IPSJ Transactions on System LSI Design Methodology Vol.5 2–13 (Feb. 2012)

[DOI: 10.2197/ipsjtsldm.5.2]

Invited Paper

A Stackable LTE Chip for Cost-effective 3D Systems

Walid Lafi[1,a) Didier Lattard[1,b) Ahmed Jerraya[1,c)

Received: September 1, 2011, Released: February 21, 2012

Abstract: To address the problem of prohibitive cost of advanced fabrication technologies, one solution consists in reusing masks to address a wide range of ICs. This could be achieved by a modular circuit that can be stacked to build TSV-based 3D systems with processing performance adapted to several applications. This paper focuses on 4G wireless telecom applications. We propose a basic circuit that meets the SISO (Single Input Single Output) transmission mode. By stacking multiple instances of this same circuit, it will be possible to address several MIMO (Multiple Input Multiple Output) modes. The proposed circuit is composed of several processing units interconnected by a 3D NoC and controlled by a host processor. Compared to a 2D reference platform, the proposed circuit keeps at least the same performance and power consumption in the context of 4G telecom applications, while reducing total cost. More generally, our cost analysis shows that 3D integration efficiency depends on the size of the circuit and the stacking option (die-to-die, die-to-wafer and interposer-based stacking).

Keywords: TSV-based 3D systems, cost model, 4G telecom applications

1. Introduction

Today's complex SoC designers are facing several problems that limit the economic benefits of advanced technology nodes. In addition to the prohibitive cost of masks (which has already exceeded 1 million euro according to the ITRS), wafer fabrication is becoming more and more expensive mainly due to huge circuit size that reduces manufacturing yield. A good solution to develop economically competitive products is to reuse masks to address a wide range of systems and to fabricate small-sized circuits to increase yield. To do so, our proposal is to design a modular circuit that could be stacked using 3D integration technologies in order to build 3D systems with processing performance adapted to several application requirements. This type of circuits is referred to as 3D same-die stacked architectures in this thesis.

In this work, we focus on modular architectures for 4G telecom applications, which are the latest standard in the mobile network technology with important performance requirements. In the 4G standard, the MIMO mode of transmission may be used to enhance either robustness or throughput within wireless communications. In this work, we propose a reconfigurable circuit that meets the SISO (Single Input Single Output) mode of transmission (1 antenna) in a stand-alone. By stacking multiple instances of this same circuit, it would be possible to boost overall system performance and address several MIMO (Multiple Input Multiple Output) modes.

The reminder of this paper is organized as follows. Section 1 explains the motivations for 3D integration and describes some of its technological options. Section 2 presents the reconfigurable and stackable circuit, and performs a case study of the 4×2 LTE mode of transmission with performance and power evaluation. Finally, Section 6 presents a cost analysis of the proposed circuit.

2. 3D Integration Technologies

3D integration is to stack many circuits vertically and interconnecting them using different technologies such as inductive coupling [1] and Through-Silicon-Vias (TSVs). In this work, we focus on the TSV technology. This results in smaller circuit footprint and shorter vertical interconnections, which improves system performance and power. Besides, heterogeneous systems can be built easily, since each layer can support diverse technology. **Figure 1** depicts an example of a 3D system. It is composed of several dies and a 3D chip, stacked on top of an interposer. The silicon interposer may be fabricated in a mature technology such as 130 nm. The interposer may be active (including active components such as network-on-chip routers) or passive (containing only wires to ensure interconnection between the stacked dies). The stacked dies have different sizes and different functionalities. They are fabricated in aggressive technology such as 28 nm. In the case of passive interposer, these dies may be set side by side and interconnected by a very large number of connections in order to provide high inter-die interconnect bandwidth [2]. By using both TSVs and solder bumps, it is possible to mount the interposer-based stack (IbS) on a package substrate using classic flip-chip assembly techniques (Fig. 1). The coarse-pitch TSVs provide the connections between the package and the interposer for the parallel and serial I/O, power/ground, clocking, data signals... Interposer-based stacking allows avoiding the reliability issues that can result from stacking multiple dies (fabricated on an aggressive technology) on top of each other.

[1] CEA-LETI, Grenoble, France
[a) walid.lafi1@gmail.com
[b) didier.lattard@cea.fr
[c) ahmed.jerraya@cea.fr

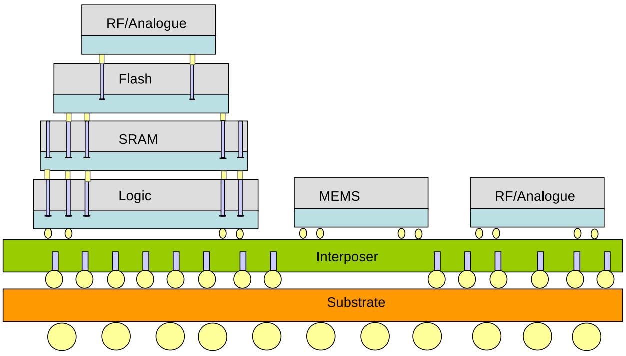

Fig. 1 An example of a 3D system.

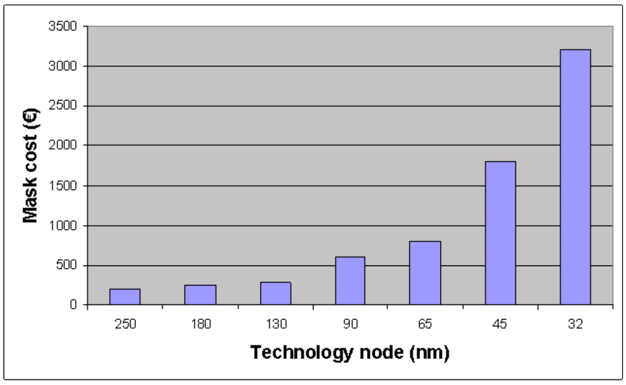

Fig. 2 Cost of a mask set according to technological nodes.

2.1 Reducing Cost

2.1.1 Heterogeneous 3D Integration

Present SoCs usually integrate heterogeneous functions (digital, memories, DSPs, analog and RF). These functions are initially designed for different manufacturing technologies. Although it is possible to fabricate all these devices on a single die using the same technology, this would be suboptimal in terms of performance, area, and power. Besides, this further complicates the fabrication process and increases manufacturing cost. Indeed, advanced digital technologies are not well adapted to realize functions such as analog or RF circuits. Past attempts to converge these different functions onto a single monolithic circuit resulted in many issues related mainly to cost and performance. For example, in a RF-CMOS process, the total price of a final wafer exceeds that of pure CMOS by more than 15% [3]. It is preferable then to make each function in its own mature technology node in order to get higher performance and lower cost.

A significant advantage of 3D integration is the possibility to integrate heterogeneous technology dies built with different processes on the same 3D circuit. This means manufacturing independently different functions such as analog, digital, or memory and integrating them in the same final system. It is then possible to manufacture each type of circuit using the most adapted technology [4], [5], [6].

2.1.2 Same-die 3D Stacking

Currently, a general VLSI application without regular system architecture requires multiple sets of masks. This can be extremely expensive since mask prices for cutting-edge processes have been increasing steadily (**Fig. 2**). According to the ITRS, the cost of only one mask set has already exceeded one million euro. For this reason, reusing mask and reducing the number of mask layers are becoming highly recommended.

In order to develop cost-competitive products, a potentiel solution is to reuse masks to address a wide range of systems. To do so, it is possible to design a modular circuit that could be stacked using 3D integration technologies to build 3D systems with processing performances adapted to several application requirements. Therefore, it would be possible to design several different systems using always the same mask set, thanks to 3D integration technology. Stacking many instances of the same circuit is referred to as same-die 3D stacking in this thesis.

Heterogeneous 3D integration and 3D same-die stacking approaches are not conflicting, but complementary. Indeed, it is possible to design a 3D SoC that includes several functions such as digital, memory, RF, analogue... Thanks to the 3D heterogeneous integration approach, each function may be fabricated using the most adapted technology in order to get better cost-performance tradeoffs. The 3D same-die stacking approach could be used for the digital part to boost the computational performance of the 3D system as needed by the targeted application (**Fig. 3**).

