Speed/Power/Area Trade-offs for High Speed Inter Layer Data Transmission in 3D Stacked ICs

THÈSE N^{O} 6278 (2014)

PRÉSENTÉE LE 26 AOÛT 2014

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR
LABORATOIRE DE SYSTÈMES MICROÉLECTRONIQUES
PROGRAMME DOCTORAL EN MICROSYSTÈMES ET MICROÉLECTRONIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Giulia BEANATO

acceptée sur proposition du jury:

Dr G. Boero, président du jury
Prof. Y. Leblebici, Prof. G. De Micheli, directeurs de thèse
Prof. A. P. Burg, rapporteur
Dr F. Clermidy, rapporteur
Prof. A. K. Coskun, rapporteuse

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2014

Considerate la vostra semenza:
fatti non foste a viver come bruti,
ma per seguir virtute e canoscenza.
— Dante Alighieri

To the memory of my grandmas...

Acknowledgements

This dissertation is the final snapshot of four and a half years of days, nights and weekends digging into some of the many challenges someone faces in order to stack multiple silicon layers. None of this work would have been accomplished if it wasn't for Prof. Yusuf Leblebici and Prof. Giovanni De Micheli, who gave me the opportunity to pursue this research and guided me throughout it. Research is an endless task that rapidly becomes an addiction. I have to thank the member of the thesis committee Prof. Burg, Prof. Coskun, Dr. Clermidy and Prof. Boero, for evaluating my work putting an end to my PhD dependence.

During these years I had the chance to share the lab with both amusing and talented people. Among them I have to particularly acknowledge several persons that marked my life. My officemate Ale who supported me both technically and as a friend, we never stopped arguing and fighting, it was fun and I learned a lot. Hossein, a brilliant officemate for 3 years and a remarkable person. Kiarash and Mahsa who provided for the "tuc of the day" and great support while I was struggling with endless problems, I really hope I've helped them at least half as much as they've helped me. Davide who listened to me when I really needed. Yuksel, Michael, Paolo, Igor and all the people that shared their knowledge and experience with me. I am extremely grateful to Alain for his everlasting patience regarding all the times I knocked on his door with some linux-related issue. I would not dare imagining how could I have endured without the invaluable support of Marie and Melinda.

Then, I should admit I owe a big debt of gratitude to my PhD, for giving me the opportunity to run into amazing people like Sara, Francesco, Alessia, Romain, Erifili, Marco(s), Mauro, Andrea(s), Luca, Alessandro(s), Stefano, ... and all the friends I was so lucky to meet but I cannot mention one by one without occupying too many pages.

The last year of a PhD is supposed to be the most stressful, but thanks to my sister who had the courage to come live with me and the theatre groups I joined I will remember it as an amazing year. I have to thank the ImproPas for all the amusement and laughter we had while bringing new people and situations into being. At the same time I'm also extremely grateful to the Catalyst and all the "enemies of the people" who accepted me as part of the group. I had a wonderful time and a lot of fun rehearsing almost everyday with all of you, we had a great run.

I do owe a really special thank to Paolo, Enrico e Wolty, although the kilometres are separating us, they've always been there for me and I know I can always count on them when I need.

Finally, the deepest thank should go to Graziana, Gigi and Elena. Family is the most important thing in life, anything I achieved and will achieve is worth it just because I can share it with them.

Acknowledgements

Grazie!

Lausanne, 19 June 2014 G. B.

Abstract

Technology scaling is now struggling to meet the expectations dictated by Moore's Law due to complexity and cost. In particular, interconnects have become a major performance bottleneck for planar chip multi-processors and are rapidly dominating die area and power dissipation. These limiting factors need to be addressed at the architectural level. A promising option explored in the last few years are *Three Dimensional (3D)* architectures exploiting *Through Silicon Via (TSV)*. In fact, TSV interconnects have the potential to reduce the interconnect wire length and, at the same time, offer high vertical connection density.

In the last decade several groups in both academia and industry have developed their own TSV technology exploiting different designs, fabrication techniques and materials. Given the diversity of the parameters characterizing TSV technology, the choice of the most appropriate 3D interconnect for a specific design is of utmost importance. Nevertheless, the silicon area overhead due to the TSV insertion and the non-negligible TSV capacitance might hinder the performance improvements offered by 3D-ICs. Hence, the IC designers are now required to find circuit design techniques that fully exploit the potential of the available 3D interconnects.

This dissertation discusses circuit design solutions for cross-chip communication through state-of-the-art TSV channels. Different TSV channels and interconnect topologies have been explored with the aim of achieving the area, power and performance requirements of 3D-stacked systems. More specifically, the contribution of this thesis consists of proposing design solutions to leverage the high bandwidth and low-delay connection provided by TSV links minimizing the cost in terms of silicon area and capacitance.

A silicon-proven analytical model of the TSV channel is presented, which provides good correlation between simulations and chip measurements. The model is then extensively utilized to emulate the 3D I/O links while designing the interconnect circuits.

The following parts of the thesis aims to demonstrate the efficiency of different 3D interconnection topologies in complete systems. A *configurable network architecture exploiting a fully parallel TSV bus* is proposed. The goal is to create a system composed of a cluster of processing elements, placed on a logic layer, and multiple layers of SRAM modules constituting a single shared L1 memory. The performance of a shared-L1 memory critically depends on the architecture of the interconnect between processors and memory banks. The required low-latency communication is achieved by a 3D logarithmic network structure. This architecture shows significant improvements in both area and latency compared to the planar 2D implementation.

Nevertheless, a parallel TSV bus is still extremely area consuming. In order to balance inter-

layer bandwidth offered by TSVs and their silicon area occupation, a *high data-rate 3D serial link* is proposed. The high-speed serial 3D link optimizes both the inter-layer communication speed and the interconnect area occupation, still guaranteeing low-power communication. The proposed *serializer-deserializer (SERDES)* circuits have been explored across a variety of state-of-the-art TSV technologies.
A modular 3D stacked multi-processor platform, 3D-MMC, featuring the 3D serial connection macro has been designed to demonstrate the potential of the proposed interconnect topology. The analysis of the systems' wiring characteristics demonstrates that the reduction in the number of TSVs obtained with the adoption of the serial vertical connection improve the routing congestion of the 3D system.
In the final part of this dissertation, a test vehicle based on the 3D-MMC architecture is presented. A prototype of the 3D multi-processor system has been designed, fabricated and tested. The final 3D system has been obtained by stacking *Known Good Dies(KGD)* using an in-house via-last TSV process. The experimental results obtained from simulations and measurements on the fabricated samples demonstrate that the system exhibits multiple Gbps vertical data bandwidth while limiting the number of TSVs.
Finally, the thesis is concluded with a summary of the contributions and a discussion on the future work.

Key words: 3D ICs, Through Silicon Vias TSVs, Chip Multi-Processors CMPs, high speed serial links.

Sommario

L'aumento della complessità e del costo legato alla riduzione delle dimensioni della tecnologia CMOS stanno rendendo sempre piú difficile rispettare le aspettative dettate dalla legge di Moore. In particolare, le interconnessioni sono diventate una delle maggiori cause limitanti per le prestazioni dei multiprocessori planari e stanno rapidamente dominando sia l'area che il consumo di potenza dei chip. Questi fattori limitanti devono essere affrontati a livello architetturale. Una promettente opzione esplorata negli ultimi anni sono le architetture tridimensionali che utilizzano le interconnessioni attraverso il silicio, TSVs. Infatti, le TSV hanno la caratteristica di ridurre la lunghezza delle interconnessioni offrendo allo stesso tempo un'alta densità di connessione verticale.

Nell'ultimo decennio diversi gruppi sia in accademia che in industria hanno sviluppato la propria tecnologia di fabbricazione delle TSV utilizzando differenti design, techniche e materiali. Data la diversità dei parametri che caratterizzano le tecnologie TSV, la scelta della tecnologia piú adatta per un particolare design é di primaria importanza. Nonostante ció, l'aumento dell'area dovuta all'instanziamento delle TSV e la loro capacità parassita non trascurabile potrebbero offuscare i benefici offerti dalla tecnologia 3D. Di conseguenza, nuove tecniche di design devono essere esplorate per sfruttare al massimo le potenzialità della tecnologia 3D. Questa tesi propone soluzioni circuitali per la comunicazione tra diversi chip attraverso le TSV disponibili oggi. Varie TSV e topologie di interconnessione sono state esplorate con lo scopo di raggiungere le aspettative sulle prestazioni promesse dalla tecnologia 3D. In particolare, il contributo di questa tesi consiste nel proporre soluzioni per ottimizzare il compromesso tra l'alta banda e la velocità delle TSV riducendo al minimo il costo in termini di utilizzo di area e capacità parassita.

Come prima cosa viene presentato un modello analitico delle TSV validato da misure su silicio, garantendo quindi una buona correlazione tra simulazioni e misure. Il modello viene quindi ampiamente utilizzato per emulare il link 3D durante il design dei circuiti di interconnessione.

La parte seguente della tesi ha come scopo dimostrare l'efficienza di diverse topologie di interconnessione inserite in sistemi completi. Viene proposta un'architettura di network combinatoria che utilizza un bus parallelo di TSV. Lo scopo è di interconnettere un sistema composto da un gruppo di processori localizzati su un layer di logica, ad una memoria condivisa di primo livello composta da vari layer occupati da blocchi di SRAM. L'architettura dell'interconnessione tra processori e memoria è un fattore critico per le prestazioni di una memoria di livello 1 condivisa. La comunicazione a bassa latenza viene ottenuta tramite una struttura di network logaritmica in 3D. Questa architettura mostra un incremento significativo

sia in termini di area che latenza se comparata all'implementazione planare in 2D.
Ció nonostante un bus parallelo di TSV occupa una porzione significante di area. Un 3D link seriale ad alta velocità viene quindi proposto per bilanciare la bandwidth offerta dalle TSV e la loro area. Il 3D link seriale ad alta velocità ottimizza sia la comunicazione tra layer che l'occupazione di area, garantendo allo stesso tempo un basso consumo in potenza. I circuiti di serializzazione-deserializzazione sono stati esplorati per diverse tecnologie di TSV.
Un multi-processore modulare tridimensionale, 3D-MMC, che utilizza una connessione inter-layer seriale è stato progettato per dimostrare le potenzialità della topologia proposta. L'analisi delle interconnessioni del sistema dimostra che la riduzione del numero di TSV ottenuta attraverso la serializzazione riduce la congestione del routing.
Nella parte finale della tesi viene presentato un prototipo che si basa sull'architettura 3D-MMC. Il prototipo del multi-processore 3D è stato progettato, fabbricato e testato. Il sistema 3D finale è stato ottenuto mettendo uno sopra l'altro i chip testati funzionanti, interconnettendoli con un processo di fabbricazione delle TSV sviluppato in-house. I risultati sperimentali ottenuti dalle simulazioni e dalle misure sui chip fabbricati dimostrano che il sistema puó arrivare a funzionare a diversi Gbps di bandwidth verticale limitando il numero di TSV.
Infine, la tesi si conclude con un sommario dei contributi scientifici ed una discussione sul lavoro da sviluppare in futuro.

Parole chiave: 3D ICs, Through Silicon Vias TSVs, Chip Multi-Processors CMPs, high speed serial links.

Contents

List of Figures

List of Tables

1 Introduction

Since 1965, with the postulation of Moore's law [1], the integration density has been increasing continuously thanks to the aggressive scaling of process technology. At the same time, the performance gains enabled by scaling have been gradually challenged by on-chip interconnects. As the request for increased performance and computing power rise, hundreds of millions of transistors are placed on a single chip occupying more and more area, but the realization of large dies impacts the global interconnects length, as well as reliability and manufacturing yield. At the same time, as feature size decreases, the cross section of the metal wires also shrinks.

Despite the advent of low-resistivity materials for the on-chip wires, the decrease of their cross-section area, A, due to scaling and the increased wirelength, l, are causing a rise in the

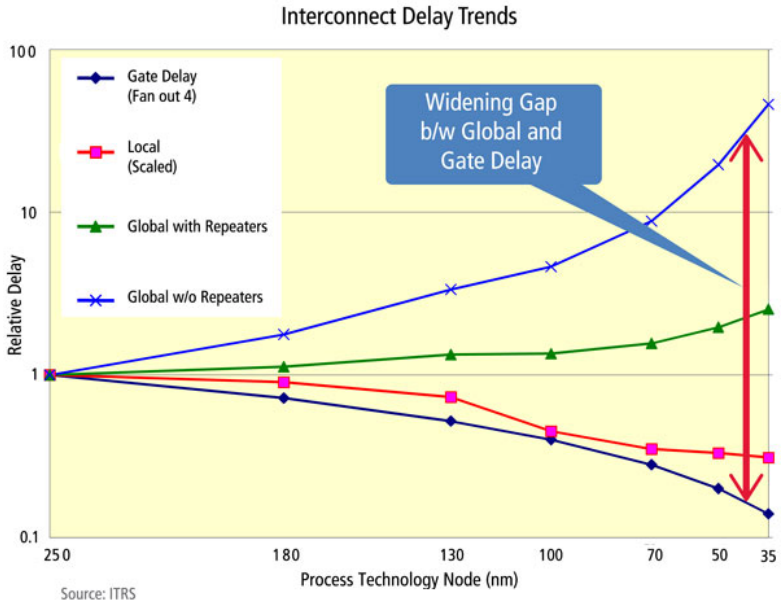


Figure 1.1: Delay of Metal 1 and global wiring vs. technology node [2].

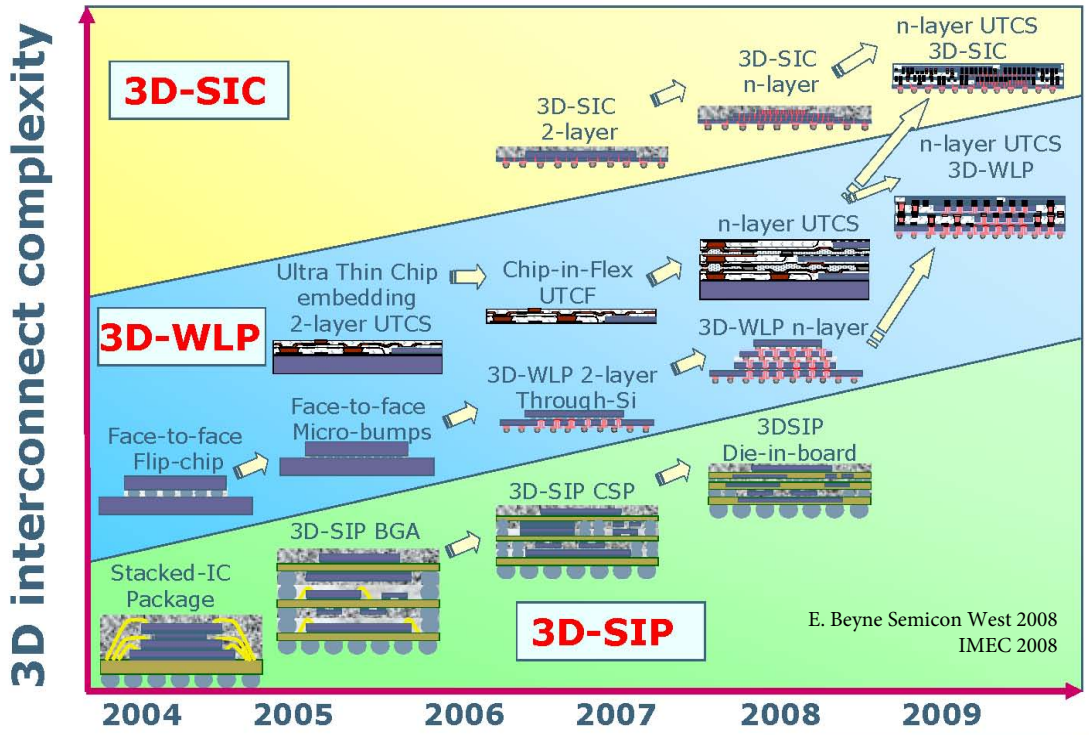


Figure 1.2: 3D interconnect roadmap by IMEC.

resistance $R = \rho \frac{l}{A}$. In addition, as the aggregated interconnect length increases since more wires are being used in each layer and more metallization layers are being added, their delay $t = \frac{l}{2} RCl^2$ becomes dominant over the gate delay. Figure 1.1 depicts the delay of local and global wiring in future generations. The length of local wires usually shrinks with traditional scaling, hence the impact of the lower metal layers delay on performance is minimal. On the other hand, global interconnects are impacted the most by the degraded delay. Even though the insertion of repeaters can improve the delay in global wiring, this approach causes a significant increase in power consumption as well as the need for increased chip area. Furthermore, interconnect delay is just a part of the problem: as the clock frequencies continue to climb, any increase in interconnect loading significantly increases the power consumption of the ICs.

This dissertation focuses on the advantages of 3D stacking applied to microprocessors and related integrated microprocessor systems where more than 30% of the power can be consumed in backend interconnect wire [3]. In general, microprocessors are driving towards lower power consumption, increased performance, reduced form factor and increased integration. With 3D integration, multiple strata of different types can be stacked with a high bandwidth, low latency and low power interface. Additionally, wire reduction using 3D provides new microarchitecture opportunities to trade off performance, power and area.

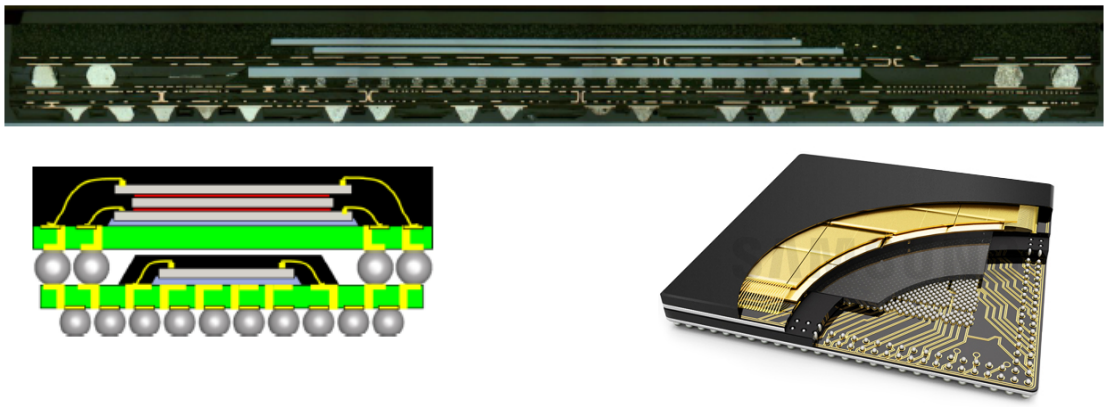


Figure 1.3: Samsung PoP technology.

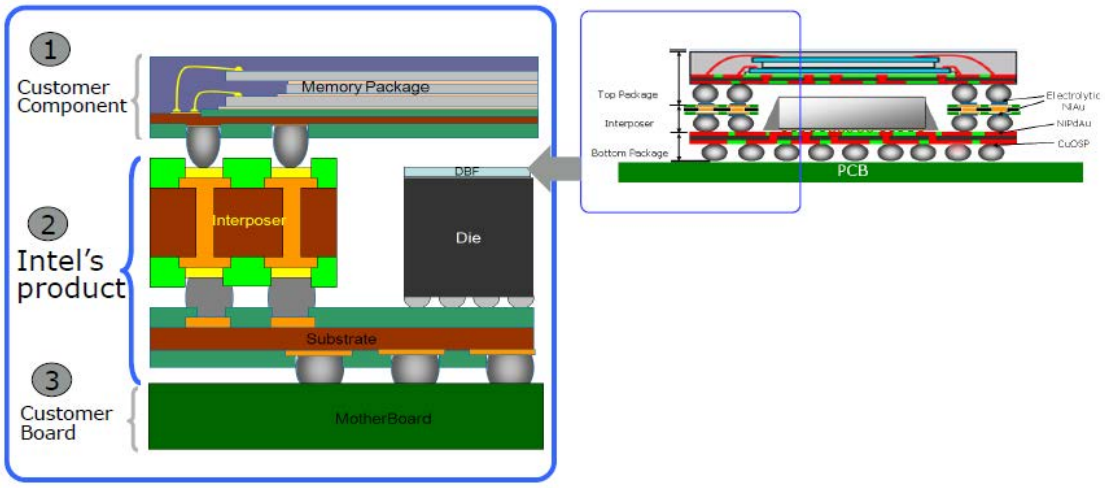


Figure 1.4: Intel Co-PoP technology.

1.1 From 2D to 3D ICs

During the last years several options have been studied as promising solutions for overcoming the interconnect bottleneck and continue the "More-than-Moore" trend [4]. An overview of the evolution of the interconnect and packing solutions studied by IMEC is shown in Figure 1.2.

1.1.1 System in package

A first step was taken with the advent of the *System-in-Package (SiP)* assemblies, which can be defined as an assembly of naked or packaged dies mounted on the same main package either in a 2D or a 3D manner. Integrating all the critical components within a package rather than on a *Printed Circuit Board (PCB)* significantly contributes to system miniaturization and communication bandwidth. Additionally, another advantage of the SiP is that each die can be implemented using the most appropriate technology process.

For 2D SiP, the dies are connected through the metal interconnects on the SiP substrate. In case of 3D SiP, the interconnections through the z-axis can be implemented in different ways, such as wire bonding, vertical interconnects along the periphery, or long and wide, low density vertical interconnects.

In the case of a SiP mounted on another SiP, the system is referred to as a *Package-on-Package (PoP)*. Samsung devices, such as Samsung's Exynos 4210, already exploit PoP technique in order to integrate a logic layer at the bottom with memory layers at the top, as in Figure 1.3. Each die can be selected, tested independently and then assembled. This approach offers significant benefits, such as reduced physical size, high-density I/Os and high heat dissipation.

Also Intel has officially disclosed his Co-PoP solution, shown in Figure 1.4, at the Intel Developer Forum of 2012. Intel supplies bottom package, while customers can select the desired memory to be placed on top. A product already using this approach is the Atom Z2460.

1.1.2 2.5D IC

A further step towards a more compact integration was enabled by the usage of a silicon interposer. A Si-interposer is a double sided die with no active devices that is used to connect the active dies among each other, while the metallization layers on the top and bottom faces are connected with *Through Silicon Vias(TSVs)*. 2.5D ICs can be defined as an assembly of dies placed on a silicon interposer.

A successful product exploiting this technology was put on the market by Xilinx in 2011. The Virtex7 2000T is still the world's highest density FPGA, delivering greater than 2X the capacity and bandwidth offered by the largest monolithic devices [5]. A simplified cross-section of the Virtex-7 2000T FPGA is depicted in Figure 1.5: 4 FPGA are flip-chip bonded to the Si-interposer integrating 2 million logic cells, 6.8 billion transistors and 12.5 Gb/s serial transceivers on a single device. The system is optimized to reduce power dissipation, for noise isolation and to achieve high yield.

1.1.3 3D IC

Although 2.5D integration gained popularity because of the and its technological feasibility, it might still not be sufficiently effective for several high performance applications. With 3D integration, multiple dies can be directly stacked in top of each other and interconnected through TSVs so that they function as a single device. 3D ICs have the potential to provide significant advantages over traditional planar circuits without the need of any interposer.

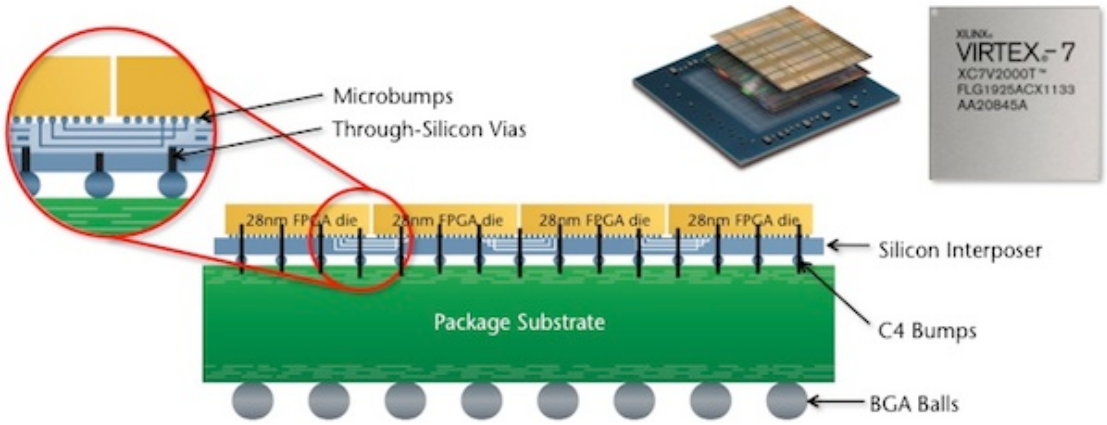


Figure 1.5: Virtex-7 2000T FPGA from Xilinx.

3D opportunities

The motivations driving the extensive research on 3D ICs are basically three: form factor, performances and heterogeneous integration. Folding a chip into a 3D configuration drastically improves the **form factor** since the same number of transistors can be integrated within a smaller area with respect to a 2D implementation. As a consequence of the higher packing density, the average interconnect length improves [6], which directly translates into a reduction of the RC delay boosting the system **performance**. Reducing the interconnect length also lowers the number of repeaters along the lines, hence the power dissipated to ensure signal integrity and propagation.

While traditional 2D *System-on-Chip(SoC)* struggles to reach the bandwidth demand of next generation computing device due to pad number limitation, the low parasitic, high vertical connect density provided by TSVs can potentially provide TB/s **data bandwidth**. Hence 3D technology has the potential to replace traditional off-chip signalling technology, in particular improving memory communication bandwidth.

The possibility of integrating **heterogeneous technologies** in the same system is also a main advantage of 3D ICs. For instance, DRAM and processors can be integrated in the same system helping to overcome the memory to processor performance bottleneck that plagues 2D ICs. Also the possibility of fabricating analog and digital circuits separately with different technologies and then stacking them is an interesting opportunity to optimize the functionality of each block and avoid problems related to noise.

3D challenges

The benefits offered by the third dimension still have to be unlocked by overcoming several challenges, starting from design methodologies at the front-end until the mature manufacturing processes at the back-end.

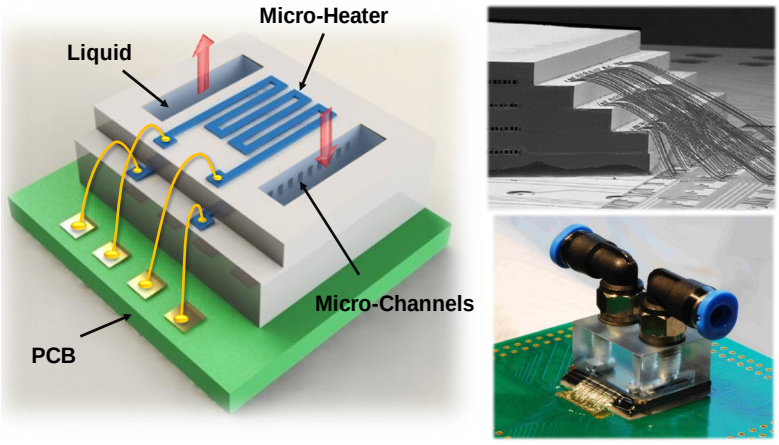


Figure 1.6: Micro-channel based liquid cooling test vehicle from a collaboration between EPFL and IBM, courtesy of LSM.

- **Manufacturing technology** - Since TSV fabrication technologies are not yet mature, reliability is expected to be a limiting factor for 3D IC performance and yield [7]. For instance, unsuccessful wafer alignment during the bonding process and handling of very thin silicon wafers are the primary mechanisms of TSV failure, with a minimum TSV diameter of 1.2 μm reported in the open literature [8]. The characteristics of the vertical interconnects are fundamental to enhance the performance enough to compensate the cost of this new technology.

- **CAD tools** - Although it is possible to design 3D ICs with the available *Computer Aided Design(CAD)* tools by ad-hoc techniques which mostly consist of a divide-and-conquer approach. The design of a 3D IC is a challenging and risky task due to the lack of industry-standard CAD tools to automate the process. The CAD community is required to develop automated solutions for partitioning, floorplanning, placement and routing of 3D designs. Lately, leading companies in the CAD domain, such as Cadence [9] and Synopsys [10] have started proposing the first solutions for 3D IC development.

- **Power density** - Even though the system power consumption is expected decrease, the high packing density and the lack of heat conduction paths will translate in a great increase of the power density. The thermal dissipation of the tiers far from the heat sink is expected to be a major concern, especially for power consuming applications. Several options have been studied to mitigate the effect of vertical integration, such as liquid cooling using micro-channels [11] [12] shown in Figure1.6, thermal vias [13] [14] [15] or, in case of multi-processor systems, software solutions like scheduling algorithms performing thermal-aware task migration [16].

- **Testing** - The testing strategy required by 3D stacked system is significantly more complicated than the traditional testing methodology for 2D ICs. Ideally, each die, or system, should be tested functional at each step. First, a pre-bonding test should identify *Known Good Dies(KGD)*. Then, after stacking each layer, a mid-bond test of the system should

be performed to guarantee that the system is still working before stacking another layer. Finally, the post-bond test should identify the functional system. Depending on the design, it may be extremely challenging to test all the layers before the bonding process. For Example some of the dies may not have real signal pads since their *Input/Output(I/O)* signals are transmitted through TSVs. Missing testing steps can cause a noticeable decrease of the final yield. R&D groups have, and still are, dedicating a lot of effort to find testing solutions and several design-dependent solutions have been proposed, however, the definition of a testing policy for 3D ICs is still in progress.

3D architectures

Depending on the type of layers stacked, the 3D system can be categorized within one of the following topologies.

- **Memory on Logic** generically includes stacking cache, main memory or strata with similar functions onto a logic device and has been one of the most studied architectural choice for 3D ICs. Some of the advantages of increasing the on-die cache capacity by stacking are increased performance by capturing larger working sets, reduction of off-die band-width requirements by accessing more data on die instead of externally, and reduction of system power by reducing bus activity through fewer main memory accesses.

  In December 2011, JEDEC Solid State Technology Association announced a new standard for Wide I/O mobile DRAM which uses chip-level 3D stacking with TSV interconnects and memory chip directly stacked upon a *System on Chip(SoC)*. The Wide I/O standard was released in order to meet the industry demands for increased level of integration, as well as improved bandwidth, latency, power and form-factor [17].

  Based on the JEDEC standard, the *WideIO Memory Interface Next Generation (WIOMING)* was presented in 2013. The 3D system, shown in Figure 1.7, was developed from the cooperation among ST-Microelectronics for SoC wafers manufacturing and assembly, CEA-Leti for middle end process steps and ST-Ericsson for electrical test and project management. The stacked dies are connected by TSVs with a diameter of 10μ and a pitch of 40μ, drilled in the SoC die.
  At the end of 2013, also Samsung moved from PoP to TSV based 3D technology by presenting its Widcon technology. The structure, depicted in Figure 1.8, brings better energy efficiency and higher bandwidth, up to 17GB/s.
  Again Samsung, as part of a consortium with leading companies like Micron and IBM, released the DRAM Hybrid Memory Cube (HMC)in 2013. The goal is to break the memory wall by removing the logic transistors from each DRAM die and then stack them on top of a small logic die, which takes care of buffering and routing from/to the memory banks. This centralized logic driving a stack of 8 memory dies allows for higher and more efficient data rates, up to 320GBps, while consuming 70% less energy than DDR.

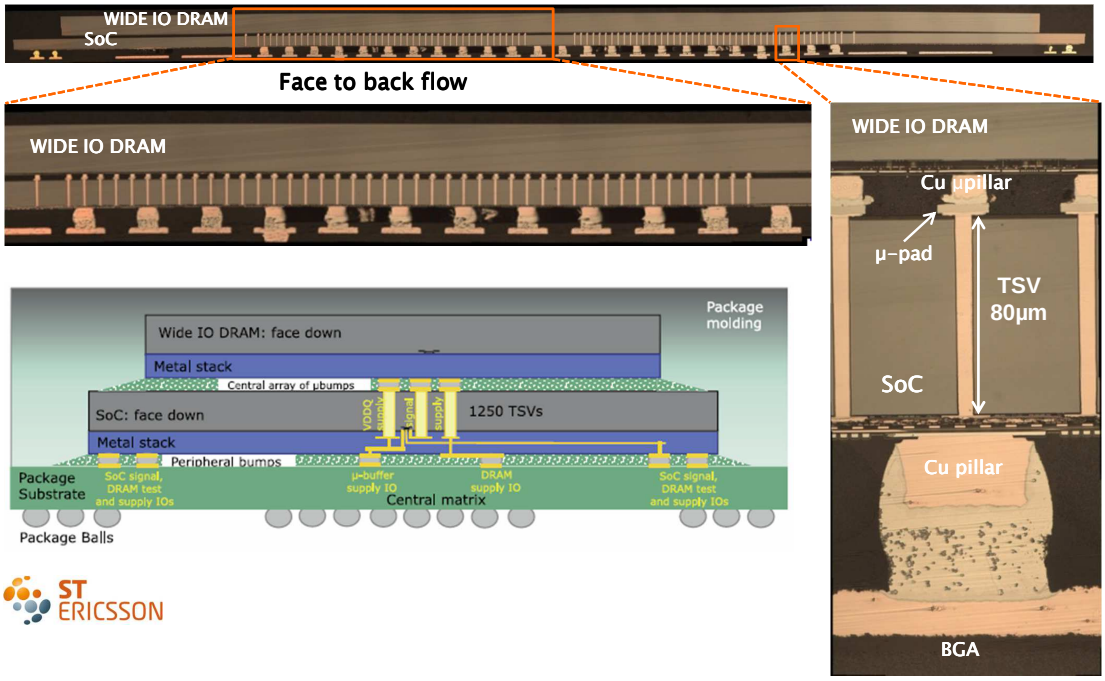


Figure 1.7: WIOMING chip from ST-Ericson, ST-Microelectronics and CEA-Leti.

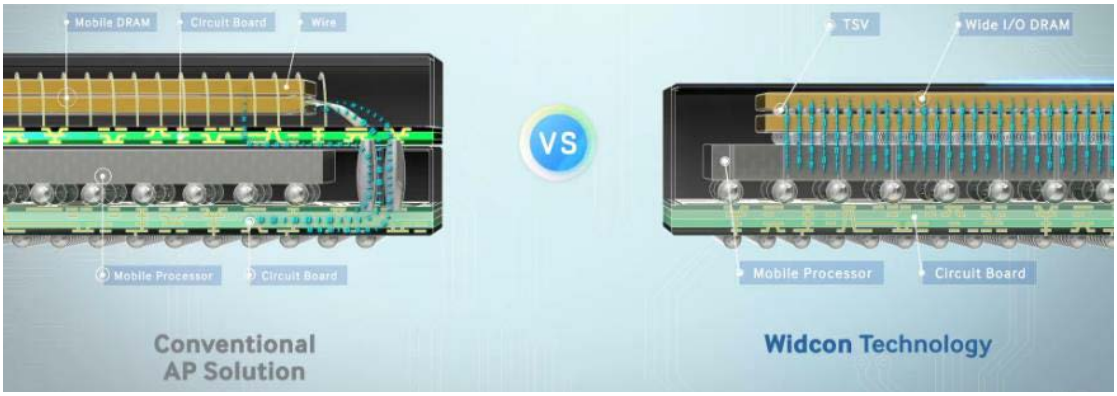


Figure 1.8: Widcon technology from Samsung.

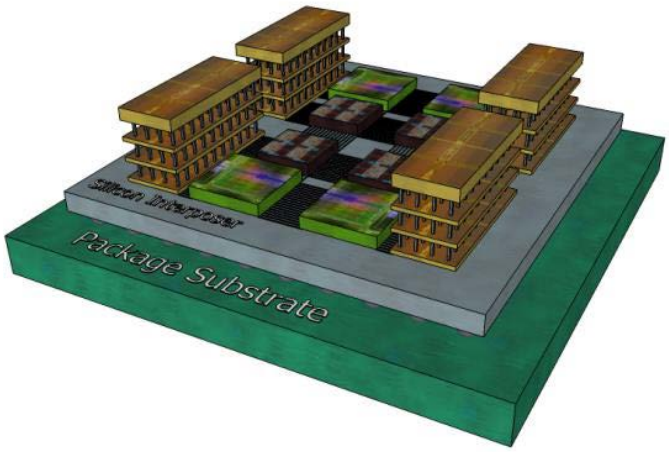


Figure 1.9: 2.5D+3D IC. Courtesy of Hsien-Hsin Lee.

- **Logic on Logic** A different approach is called logic-on-logic stacking and involves splitting a logic area between two or more strata. In particular, partitioning a microprocessor into multiple layers to reduce interconnect length. Block level partitioning is the simplest and most effective solution to rearrange a design onto multiple layers. Since the performance improvements linked to the 3D arrangement are directly dependent on the inter-strata via pitch and size, the via density limits the granularity at which a system can be divided. Block level of logic-on-logic stacking enables blocks to be moved closer in proximity, thus reducing the inter-block latency and power. The potentiality of this approach was explored by Morrow et al. [3] by splitting an Intel Pentium 4 family processor into two layers. Focusing on the layout arrangement of known performance sensitive pipelines several pipestages can be removed, eventually resulting in removal of approximately 25% of all pipestages. A 15% performance improvement from reduction of instruction execution latency and a 50% footprint reduction were demonstrated.

A further optimization can be envisaged by exploiting both the benefit of 3D integration and the interposer technology. This approach can be defined as $2.5+3D$ and is depicted in Figure 1.9. The Si-interposer can be used in order to integrate multiple 3D stacked system and 2D ICs. This approach is promising since it combines the extreme integration density and bandwidth offered by the 3D stacking technology with the ones of the 2.5D solution, such as better thermal management.

1.2 Objectives and contributions

3D-ICs can offer superior performance improvements over 2D equivalent. Nonetheless there are still several issues related to 3D design that need to be addressed before this technology can be widely adopted from the IC industry. This thesis work focuses on the two major TSV-related

challenges that can limit the benefits offered by 3D stacking technology: TSV area and delay overhead. According to the ITRS roadmap [2], TSV diameter will drop below the μm scale, remaining confined to 2-4 μm for 3D-SoC. A TSV occupies a huge amount of silicon area compared to metal via in sub-micron technologies, which can turn into a reduction of the wirelength benefit of 3D-ICs. TSV parasitic capacitance is small compared to long on-chip interconnects. However, whenever a design is simply partitioned and folded in 3D without further optimizations, just a part of the intra-layer signals experience a real length reduction, while the rest may cope with a delay overhead compared to the 2D implementation. The eventual delay overhead on the 3D signal paths should be compensated by the insertion of buffers. Nevertheless, the additional silicon area required for buffer insertion and the related power overhead would be the price to pay. The TSV capacitance is dependent not only on the TSV diameter and height, but also on parameters related to the fabrication process, such as the oxide thickness and the doping concentration of the substrate. Hence smaller TSVs do not necessarily have smaller capacitance.

The aim of this thesis is to find design solutions that leverage the high bandwidth provided by TSV links minimizing the cost in terms of silicon area and capacitance load in order to meet the design requirements. Depending on the target application, different interconnection topologies have been explored and several solutions have been envisaged in this thesis.

First, a configurable network architecture exploiting a fully parallel TSV bus is presented. The architecture of the 3D interconnect has been optimized to be integrated in a 3D stacked *Chip Multi Processor(CMP)* featuring a shared L1 memory which provide a convenient shared memory abstraction while avoiding cache coherence overheads. The performance of the tightly-coupled processor cluster critically depends on the architecture of the interconnect between the processors and the memory banks, which should provides ultra-fast access to the largest possible L1 working set. The proposed 3D network guarantees single-cycle communication . The 3D implementation and the reconfigurability of the proposed network expand the storage capability of the system while still guaranteeing single cycle memory access time. The exploration of the trade-off between memory size and network latency for different partitioning choices demonstrates the potential of the proposed solution.

Using a TSV for each inter-layer signal may be extremely expensive in terms of silicon area for designs that require a high number of 3D connections. Moreover, since the TSV fabrication technologies are not yet mature, it may be required to implement hardware redundancy in order to improve yield [18], [19], while thermal TSVs have been already proposed to dissipate the internal heat of the 3D system [13] [14] [15]. In order to reduce the silicon area occupation exploiting the TSVs' excellent frequency properties, a high data-rate 3D serial link has been envisaged. Since 3D interconnects offer a reduced load compared to off-chip channels, high speed serial transmission through TSVs does not require complex and power-hungry equalization techniques, achieving high bandwidth with low silicon area and power. Low power *Serializer-Deserializer (SERDES)* circuits for inter chip 3D links have been designed and characterized for a variety of state-of-the-art TSV channels. The proposed serial 3D link has

no performance loss compared to a low frequency parallel 3D link. The effect of serialization on both area and energy for different TSV technologies has been analysed.

