Analog/RF+Sensor)				Four Chips (Logic,DRAM,Analog/RF,Sensor)			
		2D-SoP	3D-SiP	3D-W2W	3D-D2W	2D-SoP	3D-SiP	3D-W2W	3D-D2W
Norm. Area	1.00	2.82	0.89	1.14	1.14	4.91	1.59	1.15	1.15
Yield$_{overall}$	0.95	0.98	0.98	0.96	0.98	0.98	0.98	0.92	0.94
Norm. Cost	1.00	2.21	3.01	1.30	2.52	4.60	4.48	1.25	4.01
Delay (ps)	132	170	151	155	155	187	158	156	156
$T_{top}(^oC)$	65	46	70	125	125	41	81	125	125

traces and TSVs exhibit fast transmission-linelike behavior whereas even the relatively wide global level on-chip lines are much more resistive, and exhibit diffusive (i.e., RC) behavior. Additionally, taking a signal off-chip and bringing a signal on-chip entail chip-to-package parasitics that include the pad capacitance, and bond wire or ball-grid solder ball. Finally, the layer-to-layer TSV connection includes a pad capacitance in the signal path.

Even taking into account the off-chip drivers and chip-to-package parasitics, off-chip wires are much faster than on-chip wires for transmitting a signal for the length of a die edge, for a relatively large die. This is because the fast off-chip traces more than make-up for the chip-to-package parasitics by outperforming the RC lines. In 2D-SoP and 3D-SiP, the opportunity exists to take advantage of this phenomenon by running wires off-chip and bypassing long chip-edge to chip-edge length RC lines. This is shown in Fig. 6(b). The actual saving will of course depend on the specific layout, but in [57], for example, this technique of avoiding long on-chip wires by running them off-chip to realize Package-Intermediate Interconnects, is reported to yield a saving of up to 40%, even considering the chip-to-package parasitics.

The layout and die sizes are a critical factor in determining the relative speed in different implementation technologies. If the layout permits communicating blocks to be placed vertically close to each other, for example, vertical integration does provide an excellent opportunity to substantially reduce the communication delay. For the specific cases considered in this paper, the quantitative results based on accurate parasitics show that signal transmission from the corner of one chip to the diagonally opposed corner of another (A to B in Fig. 6) is faster in the 2-D SoP- and 3-D SiP-type implementations due to the outperformance of the long on-chip wires.

Another contributing factor is the increased temperatures in the higher level layers, which has an adverse effect on device and interconnect performance, although this is of less significance. This difference is on the order of tens of picoseconds, as compared with previous values in [14], which did not include this refinement.

VI. CONCLUSION AND FUTURE WORK

In this paper, we presented detailed and quantitative area, yield, and cost models, and simple yet useful thermal models, as well as performance metrics for evaluating 3-D integration options. We combined these models in a cohesive cost and performance tradeoff analysis methodology which is suitable for early analysis and design space explorations of future nanoscale electronic systems. Using example contemporary mixed-signal systems, we demonstrated the use of the proposed methodology and models in analyzing the impact of different implementations, and conclude that the implementation strategy must be carefully selected depending on the circuit complexity and architecture, as otherwise 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 it is expected that the models and methodology presented in this paper will be a useful aid in this choice. Topics earmarked for future work include the addition of more technology options as well as the use of a more sophisticated heat flow model.

REFERENCES

[1] F. Catthoor, N. D. Dutt, and C. E. Kozyrakis, “How to solve the current memory access and data transfer bottlenecks: At the processor architecture or at the compiler level,” in *Proc. Conf. Des., Autom. Test Eur.*, New York, 2000, pp. 426–435.

[2] *The International Technology Roadmap for Semiconductors (ITRS).* [Online]. Available: http://www.itrs.net/Links/2003ITRS/Home2003.htm

[3] E. Beyne, “3D interconnection and packaging: Impending reality or still a dream?” in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2004, vol. 1, pp. 138–139.