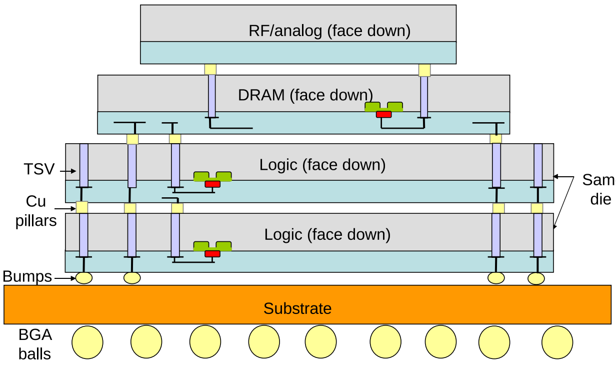

Fig. 3 A 3D SoC designed with heterogeneous integration and same-die stacking approaches.

2.2 Enhancing performance and form factor

One of the most obvious advantages of 3D integration is to replace long horizontal wires with short vertical interconnects (TSV-Trough Silicon Via). **Figure 4** illustrates the overall reduction of interconnections. The global inter-block wiring in 2D circuits (the longest wires) are replaced here by short vertical interconnections.

These shorter wires will decrease the average load capacitance and resistance (wire's capacitance and resistance are proportional to wire length) and reduce the number of repeaters needed by long wires. Since interconnect wires with their supporting repeaters consume a significant portion of total active power, the average interconnect length reduction in 3D IC will significantly reduce overall power consumption [7].

Moreover, shorter interconnects in 3D ICs (with consequent reduction of load capacitance and reduced numbers of repeaters) will reduce the noise resulting from simultaneous switching events and coupling between signal lines. This should provide better signal integrity.

Another major consequence of the reduced wire resistance and

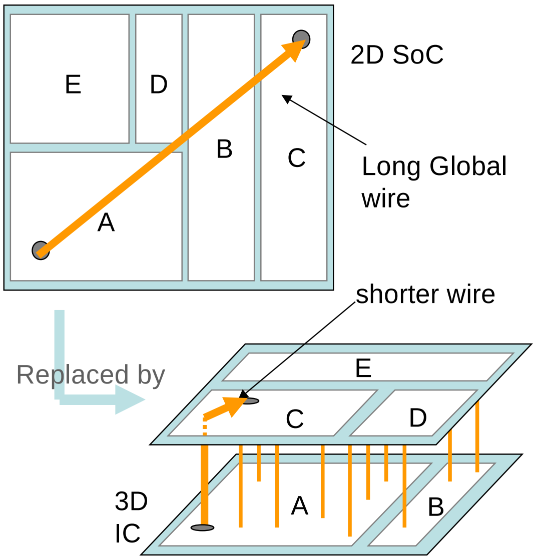

Fig. 4 Interconnects' length shortening in 3D ICs.

capacitance is the significant reduction of signal propagation delay (proportional to the product resistance times capacitance), which results in significant system performance gain.

As shown in Fig. 4, 3D integration technologies allow reducing chip area and thus enhancing form factor. Therefore, it would be possible to continue chip miniaturization without necessarily following Moore's Law.

In conclusion, 3D stacking allows having shorter global interconnects within the 3D circuit, and thus reducing its total active power, coupling noise, and signal propagation delay . Besides, 3D integration technologies allow enhancing the circuit form factor.

2.3 3D Circuit Manufacturing Technologies

A 3D integrated circuit may be fabricated according to several technological options [8]. A critical issue is to choose the way to assemble chips (**Fig. 5**):

- Die-to-die (D2D): This approach requires a stringent alignment effort since it seems difficult to handle small dies. Besides, chips' assembly is time consuming (Pick and Place).
- Wafer-to-wafer (W2W): In this case, the time necessary for chips' assembly is much shorter. Further, alignment is easier since assembly is performed on bigger objects. All dies of on these stacked wafers must have the same size to be separated after assembly.
- Die-to-wafer (D2W): The time needed for stacking is less significant than in the D2D option. Alignment issue is also less critical than in the die-to-die approach. Another advantage of this technique is that the stacked chips can be different sizes: it is possible to stack a small-sized circuit on top a larger circuit. Although the D2W approach is more expensive the W2W approach, it could be used when circuit fabrication and stacking are not performed by the same foundry.

To achieve 3D stacking, some key technological steps must be fully involved: wafer thinning, wafer bonding and TSV forming.

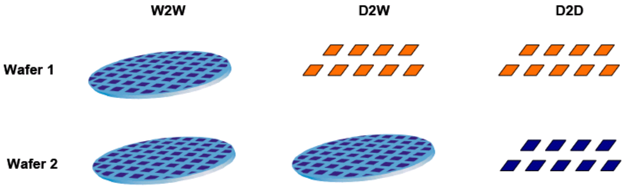

Fig. 5 3 ways to assemble the circuits vertically.

2.4 A System-level Cost Model for 3D ICs

3D integration is made in several steps, each of which includes a wide range of technological choices. Choosing the optimal process flow depends mainly on cost. Therefore, cost estimation of 3D ICs in the early stages of the design cycle is so important. This section deals with 3D ICs cost challenges.

Several works have focused on 3D integration cost analysis. Their approaches range from system level to technology detailed assessment. X. Dong and Y. Xie present a system-level analysis for 3D ICs [9]. Based on Rent's rule, they perform an estimation of the number of wires inside a 2D chip, and deduce the number of TSVs within the resulting 3D stack after partitioning. Then, they propose parameterized cost models, which take into account several aspects of 3D ICs such as bonding yield, known-good-die test, assembly options... Based on these models, some important trends about 3D ICs cost are found out. This helps the authors of [9] to propose a cost driven design flow for 3D ICs.

In this work, we propose a system-level cost analysis model that allows having a preliminary cost estimation of a 3D system, and deciding on the best options to choose in order to optimize cost. The proposed model allows making comparison between a 2D system and its 3D versions in terms of cost. We focus on 2 stacking approaches: the W2W, the D2W. The D2D approach is not investigated as it is not widely used by foundries. As depicted in **Fig. 6**, the D2W approach comes in 2 versions: the first one is to stack active dies on top of each others, and the second one is to stack the active dies on top of a passive interposer. The second version is called the interposer-based stacking (IbS). The IbS approach is to stack several dies (fabricated in an aggressive technology such as 32 nm) on top of a silicon interposer (fabricated in a mature technology such as 130 nm) in order to improve fabrication yield. When using the W2W approach, it is necessary that the stacked dies have the same size. Otherwise, it would be impossible to slice the wafer to obtain the final 3D circuits. This problem does not arise when using the D2W or the IbS approach. In this comparison, we assume that all the stacked dies are different (each die has its own mask set).

3D integration requires extra-fabrication including TSV forming, wafer/die thinning, and wafer/die bonding. In order to separate the die cost model and the 3D stacking cost model, we assume TSV-last approach is used in 3-D IC fabrication process. In order to support multiple-layer stacking, the chosen stacking mode is F2B. In addition, the entire 3D-stacked chip cost depends on whether die-to-wafer (D2W), wafer-to-wafer (W2W) or interposer-based (IbS) stacking is used. If D2W or IbS approaches are selected, cost of known-good-die (KGD) test should also be included. To get rid of fabrication details common to 2D and 3D, we assume that the wafer fabrication cost is constant for

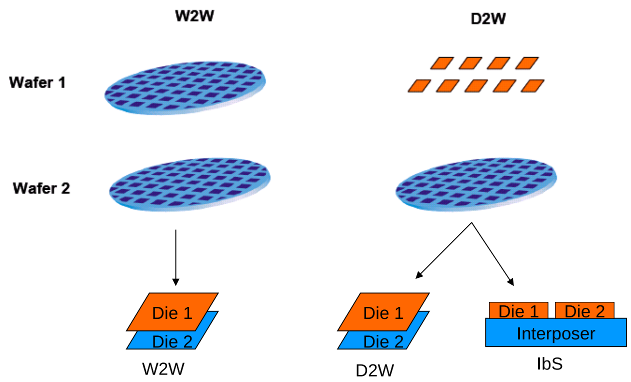

Fig. 6 The investigated stacking approaches.

a specific foundry using a specific technology node. The number of dies per wafer may be approximated by:

$$N_{die/wafer} = \frac{\pi \times (D_{wafer}/2)}{A_{die}}$$

where D_{wafer} is the wafer diameter and A_{die} is the die area. Hereafter, we present our cost models for the 2D and the different 3D approaches.

