INTEGRATED CIRCUITS AND SYSTEMS

Abbas Sheibanyrad
Frédéric Pétrot
Axel Jantsch *Editors*

3D Integration for NoC-based SoC Architectures

Bahan dengan hak cipta

Editors
Abbas Sheibanyrad
TIMA Laboratory
46, Avenue Felix Viallet
38000 Grenoble
France
abbas.sheibanyrad@imag.fr

Axel Jantsch
Royal Institute of Technology
Forum 120
SE-16440 Kista
Sweden
axel@kth.se

Frédéric Pétrot
TIMA Laboratory
46, Avenue Felix Viallet
38000 Grenoble
France
frederic.petrot@imag.fr

ISSN 1558-9412
ISBN 978-1-4419-7617-8 e-ISBN 978-1-4419-7618-5
DOI 10.1007/978-1-4419-7618-5
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number:

© Springer Science+Business Media, LLC 2011
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Bahan dengan hak cipta

Contents

I 3DI Promises and Challenges ... 1

1 Three-Dimensional Integration of Integrated Circuits—an Introduction ... 3
Chuan Seng Tan

2 The Promises and Limitations of 3-D Integration ... 27
Axel Jantsch, Matthew Grange and Dinesh Pamunuwa

II Technology and Circuit Design ... 45

3 Testing 3D Stacked ICs Containing Through-Silicon Vias ... 47
Erik Jan Marinissen

4 Design and Computer Aided Design of 3DIC ... 75
Paul D. Franzon, W. Rhett Davis and Thor Thorolfsson

5 Physical Analysis of NoC Topologies for 3-D Integrated Systems ... 89
Vasilis F. Pavlidis and Eby G. Friedman

6 Three-Dimensional Networks-on-Chip: Performance Evaluation ... 115
Brett Stanley Feero and Partha Pratim Pande

III System and Architecture Design ... 147

7 Asynchronous 3D-NoCs Making Use of Serialized Vertical Links ... 149
Abbas Sheibanyrad and Frédéric Pétrot

8 Design of Application-Specific 3D Networks-on-Chip Architectures ... 167
Shan Yan and Bill Lin

Bahan dengan hak cipta

9 3D Network on Chip Topology Synthesis: Designing Custom Topologies for Chip Stacks 193
Ciprian Seiculescu, Srinivasan Murali, Luca Benini and Giovanni De Micheli

10 3-D NoC on Inductive Wireless Interconnect 225
Hiroki Matsutani, Michihiro Koibuchi, Tadahiro Kuroda and Hideharu Amano

11 Influence of Stacked 3D Memory/Cache Architectures on GPUs 249
Ahmed Al Maashri, Guangyu Sun, Xiangyu Dong, Yuan Xie and Narayanan Vijaykrishnan

Index 273

Bahan dengan hak cipta

Contributors

Ahmed Al Maashri The Pennsylvania State University, University Park, PA, USA

Hideharu Amano Keio University, Yokohama, Japan

Luca Benini The University of Bologna, Bologna, Italy

W. Rhett Davis The North Carolina State University, Raleigh, NC, USA

Giovanni De Micheli EPFL, Lausanne, Switzerland

Xiangyu Dong The Pennsylvania State University, University Park, PA, USA

Brett Stanley Feero ARM Inc., Austin, TX, USA

Paul D. Franzon The North Carolina State University, Raleigh, NC, USA

Eby G. Friedman University of Rochester, Rochester, NY, USA

Matthew Grange Lancaster University, Lancaster, UK

Axel Jantsch Royal Institute of Technology, Sweden

Michihiro Koibuchi Japanese National Institute of Informatics, Tokyo, Japan

Tadahiro Kuroda Keio University, Yokohama, Japan

Bill Lin University of California, San Diego, CA, USA

Erik Jan Marinissen IMEC, Leuven, Belgium

Hiroki Matsutani The University of Tokyo, Tokyo, Japan

Srinivasan Murali EPFL, Lausanne, Switzerland

Dinesh Pamunuwa Lancaster University, Lancaster, UK

Partha Pratim Pande Washington State University, Pullman, WA, USA

Vasilis F. Pavlidis EPFL, Lausanne, Switzerland

Frédéric Pétrot TIMA Laboratory, Grenoble, France

Bahan dengan hak cipta

Ciprian Seiculescu EPFL, Lausanne, Switzerland

Abbas Sheibanyrad TIMA Laboratory, Grenoble, France

Guangyu Sun The Pennsylvania State University, University Park, PA, USA

Chuan Seng Tan Nanyang Technological University, Singapore

Thorlindur Thorolfsson The North Carolina State University, Raleigh, NC, USA

Narayanan Vijaykrishnan The Pennsylvania State University, University Park, PA, USA

Yuan Xie The Pennsylvania State University, University Park, PA, USA

Shan Yan University of California, San Diego, CA, USA

Bahan dengan hak cipta

Part I
3DI Promises and Challenges

Bahan dengan hak cipta

Chapter 1
Three-Dimensional Integration of Integrated Circuits—an Introduction

Chuan Seng Tan

1.1 Background and Introduction

Imagine a situation where you need to travel between your home and office every day. You need to put up with time lost during commute as well as paying for the fuel. One possible solution is to have your home in another floor of your office building. In this way, all you need to do is to go up and down between floors and you can save time and cost. This simple idea can similarly be applied to boost the overall performance in future integrated circuits.

For the past 40 over years, higher computing power was achieved primarily through commensurate performance enhancement of transistors as a result of continuously scaling down the device dimensions in a harmonious manner. This has resulted in a steady doubling of device density from one technology node to another as famously described by Moore's Law. Improvement in transistor switching speed and count are two of the most direct contributors to the historical performance growth in integrated circuits (particularly in silicon-based digital CMOS). This scaling approach has been so effective in many aspects (performance and cost) that integrated circuits have essentially remained a planar platform throughout this period of rigorous scaling. As performance enhancement through geometrical scaling becomes more challenging and demand for higher functionality increases, there is tremendous interest and potential to explore the third dimension, i.e., the vertical dimension of the integrated circuits. This was rightly envisioned by Richard Fenyman, physicist and Nobel Laureate, when he delivered a talk on "Computing Machines in the Future" in Japan in 1985. His original text reads: "Another direction of improvement (in computing power) is to make physical machines three dimensional instead of all on a surface of a chip. That can be done in stages instead of all at once—you can have several layers and then add many more layers as time

C. S. Tan (✉)
School of Electrical and Electronic Engineering, Nanyang Technological University,
50 Nanyang Avenue, 639798 Singapore
Tel.: +65-67905636
e-mail: tancs@ntu.edu.sg

A. Sheibanyrad et al. (eds.), *3D Integration for NoC-based SoC Architectures*,
Integrated Circuits and Systems, DOI 10.1007/978-1-4419-7618-5_1,

© Springer Science+Business Media, LLC 2011

Bahan dengan hak cipta

goes on" [1]. The need for 3D integration has become clear going forward and it was reiterated by Dr. Chang-Gyu Hwang, President and CEO of Samsung Electronics' Semiconductor Business, when he delivered a keynote speech at the 2006 International Electron Devices Meeting (IEDM) in San Francisco entitled "New Paradigms in the Silicon Industry" [2]. Some important points of his speech are quoted: "The approaching era of electronics technology advancement—the Fusion Era—will be massive in scope, encompassing the fields of information technology (IT), bio-technology (BT), and nano-technology (NT) and will create boundless opportunities for new growth to the semiconductor industry. The core element needed to usher in the new age will be a complex integration of different types of devices such as memory, logic, sensor, processor and software, together with new materials, and advanced die stack technologies, all based on 3-D silicon technology."

1.2 Motivations and Drivers

This section examines the role of 3D integration in ensuring that performance growth enjoyed by the semiconductor industry as a direct result of geometrical scaling, coupled with the introduction of performance boosters in more recent nodes, as predicted by Moore's Law can continue in the future. Scaling alone has met with diminishing return due to fundamental and economics scaling barriers (non-commensurate scaling). 3D integration explores the third dimension of IC integration and offers new dimension for performance growth. 3D integration also enables integration of disparate chips in a more compact form factor and it is touted by many as an attractive method for system miniaturization and functional diversification commonly known as heterogeneous integration.

1.2.1 Sustainable IC Performance Growth

Beginning with the invention of the first integrated circuit in 1958 by Kilby and Noyce, the world has witnessed sustainable performance growth in IC. The trend is best exemplified by the exponential improvement in computing power (measured in million instructions per second, MISP) in Intel's micro-processors over the last 40 years as shown in Fig. 1.1 [3].

This continuous growth is a result of the ability to scale silicon transistor to smaller dimension in every new technology nodes. The growth continued, instead of hitting a plateau, in more recent nodes thanks to the addition of performance boosters (e.g., strained-Si, high-κ and metal gate, etc) on top of conventional geometrical scaling. Scaling doubles the number of transistors on IC (famously described by "Moore's Law") in every generation and allows us to integrate more functions on IC and to increase its computing power. We are now in the Giga-scale integration era featured by billions of transistors, GHz operating frequency, etc. Going forward to Tera-scale integration, however, there are a number of imminent show-stoppers (described in the next section) that pose serious threat to continuous performance

Bahan dengan hak cipta

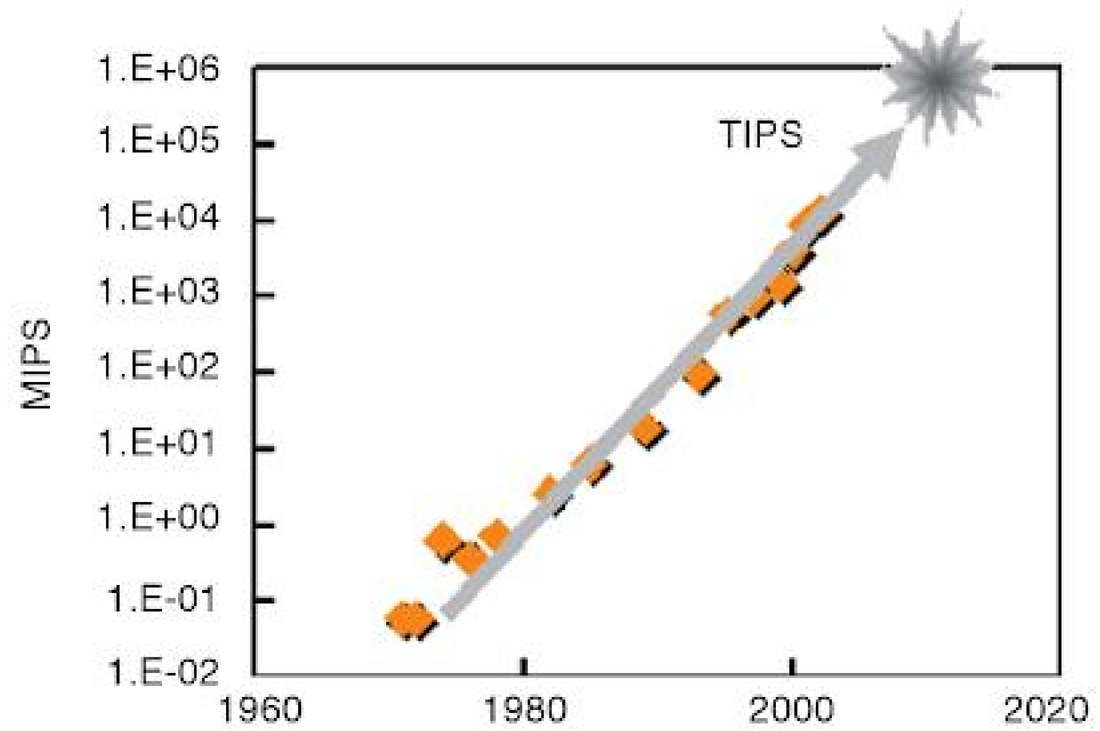

Fig. 1.1 The evolution of computing performance. (Source: Intel)

enhancement in IC and a new paradigm shift in IC technology and architecture is needed to sustain the historical growth. It is widely recognized that the growth can be sustained if one utilizes the vertical (i.e., the third) dimension of IC to build three-dimensional IC, a departure from today's planar IC as projected in Fig. 1.2.

Three-dimensional integrated circuits (3D IC) refers to a stack consists of multiple ultra-thin layers of IC that are vertically bonded and interconnected with through silicon via (TSV) as shown in Fig. 1.3. In 3D implementation, each block can be fabricated and optimized using their respective technologies and assembled to form a vertical stack. 3D stacking of ultra-thin ICs is identified as an inevitable solution for future performance enhancement, system miniaturization, and functional diversification.