Once integrated in a complete system, the reduction in area demonstrated by the 3D serial solution also affects the chip routing. A TSV interferes with cell placement and, depending on the adopted technology, may become a routing obstacle. The reduction in the number of 3D vias obtained with the adoption of the serial vertical connection can relieve the routing congestion of the 3D system. In order to quantify the benefit of serialization on the routing of a 3D system, a *3D Modular Multi-Core (3D-MMC)* architecture has been designed and used as a test case. This innovative multi-processor platform is composed of completely identical stacked chips following a logic-on-logic stacking approach and creating an expandable 3D network of processing cores to improve performance. Analysis of the routing characteristic of the placed and routed layouts reveals an improvement in the average wirelength due to the serialization.

Finally, we present a test vehicle demonstrating the efficiency and applicability of a 3D serial link in a complete multi-processor system based on the 3D-MMC architecture. The prototype has been designed, fabricated using a UMC 90nm CMOS foundry process, tested, and the KGD has been vertically stacked using an in-house via-last TSV process. A comprehensive study of the system is presented together with a software approach to optimize the applications execution time. The system exhibits multiple Gbps vertical data bandwidth while limiting the number of TSVs. The experimental results obtained from simulations and measurements on the fabricated samples are provided.

1.3 Thesis organization

The remainder of this thesis is organized as follows.

- In Chapter 2, details of TSV fabrication technologies including the state-of-the-art TSV available from both foundries and research laboratories are presented. The analytical model of the TSV channel is then proposed and validated with measured in-house and literature data. This model will be fundamental to designing functional 3D circuits that meet the expected performance.
- Chapter 3 explores how to interconnect a Memory-on-Logic 3D stacked system. We propose a configurable 3D-Logarithmic Interconnection Network (3D-LIN) that exploits a parallel bus of small TSVs connecting multiple memory dies to a logic layer. The network is based on the 2D mesh-of-tree topology [20] and guarantees low-latency connection among multiple processing elements and a multi-banked memory.
- Chapter 4 investigates the possibility of producing high bandwidth, low delay inter-chip connection reducing the cost in silicon area due to the area footprint of the TSVs. An low power SerDes circuits for inter-chip 3D links is proposed. The area, performance

and energy efficiency of different serialization levels are explored for a variety of state-of-the-art TSV technologies and 3D systems.

- In Chapter 5 we explore the benefits of the proposed serial vertical interconnection on the routing congestion for a 3D chip multi-processor systems. To this end, a homogeneous multi-core architecture, 3D-MMC, has been designed. We first describe the architectural features of the 3D multiprocessor platform; then, we present a comparison between the routing characteristics of the parallel and the serial 3D communication solutions.
- In Chapter 6 we present a test vehicle based on the 3D-MMC architecture, MIRACLE. The test vehicle has been fabricated in 90nm UMC CMOS technology and stacked using an in-house TSV process. The results from a fabricated and tested test vehicle are presented.
- The conclusions of this thesis work are drawn in Chapter 7.

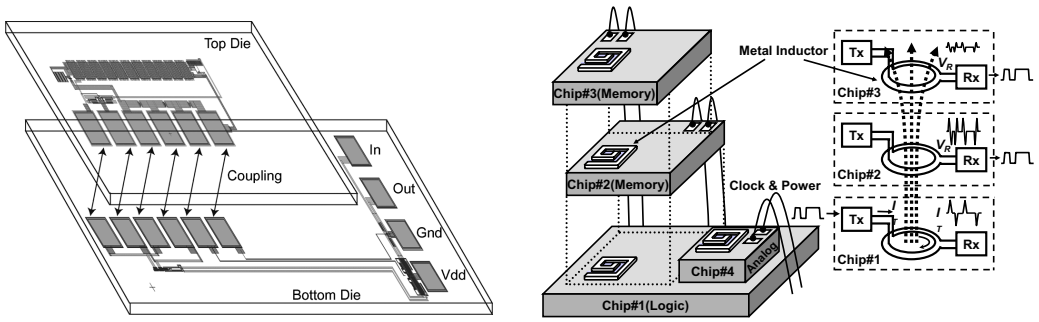
2 Through Silicon Vias Technology

With continued technology scaling, interconnect has emerged as the dominant source of circuit delay and power consumption. As discussed in Chapter 1, 3D stacking technology offers a promising architectural solution to overcome the on-chip interconnect bottleneck and integrate more functionality on the same chip. Since 3D architectures started gaining the attention of the IC community, several vertical communication topologies have been explored, e.g., TSV technology and contactless solutions like capacitive or inductive coupling.

In capacitive coupled signaling [21], the interplane communication is provided by small metal plates on different stacked silicon dies which create a capacitive channel. While the capacitor can be driven by a simple buffer, the receiving circuit requires more complex circuits. The received low voltage signal needs to be amplified to produce a full swing output. The advantage of capacitive coupling methods is the simple channel modelling and the low crosstalk due to a more confined electrical field. However, the communication distance is limited to several microns. In order to extend the communication range, the voltage across the capacitor should be increased, hence connecting more than two layers may reveal itself as a challenging task.

Wireless inductive coupling [22] methods rely on the coupled magnetic field between spiral inductors on the two stacked layers located at the same horizontal coordinates.The signal propagation is achieved through current pulses which generate magnetic flux inducing an electromagnetic induction in the receiving coil. With inductive coupling, the communication distance is not as constrained as for capacitive coupling, and signals can be propagated through more than two planes. However, the communication strength depends on the current in the transmitting coil and the coupling coefficient k between the coils, which is proportional to the coil size of the inductor. Hence, connecting more layers can be achieved either by consuming higher power consumption or by occupying a larger area.

Nonetheless, TSV-based 3D ICs have emerged as the most promising to achieve the desired vertical interconnect density. A TSV is a conductive connection between the two sides of a silicon waver or chip which can transmit a signal from a tier to another. The fabrication of 3D ICs involves three major processing steps: TSV fabrication, wafer/chip thinning and bonding.



(a) (b)

Figure 2.1: (a) Capacitive coupling [21] and (b) Inductive inter-chip signaling [22] for 3D ICs.

The order of these steps may vary depending on the fabrication choices.

2.1 TSV fabrication technologies

TSVs can be categorized in different groups according to the process flow.

- **Via First** TSVs are fabricated before the *Front-End-Of-Line (FEOL)* CMOS processing;
- **Via Middle** TSVs are fabricated after the Si front-end (FEOL) device processing but before the *Back-End-Of-Line (BEOL)* interconnect process;
- **Via Last** TSVs are fabricated after the *Back-End-Of-Line (BEOL)* interconnect process.

In the case of via first TSVs, the advantage is that tiny and dense TSVs can be fabricated. However, the conductive material for the filling can just be polysilicon, since metal cannot withstand high temperatures and would not be compatible with the subsequent steps of the technological flow. The drawback of polysilicon vias is the higher resistivity compared to metal vias, which can be a major obstacle for certain applications. On the other hand, via last TSVs occupy more Si area, but can be filled with low resistive material. Moreover, choosing a via-last approach is extremely interesting since it allows the fabrication of the planar chip in a conventional foundry, then the stacking process can be performed separately. TSVs can be also discriminated depending on the order of TSV processing and 3D-bonding.

In terms of 3D-stacking process, there are 3 possibilities:

- **Die-to-Die (D2D)** bonding: complete wafers are stacked together before the the single 3D chip are sliced;
- **Die-to-Wafer (D2W)** bonding: a single die is stacked on top of another die integrated in a wafer;

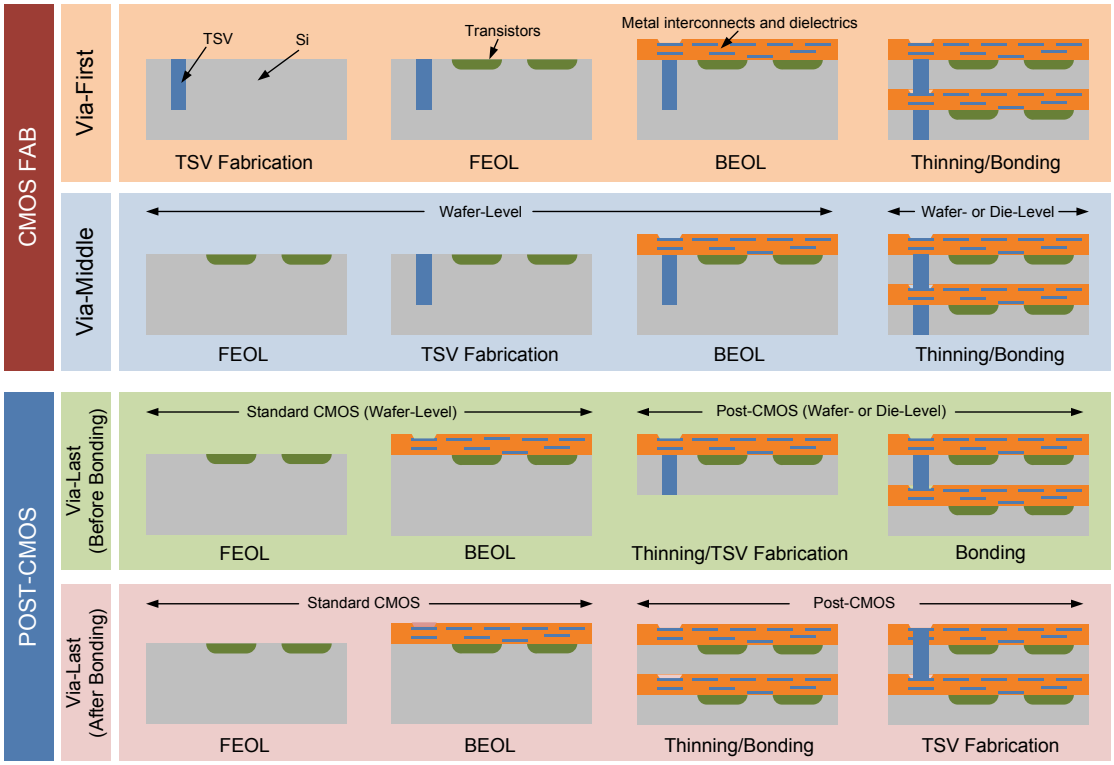


Figure 2.2: Summary of the 3D integration scenarios based on the TSV type [23].

- **Wafer-to-Wafer (W2W)** bonding: single dies are stacked together.

W2W bonding is the simplest in terms of processing, however, integrating more layers leads to a yield loss due to stacking untested faulty dies. D2W bonding allows testing the individual dies before the assembly. D2D bonding is the most expensive solution in terms of fabrication, nevertheless, it is extremely interesting since the KGD from different foundries can be integrated in a 3D system allowing more freedom in the design. The bonding can be performed *Face-to-Face (F2F)* or *Face-to-Back (F2B)*. In F2F processing, the two wafer, or dies, are stacked so that the top metal layers are connected while TSVs are used for the external I/Os. This solution provides the shortest interconnects between the dies, nonetheless it is restricted to maximum two layers. With F2B bonding, on the other hand, multiple device layers can be stacked together with the top metal layer of one die bonded with the substrate of the other die. TSVs are used for the inter-layer connections and the system is no more restricted in the number of stacked dies. The different 3D integration scenarios are summarized in Figure2.2.

Depending on the fabrication process, the physical and electrical characteristics of a TSV can vary substantially. Although using smaller vias is desirable in order to reduce the chip footprint, the minimum TSV diameter is limited by the process yield. Unsuccessful wafer alignment during the bonding process and handling of very thin silicon wafers are the primary mechanisms of TSV failure, with a minimum TSV diameter of 1.2 μm reported in the open

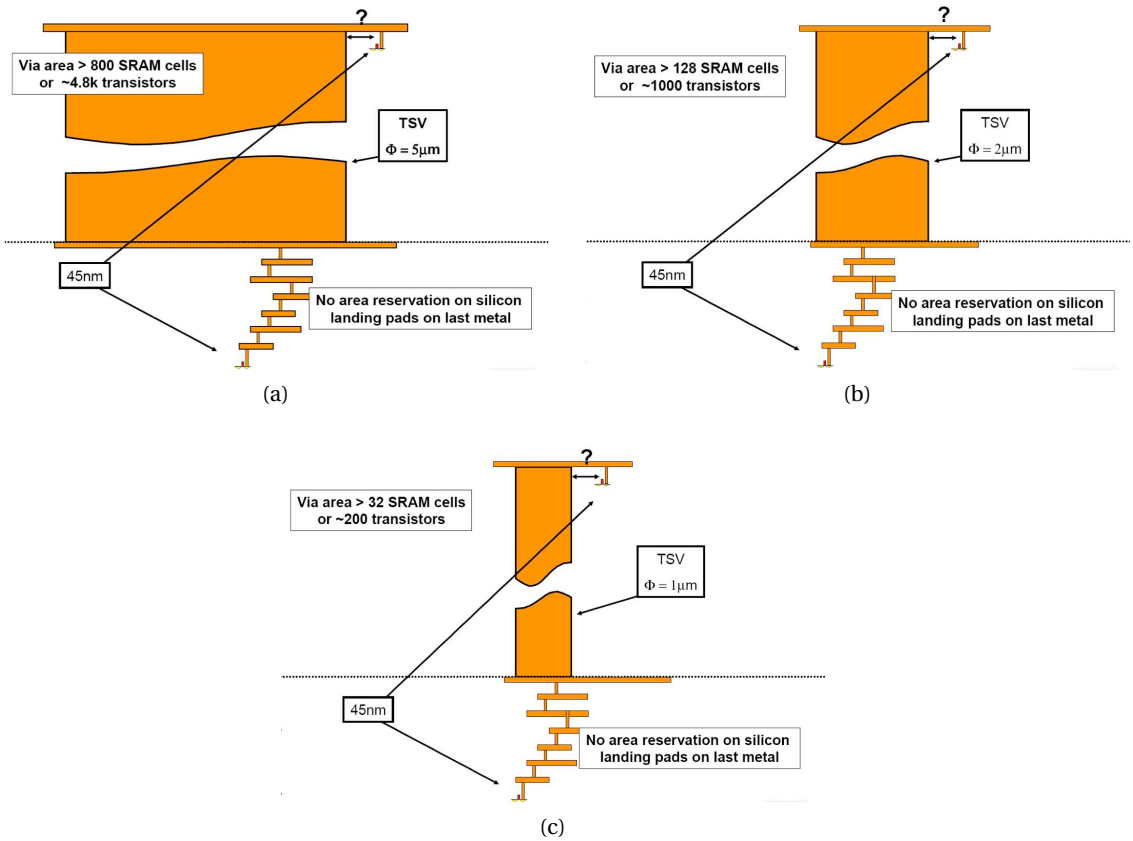
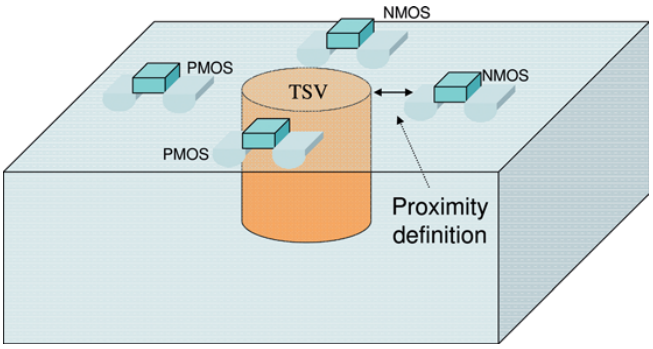


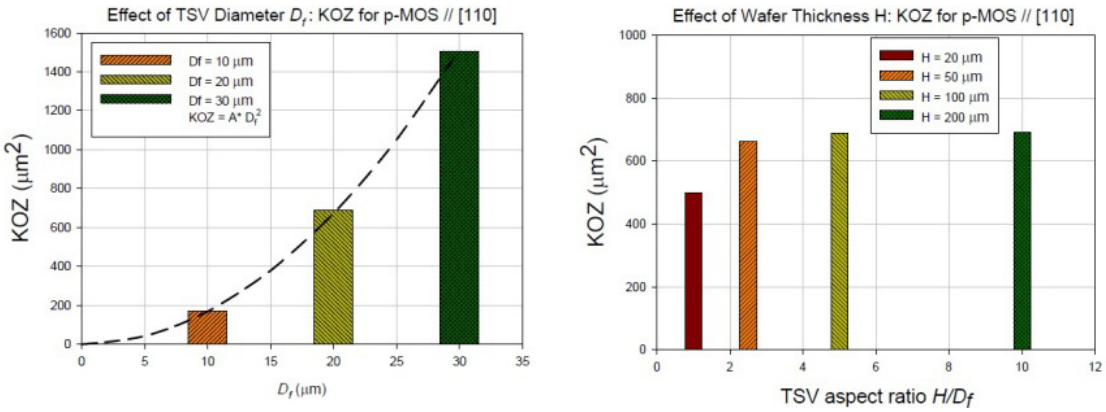
Figure 2.3: Typical TSV footprint compared to FEOL structures in a 45 nm CMOS process for 5 μm, (a), 2 μm (b) and 1 μm (c) TSV diamater [24].

TSV parameters	Intermediate Level		Global level	
	2011-2014	2015-2018	2011-2014	2015-2018
Minimum diameter [μm]	1-2	0.8-1.5	4-8	2-4
Minimum pitch [μm]	2-4	1.6-3.0	8-16	4-8
Minimum depth [μm]	6-10	6-10	20-50	20-50
Maximum aspect ratio [μm]	5:1-10:1	10:1-20:1	5:1-10:1	10:1-20:1

Table 2.1: ITRS roadmap 2011 [2]



(a)



(b)

Figure 2.4: a)Simplified TSV interconnect KOZ requirements and b) KOZ for different TSV diameters (on the left) and TSV height (on the right) [25].

literature [8]. Note that even a 1 μm^2 TSV fabricated in a 45 nm CMOS process occupies a silicon real estate corresponding to 32 SRAM cells or ≈200 transistors [24] as depicted in Figure 2.3. According to the ITRS roadmap [2], the diameter of TSVs connecting at the global interconnect level for the 3D stacking of IP-blocks, will not shrink below 2-4 μm. In the case of the stacking of smaller circuit blocks, for interconnects at the intermediate level, the minimum diameter will be limited to 0.8-1.5 μm. A summary of the ITRS roadmap for the TSV technology is depicted in Table 2.1.

Since TSV fabrication causes tensile mechanical stress around the via hole [26] because of the mismatch in the thermal expansion coefficient between silicon (2.6 ppm/°C at 20°C) and the metal filling the via, usually copper (16.7 ppm/°C at 20°C). After cooling down to room temperature, Copper contracts much faster pulling the surface of the surrounding silicon substrate. Hence, the caused stress can result in hole and electron mobility variation, which may cause performance degradation of the closest devices. For this reason, transistors should be placed at a safe distance from the TSV so that they are not influenced by the TSV-induced stress. The logic-forbidden area surrounding each TSV is called *Keep Out Zone (KOZ)*. The

	CEA-LETI [28]	IMEC [29]	Samsung [30]	IBM [31] [32]
Diameter [μm]	4	5	7.5	20
Haight [μm]	15	25	-	100
KOZ [μm]	3	-	-	2
Pitch [μm]	12	10	40	50
R_{TSV} [Ω]	0.3	0.2	0.22-0.24	-
C_{TSV} [fF]	30	37	47.4	200
	via-last	via-first	via-last	via-middle

Table 2.2: Fabricated TSV details

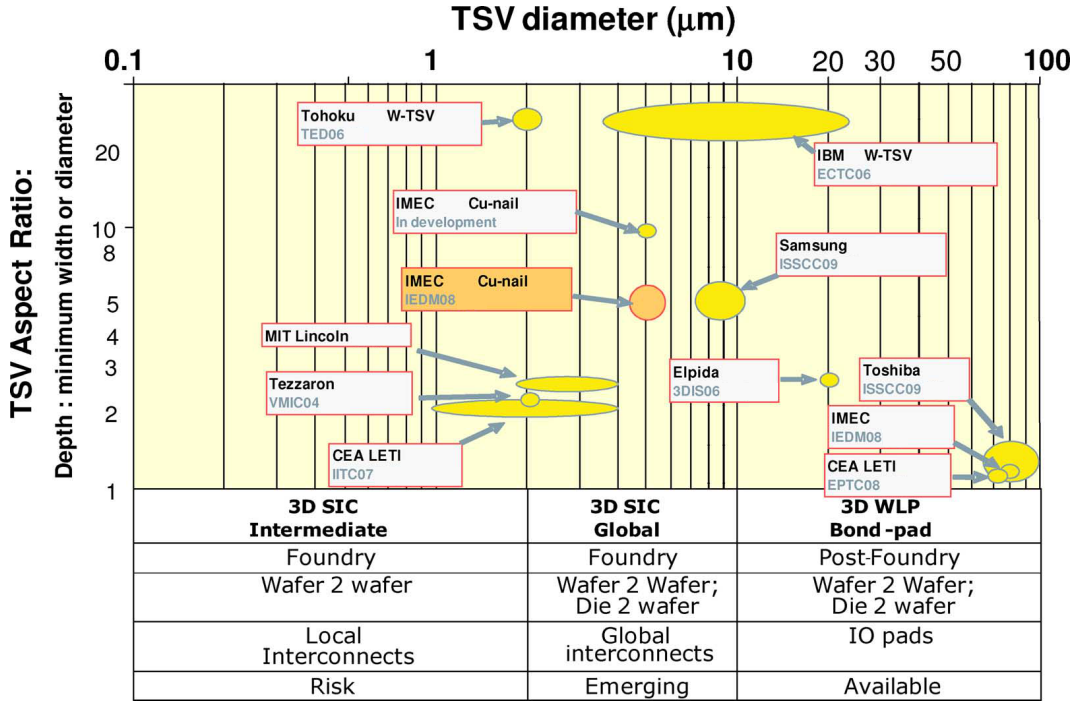


Figure 2.5: Overview of the 3D TSV technologies as function of the TSV diameter and aspect ratio [33].

KOZ is mainly defined by the TSV diameter, rather than the TSV height [25] as depicted in Figure 2.4b. Previous experiments demonstrated that a CMOS transistor should be kept around 6.7 μm away from a 20 μm TSV [27].

Table 2.2 summarize the main characteristics available 3D TSV technologies developed by some of the leading IC companies and research centres such as IBM, Samsung, IMEC, and CEA-LETI. In particular, in Figure 2.5 [33] the different TSVs are classified among three categories. The large size TSVs with diameter larger than 10μm are typically manufactured post-foundry and are compatible with both W2W, D2W and D2D stacking schemes. The medium size TSVs,

with diameter ranging from 2 to 10μm are manufactured at the foundry and are compatible with W2W and D2W stacking schemes. The smallest size TSVs, with diameters less than 2μm are an emerging technology. Even with high aspect ratios, the wafer and die handling are extremely challenging due to the extremely low thickness. Hence, the stacking is typically done W2W and the thinning is performed after bonding. Small TSVs are interesting in order to reduce the silicon area assigned to the vertival connections, however the fabrication challenges that arise with the shrinking of the TSV dimensions cause a drop in the process yield.

2.2 TSV analytical model

A single TSV can be described by by the metal core resistance R_{tsv} , the self-inductance L_{self} and the parasitic MOS capacitor between the TSV and the substrate, as shown in Figure 2.6. The set of equations, for the model, are given for a cylindrical TSV [34] [35].

The resistance of a TSV is composed by a DC-term and an AC-term due to the skin effect and can be modeled as:

$$R_{tsv} = R_{dc} + R_{ac} \tag{2.1}$$

where

$$R_{dc} = \frac{\rho l_{tsv}}{\pi r_m^2} \tag{2.2}$$

and

$$R_{ac} = l_{tsv} \frac{\sqrt{\pi \mu f \sigma}}{r_m \sigma} \tag{2.3}$$

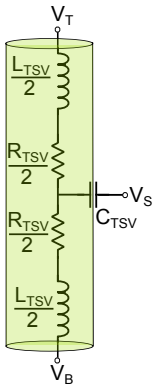


Figure 2.6: Electrical model of a single TSV channel.

and l_{TSV} is the TSV length, r_m the metal radius, f the frequency, μ the magnetic permeability and ρ the resistivity of the metal.σ the electrical conductivity of the metal. In Equation 2.2 the bulk Cu resistivity, ρ is process dependent and thus needs to be taken from the TSV process parameters.

The self-inductance of the TSV, L_{self} [36], is expressed in as a combination of two terms: l_tsv , and $f(b)$:

$$L_{self} = l_{tsv} f(b) \tag{2.4}$$

where

$$f(b) = \frac{\mu}{2\pi} ln\left(2b^{-1} + \sqrt{(0.5b)^{-2}+1} + 2b^{-1} - \sqrt{(0.5b)^2+1}\right) \tag{2.5}$$

and

$$b = \frac{2r_m}{l_{tsv}} \tag{2.6}$$

Although the TSV resistance and self-inductance are small, the terms are directly added to the complete resistance and inductance of the *Redistribution Layer (RDL)* used to route signals between tiers in a 3D stack.

The characteristic of the TSV capacitance versus voltage can be derived from the parasitic MOS capacitor model. The TSV MOS capacitor is obtained by solving Poisson's equation in a cylindrical coordinate system. The electric charge distribution does not vary with the angle around the TSV, nor along the length of the TSV, hence, taking into account the symmetry of the system, it is sufficient to solve a 1-D Poisson's equation in the radial direction. The Poisson Equation in cylindrical coordinate system with p-type substrate can be reduced to:

$$\frac{d^2\psi}{dr^2} + \frac{1}{r}\frac{d\psi}{dr} = -\frac{\rho}{\epsilon_{si}} \tag{2.7}$$

where r is the radial distance, ψ the potential, ρ the charge density and ϵ_{si} the dielectric constant of silicon. It is also assumed that all doping atoms are ionized and that they are the sole source of electric charge in the depletion layer. Considering the work function of the metal and the semiconductor equal, the equation becomes:

$$\frac{1}{r}\frac{d}{dr}\left(r\frac{d\psi}{dr}\right) = \frac{qN_a}{\epsilon_{si}} \tag{2.8}$$

where N_a is the doping level around the TSV and q the elementary charge. This approximation provides a trade-off between simplicity and accuracy. Using the exact charge density improves the accuracy of the model, but is complexity increases significantly. The surface potential of the silicon is obtained by integrating Equation 2.8 with the boundary condition that the surface potential (ψ) and the electric field (E) at the depletion radius R_{dep} is zero.

$$\psi(r) = \frac{qN_a}{2\epsilon_{si}} \left[(r + w_d)^2 ln\left(\frac{r + w_d}{r}\right) - \frac{(r + w_d)^2 - r^2)}{2} \right] \tag{2.9}$$

Once the surface potential has been calculated, the equations for the MOS capacitor can be derived. In the accumulation region, the capacitance is constant and can be expressed by the logarithmic expression:

$$C_{acc} = C_{ox} = \frac{2\pi\epsilon_{si} l_{tsv}}{ln\left(\frac{r_m + t_{ox}}{r_m}\right)} \tag{2.10}$$

where t_{ox} is the dielectric thickness. As the TSV gate bias increases, the depletion capacitance acts in series with the oxide capacitance and the total capacitance is the series combination of the oxide capacitance (Equation 2.10) and depletion capacitance (Equation 2.11):

$$C_{dep} = \frac{2\pi\epsilon_{si} l_{tsv}}{ln\left(\frac{r_m + t_{ox} + w_d}{r_m + t_{ox}}\right)} \tag{2.11}$$

$$C_{tsv} = \frac{C_{dep} C_{ox}}{C_{dep} + C_{ox}} \tag{2.12}$$

The accumulation and depletion capacitance expressions show that the C_{tsv} is directly proportional to the length of the TSV and inversely proportional to the TSV dielectric thickness. In fact, for a thinner oxide C_{ox} increases, hence becoming less significant in the series with the depletion capacitance. C_{tsv} can also be reduced by using highly resistive substrate, in fact lower doping concentrations increase the depletion width, hence reducing C_{tsv}.

The minimum depletion capacitance is reached when the depletion radius reaches is maximum, hence $w_{dep} = w_{dmax}$. In order to calculate w_{dmax} the maximum surface potential should be calculated first:

$$\psi_s = 2V_T ln\left(\frac{N_a}{n_i}\right) \tag{2.13}$$

Then ψ_s can be substituted in Equation 2.9 to obtain w_{dmax}.

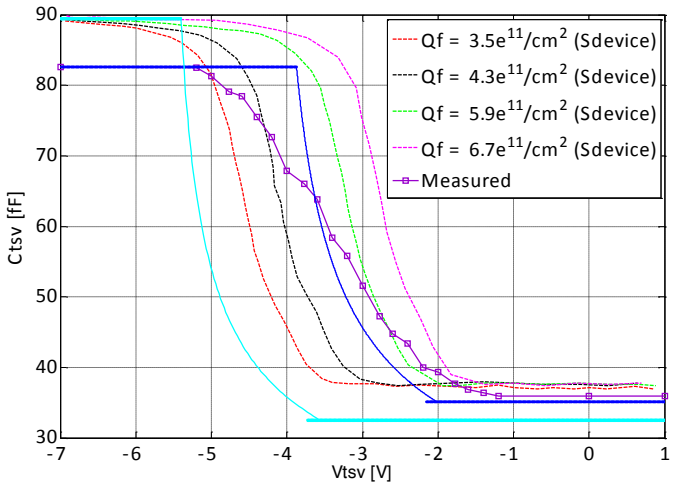


Figure 2.7: TSV C-V curve computed with Sdevice simulator (dashed lines), measurements [34] (squares) and the proposed analytical model (solid lines).

	$r_m[\mu m]$	t_{ox}[nm]	$l_{TSV}[\mu m]$	$N_A[cm^{-3}]$	$Q_{TOT}[cm^{-3}]$
without process variations	2.38	128	20	$2{\cdot}10^{15}$	$6.7{\cdot}10^{11}$
with process variations	2.38	118.2	20	$3{\cdot}10^{15}$	$6.7{\cdot}10^{11}$

Table 2.3: Process variations

By definition, the MOS capacitor enters the depletion region when the gate voltage pass the flat-band voltage, while the maximum depletion is achieved at the threshold voltage. The ideal flat-band voltage is:

$$V_{fb} = \phi_{ms} \tag{2.14}$$

where ϕ_{ms} is the work function difference between the metal and the silicon. However non-idealities, such as dielectric charges, trapped charges and interface traps, modify the V_{fb}:

$$V_{fb} = \phi_{ms} - \frac{2\pi r_{ox} l_{TSV} q Q_{tot}}{C_{ox}} - \frac{2\pi r_{ox} l_{TSV} q Q_{tot} Q_{it} \phi(r)}{C_{ox}} \tag{2.15}$$

The expression for the TSV MOS capacitor threshold voltage V_T can be calculated from the link between the surface potential and the applied voltage:

$$V_T = V_{fb} + \psi_s + \frac{q N_a \pi l_{tsv}[(m + t_{ox} + w_d)^2 - (r_m + t_{ox})^2]}{C_{ox}} \tag{2.16}$$

Figure 2.7 shows the effect of process variations on the MOS V_T, which can cause a shift in the C-V characteristic of the TSV. In particular, the dark blue solid line corresponds to a TSV

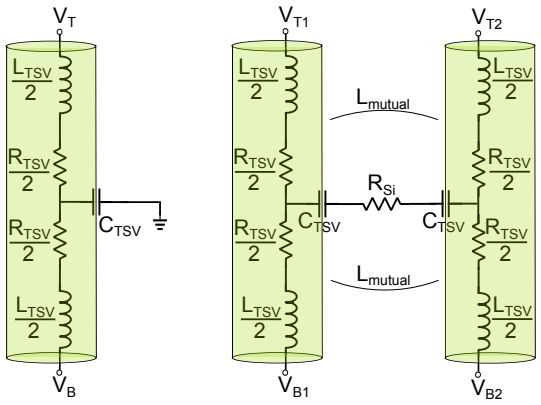


Figure 2.8: TSV coupling a) directly to ground, b) to a neighbouring TSV.

with the process variations summarized in Table 2.3. Compared to the nominal case reported by Katti et al. [34], the deviation of the oxide capacitance C_{ox} and the maximum depletion capacitance $C_{dep_{max}}$ are less than 5%. V_{fb} and V_T are within 10% and 40% of the measurement respectively, which is due to the charge distribution approximation used [37].

The equivalent electrical circuit of the TSV is depicted in Figure 2.6: the resistance R_{TSV} (Equation 2.1) and self-inductance L_{TSV} (Equation 2.4) are connected in series causing the voltage drop along the interconnected nodes between the top and the bottom tier. The parasitic MOS capacitor C_{TSV} (Equation 2.12) is connected between the TSV metal core and the substrate. The potential of the substrate, V_S, depends on the nearby structures. Assuming that the MOS capacitor terminal connected to the substrate is grounded, as in [34], the final TSV model is depicted in Figure 2.8a.

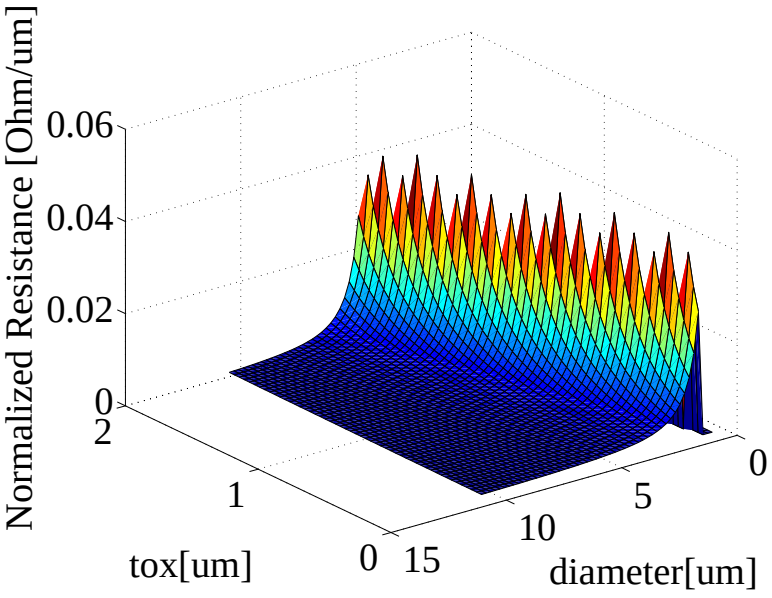
The model extension including coupling between two TSVs is also shown in Figure 2.8b. The coupling occurs over the TSVs mutual inductance L_{mutual} and through the silicon substrate. The resistance of the substrate can be calculated as:

$$R_{Si} = \frac{\rho d}{2 r_m l_{TSV}} \tag{2.17}$$

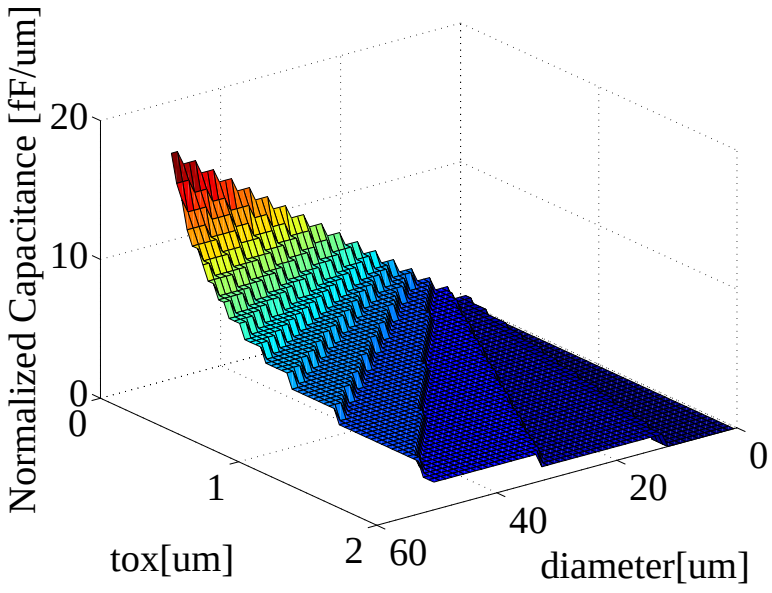
being ρ the resistivity of the bulk. The TSV mutual inductance can be estimated using Equation 2.4, where

$$b = \frac{2d}{l_{TSV}} \tag{2.18}$$

Since the Maxwell equations have not been solved, the magnetic induction was not taken in consideration, hence the model is valid up to a dozen of GHz [38]. Nevertheless, several groups have demonstrated TSVs working at 20GHz [39], 60GHz [40] and 170GHz [41].



(a)



(b)

Figure 2.9: Normalized TSV parasitics a) resistance and b) capacitance.

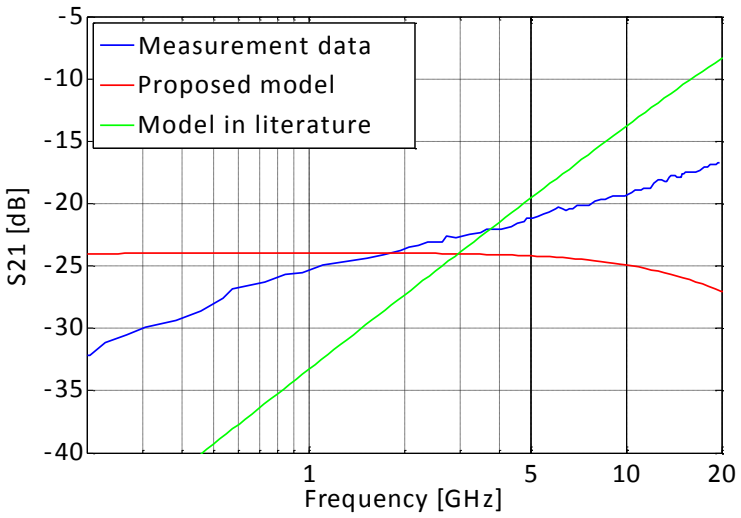


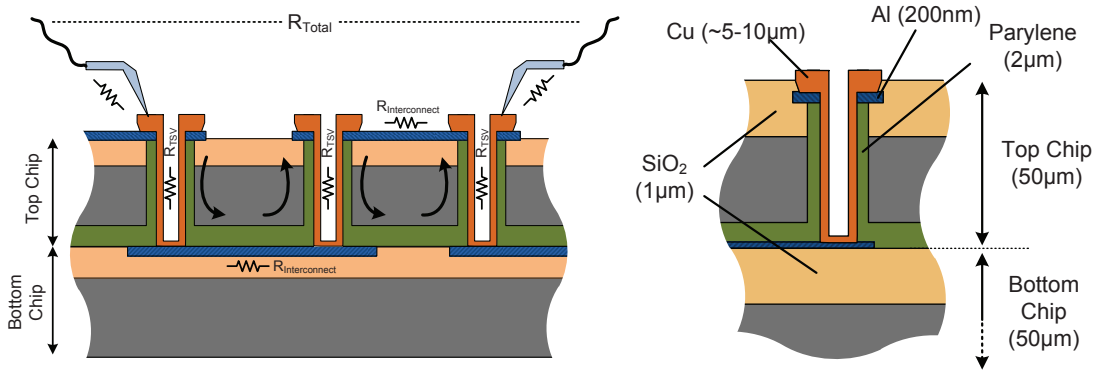
Figure 2.10: S21 parameter from the proposed model.

The analytical model demonstrates how the circuit parasitic elements depends on the physical parameters of the TSV and on the material characteristics. In particular, the effect of varying diameter and oxide thickness on R_{TSV} and C_{TSV} normalized to the TSV height is shown in Figure 2.9 for highly doped substrates. We notice that each of the three parameters, r_m, t_{ox} and l_{TSV}, significantly affects the TSV parasitics. The capacitance increases almost linearly for a thinner oxide and for larger diameters. For large diameter TSVs the resistance is almost constant, but as the diameter approaches the nanometer scale, R_{TSV} has a sharp increase. The diameter at which this abrupt increase is observed depends on the t_{ox} value: for large t_{ox} the resistance increases significantly for TSV diameters smaller than $7\mu m$, while for thinner oxide R_{TSV} increases significantly for diameters smaller than $2.5\mu m$. The spikes in Figure 2.9a are just artefacts due to the resolution of the simulation.

2.2.1 Model validation

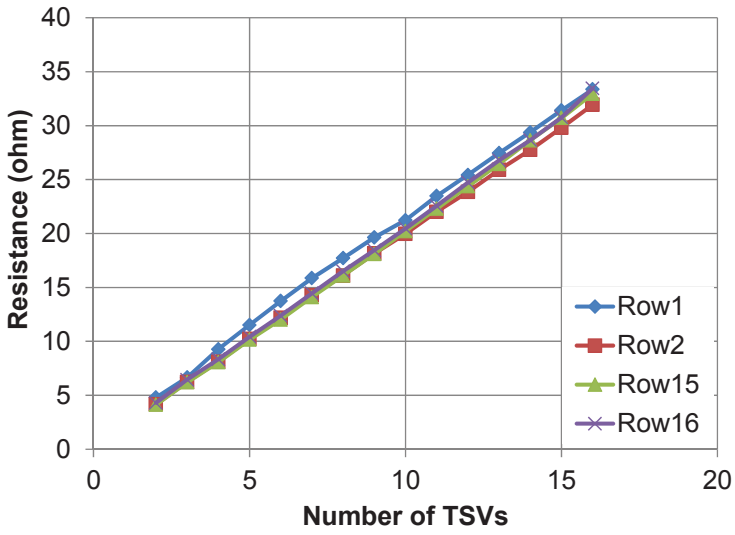
The validity of the proposed model was first verified comparing it to measured results in the literature. The MOS capacitance C_{ox} from the proposed model is depicted in Figure 2.7 in comparison with the measured results and Sdevice simulations from [34]. The proposed model gives the same estimation as the Sdevice model and is within 10% of the measurements. The maximum depletion capacitance is underestimated by 10% compared to the measurements.