2.4.1 2D IC Cost Model

The final 2D IC cost my be given by:

$$C_{2D} = \frac{C_{fab} + C_{test}}{Y_{2D}}$$

where:

- C_{fab} is the 2D IC fabrication cost,
- C_{test} is the final test cost,
- Y_{2D} is the fabrication yield.

The fabrication yield and the 2D IC fabrication cost may be given by:

$$Y_{2D} = (1 + \frac{A_{2D} \times D_0}{\alpha})^{-\alpha}$$

$$C_{fab} = \frac{C_{wafer}}{N_{die/wafer}} + \frac{C_{mask}}{N}$$

where:

- A_{2D} is the 2D IC area,
- C_{mask} is the mask cost,
- C_{wafer} is the wafer fabrication cost for a specific foundry using a specific technology node,
- N is the total number of 2D ICs (or total production volume),
- D_0 is the density of point-defects per unit area,
- α is a model parameter, and typically ranges from 1.0 to 5.0.

2.4.2 3D IC Cost Model

$C_{3D,W2W}$, $C_{3D,D2W}$ and $C_{3D,IbS}$ are the final 3D IC costs when using the wafer-to-wafer, the die-to-wafer and the interposer-based stacking approaches respectively. They are given by:

$$C_{3D,W2W} = \frac{\sum_{i=1}^{L} C_{die_i} + (L-1) \times C_{stacking,W2W} + C_{test,W2W}}{(Y_{stacking,W2W})^{L-1} \times (\prod_{i=1}^{L} Y_{die_i})}$$

$$C_{3D,D2W} = \frac{\sum_{i=1}^{L} (C_{die_i} + C_{test,die_i})/Y_{die_i} + (L-1) \times (C_{stacking,D2W} + C_{test,stacking,D2W})}{(Y_{stacking,D2W})^{L-1}}$$

$$C_{3D,IbS} = \frac{\sum_{i=1}^{L} (C_{die_i} + C_{test,die_i})/Y_{die_i} + (C_I + C_{test,I})/Y_I + L \times (C_{stacking,IbS} + C_{test,stacking,IbS})}{(Y_{stacking,IbS})^{L}}$$

where:

- $C_{stacking,W2W}$, $C_{stacking,D2W}$ and $C_{stacking,IbS}$ are the stacking costs (including all the steps of the 3D fabrication process) when using the wafer-to-wafer, the die-to-wafer and the interposer-based stacking approaches respectively,
- $Y_{stacking,W2W}$, $Y_{stacking,D2W}$ and $Y_{stacking,IbS}$ are the stacking yields when using the wafer-to-wafer, the die-to-wafer and the interposer-based stacking approaches respectively
- $C_{test,W2W}$ is the final test costs of the 3D IC when using the wafer-to-wafer approach,
- $C_{test,die}$ and $C_{test,I}$ are the costs of testing the die and the interposer respectively,
- $C_{test,stacking}$ is the cost of testing the the vertical interconnections (TSVs or bumps),
- C_{die} and C_I are the fabrication costs of the die and the interposer respectively,
- Y_{die} and Y_I are the fabrication yields of the die and the interposer respectively,
- L is the number of stacked dies,
- A_{3D} is the final die area (including TSV area overhead),
- A_{TSV} is the TSV area overhead,

C_{die} and C_I are given by:

$$C_{die} = \frac{C_{wafer}}{N_{die/wafer}} + \frac{C_{mask}}{N}$$

$$C_I = \frac{C_{wafer}}{N_{I/wafer}} + \frac{C_{mask}}{N}$$

Y_{die} and Y_I are given by:

$$Y_I = \left(1 + \frac{A_I \times D_0}{\alpha}\right)^{-\alpha}$$

$$Y_{die} = \left(1 + \frac{A_{3D} \times D_0}{\alpha}\right)^{-\alpha}$$

A_{TSV} is given by:

$$A_{3D} = \frac{A_{2D}}{L} + A_{TSV}$$

2.4.3 Test Cost Model

According to [10], test cost may be modelled as the product of the cost of tester use per second and the average IC test time. Tester-use cost (per die) may be then given by:

$$C_{test} = R.T_{test}$$

where R is the cost rate (euros per second) for a tester, and T_{test} is the average IC test time. According to [10], test execution time may be considered as proportional to the die area. Besides, the average test time may be considered as depending on yield, because test time is shorter for a failing die than for a good die. Indeed, testing usually terminates upon first failures. As a result, the average time required to test a single IC is:

$$T_{test} = T_{setup} + [Y + \beta(1 - Y)]K.A$$

where T_{setup} is the setup time for an IC on the tester, β is the average ratio between good-die test time and defective IC test time, and K is a constant multiplier that relates test time to IC die area A. Both β and K may be extracted based on regression analysis

Table 1 Technological parameters for 32 nm-technology die.

Parameter	Value
α	1
Defect density	2.10^{-2}/mm^2
mask cost (Euro)	3,500,000
300 mm Wafer cost (Euro)	8,000

Table 2 Technological parameters for 130 nm-technology interposer.

Parameter	Value
α	1
Defect density	2.10^{-4}/mm^2
Mask cost (Euro)	400,000
300 mm Wafer cost (Euro)	2,000

of historical data on test execution times of various products. A less-than-1 β means that the entire test sequence needs not to be applied to a failing die.

In the case of 3D ICs, test time includes also time required to test vertical interconnects (TSVs). This time may be considered as proportional to the number vertical interconnects per die. Consequently, TSVs test time may be given by:

$$T_{stacking,test} = H.N_{TSV}$$

Where N_{TSV} is the number of TSVs per die, and H is a constant multiplier that relates TSV test time to their number.

2.4.4 3D Stacking Cost Model

Unlike the test cost, it is a complicated task to elaborate a system-level cost model for 3D stacking. This is due to the variety of 3D technological options, which makes the final cost of 3D stacking depending directly on the 3D fabrication process used. Besides, it is quite hard to find realistic information about the cost of the different steps of a particular 3D fabrication process, since these information are rarely published by the industrial community. Our 3D stacking cost model is based on information obtained either from our industrial partner, or from some publications that deals with 3D IC cost analysis such as Ref. [11] of John H. Lau from the Industrial Technology Research Institute of Taiwan.

2.4.5 Implementation of the Cost Model

The previously described cost models were implemented using the Excel tool. Excel is a well-known widely used tool that provides a simple and easy-to-use interface. As it offers all the mathematical functions needed by our cost models, it is simpler and quicker to use compared to standard programming languages (JAVA, Visual Basic...). In order to allow the exploration of the economical trends of 3D integration, the developed Excel sheet may be easily configured by new values for the technological parameters.

2.5 Cost Analysis of 3D ICs

This analysis is based on data of **Table 1** and **Table 2**. The stacking yields for all the 3D approaches are set to 99%, and production volume is set to 1 million.

2.5.1 Cost Variation According to Area

Figure 7 shows the variation of unit cost for a 2D chip and its equivalent W2W, D2W and IbS 3D 2-layer circuits, when die area increases.

As depicted in Fig. 7, for small-sized circuit, the 2D approach is more economical than any of the 3 stacking approaches. For example, considering a 50 mm^2 design, the W2W, D2W and IbS 3D schemes increase cost by 90%, 56% and 120% respectively compared to the 2D approach. This may be explained by the very high yield of small-sized circuits that makes inefficient any further reduction of the circuit size. 3D stacking includes additional technological steps in manufacturing process (and then extra-fees), without significantly improving fabrication yield.

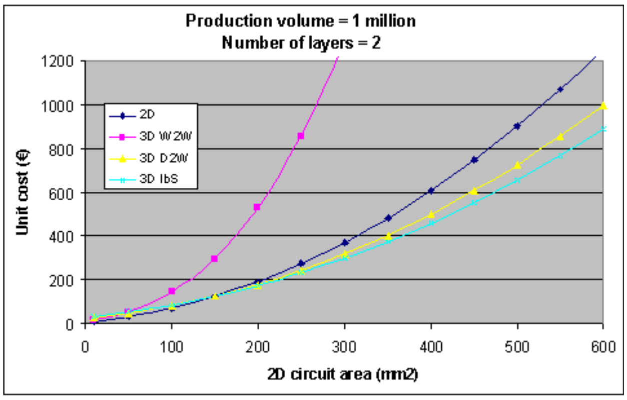

Fig. 7 Unit cost for a 2D circuit and its W2W, D2W and IbS 3D versions for different die areas.