[4] M. Bamal, S. List, M. Stucchi, A. Verhulst, M. Van Hove, R. Cartuyvels, G. Beyer, and K. Maex, “Performance comparison of interconnect technology and architecture options for deep submicron technology nodes,” in *Proc. Int. Interconnect Technol. Conf.*, 2006, pp. 202–204.

[5] S. Kim, C. Liu, L. Xue, and S. Tiwari, “Crosstalk reduction in mixed-signal 3-D integrated circuits with interdevice layer ground planes,” *IEEE Trans. Electron Devices*, vol. 52, no. 7, pp. 1459–1467, Jul. 2005.

[6] 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.

[7] B. Goplen and S. S. Sapatnekar, "Placement of thermal vias in 3-D ICs using various thermal objectives," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 25, no. 4, pp. 692–709, Apr. 2006.

[8] Z. Li, X. Hong, Q. Zhou, S. Zeng, J. Bian, W. Yu, H. H. Yang, V. Pitchumani, and C.-K. Cheng, "Efficient thermal via planning approach and its application in 3-D floorplanning," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 26, no. 4, pp. 645–658, Apr. 2007.

[9] J. M. Koo, S. Im, L. Jiang, and K. E. Goodson, "Integrated microchannel cooling for three-dimensional circuit architectures," *J. Heat Transf.*, vol. 127, no. 1, pp. 49–58, Jan. 2005.

[10] B. Dang, M. Bakir, and J. Meindl, "Integrated thermal-fluidic i/o interconnects for an on-chip microchannel heat sink," *IEEE Electron Device Lett.*, vol. 27, no. 2, pp. 117–119, Feb. 2006.

[11] M. Shen, L.-R. Zheng and H. Tenhunen, "Cost and performance analysis for mixed-signal system implementation: System-on-chip or system-on-package?" *IEEE Trans. Electron. Packag. Manuf.*, vol. 25, no. 4, pp. 262–272, Oct. 2002.

[12] C. C. Liu, J.-H. Chen, R. Manohar, and S. Tiwari, "Mapping system-on-chip designs from 2-D to 3-D ICs," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2005, pp. 2939–2942.

[13] P. Mercier, S. R. Singh, K. Iniewski, B. Moore, and P. O'Shea, "Yield and cost modeling for 3D chip stack technologies," in *Proc. IEEE Custom Integr. Circuits Conf.*, Sep. 2006, pp. 357–360.

[14] R. Weerasekera, L.-R. Zheng, D. Pamunuwa, and H. Tenhunen, "Extending systems-on-chip to the third dimension: Performance, cost and technological tradeoffs," in *Proc. IEEE/ACM Int. Conf. Computer-Aided Design*, 2007, pp. 212–219.

[15] R. K. Ulrich and W. D. Brown, Eds., *Advanced Electronic Packaging*, 2nd ed., ser. IEEE Press Series on Microelectronic Systems. New York: Wiley-Interscience, Sep. 2005.

[16] M. Dreiza, A. Yoshida, J. Micksch, and L. Smith, "Stacked package-on-package design guidelines," *Chip Scale Rev.*, no. 7, Jul. 2005. [Online]. Available: http://www.chipscalereview.com/issues/0705/article.php?type=feature&article=f3

[17] 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," *Jpn. J. Appl. Phys.*, vol. 45, no. 4B, pp. 3030–3035, 2006.

[18] E. Culurciello and A. G. Andreou, "Capacitive inter-chip data and power transfer for 3-D VLSI," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 53, no. 12, pp. 1348–1352, Dec. 2006.

[19] A. Fazzi, L. Magagni, M. Mirandola, R. Canegallo, S. Schmitz, and R. Guerrieri, "A 0.14 mW/Gbps high-density capacitive interface for 3D system integration," in *Proc. IEEE Custom Integr. Circuits Conf.*, 2005, pp. 101–104.

[20] J. Xu, J. Wilson, S. Mick, L. Luo, and P. Franzon, "2.8 Gb/s inductively coupled interconnect for 3D ICS," in *Proc. Symp. VLSI Circuits, Dig. Tech. Papers*, 2005, pp. 352–355.