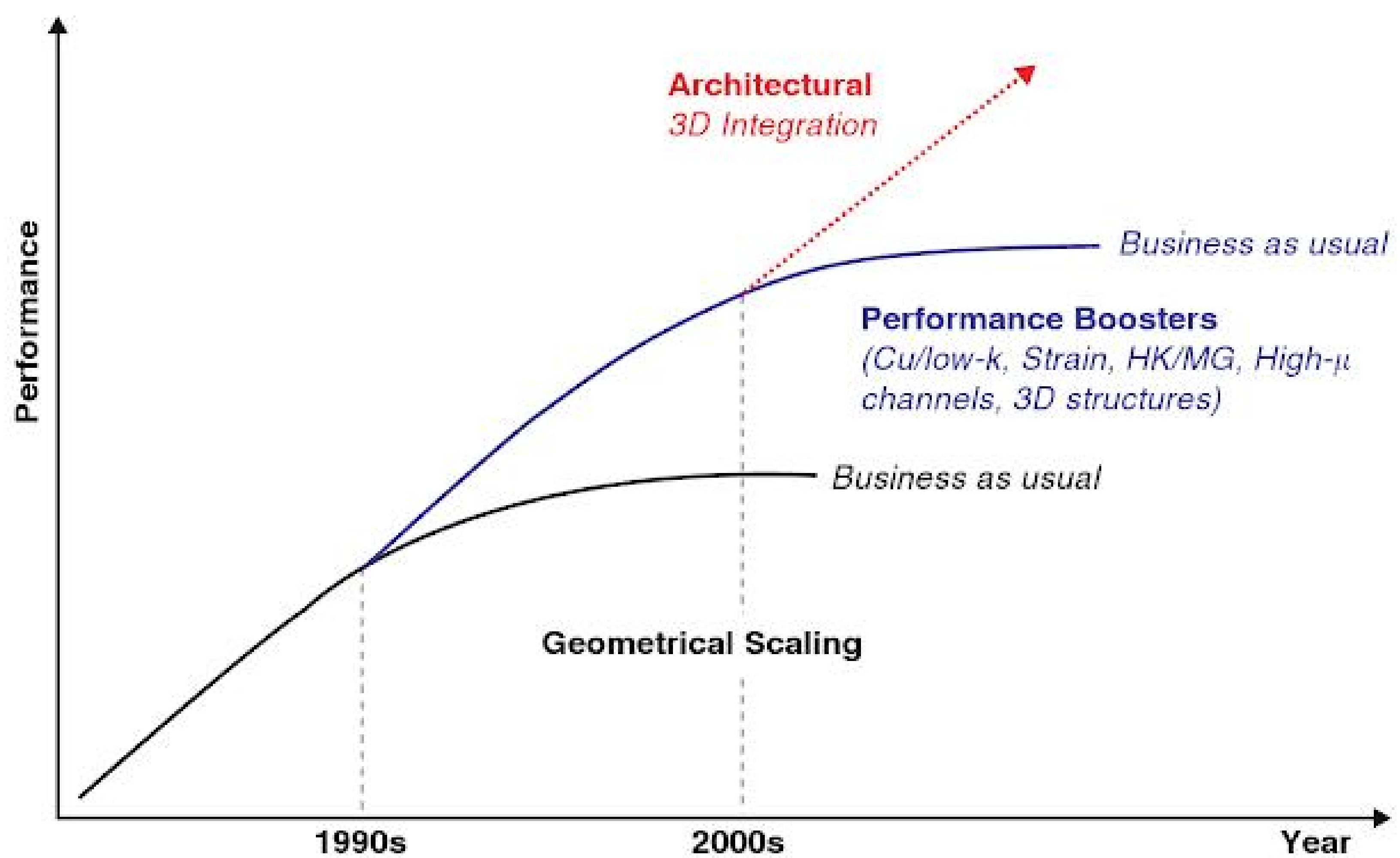

Fig. 1.2 Historical IC performance growth can be sustained with a new paradigm shift to 3-D integration

Bahan dengan hak cipta

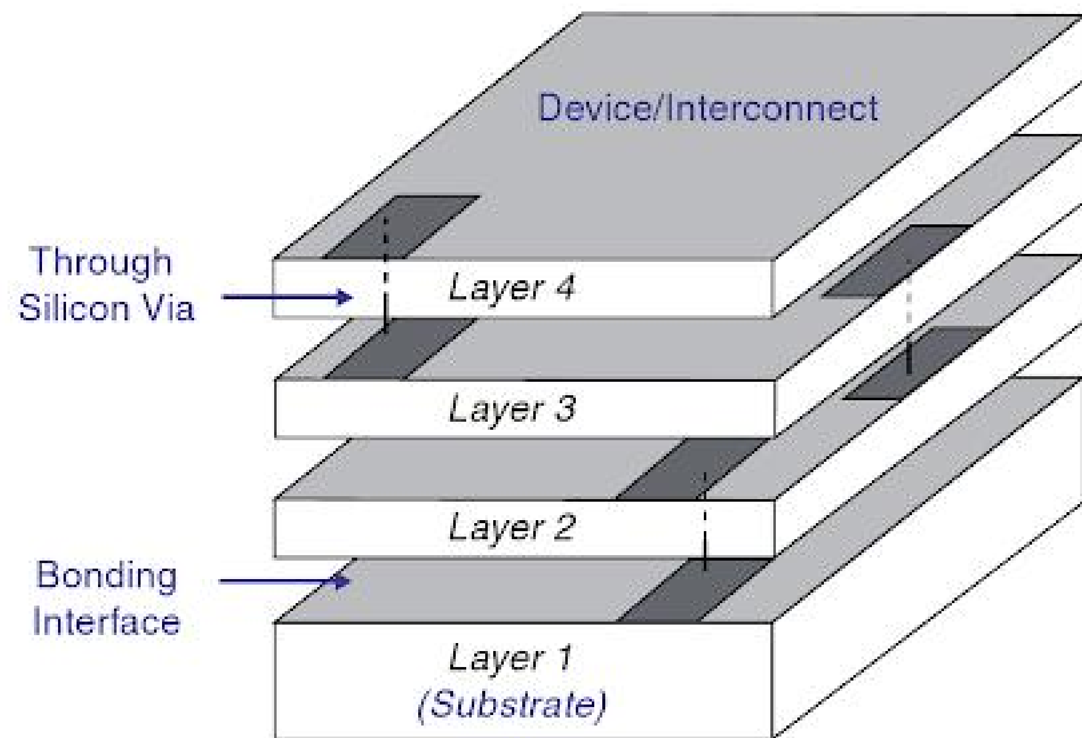

Fig. 1.3 A conceptual representation of 3D IC

1.2.2 *Show-Stoppers and 3-D Integration as a Remedy*

1.2.2.1 Transistor Scaling Barriers

There are at least two barriers that will slow down or impede further geometrical scaling. The first one relates to the fundamental properties of transistor in extremely scaled devices. Experimental and modeling data suggest that performance improvement in devices is no longer commensurate with ideal scaling in the past due to high leakage and parasitic hence they consume more power. This is shown in Fig. 1.4 by Khakifirooz and Antoniadis [4]. The intrinsic delay of n-MOS is shown to increase beyond the 45 nm despite continuous down scaling of the transistor pitch.

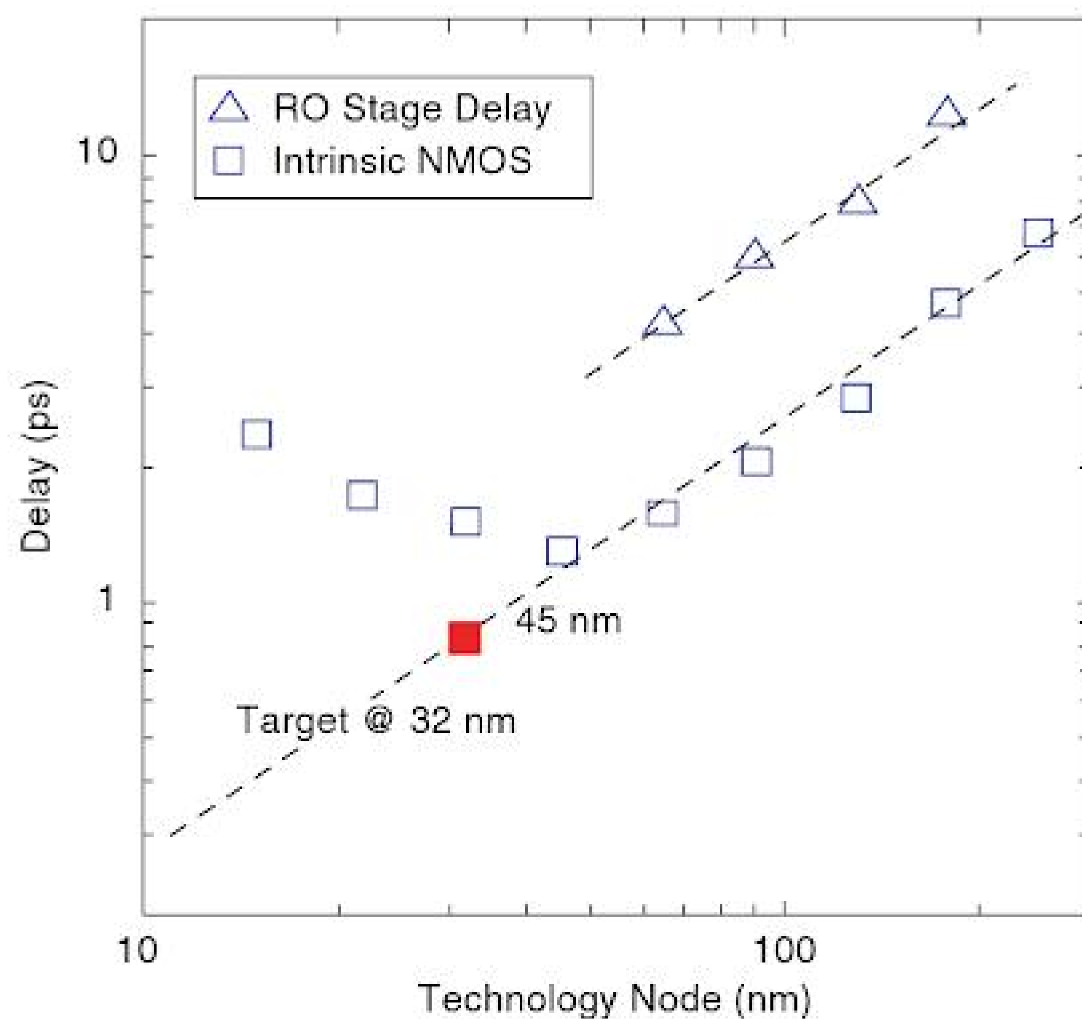

Fig. 1.4 The intrinsic delay in n-MOS transistor is projected to increase in future nodes despite continuous down scaling of the device pitch [4]

Bahan dengan hak cipta

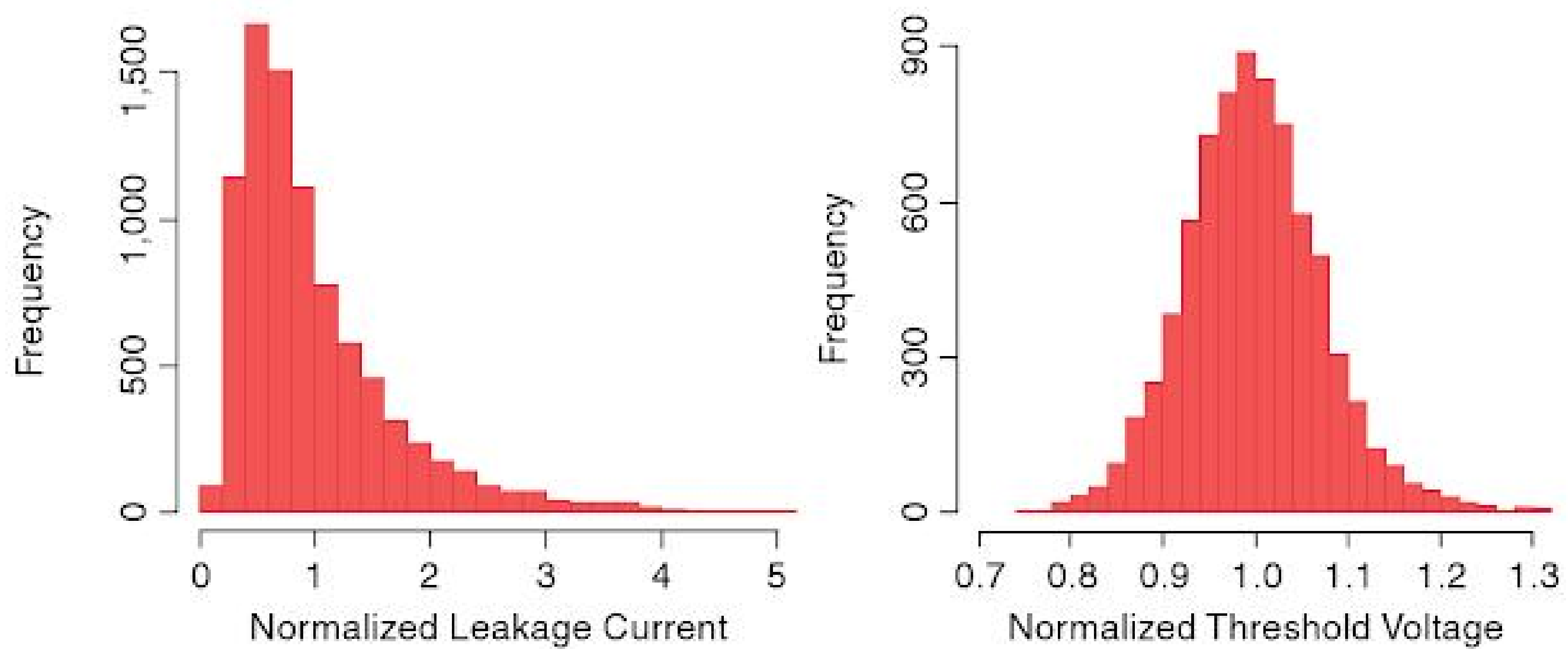

Fig. 1.5 Measured leakage current and threshold voltage in 65 nm devices reported by IBM [6]

Another issue related to scaled devices is variability [5]. Variability in transistor performance and leakage is a critical challenge to the continued scaling and effective utilization of CMOS technologies with nanometer-scale feature sizes. Some of the factors contributing to the variability increase are fundamental to the planar CMOS transistor architecture. Random dopant fluctuations (RDFs) and line-edge roughness (LER) are two examples of such intrinsic sources of variation. Other reasons for the variability increase are the advanced resolution-enhancement techniques (RETs) required to print patterns with feature sizes smaller than the wavelength of lithography. Transistor variation affects many aspects of IC manufacturing and design. Increased transistor variability can have negative impact on product performance and yield. Figure 1.5 shows measurement data reported by IBM on its 65 nm devices that clearly show variation in leakage current and threshold voltage. Variability worsens as we continue to scale in future technology nodes and it is a severe challenge.

The second barrier concerns the economic aspect of scaling. The development and manufacturing cost has increased from one node to another making scaling a less favorable option in future nodes of IC. 3D integration on the other hand achieves device density multiplication by stacking IC layers in the third dimension without aggressive scaling. Therefore it can be a viable and immediate remedy as conventional scaling becomes less cost effective.