Also the cross talk model simulation results are compared to the ones presented in the literature [42]. Figure 2.10 shows the S21 parameter for the cross-talk of two TSVs. The s-parameter of the proposed model (red line) can be compared against the one from a model in literature (green line) [42] and the measured data of a crosstalk emulation model (blue line) [42]. The plot shows that below 5GHz the proposed model is more accurate.



(a) (b)

Figure 2.11: a) Illustration of the two-tier chip stack used for the daisy-chain resistance measurements and b) illustration of the TSV used [23].



(a)

Figure 2.12: a) Resistance of 4 rows after 64 resistance measurements [23].

A test chip was designed to allow the measurement of daisy chains of arbitrary length, as illustrated in Figure 2.11. The test chip composed by blocks of 900 TSVs in a daisy chain was fabricated using in-house TSV process [23]. TSVs with 60μm diameter, 50μm length filled with copper were fabricated and interconnected through Al metal lines. The total measured resistance is the sum of the series-connected TSVs, the Al interconnections and the off-chip connections. In order to extract the average TSV resistance, the Al interconnect resistance is estimated around 1.5Ωbased on its geometry. In Figure 2.12, 64 TSVs in 4 rows are measured and plotted. Subtracting the total contribution of the interconnections resistance, the total resistance of a single TSV is calculated as 0.5Ωon average. Impedance and phase were also measured for frequency up to 30 MHz, due to limitation of the measurement setup. For the measured frequency range the proposed model parameters are within 10% of the measurements [37].

2.3 Summary

This chapter sets the background with the state of the art TSV fabrication technologies. The TSVs physical and electrical characteristics strongly depend on the adopted stacking technology. The different TSV fabrication processes are introduced and an overview of the fabricated TSV from major companies and research institutes is presented.

In the second part of the chapter, compact analytical model is proposed and validated against simulated and measured results from both literature and in-house fabricated TSVs. The model allows a quick and accurate exploration of the TSV performances and their impact on the electronic circuitry of 3D systems.

3 3D-LIN: a Logarithmic Network for Inter-Layer Memory to Processor Communication

A promising option to overcome the barrier in interconnect scaling is the 3D integration of integrated circuits (3D ICs)[43]. Stacking multiple chips and connecting them by *Through Silicon Vias (TSVs)* has the potential to reduce the interconnect wirelength while offering high vertical connection density. Multi-cores and many-cores processors can benefit from several characteristics of 3D devices: (a) Wirelength reduction improves the latency of core to memory interconnects; (b) High TSV density and their small length can be exploited for improving memory bandwidth when stacking memory layers on top of logic layers; (c) The smaller form factor due to the addition of a third dimension is essential for moving on-chip the memory required by the processing elements, therefore avoiding slow off-chip connections.

This chapter presents a configurable network architecture exploiting a fully parallel TSV bus. The architecture of the 3D interconnect has been optimized to be integrated in a 3D stacked *Chip Multi Processor(CMP)* featuring a shared L1 memory, which provides a convenient shared memory abstraction while avoiding cache coherence overheads. The performance of the tightly coupled processor cluster critically depends on the architecture of the interconnect between the processors and the memory banks, which should provide ultra-fast access to the largest possible L1 working set.

The proposed 3D network guarantees single-cycle communication. The 3D implementation and the reconfigurability of the 3D network expand the storage capability of the system still guaranteeing single cycle memory access time. The exploration of the trade-off between memory size and network latency for different partitioning choices demonstrates the potential of the solution.

3.1 Problem formulation

Following Moore's law, the scaling to nanometer technologies has led to a transition from single-core to multi-core processors, and is now moving towards many-cores architectures [44]. Whereas hundreds of millions of transistors can now be placed on a single chip leading

to increased computing power, they cannot be fully exploited due to interconnect latency. In nanometer-scale technologies, interconnect latency and power do not scale as much as device geometries, thus becoming a performance bottleneck, as explained in Chapter 1. These limiting factors need to be overcome at the architectural level.

For many applications, the exploitation of customized accelerators will be the way to obtain the highest performance, together with more efficient types of interconnect and memory hierarchies [45]. Shared *Level 1(L1)* memories are of interest for tightly-coupled processor clusters in programmable accelerators as they provide a convenient shared memory abstraction while avoiding cache coherence overheads. *Tight Coupled Data Memories (TCDMs)* are used since they yield much higher storage density per unit area, lower power consumption and lower access latency compared to cache memories [46]. Nevertheless, the performance of a shared-L1 memory critically depends on the architecture of the low-latency interconnect between processors and memory banks, which needs to provide ultra-fast access to the largest possible L1 working set.

For this reason, new interconnect architectures have already been envisaged. For instance, *Network-on-chip (NoC)* [47] has been adopted to substitute conventional bus-based systems when high bandwidth and high speed are required. When ultra-low latency processor to memory interconnection is requested for parallel computing, novel fast interconnect topologies are imperative to guarantee the access to the memory in few clock cycles. Several research efforts are already focused on low-latency, high-bandwidth connection between the processing elements and multi-banked on-chip memories. The *Mesh-of-Trees (MoT)* Interconnect Network proposed in [48], the Hyper-core architecture [49] and the single-cycle interconnect network presented in [20] are just few examples of low-latency networks.

At the same time, the increasing demand for storage capacity is challenging chip designers, which already make an extensive use of off-chip memories. Nonetheless, the bandwidth of future processors will continue to be restricted by the limited number of I/Os. According to the ITRS roadmap, the increase in the number of package pins will be limited by cost and power constraints, and the additional pins will be mainly dedicated to power delivery. Hence, future generations of *Chip Multi-Processor (CMP)* require a major innovation in both integration technology and on-chip communication infrastructure.

In this chapter, 3D integration is exploited to increase the shared L1 memory size in a modular fashion by stacking multiple layers of SRAM modules on top of a logic die, hence increasing the on-chip storage capacity. A fully synthesizable *3D Logarithmic Interconnect Network (3D-LIN)* is presented, the network takes advantage of the 3D configuration and the TSV channels to guarantee single cycle processor to memory communication. The network is conceived for connecting a cluster of processing elements, placed on the logic layer, to the SRAM modules placed on the memory layers. These modules constitute a single, on-chip, shared L1 memory that can enable fast communication among the tightly coupled processing elements avoiding cache coherence overheads. The network is configurable in both 2D and 3D-domains and is

automatically divided between the chosen number of memory layers. In order to reduce the chip cost, all the memory layers have the same layout and can all be produced exploiting the same mask. Design automation and configuration of the network allow us to experiment with different 3D structures, in the search for the trade-off points between speed, footprint and number of layers.

3.2 State of the art

In the last few years, several studies have been published exploring 3D integration technology in order to address the high area overhead of SRAM. A proposal from Li et al.[50], focuses on the L2 cache design and management in a 3D chip. They propose a network architecture embedded into the L2 NUCA cache memory for connecting it to a collection of cores. A different approach is followed by Loh, that in [51] considers 3D-DRAM stacked on top of multi-processors and revises the memory system organization in a 3D context. More recently, also Woo et al.[52], have explored a memory architecture that exploits TSVs for connecting the last level cache to the 3D stacked DRAM. The work of Madan et al.[53] instead, takes in consideration a 3D system composed by a DRAM layer and an SRAM cache banks layer on top of a processing layer. Considering emerging memory technologies, Mishra et al.[54] study the integration of STT-RAM in a multi-core system, together with a network level solution for decreasing the write latency associated with these novel memories.

In order to connect memory and logic placed on different layers, several groups already explored a methodology to extend NoC design into a 3D setting. The simple extension of traditional NoC fabrics to the third dimension adding routers at each layer (Symmetric NoC), does not pay in performance due to the different delay between fast vertical TSV and the horizontal interconnects. A first proposal has been done by Li et al. [50], with a network architecture embedded into the L2 cache memory. The use of Time-Division Multiple Access (dTDMA) buses as ”Communication Pillars” between the wafers is proposed in order to have single-hop communication amongst the layers. The 3D Dimensionally-Decomposed(DimDe) Router [55], focus on optimizing of the inter-strata communication with single hop connection between any two layers. Park et al. [56] propose a Multi-layered on-chip Interconnect Router Architecture (MIRA) divides the NoC between the multiple layers optimizing the micro-architecture for Non Uniform Cache Architecture (NUCA)-based CMP. A Low-Radix Low-Diameter 3D Interconnect Network is proposed by Xu et al. [57] which adopts long wires to connect remote intra-layer nodes and results in a 3 hops diameter network. More recently, Xue et al. [58] uses long range links to replace multiple short links in order to build a 5 hops 3D interconnect network for many core processors that exploits the DimDe router. While Ben Ahmed et al.[59] focus on overcoming the limitations in power, communication cost and throughput of their 2D OASIS-NoC by extending it to 3D.

3.3 3D parallel computing

Traditionally, the performance of a computer depends on the time required to perform a basic operation, which is ultimately limited by the clock frequency. Alas, the decrease of the clock cycle time is slowing down. In order to sustain the performance growth demanded by the market, the computer architecture is moving towards higher levels of parallelism. Parallel computing relies on the simultaneous use of multiple processing elements to solve a computational problem. Each processor works on its section of the problem and exchange informations with the other processing units. Multiprocessor system are widely adopted nowadays, while many-core systems such as programmable accelerators like *Graphic Processing Units (GPUs)* are attracting always more attention.

In parallel architectures, the processor to processor communication can be achieved either by message passing for distributed memory architectures or through a shared memory, the two topologies are sketched in Figure 3.1. Shared memory parallel computers vary widely, but generally have in common the ability for all processors to operate independently but share the same memory resources. Accessing all memory as global address space is extremely interesting since it is simplify the memory management and allow the system to have a uniform memory access time.

The hierarchical level of the shared memory can be chosen depending on the system requirements. Scratchpad memories yield higher storage density per unit area, lower power consumption and lower access latency compared to cache memories [46]. Hence, for computationally intensive applications, using a scratchpad memory can reduce the power consumption of the system while avoiding the overhead due to cache coherency problems.

Nevertheless, the disadvantage of a shared memory is the lack of scalability between memory and CPUs. Adding more CPUs or memories can geometrically increases traffic on the shared memory-CPU path reaching an interconnect bottleneck.

3D integration technology has the potential to overcome this interconnect problem by dividing the system into multiple layers and using TSVs as vertical connections. In addition, 3D integration technology offers a promising solution to increase the shared memory size in a modular fashion.

In this chapter, a shared L1 memory system is folded into multiple layers. Since the system

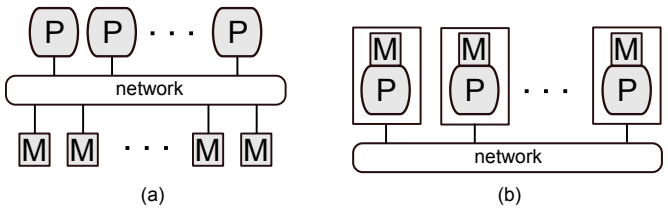


Figure 3.1: (a) Shared memory and (b) distributed memory architectures.

performances critically depends on the memory-to-processor communication, the following sections focus on the architectural and design aspects to extend a 2D network to a 3D low-latency network infrastructure.

3.4 2D network

The basic 2D-LIN is a low-latency and flexible crossbar that connects multiple masters to multiple SRAM *memory modules (MMs)*, as depicted in Figure 3.2. The IP is designed and optimized for sustaining full bandwidth and supporting non-blocking communication between the *Processing Elements (PEs)* and the MMs within a single clock cycle. The architecture of 2D-LIN avoids data replication providing also a simple and fast way for inter-processors communication and multi-core synchronization. These features makes LIN an interesting option for interfacing multi-processors to a shared TCDM constituted by multiple identical memory banks.

In order for the design to be simple and efficient, the interconnect is built following the *Mesh of Trees(MoTs)* approach, where the network is created combining binary trees. Each tree provides a unique combinational path between the processing element cluster and one memory module, and vice-versa. Aiming to sustain non blocking communication, the request and the response path must be decoupled, hence 2D-LIN features independent request and response network.

The key property of this soft IP is the reconfigurability. The user has control on a number of parameters:

- Number of processor channels, P;
- Number of *Direct Memory Access (DMA)* channels, D;
- Number of masters (PEs), N=P+D, that is a power of two;
- Number of memory cuts, M, that is a power of two. With a number of MMs at least double the number of PEs, access collision can be drastically reduced;
- Size of the memory cuts, all the banks should have the same size;
- Data and Address width;
- Enable for word level interleaving, for spreading transactions among all banks drastically reducing access collision.
- Test and Set bit. This bit act as enable for a test-and-set instruction used to write to a memory location and return the old value as a single atomic operation.

For the sake of simplicity, in the remainder of the chapter both the processing elements and the DMAs will be addressed as PEs.

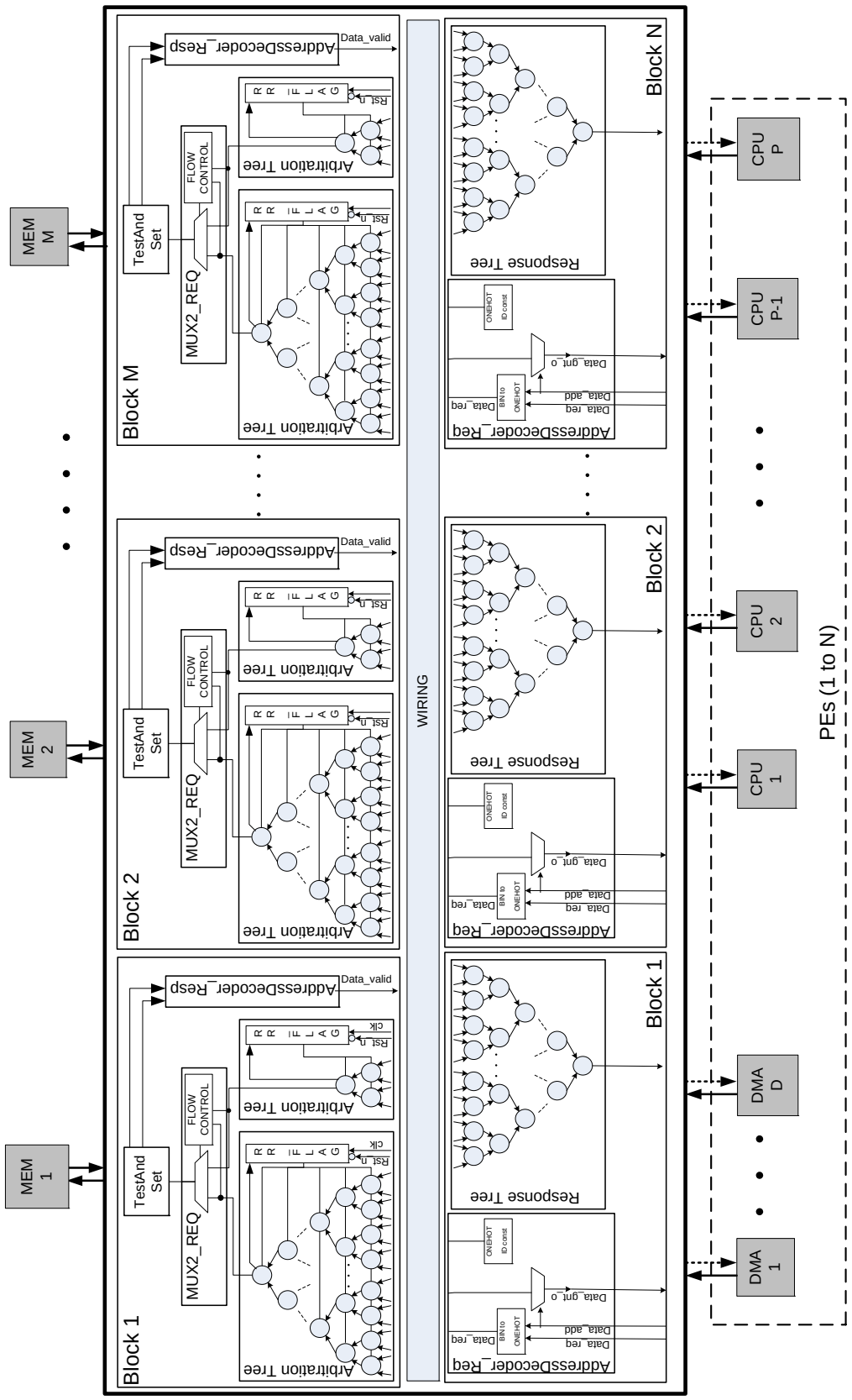


Figure 3.2: Block schematic of the 2D-LIN

3.4.1 Network architecture protocol

The network is created by independent and decoupled Request and Response channel. A memory access starts with a request issued by a PE through a master port, then, the master is kept updated on the status of the request by a simple and lean protocol based on a credit based flow control. Each clock cycle, all the requests from the masters are propagated through the binary trees. Collisions due to multiple requests directed to the same memory bank are avoided by Round Robin arbitration performed at each node. The processors losing the arbitration are stalled. The PE winning the arbitration concludes the transfer in a single clock cycle in case of a store, whereas, in case of a load, the read data is returned the next clock cycle. Even though the read operation takes two cycles to complete, it is possible to achieve an average performance of one clock cycle by pipelining.

3.4.2 Request block

The request block is in charge of collecting all the PE's requests directed to a specific memory module (see Figure 3.2). In the simplest case of two PEs, the block is built out of a single binary tree where the request block is composed by 1 node, being a routing-arbitration primitive. The number of stages of the Arbitration Tree is a function of the number of masters attached to it: NUM_{stage}=$\log_2$(N), N being the number of PEs. Combining several binary trees, the network can support both generic number of ports and different priorities. Consequently, a high priority channel for the processors and a low priority channel for eventual peripherals can be supported. The primitives composing the request block first arbitrate among eventual requests through a Round Robin policy, then the request from the winning master is routed to the MM in a combinational way. At the same time, the flow control signals travelling from MMs to PEs are also managed. Both normal read/write operation and atomic test and set are supported.

3.4.3 Response block

The response block (see Figure 3.2) is in charge of collecting all the responses from memory modules which are directed to a specific PE. Therefore, the response block can be considered as a specular version of the request block. Nevertheless, since the response network is only used for read operations and the read latency is deterministic (1 cycle), no response collisions are possible. Hence, the response path does not need any arbitration, and it can be simplified replacing round robin arbiters with simpler decoders.

3.5 3D interconnect network

Within a standard planar(2D) architecture, when more storage capability or more processing power are needed, the network size increases, and the single-cycle communication becomes

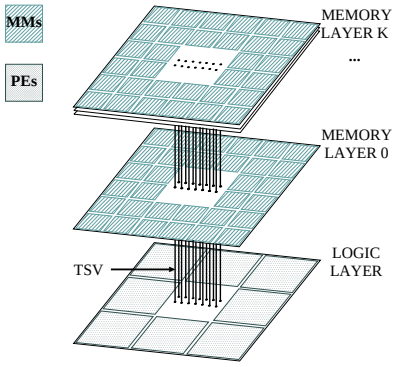


Figure 3.3: 3D chip architecture.

Parameter	Definition
P	Number of processor channels
D	Number of DMA channel
N	Number of masters (N=P+D)
M	Number of memory banks
K	Number of memory layers
Nb_{addr}	Width of the address bus
Nb_{data}	Width of the data bus
$Nb_{layerID}$	Width of the layer identification signal

Table 3.1: Network main parameters.

the limiting factor for the maximum achievable operating frequency.

3D-LIN is the extension of the 2D structure presented in the previous section, to be integrated in a 3D-stacked CMP. This network topology allows designers to overcome the limitation in frequency by automatically splitting the 2D floorplan into one logic layer and several memory layers and stacking them one on top of the other as in Figure 3.3. Moreover, the possibility of stacking multiple memory layers increases the available storage capability of the system.

All the power-hungry processing elements are placed on a logic layer, close to the heat sink, while the memory banks, are divided among the memory layers. The network is partitioned among the layers in an automated way following the assumption that all the memory layers should have the same identical layout. Each layer automatically auto-configures during runtime. This permits to reduce the chip cost and the design effort. The M memory banks are equally divided among K memory layers, where K is a power of 2. Each memory layer contains M_L=M/K memory banks.

Table 3.1 summarizes the main parameters of 3D-LIN.

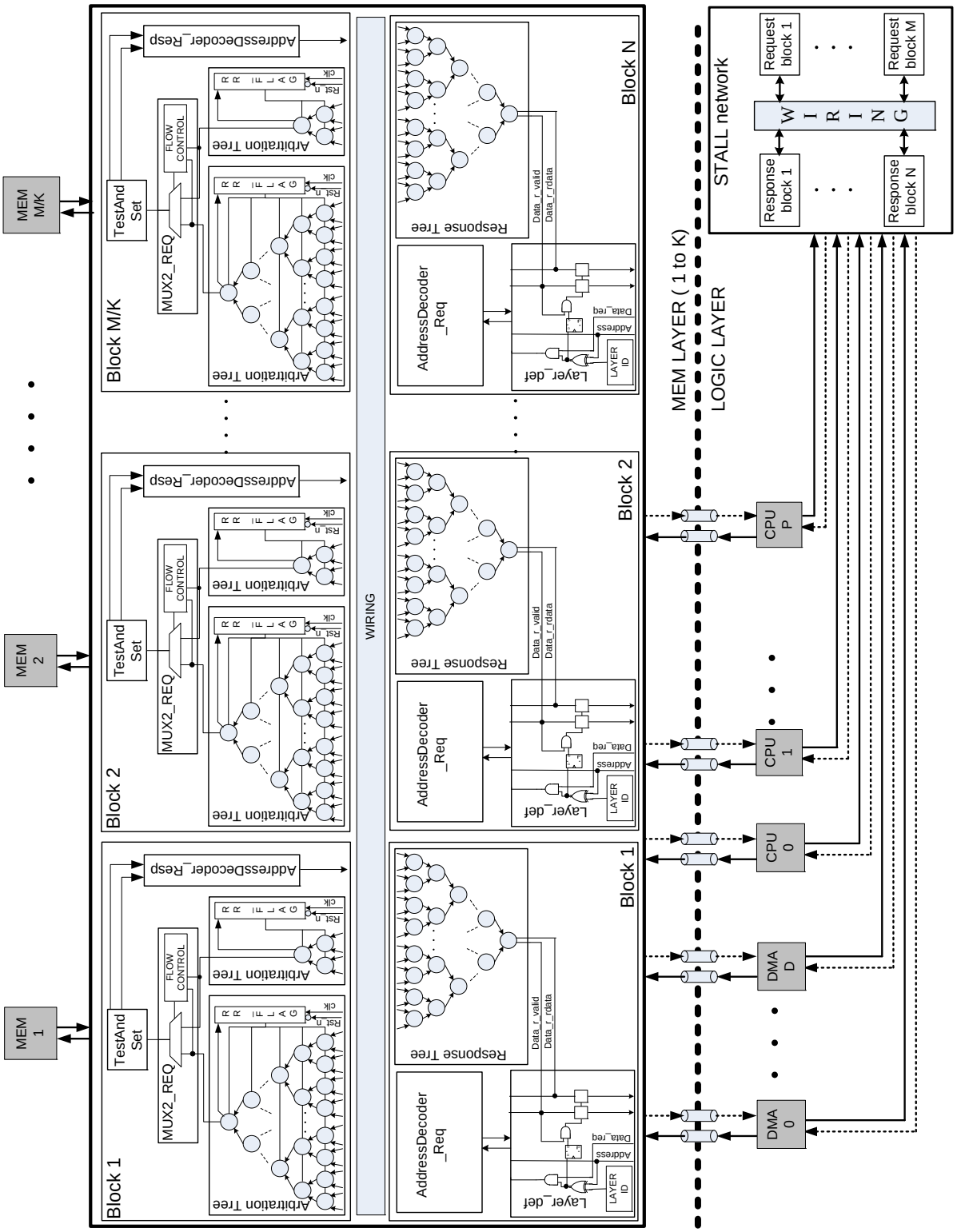


Figure 3.4: Block schematic of the 3D-LIN.

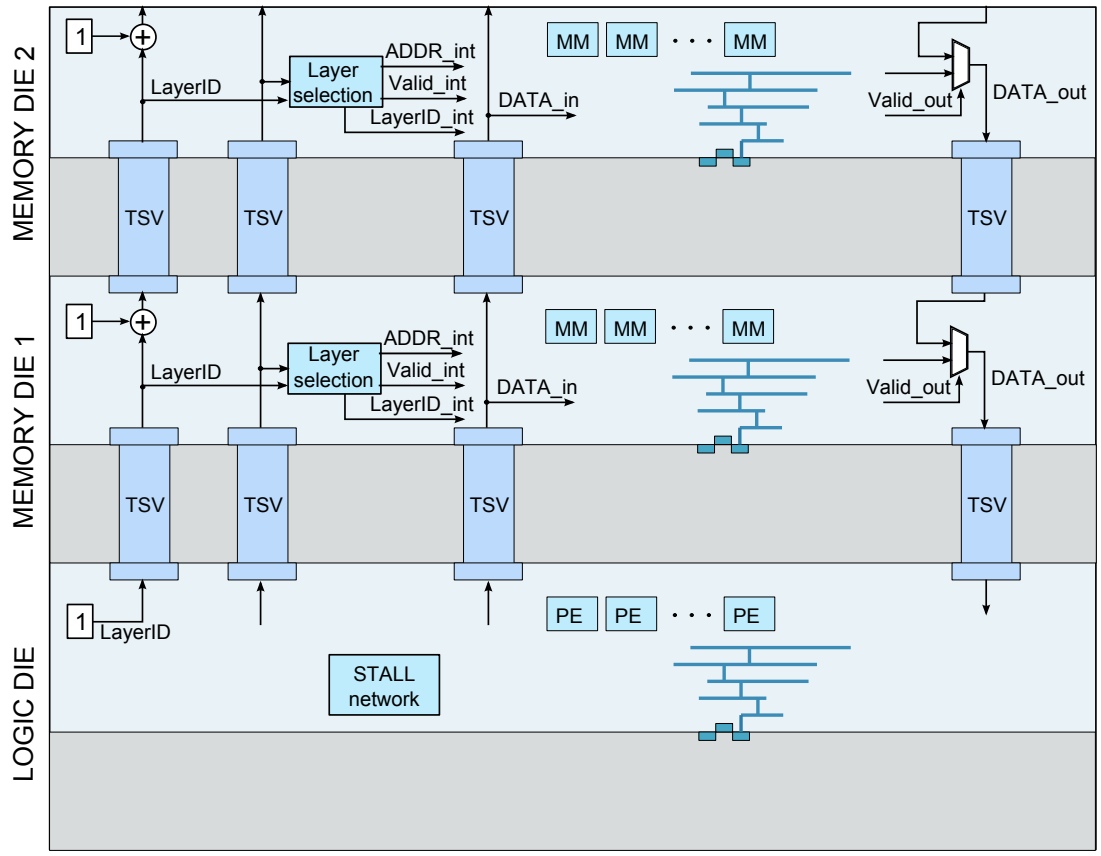


Figure 3.5: Cross section schematic of the 3D stacked system.

3.5.1 Network architecture

The possibility of stacking multiple identical memory layers is beneficial in terms of cost, nevertheless it requires circuit solutions in order configure the stacked memory dies after the fabrication. For this reason, a unique layer identification signal (layerID) is provided, which is used by each layer to automatically auto-configures at the system power-up. The layerID signal is sent from the logic layer, and is composed of $\text{Nb}_{layerID}$=$\log_2$K bits. Each memory layer takes the incoming layerID as its own identifier, and send to the next memory layer the received signal incremented by one.

Due to the modular architecture, the address space is equally divided between all the identical memory layers. As an example, if the processor cluster sees a memory space of 1MB, the address space is addressed by a 10bit signal (Nb_{addr} = 10). Each of the K memory layers, in this example K=2, contains $\frac{1MB}{K} = 500kB$. Therefore, on each memory layer, 9bit are needed to address the stored data. Since $\log_2$K=1, the *Most Significant Bit (MSB)* can be used as a layer selector by comparing it with the layerID, as depicted in Figure 3.6. In the case of a 4 memory layers stack, the two MSB of the request address will be compared with the two bits of the layerID and so on.

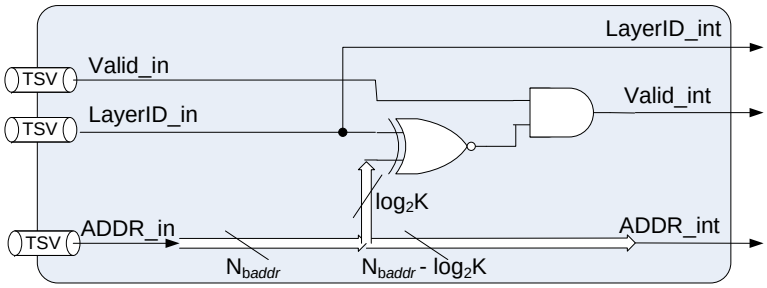


Figure 3.6: Schematic of the layer selection block.

As described in the previous section, since the response network is only used for read operations and the read latency is deterministic, no response collisions are possible. As shown in Figure 3.5, a 2:1 multiplexer controlled by an internally generated selection signal (Valid_out) is used to forward the incoming signal from the upper layer or transmit the read data from the MMs to the TSV channel. Since just one layer requires to drive the response TSV bus at each cycle, no data are lost.

In the 2D network, the Stall signal is critical, because it needs to be asserted much in advance with respect to the next clock rising edge. Hence, in order to optimize it, the logic that computes the Stall signals is detached from the main Network connecting PEs to MMs and placed on the logic layer as a small independent Network, as depicted in Figure 3.4.

During the fabrication process of the 3D-IC, while stacking the dies, the interfaces of the 3D chip may be vulnerable to over voltage stress induced by electrostatic discharge. However, experimental results from IMEC [33] indicate that there is no need of ESD protection. This avoids additional capacitive load on the vertical channel.

TSVs connecting the stacked dies have good electrical characteristics, but their area footprint is bigger compared to the on-chip metal lines. For this reason it is important to place the minimum number of TSVs, while still guaranteeing the maximum possible bandwidth. When the signals traversing the tiers are the direct input and output of the processor, it is possible to place the minimum number of TSVs dedicated to signal propagation:

$$TSV = (Nc + 1 + log_2K) + N(Nb_{addr} + 2Nb_{data} + Nb_{byteEN} + 2) \tag{3.1}$$

where Nc is the number of TSVs for clock propagation, summed to one TSV for the reset signal, $\log_2$K is the number of bits needed for the layer ID. Nb_{addr}, Nb_{data} and Nb_{byteEN} are respectively the number of TSVs for propagating the address, the data and the byte enable signals. The maximum bandwidth of the 2D system is:

$$BW_{max} = f(\frac{Nb_{data}}{8})N \tag{3.2}$$

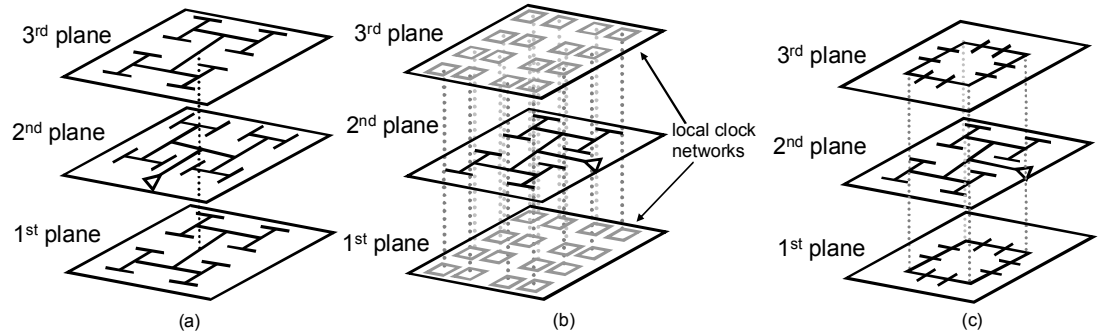


Figure 3.7: Solutions for the 3-D clock distribution network [60], (a) H-trees, (b) H-tree and local rings/meshes, (c) H-tree and global rings.

Hence, the PEs and the small Network for the stall are placed on the logic layer, while each memory layer has the same layout and contains a Network of cardinality $N \times \frac{M}{K}$ and $\frac{M}{K}$ memory banks (see Figure 3.4). This configuration that minimize the number of TSVs needed for the signals, still guarantees BW_{max} also for the 3D implementation.

Stacking multiple dies enhance the issue related to the reliability of the clock distribution, since sequential blocks belonging to the same clock domain are located on different planes. The challenges related to 3D clock networks design have already been extensively studied by Pavlidis et al., [60] [7], which propose several solutions depicted in Figure 3.7. The H-tree solution has been chosen (Figure 3.7a), which is the topology that provides the lowest skew [7]. Nevertheless, in order to guarantee the reliability of the clock distribution and maintain the combinational nature of 3D-LIN, the clock margins have been increased during the clock tree synthesis phase.

Table 3.2 summarize the main parameters of 3D-LIN versus 2D-LIN. We can notice that in terms of number of levels of the trees, the first strongly depends on the number of PEs, while the second is related to the number of MMs. The number of levels directly affects the latencies of the request network path (PE to MM), and the response path (MM to PE). When connecting the memory banks, the access time to read the data from the memory is added to the latency of the response path. 3D-LIN allows us to decrease the number of arbitration levels of the response tree when implemented on 2 or more memory layers, thus allowing the system to run at higher frequencies. The number of primitives per layer and in the system give an estimation on how the area of the network can be reduced by moving to 3D. The main reduction is encountered for the primitives of the Response Tree, but also the Arbitration Tree diminish.

3.5.2 Network operation

During a read/write operation, the master asserts data and control signals that are sent as a packet. Some control signals go to the Stall Network that arbitrates possible collision and eventually sends the Stall signal to the PE within the same clock cycle. The full packet, data

Table 3.2: 3D-LIN vs. 2D-LIN

	2D-LIN	**3D-LIN**
Number of levels Response Tree	$log_2 M$	$log_2 \frac{M}{K}$
Number of levels Arbitration Tree	$log_2 N$	$log_2 N$
Number of primitives on each memory layer - Response Tree	$\sum_{i=1}^{log_2 M} \frac{M}{2^i} \times N$	$\sum_{i=1}^{log_2 \frac{M}{K}} \frac{\frac{M}{K}}{2^i} \times N$
Number of primitives on each memory layer - Arbitration Tree	$\sum_{i=1}^{log_2 N} M \times \frac{N}{2^i}$	$\sum_{i=1}^{log_2 N} \frac{M}{K} \times \frac{N}{2^i}$
Number of primitives in the system - Response Tree	$\sum_{i=1}^{log_2 M} \frac{M}{2^i} \times N$	$\sum_{j=1}^{K} \sum_{i=1}^{log_2 \frac{M}{K}} \frac{\frac{M}{K}}{2^i} \times N$
Number of primitives in the system - Arbitration Tree	$\sum_{i=1}^{log_2 N} M \times \frac{N}{2^i}$	$\sum_{j=1}^{K} \sum_{i=1}^{log_2 N} \frac{M}{K} \times \frac{N}{2^i}$

and control signals, are also sent through the TSVs to the memory layers. Each memory layer receives the packet and checks if the request is for a position in its address range. The layer containing the address lets the packet enter, while the other layers invalidate the request. When a packet accesses the memory layer containing the requested address, the network routes and arbitrates the packet among the other simultaneous requests, allowing the higher priority request to access the memory bank. Write operations are performed in the same clock cycle, while for Read operations and Test and Set operations, the read data is propagated back to the related PE in the next clock cycle.

3.6 Simulations and results

This section provides the evaluation of 3D-LIN in terms of area, power and delay. The Network is implemented in System-Verilog and synthesized with Synopsys Design Compiler in topographical mode using 65nm CMOS technology library from ST-Microelectronics. The physical synthesis has been chosen to extract the results because it allows the user to floorplan the design and accurately predict post-layout timing using real net capacitances during RTL synthesis [61]. The functionality has been verified using Mentor Graphics' Modelsim.

In this experiment we considered 5μm wide TSV with 10μm minimum pitch and a length of 25μm, which represents the state-of-the-art for high density through silicon vias [33]. The

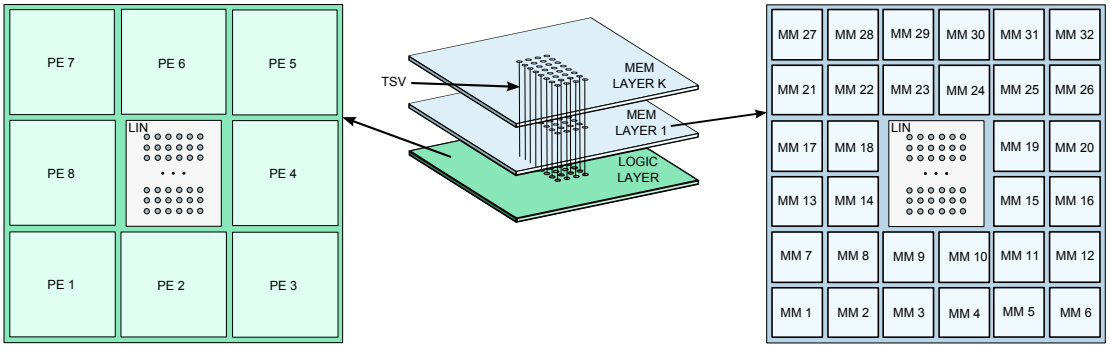


Figure 3.8: Floorplan of a 3D system hosting 8 processing elements and 32 memory banks.

TSVs are fabricated after the FEOL processing and prior to BEOL processing. The chosen technology allows TSVs from the bottom layer to be connected to the lowest metal layer, while the TSVs to the upper layer are connected to the top metal layer. According to the chosen dimensions, the TSV's parasitics have been obtained through the analytical model proposed in Chapter 2. For the experiments, the parasitics values have been rounded to 20mΩ for the resistance and 30fF for the capacitance.

The memory banks size should be chosen depending on the multi-core application requirements. For the experiments, we chose a case study with memory modules chosen to be SRAM banks of 8kB, which timing and physical information are provided by the lib file and the Milkyway database. Each MM occupy 0.06mm^2. Regarding the processing elements, dummy hard macros are used in order to emulate their area occupation. Each PE is considered to be an ARM CortexM3, which estimated area is around 0.07mm^2 for 65nm technology.

All the TSVs for signal propagation are placed in a TSV array located in the center of the chip, as depicted in Figure 3.8 In view of the fact that any of the PEs can access randomly any of the memory banks, this solution minimize the standard deviation on the average distance between PEs and MMs.

Since the memory layers are stacked on top of the logic, the only die accessible from the outside is the topmost memory layer, therefore the proposed 3D system uses TSVs for I/O as depicted in Figure 3.9. The I/O signals should be delivered to the processor cluster. This approach also allows the power to be delivered to all the layers. Since the pad ring is present on both 2D and 3D design, its area overhead has been omitted for the area analysis.

Unfortunately, the current version of Synopsys DC does not support TSV and 3D stacking, as a consequence, in the absence of established design kits, the synthesis flow is performed in several main steps. Starting from the synthesizable RTL description of the network, already configured with the user constraints, the floorplanning of memory layer is performed. In order to emulate the TSVs, the time and physical constraints are added. After the physical synthesis of the memory layer, the back-annotated delays are used to perform the physical synthesis of the logic layer. Once the floorplan is defined, the logic layer is synthesized considering

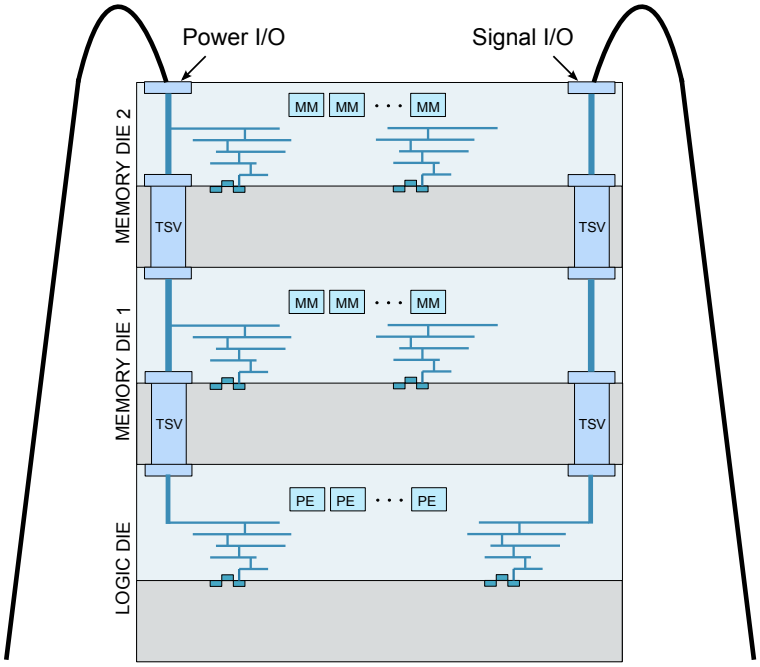


Figure 3.9: Input/Output connections: I/O signals are connected to the logic layer, I/Os for power delivery are connected to all layers.

the latencies of the stacked dies. These steps are then iterated to meet the desired timing constraints for the complete 3D-stacked system.