[21] A. Iwata, M. Sasaki, T. Kikkawa, S. Kameda, H. Ando, K. Kimoto, D. Arizono, and H. Sunami, "A 3D integration scheme utilizing wireless interconnections for implementing hyper brains," in *Proc. IEEE Int. Solid-State Circuits Conf.*, 2005, pp. 262–597.

[22] H. B. Backoglu, *Circuits, Interconnections and Packaging for VLSI*. Reading, MA: Addison-Wesley, 1990.

[23] B. Geuskens and K. Rose, *Modeling Microprocessor Performance*. Norwell, MA: Kluwer, 1998.

[24] D. Liu and C. Svensson, "Power consumption estimation in CMOS VLSI chips," *IEEE J. Solid-State Circuits*, vol. 29, no. 6, pp. 663–670, Jun. 1994.

[25] V. Garg, D. Stogner, C. Ulmer, D. Schimmel, C. Dislis, S. Yalamanchili, and D. Wills, "Early analysis of cost/performance trade-offs in MCM systems," *IEEE Trans. Compon., Packag., Manuf. Technol. B, Adv. Packag.*, vol. 20, no. 3, pp. 308–319, Aug. 1997.

[26] S. Takahashi, M. Edahiro, and Y. Hayashi, "Interconnect design strategy: Structures, repeaters and materials with strategic system performance analysis (S^2PAL) model," *IEEE Trans. Electron Devices*, vol. 48, no. 2, pp. 239–251, Feb. 2001.

[27] R. Venkatesan, J. A. Davis, K. A. Bowman, and J. D. Meindl, "Optimal n-tier multilevel interconnect architectures for gigascale integration (GSI)," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 9, no. 6, pp. 899–912, Dec. 2001.

[28] G. Sai-Halasz, "Performance trends in high-end processors," *Proc. IEEE*, vol. 83, no. 1, pp. 20–36, Jan. 1995.

[29] Q. Chen, J. A. Davis, P. Zarkesh-Ha, and J. Meindl, "A compact physical via blockage model," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 8, no. 6, pp. 689–692, Dec. 2000.

[30] P. A. Sandborn and H. Moreno, *Conceptual Design of Multichip Modules and Systems*. Norwell, MA: Kluwer, 1994.

[31] W. Donath, "Placement and average interconnection lengths of computer logic," *IEEE Trans. Circuits Syst.*, vol. CAS-26, no. 4, pp. 272–277, Apr. 1979.

[32] M. Feuer, "Connectivity of random logic," *IEEE Trans. Comput.*, vol. C-31, no. 1, pp. 29–33, Jan. 1982.

[33] W. R. Heller, C. G. Hsi, and W. F. Mikhaill, "Wirability—Designing wiring space for chips and chip packages," *IEEE Des. Test Mag.*, vol. 1, no. 3, pp. 43–51, Aug. 1984.

[34] J. A. Davis, V. K. De, and J. D. Meindl, "A stochastic wire-length distribution for gigascale integration (GSI). I. Derivation and validation," *IEEE Trans. Electron Devices*, vol. 45, no. 3, pp. 580–589, Mar. 1998.

[35] P. Sandborn, M. Abadir, and C. Murphy, "The tradeoff between peripheral and area array bonding of components in multichip modules," *IEEE Trans. Compon., Packag., Manuf. Technol., Part A*, vol. 17, no. 2, pp. 249–256, Jun. 1994.

[36] P. Zarkesh-Ha, J. A. Davis, and J. D. Meindl, "Prediction of net-length distribution for global interconnects in a heterogeneous system-on-a-chip," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 8, no. 6, pp. 649–659, Dec. 2000.

[37] A. George, J. Krusius, and R. Granitz, "Packaging alternatives to large silicon chips: Tiled silicon on MCM and PWB substrates," *IEEE Trans. Compon., Packag., Manuf. Technol., B: Adv. Packag.*, vol. 19, no. 4, pp. 699–708, Nov. 1996.

[38] Y. Deng and W. P. Maly, "2.5-dimensional VLSI system integration," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 13, no. 6, pp. 668–677, Jun. 2005.