1.2.2.2 On-Chip Interconnect

While dimensional scaling has consistently improved device performance in terms of gate switching delay, it has a reverse effect on global interconnect latency [7]. The global interconnect RC delay has increasingly become the circuit performance limiting factor especially in the deep sub-micron regime. Even though Cu/low-κ multilevel interconnect structures improve interconnect RC delay, they are not a long-term solution since the diffusion barrier required with Cu metallization has a finite thickness that is not readily scaled. The effective resistance of the interconnect

Bahan dengan hak cipta

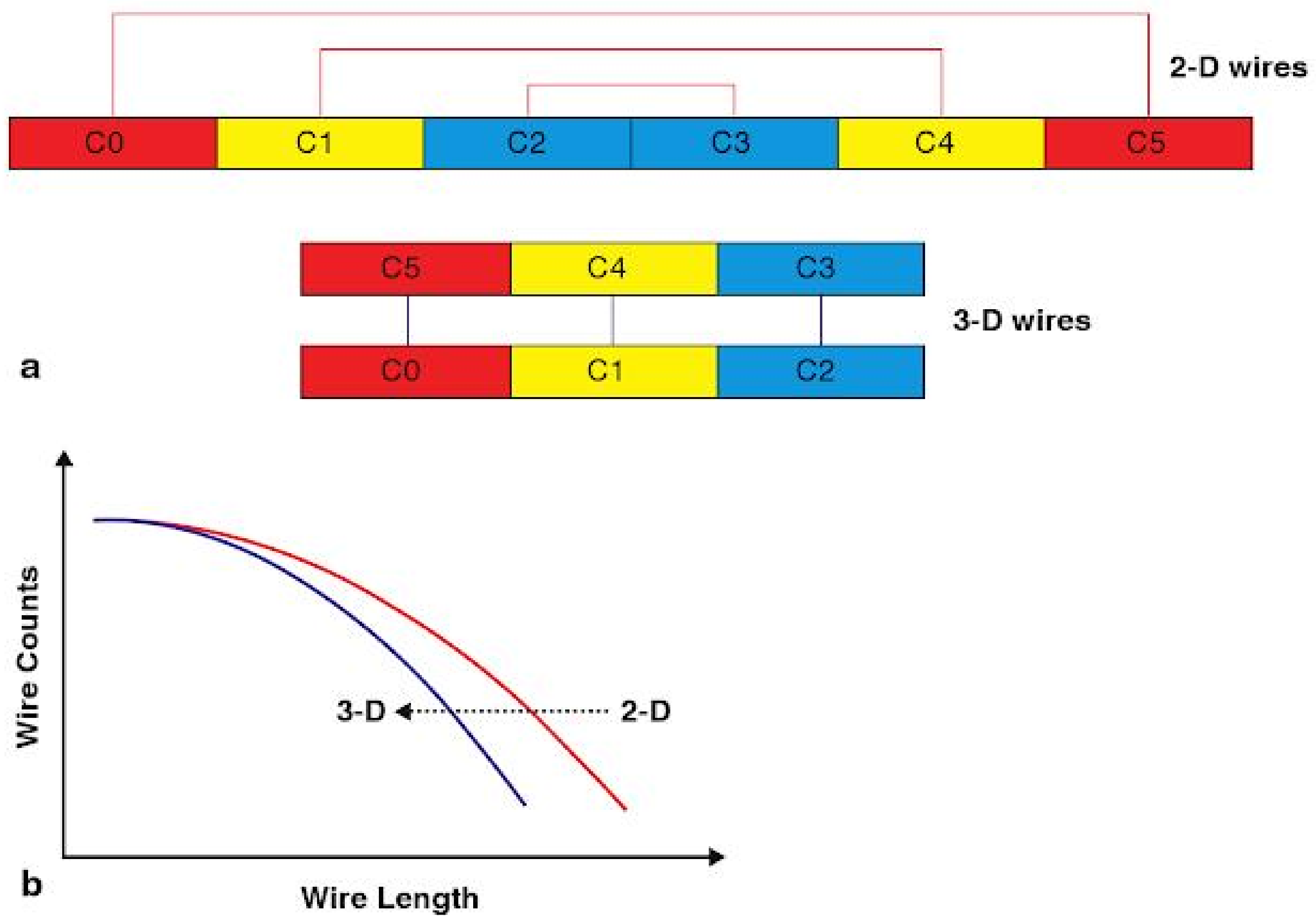

Fig. 1.6 **a** Long global wires on IC can be shortened by chip partitioning and stacking. **b** 3D integration reduces the number of long wires on IC

is larger than would be achieved with bulk copper, and the difference increases with reduced interconnect width. Surface electron scattering further increases the Cu line resistance, and hence the RC delay suffers [8]. When chip size continues to increase to accommodate for more functionalities, the total interconnects length increases at the same time. This causes a tremendous amount of power to be dissipated unnecessarily in interconnects and repeaters used to minimize delay and latency. On-chip signals also require more clock cycles to travel across the entire chip as a result of increasing chip size and operating frequency.

Rapid rise in interconnects delay and power consumption due to smaller wire cross-section, tighter wire pitch, and longer lines that transverse across larger chips is severely limiting IC performance enhancement in current and future nodes. 3D IC with multiple active Si layers stacked vertically is a promising method to overcome this scaling barrier as it replaces long inter-block global wires with much shorter vertical inter-layer interconnects as shown in Fig. 1.6.

1.2.2.3 Off-Chip Interconnect (Memory Bandwidth Gap)

Figure 1.7a depicts the memory hierarchy in today's computer system in which the processor core is connected to the memory (DRAM) via power-hungry and slower off-chip buses on the board level. Data transmission on these buses experiences

Bahan dengan hak cipta

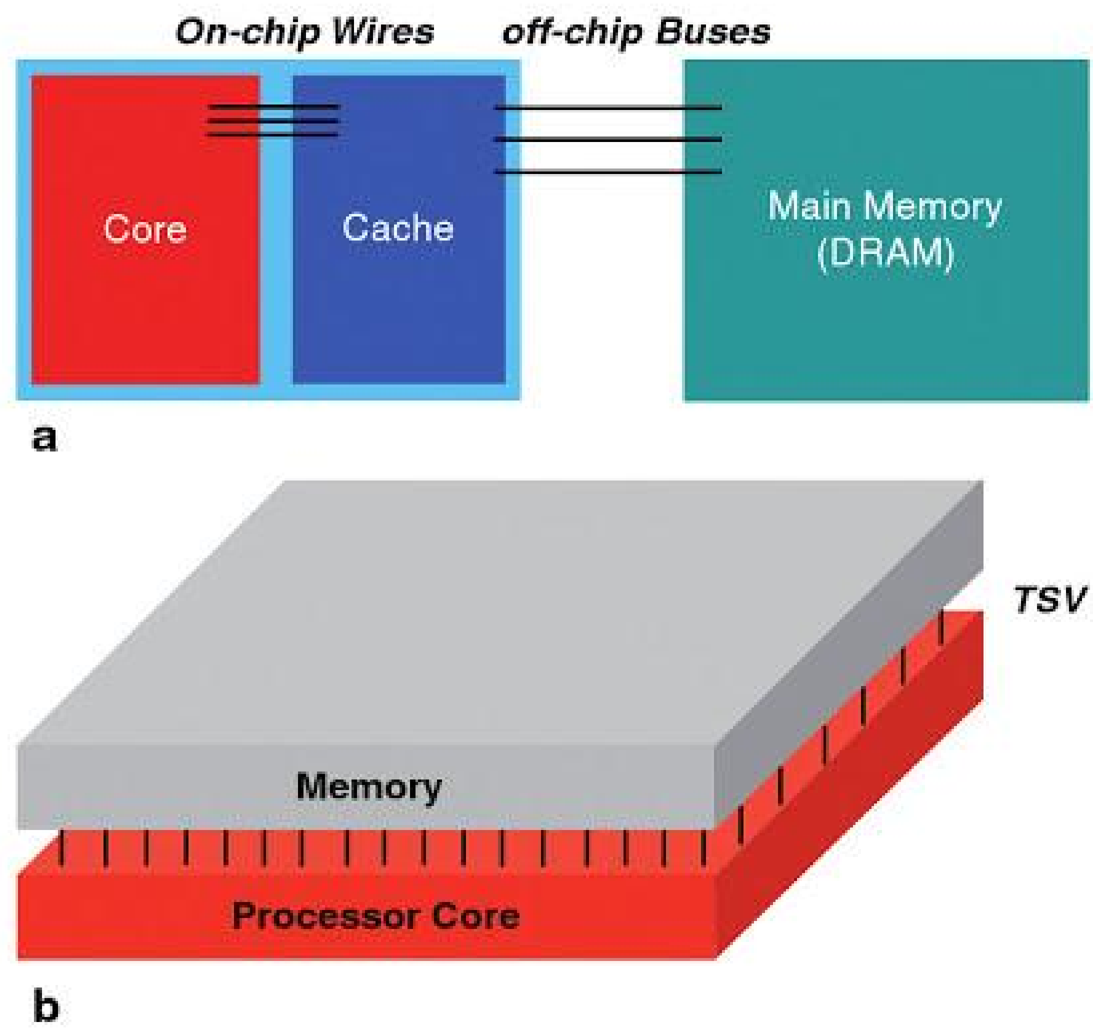

Fig. 1.7 **a** Memory Hierarchy in today computer system. **b** Direct placement of memory on processor improves the data bandwidth

severe delay and consumes significant amount of power. The number of available bus channels is also limited by the amount of external pin count available on the packaged chips. As a consequence, the data bandwidth suffers. As the computing power in processor increases in each generation, the limited bandwidth between processor core and memory places a severe limitation on the overall system performance [9]. The problem is even more pressing in multi-core architecture as every core will demand for data supply. To close this gap, the most direct way is to shorten the connections and to increase the number of data channels. By placing memory directly on processor, the close proximity shortens the connections and the density of connections can be increased by using more advanced CMOS processes (as opposed to packaging/assembly processes) to achieve fine-pitch TSV. This massively parallel interconnection is shown in Fig. 1.7b. Table 1.1 is a comparison between 2D and 3D implementations in terms of connection density and power consumption. Clearly, 3D can provide bandwidth enhancement (100X increment at the same frequency) at lower power consumption (10X reduction). Effectively, this translates into 1,000X improvement in bandwidth/power efficiency, an extremely encouraging and impressive number.

Table 1.1 Comparison of 2D and 3D implementations

	2D	3D	Remark
Connections density	<1e3 per cm^2	~1e5 per cm^2	100X increment
*Power consumption per pin/via	30–40 mW	~25 μW	
Total power consumption (per cm^2)	30–40 W	2.5 W	10X reduction

*Data from Tezzaron [10].

Bahan dengan hak cipta

1.3 Options of 3D IC

1.3.1 System Integration Landscape

System integration, that is, the integrating together of circuits or intellectual property (IP) blocks, is one of the major applications of 3D integration. As such, 3D integration must compete against a number of established technologies. Figure 1.8 compares the relative capability of several system integration methods (board, 2D multi-chip module—2D-MCM, package-on-package—PoP, system-in-package—SiP, and 2D system-on-chip—2D-SoC) in terms of form factor and interconnects density between circuit blocks [11]. 3D integration offers more compact form factor and higher chip-to-chip interconnects density [11, 12]. Comparing with 2D-SoC, 3D integration shortens time-to-market and lowers the system cost. By using larger number of smaller and shorter through silicon via (TSV) as compared to wire bonding in SiP, performance is enhancement via 3D integration due to smaller latency and higher bandwidth, as well as smaller power consumption.

1.3.2 Classification

There are a number of technology options to arrange integrated circuits in a vertical stack. It is possible to stack ICs in a vertical fashion at various stages of process-

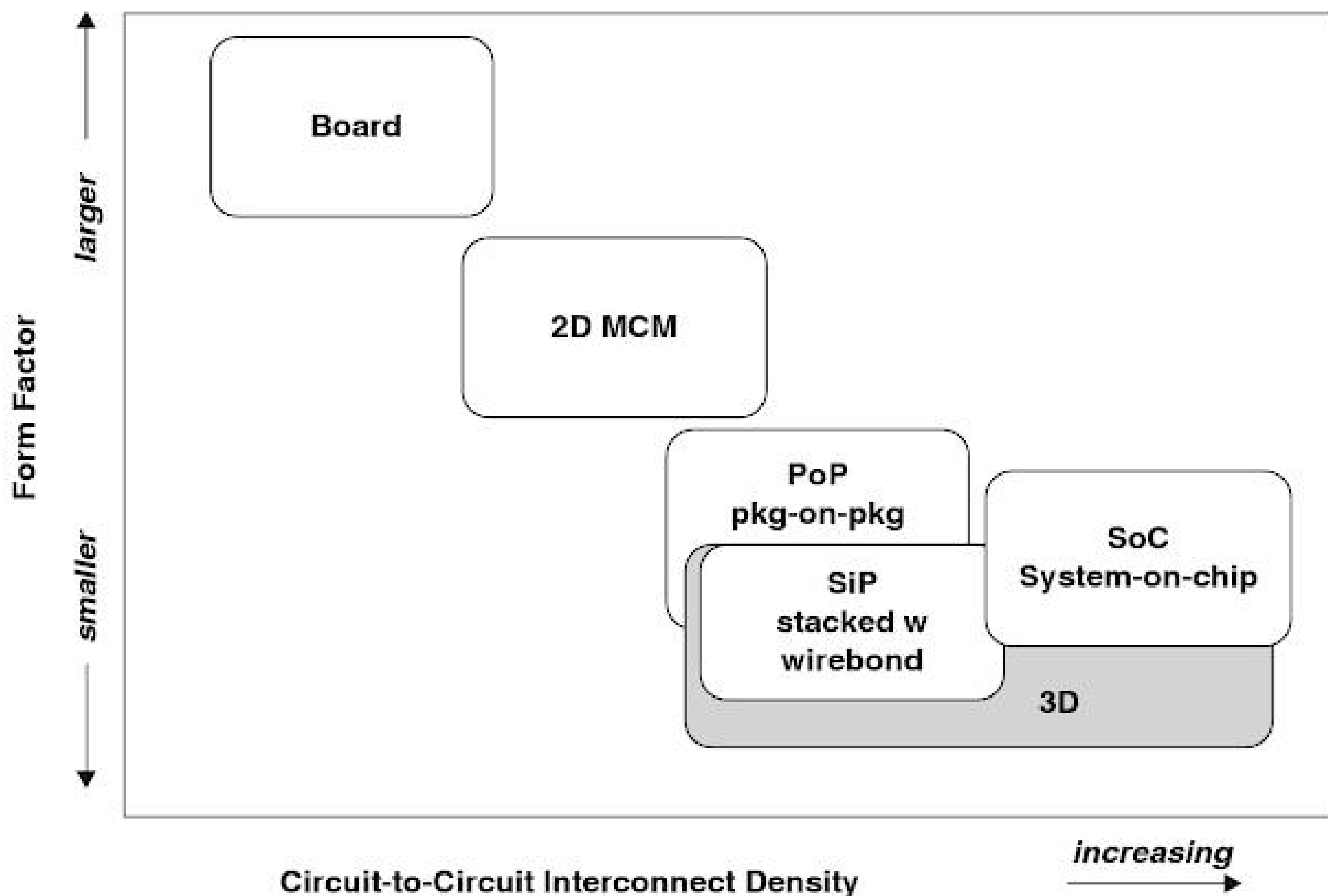

Fig. 1.8 Comparison of various system integration technologies in terms of form factor and circuit-to-circuit interconnect density [11, 12]

Bahan dengan hak cipta

ing: (1) post-singulation 3D packaging (e.g., chip-to-chip), and (2) pre-singulation wafer level 3-D integration (e.g., chip-to-wafer, wafer-to-wafer, and monolithic approaches). Active layers can be vertically interconnected using physical contact such as bond wire or interlayer vertical via (including TSV). It is also possible to establish chip to chip connection via non-contact (or wireless) links such as capacitive and inductive couplings [13]. Capacitive coupling utilizes a pair of electrodes that are formed using conventional IC fabrication. The inductive-coupling I/O is formed by placing two planar coils (planar inductors) above each other and is also made using conventional IC fabrication. The advantages of these approaches are fewer processing steps hence lower cost, no requirement for ESD protection, low power, and smaller area I/O cell.