3.6.1 Physical analysis

When moving to a 3D configuration, the original NxM network is divided among the layers: a small NxM network for the Stall signal is placed on the logic layer, while the rest of the network that communicates with the memory banks is divided in Nx$\frac{M}{K}$ smaller networks distributed on each memory layer. We first explore the impact of the 3D partitioning on the network area, measured as equivalent kgates (nand2), for the following systems:

- 16 PEs and 64MMs.
- 16 PEs and 128MMs.
- 8 PEs and 64MMs.
- 8 PEs and 128MMs.

Figure 3.10 depicts the trend of the total area, that is the sum of the area occupied by the partitioned network on each layer, for different network cardinalities. We can notice that for 3D-systems composed of 1 memory layer, the total area has a slight increase. This is due to

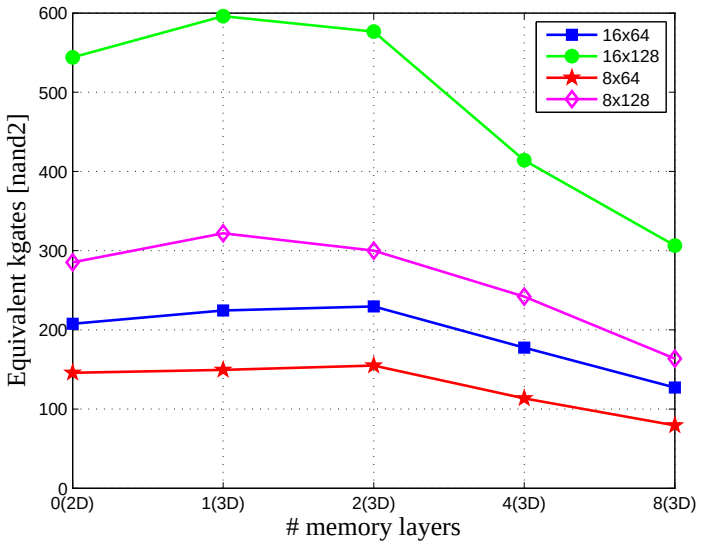


Figure 3.10: Area occupied by the network in the 3D system.

the fact that moving from a 2D-system to a 3D-system, the small stall network is added on the logic layer. Once we reach 3 or more layers, even if the network is replicated on each memory layer, the area reduction per layer dominates. Since the total number of primitives constituting 3D-LIN is equal to

$$N_{primitives} = \sum_{j=1}^{K} \sum_{i=1}^{log_2 \frac{M}{K}} \frac{\frac{M}{K}}{2^i} \times N + \sum_{j=1}^{K} \sum_{i=1}^{log_2 N} \frac{M}{K} \times \frac{N}{2^i} \tag{3.3}$$

the area reduction is expected to be more accentuated for networks connecting a higher number of MMs.

In a 3D system, however, is important to consider the reduction for each layer, since the form factor is influenced by the single layers dimension. The area occupied by the network on the logic layer and the ones on each memory layer is shown in Figure 3.11. Once adding more memory layers, there is a strong decrease in the per-layer network area. Figure 3.12 shows the trend of the ratio between the network area and the memory area both per layer and in the full 3D system composed of 16 PEs interfaced to 64 MMs. When moving from a planar design to a stacked system, the sum of the network areas on each layer is higher than the 2D counterpart, nevertheless the area per layer decreases.

The configurability of the network gives the possibility to explore the form-factor trend for the 3D multi-core systems with shared L1 memory on top of logic. Given the specification of the system, the best trade-off can be found in terms of number of layers. In particular, we chose to analyse the area of the chip(A_{3D}) normalized to the area of the same chip implemented

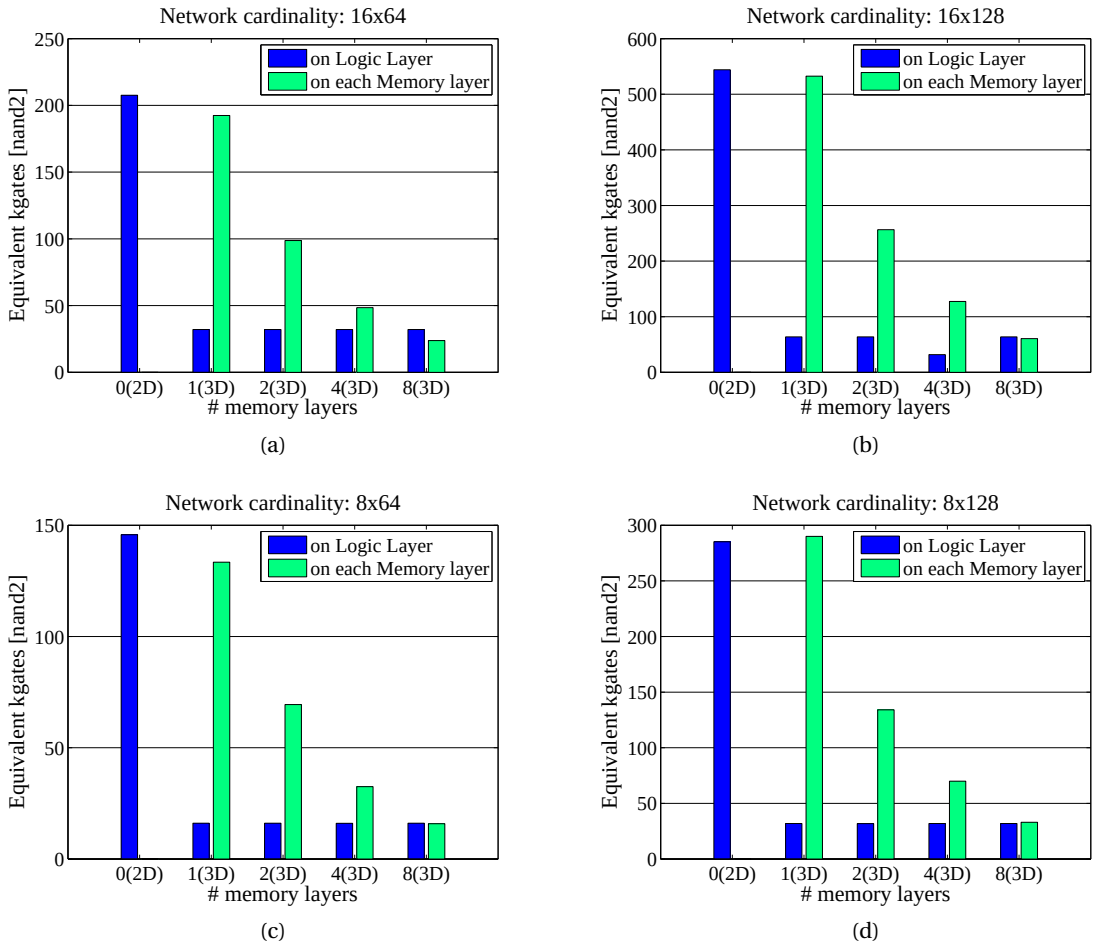


Figure 3.11: Area of the Stall/Valid Network on the logic layer (blue) and area of the data Network on each memory layer (green) for different number memory layers stacked on top of the logic layer.

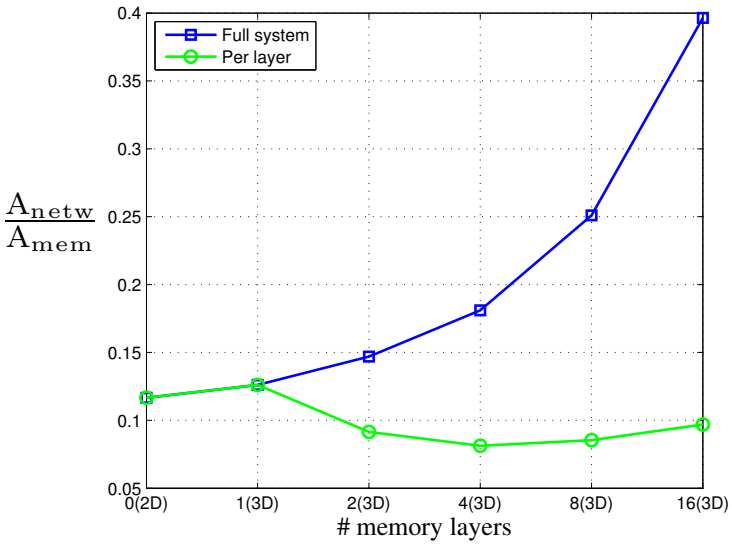


Figure 3.12: Area of the network over the area of the memory for each memory layer(green), and for the whole system(blue).

on a single silicon layer(A_{2D}) for the following configurations and area occupation of the memory(A_{mem}) over the area of the planar chip(A_{2Dchip}):

- 16 PEs and 16 MMs : $\frac{A_{mem}}{A_{2Dchip}}$=43% ;
- 16 PEs and 32 MMs : $\frac{A_{mem}}{A_{2Dchip}}$=58%;
- 16 PEs and 64 MMs : $\frac{A_{mem}}{A_{2Dchip}}$=70% ;
- 16 PEs and 128 MMs : $\frac{A_{mem}}{A_{2Dchip}}$=79% .

Figure 3.13 depicts the reduction of the area when the chip is designed to stack different numbers of memory layers on top of the logic layer. When moving from the planar structure, to a 2-layer structure, the memories and the network are moved to the upper layer, and we can notice a decrease in the form factor. However, this reduction is still limited due to the size of the network that, as explained before, does not shrink effectively. In additions, the TSV area occupation increases the size of both layers. Considering the stacking of two or more layers on top of the logic, the network cardinality start changing depending on the number of memory layers, leading to a decrease in its area occupation, while the TSV occupation remains the same as for the 3D, single memory layer, case. The best trade-off point can be found when the area of the memory layer is almost equal to the area of the logic layer. When reaching the best trade-off, the stacking of any more memory layers does not affect the form factor that is now defined from the area of the logic layer.

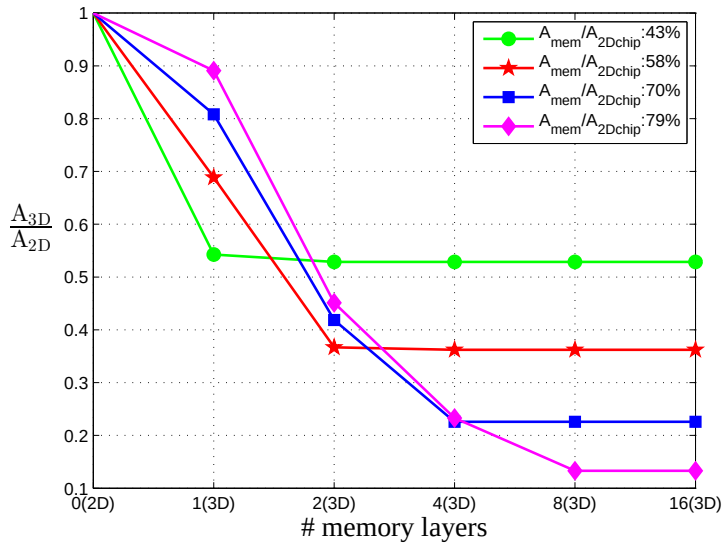


Figure 3.13: Area of the 3D chip normalized to the area of the 2D implementation.

3.6.2 Power analysis

One of the main challenges faced by 3D stacking technology is the power management. Stacking more layers arise new challenges due to an increased power density per footprint, which may cause temperature to increase beyond the limits, thereby affecting the system reliability. At the design level, a careful floorplan definition and thermal management techniques such as *dynamic voltage and frequency scaling (DVFS)* can help, but are not sufficient. There is a significant research effort to tackle the power issue at different levels. At the software level thermal-aware task scheduling policies [16] can be implemented. At the fabrication level, cooling techniques such as inter-layer micro-channel liquid cooling [11] and *Thermal-TSVs(TTSV)* [62], [63] can be exploited to remove the excessive heat. Nonetheless, while designing 3D interconnects, the power consumption should be considered in order to avoid power hungry solutions.

In this chapter, we do not propose any cooling or thermal management techniques, but we focus on exploring the power dissipation of 3D-LIN to ensure reliability. The total dynamic power consumed by the network is depicted in Figure 3.15. We can observe how the trend for power is correlated to the network area. As the number of blocks to be interconnected increases, the size of the die affect the wirelength and the power related to wiring start dominating the cell internal power. Hence, the gain in power consumption is more pronounced for systems with higher cardinality and appears once stacking more memory layers which reduces both the per-layer network cardinality, and the single layer size.

The power contribution of the different single layers is shown in Figure 3.15. The power consumed by the stall network on the logic layer is small compared to the consumption of the

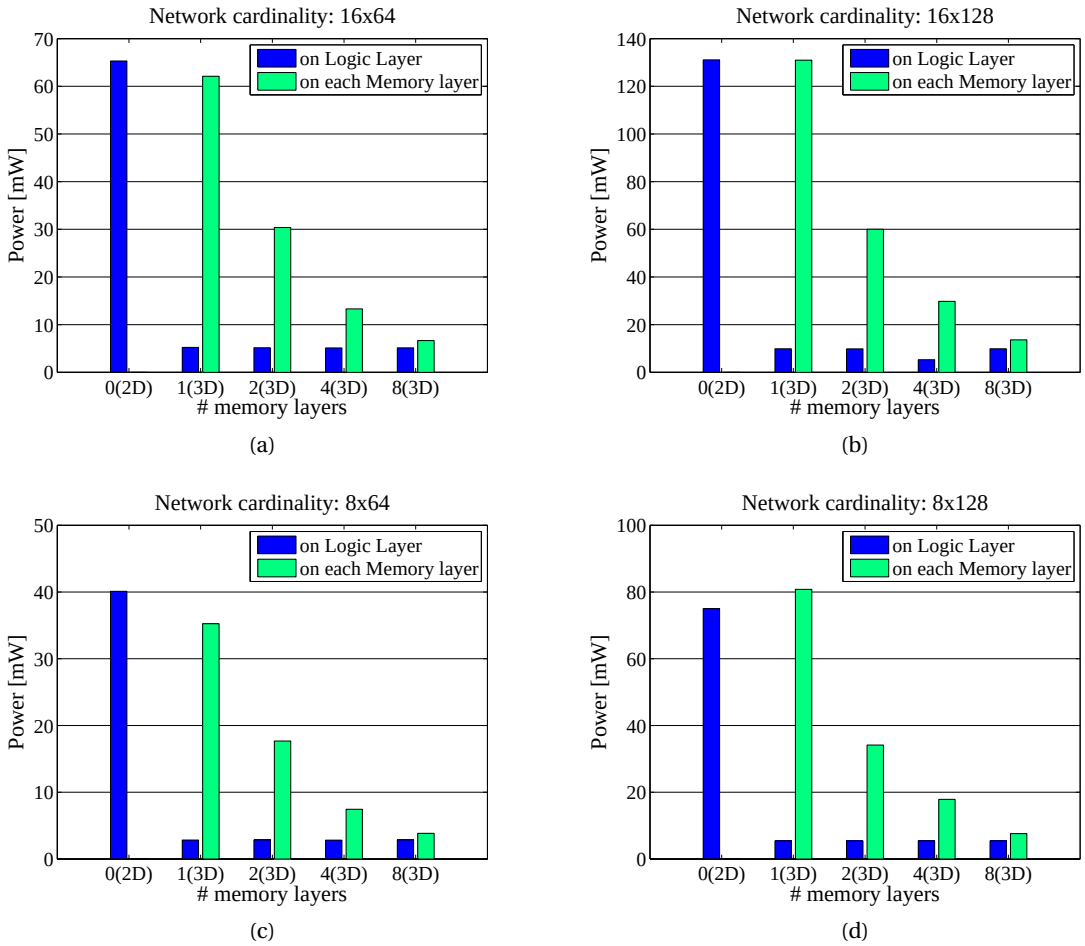


Figure 3.14: Dynamic power consumed by the Stall/Valid Network on the logic layer (blue) and dynamic power consumed by the data Network on each memory layer (green) for different number memory layers stacked on top of the logic layer.

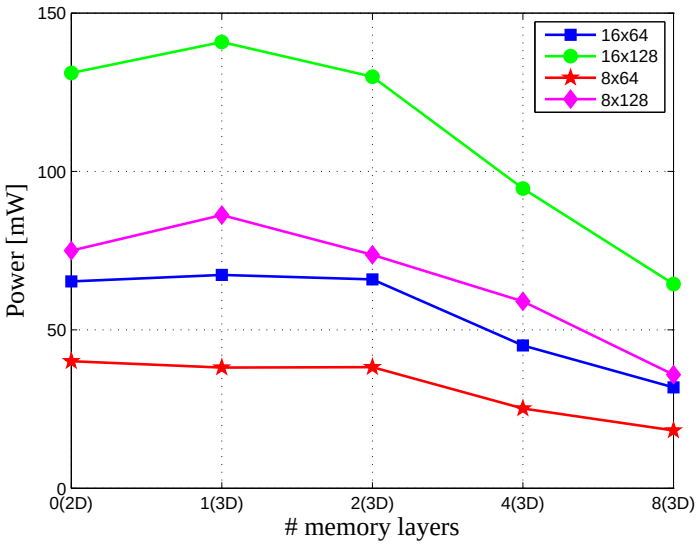


Figure 3.15: Total dynamic power consumption of the network in the 3D system.

network on each memory layer, which is the dominant contribution. As the number of stacked memory layers increases, the cardinality of the network on each layer is reduced, leading to a significant gain in power.

3.6.3 Timing analysis

Exploring 3D-LIN in term of latency the following configurations are considered:

- 16 PEs and 32 MMs;
- 16 PEs and 64 MMs;
- 16 PEs and 128 MMs.

As previously discussed, the frequency of the network is limited by the response path that includes the access time to read a data from the memory bank. However, depending on the size of the memory module, this access time changes. In our experiments, we explored the latency of the network when connecting memory banks of 8kB.

In Figure 3.16 and 3.17, both the system latency and the network latency are shown. We can notice that moving from the planar system to one stacked memory layer, the latency slightly decreases due to the shorter interconnect. The reduction in delay is more evident for systems with two or more memory layers, due to the changes in the network topology. The reduction in delay is more evident in Figure 3.17 considering the network itself, independently from the

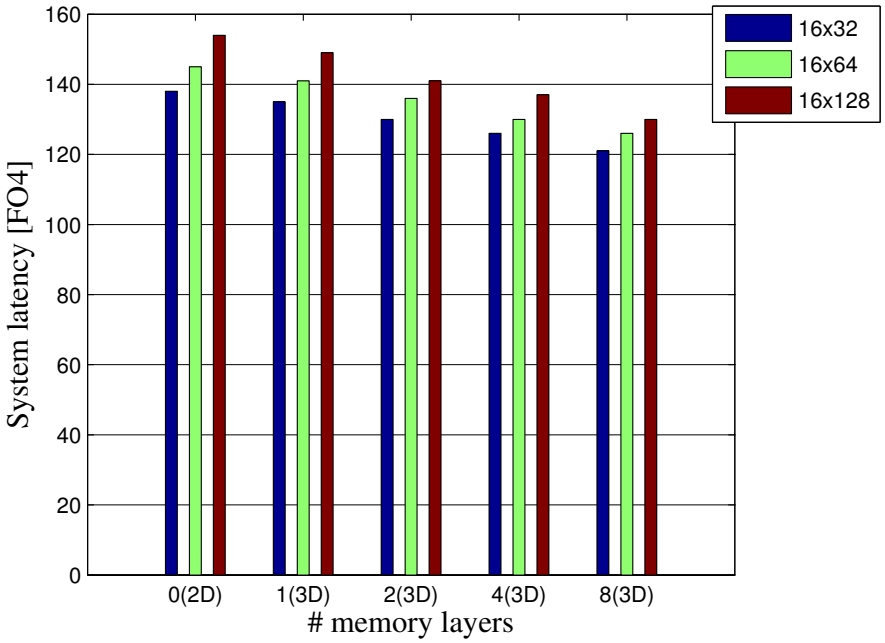


Figure 3.16: System latency: Network delay plus memory access time.

Table 3.3: Latency improvement

	16x32		**16x64**		**16x128**	
	system	network	system	network	system	network
1 memory layer	2%	9%	2%	7%	3%	10%
2 memory layers	6%	22%	6%	20%	8%	24%
4 memory layers	8%	32%	10%	35%	11%	31%
8 memory layers	12%	46%	13%	44%	16%	46%

attached memory banks. The latency of the network shows significant improvement, in the case of 16PEs connected to 64MMs, the 2D latency of ~42FO4 is reduce down to ~23FO4.

Table 3.3 shows the latency improvements in percentage. The results show that stacking a single memory layer, the memory access time dominates the decreased latency of the interconnect and the improvement is only a few percents. However, when moving to two memory layers, we can obtain already around 8% improvement, reaching 11% improvement with four memory layers for a network cardinality of 16x128. The benefits are more evident considering the network alone, independently from the attached memory, achieving 35% improvement for four memory layers stacked on top of the logic layer.

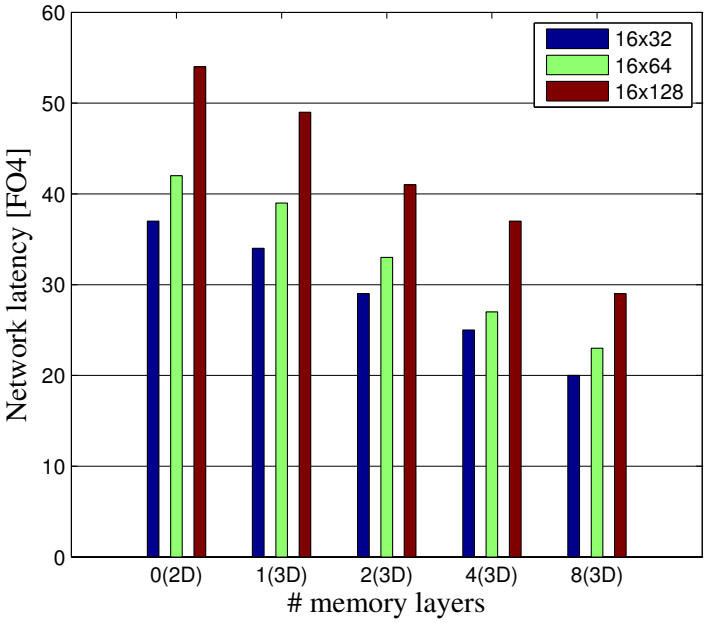


Figure 3.17: Network latency.

3.7 Summary

In this chapter, we present a configurable network architecture that can be integrated in a 3D stacked CMP featuring a shared L1 SRAM memory. A TCDM composed of multiple SRAM memory modules provides a fast and convenient method for inter-processor communication, avoiding cache coherence overheads. The network guarantees, single cycle, low-latency communication, therefore, it is well suited for tightly coupled processor cluster which performances critically depends on the architecture of the interconnect between the processors and the memory banks.

The network and the multi processor system has been explored in terms of area, form factor, power and latency. The benefits obtained by exploiting 3D integration are evaluated. Moreover, the study also focuses on exploring the performances for different 3D structures, studying the effects of stacking different number of memory layers on top of the multi-processor logic die.

The physical synthesis results show the best trade off point between the amount of memory needed in the system and the number of stacked layers. In case of a memory occupation of 60% of the planar chip, by moving to a system that integrates two memory layers on top of a logic layer, the form factor is improved more than 60%. In terms of latency, the 16x128 configuration of the network can be improved up to around 24% in case of 2 memory layers, and 31% in case of four memory layers, leading to a latency reduction for accessing 8kB memory banks of 8% and 11% respectively. Latency and area improvements come without a worsening in terms of power. Stacking 2 or 3 layers, the power consumption is kept almost the same as for the 2D implementation, while starts improving as the number of layer increases.

3.8 Acknowledgements

The author would like to thank Prof Luca Benini for his guidance during this project and Igor Loi for his work on the original 2D logarithmic interconnect network.

4 Design and Analysis of High Speed Serial Vertical Links

Chapter 3 presents a 3D network interconnecting a system composed of multiple memory layers stacked on top of a logic layer hosting a cluster of processing elements. The proposed solution achieves low-latency, single cycle processor to memory communication. Nevertheless, the use of a parallel vertical bus requires the placement of one TSV for each inter-layer signal. Adopting this approach for designs that require a high vertical bandwidth may results extremely expensive in terms of silicon area.

In this chapter, we present circuit-level design and analysis of low power high-data-rate 3D serial TSV links. A design space exploration is performed and trade-offs in terms of area, power and performance are presented. Circuit simulations of RC-extracted layouts in 40nm CMOS-technology revealed that 8:1 serialization efficiently balances area consumption and energy efficiency.

4.1 Problem formulation

3D stacking technology is offering many product benefits to SoC and memory: performance enhancement, product miniaturization and lower power consumption [64]. Dense vertical 3D interconnects, with potentially thousands of multi Gb/s 3D I/Os, can support very high data bandwidth. Hence, 3D technology has the potential to replace traditional off-chip signalling technology to satisfy the bandwidth demand of next generation computing devices.

3D integration is also a promising solution to solve the fundamental challenge of planar *chip multiprocessors (CMP)* due to on-chip interconnects not scaling well with technology. In conventional process scaling, the signal delay time (RC) is expected to increase with technology node mostly from the increasing resistance of the wires. In addition, the aggregated interconnect length increases since more wires are being used in each layer and more metalization layers are being added. Hence, exploiting 3D stacking to reduce wiring [6] [65] [66] in microprocessor systems can lead to a considerable enhancement in performance. Moreover, wire length reduction in 3D ICs also translates into lower power dissipation in interconnects

and repeaters.

Nevertheless, 3D integration also comes with many challenges. Since TSV fabrication technologies are not yet mature, reliability is expected to be a limiting factor for 3D IC performance and yield [7]. Hardware redundancy [18], [19] is a potential solution to improve yield, yet, the drawback of connecting multiple TSVs to the same signal is that it increases the silicon area and RC delay. Furthermore, the need of thermal TSVs to dissipate the internal heat of 3D ICs will lead to an even greater TSV footprint on each layer [15][67].

Although advanced wafer level 3D stacking manufacturing technology has been demonstrated featuring 7 μm TSV diameter with an estimated *keep out zone (KOZ)* of 2um [68] , such TSV footprint still occupies large silicon area compared to nanometer *Back end of the line (BEOL)* structures. Indeed, a 25 μm^2 TSV with 5 μm^2 diameter it is equivalent to ~45K transistors in a 45nm CMOS process [24].

In this chapter, high speed serialization over TSV interconnects is proposed. Since 3D interconnects offer a reduced load compared to off-chip channels, high speed serial transmission through TSVs do not require complex and power hungry equalization techniques, achieving high bandwidth with low silicon area and power. With serialization, TSV footprint can be kept relatively large, allowing for a thicker substrate and relaxed alignment accuracy making the fabrication yield viable for industrial exploitation. A low power *Serializer-Deserializer (SERDES)* circuits for inter chip 3D links has been investigated running spice level simulations on RC-extracted designs and silicon-proven TSV electrical models, thus providing good correlation with chip measurement.

4.2 State of the art

Test chips with 3D IC technology have already been discussed in literature. Liu et al. [69] proposed a compact and low power 3D-IO fabricated in 45nm SOI CMOS which dissipates only 0.11 mW/Gb/s. Another example is the 3D-DRAM developed by Kim et al. [30] which adopts a large number of TSVs operating at low frequency to achieve high vertical bandwidth.

The drawback is that parallel vertical inteconnects do not scale well with future CMOS process generations, due to the poor scalability of TSVs. To address this issue, a TSV serialization scheme has been proposed by Pasricha [70]. In [70] 3D multicore benchmarks showed that 4:1 serialization on TSV interconnects can save 70% area footprint with negligible performance and power overhead. More recently, Lasio et al. [71] outlined the benefits of TSV serialization for 3D *Network on chip (NOC)* using a cycle accurate NoC simulator. Moreover, Lasio et al. [71] demonstrated that TSV serialization results in low performance overhead in 3D NoCs based *Multiprocessor System-on-Chip (MPSoC)*

Nevertheless, in [71] and [70] the parallel to serial conversion scheme utilizes the same clock as the parallel data stream, obtaining a reduction of the vertical data throughput. Moreover,

such reduction increases with the serialization level. The contribution of this chapter lies on the proposition of a serialization scheme aimed to maintain the same aggregate bandwidth as in the case of low frequency fully parallel bus. While [71] and [70] are mostly focused on architectural studies of TSV serialization targeting multiprocessors and 3D-NoCs while this work presents a circuit level metric to characterize 3D serial links.

4.3 SERDES circuit design

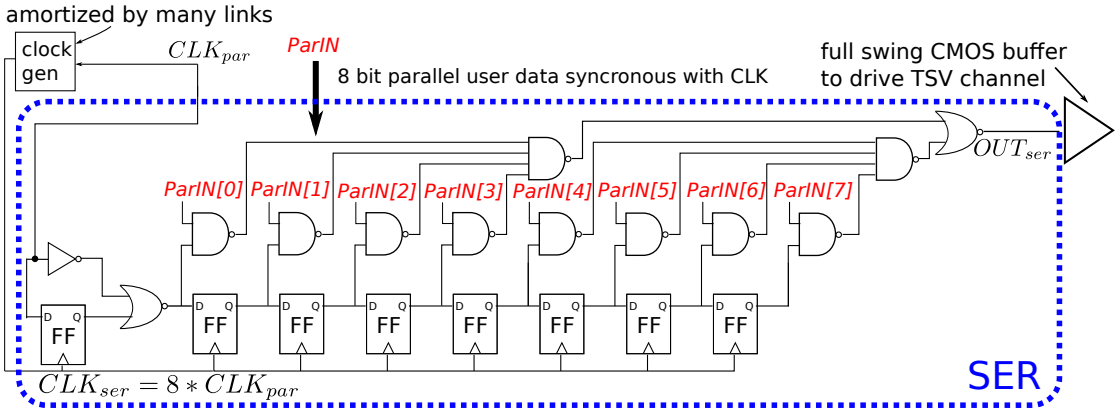
TSVs have excellent high frequency properties, but their area footprint is bigger compared to the BEOL vias, therefore it is important to place the minimum number of TSVs without sacrificing vertical data-bandwidth. For this reason, a serializer-deserializer module can be introduced to optimize the trade-off between bandwidth and number of TSVs in the array.

Serial links have gained attention as promising alternative to standard parallel links for both on-chip and off-chip communication. Different SERDES connections topologies have been developed to fit the requirements of different circuits. Table 4.1 summarizes the SERDES designs available in the literature. Asynchronous data link [72] [73] are used to transfer data across different clock domains and employ handshake instead of clock signal for operation control. An acknowledgement signal should be added for each serial channel. System synchronous clocks or source-synchronous clock topologies requires the clock to be transmitted, while another common timing mechanism for serial interconnects injects a clock into the data stream at the transmitting side and recovers the clock at the receiver.

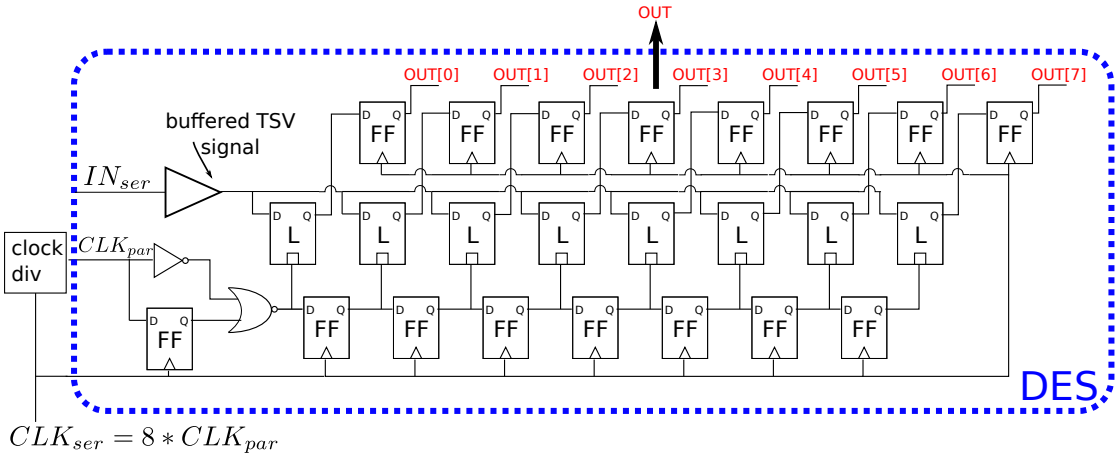
The goal of this work is to reduce the number of TSVs in the system. In the proposed high speed 3D serial vertical link, each serial vertical data connection transmit a single-ended signal using a single TSV channel. N slow speed channels at a frequency f_{par} are serialized into a high speed data stream at $f_{ser} = N \times f_{par}$, where N is the serialization level. The high speed stream is sent through a TSV and deserialized on the receiving layer. Asynchronous connections would require the addition of a second acknowledgement TSV for each connection, therefore a synchronous transmission has been chosen. Since the delay due to the TSV is negligible the clock phases are effectively matched [69], hence the clock can be distributed on each stacked die with identical distribution networks.

	Throughput (Gb/sec)	Technology	Protocol	Transmission
[72]	3.9	180nm	asynchronous	single ended/differential
[73]	67	65nm	asynchronous	differential
[74]	8	180nm	synchronous	differential
[75]	3.2	180nm	synchronous	differential
[76]	10	45nm	synchronous	differential
[77]	12	90nm	synchronous	differential

Table 4.1: SERDES state of the art.



(a)



(b)

Figure 4.1: (a) 1:8 SER circuit and (b) 1:8 DES circuit.

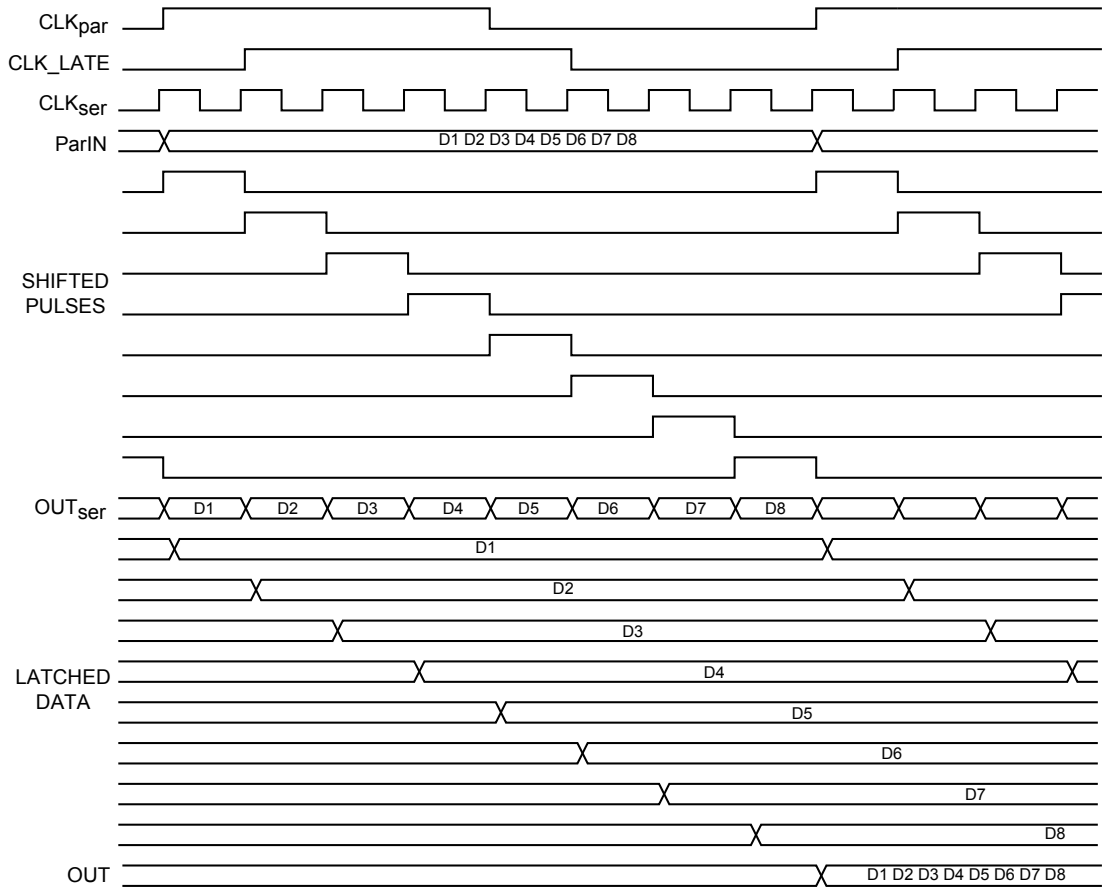


Figure 4.2: Waveforms describing the SERDES functionality.

The transmitter circuit consists of a *serializer (SER)* followed by a buffering stage to drive the TSV. The architecture chosen for the serializer is based on the design proposed by Kurisu et al. [78]. Figure 4.1a depicts the circuit diagram of the 8:1 SER: the circuit creates a pulse of width $1/f_{ser}$ and period $1/f_{par}$, which is then passed through a shift register synchronously to the fast clock CLK_{ser} in order to produce N shifted pulses. These pulses are then used as enable signals for a combinational circuit that converts the N parallel signals, ParIN[7:0], into a serial stream, OUT_{ser}, that goes to the buffering stage driving the TSV.

The receiver architecture is depicted in Figure 4.1b for a 1:8 deserialization scheme. The *deserializer (DES)* also exploits a shift register in order to produce N shifted pulses of width $1/f_{ser}$ and period $1/f_{par}$ that act as trigger for N latches. Each latch stores one bit of the incoming serial stream, IN_{ser}, until they are resynchronized with the system slow clock, CLK_{par}, through the output registers.

The maximum data-rate of the SERDES topology is limited by *D-Flip-Flop (D-FF)* performance, hence the D-FF structure should be carefully chosen in order to achieve the required speed with a reasonable energy consumption. As demonstrated by Alioto et al. [78], *Master-Slave(MS)*

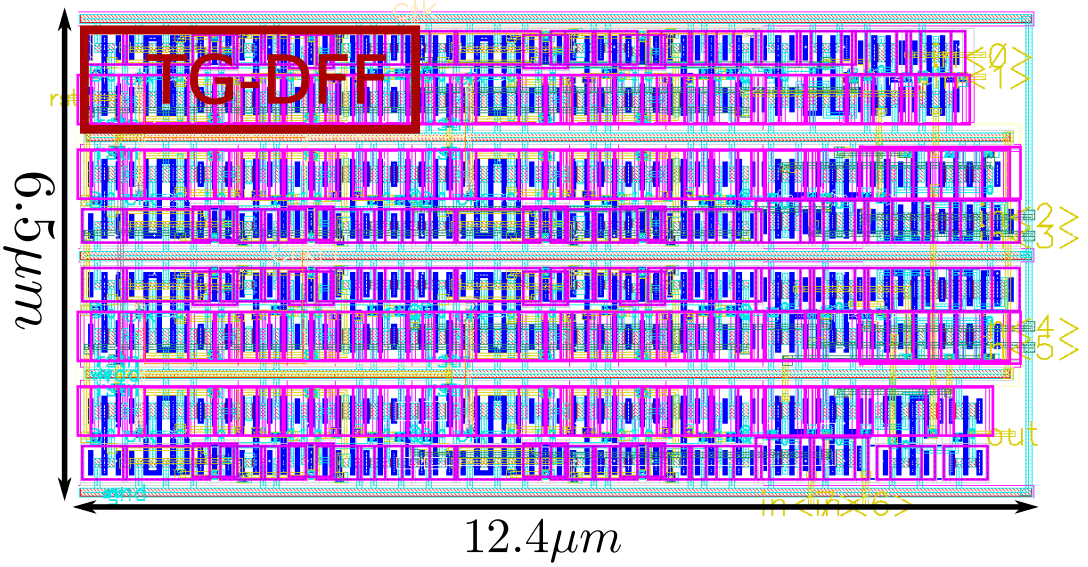


Figure 4.3: Full custom layout view of the 8-bit SER in 40nm TSMC technology.