[39] D. Ragan, P. Sandborn, and P. Stoaks, "A detailed cost model for concurrent use with hardware/software co-design," in *Proc. 39th IEEE/ACM Des. Autom. Conf.*, 2002, pp. 269–274.

[40] *The International Technology Roadmap for Semiconductors (ITRS)*, 1999. [Online]. Available: http://www.itrs.net

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

[42] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. Stan, "HotSpot: A compact thermal modeling methodology for early-stage VLSI design," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 14, no. 5, pp. 501–513, May 2006.

[43] G. Luca, B. Agrawal, N. Srivastava, S.-C. Lin, T. Sherwood, and K. Banerjee, "A thermally-aware performance analysis of vertically integrated (3-D) processor-memory hierarchy," in *Proc. Des. Autom. Conf.*, 2006, pp. 991–996.

[44] H. Hua, C. Mineo, K. Schoenfliess, A. Sule, S. Melamed, R. Jenkal, and W. Davis, "Exploring compromises among timing, power and temperature in three-dimensional integrated circuits," in *Proc. Des. Autom. Conf.*, 2006, pp. 997–1002.

[45] L.-R. Zheng, D. Pamunuwa, and H. Tenhunen, "Accurate a priori signal integrity estimation using a multilevel dynamic interconnect model for deep submicron VLSI design," in *Proc. Eur. Solid-State Circuits Conf.*, 2000, pp. 352–355.

[46] D. Pamunuwa, L.-R. Zheng, and H. Tenhunen, "Maximizing throughput over parallel wire structures in the deep submicrometer regime," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 11, no. 2, pp. 224–243, Apr. 2003.

[47] R. Weerasekera, D. Pamunuwa, L. Zheng, and H. Tenhunen, "Minimal-power, delay-balanced smart repeaters for global interconnects in the nanometer regime," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 16, no. 5, pp. 589–593, May 2008.

[48] J. Ku and Y. Ismail, "Thermal-aware methodology for repeater insertion in low-power VLSI circuits," in *Proc. Int. Symp. Low Power Electron. Des.*, 2007, pp. 86–91.

[49] Y. Ku and J. C. Ismail, "On the scaling of temperature-dependent effects," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 26, no. 10, pp. 1882–1888, Oct. 2007.

[50] J. E. Sergent and A. Krum, Eds., *Thermal Management Handbook for Electronic Assemblies*. New York: McGraw-Hill, 1998.

[51] Micron Technol., Inc., *Micron CMOS Image Sensor Part Catalog*, Boise, ID, Mar. 2007. [Online]. Available: http://www.micron.com

[52] Micron Technol., Inc., *Micron 128 MB SDRAM Part Catalog*, Boise, ID, 2007. [Online]. Avaialble: http://www.micron.com

[53] ARM Holdings PLC, *ARM Cortex-A8 Processor Product Brief*, Cambridge, NJ, Mar. 2007. [Online]. Avaialble: http://www.arm.com

[54] E. Beyne, "The rise of the 3rd dimension for system intergration," in *Proc. Int. Technol. Conf.*, 2006, pp. 1–5.

[55] R. Patti, "Three-dimensional integrated circuits and the future of system-on-chip designs," *Proc. IEEE*, vol. 94, no. 6, pp. 1214–1224, Jun. 2006.

[56] R. Waiser Ed., *Nanoelectronics and Information Technology: Advanced Electronic Materials and Novel Devices*. Hoboken, NJ: Wiley-VCH, Sep. 2004.

[57] S. Afonso, L. Schaper, J. Parkerson, W. Brown, S. Ang, and H. Naseem, "Modeling and electrical analysis of seamless high off-chip connectivity (SHOCC) interconnects," *IEEE Trans. Adv. Packag.*, vol. 22, no. 3, pp. 309–320, Aug. 1999.

[58] P. Mehrotra, V. Rao, T. M. Conte, and P. D. Franzon, "Optimal chip-package codesign for high-performance DSP," *IEEE Trans. Adv. Packag.*, vol. 28, no. 2, pp. 288–297, May 2005.