In the case of large-sized design, the D2W and IbS 3D stacking become the most cost-effective approaches. Besides, cost gain increases when the design area increases. As an illustration, the D2W stacking allows reducing cost (compared to the 2D approach) by 12% and 20% when the design area is 300 mm^2 and 600 mm^2 respectively. This can be explained by the low yield of large circuits, which makes so advantageous to partition the 2D circuit into smaller dies, and to test these obtained dies before stacking (known-good die test). However, the W2W scheme remains more expensive than the 2D approach. Indeed, because of low yield (due to large circuit area), die stacking without testing turns out to be economically inefficient.

It could be concluded that 3D stacking involves extra-fees due to additional steps of the 3D fabrication process (such as bonding, TSV formation, known-good-die test...), but also allows cost reduction by reducing the area of the stacked dies and then improving yield. Therefore, the 3D approach is more cost effective in the case of large-sized circuits, when using the D2W or the IbS integration schemes.

2.5.2 Cost Variation According to the Number of Layers

Figure 8 depicts cost variation for different numbers of layers when using the W2W, D2W and IbS 3D schemes.

For the W2W approach, the 3D cost increases considerably when the number of layers increases. For example, considering a 200 mm^2 design, the cost of the W2W 3D IC increases by 2 times, 7 times and 12 times (compared to 2D approach) when the 2D initial design is partitioned across 2, 4 and 6 layers respectively. This is due to low yield of aggressive technologies that makes it indispensable to test the dies before stacking them, especially for large-sized circuits.

For the D2W and IbS approaches, the 3D cost increases with the number of layers in the case of small-sized circuit, due to their high fabrication yield. For the D2W approach, the cost of a 50 mm^2 design increases by 55% for the 2-layer 3D version,

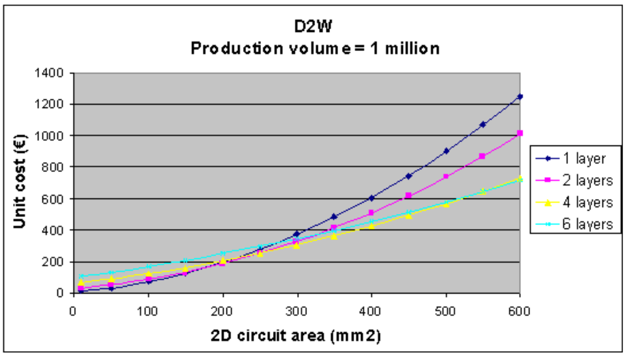

Fig. 8 Unit cost for different numbers of layers using the D2W scheme.

and by 300% for the 6-layer 3D version (compared to the 2D approach). However, in the case of large-sized design, the cost of the D2W or IbS decreases when the number of layers becomes more and more important. When using the D2W approach, the cost of a 600 mm^2 circuit decreases by 20% for the 2-layer 3D version, and by 40% for the 6-layer 3D version (compared to the 2D approach). This could be explained by the low yield of large designs, which makes it so beneficial to partition the circuit into several small dies, and to test them before stacking.

It could be concluded that the optimal number of 3D layers (in terms of cost) depends on how large the design it is.

3. A Stackable Chip for 4G Telecom Applications

3.1 Implementation of 4G Terminals

The 4G standard is the latest standard in the mobile network technology. It is known also as 3GPP LTE: 3rd Generation Partnership Project Long Term Evolution. The 3GPP is collaboration between groups of telecommunications associations that aims to make a globally applicable third-generation (3G) mobile phone system specification within the scope of the International Mobile Telecommunications-2000 project of the International Telecommunication Union (ITU). Long Term Evolution (LTE) is a project of the 3GPP that produces the latest standard in the mobile network technology tree in order to move forward from the cellular 3G services to the 4G services. The main objectives for 3GPP LTE (or 4G) are to increase downlink and uplink peak data rates (100 Mbps for DL with 20 MHz, 50 Mbps for UL with 20 MHz), to improve spectral efficiency (5 bps/Hz for DL and 2.5 bps/Hz for UL), to reduce latency, to improve bandwidth scalability, and to establish a standard's based interface that can support a multitude of user types [12].

The LTE physical layer is in charge of transmitting data and control information between an LTE base station and the user equipment (a mobile phone typically). The LTE physical layer is based on Orthogonal Frequency Division Multiplexing scheme OFDM to meet the targets of high data rate and improved spectral efficiency. OFDM makes use of a large number of closely-spaced orthogonal sub-carriers to carry data. The data is divided into several parallel data streams or channels, one for each sub-carrier. Each sub-carrier is modulated with a conventional modulation scheme (such as quadrature amplitude modulation or phase-shift keying). The modulation schemes supported in the downlink and uplink are QPSK, 16 QAM and 64 QAM. In order to improve communication robustness and throughput, the LTE physical layer supports several Multiple Input Multiple Output (MIMO is the use of multiple antennas at both the transmitter and receiver) options with 1, 2 or 4 antennas.

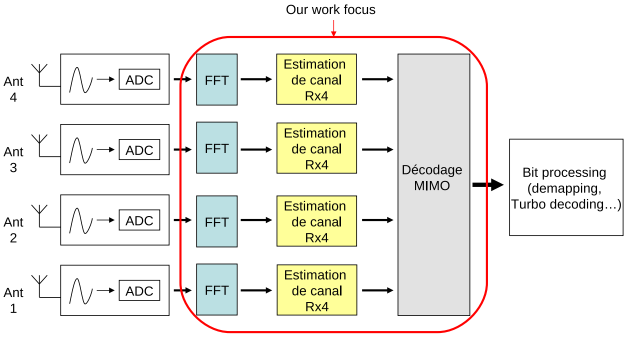

Fig. 9 A functional block diagram of an LTE UE reception chain with 4 receive (Rx) antennas.

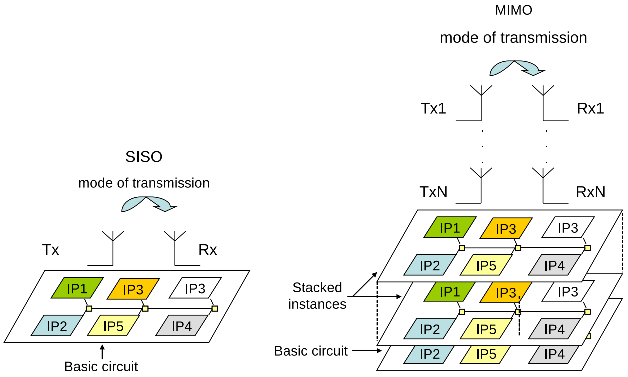

Fig. 10 3D reconfigurable circuit obtained by stacking multiple instances of a same basic circuit.

In this section, we focus only on the baseband processing of the downlink part (reception chain) and omit all other components of LTE user equipment UE such as radio-frequency and analogue functions, higher layer protocols and multimedia processing. **Figure 9** depicts a functional block diagram of the internal data flows of the downlink part within an LTE UE with four receive (Rx) antennas [13]. First, the RF signal is received by the receiver antennas, converted to an electrical quantity, and digitized by an analogue to digital converter (ADC). Then, the baseband processor receives the digitized signal as complex samples and performs OFDM demodulation, channel estimation and finally MIMO decoding. In the following subsections, we will analyse each of these processing steps in order to determine their computational efforts.

3.2 A Stackable Chip for 4G Terminals

In this section, we present a reconfigurable NoC-based circuit for LTE applications. When used alone, the proposed circuit (henceforth called basic circuit) can meet the requirements of the SISO transmission mode. By stacking multiple instances of this basic circuit and performing some software reconfigurations, it will be possible to boost system performance and address several MIMO modes (**Fig. 10**). This section presents the hardware components of the basic circuit such as processing units and the NoC-based communication structure, and provides synthesis results in 65 nm technology.

To be able to satisfy the needs of several telecom applications, the basic circuit has to support data manipulation and data processing at the same time. To do so, three reconfigurable units are designed.

3.2.1 Smart Memory Engine (SME)

7The SME unit is a Micro-programmable Memory Controller (MMC) designed to perform data synchronization and distribution in dataflow systems [14]. The SME allows separating data synchronization from data processing, and thus reducing the complexity of processing units and helping their reuse. A subset of the C programming language and a dedicated compiler are used for flow programming in the SME.