Serialization Level	4	8	16	32
Area Serializer [μm^2]	40.57	80.62	160.53	301.69
Area Deserializer [μm^2]	86.82	165.63	260.53	637.51

Table 4.2: SERDES area.

FFs are the most energy efficient FFs in the low energy region. Among the different topologies, the *Transmission Gate Flip-Flop (TGFF)*, offers one of the best energy-delay product. This static CMOS design provides robust operation in nanometer CMOS technologies. Therefore, the TGFF has been used to implement the SERDES.

SERDES circuits have been custom designed in 40nm TSMC CMOS technology. Both SER and DES have been implemented for different serialization levels: 4, 8, 16 and 32. Table 4.2 reports the area of the implemented circuits. In particular, Figure 4.3 shows the layout view of the 8:1 serializer circuit.

4.4 Design exploration

In this section, we first explore the trade-off of the proposed technology in order to find the most convenient serialization level for different TSV channels. The simulation environment for the evaluation of the 3D serial link includes RC-extracted layouts of the SERDES circuits implemented in 40 nm TSMC CMOS technology, and the TSV electrical circuit model presented in Chapter 2. The 3D serial communication circuit was simulated in Cadence Virtuoso at

Diam [μm]	Height [μm]	Res [$m\Omega$]	Cap [fF]	Ind [pH]
5	50	147	18	46
10	50	36	41	36
40	50	7	174	20

Table 4.3: TSV parasitics for different geometries

a supply voltage of 0.9V. Transient simulations on the RC-extracted layouts of the different SERDES circuits prove functionality of the 3D link at a serial data-rate of 8 Gb/s for all the circuit topologies. The area and power consumption of the proposed 3D serial link, Figure 4.4(a), are analysed for different serialization levels and compared to a parallel 3D link, Figure 4.4(b), with the same aggregate data-rate.

4.4.1 TSV channel

Since the delay due to the TSV is negligible, the clock phases are effectively matched [69]. Nevertheless, the serialized data stream coming from the TSV channel is re-synchronized to the layer clock domain by the deserializer. This avoids problems due to an eventual inter-layer skew between the clocks. As described in Chapter 2, the wide range of available TSVs can offer trade-off options that may be matched for various product requirements. Unfortunately, TSV fabrication technology is still in its early stage: with the decreasing of the TSV dimensions, reliability issues limiting the yield are becoming more and more critical. Hence, the impact of the proposed serialization scheme has been explored for 3D ICs based on the TSV technologies presented in Table 4.3. 5μm TSVs represent state-of-the-art for high density through silicon vias [33], 40μm TSVs are a more established technology which guarantees better reliability [2] while 10μm TSVs provide a fair compromise between the two extreme.

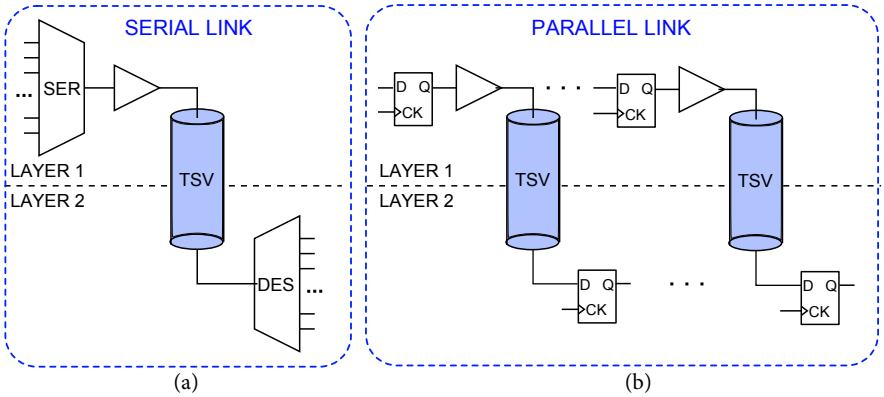


Figure 4.4: (a) 3D serial link and (b) 3D parallel link scheme for a 2-layers stack.

4.4.2 Area analysis

In a standard parallel 3D link, N-TSVs are used to transmit a N-bits data stream. The proposed serial link allows the data to be sent through a single TSV channel, however, the area overhead of the N-bit serializer and deserializer circuits should be explored. Moreover, since the serial communication works at a frequency N-times faster than f_{par}, the buffer used to drive the TSV should be considerably larger than the one used to drive a TSV in a parallel configuration with a low frequency signal. In order to quantify the advantage of TSV serialization, we introduce the serialized 3D link *area gain* (A_g) as the ratio

$$A_g = \frac{A_{par}}{A_{ser}} \tag{4.1}$$

where A_{ser} represents the area occupied by the serialization and buffering circuits on the transmitting layer, the deserialization circuit on the receiving layer and the corresponding TSV. A_{par} represents the area of N TSVs and corresponding transmitter and receivers, as depicted in Figure 4.4(b).

Figure 4.5 shows the trend of A_g serializing 4, 8, 16 or 32-bit for different TSV technologies for a 2-layer 3D-IC. Experimental data clearly show that the serialization scheme is beneficial for all the considered TSV technologies. Even for a serialization level of 4-bit, the area of the serial link is 3-4 times smaller than the area of the parallel counterpart for all the TSV technologies considered. As expected, the impact of using serialization consistently varies depending on the TSV dimension. For small footprint TSVs, e.g. $5\mu m$, A_g slowly increases with serialization rate since the areas of the SERDES circuits are comparable to the area of the TSV.

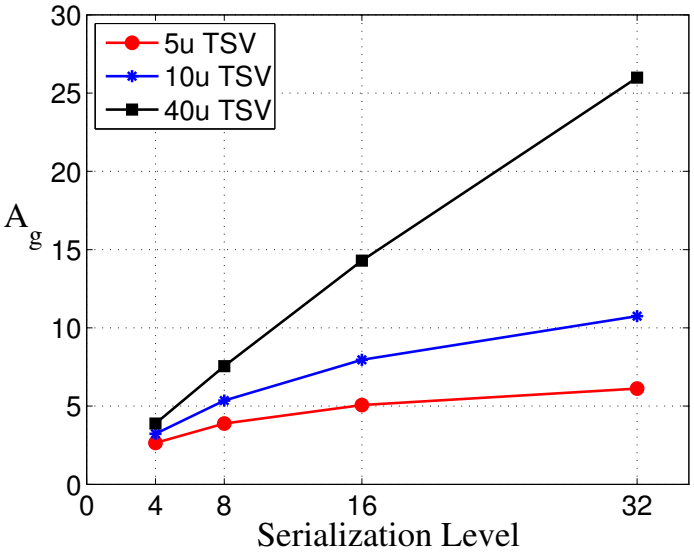


Figure 4.5: A_g of the 3D link for different serialization levels.

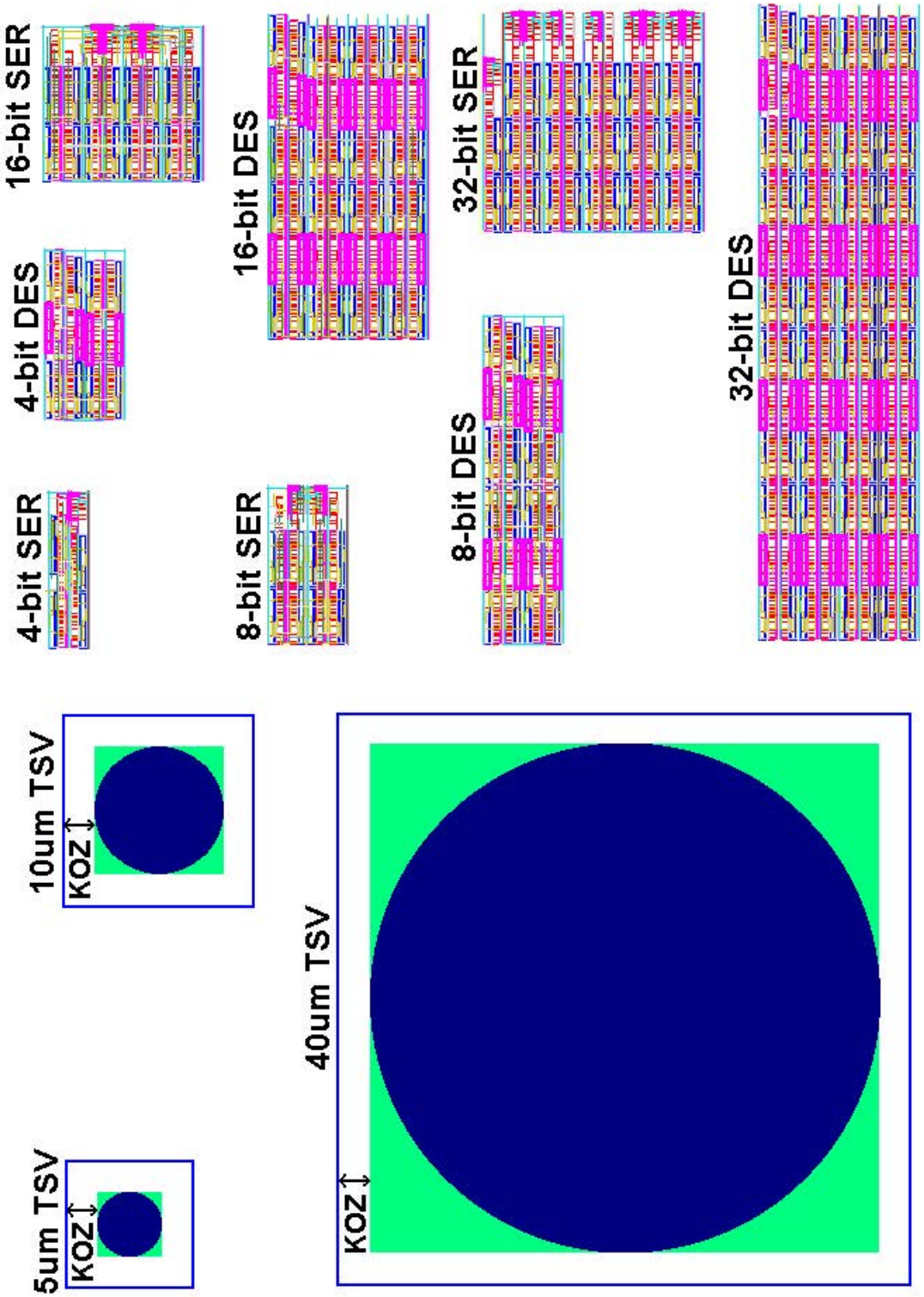


Figure 4.6: Layout of the implemented SER and DES circuits; TSVs are represented as a square pad with a KOZ of 2.5μm .

As the TSV diameter increases, its footprint becomes much larger compared to the area of the transmitter and receiver logic. We can notice that the area saving grows linearly with the serialization level.

Figure 4.6 depicts the layout of the SERDES circuits together with TSVs of different dimensions, clearly showing the larger area occupied by the TSVs with respect to the SERDES circuits. The observed area reduction for serial connections can significantly reduce routing congestion. Moreover, the reduction of the number of TSVs in a 3D-IC improves both crosstalk and fabrication yield.

4.4.3 Energy analysis

The observed area reduction comes at a cost: an increase in power dissipation due to the SERDES circuits. Hence the energy efficiency of the serial 3D link should be evaluated and compared to the energy of a parallel link. A *pseudo random bit sequence (PRBS)* was used to generate the parallel input stream. The energy efficiency of the circuit was extracted averaging the power consumption over 80.000 received bits. We define the *energy cost (E_c)* of serialization as the energy efficiency of the serial link over the energy efficiency of the parallel link

$$E_c = \frac{E_{ser}}{E_{par}}. \tag{4.2}$$

Figure 4.7 shows the energy trend for different TSV technologies: the higher the serialization rate, the more power is consumed by the SERDES circuits. The energy efficiency of the studied links are reported in Table 4.4. Even though the power consumption of the serial 3D link

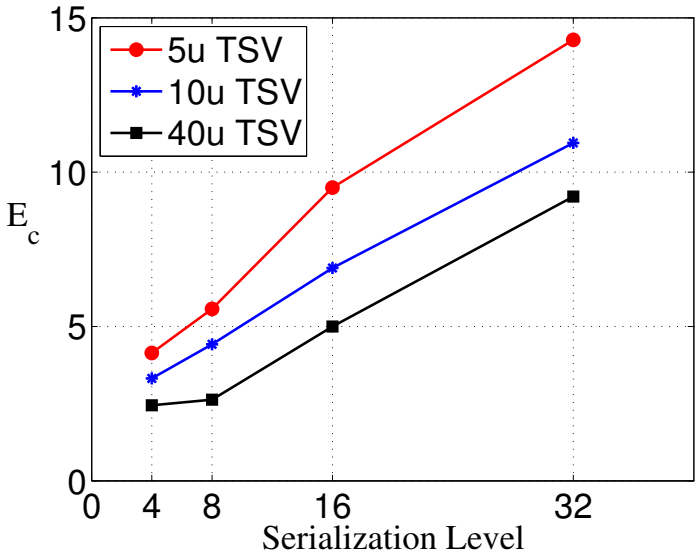


Figure 4.7: Energy cost E_c of the 3D link for different serialization levels for a 2-layers system.

is higher than its parallel counterpart, it is still relatively low once compared to off-chip signalling. For instance, serializing 8-bits over a 10 μm TSV consumes 84 fJ/b, achieving 10 Tb/s bandwidth within a power budget of only 8.4 W. Considering 95 W *Thermal Dissipation Power (TDP)* of an high-end desktop processor as Sandy Bridge form INTEL [79], our serial 3D link would provide ultra low power signalling within the 10% of the total TDP as required by ITRS [80].

Up to now, we have always assumed a 2-layers 3D stack, however in many applications, such as vertically stacked DRAM, it is desiderable to transmit signals over multiple layers. We performed an energy analysis on systems composed by M+1 layers, being M the number of TSVs that a transmitted signal should cross to reach the receiver. In a 3D-IC with more than two layers, the TSVs are connected serially. When the transmitter drives several stacked TSVs, the buffer should be replaced by a larger one to compensate for the increased TSV load, which means a more power hungry circuit. This trend is displayed in Figure 4.8 showing the energy efficiency for a 8bit SERDES 3D link and how it changes when the signal should cross more stacked TSVs. Interestingly, a negligible energy cost has been observed for 5 and 10μm TSVs crossing multiple layers. On the contrary, for 40μm TSVs, the energy grows linearly with the

TSV diameter	$5\mu m$	$10\mu m$	$40\mu m$
Parallel link energy [fJ/bit]	14	19	38
4-bit SERDES energy [fJ/bit]	58	63	93
8-bit SERDES energy [fJ/bit]	78	84	100
16-bit SERDES energy [fJ/bit]	133	131	190
32-bit SERDES energy [fJ/bit]	200	208	350

Table 4.4: Energy efficiency at 8Gb/s per channel for a 2-layers 3D stack.

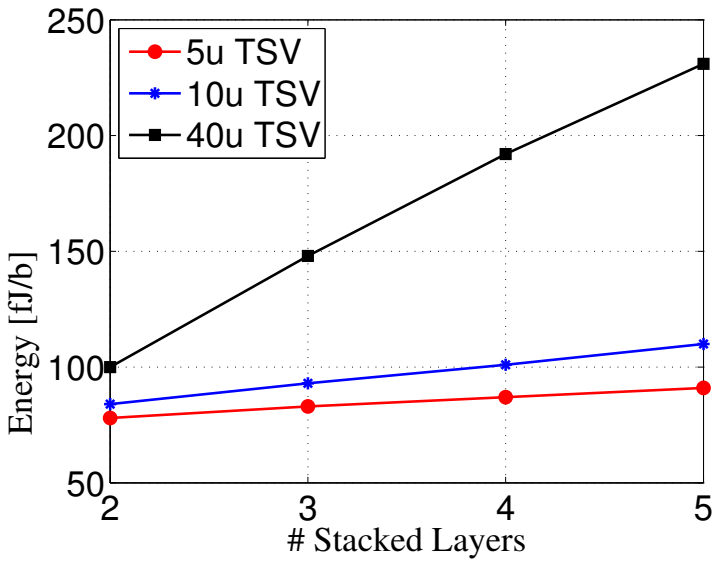


Figure 4.8: Energy efficiency of a 8bit 3D SERDES vs the number of crossed layers.

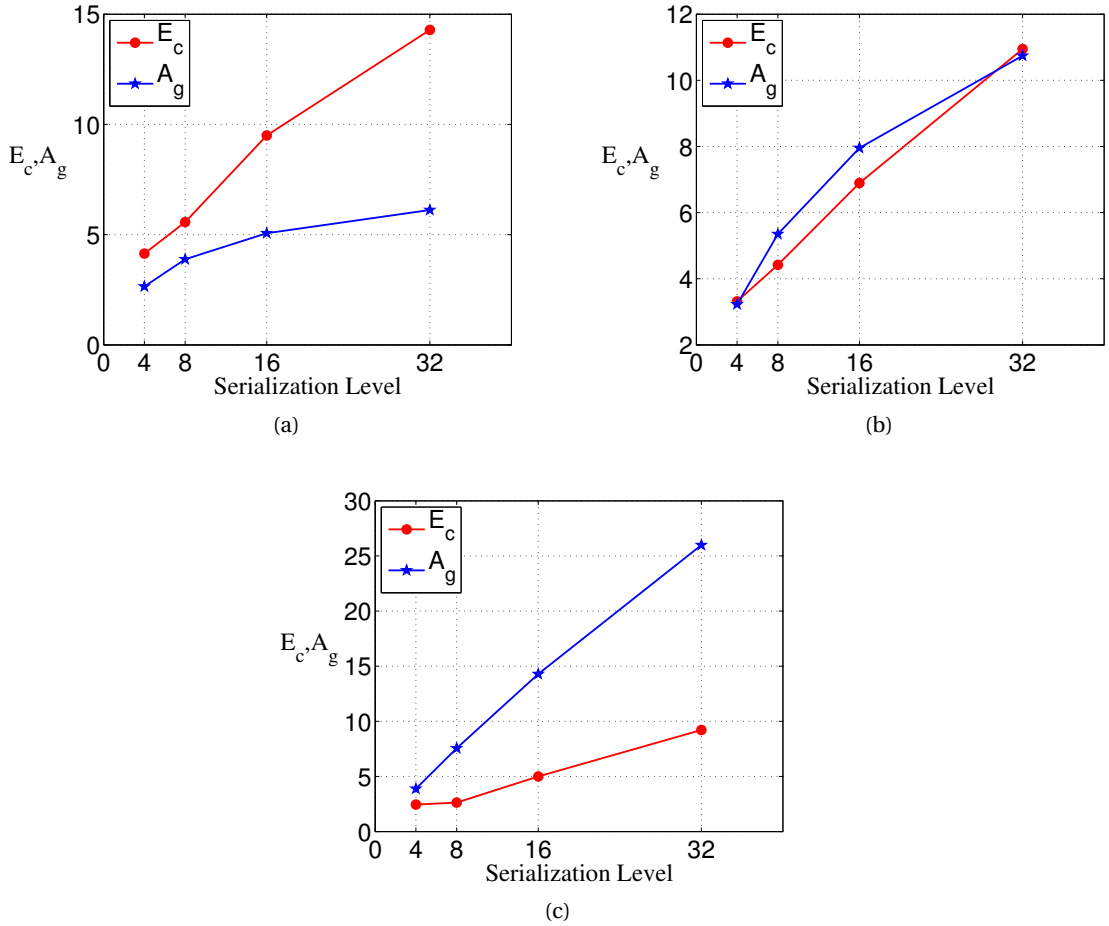


Figure 4.9: A_g and E_c for (a) 5μm TSVs (b) 10μm TSVs and (c) 40μm TSVs.

number of crossed layers.

4.4.4 Trade-off analysis

In the previous sections we have explored the area and energy impact of the proposed 3D serial link compared to a standard parallel topology. In order to fully evaluate the benefit of a serial link with respect to the parallel topology we should consider both A_g and E_c. Since the proposed scheme guarantees no performance loss, the benefits of serialization for a specific design can be easily evaluated without advanced architectural studies. Given the specifications for the power consumption, we can estimate the maximum energy per bit allowed for the intra-chip communication. Hence, for each design, the best serialization level can be chosen such that the power consumption of the vertical link does not exceed the one required by the design specifications.

This study can be considered as a reference for 3D -ICs architects. The graphs in Figure 4.9a,

4.9b and 4.9c depicts the area-energy trends for each TSV technology in a 2-layer system. For large TSVs the energy cost due to the serialization is almost negligible with respect to the area saved by reducing the number of TSVs. Reducing the TSV diameter we can see that A_g lowers while E_c increases. As discussed in the previous section, despite the power overhead, the energy per bit still remains low enough to consider the link as low-power. We define the energy gain as

$$E_g = \frac{1}{E_c} \tag{4.3}$$

The best trade off between area and energy can be evaluated by finding the serialization level that gives the most area saved at the minimum energy cost. To the end, the product $A_g xE_g$ plotted in Figure 4.10. The best area-energy trade-off corresponds to the maximum of the $A_g xE_g$ plot. It is interesting to notice that for the TSV technologies considered, this is independent from the TSV size, with 8-bit serialization level being the optimum point.

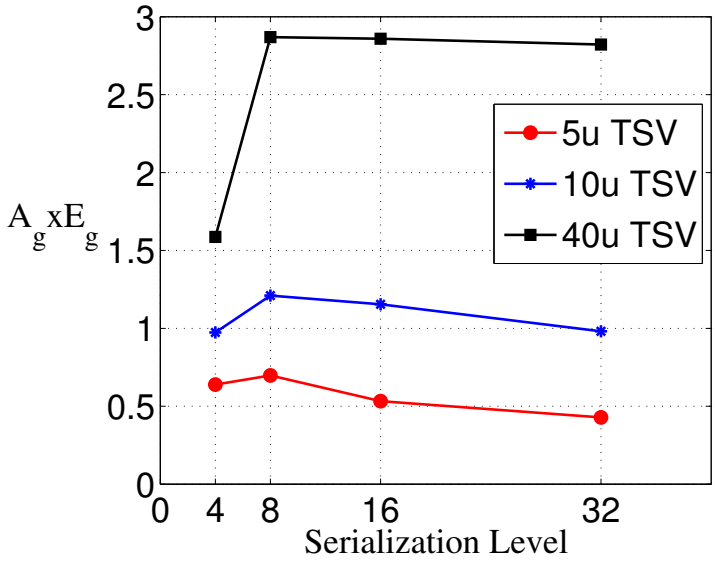


Figure 4.10: Area gain - energy gain product.

4.5 8-bit serial link

The previous section analyses the different serial links versus the parallel correspondent. A serialization level of 8 has emerged as the best compromise between power and area. In this section, we analyse the behaviour of the 8-bit serial 3D-transceiver.

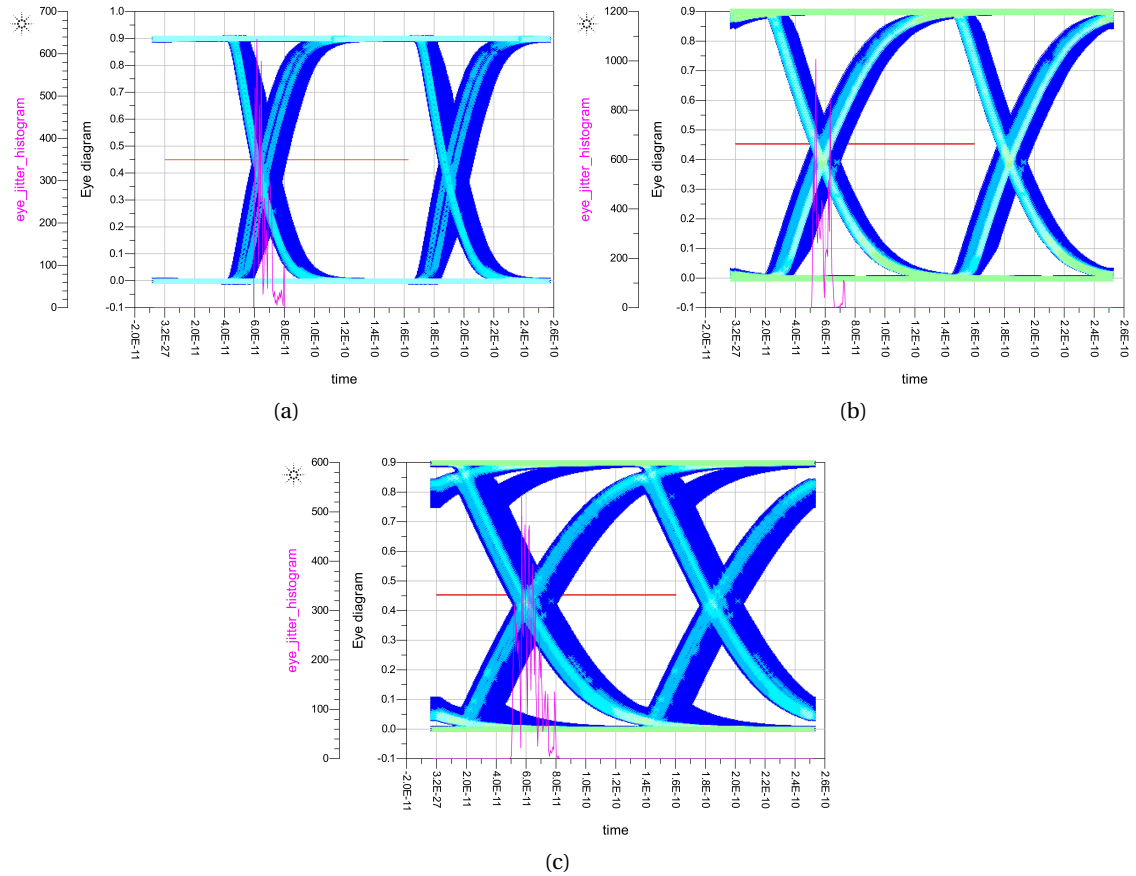


Figure 4.11: Eye diagram and jitter histogram for the serializer driving a (a) 5μm (b) 10μm and (c) 40μm TSV channel .

4.5.1 Jitter analysis

When transmitting digital information, the *Bit Error Rate/Bit Error Ratio (BER)* defines the quality of the communication system. The BER can be defined as the number of bits received in error (N_{berr}) over the total number of received bits (N_b).

$$BER = \frac{N_{berr}}{N_b} \tag{4.4}$$

The test time required to directly measure a low BER is excessively long. Instead, the jitter can be used to determine the quality of the link. Jitter is defined as the short term variation of the significant instants of a digital signal from their ideal position in time [81] and is fundamentally an expression of phase noise. Noise-related issues become particularly critical in high speed serial data links which have to keep up with the increase in desired data rate.

The jitter generation, or intrinsic jitter, is the jitter generated by a component when the input has no jitter [82]. In order to track the intrinsic jitter caused by the proposed system and how

TSV diameter	5μm	10μm	40μm
Eye amplitude [mV]	900	870	816
Eye height [mV]	890	770	550
Eye width [ps]	106	106	100
Eye opening factor	0.997	0.961	0.890
Average rise time [ps]	31	43	58
Average fall time [ps]	33	40	54
Jitter rms [ps]	4.67	4.84	6.8
Jitter pp [ps]	20.5	21.4	32.2

Table 4.5: Eye diagram measurements.

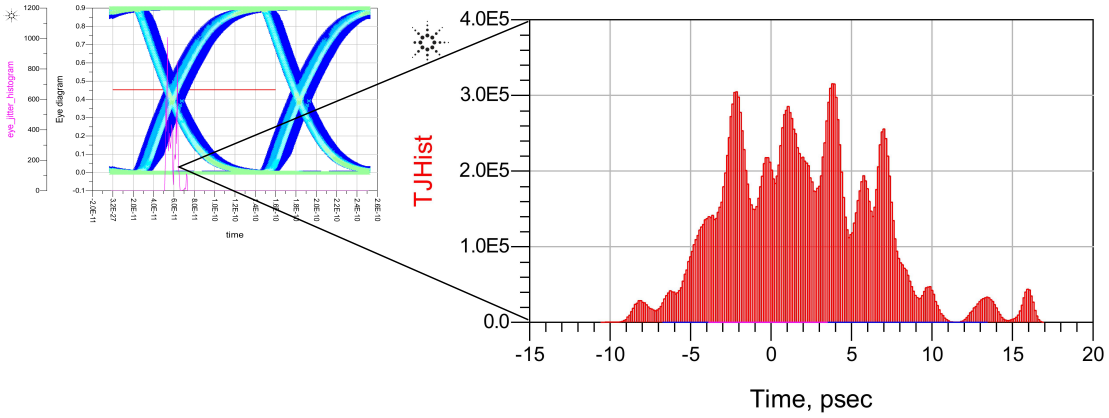


Figure 4.12: Histogram of the total jitter for the serializer driving a 10μm TSV channel.

it affects the BER, the netlist of the serial link has been simulated in *Agilent ADS (Advanced Design System)*. The eye diagram of the serialized data stream at the output of the TSV channel are shown in Figure 4.11 for different the TSV technologies in a 2-layers 3D system. 10^4 cycles have been simulated at a serial data rate of f_s=8Gb/s in order to plot the eye diagram. As expected, the eye opening worsen as the TSV dimension increases due to the higher parasitic load. The received eye diagrams clearly show an open eye in absence of equalization. The eye measurements are summarized in Table 4.5.

The jitter in a system can be described by its *Probability Density Function (PDF)*, Figure 4.12 depicts the crossing point histogram of the total jitter in the proposed serial link. In order to determine the cause of the jitter, we need to separate the different components. The total jitter is the convolution of *Random Jitter(RJ)* and *Deterministic Jitter (DJ)*. From the peaks of the *Total Jitter(TJ)* histogram, is immediately clear the presence of both RJ and DJ.

DJ is caused by systematic problems within the system and is bounded. the three sources of DJ are *Periodic Jitter(PJ)*, *Duty Cycle Distortion (DCD)* and *Inter Symbol Interference(ISI)*. Considering that the system has been simulated without any supply noise and stand alone, there are no coupling phenomena with adjacent TSVs causing PJ. Performing the transient

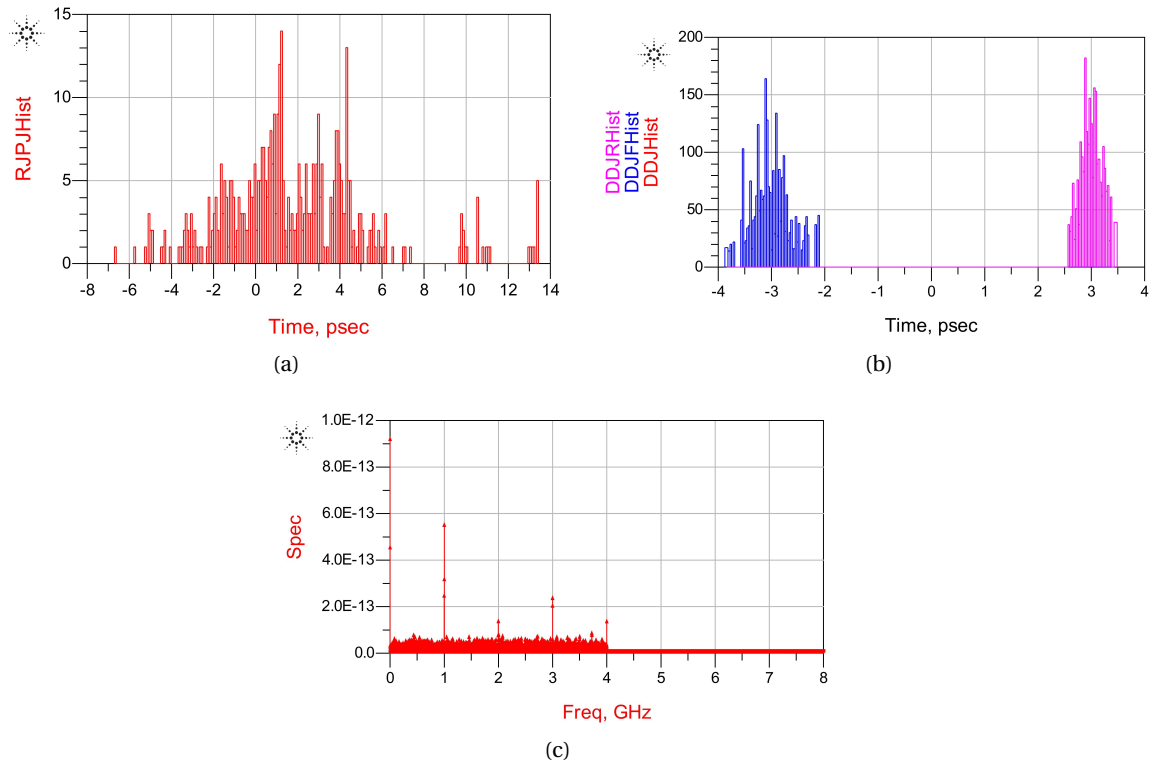


Figure 4.13: (a) Histogram of the random and periodic jitter, (c) histogram of the data dependent jitter, (b) BER for the serializer driving a 10μm TSV channel.

noise analysis of the system, which generate pseudo-random noise at each step, the results from the accumulation of random processes can be observed in Figure 4.13a.

The Data Dependent Jitter (DDJ), shown in Figure 4.13b, depicts the histogram of correlated jitter, which can be caused by ISI and DCD. Nonetheless, the bandwidth of the TSV channel is well beyond the Nyquist frequency ($\frac{1}{2}f_s$=4GHz), hence, the contribution of ISI to the jitter is negligible. Therefore, the DDJ in the system is primarily caused by asymmetric rising/falling edges and offset in the transmitter sampling threshold. A confirmation comes from the noise spectrum in Figure 4.13c. In presence of significant DCD, the jitter spectrum presents a component at $\frac{1}{2}$ the data rate, hence at 4GHz. Since the input pattern is random, we can also see components at 1, 2 and 3GHz.

The graph in Figure 4.14 depicts the bathtub plot: the traces in red represent the bit error rate calculated directly from the data, while the blue traces represent the BER extrapolated from the calculated values of RJ and DJ. At a BER of 10^{-12} the stand alone link shows an eye aperture of 62% for the 10μm TSV channel, that reduces to 50% for the 40μm TSV channel. Using the serial link in a real design, the eye aperture will worsen due to cross-talk, clock jitter and supply noise. In the next section we analyse the effects of the high speed clock distribution.

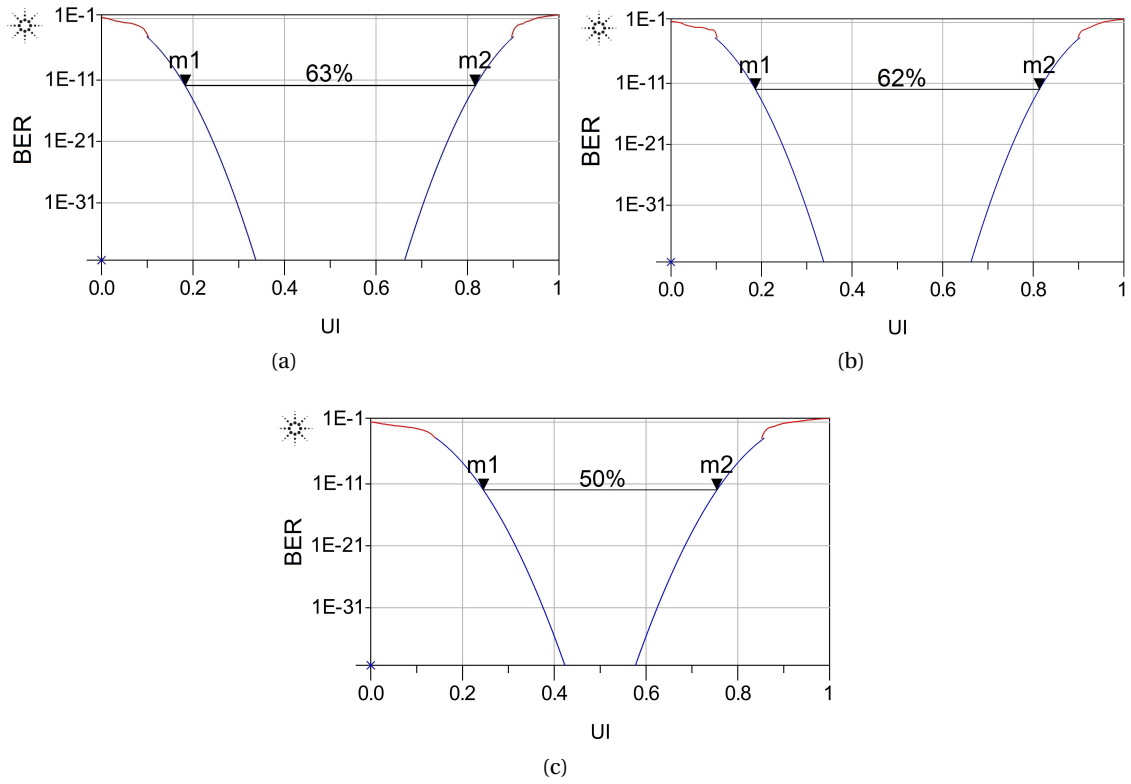


Figure 4.14: BER for the serializer driving a (a) 5μm, (b) 10μm and (c) 40μm TSV channel.

4.5.2 Clock distribution

The overhead of the high speed clock distribution has not been taken into account for the area-power exploration in Section 4.4 since it should be shared among all the 3D links in a design.

In this section, a test case has been defined in order to evaluate the power overhead caused by the additional high speed clock of the serial connection. Considering that an iPhone5 from Apple utilizes a memory bandwidth that has a maximum speed of 8528MB/s [83], we assume a memory on logic system with a bandwidth requirement of 10GB/s. Choosing a serialization level of 8 working at 8GHz, 10 TSVs are needed for the inter-layer data transmission.

The clock distribution scheme chosen for this study is depicted in Figure 4.15. The high speed clock CLK_in is propagated to the next layer through a TSV. On each layer the clock is first decoupled and the square wave is restored by an analog buffer with a feedback which guarantees a 50% duty cycle. The restored clock signal is than propagated through the clock tree distribution network. A frequency divider is used to create the system clock. Both the high speed clock and the system clock are distributed on each stacked die with identical distribution networks.

The post-layout simulations of the system show an overhead of 75%-80% in the energy efficiency due to the high speed clock distribution. Table 4.6 summarizes the power consumption of the system for each TSV technology considered. In case of 10μm TSV, around 63fJ/bit are required by the clock distribution network.

Figure 4.16 displays the eye diagram of the serial data stream for one of the TSV channels in the

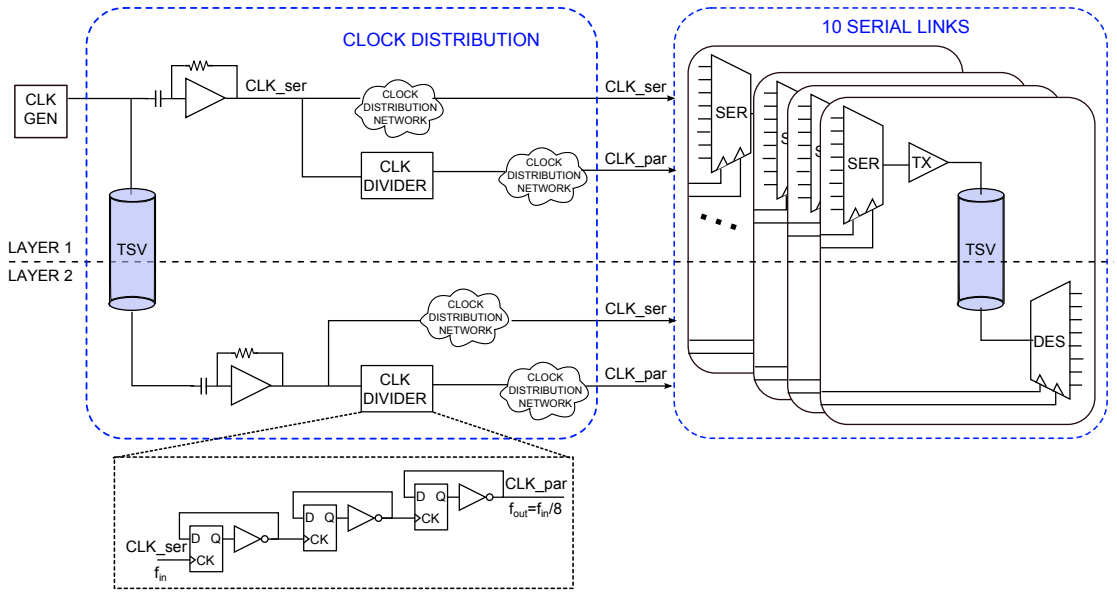


Figure 4.15: Clock distribution scheme for the 10GB/s system.