Since there is substantial overlap between various options and lack of standardization in terms of definition, classification of 3D IC technology is often not straight forward. This chapter makes an attempt to classify 3D IC based on the processing stage when stacking takes place.

1.3.3 Monolithic Approaches

In these approaches, devices in each active layer are processed sequentially starting from the bottom-most layer. Devices are built on a substrate wafer by mainstream process technology. After proper isolation, a second device layer is formed and devices are processed by conventional means on the second layer. This sequence of isolation, layer formation, and device processing can be repeated to build a multilayer structure.

The key technology in this approach is forming a high quality active layer isolated from the bottom substrate. This bottom-up approach has the advantage that precision alignment between layers can be accomplished. However, it suffers from a number of drawbacks. The crystallinity of upper layers is usually low and imperfect. As a result, high performance devices cannot be built in the upper layers. Thermal cycling during upper layer crystallization and device processing can degrade underlying devices and therefore a tight thermal budget must be imposed. Due to the sequential nature of this method, manufacturing throughput is low. A simpler FEOL process flow is feasible if polycrystalline silicon can be used for active devices; however, a major difficulty is to obtain high-quality electrical devices and interconnects. While obtaining single-crystal device layers in a generic IC technology remains in the research stage, polycrystalline devices suitable for non-volatile memory (NVM) have not only been demonstrated but have been commercialized (for example by SanDisk). A key advantage of FEOL-based 3-D integration is that IC BEOL and packaging technologies are unchanged; all the innovation occurs in 3-D stacking of active layers.

A number of FEOL techniques include: laser beam recrystallization [14, 15], seeding-assisted recrystallization [16, 17], selective epitaxy and over-growth [18], and grapho-exitaxy [19].

Bahan dengan hak cipta

1.3.4 Assembly Approaches

This is a parallel integration scheme in which fully processed or partially processed integrated circuits are assembled in a vertical fashion. Stacking can be achieved with one of these methods: (1) chip-to-chip, (2) chip-to-wafer, and (3) wafer-to-wafer. Vertical connection in chip to chip stacking can be achieved using wire bond as shown in Fig. 1.9 or through silicon via (TSV) as shown in Fig. 1.10.

Wafer level 3D integration, such as chip-to-wafer and wafer-to-wafer stacking, use TSV as the vertical interconnect. This integration approach often involves a sequence of wafer thinning and handling, alignment, TSV formation, and bonding. The key differentiators are:

- Bonding medium—metal-to-metal, dielectric-to-dielectric (oxide, adhesive, etc) or hybrid bonding;
- TSV formation—via first, via middle or via last;
- Stacking orientation—face-to-face or back-to-face stacking;
- Singulation level—chip-to-chip, chip-to-wafer or wafer-to-wafer.

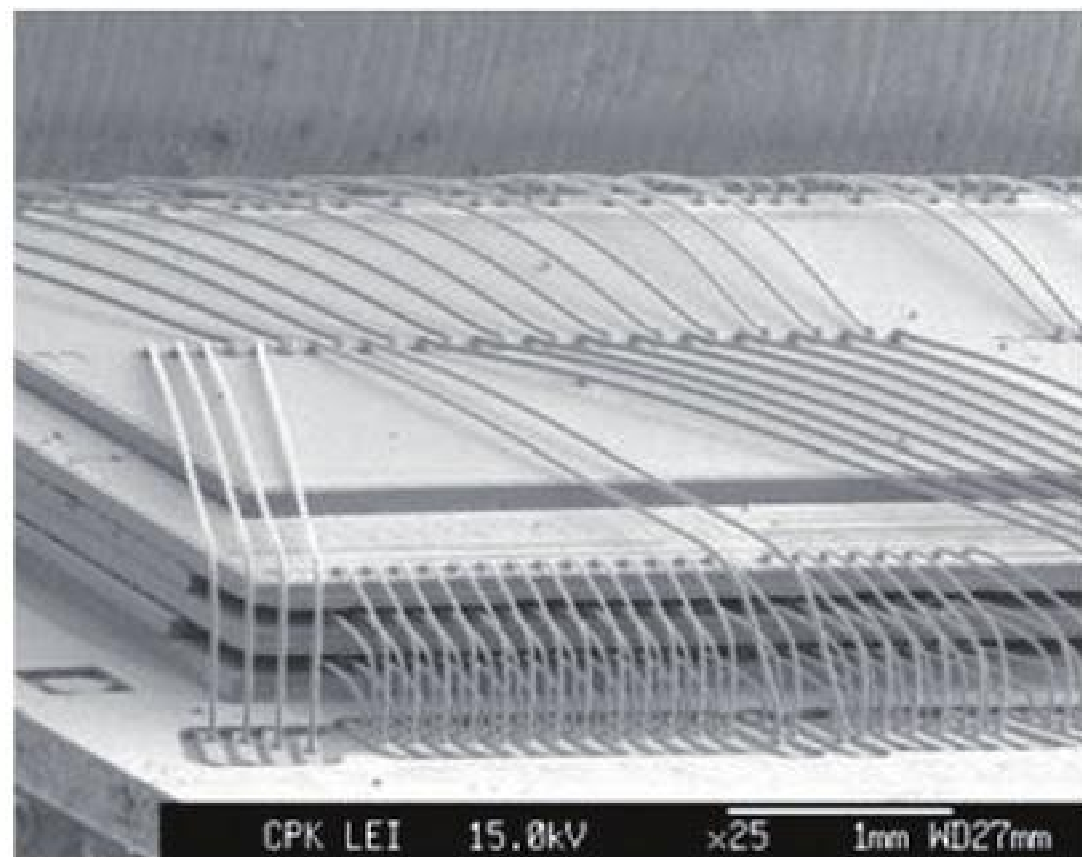

Fig. 1.9 Stacked die with wire bond interconnections in a chip-scale package [20]

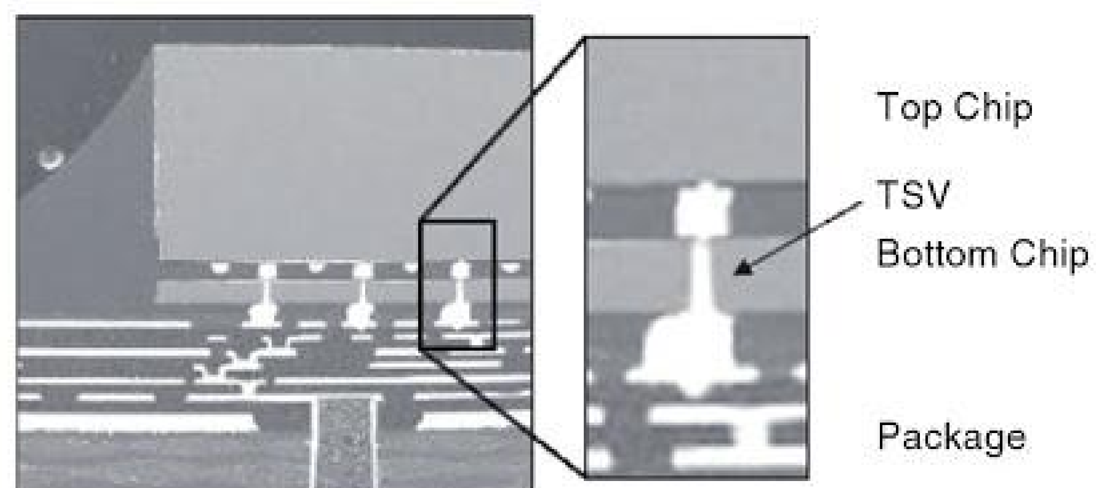

Fig. 1.10 3D chip stacking using through silicon via [21]

Bahan dengan hak cipta

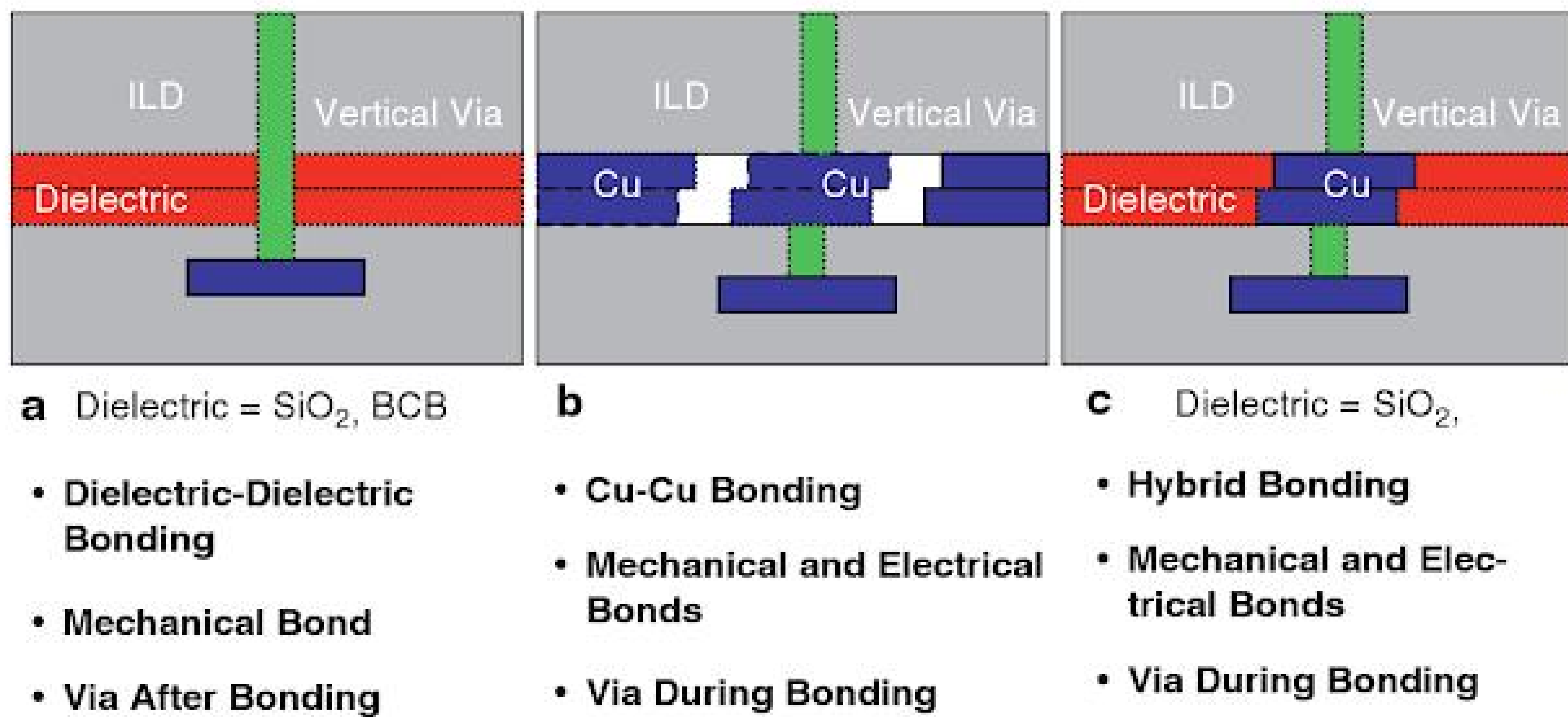

Fig. 1.11 Wafer bonding techniques for wafer-level 3-D integration: **a** dielectric-to-dielectric, **b** metal-to-metal, and **c** dielectric/metal hybrid

The types of wafer bonding potentially suitable for wafer-level 3D integration are depicted in Fig. 1.11. Dielectric-to-dielectric bonding is most commonly accomplished using silicon oxide or BCB polymer as the bonding medium. These types of bonding provide primary function as a mechanical bond and the inter-wafer via is formed after wafer-to-wafer alignment and bonding (Fig. 1.11a). When metallic copper-to-copper bonding is used (Fig. 1.11b), the inter-wafer via is completed during the bonding process; note that appropriate interconnect processing within each wafer is required to enable 3D interconnectivity. Besides providing electrical connections between IC layers, dummy pads can also be inserted at the bonding interface at the same time to enhance the overall mechanical bond strength. This bonding scheme inherently leaves behind isolation gap between Cu pads and this could be a source of concern for moisture corrosion and compromise the structural integrity especially when IC layers above the substrate is thinned down further. Figure 1.11c shows a bonding scheme utilizing a hybrid medium of dielectric and Cu. This scheme in principle provides a seam-less bonding interface consists of dielectric bond (primarily a mechanical bond) and Cu bond (primarily an electrical bond). However, very stringent requirements with regards to surface planarity (dielectric and Cu) and Cu contamination in the dielectric layer due to misalignment are needed.