3.2.2 Mephisto Digital Signal Prosessor

The second unit used in the proposed basic circuit is a digital signal processor (DSP) called MEPHISTO [15]. It is a high-performance reconfigurable core designed by the LISAN team to perform complex matrixs computation, useful for channel estimation, advanced MIMO coding/decoding... MEPHISTO is designed as a 32-bit data path Very-Long-Instruction-Word (VLIW) structure composed of a MAC (Multiplier/Accumulator) unit dedicated to complex arithmetic operations, a compare/select unit for branch operations, and a cordic/divider unit for special computations.

3.2.3 OFDM Core

The OFDM core is designed to perform direct and inverse fast Fourier transform (FFT and IFFT). It also incorporates features to achieve a formatting of incoming OFDM symbols (framing i.e insertion of pilots and zeros) and a separation of outgoing pilot and data symbols (deframing). Therefore, this block can be used for both OFDM transmission (framing + IFFT + inserting guard interval) and reception OFDM (FFT + deframing).

3.2.4 MIPS-based Semi-distributed Control

The previously described units require an efficient control mechanism to deal with scheduling and configuration. In this work, we use a semi-distributed control for the whole basic circuit. This allows alleviating the load of the host processor. In addition to the local configuration and communication controller performed by the NI, a global control is performed by a 32-bit MIPS processor, by means of direct addressing and interrupts mechanisms. The MIPS is chosen for its compactness. It is in charge of dynamic reconfigurations, real time scheduling and synchronizations. As depicted in **Fig. 11**, the MIPS processor has several extensions useful to interact and communicate with other basic circuit's components. These extensions include:

- an output extension managing the generation of data and configuration packets from the MIPS,
- an input extension allowing to read (to dump) data and configuration values from any of the circuit's units at the request of the MIPS,
- an interrupt controller in charge of handling interrupts generated in the NoC such as end-of-task notifications,
- and finally a local 16 KB RAM (32-bit word) used to store both instructions (the embedded control software), and data (configurations). An arbiter is used to allow the NoC to write the embedded control code in the MIPS's memory.

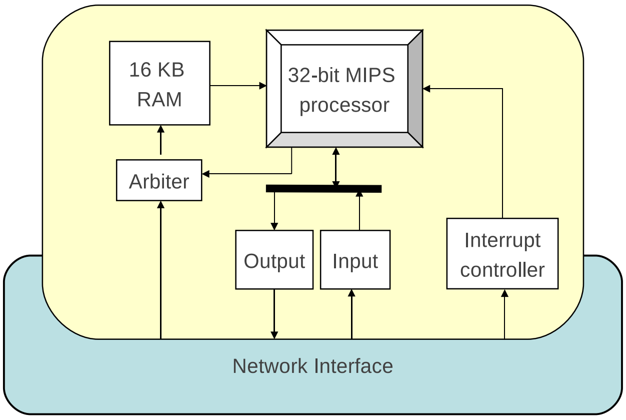

Fig. 11 The MIPS-based global control unit.

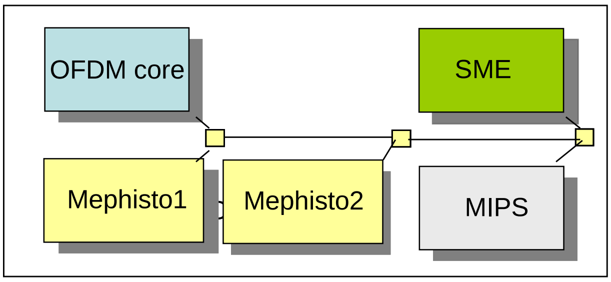

Fig. 12 The basic circuit.

The global control is software-based to allow more flexibility. When multiple basic circuits are stacked (to build a 3D system), global control is distributed between all the MIPS processors of the resulting 3D stack. Each one is in charge of controlling the 4 components of its layer and communicates with other host processors to exchange information about scheduling. Compared to a centralized host processor approach, this approach allows distributing and then alleviating the global control load. Moreover, such a distributed approach improves scalability by avoiding control bottleneck problem when the number of processing cores increases.

3.2.5 3D Asynchronous Mesh NoC

All the units of the basic circuit are interconnected via a NoC. This NoC is also used to connect all the components of a 3D system resulting from stacking 2 or more basic circuits. The basic circuit is designed as Globally Asynchronous Locally Synchronous (GALS) system. Processing units are synchronous (each one has its own clock frequency), while NoC routers are implemented in Quasi-Delay Insensitive asynchronous logic. The NI performs synchronization between the synchronous and asynchronous domains. The implementation details of this asynchronous router are not the focus of this paper. In this work, we use only 5×5 routers (5 Input ports × 5 output ports) to deal with all intra-layer and inter-layer communications as depicted in figure 5. The down ports of the routers located at the bottom layer are used to communicate with the external world.

Figure 12 depicts the resulting basic circuit. It is composed of 1 OFDM core, 1 SME and 2 MEPHISTOs (to meet real time constraints) interconnected via 3 routers. Each unit is plugged on the NoC via a network interface (NI). The NI deals with packetization, depacketization and flow control using credits, handled

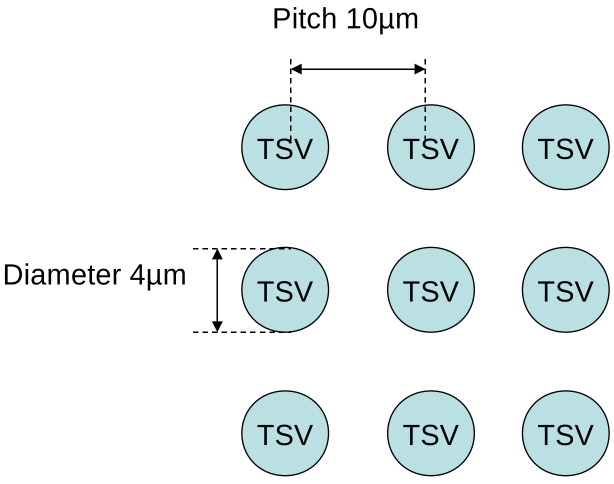

Fig. 13 TSV characteristics.

Table 3 Synthesis results in 65 nm technology

Bloc	Frequency (MHz)	Area (mm^2)
MIPS	300	0.175
MEPHISTO	400	0.455
SME	400	1.274
OFDM core	400	0.58
184 TSVs/router	-	0.0128
Router	-	0.20
Basic circuit	-	3.58

by input/output communication controllers.

3.2.6 Design Results

The asynchronous router and all the previously described units were synthesised with 65 nm low-power CMOS technology. All units' designs include the NI and test mechanisms like scan chains and memory BIST. TSVs' area overhead is a key challenge limiting the viability of 3D circuits. The asynchronous router needs 184 links at each port to communicate with its neighbors. Considering high-density TSV with 4 μm diameter and 10 μm pitch (**Fig. 13**), total TSVs' area would be 12,800 μm^2, which corresponds to 6.5% of the router size.

Table 3 depicts area results with the corresponding frequency of each synchronous core. The basic circuit has a total area of 3.6 mm^2. The TSVs' area overhead corresponds to 1.1% of the whole basic circuit size. We consider this overhead as negligible in this work. The MIPS host processor induces a 4.8% silicon impact. The NoC infrastructure represents 17% of the whole circuit area.

3.3 Technological Considerations

When using the macro-block partitioning granularity, each processor core is identical to its original 2D version and therefore has the same performance and power characteristics. Benefits of 3D stacking in terms of performance and power consumption are limited to vertical interconnections (TSVs). In this work, we consider the 3D integration process presented by Cadix et al. [16] from CEA-LETI. The TSV physical model [16] was realized using the Cadence Opus tool and simulated using the analog simulator ELDO. With a buffer driver 13 in 65 nm CMOS technology, signal propagation delay is 50 ps. The delay model is used in the next section to perform a comparison between a 3D architecture and its equivalent 2D version in terms of performance.