TSV diameter	$5\mu m$	$10\mu m$	$40\mu m$
Power [mW]	11.3	11.7	14.4
Energy efficiency [fJ/bit]	141	147	180
Energy overhead due to clocking	81%	75%	80%

Table 4.6: Energy efficiency of the system composed by 10 8bit SERDES TSV links for a 2-layers 3D stack delivering a total aggregate bandwidth of 10GB/s.

considered system for a serial data rate of 8Gb/s. The main parameters extracted from the are summarized in Table 4.7. Compared to the single channel system described in Section 4.5, the eye width has reduced from 106mV to 85mV, while the jitter rms value has increased from 4.8ps to 20ps. From the eye diagram we can notice that the clock distribution network significantly add DDJ to the system's jitter, due to the non ideal clock rise and fall times. Nonetheless, the eye aperture is still enough for a correct data recovery in absence of equalization from the receiving deserializer.

4.6 Double data rate TSV serial link

Assuming that the input clock CLK_in is delivered by an on-chip PLL, for the parallel system the PLL should create a 1GHz clock signal, while for the serial system the clock should have a frequency of 8GHz. A PLL in 40nm technology is expected to consume 4.5mW/GHz according to the ISSCC 2013 trends [84]. Therefore the PLL power for the serial link would be around 36mW, that added to the system would worsen the energy efficiency of the serial vertical

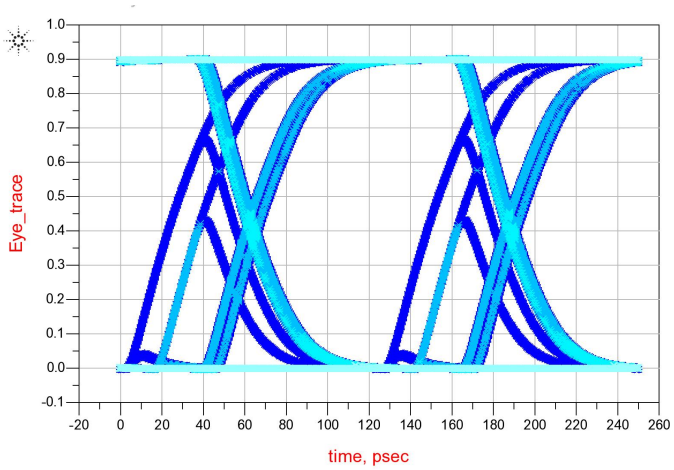


Figure 4.16: Simulated eye diagram for one of the TSV channel in the system.

Eye amplitude [mV]	Eye height [mV]	Eye width [ps]	Jitter rms [ps]	Jitter pp [ps]
888	792	85	20	102

Table 4.7: Eye diagram measurements.

TSV diameter	$5\mu m$	$10\mu m$	$40\mu m$
8-bit SDR SERDES energy [fJ/bit]	78	84	100
8-bit DDR SERDES energy [fJ/bit]	132	138	172
Energy overhead	70%	64%	72%

Table 4.8: Energy efficiency at 8Gb/s per channel for a 2-layers 3D stack.

connection. Depending on the design requirements, the area reduction may not be sufficient to justify an excessive cost in power consumption from the insertion of the high speed clock in the design. Hence, in view of the fact that a significant part of the power is consumed by the high speed clock generation and distribution, reducing the working frequency is desirable in order to improve the energy efficiency.

A solution to this problem is to use a serialization scheme which utilizes the same clock as the parallel data stream. However, this approach reduces power by increasing the transmission latency. Since our goal is to improve the energy efficiency without sacrificing the bandwidth, a *Double Edge Trigger Flip-Flop (DET-FF)* has been implemented to trigger both at the rising edge and falling edge of the clock. By replacing the standard flip-flop with DET-FF, the high speed clock frequency can be reduced by a factor of two, still achieving the desired data rate. Figure 4.17 depicts the full-custom layout views of the 8-bit DDR SERDES circuits. Compared to a fully parallel vertical interconnection, the 8-bit DDR serial link has an area gain A_g=4.9, slightly lower than the A_g=5.2 of the *single data rate(SDR)* studied in Section 4.4.

In terms of power, the DDR single channel has an energy overhead around 70% compared to the SDR solution, as summarized in Table 4.8. Nevertheless, including the clock distribution

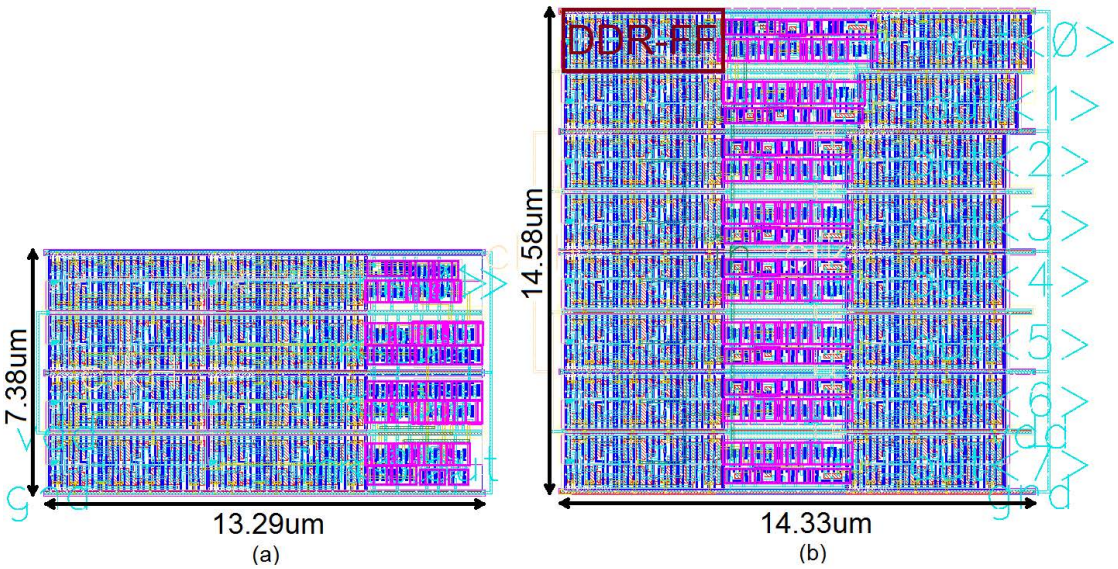


Figure 4.17: Full custom layout views of the 8-bit double data rate (a)serializer and (b) deserializer in 40nm TSMC technology.

	Energy efficiency full system [fJ/bit]
SDR	597
DDR	376

Table 4.9: Energy efficiency of the system composed by 10 8-bit SERDES TSV links for a 2-layers 3D stack delivering a total aggregate bandwidth of 10GB/s.

network for the considered 10GB/s system described in Section 4.5.2 using 10μm TSV channels, the energy efficiency becomes almost equivalent to the efficiency of the SDR system. Including the power required by an on-chip PLL to generate the high speed clock, the energy efficiency of the DDR system is 37% less than the SDR counterpart. Table 4.9 summarizes the energy efficiency of the SDR and DDR systems including the clock distribution network and the high speed clock generation circuitry.

In our study we have used a standard dual edge triggered flip-flop composed by two FFs, one triggering on the rising edge and the other on the falling edge, each of them connected to an input of a multiplexer that then selects the output depending on the clock level. Although this FF topology helps reducing the clock frequency by half, it almost doubles the area and increases the load on the data and clock inputs, which worsen the power consumption. A further optimization of the energy efficiency can be therefore achieved by re-designing the DET-FF topology targeting low-power consumption.

4.7 Summary

The potential of 3D IC is limited by the large area footprint of its vertical interconnects. In this chapter we propose a 3D serial link that reduces the number of TSVs maintaining the performance unvaried. A serialization scheme is proposed in order to exploit the TSVs' high bandwidth. We show how the serialization level can affect both area and energy for different TSV technologies.

For a mature TSV fabrication technology, such as 40μm TSVs, 15X area reduction can be achieved within a low power budget for 16-bits serialization. A serialization level of 8-bit guarantees a good balance between area consumption and energy efficiency across all the explored TSVs. Considering a 2-layers system with an aggregate bandwidth requirement of 10GB/s, an 8-bit data serialization over 10μm TSVs consumes just 147fJ/bit including the clock distribution network. Finally, we show that for systems dominated by the power consumption of the clock generation and distribution, a double data rate topology can improve the system's energy efficiency.

5 TSV Serialization Impact on a 3D Modular Multi-Core Processor Platform

As demonstrated in Chapter 4, the use of serial vertical interconnects can significantly reduce TSV area footprint with a reasonable power overhead and no performance loss. An optimal serialization level of 8-bit has been found as the best trade-off between area and power consumption for different TSV technologies.

Although 3D integration can alleviate routing congestion, reducing the wirelength and improving performances, a TSV still occupy non-negligible silicon area. As the number of TSV increases, their effect on the chip routing is detrimental. The reduction in the number of 3D vias obtained with the adoption of the serial vertical connection can relieve the routing congestion of the 3D system by reducing the average wirelength. In this chapter we explore the impact of the serial approach on the chip routing to quantify the achievable wirelength reduction. To this end, a modular multi processor platform, 3D-MMC, composed by completely identical stacked chips has been designed. The proposed system uses TSVs to transfer data across layers, creating an expandable 3D network of processing cores to improve performance.

5.1 Problem formulation

With the advent of deep-submicron CMOS technologies, on-chip interconnect wires have rapidly gained a lot of attention. With the scaling of the CMOS technologies, the parasitics effects due to the wiring does not exhibit the same scaling behaviour as the CMOS logic. As the IC feature sizes shrink, device area shrinks roughly as the square of the scaling factor while the device propagation delay improves almost linearly with the deacrease in feature size under constant field assumption. On the other hand, interconnect delay does not scale with feature size, and tend to gain importance as device dimensions are reduced and circuit speed is increased. Worsening the situation, as the silicon dies get larger, also the average length of the interconnects increases, hence their associated parasitic effects. As a consequence, the interconnects start dominating some of the most important metrics of digital ICs, such as speed, power consumption and reliability.

A promising solution to break through the interconnect wall emerged with the advent of 3D integration featuring TSVs. As discussed in Chapter 1, 3D ICs have the potential to reduce interconnects length and improve the system performance. Nevertheless, as discussed in Chapter 2, technological processes necessary for fabricating the TSVs connecting the superimposed layers can not yet be regarded as mature. According to the ITRS roadmap [2], the TSV diameter will not shrink below 2-4 μm for global interconnects. Using small vias is desirable to reduce the chip footprint, yet, as the diameter decreases, the TSV fabrication yield worsen. Since the silicon area occupied by a TSV is quite significant, it interferes with cells placement, spreading them out and limiting the achievable reduction of the average routing distance. In case of most via-last TSVs, the impact becomes even more severe since this TSVs occupy all metal layers, becoming a routing obstacle. Hence, as the number of TSVs increases, the wirelength and form-factor benefits of 3-D ICs significantly reduce, as demonstrated by Kim et al. [85].

Serial vertical TSV interconnects have been proposed as an effective solution to keep under control the TSV count. This chapter aims to explore the impact of serialization on the routing congestion of a 3D-CMP. A *3D Modular Multi-Core (3D-MMC)* architecture has been designed and used as test case. Differently from the architectural approach explored in Chapter 3, where we proposed a memory on logic CMP system, 3D-MMC is based on the integration of identical multi-processor layers. Thanks to the homogeneous approach, the system performance can be augmented with minimal design cost compared to conventional planar IC designs.

5.2 State of the art

A significant amount of recent work has been focused on exploring the potential benefits of 3D stacked processor architectures. In 2006, Black et al. [65] have proposed to arrange the logic modules of an Intel Pentium 4 microprocessor in clusters and re-organized them in two stacked layers, resulting in 15% performance gain and 15% power savings at constant frequency. In the same study, a memory-on-logic solution is also presented, using as a simulation vehicle an Intel Core 2 Duo unit. The implemented architecture aims to increase the cache capacity by stacking a memory layer on top of the dual-core die, highlighting the reduction of both latency and access memory time.

Many other examples of 3D processors, involving a heterogeneous partitioning, have been presented. An early approach [50] was based on the superimposition of layers containing both cores and cache banks, interconnected by a Network-in-Memory. A placement algorithm was used for placing the processing units with a three-dimensional offset to avoid thermal problems. One of the first solutions implementing multiple memory layers on top of processors was presented by Kgil [86], modelling a web server as a CMP built of 4 DRAM layers stacked on top of a processing die hosting up to 8 parallel cores. Other CMPs have been designed in later years exploiting multiple 3D-DRAM layers [87] [52]; these solutions showed the possibility to re-organize modules interconnect fabrics in order to have a significant bandwidth increase,

resulting in a relevant speed-up in the routine execution. Loh's [51] solution demonstrated an achievable speed-up of 280% with respect to the baseline CMP (an Intel QuadCore) connected to off-chip DRAM.

Nevertheless, the TSV size and count strongly affect the performance of 3D ICs. The wire-length reduction varies depending on the number of TSVs. The impact of TSV size on the 3D wirelength distribution was first studied by Kim et al. [85] demonstrating that the wirelength increase due to TSV placement is not negligible. In high density 3D-ICs, routing congestion can cause routing failure or re-design from the beginning, to tackle this issue Ahn et al. [88] have proposed a precise routing congestion estimation method at the floorplan stage, which is beneficial to reduce the total design cost. A different approach focused on co-optimizing the TSV count and wirelength at the placement level was studied by Tsai et al. [89] and Cong et al. [90], while Lee et al. [91] have developed an algorithms for TSV resource sharing and optimization. While these previous works mostly focus on placement algorithms to reduce the TSV area impact on the design, this chapter analyses the benefits of the serial approach on the routing congestion of a 3D chip multi-processor system.

5.3 3D modular multi-core architecture

In order to explore the potential of the serial data transmission through TSV, a modular 3D architecture, 3D-MMC, built by stacking identical layers has been developed. Figure 5.1 presents a basic block diagram of the stacked structure with a 2-layer configuration for the sake of simplicity.

The novel and unique architecture has been specifically designed for stacking identical dies in order to form the 3D system. Without loss of generality, the proposed architecture can be expanded to include multiple identical layers that can communicate with each other. Each die can be considered as a planar multi-core architecture, composed by multiple *Processing Elements* (PEs), working in parallel. The cores exchange data through a shared memory implemented in the *Peripheral Subsystem* (PS) unit; the access of PE to the shared memory is arbitrated by a system of semaphores to avoid contention. The interaction between cores occurs through a specific source-routed NoC, composed of a 36-bit switch, in charge of the effective signals routing to and from 6 directions (North, South, East, West, Up, Down), and a Network-Interface (NI) for each logic block present on the layer. The network system has a 3D folded architecture in order to enable the management of the signals in both the horizontal and vertical directions. The intra-layer communication is achieved through the introduction of a 3D connection macro, exploiting arrays of TSVs as vertical data bus.

5.3.1 2D layer architecture

A single layer consists of four *Processing Elements (PE)* that exchange data through a shared memory, which is placed in the *Peripheral Subsystem (PS)* unit. The routing between each PE

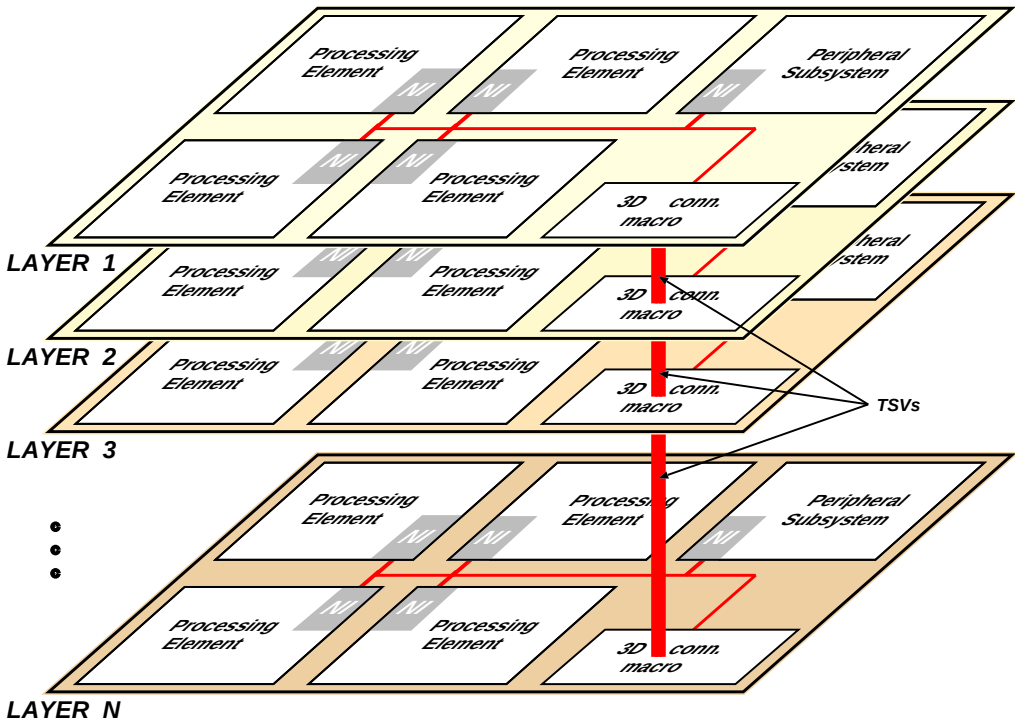
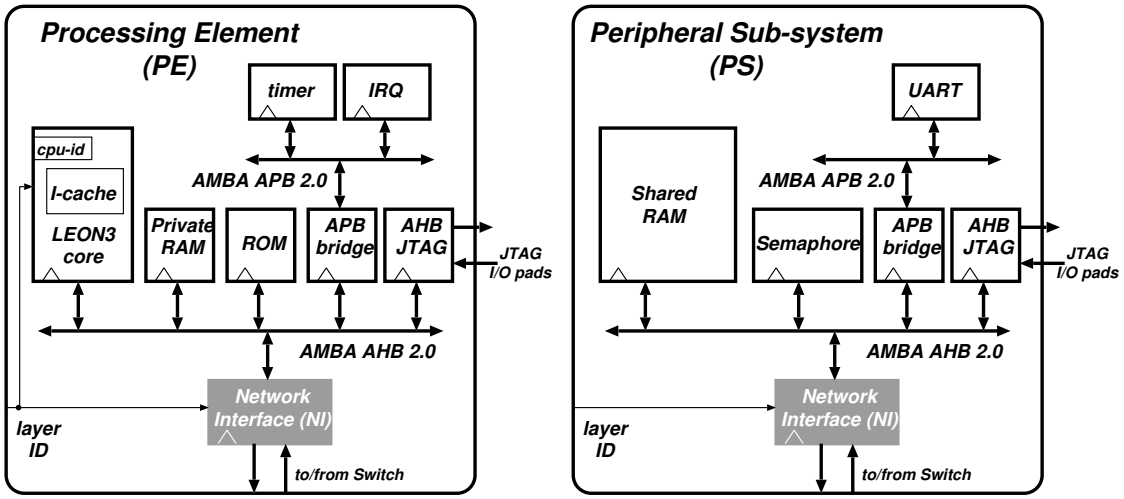


Figure 5.1: Block diagram of the 3D-MMC architecture. The generic 3D connection macro block on each identical layer allows the inter-layer communication among multiple layers, with serial multiplexed TSV arrays.



(a) (b)

Figure 5.2: (a) Processing element (PE) internal architecture, with the LEON3 core and its private modules. Each unit is accessible through JTAG ports for debugging purposes. The network interface (NI) routes packets from PE to the shared memories in the (b) Peripheral Subsystems (PS).

and the shared memory occurs through a specific source-routed NoC. In the 3D stack, NoC on different layers are interconnected to enable the management of the signals in both the horizontal and vertical directions.

Figure 5.2a and Figure 5.2b illustrate the internal architecture of a PE and a PS, respectively. Each PE is built out of a 32-bit RISC processor, the open-source LEON3 unit from Aeroflex Gaisler, a general-purpose unit able to perform a wide range of applications, making the designed architecture eligible for several market segments. The LEON3 unit is connected to slave modules through an AMBA bus. The PE is accessible by off-chip components through a JTAG port for debugging and pre-loading of each core's memories with desired data. Each core utilizes privately addressable memory space, composed of a 32 KB ROM, containing the boot sequence, and a 32 KB RAM, as well as a common memory space composed by the system shared memories. The access of the cores to the shared space is regulated by a semaphore module present in the PS, able to avoid conflicts in case of simultaneous requests.

Multi-core interactions are managed by the *Network-Interface (NI)*. The NI block is a master located within both PE and PS. It interfaces the AMBA bus to the NoC, and is responsible of forwarding/receiving data packets to/from the shared memory, which has an addressing space visible by each core. This NoC has been specifically adapted from [33] for the proposed CMP architecture. The 7x7 Switch is characterized by 5 horizontal interfaces (one for each PEs plus one for the PS) and 2 vertical ports (for the upper and lower dies), through which 36-bit FLIT (*FLow control unITs*) packets are transmitted. Similar to PEs, the PS contains NI and AHB JTAG acting as master modules whereas all the remaining units (semaphore, shared RAM) act as slaves.

Each PE in an N-layer system has access to N+1 different memory modules that can be accessed in parallel: a private-RAM contained in his own PE, a shared local-RAM located in the PS of its layer, and N-1 shared remote-RAMs situated in the PS of the other stacked layers. In the same way as proposed by Benini et al. **[92]**, the proposed memory hierarchy with shared data memory for inter-processor communication simplifies the hardware complexity and avoids memory coherency overhead. The multi-core synchronization is handled at the software level.

5.4 Serial vs. parallel vertical link

In order to include the serial vertical connection into the semi-custom digital design flow, the SERDES circuits presented in Chapter 4 have been implemented in RTL and synthesized with the UMC 90nm CMOS technology library using Synopsys Design Compiler. The functionality has been verified using Mentor Graphics ModelSim.

Table 5.1 summarizes the maximum working frequency, the gate area and the power consumed by the SERDES circuits. As expected, the maximum working frequency achievable by the semi-custom solution is limited by the clock to Q delay of the flip-flop available in the library. The

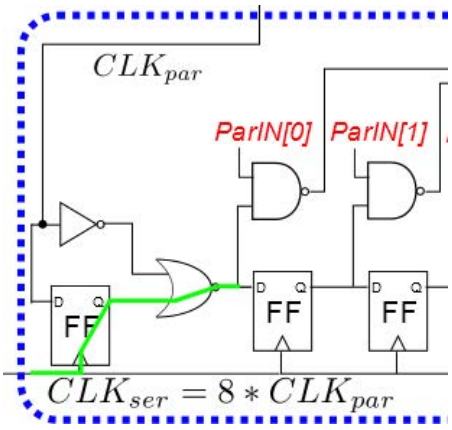


Figure 5.3: Critical path of the serializer circuit.

serializer critical path is highlighted in Figure 5.3.

As discussed in Section 5.3, the network system has a 3D folded architecture which extends its capability of managing the signal transmission also in the vertical direction through a 3D connection macro. Figure 5.4a depicts the traditional parallel configuration exploiting one TSV for each inter-layer signal. Instead, Figure 5.4b depicts the serial configuration of 3D-MMC. The SERDES circuits have been integrated in the 3D connection macro at the vertical interface of the NoC. The 174 signal TSVs required for the parallel configuration are reduced to 34 after serialization: specific signals, like the clock, reset, layerID (2-bits) and JTAG debugging signals (TCK, TRST, TDI, TDO, TRST) are directly sent to the above and bottom layers, while the rest of the NoC data (144) and control signals (12) are grouped into bytes, serialized and then sent through the TSV channels to the upper and bottom layers. The number of TSVs marked in Figure 5.4 consider both the TSV to the lower layer and the TSVs to the upper layer.

5.4.1 Physical design

Both the serial and parallel version of 3D-MMC have been implemented in RTL. The designs have been synthesized with the UMC 90nm CMOS technology library using Synopsys Design Compiler. The layouts have been placed and routed with Cadence Encounter. The functionality has been verified using Mentor Graphics ModelSim. The multiprocessor has been constrained to work at 200MHz, with the serial vertical interconnect working at 1.6GHz.

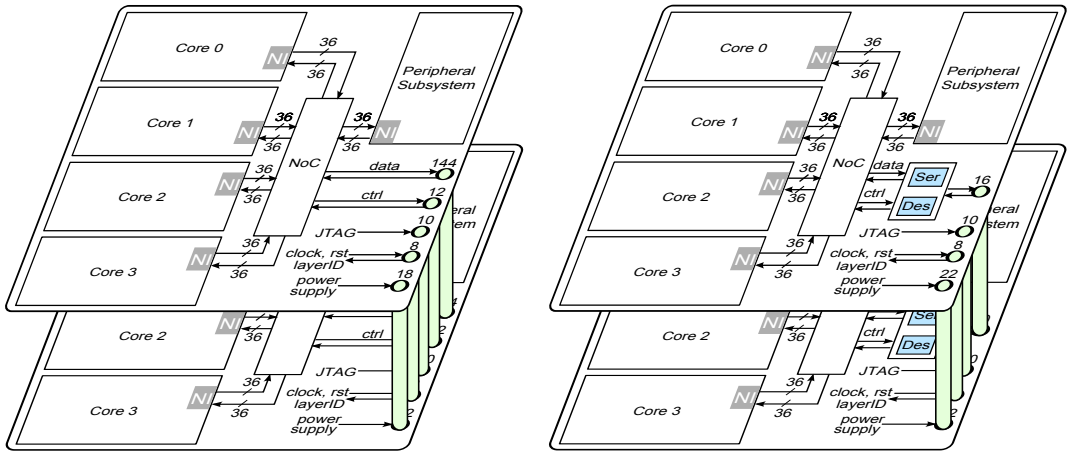
The routing analysis has been performed for 3D ICs based on the TSV technologies presented

	$F_{parallel}$ [MHz]	F_{serial} [GHz]	Area [μm^2]	Power [uW]
Serializer 8:1	312	2.5	154	262
Deserializer 8:1	390	2.3	406	608

Table 5.1: SERDES characteristics.

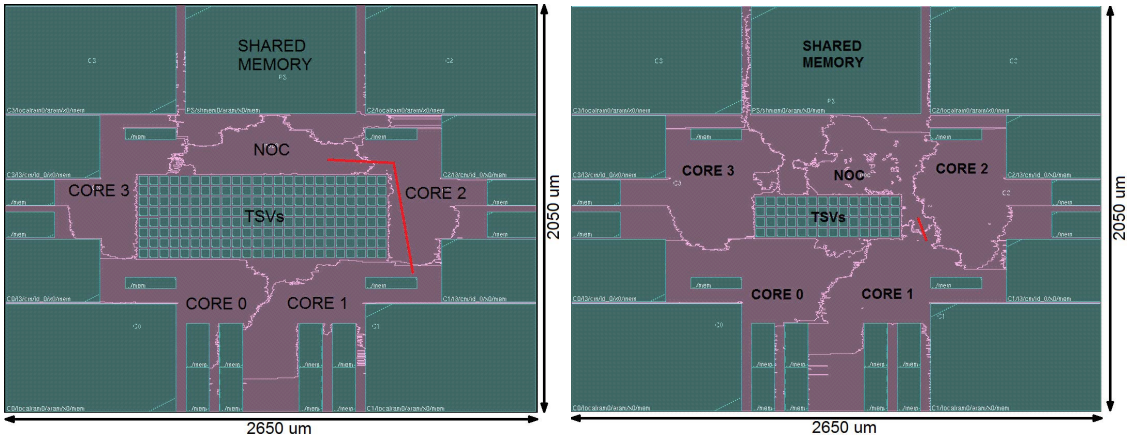
in Table 4.3 of Chapter 4 assuming a via-last process. 5μm TSVs represent state-of-the-art for high density through silicon vias [33], 40μm TSVs are a more established technology which guarantees better reliability [2] while 10μm TSVs provide a fair compromise between the two extreme. All the TSVs are placed in a TSV array located in the center of the chip, as depicted in Figure 5.6. A 5μm keep out zone and a minimum distance of 2.5μm from the metal interconnects has been used. Beyond the vertical signal connections, 22 TSVs has been added for the power and ground delivery in each design.

All designs have been constrained within a chip area of 2050μmx2650μm. The serial configuration includes 10 serializers and 10 deserializers. The design parameters are summarized in



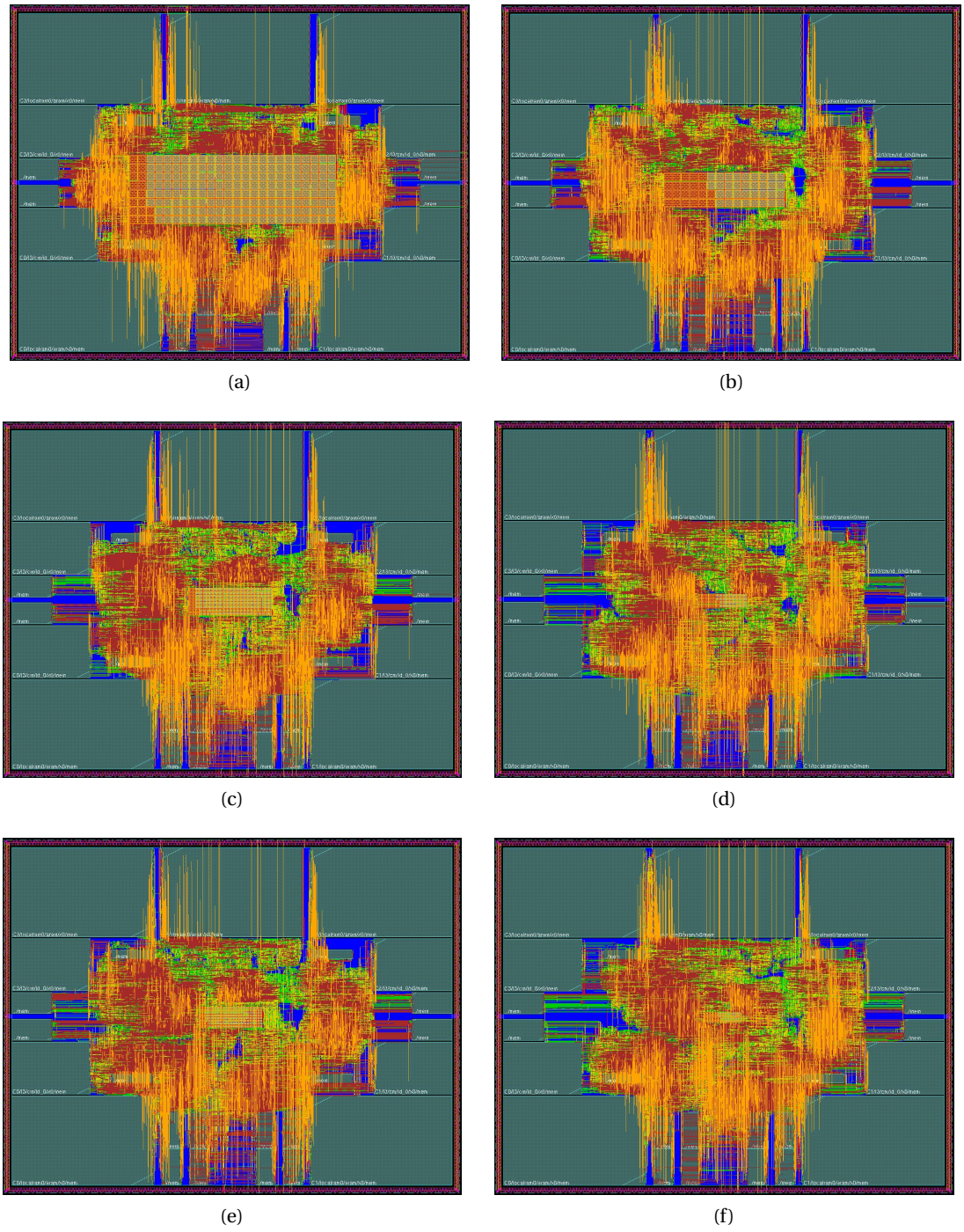
(a) (b)

Figure 5.4: View of the 3D-MMC (a) parallel and (b) serial configurations.



(a) (b)

Figure 5.5: View of the 3D-MMC (a) parallel and (b) serial configurations design with 40μm TSV channels. The red line depicts a path from CORE1 to the NoC.



(a) (b) (c) (d) (e) (f)

Figure 5.6: Layouts of the 3D-MMC design with (a) a parallel vertical bus and (b) with serialization through 40μm TSV channels. Layouts of the 3D-MMC design with (c) a parallel vertical bus and (d) with serialization through 10μm TSV channels. Layouts of the 3D-MMC design with (e) a parallel vertical bus and (f) with serialization through 5μm TSV channels.

	Parallel configuration			**Serial configuration**		
Chip size[μm]	2050x2650			2050x2650		
Signal TSVs	174			34		
Power TSVs	18			22		
KOZ [μm]	5			5		
TSV size [μm]	5	10	40	5	10	40
TSV array dimension [μm] (TSVs+KOZ)	45x145	65x215	185x635	85x245	125x365	365x1085

Table 5.2: Physical design parameters.

Table 5.2. The design placement of 3D-MMC in Figure 5.5 clearly show the TSV area reduction, and the consequent effect on the cell placement achieved by serializing the signal delivered by the 40μm TSV channels.

The parallel and serial placed and routed designs for each considered TSV technology are depicted in Figure 5.6 (the two top metal layers are not visible to better display the routing congestion around the TSVs). The memory macros of each core, such as private RAM, register file, instruction cache, are placed all around the core area, while the TSVs are placed in a matrix in the center of the chip. As the TSV diameter increases, the area occupied by the TSV becomes comparable to the area occupied by the active devices.

5.5 Routing analysis

In Chapter 4 we presented the exploration of the area-power trade-off for the serial configuration versus the parallel one. Nevertheless, the impact of the TSV footprint on the chip area is not the only issue related to the TSV size. TSVs also contribute to routing congestion of each layer since they both interfere with cell placement and, in case of via-last TSVs, become a routing obstacle.

As the CMOS technology scales down, semi-global and global wires are becoming an increasingly important performance bottleneck since their typical wirelength does not scale [2]. Generally, the wire's parasitics define its performances, and both the resistance and the capacitance of a wire directly depend on its length l, as discussed in Chapter 1. Consequently the RC delay is proportional to l^2, which becomes unacceptably great for long wires. Moreover, the switching of the interconnecting wires' capacitance causes dynamic power consumption following the relationship:

$$P = A_f C V^2 f. \tag{5.1}$$

where f is the frequency of digital signal, A_f is wire activity factor, V stands for voltage swing between the two digital levels, and C is the total interconnect capacitance of a certain wire

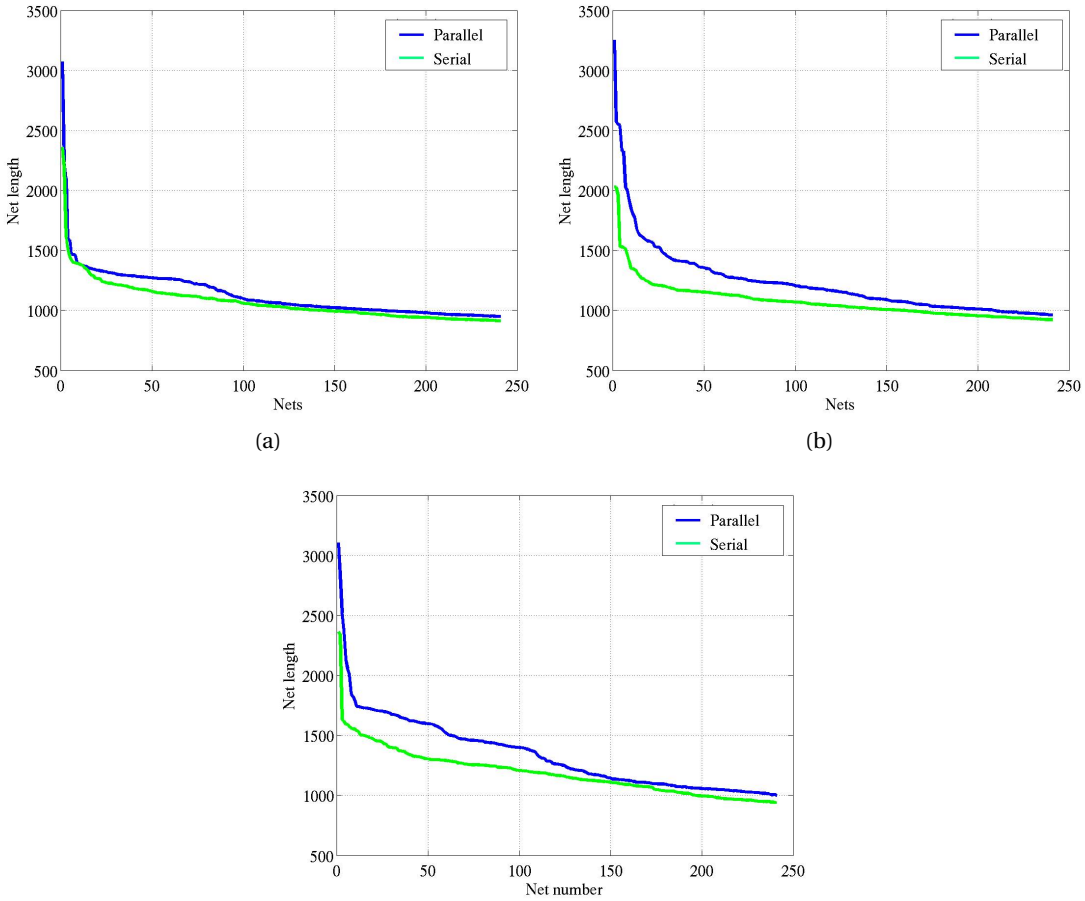


Figure 5.7: Length trend of the longest 240 nets in the design for (a) $5\mu m$ (b) $10\mu m$ (c) $40\mu m$ TSVs for the parallel (blue) and serial (green) configuration.

length [2]. Since dynamic power is currently the main component of the power dissipation, with approximately 50% of microprocessor power consumed by the interconnects [93], the designers should struggle to keep the routing congestion of a chip under control.

In this section we show how the proposed serial approach can reduce routing congestion improving the design performance using the 3D-MMC architecture as test vehicle and via-last TSVs for the inter-layer connections. As an example, we can consider the net depicted as a red line in Figure 5.5 which connects CORE 1 to the NoC. A lower TSV count translates into a significant reduction of routing obstacles in the design, and allows the logic gates to be placed closer to each other. Hence in the serial configuration we can notice that the length of the net connecting CORE 1 to the NoC can be drastically reduced.

The following analysis has been performed on the routing of each placed and routed design depicted in Figure 5.6. First we extract the length of each net in the designs focusing on the

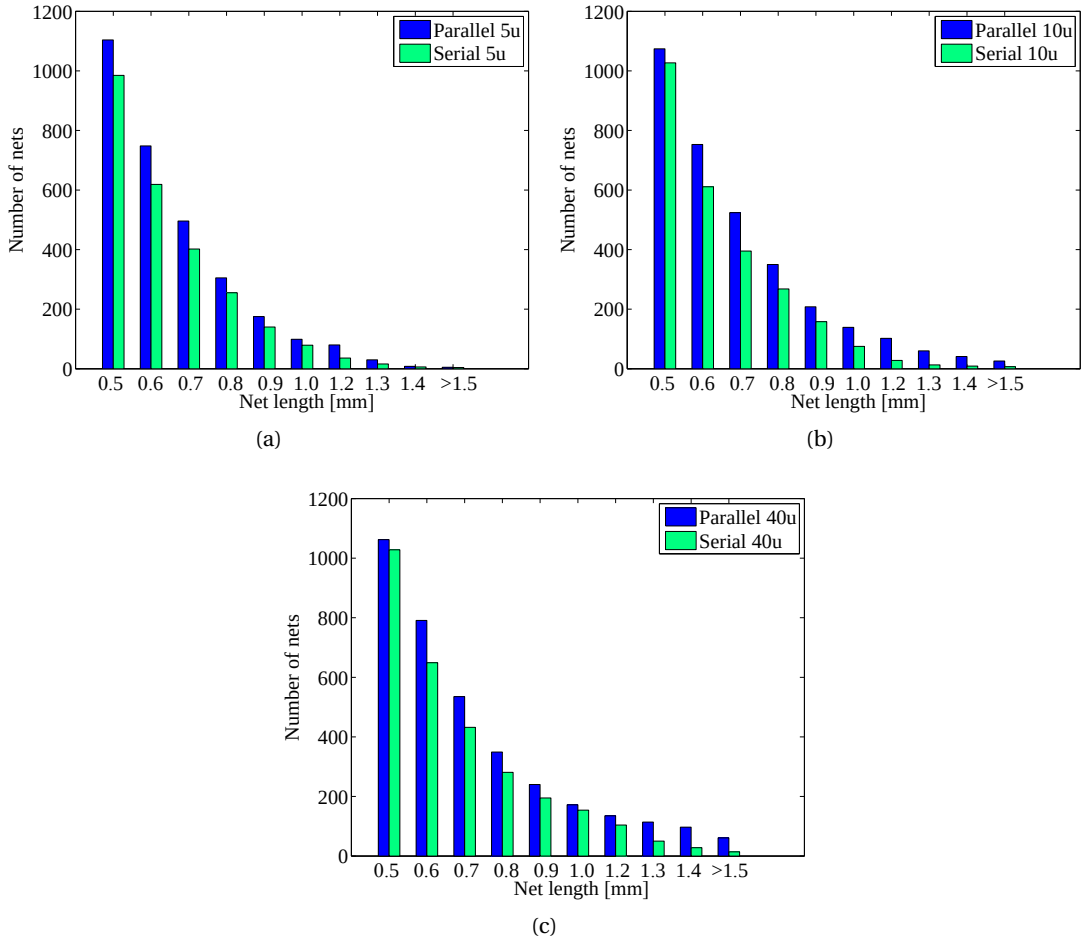


Figure 5.8: Net statistics.

first 240 longest connections. Figure 5.7 shows the trend of the considered nets for both the parallel (in blue) and the serial configuration (in green). The trends clearly show that the serialization causes a reduction of the wirelength, which is limited for the design utilizing the small $5\mu m$ TSVs, while becomes more marked for the designs featuring larger TSVs.