The selection of the optimum technology platform is subject to ongoing development and applications. Cu-to-Cu bonding has significant advantages for highest inter-wafer interconnectivity. As a result, this approach is desirable for microprocessors and digitally-based system-on-a-chip (SoC) technologies. Polymer-to-polymer bonding is attractive when heterogeneous integration of diverse technologies is the driver and the inter-wafer interconnect density is more relaxed; benzocyclobute (BCB) is the polymer most widely investigated. Taking advantage of the viscosity of the polymer, this method is more forgiving in terms of surface planarity and particle contamination. Oxide-to-oxide bonding of fully processed IC wafers requires atomic-scale smoothness of the oxide surface. In addition, wafer distortions

Bahan dengan hak cipta

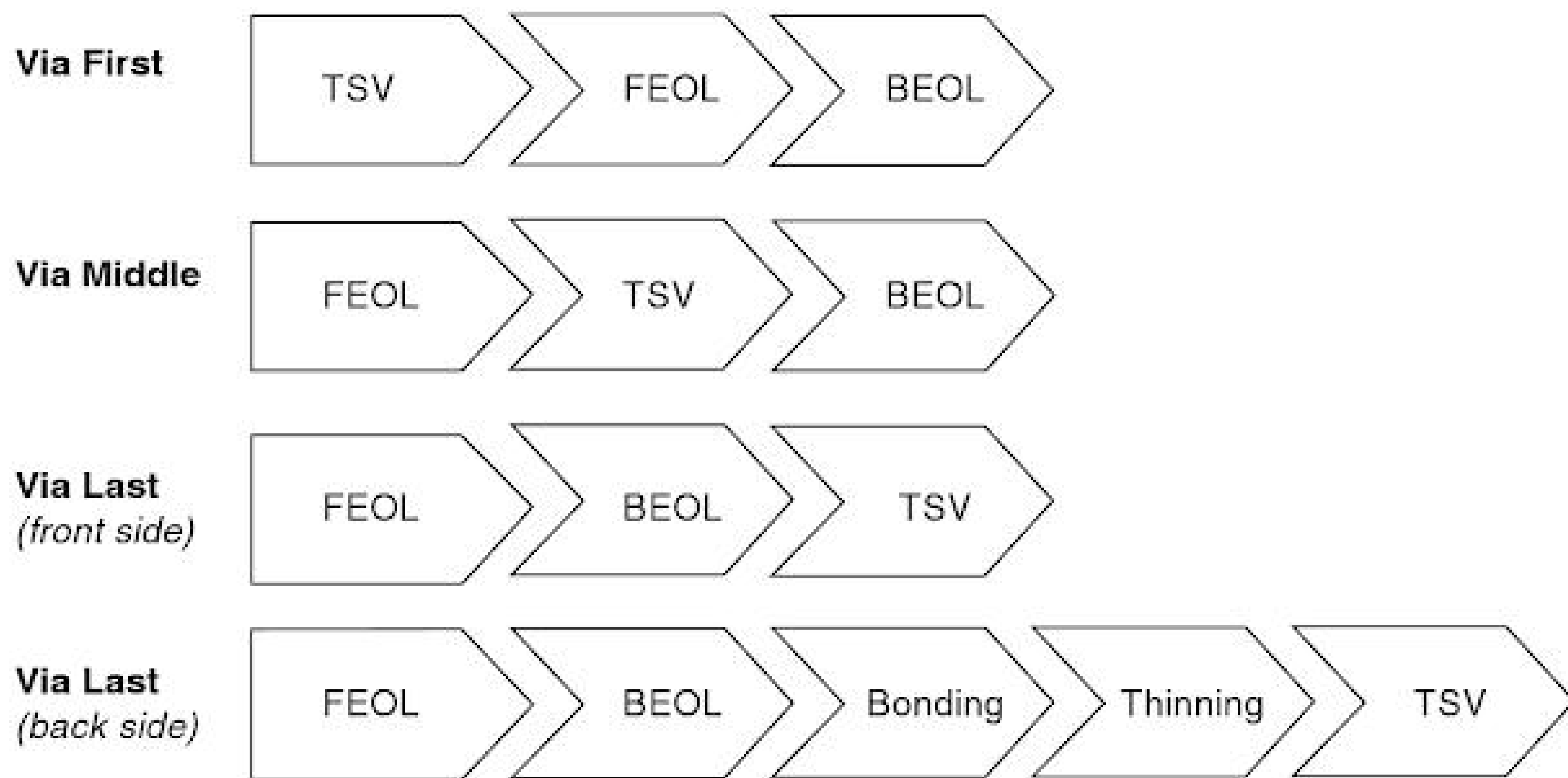

Fig. 1.12 TSV can be formed at various stages of IC processing

introduced by FEOL and BEOL processing introduces sufficient wafer bowing and warping that prevents sufficient contact area to achieve the required bonding strength. While oxide-to-oxide bonding after FEOL and local interconnect processing has been shown to be promising (particularly with SOI wafers that allows for extreme thinning down to the buried oxide layer) the increased wafer distortion and oxide roughness after multilevel interconnect processing require extra attention during processing.

TSV can be formed at various stages during the 3D IC process as shown in Fig. 1.12. When TSV is formed before any CMOS processes, the process sequence is known as "via first". It is also possible to form the TSV when the front end processes are completed. In this "via middle" process, back end processes will continue after the TSV process is completed. When TSV is formed after the CMOS processes are completed, it is known as "via last" process. TSV can be formed from the front side or the back side of the wafer. The above schemes have different requirements in terms of process parameters and materials selection. The choice depends on final application requirements and infrastructures in the supply chain.

Another key differentiator in 3D IC integration is related to the stacking orientation. One option is to perform face-to-face (F2F) alignment and bonding with all required I/Os brought to the thinned backside of the top wafer (which becomes the face of the two-wafer stack). Another approach is to temporarily bond the top wafer to a handling wafer, after which the device wafer is thinned from the back side and permanently bonded to the full-thickness bottom wafer; after this permanent bonding the handling wafer is removed. This is also called a back-to-face (B2F) stacking. These two stacking orientations are shown in Fig. 1.13.

F2F stacking allows a high density layer to layer interconnection which is limited by the alignment accuracy. Handle wafer is not required in F2F stacking and this imposes more stringent requirement on the mechanical strength of the bonding interface in order to sustain shear force during wafer thinning which is often

Bahan dengan hak cipta

image
not
available

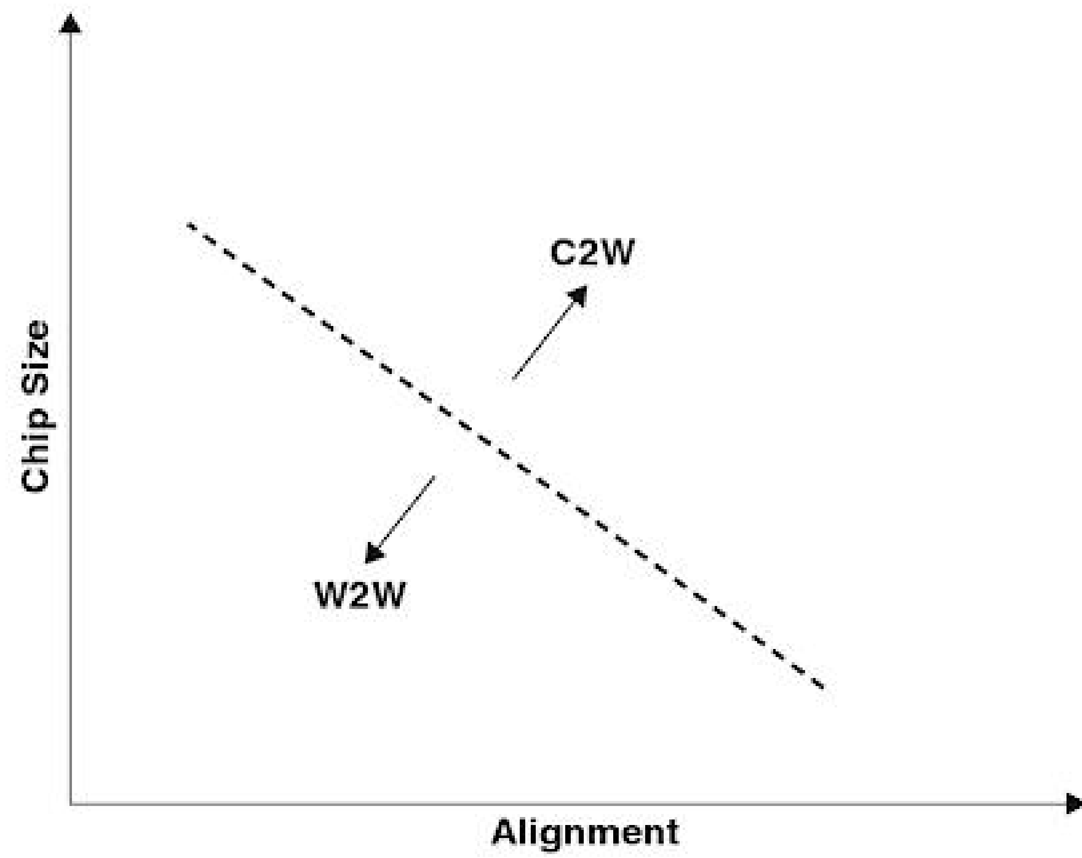

Fig. 1.14 The choice between C2W and W2W depends on the chip size and the required alignment accuracy

interconnections, W2W is a preferred choice to maintain acceptable throughput by performing a wafer level alignment. W2W is also preferred when chip size gets smaller.

1.4 Technology Platforms and Strategies

A number of new enabling technologies must be developed and introduced into the existing fabrication process flow to make 3D integration a reality. Depending on the level of granularity, new capabilities include wafer bonding (permanent or temporary), through silicon/strata via (TSV), wafer thinning and handling, and precision alignment. There are a number of references on technology platforms available in the literature and the references therein [22, 23]. A brief introduction to TSV process flow is given. This section will primarily discuss low temperature Cu–Cu permanent bonding which is the author's core research expertise.

1.4.1 Through Silicon Via

Figure 1.15 is a generic process flow of TSV fabrication flow using Cu as the core metal. It begins with high aspect ratio deep etching of Si. Dielectric liner layer is then deposited on the via sidewall followed by barrier and Cu seed layers deposition. Liner layer, which is made of dielectric layer such as silicon dioxide, provides electrical isolation between Cu core and Si substrate. The liner thickness must be chosen appropriately to control leakage current and capacitance between Cu core and Si substrate. Cu super conformal filling is then achieved

Bahan dengan hak cipta

image
not
available

image
not
available

bonding parameters, i.e., temperature, duration, and contact pressure, are frequently considered. In the bonding procedures described in above, thermo-compression bonding of Cu is accomplished in two steps, i.e., an initial bonding step to establish bond between pairing wafers and a post-bonding anneal to enhance the bond. Since the bonding step is a single wafer pair step, long bonding duration will decrease through-put in a manufacturing environment. On the other hand, annealing can be accomplished in an atmospheric furnace and it is possible to process batches of wafers during annealing. Therefore, a better way to achieve high through-put Cu wafer bonding is to initiate a preliminary bond with a short bonding step and to enhance the bonding strength with a post-bonding anneal. A number of references that discuss process parameters during thermo-compression bonding of Cu can be found in [27, 28].

1.4.2.5 Observation of Interfacial Voids

In order for the bonded Cu layer to act as a reliable electrical or mechanical bond, a defect-free and uniformly bonded Cu layer is desired. Careful examination by SEM analysis across a length of 20 μm reveals large voids in the bonded Cu layers as shown in Fig. 1.17. These voids are located at and near the location of the original bonding interface. These voids can provide nucleation sites for electromigration failure and can lead to open circuit failures. When the void density is too high, the voids can also cause thin film delamination. Therefore, this observation requires

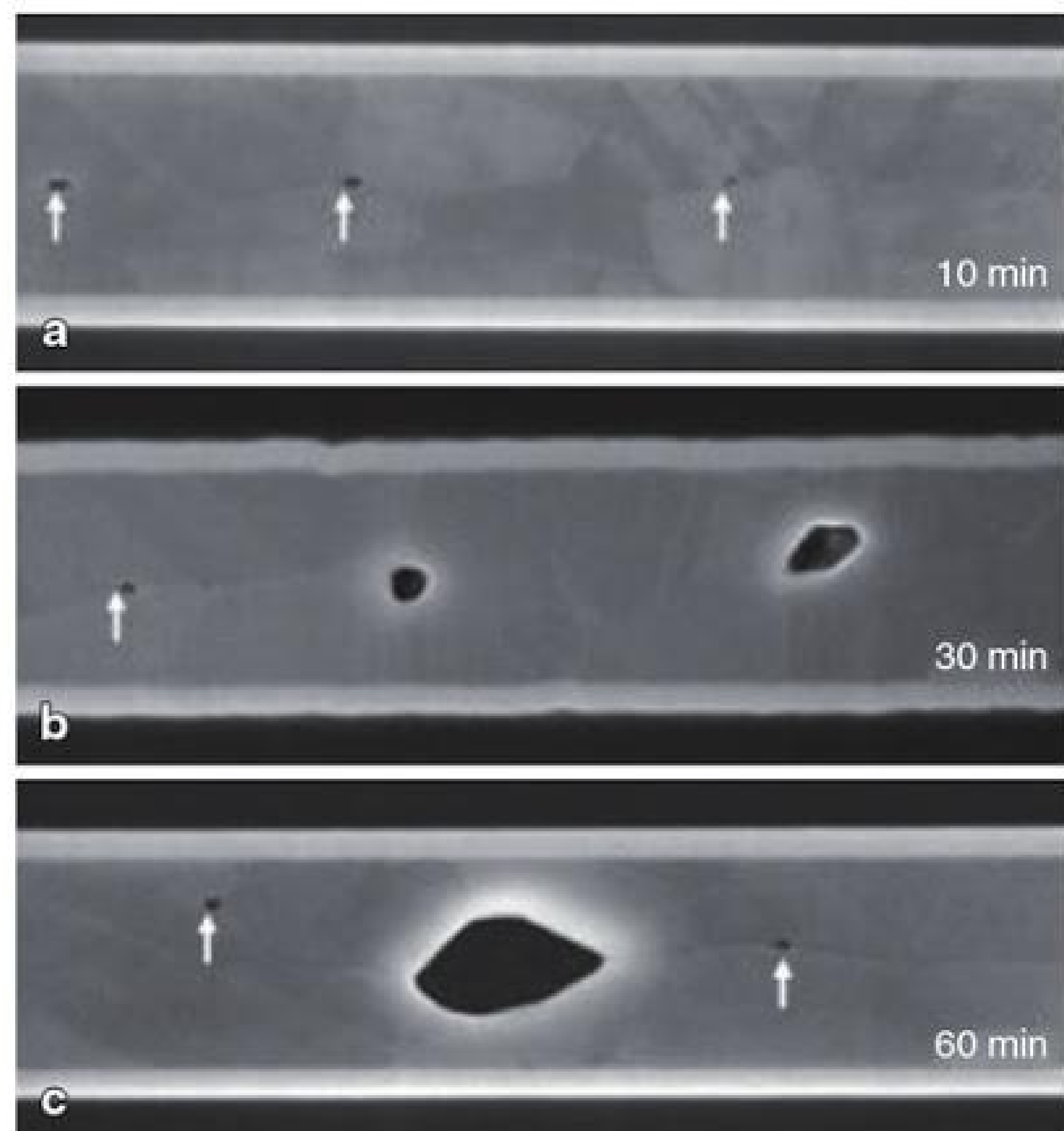

Fig. 1.17 Interfacial void growth during Cu thermo-compression bonding at 300°C for **a** 10 min **b** 30 min and **c** 60 min