3.4 Case Study: Downlink Part of the 4×2 LTE Mode

To assess the performance and the power consumption of the proposed platform, we choose to deal with the downlink part of the LTE standard, and more specifically with the receiver side. The system is designed to transmit on 4 antennas and to receive on 2 antennas (4×2 MIMO), which requires a high performance processing, because of the implementation of diversity and spatial multiplexing schemes. Data are transmitted in 10 ms frames equally divided in 10 sub-frames also called TTIs (Time Transmission Intervals). A TTI is composed of 14 OFDM symbols and lasts 1 ms (at a sampling frequency of 15.36 MHz) [17]. 4 OFDM symbols contain pilot subcarriers with predetermined values that are used to estimate the transmission channel.

As said previously, the benchmark application is composed of 3 tasks:

(1) OFDM demodulation,
(2) Channel Estimation for each RX antenna,
(3) MIMO MMSE decoding based on a 4×2 double-Alamouti algorithm.

Data processing after MIMO decoding is performed by several demodulation operators (de-mapping, de-interleaving, channel decoding ...) to move from frequency samples (represented as complex numbers) to a stream of binary data. In this work, we consider a 2-layer 3D system resulting from stacking 2 instances of the basic circuit. The targeted application is mapped to this 3D platform as shown in **Fig. 14**.

In this work, we take as 2D reference architecture a platform called MAGALI devoted to wireless telecom applications. A silicon prototype is fabricated using the STMicroelectronics CMOS 65 nm LP technology [18]. MAGALI platform focuses mainly on 3GPP LTE standard and aims multi antennas schemes (MIMO). It supports OFDMA/MIMO TX/RX baseband algorithms. MAGALI architecture consists of the same units and the same NoC used in our basic circuit. The MAGALI platform control (scheduling and configuration) is semi-distributed. The global control is centralized and performed by an ARM11 host processor. In order to make comparison with our 3D platform, we use only a sub-set of the MAGALI platform as shown in **Fig. 15**. This sub-set will be referred to as MAGALI in the rest of the paper. It has a total area of 7.18 mm^2. The 3D system resulting from stacking 2 instances of the basic circuit is functionally equivalent to the MAGALI 2D platform. Thus, the 3D approach achieves up to 50% enhancement in form factor compared to the planar version.

3.4.1 Performance Results

Simulation environment includes 2 SystemC-TLM data generators to emulate the incoming data-flow from the 2 antennas. To speed up simulation, the asynchronous NoC is modelled in SystemC-TLM with post-layout parameters. All processing units of the 2 platforms are modelled at RTL level to provide cycle-accurate results. During simulation, a SystemC-TLM unit, called the recorder, records a trace of the output data-flow. This trace is compared to a reference file to check the obtained values and guarantee the right execution. **Figure 16** depicts an abstract view of the simulation platform.

Table 4 summarizes the performance results corresponding

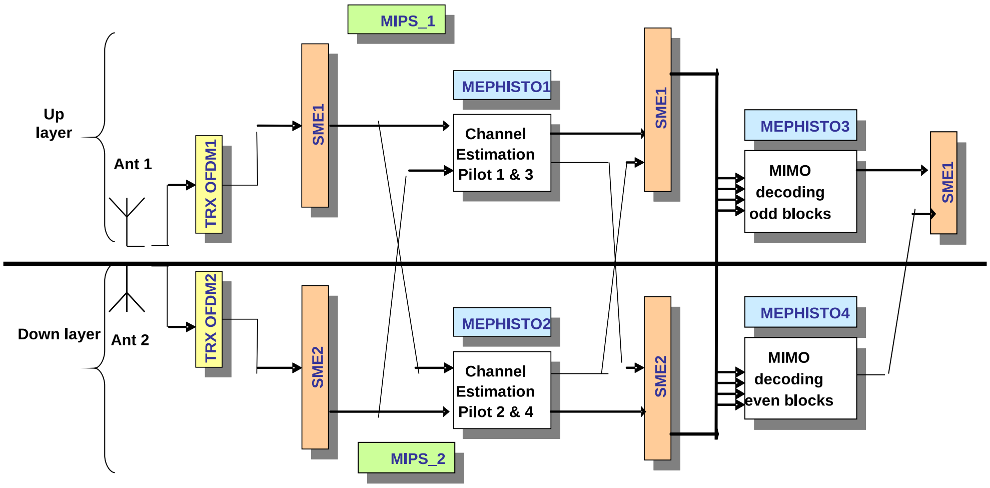

Fig. 14 Mapping of the 3GPP LTE application on our 3D platform.

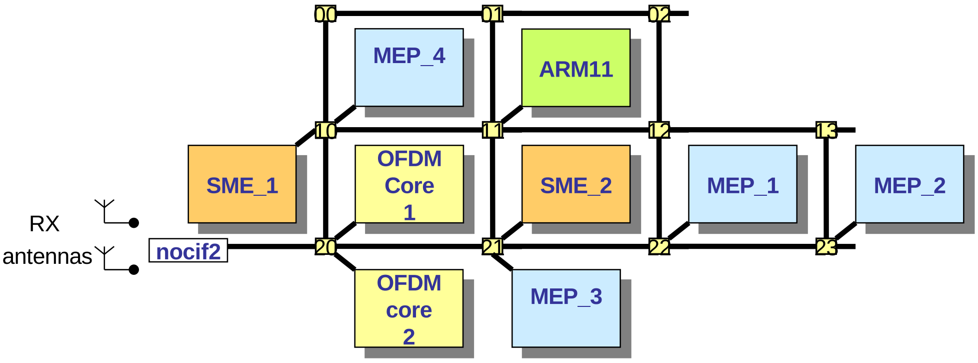

Fig. 15 2D reference architecture MAGALI.

Table 4 Time to process a TTI.

Platform	2D MAGALI	3D platform
Execution time	500.3 μs	481 μs
performance speed up	-	4%

Table 5 Power consumption of the processing units.

	Power consupmtion (mW)
FFT	172
Processing	207
Data manipulation	258

Table 6 Power consumption due to control.

Platform	2D MAGALI	3D platform
Power consumption due to control (mW)	22	20

to the processing time of a complete TTI including scheduling and reconfiguration phases. Host processors of the 2 platforms run at 300 MHz clock frequency, while processing units run at 400 MHz. In this work, our challenge was to keep at least the same performance to guarantee hard real-time constraints.

As depicted is Table 4, the execution time is almost the same in the two platforms. This is expected since we are using the same processing units on the 2 platforms. A speed up of 4% is achieved by the 3D approach thanks to the use of short vertical TSV interconnects. From 2D MAGALI to the 3D 2-layer system, the scheduling and reconfiguration management are transferred from centralized host CPU to 2 distributed MIPS processors. By efficiently using the smaller MIPS processor, there is no time overhead due to the communication between the 2 control processors of the 2 layers.

These results based on RTL simulation, confirm that the 3D platform with a distributed control is as efficient as the 2D MAGALI platform with a centralized controller.

3.4.2 Power Consumption Results

To provide a full comparison, we have evaluated the power consumption of the 2 platforms. Each platform was placed and routed in 65 nm low-power CMOS technology. A complete TTI processing was simulated with the placed and routed netlist.

Table 5 presents the average power consumption of the different processing units of the 2 platforms at gate level. The contributions of the FFT, data processing and data manipulation are the same for the 2 platforms since we are using the same units.

The control in the 3D platform is performed by two MIPS processors, which deal only with control. The control in the 2D MAGALI platform is performed by one ARM11 processor, which deals also with the MAC (Medium Access Control) layer. Its total power consumption is 150 mW. Power consumption due to control is estimated to be 22 mW. **Table 6** depicts power consumption due to control within the 2D and the 3D layer. The MIPS processors are as efficient as the ARM11 processor in term of power consumption.

Finally, it could be concluded that, based on post place and route simulation results, sthe distributed control approach introduce a low power consumption overhead to perform the control of the whole 3D platform.

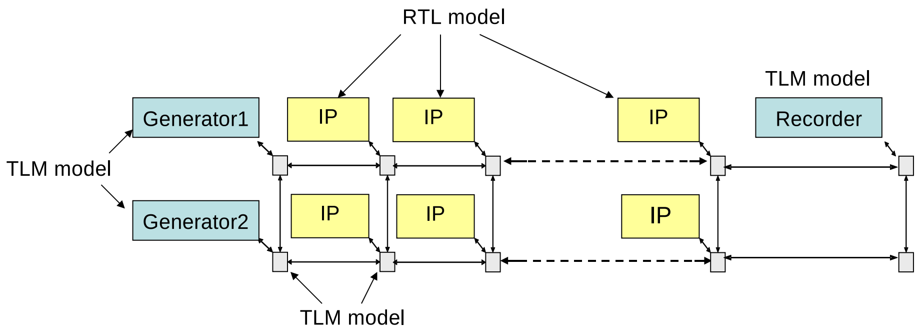

Fig. 16 An abstract view of the simulation platform.