Focusing more on the length of the interconnects, we first define lower bound to focus on the nets longer than that threshold. A typical $500\mu m$ long on-chip copper connection in 90nm CMOS technology at the intermediate level, metal2-metal6, is characterized by a resistance R~80mΩ and a capacitance that exceeds 70fF. Consequently, the net delay can be approximated as $\frac{1}{2}RCl^2$=717ps. The plots in Figure 5.8 depict the histograms of the design's interconnects starting from the $500\mu m$ threshold. In particular, it shows the number of nets in the design for each range of lengths. We can notice that for all the TSV technologies considered the number of long nets decreases in the case of the serial configuration (green bars). As expected, there are few nets longer than $1300\mu m$ in the design with $5\mu m$ TSVs, while they are more numerous as the TSV size increases.

The average length reduction for the considered designs are summarized in Table 5.2. For the design using $5\mu m$ TSVs, the reduced number of TSVs needed after serializing leads to a 5.3% wirelength improvement. The benefits are more pronounced in case of larger TSVs: for $40\mu m$ and $10\mu m$ TSVs, the wirelength improvement reaches respectively 11.2% and 12.4%.

TSV size	Average length reduction	Wirelength improvement
$5\mu m$	$60\mu m$	5.3%
$10\mu m$	$152\mu m$	12.4%
$40\mu m$	$150\mu m$	11.2%

Table 5.3: Routing results.

5.6 Summary

This chapter explores the impact of data serialization for inter-layer communication on the chip's routing congestion. Adopting a serial vertical data communication approach allow for a reduction of the overall number of TSVs, therefore reducing the the average on-chip wirelength.

A modular 3D stacked multi-processor platform, 3D-MMC, consisting of identical dies has been introduced. The 3D-MMC architecture has been implemented using UMC 90nm CMOS technology, and used as test vehicle. An 8-bit serialization of the 3D-MMC's inter-layer signals has been implemented and compared to the fully parallel solution for different TSV technologies. The wiring characteristics of each solution have been extracted from the placed and routed design.

Results show that the serial approach reaches up to 12.4% wirelength improvement compared to the fully parallel counterpart when using $10\mu m$ TSVs. Even for high end TSV technologies such as $5\mu m$ TSVs, the wirelength undergoes an average reduction of 5.3%.

6 MIRACLE: a 3D Multi-core Processor Test Chip

In chapter 5 we explored the benefits of serial vertical interconnections for 3D systems. A homogeneous multi-core architecture, 3D-MMC, has been designed and used to compare the parallel and serial 3D communication solutions.

In this chapter, we aim to demonstrate the efficiency and applicability of a 3D serial link implemented in a test vehicle. A vertically stacked multi-processor platform based on the 3D-MMC architecture has been designed and fabricated using conventional UMC 90nm CMOS technology. The fabricated 2D dies of the 3D multi-processor prototype have been tested and the KGD has been vertically stacked using an in-house via-last TSV process. Initial results show that the proposed multi-core system is capable of operating at a working frequency of 400MHz, supporting a vertical data bandwidth of 3.2Gb/s.

6.1 Problem formulation

Latest embedded application implementations [94] demonstrate a very high degree of achievable parallelization, thus providing near-optimal "linear" speed-up as specified by Amdahl's Law [95]. Parallel computing is also the most potent alternative to traditional frequency scaling techniques used extensively throughout the past decades. Additional effort is exerted to reduce the inherent level of complexity involving the parallelization of applications, improve the standardization of the flow and reduce the immense fallibility of the parallelization process [96].

3D stacked chip multi-processors (3D-CMPs) are expected to increase the overall core count, while improving core-to-core communication [97]. Hence, 3D benefits can fit the requirements of the embedded processor market. Previous architectural proposals for *3D-CMPs* focus either on stacking memory layers on top of core logic to boost memory bandwidth, or on augmenting the capabilities of planar CMPs including additional logic layers. Nevertheless, 3D-ICs has still numerous challenges to face to become commercially attractive.

This chapter presents a fabricated test vehicle, *MIRACLE*, based on the 3D-MMC architecture

presented in Chapter 5. Due to the modular approach, the 3D-MMC platform has several advantages. First, it dramatically simplifies the chip design process and reduces *Non Recurring Engineering* (NRE) costs, creating a portfolio of architectures with the same mask set. For instance, the stand-alone die (2D-CMP) can be used directly as a final product or integrated on top of identical chips with no additional design effort, thereby creating a high performance version of the same device. Second, homogeneity of the system allows using the same testing protocol for each die within the stack, leading to pre-bond testability without any additional effort for test engineers. A similar approach has already been adopted for 3D-DRAM, where identical memory chips are stacked to increase the overall memory capacity [98].

The multi-processor system has been designed and fabricated using a conventional UMC 90nm CMOS foundry process and vertically stacked. Fully functional microprocessor samples have been post-processed and stacked using an in-house Via-Last Cu TSV process fully developed in the EPFL center of micro-nano fabrication and specifically tailored to the microprocessor platform.

A comprehensive study of the system is presented together with a software approach to optimize the applications execution time. Results show that the proposed 3D *chip multiprocessor (CMP)* is capable of operating at a target frequency of 400 MHz, supporting a vertical data bandwidth of 3.2 Gb/s while limiting the number of TSVs. The fabricated prototype also servers as an example to show the feasibility of post-processing and vertically stacking identical CMOS chips fabricated with a standard foundry process. Since no intervention is required in the FEOL/BEOL CMOS fabrication steps, this approach can be viewed as an advanced packaging technology well suited for low energy serial data transmission between local chips.

6.2 State of the art

Although a number of experimental processes have been proposed for TSV fabrication to construct multiple stacked layers [99], [97], just few industrial examples of memory chips on top of a processor have been demonstrated so far, such as the 3D-processor system by Tezzaron [8], integrating an Intel 8051-based processing layer with an SRAM layer. Despite significant latency and bandwidth improvements that are expected [100], the heterogeneous approach requires additional design effort and costs for the realization of different layers to be stacked in the 3D system. Moreover, in previous proposals, it is extremely challenging to test all the layers before the bonding process, causing a noticeable decrease of the final yield. Recently, a solution is represented by the 3D-CMP proposed by Healy et al. [101], where dummy pads on non-accessible layers are employed for the pre-bonding verification and then buried inside the stacked structure [102].

The logic-on-logic case involves splitting a planar design's logic area into two or more layers, such as the 3D version of an Intel Pentium 4 family processor in Garrou et al.'s work **[103]**.

A processor architecture where a baseline micro-architecture can be augmented by vertically

stacking additional blocks (e.g, more caches, reservation station and so on) to target different market segments has been proposed by Loh in [51]. The novelty of this work relies on the capability of stacking multiple samples of the same design realizing a fully modular, testable and highly reusable 3D-CMP platform with a fairly limited design effort and reduced mask costs.

6.3 MIRACLE

In order to explore the potential of the serial interconnection solution, a test chip based on the 3D-MMC architecture has been designed, fabricated and tested. The test chip, called *MIRACLE*, has been built by stacking identical layers, as depicted in Figure 6.1

The 2D layer architecture is based on the 3D-MMC architecture proposed in Chapter 5 which

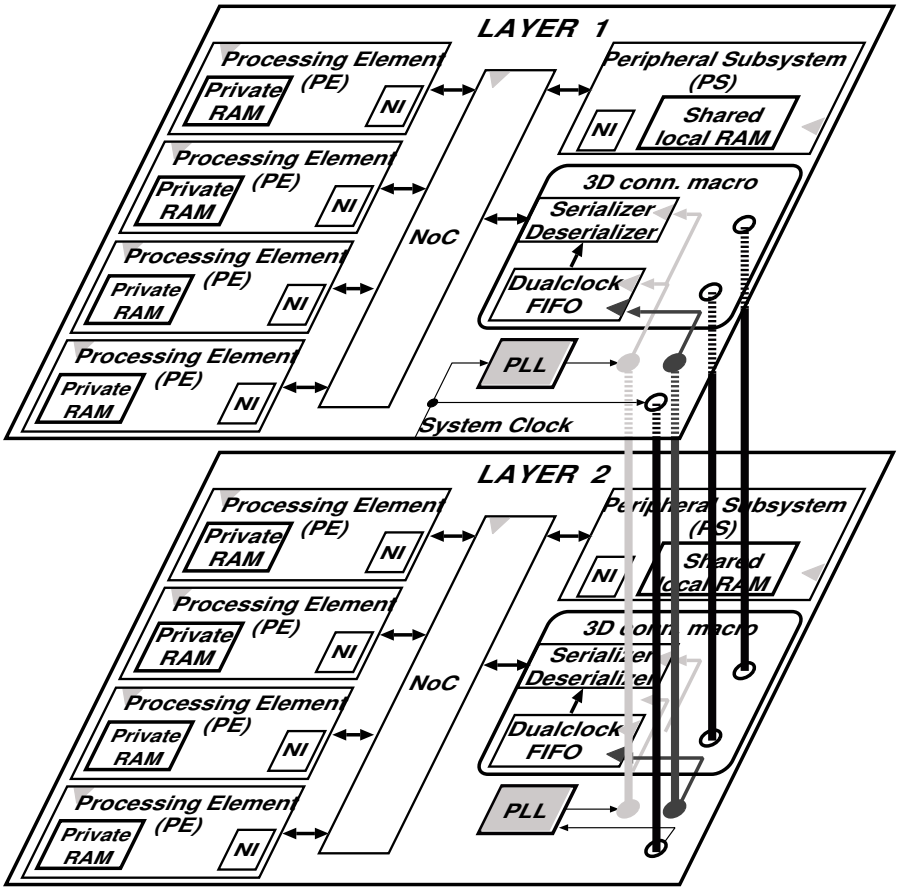


Figure 6.1: Proposed architecture for the 3D-CMP in a 2-layer configuration: Four identical Processing Elements (PE) and a Peripheral Subsystem (PS) are placed in each layer. A 3D connection macro with TSVs is responsible of inter-layer communication. Note that only main building blocks and relevant TSVs are shown in the diagram, the data TSVs are omitted for clarity.

is composed by multiple *Processing Elements* (PEs) communicating through a shared memory implemented in a *Peripheral Subsystem* (PS) unit. The interaction between cores occurs through a specific source-routed NoC.

The intra-layer communication in MIRACLE is achieved through the introduction of a 3D connection macro, exploiting arrays of TSVs as serial vertical data bus. Apart from the SERDES module, additional circuitry is introduced in the macro in order to guarantee the correct functionality of the test vehicle. Each stacked layer has an independent clock domain, provided with a PLL module to regenerate the transmitted clock signal. Data synchronization at the interface of the 3D structure is guaranteed by a Dual-Clock FIFO.

6.3.1 Homogeneous and modular approach

The homogeneous 3D integration, obtained by stacking completely identical dies, results in a cost-effective final structure. In fact, the development of a single layer guarantees a reduction in design time and fabrication costs, involving only one set of lithographic masks [97]. A traditional 2D-IC design flow is employed to design both the multi-processor units and the 3D connection macro. At this point, it is important to note that each layer is a stand-alone MPSoC IC and can function as a fully testable and operational 2D-CMP, as shown in Figure 6.2(a). Once the target 3D structure is decided, *Known-Good-Dies* (KGDs) are post-processed. The dies are stacked on top of each other and the Via-last TSVs are fabricated for inter-layer communication, leading to a homogeneous 3D-CMP structure of identical layers, as shown in Figure 6.2(b). With this approach, the overall number of cores in the system is increased, speeding up the parallel workload of the processors and improving the CMP performance. Nevertheless, the proposed design strategy is not limited to homogeneous systems. Different dies can also be integrated in the stacked platform, as long as they share the 3D connection macro. Figure 6.2(c) depicts a possible configuration where a memory layer is placed on top of two CMP dies. The proposed design approach leads to a reusable platform with high cost-effectiveness: it can target various market segments simply by selecting the appropriate number of layers to be stacked in the system. An important contribution of this work lies in the possibility to configure the stacked dies number after the fabrication: a unique identification signal (LayerID) is provided to each die in order to distinguish identical layers. The LayerID is automatically assigned at the system power-up, after all chips are fully processed and assembled. Nevertheless, this post-fabrication configurability of the number of layers eventually leads to the need of over-constraining the power I/Os. Before the design phase, a maximum number of layers that can possibly be stacked should be defined in the specs. The designer should then over-constrain in order to guarantee the correct functionality in the case of a system with the maximum number of layers, to avoid PEs starving for power.

The proposed architecture design and the homogeneous stacking solution, as the one depicted in Figure 6.2(b), offer the advantage of increasing the overall yield of the assembled structure. Even though the manufacturing yield of small footprint chips can be high, one faulty die can

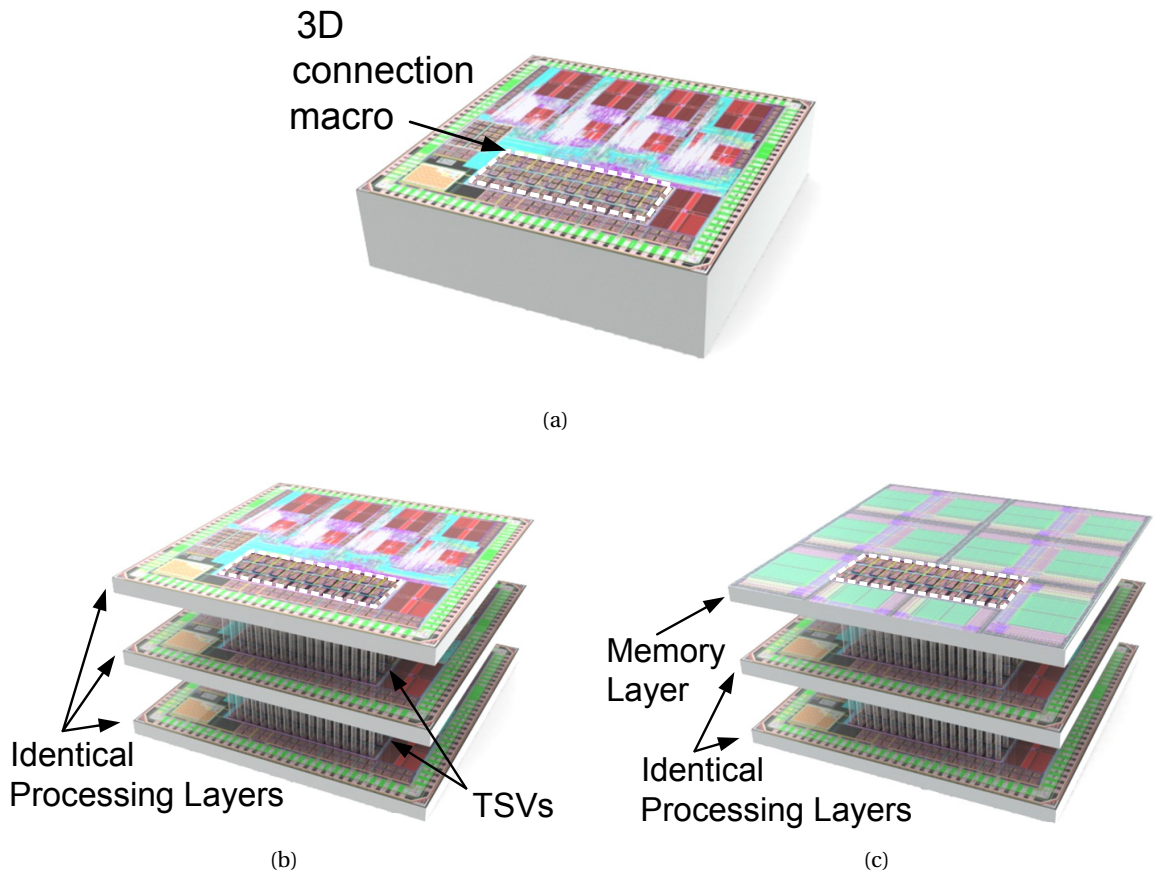


Figure 6.2: Modular re-usability of 3D MMC: (a) Single die used as stand-alone 2D-CMP. (b) Homogeneous stacking for high performance 3D-CMP (c) Heterogeneous stacking for 3D-CMP, integrating additional layers (e.g. a memory die) that shares the same 3D connection macro.

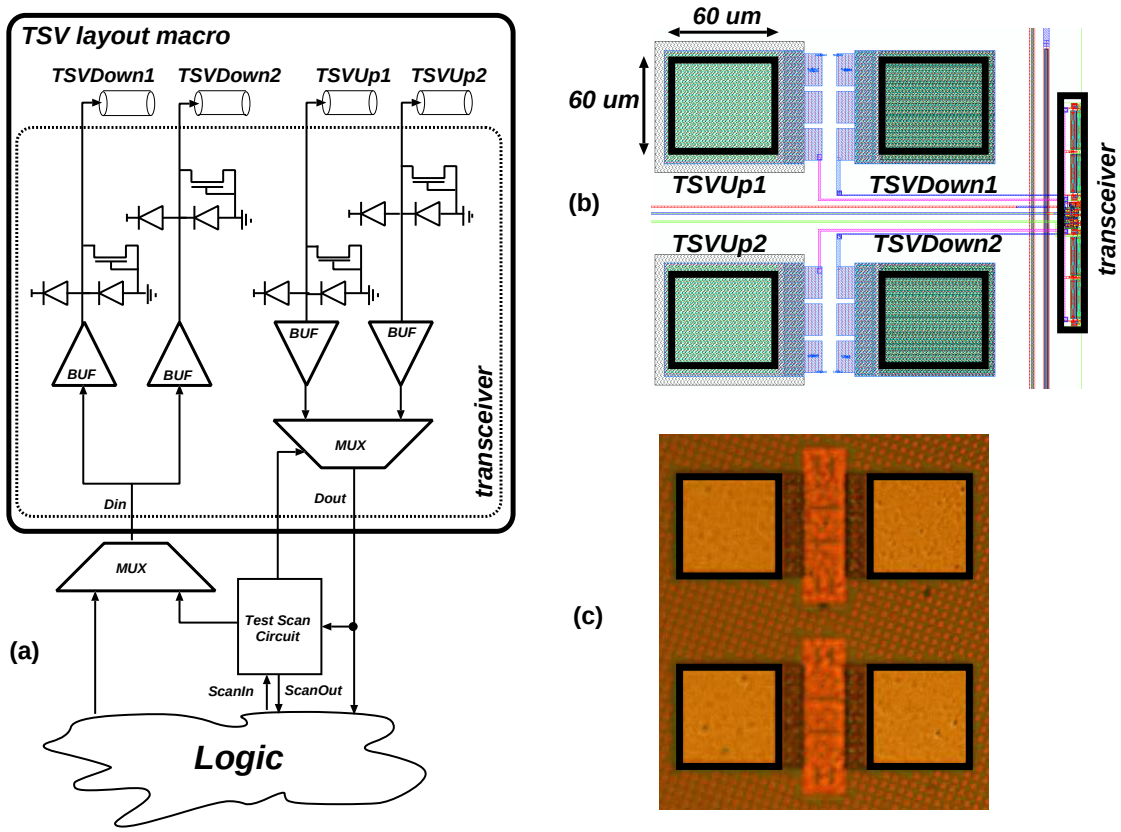


Figure 6.3: (a) Circuit schematic of the TSV macro. (b) Layout of the TSV macro, highlighting the main blocks from the corresponding circuit schematic. The effective TSV pad area is put in evidence. (c) Optical microscope image of the TSV macro on the fabricated test vehicle.

completely ruin the behavior of the entire 3D system. Identical die integration allows each stand-alone chip to be fully testable, leading to a higher assembly yield for the final 3D-CMP structure that is by definition built out of KGDs [104].

6.4 3D specific macro architecture and circuit design

6.4.1 TSV redundancy and yield collection

3D integration is still a topic of active research. In particular, the non-mature TSV fabrication processes available up to now cannot guarantee, to our best knowledge, the desired yield for the final system. In order to tackle the yield related issues, a redundancy policy has been adopted for the data transmission between layers to ensure reliable communication. Each vertical signal is simultaneously forwarded to the neighbour layer by means of two TSVs, reducing the probability of failure.

Moreover, being able to collect statistics on TSV yield on the final stacked system is desirable.

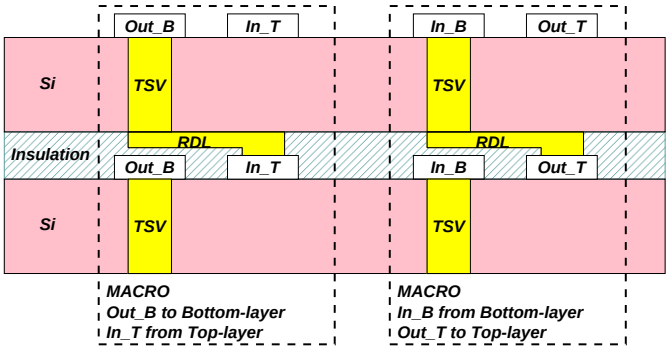


Figure 6.4: TSV macro cross-section, highlighting the use of multiple pads and redistribution layer (RDL).

For this purpose, a *Built-In-Self-Test* (BIST) engine has been implemented. The test is performed at boot time, each TSV is individually tested thanks to a multiplexer inserted in the design to select between the two redundant TSVs. The test pattern is injected through a scan chain; the reader can refer to [33] for a more detailed description of the scan chain design. The statistics are then stored in specific user-visible registers. Additionally, the aforementioned registers provide the selection value to the multiplexers that control, through the redundancy policy, which TSV is used in order to ensure reliable operation of the system.

The final TSV-macro is depicted in Figure 6.3, presenting the schematic circuit, the layout and a real optical image on a fabricated test chip. For each TSV, two adjacent pads are used, one connected to the TSV of the upper layer and the second one connected to the TSV to the bottom layer. A redistribution layer is used, as shown in the cross section of the TSV macro in Figure 6.4. The dimensions of the pad hosting the 40μm TSV are (60μmx60μm), in order to avoid alignment problems. The transceiver includes individual ESD protection, a buffer for signal integrity and a weak pull down for each TSV. For data transmission, two different TSV-macros are needed depending on signal direction: a first macro receives the signals coming from the bottom layer and transmits to the top one; the second macro receives the signals from the top layer and transmits to the bottom one. There are three main critical signals in the design that need their integrity to be guaranteed. Hence, additional safety has been incorporated for clock, reset and LayerID signals: they are transmitted over three parallel TSVs and continuously checked during runtime. A glitch-free majority voter is implemented inside each die to ensure the validity of the transmitted bits. A continuous and implicit auto-check of the TSV connection is achieved at the very minor cost of additional area and negligible combinatorial delay, ensuring the correct transmission of the critical signals.

Since only the top-most layer has access to wire-bonded I/O and power pads, the power supply (VDD and GND) of all of the bottom layers should also be provided by dedicated TSV. These supply TSVs have a simpler structure and do not incorporate self-check features. A detailed classification of the different purpose TSVs used in this first prototype is presented in Table 6.1, specifying case by case the presence of redundancy with the asterisk.

Table 6.1: TSVs features summary

Overall number of TSVs	120	(95*)
Total Power TSVs	54	
Total Signal TSVs	66	(31*)
• Total NoC TSVs	44	(22*)
• Total JTAG TSVs	10	(5*)
• Total Control Signal TSVs	12	(4*)
TSV diameter	40	µm
TSV depth	50	µm
TSV capacitance	1	pF
TSV resistance	0.7	Ω
* TSV number without redundancy.		

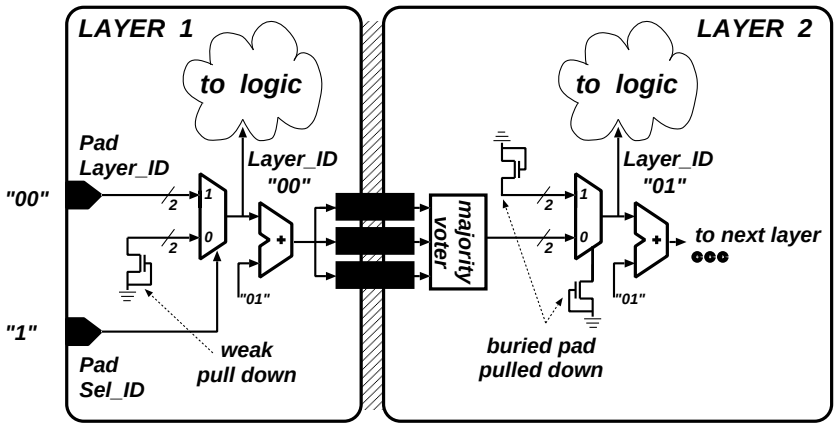


Figure 6.5: LayerID generation and propagation between two stacked layers using three redundant TSVs for the signal interface. Schematic of the configured circuit in each layer is shown, unrelated logic is not depicted.

6.4.2 Layer identification

Once the identical layers are stacked to form a 3D-IC, they need to operate as a complete system without the need for any further modification or action. Hence, it is necessary to embed specific modules enabling an effective auto-configuration of the layers. For this purpose, a dedicated control signal that provides a different n-bit digital word for each layer has been added, namely the layer identification number (LayerID). Depending on the value of the layerID, each stacked die(layer) knows its position and role in the system and auto-configure itself.

The layerID generation circuit already configured in each layer is depicted in Figure 6.5 for a case study of a two-layer system. The starting sequence ("00") is injected through the pads of the top die and is selected by a multiplexer to become the identification value for that layer. The value is also forwarded to a half adder, defining the label for the bottom layer ("01"), to which

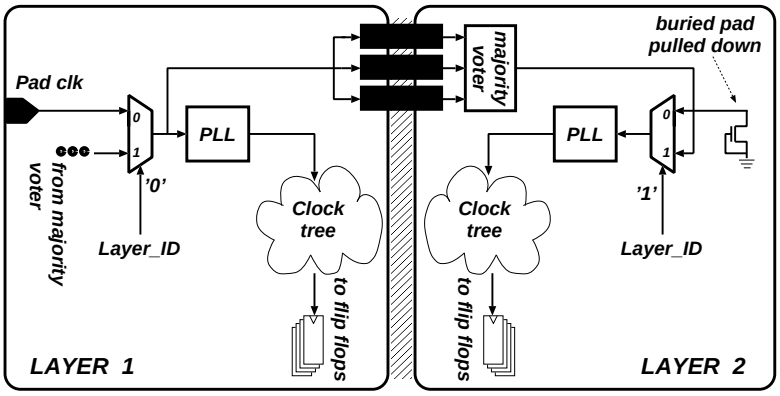


Figure 6.6: Clock distribution and propagation between two stacked layers using three redundant TSVs for the signal interface. Schematic of the circuit configured through the LayerID is shown, unrelated logic is not depicted.

it is transmitted through three redundant and parallel TSVs. Buried inside the 3D structure, the pads of the bottom tier are considered inaccessible after stacking; hence they are designed to be internally and automatically pulled-down when no signal is applied to them. As a result, the multiplexer on the second layer is forced to select the LayerID transmitted by the TSV, enabling the identification of the layer's vertical position and the relative self-configuration of the second tier circuitry.

6.4.3 Clocking scheme and data transmission

The clock distribution in 3D-ICs is a complex and challenging task, as presented by Pavlidis et al in [60]. The synchronization of sequential elements located on multiple planes by the same clock signal underlines the importance of controlling the clock skew. With the proposed architecture, each layer has its own clock generation and distribution using a layer-dedicated PLL, hence the obtained clock tree presents minimum skew, high robustness and large tolerance to any timing variation. As depicted in Figure 6.6, the clock is injected onto a pad of the top layer, after passing through a PLL module, it is both distributed in the circuit and sent to the three redundant TSVs that propagate it to the next layer. The bottom layer receives the clock from TSVs with triple redundancy which, thanks to the layerID, are selected to enter the PLL module to re-generate the clock signal for maintaining its integrity. This multi-PLL approach results in all layers operating at the same frequency, but being asynchronous from each other due to the unknown phase shift among the clocks.

Transferring signals among different clock domains requires the data to be re-synchronized. For this purpose, data signals are transmitted among layers together with their clock, used by a Dual Clock FIFO to re-synchronize them to the layer clock domain. With this approach the problem of the skew control is intrinsically reduced to a 2D clock tree synthesis.

In order to reduce the silicon area occupied by the TSVs, data signals are serialized before

the transmission through TSVs and successively de-serialized at the receiving layer. The loss in bandwidth due to the serialization can be compensated by increasing the serializer clock frequency, fully exploiting the capability of the sub-micron CMOS processes, as proposed in [105]. However, in the fabricated chip, it has been decided not to implement multi-clock domains, because of the area limitation. Each 32-bit data word that has to be transmitted to the neighbouring layer is partitioned into four bytes, and then sent serially through data TSVs that support 8 bits in parallel. The receiving layer reconstructs the original 32-bit word by means of a deserializer after having re-synchronized the data to the layer clock domain. Figure 6.1 shows the path of the data transmitted between the dies of a 2-layer 3D-MMC system.

6.4.4 Physical Design

MIRACLE has been implemented in RTL and synthesized with the UMC 90nm CMOS technology library using Synopsys Design Compiler. The layout has been placed and routed with Cadence Encounter. The functionality has been verified using Mentor Graphics ModelSim. Unfortunately, the current version of Synopsys DC does not support TSVs and 3D stacking, hence, the synthesis flow has to be performed in several steps. Starting from the synthesizable RTL description, an *ad hoc* set of timing constraints is applied to the TSV macro in order to ensure a correct timing budget between layers. The design is synthesized considering the latencies of the stacked dies. Thanks to the modularity of the design, no additional challenges due to the 3D target are added to the back-end design. The single die is placed and routed as a traditional 2D design, following the timing constraints already set up for the synthesis. The TSV macros, designed as full-custom modules with Cadence Virtuoso, are included in the top level design for placement and routing.

6.5 In-house 3D stacking process

Throughout this research, a chip-level 3D integration platform has been developed for KGDs stacking and TSV fabrication [106]. The via-after-bonding (or via-last after BEOL) integration technique [107] is employed for the proposed CMP architecture. Unlike via-first or via-middle techniques [108], where the TSVs are fabricated during the IC fabrication, via-last solution offers the benefit of decoupling the TSV process from the CMOS process, allowing the placement of TSVs after the conventional IC fabrication is completed. Moreover, in the proposed approach, the chips are first thinned and bonded, and then the TSVs are fabricated. Therefore, the technique does not require any metal-metal bonding step, which is essential in all via-first approaches. This reduces the complexity of the fabrication process and eliminates the bonding-related reliability issues. The entire TSV process has been developed and experimentally validated at EPFL Center of MicroNano Technology with test chips emulating real CMOS chips.

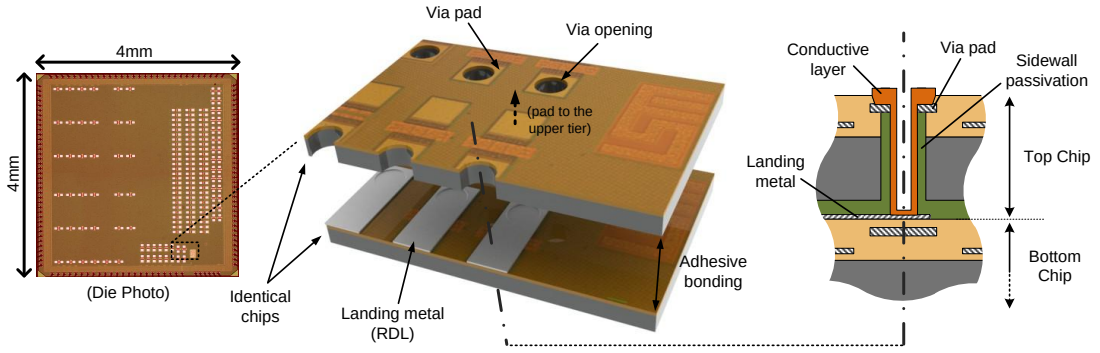


Figure 6.7: Die photo of the multi-processor chip, and the illustration of the chip stacking approach with 40 μm diamater TSVs fabricated on 60 x 60 μm^2 CMOS pads. Since the two chips are identical, the surface of the bottom chip is passivated and RDL is patterned to re-route the signal to the upper tier.

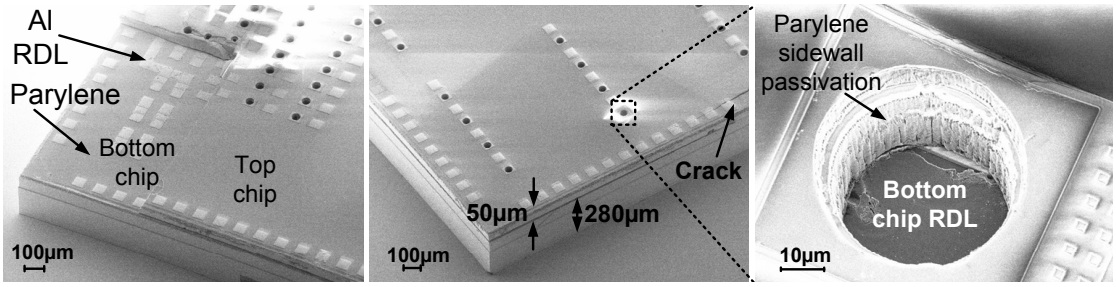


Figure 6.8: SEM photos of the bonded chips and the close-up image of the via opening showing the sidewall parylene passivation and the RDL layer on the bottom chip. (An already broken chip is used as the top chip to inspect the alignment accuracy).

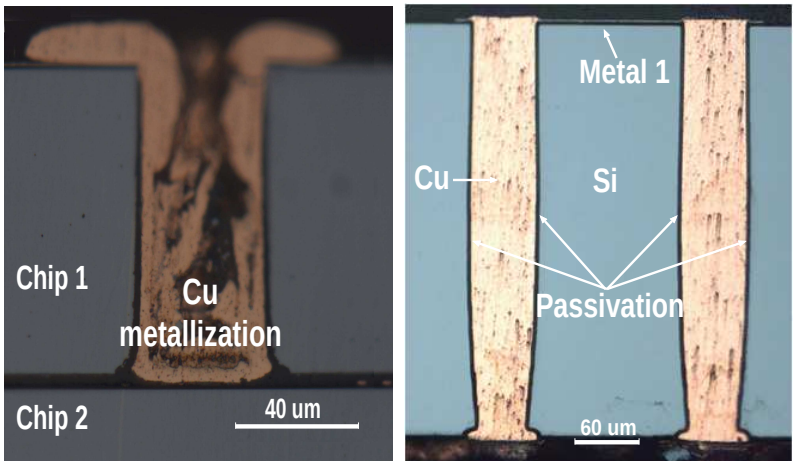


Figure 6.9: Cross section of the in-house developed TSVs: (a) Lined TSV connecting two stacked chips. (b) Fully-filled TSV developed for the characterization tests.

6.5.1 Process steps

Figure 6.7 shows the illustration of the proposed integration approach. The diced chips are tested and KGDs are post-processed for both layers. Figure. 6.10 summarizes the process flow used in the post-CMOS processing and chip-to-chip integration.

The top chip is first thinned down to 50 μm by grinding. Then, the chips are placed on a carrier wafer and the dry-film resist is laminated and patterned. For the top chip; Al pad, dielectric layers and the Si substrate are etched all the way through to the chip backside for the TSV openings with 40 μm-diameter. Alternatively, the order of the grinding and etching steps can be swapped; thus, a blind via is first formed inside the substrate with full thickness, then the chip is grinded till the via opening is reached. For the bottom chip post-processing, first a dielectric layer is deposited and patterned, then the *redistribution layer (RDL)* is fabricated. Since the chips in the stack are identical, RDL is used to re-route the signal to the upper tier. Then, the two chips are aligned and bonded by adhesive bonding with parylene as the intermediate layer. Figure 6.8 shows the SEM images after C2C bonding and parylene etching steps. Bonding is performed with a low temperature budget of below 200°C, to ensure no drift or change on the transistors characteristics. Finally, Cu-TSVs are fabricated by side-wall passivation and Cu electroplating. The electrical connection is realized between the Al pad of the top chip and RDL of the bottom chip. If required, these steps can be repeated for a multilayer stack by using the already-bonded chips as the bottom chip. Figure 6.9 shows the cross-sections of the lined **[106]** and fully-filled **[109]** TSVs developed for the preliminary characterization and verification tests.

Compared to the blind-via fabrication techniques where the TSVs are drilled from the backside till the landing metal, the through-via approach proposed in this thesis is much simpler since it eliminates several fabrication steps, such as metal-metal bonding and passivation layer

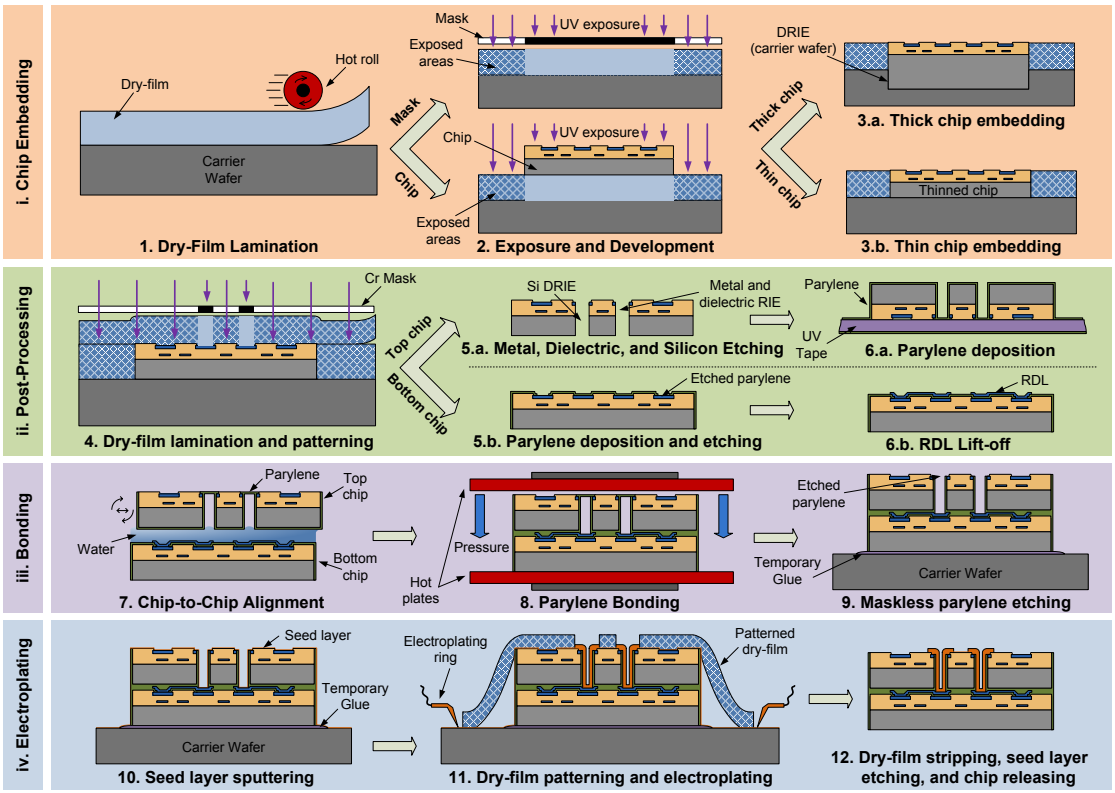


Figure 6.10: Process flow for post-CMOS processing and chip-to-chip integration [23].