Bahan dengan hak cipta

image
not
available

image
not
available

image not available

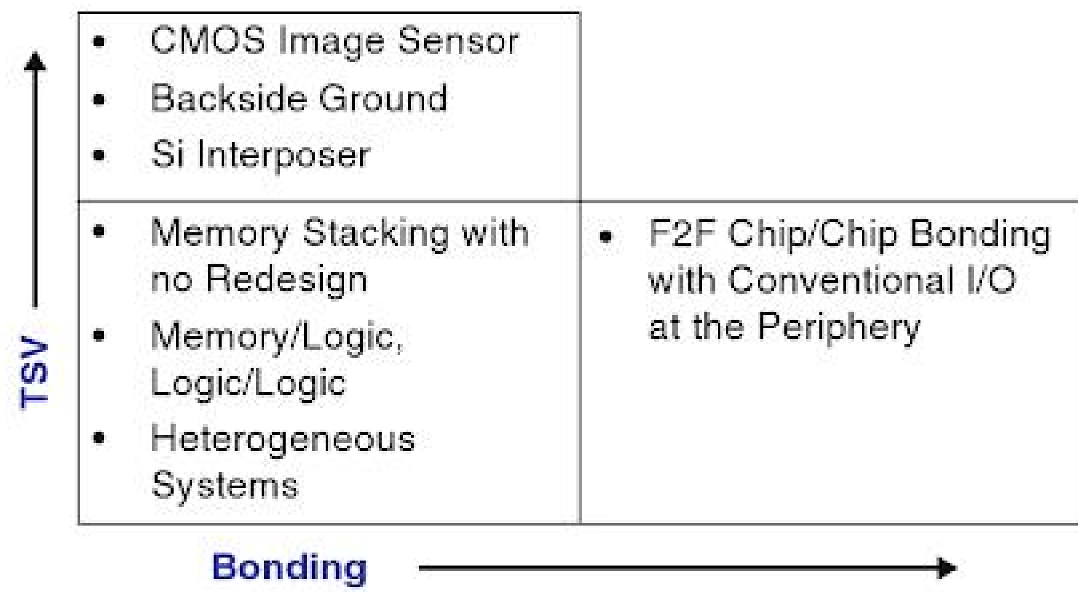

Fig. 1.19 Applications enabled by 3D technology

logic, and (4) heterogeneous integration of disparate chips. Regardless of the main driver, the feasibility and key consideration of any 3D application for consumer products has always been low cost manufacturing.

Broadly, applications enabled by 3D technology can be classified into three categories as shown in Fig. 1.19. The first group of products only utilizes TSV such as CMOS image sensor (at the time of writing, there are commercial products from companies such as ST Microelectronics, Toshiba, OKI, etc), backside ground (e.g., SiGe power amplifier by IBM), and silicon interposer (based on the silicon carrier alone). In this class of devices, chip to chip bonding is not required. In another group, 3D devices are implemented by bonding chip on chip in a face to face fashion. Unlike system shown in Fig. 1.9, chip to chip electrical connections can be established with micro-bump or bump-less metal-metal bonding. I/O is formed using conventional wire bond or flip chip at the non-bonding periphery area. One such example is the Sony Play Station featuring memory on logic. The real 3D devices that make use of both TSV and bonding include stand-alone high density memory stack, memory on logic, logic on logic, and heterogeneous systems. At the time of this writing, there has been announcement from Elpida on a multi-layer DRAM stack using 3D stacking technology.

For more than 40 years, performance growth in IC is realized primarily by geometrical scaling. In more recent nodes, performance boosters are used to sustain this historical growth. Moving forward, 3D integration is an inevitable path. There have been significant investment in 3D technology by various sectors and the development has been both rewarding and encouraging. While 3D technology is not without it challenges, it is likely to see wider adoption of 3D technology in the future when solutions for thermal management, EDA tools, testing, and standardization are made available in the IC supply chain.

Acknowledgments The author is supported by funding from the Nanyang Technological University through an award of Nanyang Assistant Professorship, Defense Science and Technology Agency (DSTA, Singapore), Semiconductor Research Corporation (SRC, USA) through a subcontract from the Interconnect and Packaging Center at the Georgia Institute of Technology, and Defense Advanced Research Projects Agency (DARPA, USA). The author thanks Professor Rafael Reif of MIT for his constructive and valuable comments on the content of this chapter.

Bahan dengan hak cipta

image
not
available

image
not
available

image
not
available

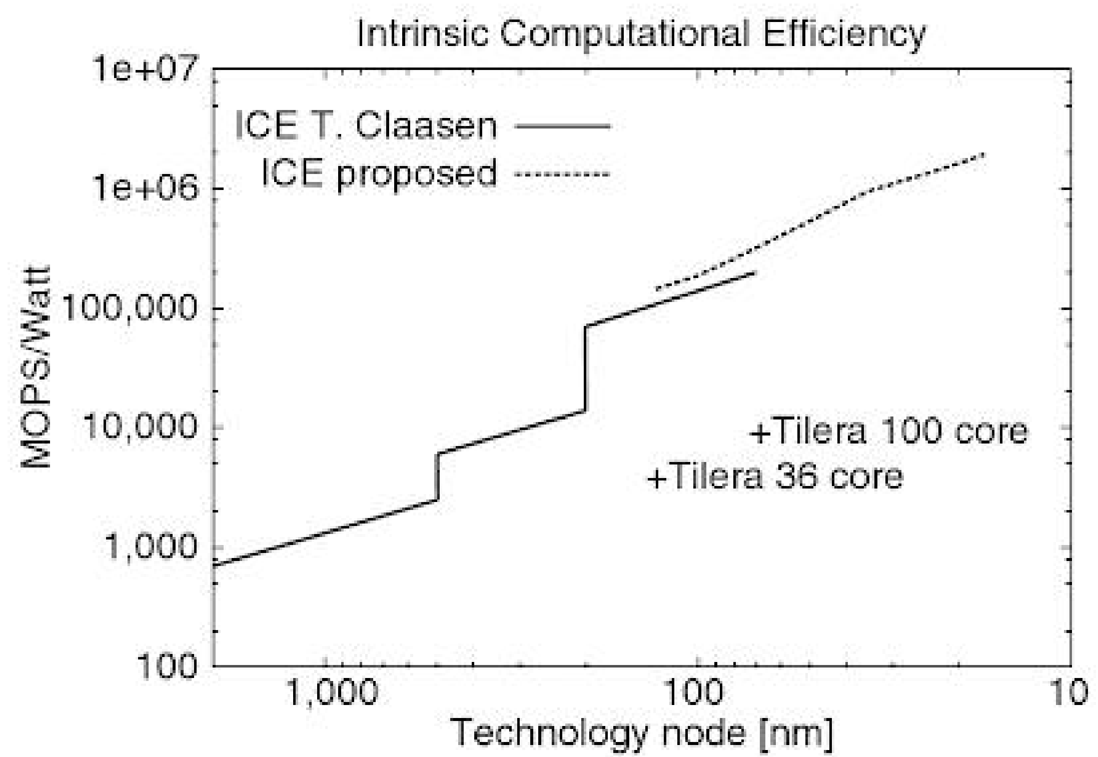

Fig. 2.1 Computational efficiency vs. minimum feature size is shown here. The discrepancy in the overlapping section between T. Classon's work and this study is due to the significant variation in the energy consumption of different adder architectures. The performance of the two processors shown fall outside of this range due to the overhead of the control circuitry and interconnect which is not accounted for by this metric

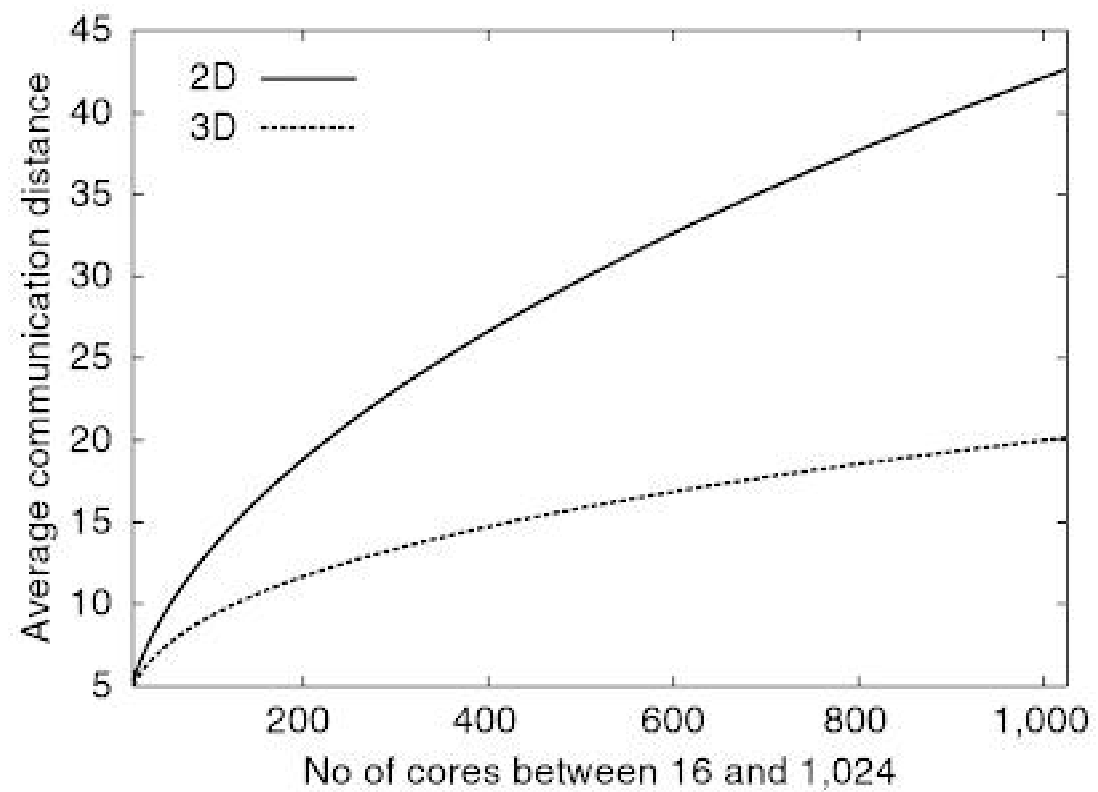

Fig. 2.2 The average geometric distance for a multi-core system for a 2-D and a 3-D realization

ric distance translates linearly to latency, we can expect to cut communication latency by 50%. A number of recent studies of communication performance in 3-D structures [3–7] demonstrate the significant potential of 3-D integration technology for reducing power consumption and increasing performance.

3-D integration enables stacking of memory on top of processors, thus realizing a direct low latency and high bandwidth memory access link. However, to exploit the benefits, the memory architecture has to be adapted to allow for multi-port, parallel memory access. Several recent studies have explored various memory and cache architectures while exploiting the third dimension. For instance Li et al. [8] propose a 3-D distributed L2 cache and observe a 50% access latency reduction, essentially due to shorter wires within the L2 cache. Loh [9] explores the effect of parallel memory access by means of multiple memory controllers and ranks in a 3-D stacked DRAM based memory architecture and reports a performance increase of more than 280% over a conventional memory architecture for a set of benchmark applications. In our model we assume that in a 3-D topology, DRAM is used as embedded memory because it can be placed on a separate die, thus leveraging on the capability of 3-D to integrate different process technology in the same system.

Bahan dengan hak cipta

image
not
available

image
not
available

Table 2.1 Notation and metrics of comparison

Abstraction of architectural design parameters	
tn	Minimum feature size of a technology node (nm)
arch	Architecture of system (2D, 3D2, 3D4, 3D8, 3D16)
ω	Ratio of on- to off-chip memory ($\omega = 1$: all memory is on-chip, $\omega = 0$: all off-chip)
Δ	Memory distribution factor ($\Delta = 1$: all centralized memory, $\Delta = 0$: all local)
μ_T	Number of memory accesses per h/w operation ($\mu_T = 1$: one mem. access per op)
μ_s	Amount of memory per h/w operator (typically μ_s = 1,000–10,000)
σ	Interconnect sharing factor ($\sigma = 1$: no sharing, $\sigma = 0$: completely shared)
n	Number of die layers for 3-D architectures
area	Area of die
Technology and architecture dependent parameters	
E_{32}	Energy for a 32-bit add operation
e_1	Energy for a 32-bit read/write to local SRAM
e_2	Energy to transport a 32-bit word over 1 mm on a planar on-chip bus
e_3	Energy to transport a 32-bit word over one vertical layer across TSVs
a_1	Area for a 32-bit memory word in SRAM or DRAM
a_2	Area for a 1 mm long 32-bit planar on-chip bus
a_3	Area for 32 TSVs
$E_{offchip}$	Energy to read/write to off-chip memory. Includes I/O drivers, inter-chip communication and memory chip energy consumption
Primary comparison metrics	
ICE	Number of 32-bit add operations per Joule
ICD	Number of 32-bit adders per mm^2
$A_int^{tn}_{arch}$	Interconnect area required to transport a 32-bit word from a non-adjacent on-chip memory to the local cache: $\sqrt{\frac{area}{n}}a_2 + \frac{n}{2}a_3$
$E_int^{tn}_{arch}$	Interconnect energy required to transport a 32-bit word from a non-adjacent on-chip memory to the local cache: $\sqrt{\frac{area}{n}}c_2 + \frac{n}{2}c_3$
EE^{tn}_{arch}	Effective Energy for a 32-bit addition: $EE^{tn}_{arch} = E^{tn}_{32} + \mu_T(\omega(e_1 + \Delta E_int^{tn}_{arch}) + (1-\omega)(e_1 + \mathrm{E}_int^{tn}_{arch} + E_{offchip}))$
ECE^{tn}_{arch}	Amount of computation achieved with 1 J: $\frac{1}{EE^{tn}_{arch}}$
EA^{tn}_{arch}	Effective area for a 32-bit addition without off-chip memory: $EA^{tn}_{arch} = A^{tn}_{32} + \mu_S\ \omega\ a_1 + \sigma A_int^{tn}_{arch}$

For example the special case of $\mu_T = 0$ (no memory read or write) and 2-D in a 130 nm technology we get

$$ECE^{130}_{2D} = \frac{1}{EE^{tn}_{arch}} = \frac{1}{E^{130}_{32}} = ICE^{130} = 144.3\ \mathrm{GOPS/W}$$

2.2.4 *Effective Computational Density*

Similarly, the area cannot be filled with computational units only. We need to take memory and interconnect into account as well. We define the *Effective Area* (*EA*) as follows.