4. Cost Analysis

In the previous sections, we present a reconfigurable and stackable circuit for LTE telecom applications. We perform a rigorous comparison between a 3D same-die stacked system (built by stacking 2 instances of the proposed basic circuit) and a 2D reference architecture. Results confirm that the same-die stacking approach is possible in the case of LTE applications, and that overall system performance and power consumption are not affected compared to the 2D approach.

As explained previously, 3D same-die stacked architectures are obtained by stacking several instances of the same die. This allows building a wide range of systems while reusing always the same mask set. Therefore, it would be possible to avoid designing a new mask set for each new application. As mask cost is prohibitive for aggressive technologies, 3D same-die stacking would reduce considerably final cost in some cases. In this section, we perform a quantitative investigation of the economical benefits of 3D same-die stacking, based on a system level cost model that will be detailed hereafter.

3D stacking (of different dies) is cost-effective only in the case of large-sized design. However, in this section, it is proven that 3D same-die stacking is beneficial even for small-sized circuits in some cases. To illustrate the economic benefits of the same-die stacking approach, assume that a semiconductor company is designing 3 digital circuits: a low-range, a medium-range and a high-range circuit, to meet market demands in terms of electronics dedicated to LTE applications. Each one of these circuits has its own processing performance. The low-range circuit corresponds to the basic circuit defined earlier in this paper. The proposed basic circuit has a small area (no more than 4 mm^2). The high range-circuit corresponds to the 4×4 transmission mode. Its computational performance is 10 times stronger than the basic circuit. Therefore, its 2D version is 10 times larger (in term of area) than the low range-circuit, and its 3D version may be obtained by stacking 10 instances of the low-range circuit. Similarly, the mid-circuit corresponds to the 4×2 transmission mode. Its computational performance is 2 times stronger than the basic circuit. Therefore, its 2D version is 2 times larger (in term of area) than the low range-circuit, and its 3D version may obtained by stacking 2 instances of the low-range circuit. When using the classical 2D approach, a new mask set has to be designed for each circuit. When using the same-die stacking 3D approach, only the mask set of the low-range circuit (which is the basic circuit) is needed to be designed. All other circuits can be built using this same mask set, by stacking multiple instances of the same low range circuit. Total production volume of the 3 circuit types (high, mid and low-range) is given by:

$$P_{total} = P_H + P_M + P_L$$

where P_H, P_M and P_L are the production volumes of the high, mid and low-range circuits respectively. h, m and l and are the fractions of P_{total} corresponding to production volumes of the high, mid and low-range circuit respectively. They are given by:

$$h = \frac{P_H}{P_{total}}$$

$$m = \frac{P_M}{P_{total}}$$

$$l = \frac{P_L}{P_{total}}$$

Total cost of the 3 types of circuits is given by:

$$C_{total} = P_H \times C_H + P_M \times C_M + P_L \times C_L$$

where C_H, C_M and C_L are unit costs of high, mid and low-range circuits respectively. In this analysis, we keep using the same data as in Section 2. **Figures 17** and **18** depict total cost of the 3 digital circuits, for 2 different production volumes (1 million and 10 million respectively) and different values of (h,m,l).

In the case of low total production volume (P_{total}=1 million), the W2W same-die stacking approach is more economical than the D2W and the IbS approaches. As an illustration, for (h,m,l)=(90%,5%,5%), using the W2W approach allows reducing total cost by 3.5 and 4.5 times compared to the D2W and IbS approaches (respectively). This may be explained by the high-yield of the basic circuit (thanks to its small area). Therefore, there is no need to test dies before stacking them.

Besides, the W2W approach is more beneficial than the 2D approach for all (h,m,l) combinations. For example, for (h,m,l)=(5%,90%,5%), total cost when using the 2D approach is twice higher than total cost when using the W2W approach. This is can not be due to 3D stacking since it is beneficial only in the case of large-sized circuit. The real reason is the mask reuse allowed by the same-die stacking approach.

Moreover, another interesting idea is to design only a 2D high-range circuit that is able to address all the LTE modes ((h,m,l)=(100%,0,0)). In this case, only one mask set would be

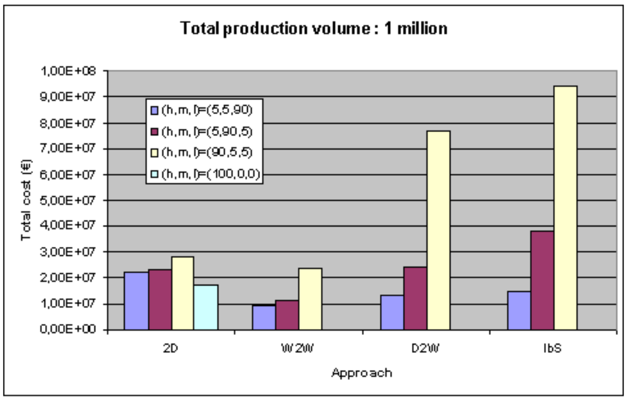

Fig. 17 Total cost of the 3 digital circuits for a total production volume of 1 million.

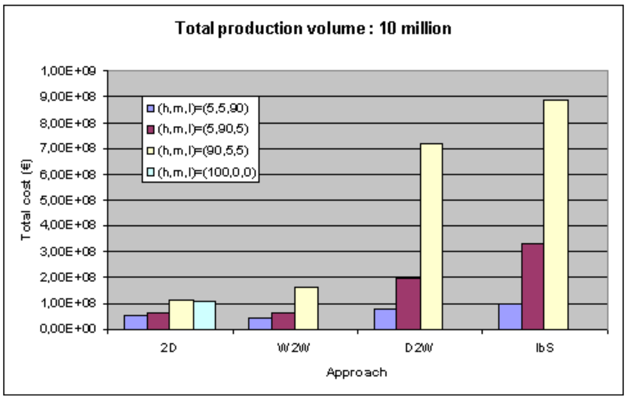

Fig. 18 Total cost of the 3 digital circuits for a total production volume of 10 million.

required (that of the high-range circuit). Figure 17 shows that this idea may give better economical results than the W2W same-die stacking approach in some cases, where the high-range circuit is the most produced (h > 50%). For instance, for (h,m,l)=(90%,5%,5%), this approach allows reducing total cost by 1.5 times compared to the W2W approach.

To summarize, in the case of small total volume production, and for low volume production of high-range circuits, the W2W same-die stacking provides the best results in term of cost.

In the case of high total production volume (P_{total}=10 million), classic 2D approach is the most cost-effective, compared to all 3D same-die stacking schemes. As an illustration, for (h,m,l)=(90%,5%,5%), the 2D approach allows reducing total cost by 1.5 times compared to the W2W approach. Indeed, the large number of produced circuits allows amortizing the cost of the 3 mask sets required by the 2D approach. Therefore, the contribution of mask cost to the final circuit cost is drastically reduced. As a result, the cost of the 2D circuits becomes less than any of the 3D circuits.

To conclude, 3D stacking of different dies in cost-effective only in the case of large-sized design. For small-sized circuits, 3D integration becomes cost-effective when using the 3D W2W same-die stacking approach, in the case of low total production volume, and for low production volume of high-range circuits. **Figures 19** and **20** depict the cost-effective options (among the 2D, 3D W2W, 3D D2W and 3D IbS approaches) for different circuit-sizes and production volumes, in the case of 3D stacking of different dies, and 3D same-die stacking respectively.

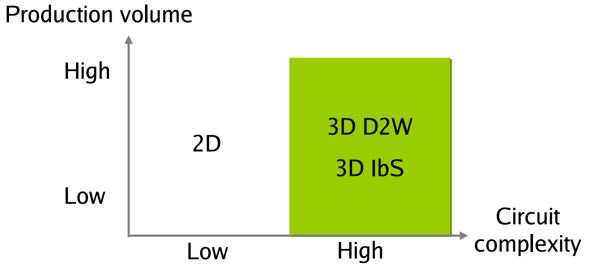

Fig. 19 Cost-effective approaches in the case of 3D stacking of different dies.