Roshan Weerasekera (S'01–M'04) received the B.Sc.Eng. degree from the University of Peradeniya, Peradeniya, Sri Lanka, in 1998 and the M.Sc. and Ph.D. degrees in electronic system design from the Royal Institute of Technology, Stockholm, Sweden, in 2002 and 2008, respectively.

During October 2001 to December 2003, he worked as a Lecturer with the Department of Electrical and Electronic Engineering, University of Peradeniya, Sri Lanka. Since February 2008, he was appointed a research associate in Lancaster University, Lancaster, U.K., to lead a 3-D Flash memory integration project (ELITE) under a European Union FP7 research grant. His research interests include design, analysis and modeling of packaging and on-chip interconnections, 3-D integration and packaging-related issues, and architectures for silicon nanoelectronics.

Dinesh Pamunuwa (M'04) received the B.Sc. degree (with honors) in engineering from the University of Peradeniya, Peradeniya, Sri Lanka, in 1997 and the Ph.D. degree in electronic system design from the Royal Institute of Technology, Stockholm, Sweden, in 2003.

In 2002, he was with Cadence Berkeley Laboratories, Berkeley, CA. He has worked extensively on interconnect design and signal integrity issues and methodologies for electronic system design, and is the author or coauthor of many papers in this area. In the past, he has been the Cofounder of an electronics and software consultancy company based in Sweden and Sri Lanka. He is currently a Faculty Member of the Department of Engineering, Lancaster University, Lancaster, U.K. His research interests include on-chip communication issues and architectures, system level modeling of interconnects, devices, and circuits, and architectures for terascale integration in the nanometer regime.

Li-Rong Zheng (M'01) received the Tech.D. degree in electronic system design from the Royal Institute of Technology (KTH), Stockholm, Sweden, in 2001.

He became an Associate Professor in electronics system design with KTH in 2003 and a Full Professor in media electronics in 2006. Since then, he was with KTH as a Research Fellow and Project Leader in the Laboratory of Electronics and Computer Systems. His research experience and interest includes electronic circuits and systems for ambient intelligence and media applications, radio and mixed-signal integrated circuits, wireless system-in-package, and signal integrity issues in electronic systems. He has authored and coauthored over 200 international reviewed publications, covering areas from electronic devices and thin-film technologies, very large scale integration circuit and system design, to electronics system integration and wireless sensors.

Hannu Tenhunen (S'83–M'90) received the Diploma Engineer degree in electrical engineering and computer sciences from Helsinki University of Technology, Helsinki, Finland, in 1982 and Ph.D. degree in microelectronics from Cornell University, Ithaca, NY, in 1986.

During 1978–1982, he was with Electron Physics Laboratory, Helsinki University of Technology. From 1983 to 1985, he was with Cornell University as a Fullbright Scholar. From September 1985, he was with the Signal Processing Laboratory, Tampere University of Technology, Tampere, Finland, as an Associate Professor. He was also a Coordinator of the National Microelectronics Program of Finland during 1987–1991. Since January 1992, he has been with the Royal Institute of Technology, Stockholm, Sweden, where he is a Chair Professor in electronic system design. His current research interests are very large scale integration (VLSI) circuits and systems for wireless and broadband communication, and related design methodologies and prototyping techniques. He has contributed significantly in microelectronics research and education in Europe. He has been actively involved in several EU programs on VLSI/system-on-a-chip design and education either as a coordinator or as a contributor. He has made over 500 presentations and publications on IC technologies and VLSI systems worldwide, and has over 16 patents pending or granted.

Dr. Tenhunen was awarded honorary doctor degree (Dr.h.c.) from Tallinn Technical University, Estonia, in 2003 for advancement of advanced education and research in microelectronics area. In 2003, he was also nominated to an honorary Turku Centre for Computer Science (TUCS) Research Fellow by the University of Turku and Åbo Akademien University in Finland. From 2001 to 2005, he was the Dean of IT-University, Stockholm, Sweden, responsible for faculty and educational program development for IT-University. Since 2006, he was the Director of TUCS, Finland.