Bahan dengan hak cipta

image
not
available

image
not
available

associated with off-chip transactions, such as bus controller architecture, termination power, transaction delay, and the number of peripheral I/O devices, which cause the energy to vary over a wide range depending on these choices. In our study we have been consistent with the values we use in order to minimize the impact on comparisons between different schemes.

2.4 ECE Trends and Dependencies

Next, we study the dependency of *ECE* on parameters like Δ and ω and then we investigate the limits of *ECE* and raw performance under power, area and frequency constraints.

To see the overall trend, Fig. 2.3 illustrates how the *ECE*, the performance for a given power envelope, will develop as technology scales. As a reference the plot shows the *ECE* of two recent multi-core Tilera processors. 3-D topologies have a 3 times higher *ECE*, mainly due to lower communication power consumption in a more compact geometry. Moreover, this increased efficiency of 3-D is gained at a much smaller area and lower frequency for the same performance, as will be illustrated below in Sect. 2.4.3.

2.4.1 Distributed versus Central Memory

To study the effect of the memory distribution factor on the *ECE* we assume that for every operation on average one word has to be read or written from or to memory (or cache)[3]. Hence, $\mu_T = 1.0$.

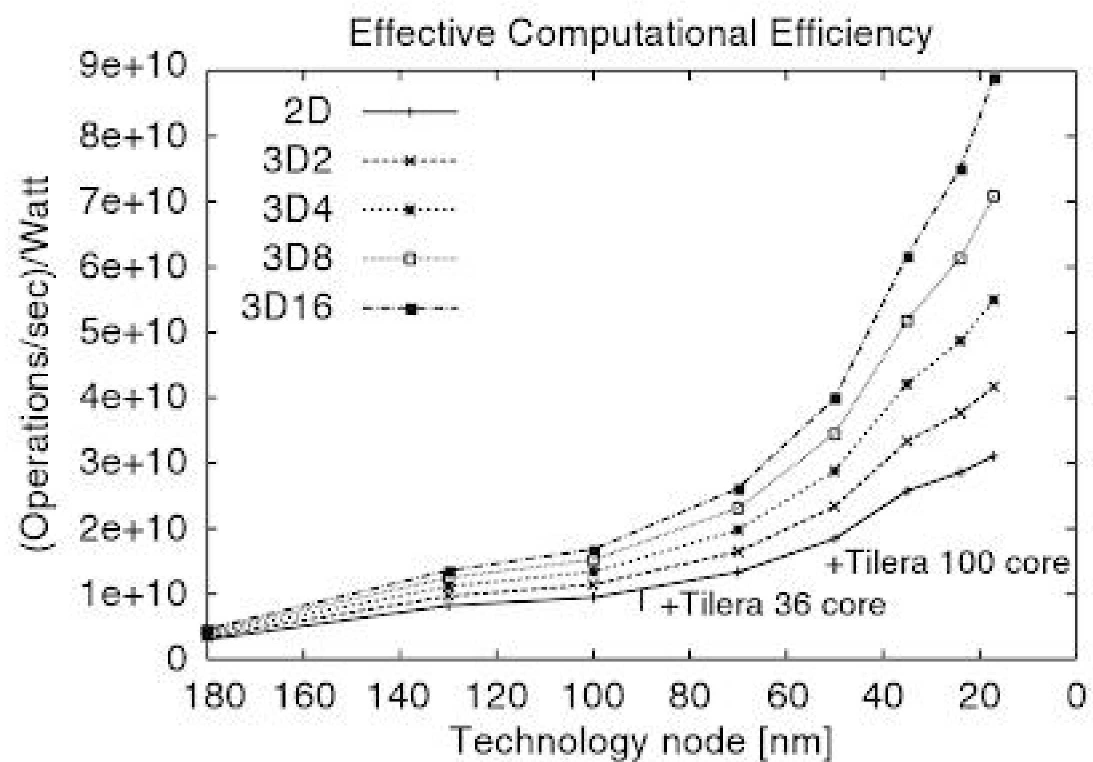

Fig. 2.3 Performance of different topologies at different technology nodes with Δ=0.05, $\omega = 1$ and $\mu_T = 1$. The data for the two Tilera processors are closer to the theoretical performance than in Fig. 2.1 due to the fact that the interconnect overhead has been accounted for, although the control circuitry overhead is neglected

[3] We assume registers and small register files close to the operators. Reading and writing of registers is not considered as memory access.

Bahan dengan hak cipta

image
not
available

image
not
available

image not available

cycle. It has 4 MB (=1 M Word) of on-chip cache resulting in $\mu_s = 2^{20}/512 = 2{,}048$. The Niagara 2 [17] processor from Sun Microsystems, which is an 8 core 64 thread processor with 4 MB of on-chip cache, falls into a similar range.

Keep in mind that our model illustrates trends and limits but does not account for control logic, decoders, arbiters, etc. The area contribution of that part is not seen in Fig. 2.7a. We usually attribute those elements to the processing units and hence, their area fraction is lower in Fig. 2.7a than we intuitively expect. However, the comparison of 2-D and 3-D topologies is interesting. Due to the higher density of memory in a 3-D architecture (DRAM vs SRAM in 2-D), the area dominance of memory in 2-D is much higher than in 3-D for the same μ_s. Consequently, more of the area in a 3-D system is filled with computation units and interconnect. (The relative ratio of the two latter is given by σ. A lower σ would reserve more of the area to computation.)

Figure 2.7b shows how the area not covered by memory, is used for computation in 2-D and 3-D topologies. For $\mu_s = 2{,}000$ we can afford 684 operators in a 2-D system, while we can squeeze in 24,551 operators in a 3D16 system. The reason 35 times more operators fit into the same area is mainly due to the much higher density of DRAM as opposed to SRAM that is common in 2-D based systems. This naturally translates to a similar increase of performance as Fig. 2.7c illustrates. It also results in a prohibitively high power consumption since the computation consumes much more power than the memory. Apparently, we cannot power all these computations in reality, but we can translate the increased potential that 3-D offers into either smaller chips, or lower frequency, or higher memory content.

Figure 2.8 shows performance and power consumption for a smaller system (100 mm²) clocked at a somewhat lower frequency and at the 35 nm technology node. With $\mu_s > 4{,}000$ we get a practical power consumption and still a very respectable tera-scale performance.

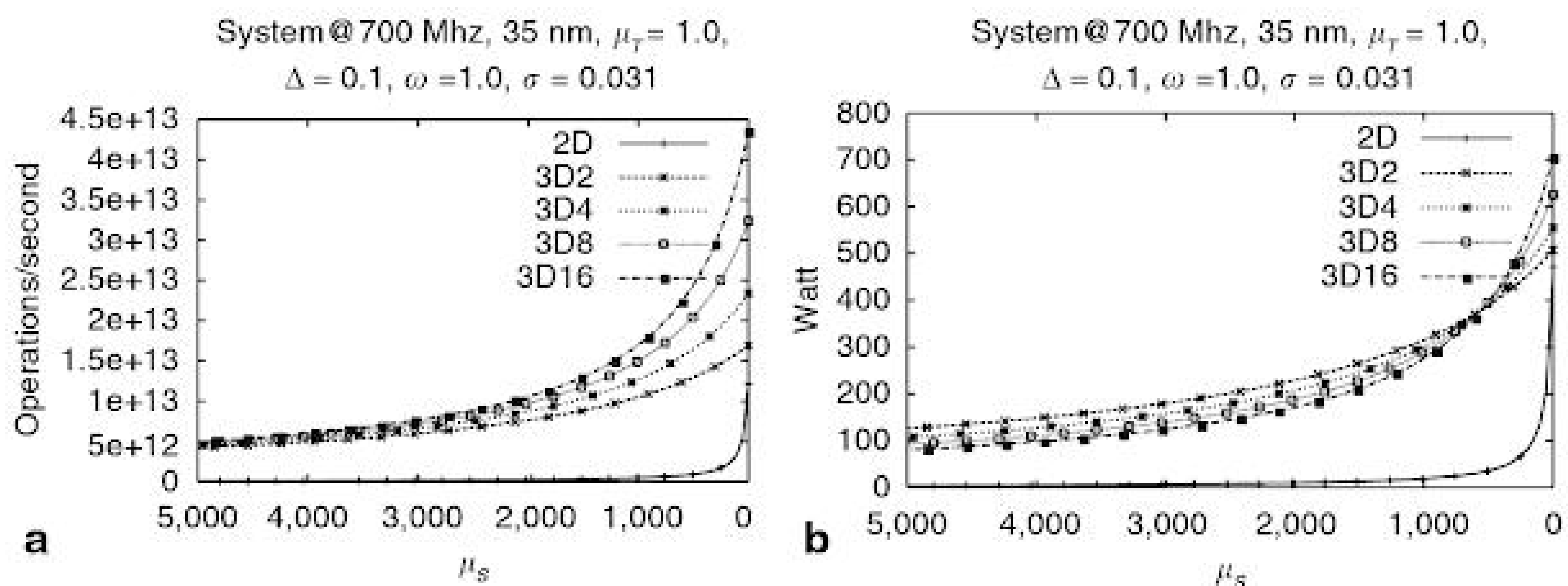

Fig. 2.8 Performance and power consumption for a concrete 100 mm² system with $\mu_T = 1.0$, $\sigma = 0.031$, $\omega = 1.0$ and $\Delta = 0.1$, clocked at 700 MHz, with a 35 nm technology. **a** Operations per second. **b** Power consumption

Bahan dengan hak cipta

image
not
available

15. S. Perri, P. Corsonello and G. Staino. A low-power sub-nanosecond standard-cells based adder. *Proceedings of the 2003 10th IEEE International Conference on Electronics, Circuits and Systems*, 2003.
16. W.J. Dally, U.J. Kapasi, B. Khailany, J.H. Ahn and A. Das. Stream processors: progammability and efficiency. *Queue*, 2(1):52–62, 2004.
17. U. Nawathe, M. Hassan, K. Yen, L. Warriner, B. Upputuri, D. Greenhill, A. Kumar and H. Park. An 8-Core 64-Thread 64b Power-Efficient SPARC SoC. *Proceedings of the Inetrnational Solid State Circuits Conference*, 2007.

Bahan dengan hak cipta

image
not
available

image
not
available

er, high-yielding dies in possibly different, dedicated process technologies [8]. It utilizes the vertical dimension for stacking thinned dies, thereby creating a smaller footprint as well as a higher transistor density per volume unit. And the TSV-based interconnects between the various dies have in number, speed, and power dissipation more similarity to on-chip wires than to off-chip wire bonds [9, 5]. On-chip interconnect wire length, especially of the global and semi-global wires, can be reduced drastically [10].

The process technology for 3D-SICs based on TSVs is becoming available now, design support tools are emerging [11], and first products start to appear on the market. Among the early applications are CMOS image sensors stacked on their corresponding read-out and digital processing circuitry, and stacks of memories [12]. Memory-on-logic [13] and logic-on-logic [14] are expected to follow suit.

Through-Silicon Vias provide, as their name indicates, an electrical connection from the active front-side ('face') of a silicon die through the silicon substrate to the back-side. This allows to interconnect multiple vertically stacked dies with each other. Below, we describe one of the TSV types developed at IMEC. These TSVs are cylindrical copper nails of 25 μm height, 5 μm diameter (i.e., an aspect ratio of 5:1), and have a minimum pitch of 10 μm [15, 16]. The TSV fabrication steps are depicted in Fig. 3.2 and consist of (1) deep silicon etching of TSV holes, (2) oxide deposition, (3) copper seed deposition, (4) copper plating, and (5) chemical-mechanical polishing (CMP).

As can be seen in Fig. 3.2, when the wafer processing is completed, the TSVs are still deeply buried within the wafer; their height is only 25 μm, while the total wafer thickness is around 750 μm. To expose the TSV tips, the wafer needs to be thinned down from the back-side to just below 25 μm thickness. In order to provide sufficient mechanical strength and prevent it from breaking or cracking, the to-be-thinned wafer is temporary bonded onto a carrier wafer, prior to thinning [17]. Subsequently, the thinned product wafer on its carrier wafer is permanently bonded to the next die, after which the temporary carrier wafer is removed. The thinning and bonding steps are depicted in Fig. 3.3. This process can be repeated in case more than two dies are stacked.

There are many different variants of TSV types, process steps, and stacking options [3, 18, 19]. Below, we describe some of them, in as far as relevant for the remainder of this chapter.

The TSVs we described above have a 5 μm diameter, 25 μm height, and 10 μm minimum pitch; TSVs can also be larger or smaller. Taller TSVs cause less problems with respect to wafer thinning, but have a larger aspect ratio and hence are more difficult to properly fill [20]. That is, unless we also make the TSV diameter larger, but that has a negative impact on the silicon area they consume as keep-out

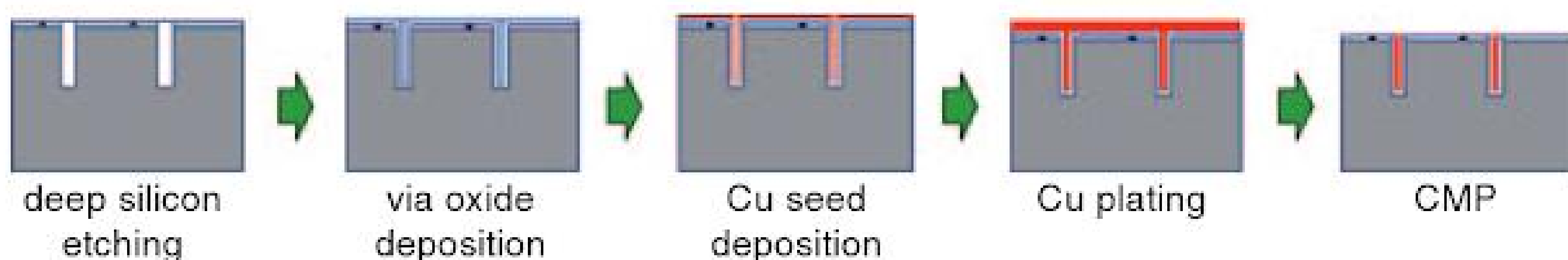

Fig. 3.2 Subsequent TSV fabrication steps

Bahan dengan hak cipta

image
not
available

image not available

package test. The wafer test typically serves merely as an economical optimization to prevent unnecessary package and packaging costs of dies that can be identified as faulty already during wafer test. Hence, the wafer test contents is typically a large subset of the package test. Whether or not wafer testing should be done for a particular product, depends on the wafer fabrication yield y, the fraction d of faulty products that the test can detect (which is determined by the test quality), the costs p that can be prevented per product (here: the cost of packaging a single die), and the cost t of executing a wafer test on a single product (which is typically directly related to the test quality d). Executing a wafer test contributes to overall cost savings if

$$(1 - y) \cdot d \cdot p > t. \tag{3.1}$$

For many conventional ICs, the benefits of wafer testing indeed exceed its execution cost and hence the wafer test is part of their test flow.