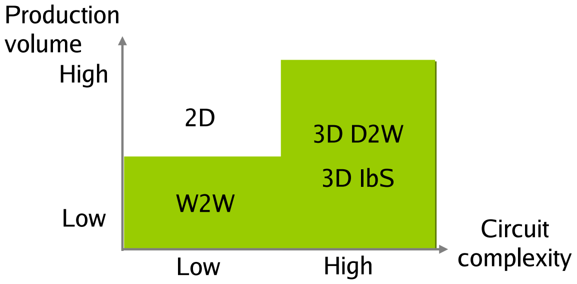

Fig. 20 Cost-effective approaches in the case of 3D same-die stacking.

5. Conclusion

In this work, we propose a reconfigurable and stackable circuit for 4G telecom applications. The proposed circuit can meet the computational requirements of the SISO (Single Input Single Output) transmission mode. By stacking several instances of this same circuit, it would be possible to boost overall system performance and address several MIMO (Multiple Input Multiple Output) modes.

The proposed reconfigurable and stackable circuit is intended to provide hardware resources that meet the requirements of the 4G telecom applications. It is composed of 5 elements interconnected thanks to a NoC. The first component is the Smart Memory Engine which is Micro-programmable Memory Controller (MMC) designed to perform data synchronization and distribution in dataflow systems. The second component is the OFDM core devoted to perform direct and inverse fast Fourier transform (FFT and IFFT). The proposed circuit includes also 2 DSPs dedicated to perform complex matrixs computation, useful for channel estimation, advanced MIMO coding/decoding. The global control of the previously described units is performed by a 32-bit MIPS processor, by means of direct addressing and interrupts mechanisms.

To assess the performance and the power consumption of the proposed platform, the 4×2 LTE mode is implemented on a 3D system resulting from stacking 2 instances of the basic circuit. A rigorous comparison between a 3D same-die stacked system and a 2D reference architecture confirms that the same-die stacking approach is possible in the case of LTE applications, and that overall system performance and power consumption are not affected compared to the 2D approach.

Besides, the same-die stacking approach for LTE applications provides good results in terms of cost (compared to the 2D approach) in some cases when using the W2W assembly scheme.

The major limitation of the same-die stacking approach is the thermal constraints due to the stacking of several highly active logic layers. High temperatures may limit the operating frequencies of vertically-stacked chip and degrades chip reliability. Future work will investigate potential solutions for this problem either technological (by inserting thermal vias) or architectural (enhancing power management by means of dynamic task mapping for example).

Reference

[1] Miura, N., Take, Y., Saito, M., Yoshida, Y. and Kuroda, T.: A 2.7 Gb/s/mm2 0.9 pJ/b/chip 1 coil/channel ThruChip interface with coupled-resonator-based CDR for NAND Flash memory stacking, *IEEE International Solid-State Circuits Conference Digest of Technical Papers* (*ISSCC*), pp.490–492 (2011).

[2] Dorsey, P.: *Xilinx Stacked Silicon Interconnect Technology Delivers Breakthrough FPGA Capacity, Bandwidth, and Power Efficiency*, Xilinx white paper (2010).

[3] Maly, W.P. and Yangdong, D.: 2.5-dimensional VLSI system integration, *IEEE Trans. Very Large Scale Integration* (*VLSI*) *System*, pp.668–677 (2005).

[4] Yu, C.H.: The 3rd dimension-More Life for Moore's Law, *Microsystems, Packaging, Assembly Conference*, pp.1–6 (2006).

[5] Kim, J.S. et al.: A 1.2 V 12.8 GB/s 2 Gb mobile Wide-I/O DRAM with 4×128 I/Os using TSV-based stacking, *IEEE International Solid-State Circuits Conference Digest of Technical Papers* (*ISSCC*), pp.496–498 (2011).

[6] Ono, T., Inoue, K. and Murakami, K.: Adaptive cache-line size management on 3D integrated microprocessors, *International SoC Design Conference* (*ISOCC*), pp.472–475 (2009).

[7] Franzon, P.D. et al.: Design for 3D Integration and Applications, *International Symposium on Signals, Systems and Electronics*, pp.263–266 (2007).

[8] Leduc, P. et al.: Enabling technologies for 3D chip stacking, *International Symposium on VLSI Technology, Systems and Applications*, pp.76–78 (2008).

[9] Dong, X. and Xie, Y.: System-level cost analysis and design exploration for three-dimensional integrated circuits (3D ICs), *Asia and South Pacific Publication Design Automation Conference*, pp.234–241 (2009).

[10] Nag, P.K., Gattiker, A., Sichao, W., Blanton, R.D. and Maly, W.: Modeling the economics of testing: a DFT perspective, *IEEE Design and test of Computers*, Vol.19, No.1, pp.29–41 (2002).

[11] Lau, J.H.: TSV Manufacturing Yield and Hidden Costs for 3D IC Integration, *Proc. Electronic Components and Technology Conference*, pp.1031–1042 (2010).

[12] Motorola: *Long Term Evolution* (*LTE*): *A Technical Overview* (2007).

[13] Berkmann, J., Carbonelli, C., Dietrich, F., Drewes, C. and Xun, W.: On 3G LTE Terminal Implementation - Standard, Algorithms, Complexities and Challenges, *International Wireless Communications and Mobile Computing Conference*, pp.970–975 (2008).

[14] Martin, J., Bernard, C., Clermidy, F. and Durand, Y.: A Microprogrammable Memory Controller for high-performance dataflow applications, *Proc. ESSCIRC*, pp.348–351 (2009).

[15] Bernard, C. and Clermidy, F.: A Low-Power VLIW processor for 3GPP-LTE Complex Numbers Processing, *Conference and Exhibition Design, Automation and Test in Europe*, pp.1–6 (2011).

[16] Cadix L. et al.: Integration and Frequency Dependent Parametric Modeling of Through Silicon via Involved in High Density 3D Chip Stacking, *The Electrochemical Society transactions*, Vol.33, No.12, pp.1–21 (2010).

[17] 3GPP TSG-RAN: *3GPP TS36.211, Physical Channels and Modulation* (*Release 8*) (2007).

[18] Clermidy, F., Bernard, C., Lemaire, R., Martin, J., Miro-Panades, I., Thonnart, Y., Vivet, P. and Wehn, N.: A 477 mW NoC-Based Digital Baseband for MIMO 4G SDR, *In Proceeding of IEEE International Solid-State Circuits Conference*, Vol.53, pp.278–279 (2010).

Walid Lafi received his B.Eng. degree from Tunis Polytechnic School Tunisia in 2007. He then received his master's degree in microelectronics from the University of Grenoble France in 2008. Since October 2008, he has been a research assistant (Ph.D. student) at the French Atomic Energy Commission CEA within the Laboratory for Electronics and Information Technology LETI. Currently, his research activities are focusing on multiprocessor architecture based on 3D integration technologies.

Didier Lattard received his Ph.D. degree in microelectronics from the National Polytechnic Institute of Grenoble, France, in 1989. In 1990, he joined the CEA-LETI Laboratory in the Center for Innovation in Micro and Nanotechnology (MINATEC), Grenoble. He was involved in the design of image and baseband processing circuits as Senior R&D Engineer and Project Leader. From 2003 to 2006, he led the development of the FAUST NoC-based telecom platform. Since 2006, he has been in charge of new projects in imaging and high-performance computing applications. He has published 29 papers in books, refereed journals and conferences. He holds 18 patents in the fields of baseband processing and NoC architectures.

Ahmed Jerraya received his B.Eng. degree from the University of Tunis in 1980 and the D.E.A., "Docteur Ingénieur", and the "Docteur d'Etat" degrees from the University of Grenoble in 1981, 1983, and 1989 respectively, all in computer sciences. In 1986, he held a full research position with the CNRS (Centre National de la Recherche Scientifique). From April 1990 to March 1991, he was a Member of the Scientific Staff at Nortel in Canada, working on linking system design tools and hardware design environments. In 2007, he became a Director of Strategic Design Programs at CEA/LETI France. He served as General Chair for the Conference DATE in 2001, Co-founded MPSoC Forum (Multiprocessor system on chip) and served as the organization chair of ESWEEK2009. He supervised 51 Ph.D., co-authored 8 Books and published more than 250 papers in International Conferences and Journals.

(Invited by Editor-in-Chief: *Hiroyuki Tomiyama*)