In recent years, a market has grown for delivery of unpackaged dies, either still in their original wafers or already singulated. These unpackaged dies are delivered by one company (the 'die provider') and subsequently used by another company (the 'die user'), for example for embedding in an MCP or SiP. Suddenly, the naked die is no longer an intermediate product; for the die provider, it is the final product. The die user wants the die provider to deliver high-quality products. This changes the role of the wafer test. Instead of consisting of an economically-justified subset of the (final) package test, it *is* now the final test. This has led to the concept of *Known-Good Die* (KGD) tests [22], which should guarantee as much as possible the quality of the outgoing dies. A typical KGD wafer test includes at-speed and burn-in tests, in order to achieve quality levels otherwise known from final tests.

3.3.2 Test Flows for 3D-SICs

For 3D-SICs, we distinguish between (1) *pre-bond* tests, (2) *post-bond* tests, and (3) the package test [23]. These tests are depicted in Fig. 3.6b. Only the package test is a test on packaged products; all other tests are tests of naked dies and die stacks. As handling and testing individual dies is rather cumbersome, we assume here that all pre-bond and post-bond tests are wafer-level tests. We distinguish between the pre-bond and post-bond wafer tests, as they are distinctly different, not only in content and purpose, but also in test access. For pre-bond tests, each die requires its own probe access points, while in post-bond tests, all test data is assumed to be pumped in and out through the bottom die of the stack.

The package test serves the purpose of guaranteeing the outgoing product quality of the final product. Depending on the relationship and agreements between the die provider(s) and die user, the pre-bond and post-bond tests might have different roles. In the case of a vertically-integrated 3D-SIC maker, wafer tests have the role of economic optimizers only. However, in case die provider and die user are distinct companies, pre-bond and/or post-bond wafer tests might have the role of final test

Bahan dengan hak cipta

image
not
available

image
not
available

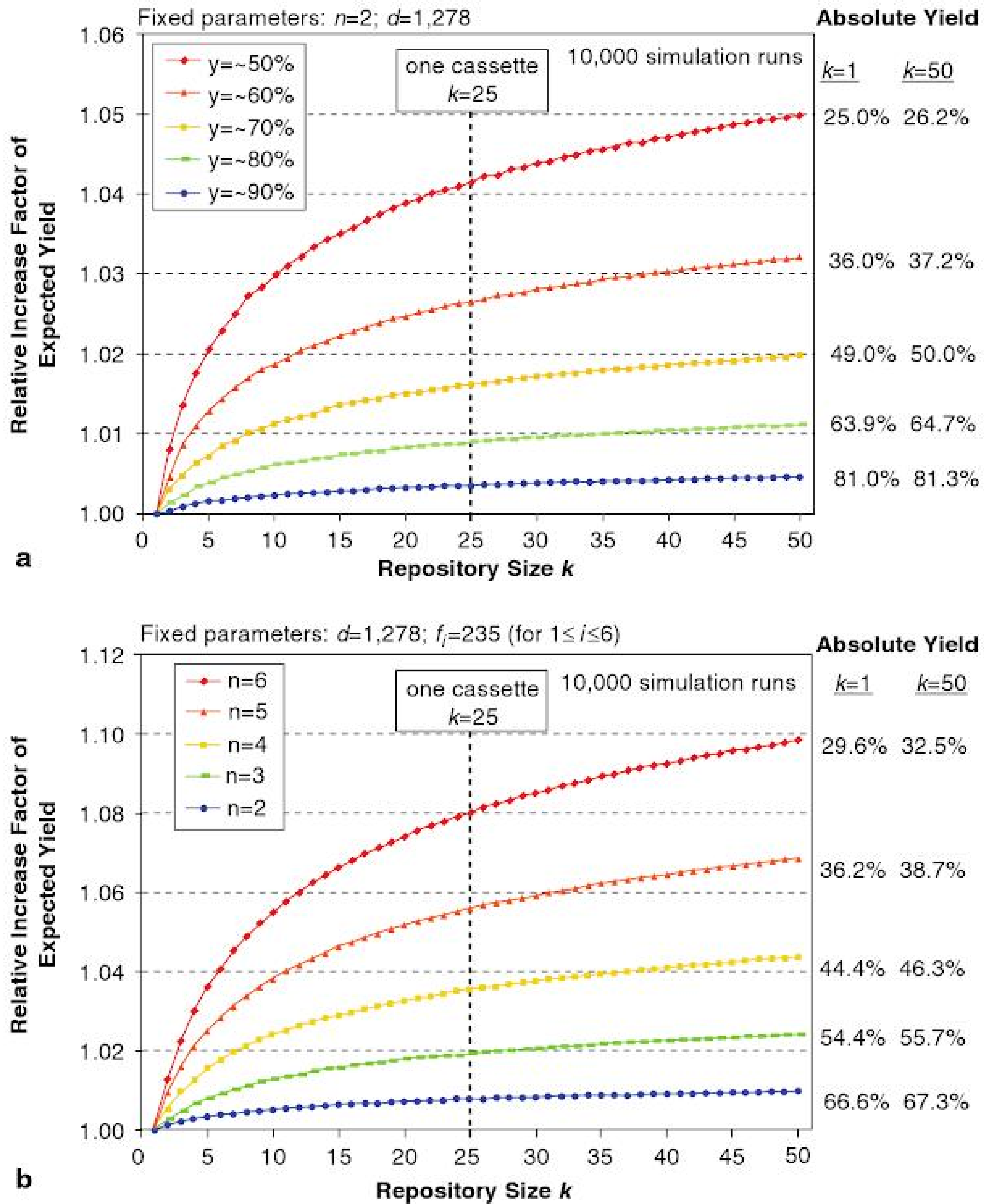

Fig. 3.8 Expected yield increases for an example 3D-SICs with **a** varying die yield y and fixed stack size $n = 2$ and **b** varying stack size n and fixed die yield $y = 81.6\%$ [26]

cantly reduces the resulting test data volume as well, as every module gets only the test patterns it requires, instead of the SoC-wide maximum [30].

- Modular testing allows for easy reuse of tests, throughout the life-cycle of one SoC, and also in subsequent derivate designs that reuse the same design module.

The modular test approach is especially suited to 3D-SICs. All four arguments listed above for modular SoC testing apply even stronger to the case of 3D-SICs. In

Bahan dengan hak cipta

image
not
available

image
not
available

image
not
available

3.6 3D Wafer Test Challenges

Compared to a conventional 2D test flow, a 3D test flow contains potentially many more wafer tests for carrying out pre-bond and post-bond tests on intermediate product stages, which prevent downstream processing of bad products and hence are necessary to keep the overall production cost down. Pre-bond and post-bond wafer tests both have their specific challenges with respect to probe access.

3.6.1 Pre-Bond Wafer Test Access

For pre-bond tests, the main wafer test access challenges stem from probing on very many, very small probe points, possibly on thinned wafers on a temporary carrier wafer.

Today's probe technology, using either cantilever or vertical probes, goes down to a minimum pitch of about 35 μm [43], has a maximum probe count of several hundreds, and makes significant scrub marks in order to achieve a proper electrical contact. This is insufficient to probe on TSV tips of 5 μm diameter and 10 μm pitch (or smaller), which might come in several thousands (the 10 μm pitch allows TSV densities up to 10 k/mm^2), are made of fragile copper, and do not tolerate scrub marks that inhibit downstream Cu–Cu bonding on the same surface. Probing on Cu–Sn micro-bumps is also challenging, but nevertheless a bit easier, as the sizes and pitches of the micro-bumps are larger, consequently their numbers smaller, and the constraints on scrub marks less strict.

For pre-bond testing of the various dies that will make up a 3D-SIC, we distinguish between the bottom die and the other (non-bottom) dies. The non-bottom dies receive all their functional signals (power, ground, clocks, control, data) exclusively through TSV connections, and hence only possess I/Os which cannot be probed with today's probe technology. The bottom die is different, in the sense that, next to its TSV connections, it also has wire-bond or flip-chip pads for the extra-connect, which provide an interface probe-able with today's probe technology and which allows us to get test data in and out of that die and test it. If we want to execute pre-bond tests on the other, non-bottom dies, new solutions need to be developed. The following solution approaches are being explored.

- *Additional probe pads*
  Providing dedicated additional probe pads (as a form of DfT) at the side to be probed and sized such that today's probe technology can handle them. Obviously, this comes with an area penalty, and hence the number of extra pads should be minimized; on-chip DfT can help with this (see Sect. 3.7).
- *Probe technology improvement*
  Significant improvement of wafer probe technology, scaling down to (in the order of) 25 μm pitch for micro-bump probing or down to (in the order of) 10 μm pitch for TSV tip and TSV landing pad probing, while at the same time increasing the maximum probe count and reducing the scrub mark damage.

Bahan dengan hak cipta

image
not
available

3.6.2 Post-Bond Wafer Handling

In post-bond testing, the wafer test access is typically through the regular functional pads of the bottom die of the stack. This is largely 'business-as-usual' as far as probe technology is concerned. The challenges for post-bond testing are in the wafer handling within the probe station, especially in the case of D2W stacking. On top of the bottom wafer, stacks consisting of one or multiple dies stick out. If that top side is the probe side of the wafer, the stacks might obstruct the contact view, making probe needle positioning difficult. Also, during probe needle movement, we need to make sure that the needles do not collide with the stacks. On the other side, if we probe on the bottom side of the wafer, the stacks create a very non-planar surface, which is difficult to keep stable on the chuck. It is still a matter of on-going research to resolve these issues.

3.7 3D On-Chip DfT Architecture

3.7.1 Hierarchical DfT Requirements

The primary role of on-chip Design-for-Testability (DfT) is to provide controllability and observability from the chip I/Os into the heart of the chip design and vice versa. We consider a hierarchical 3D-SIC consisting of multiple stacked dies, which each consist of one or multiple 'test modules' (see Sect. 3.4). DfT requirements exist at every level of the design hierarchy.

1. *Core-Level DfT*
   DfT within the test modules, e.g., internal scan chains, Test Data Compression (TDC) hardware [48], and/or Built-In Self-Test (BIST).
2. *Die-Level DfT*
   DfT to enable modular testing per die, i.e., wrappers around the test modules and TAMs [29]. Wrappers based on IEEE Std. 1,500 [32] provide controllability and observability at the module boundary, for both inward-facing tests (*InTest*) as well as outward-facing tests (*ExTest*). The TAM architecture can be optimized for test bandwidth vs. test length trade-off [33].
3. *SIC-Level DfT*
   This level constitutes the new DfT for TSV-based 3D-SICs. It includes the following.

   – A wrapper at the die boundary, to support modular testing per die, and allow for both InTest (testing the dies and cores) and ExTest (testing the TSV-based interconnect in between the dies). The wrapper cells could either be based on IEEE Std. 1,149.1 or IEEE Std. 1,500 (see Fig. 3.14). Especially if dies are designed by different teams or companies, and not necessarily 'friends' at their interfaces, it is probably a wise thing to equip them with ripple-protected wrapper cells.

Bahan dengan hak cipta

image
not
available

- **Post-Bond Wafer and Package Tests**
  For all post-bond wafer and package tests, test access is via two scan chains which are multiplexed onto already-existing functional pads on the front-side of the bottom die, viz. D1/SI1–Q1/SO1 and D2/SI2–Q2/SO2.

  4. A mode in which a post-bond wafer or package test can be applied to the bottom die (Fig. 3.16d).
  5. A mode in which a post-bond wafer or package test can be applied to the top die (Fig. 3.16e).
  6. A mode in which a post-bond wafer or package test can be applied to the TSV-based interconnects between bottom and top die (Fig. 3.16f).

- **Board-Level Test**

  7. A mode in which a board-level test is executed (Fig. 3.16g). Test access is via one boundary scan chain along the extra-connect interface of the bottom die which connects to dedicated front-side pads, viz. TDI–TDO (part of the IEEE 1,149.1 Boundary Scan standard).

3.7.3 Advanced DfT Techniques and Test Resource Partitioning

The following advanced DfT techniques play a special role in the context of 3D-SICs.

- *Reduced Pad-Count Testing (RPCT)*
  RPCT is a DfT technique to reduce the width of a scan test interface [53]. As discussed in Sect. 3.6, dedicated additional probe pads might need to be provided to execute a pre-bond test on the bottom die *after* thinning or to execute a pre-bond test on other, non-bottom dies. These extra probe pads are costly in area. RPCT can be exploited to reduce the number of extra pads required. Note that the usage of RPCT does not affect the total test data volume, which implies that a reduced test interface width comes at the expense of an increased test length.
- *Test Data Compression (TDC)*
  TDC is a DfT technique that exploits the many 'don't care' bits in ATPG-generated test patterns to compress the off-chip test data in a (near-) lossless manner. It allows to reduce the volume of both test stimuli and responses (and the corresponding test length) with one or two orders of magnitude [48]. TDC can play a role in reducing the test data volumes of 3D-SICs. When applied per test module, the TAMs that provide these test modules with stimuli and responses can be scaled down. Also, RPCT and TDC form an attractive DfT combination, as the first scales down the test interface width, while the second prevents the test length from increasing.

Bahan dengan hak cipta

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.

You have either reached a page that is unavailable for viewing or reached your viewing limit for this book.