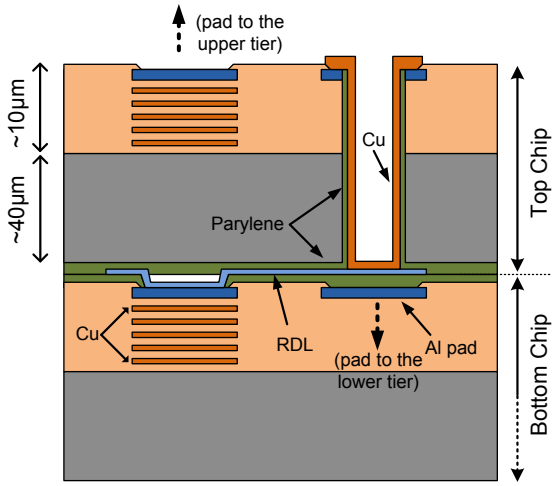


Figure 6.11: Illustration of the TSV macro, showing the parylene sidewall passivation and Cu metallization connecting RDL to the Al pad on the top chip (not drawn to scale). Each TSV macro is composed of two adjacent pads; one for routing the signal to the upper tier through the RDL, the other for the lower tier through the TSV [23].

patterning. Moreover, front-side photolithography allows higher alignment accuracy as the etching mask can easily be aligned to the patterns on the CMOS chip. On the other hand, the main drawback is that the front-side vias block the BEOL layers; thus, the signal routing on top of the TSVs is not possible.

6.6 Thermal evaluation

A significant challenge in 3D stacking is the power density increase per footprint, which may cause temperature to increase beyond reliable thresholds. This section provides the thermal analysis of 3D-MMC and demonstrates the thermal feasibility of the proposed architecture.

Power consumption of each component in a layer of the 3D stack is estimated via statistical power analysis using Encounter Power System by Cadence. We assume a switching rate of 50% for each flip flop and each input port, and we use a 100% toggling rate for the clock. The tool automatically propagates the activity through internal nodes and estimates the power consumption based on the average toggling rate of each gate. Table 6.2 provides the power consumption of all the components at 400MHz. All values include leakage power. Core power in the table includes the logic, I-Cache, ROM, and all other sub-blocks of the core except for the local RAM. Table 6.2 highlights the low power consumption of 3D-MMC, where each layer consumes 267mW.

We use HotSpot version 5.02 [110] for thermal simulations. The package and die parameters used in the simulation are provided in Table 6.2. The floorplan of each layer is identical and is shown in Figure 6.12. To take the impact of TSVs into account during thermal evaluation, we

Table 6.2: Power Consumption and Thermal Properties of 3D-MMC

Power Consumption Characteristics	
Components	**Power (mW)**
Core	37.98
Local RAM for each core	17.13
Router	10.07
Data TSV arrays	1.6 (smaller array) to 8.11 (larger array)
Shared memory	22.16
PLL	5
Package and Die Thermal Characteristics	
Die area	3.5mmx3.5mm
Die thickness (bottom layer)	280μm
Die thickness (other layers)	50μm
Die (Si) resistivity	0.01mK/W (meter-Kelvin per Watt)
Glue conductivity	0.082W/mK at 25°C
Glue thickness	2μm

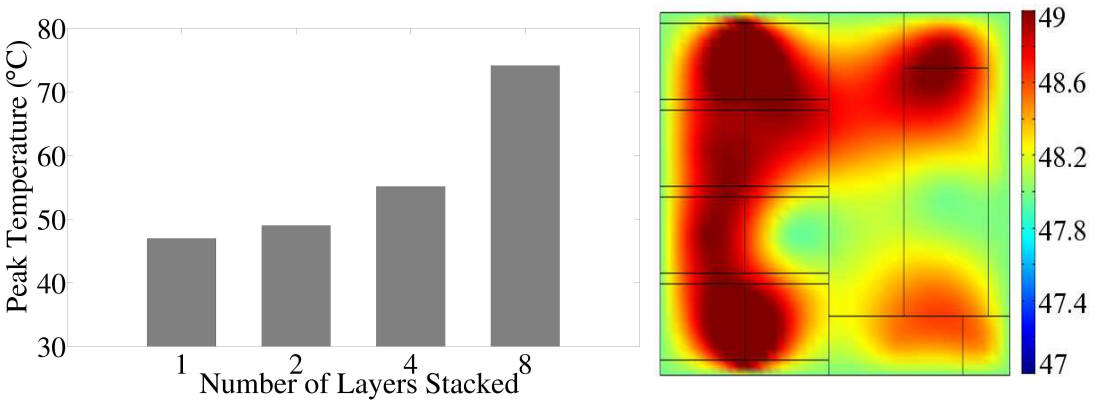


Figure 6.12: The figure demonstrates the peak temperatures at steady state for a single layer as well as 2, 4, and 8-layered stack. On the right, we show the thermal map of the top layer for the 2-layered stack. Thermal variations are similarly low (limited to a few degrees only) for 4 and 8-layered stacks.

use a modified version of HotSpot that enables modelling heterogeneity within a layer [111]. We compute joint thermal resistivity for each TSV block based on the ratio of TSV (Cu) area to overall TSV array area (including all the spacing between the TSVs). We simulate the system without a heat sink by using a very small number for the heat sink thickness in HotSpot. Layers are stacked using a glue (interface material) layer of Parylene-C.

Figure 6.12 provides the steady state peak temperatures for a single layer chip and 3D systems including 2 layers, 4 layers, and 8 layers when all cores are active. In the figure, we also provide a thermal map for the top layer of the 2-layered system. For 2 and 4-layered systems, even though cores are overlapped on top of each other and all cores are active, we do not observe high temperatures. For the 8-layered stack, peak temperature reaches 74°C, which is still below the typical 85° thermal thresholds used in most processor chips. As we focus on a 2-layered stack in this paper, we do not apply thermal management strategies.

6.7 Design verification

Post-layout simulations in ModelSim verify the correct functionality of the system and extract its performance parameters: each layer has been synthesized with a target operative frequency of 400 MHz, which results in a vertical data bandwidth of 3.2 Gbps.

Particular attention has been dedicated to the verification of the system behaviour at the interface between the two stacked layers. An emblematic case is a core's read/write request to the shared memory of the next layer. Waveforms showing a memory write operation (followed by a read verification) are presented in Figure 6.16. JTAG signals are injected requesting core 0 of the top layer to write a 32 bit word, "0XABBAABB0", in the shared memory of the bottom layer (Figure 6.16(1)). Core 0 delivers the request to the NoC (Figure 6.16(2)), which

Table 6.3: Summary of the tested functionalities

Procedure	Validation
1.	Read own IDcod
2.	Read ROM content
	Read/Write from private RAM
	Read/Write from shared RAM
3.	Binary download & execution
4	TAP controller bypass
	Scan chain of multiple cores
5.	Test of Top and Bottom layer together

encapsulates it in a frame and forwards it to the 3D interface (Figure 6.16(3)). A serializer divides the encapsulated data in 4 bytes sending them one by one through the TSVs together with a valid signal and the layer clock (Figure 6.16(4)).

In the bottom layer, the received data signals are re-synchronized to the local clock domain through a Dual Clock FIFO (Figure 6.16(5)). Then they are de-serialized and sent to the shared memory (Figure 6.16(7)). Further reading of the same location verifies the correctness of the previous operation (Figure 6.16(8)), sending out to the JTAG TDO the data packet (Figure 6.16(9)).

6.7.1 FPGA emulation

The basic functional verification is not sufficient to guarantee the correct behaviour of the multi-core structure. Hence, the full 3D system has been emulated on a Xilinx Virtex5 FPGA board in order to observe the system running. A complete testing procedure, shown in Table 6.3, has been developed in order to debug the device by an external source.

The positive results obtained confirm the correctness of the 3D-MMC design, proving the capabilities of interaction among cores located in different layers. In particular, the core's read/write request to the shared memory of the next layer, described in the previous section, has been repeated on the FPGA model, demonstrating the optimal intra-layer communication. Moreover, this procedure is able to verify the correct behavior of the NoC during the packet routing; both the NIs and the Switches present a two clock cycles latency. The FPGA emulation enables also the verification of the auto-configuration of the different layers according to their identification signal. The behavior of the self-verification strategy applied to the redundant TSV is validated emulating possible faults causing opens on the TSVs (more frequent problem in TSV technology process).

Table 6.4: Architecture details of 3D test vehicle

Process technology	90nm CMOS
Number of layers	2
Die size	4x4 mm
Core footprint	800x1650 μm
Number of cores	8
Total on-chip memory	320 KB
Max operative frequency	400 MHz
Vertical data bandwidth	3.2 Gbps
Number of I/O pads	120

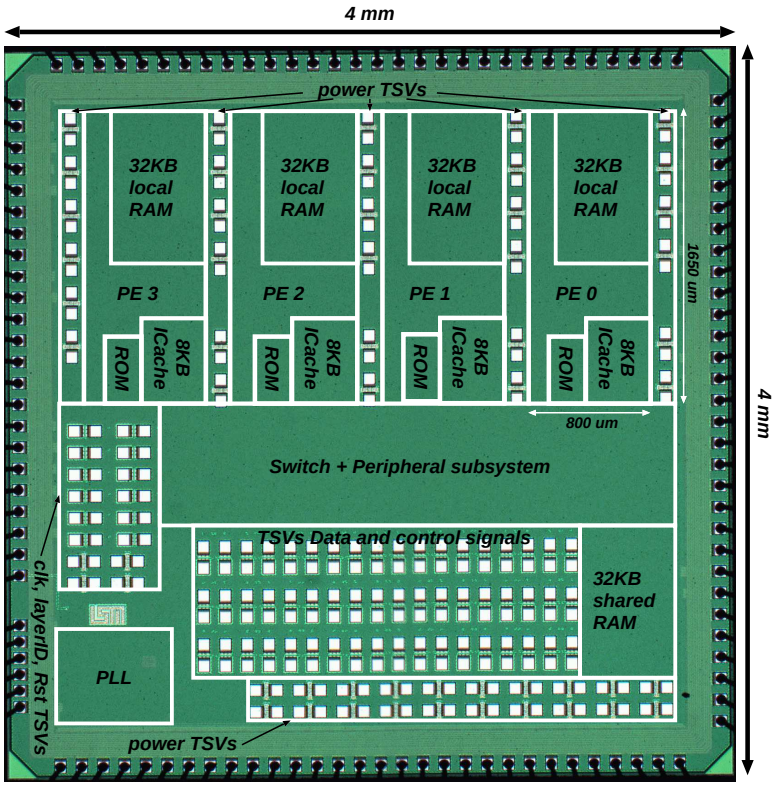


Figure 6.13: Single die microphotograph. Pads for signal and power TSVs are visible before the post-processing. Main blocks are identified in the image (PE, PS, Switch of the NoC, PLL). Size of the full-chip and of the PE footprint are also shown.

6.8 Prototype verification

2D-CMPs of the MIRACLE architecture have been realized using a standard UMC 90nm CMOS technology. The functional blocks of the test chip are identified in the die micro-graph in Figure 6.13. Table 6.4 lists the main specifications of a 2-layer test prototype.

6.8.1 3D oriented testing policy

The realization of a multi-layer 3D device poses unique challenges in terms of testability. The MIRACLE architecture enables a complete testing strategy, including both pre- and post-bonding validation, allowing the stacking of only the KGDs. Following the fabrication of 2D samples using a conventional CMOS process, each individual die is fully tested as a single entity, by accessing directly the multi-core processor through the designed frame of 120 I/O pads using a probe card assembly. After this initial testing and validation, functional dies that are destined for 3D assembly are further processed to manufacture the TSVs.

A second pre-bonding validation can be performed to screen dies with non functional TSVs. After post-processing, a performance test is performed to verify that TSV fabrication induced stress has not altered electrical properties of transistors located nearby the TSVs. Finally, a post-bond test is performed to validate the stacked system. In particular, TSV yield results are determined using the method described in [33]. In future prototypes, sensing circuitry can be integrated in each TSV macro to verify their stand-alone functionality with a capacitance measurement [112].

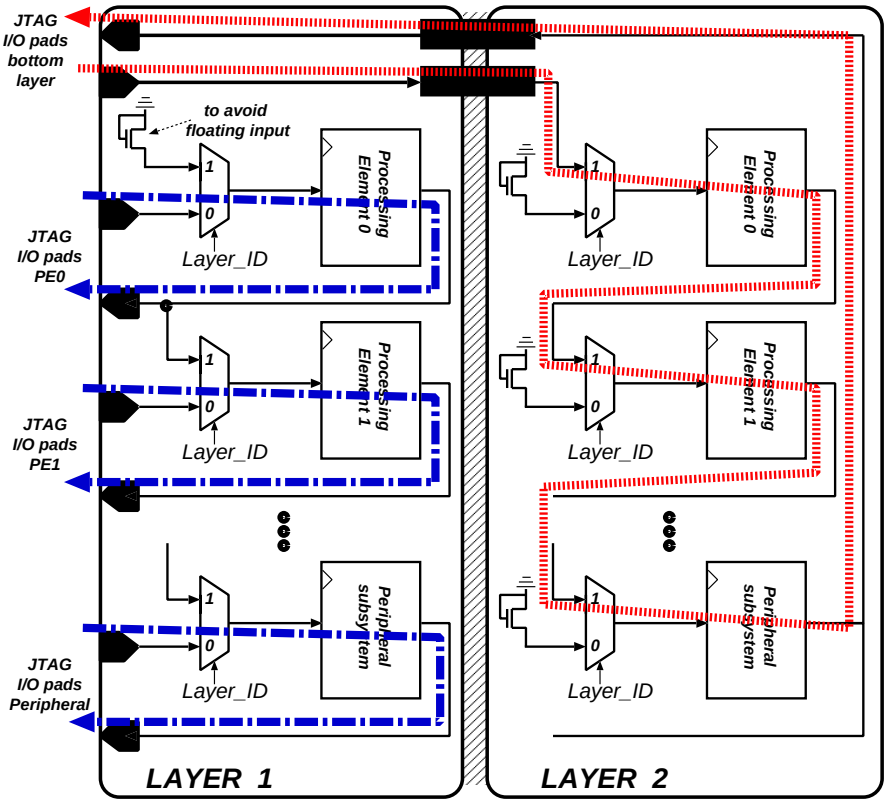


Figure 6.14: Block diagram of the multiplexers interface between two stacked layers. The represented circuit is in charge of the scan-chain configurability allowing pre- and post-bonding testability.

Once the two layers are assembled, I/O pads are no longer accessible in the bottom tier, therefore a custom testing method based on boundary scan chain has been developed. This requires embedding additional modules, enabling the communication through JTAG signals between an external debugger and the processors. In particular, each layer contains a JTAG interface for the management of the debug signals according to layer position in the stacked structure, defined by the LayerID. The interface is shown in Figure 6.14.

For the verification of a stand-alone die, the set of multiplexers is forced to assign the JTAG external signals from the I/O pads to JTAG ports dedicated to each PE. With this approach, each core is accessed in parallel. The top layer of the stacked structure exploits the same configuration. In the bottom layer, the pads are buried during the bonding process, hence the LayerID configures the multiplexers interface to receive the JTAG signals from the upper layer pads through the TSVs. Moreover, the cores are automatically arranged in a chain, that can be accessed serially through the single set of JTAG signal. A representative image of the resulting testing procedure for a two-layer configuration is presented in Figure 6.15.

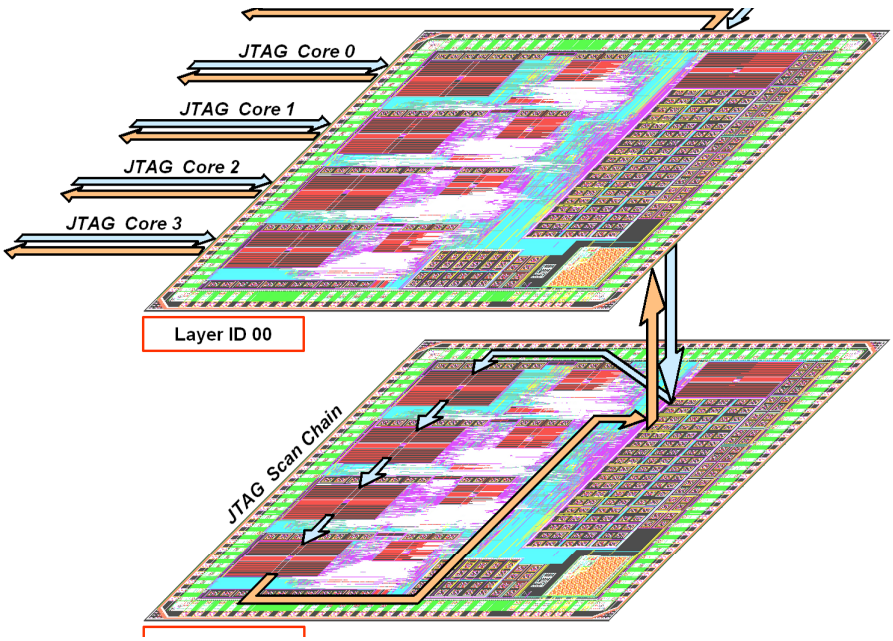


Figure 6.15: Auto-configuration for testability. Top layer cores are accessed in parallel from the pads. The processors on bottom layer are configured in a scan chain for the debug procedure: JTAG inputs are transmitted from top to bottom die, the TDO produced on the bottom layer returns up to the top one.

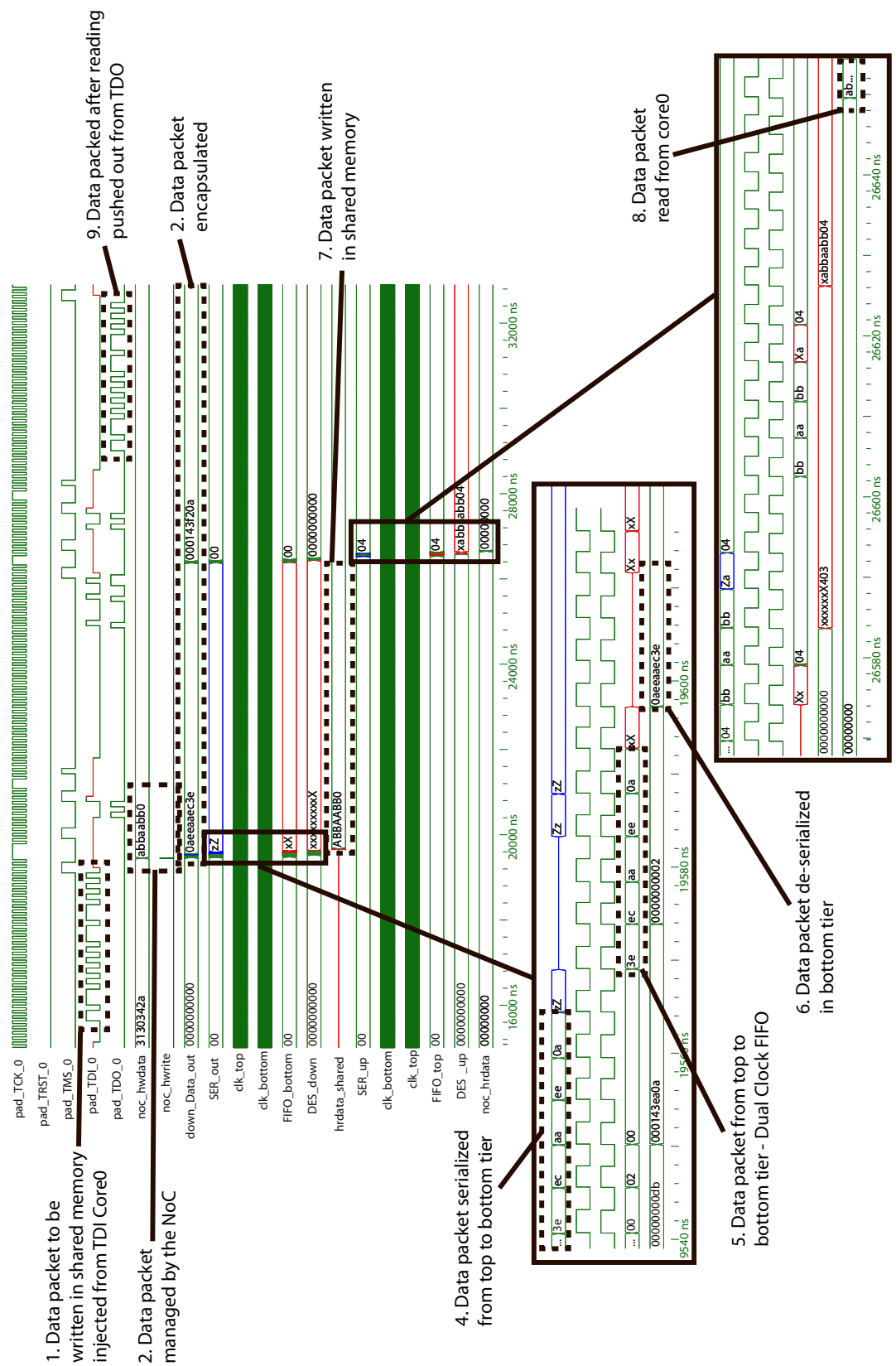


Figure 6.16: Post-layout simulation waveforms describing a inter-layer operation: One core is requesting to write a word on the shared memory of the bottom tier, via JTAG.

6.8.2 Testing setup

The stand-alone and modular philosophy of the architecture provides an increased level of testability of the system. It allows the application of a coherent validation procedure both to each of the individual layers and the stacked structure. Both components and methodology were chosen so that they could be reused between the different phases. The testing protocol, already shown in Table 6.3, has been applied both to packaged and naked single dies through a specifically designed Probe Card setup, attached to a manual Probe Station (Karl Suss PM8).

The testing source consists of a custom software running on a host computer, able to translate the user commands in input bit vectors. An open-source tool, namely OpenOCD, was chosen as the base of the software debugger infrastructure. After extensive adaptation of the original code, it has been possible to apply the testing procedure previously described.

An USB-to-JTAG converter transmits these vectors to an FPGA board that acts as an interface to the prototype dies. In particular, the programmable unit, integrated on the Probe Station, is in charge of selecting to which cores the test is addressed, setting them in an external scan chain. By utilizing such an approach, it is possible to physically link all cores of all layers with a unique path inside the FPGA board, allowing the same methodology to be used for both individual layer and stacked structure verification.

The processed JTAG debug signals are transmitted through a system of *Printed Circuit Boards* (PCB) able to generate all power supplies, clock sources and control signals for the chips. The 120 signals applied to the I/O pins of the device are then transmitted to the Probe Card's needle frame, which is contacting the chip I/O pads.

6.8.3 2D prototype testing

The set of tests applied to the FPGA emulator is then re-applied to the 2D naked dies of the prototype. Valid response sequences has been registered on the ASIC stand alone multi-core layers ensuring the expected behavior of all the specific functionalities of the single dies, prior to 3D stacking.

It has been possible to access and test each single core validating basic and complex behavior of the processors, including reading and writing from the entire addressable memory space. Checking of the boot sequence inside the ROM, reading/writing operation from both private and shared memories have been verified. Particular effort has been invested for downloading routines inside the private RAM memory of each core; with their execution it has been possible to verify the in-layer multi-core interactions through the shared memory. The complete test procedure has been validated in a wide range of frequencies, starting from 1 MHz and reaching the target frequency of 400 MHz.

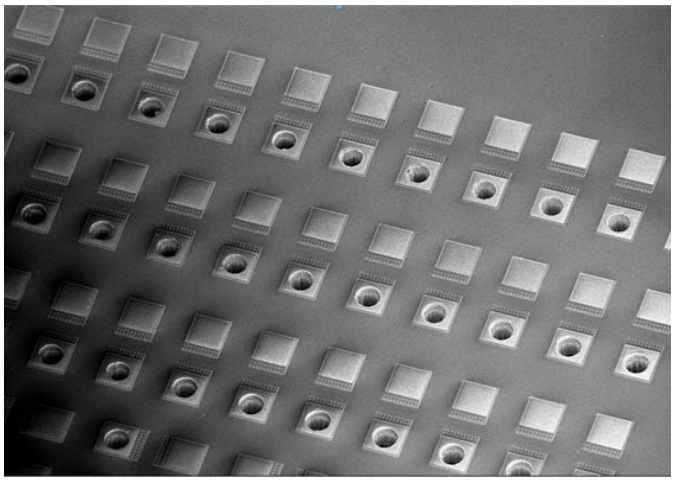


Figure 6.17: SEM image of the multi-core die with post-processed TSV openings.

6.8.4 3D prototype

The proposed testing strategy has led to the identification of KGD, that are then post-processed in the clean-room for the TSV fabrication, exploiting the Via-Last TSVs presented in Section IV. The technology has already been validated and the fabricated TSVs have been shown to be fully functional by connecting them in a daisy chain on a test wafer. A SEM image of the actual multi-core die with fabricated TSV openings is shown in Figure 6.17.

In order to increase the yield of the final stacked structure, the planned methodology includes the repetition of the same testing procedure to the single chips after the TSV etching process, verifying that no electrical or mechanical damage has occurred during the in-house post-processing and TSV fabrication.

6.9 Software approach

In traditional 2D design, as the number of cores increases, the bandwidth of the shared memory becomes a performance bottleneck. In MIRACLE, the memory bottleneck problem can be tackled using resource pooling, thanks to the possibility to utilize all layers' resources as a whole. Resource pooling refers to the sharing of resources between vertical stacked layers [113].

MIRACLE's memory subsystem has only one write port and one read port. When the memory access rate exceeds a certain level, access blocking occurs. Memory bottlenecks and resource pooling can be evaluated by a memory-intensive benchmark which performs 1000 writes of integer-length values into the shared memory. Experiments writing on the local (same layer's) shared memory and the remote (different layer's) shared memory are performed with respectively 1-Core, 2-Core, 3-Core, and 4-Core. In this group of experiments, all active cores are on the same layer. The resulting execution times for both local and remote shared memory

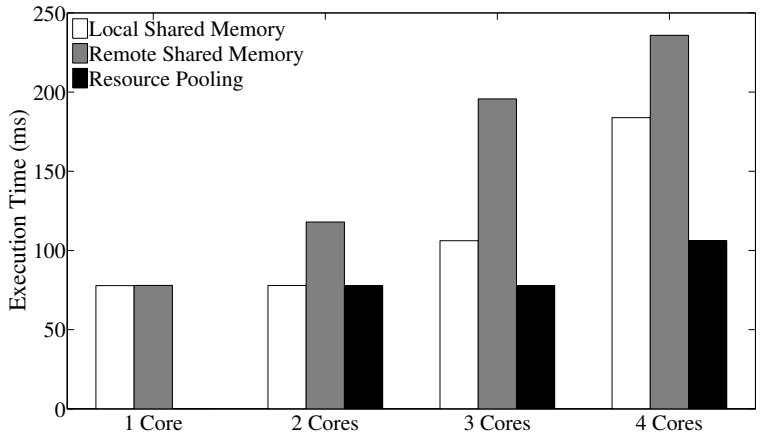


Figure 6.18: Comparison of execution time of the memory intensive benchmark when all cores access local memory, all cores access remote memory, and when memory resource pooling is applied.

cases are shown in Figure 6.18. This figure shows that as more cores attempt to access shared memory, performance penalty increases. When there are either more than two cores accessing the remote shared memory or more than three cores accessing the local shared memory, the extra cores are blocked for a while. Blocking of cores happens because of the memory bottleneck and the communication limitation between layers.

To overcome this bottleneck, the remote shared memory can be used to mitigate the access competition on the local shared memory, defining this scenario as *Memory Resource Pooling*. In order to explore the benefits of this approach, the following experiments are performed: for the multi-core cases we assign one core to access the remote shared memory while the other ones still access the local shared memory; for the 3-core case, one core is assigned to access remote shared memory while the other two write to the local shared memory. The results of the experiments are depicted in Figure 6.18. For 3-core and 4-core cases, memory resource pooling brings 26.6% and 42.3% reduction in execution time, respectively. This above memory resource pooling strategy schedules memory accesses at the core granularity; thus, it can be called *Core Level Resource Pooling (CLRP)*.

A second possibility is the *Task Level Resource Pooling (TLRP)*, which includes adjustable workload allocation and workload scheduling within each core. With TLRP, the workload of each core is divided into two parts: local memory accesses and remote memory accesses. For each core in the system, workload allocation determines the ratio of local and remote memory accesses, while workload scheduling defines their execution sequence. To perform a fair comparison, an equal amount of workload is allocated to all the cores. In Figure 6.19, each group of 4 bars represents the workload execution of four cores on the same layer. White and gray blocks stand for local and remote memory accesses, respectively, and black ones represent the memory stalls. In the same figure, *Scheduled* refers to the sequence of workload

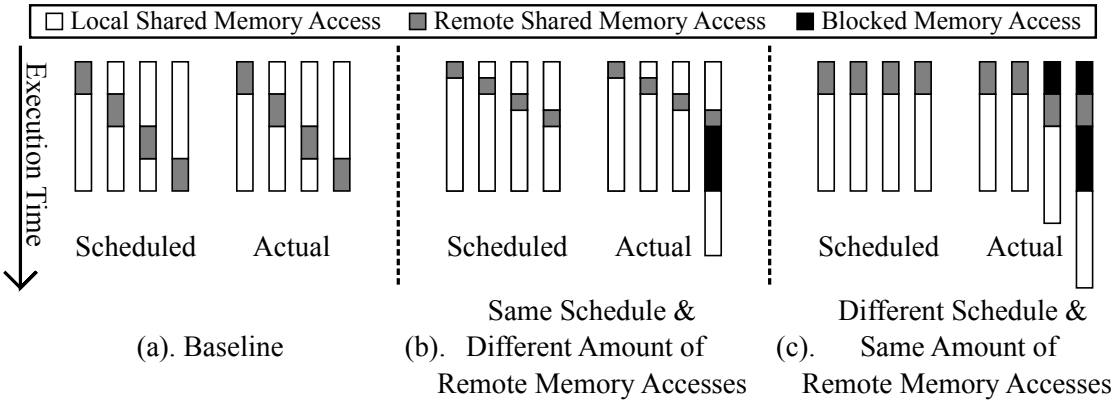


Figure 6.19: Performance under different workload allocation and scheduling combinations. (a) serves as the baseline, (b) has the same schedule as the baseline but fewer remote memory accesses, while (c) has the same number of remote memory accesses but has a different schedule.

execution planed for each core, while *Actual* shows the real execution. We consider (a) as a baseline case, where simultaneous local memory accesses from 4 cores are avoided. In this case, all cores behave as scheduled and no core is blocked because of contention. Case (b) shows the situation where the system applies the same workload schedule with (a) but with fewer remote memory accesses. Since local shared memory allows for 3-Core simultaneous access at most, there is a noticeable performance loss once all four cores access the local shared memory. Case (c) demonstrates the system's behaviour when the cores have the same amount of remote memory accesses as (a), but they are scheduled to access local and remote shared memory at the same time, which causes considerable performance loss compared to (a). Thus, combination of workload allocation and scheduling in TLRP has significant effects on performance.

To optimize performance via memory resource pooling, memory congestion should be avoided as much as possible. The following approach is used to compute the relationship between performance and the number of remote memory accesses. Three workload schedules are introduced, as shown in Figure 6.20, where each schedule is applicable to any workload allocation. In Figure 6.20, remote memory accesses increase gradually from left side to right side. Schedule (a) always makes all four cores access the remote shared memory (*4-thread-RSM*, where RSM stands for remote shared memory). Schedule (b) issues two cores to access remote shared memory at a time (*2-thread-RSM*) from (1) to (3). From (4) on, remote memory accesses are too many to be scheduled using *2-thread-RSM*, thus mixed *2+4 thread-RSM* is applied until the ratio of remote memory accesses increase to 100%. *1-thread-RSM* has only one thread accessing remote shared memory at a time to minimize simultaneous local memory access, as shown in case (c) from (1) to (3). As the remote memory accesses increase, schedule (c) uses *1-thread-RSM*, *1+2 thread-RSM*, and *2+4 thread-RSM* successively, which minimizes simultaneous local memory accesses.

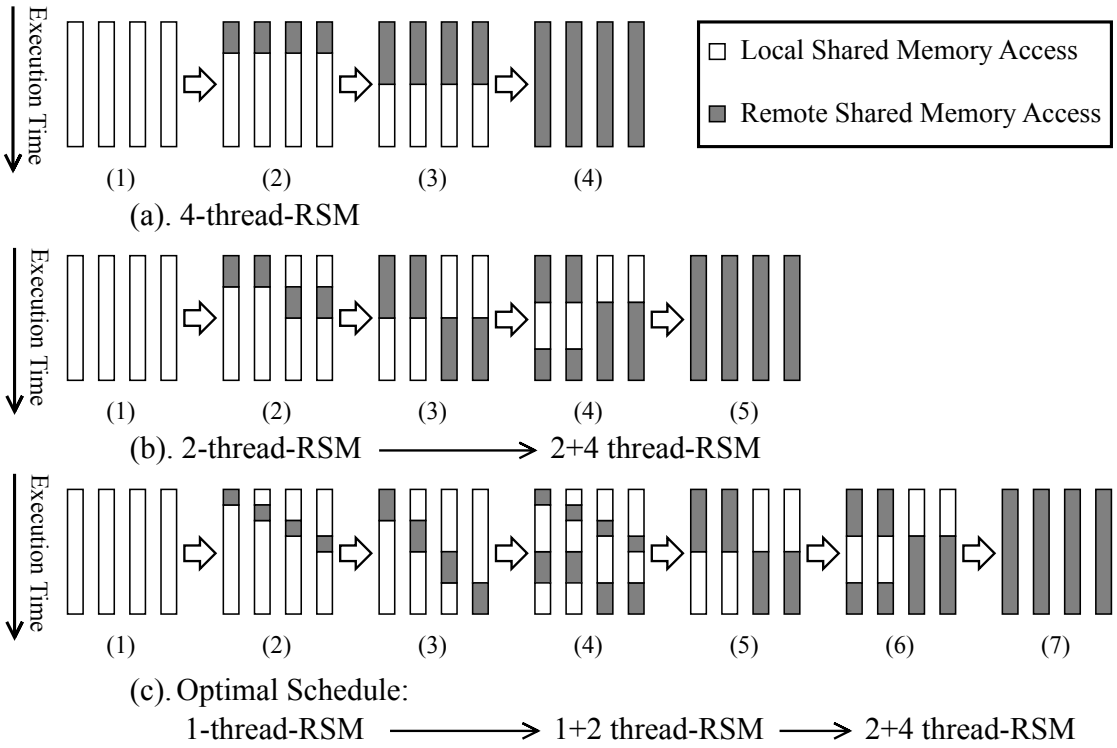


Figure 6.20: Workload schedules for task level resource pooling. (a). 4 threads accessing remote shared memory at the same time–*4-thread-RSM*; (b). (1)-(3): 2 threads accessing remote shared memory at the same time–*2-thread-RSM*; (c). (1)-(3): 1 thread accessing remote shared memory–*1-thread-RSM*.

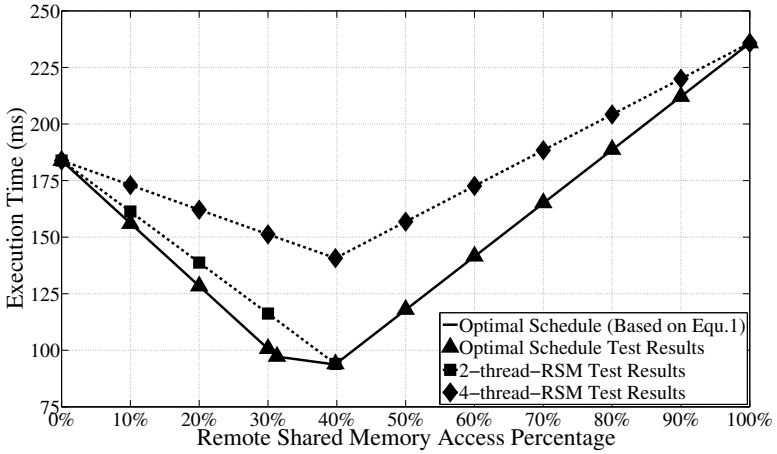


Figure 6.21: Test results of different memory resource pooling schedules and the optimal schedule's curve based on Eqn.1.

To compare the performance of these 3 schedules, we observe the execution time of the whole application, i.e., the longest execution time among all four cores. The execution time on each core is the product T_{exec} = instruction count × cycles per instruction × cycle time. Since most of the memory-intensive applications contain a large number of memory accesses, instructions can be substituted by memory accesses, which means that the execution time becomes T_{exec} = # of Memory Accesses × Cycles per Memory Access × Cycle Time. To apply TLRP, we need to split memory accesses into local memory accesses and remote memory accesses. In the following equation, T_{exec} stands for the execution time, N_{MemAcc} is the total number of shared memory accesses in the application, W_L and W_R represent the weight (i.e., ratio) of local and remote memory access, and C_{Li} and C_{Ri} refer to the number of cycles when i cores are accessing local shared memory and remote shared memory, respectively. C_{Li} and C_{Ri} can be obtained from the shared memory access test. After replacing the variables with their values we can compute the function of T_{exec} and W_R, and thus the execution time can be computed according to the ratio of remote memory accesses.

$$T_{exec} = (\sum(W_{Li} \times N_{MemAcc} \times C_{Li}) + \sum(W_{Ri} \times N_{MemAcc} \times C_{Ri})) \times T_{Cycle} \tag{6.1}$$

$$W_L + W_R = \sum W_{Li} + \sum W_{Ri} = 1 \tag{6.2}$$

Figure 6.21 shows the test results and fitted curves of the workload schedules in Figure 6.20. The theoretical curve based on Eqn. 6.1 is drawn for the optimal schedule. The experimental results fit with the theoretical curve. Although all of the three schedules can take advantage of resource pooling, the optimal one improves the performance by 48.9% at most, which coincides with the curve for 2-thread-RSM. The optimal schedule shown in the Figure 6.21 demonstrates the potential benefits of memory resource pooling and TLRP workload scheduling.

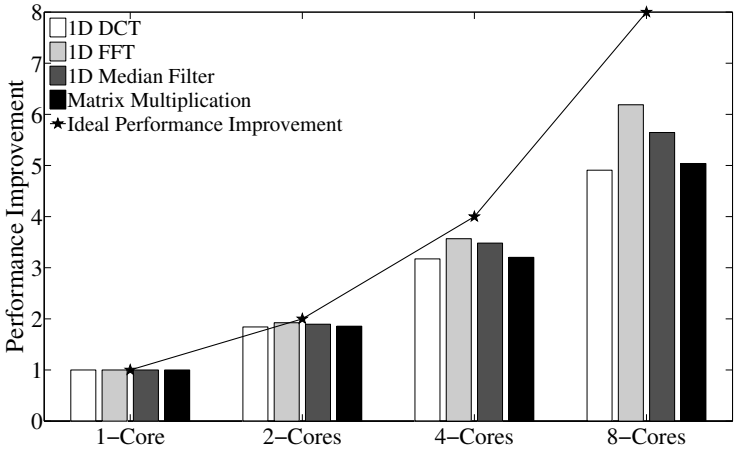


Figure 6.22: Performance improvement compared to single core.

In Eqn. 6.1, N_{MemAcc} is only related to the applications, T_{cycle} depends only on the architecture, and C_{Li} and C_{Ri} are both application and architecture related. Thus, for most of the shared-memory systems and applications, the proposed approach is applicable for deciding whether there would be potential performance improvement of memory resource pooling and for quantifying the benefits.

6.10 Performance evaluation

To evaluate the performance of MIRACLE, four benchmarks has been designed: **1D DCT**, a single dimension 8 × 8 matrix discrete cosine transformation; **1D FFT**, 8 × 8 matrix 12 butterfly Fast Fourier Transformation; **1D Median Filter** with a window size of 3 and an input array with 64 integers; **Matrix Multiplication**, 8 × 8 multiplication implemented using divide and conquer algorithm.

The test results are shown in Figure 6.22. Ideal performance improvement refers to the improvement that can be achieved by the multi-core system if the benchmarks can be fully parallelized. For 1-Core to 4-Core cases, the cores are all on one layer and thus the system can be viewed as a 2D layer only. For the 8-Core case, we have two 2D layers with four cores on each, which is the 3D-MMC architecture described in the previous chapter. It can be observed in Figure 6.22 that the performance improves significantly (61% on average) from 4-Core (2D) to 8-Core (3D). The difference of improvement among benchmarks is due to the fact that the benchmarks vary in their scalability. As an example, both DCT and FFT have the same input matrix and large amounts of computation, but FFT is more computation-bound compared to DCT. Reading the input can be viewed as the serial part of a benchmark while the computation is the parallel part since it can be done locally in each PE. Thus, the scalability of DCT is lower compared to FFT, and this difference results in different performance improvement in

Table 6.5: Execution Time of Different Shared Memory Access Scenarios When All 8 Cores are Active.

Benchmark	All cores access a single shared memory [ns]	All cores access their local shared memory [ns]
1D DCT	16716	16495
1D FFT	26260	26063
Median Filter	8674	7420
Matrix Multiplication	52838	51935

turn. Figure 6.22 demonstrates the performance benefits of using 3D stacking. Assuming a negligible area overhead is imposed by stacking, performance per area is 61% better than a 2D chip on average.

For the 8-Core case there are two ways of shared memory accesses for the cores: accessing a single shared memory in the system or each core accessing their local shared memory only. Table 6.5 shows the execution times of these two scenarios. For DCT and FFT, execution times differ by 1.3% and 0.8% only. Matrix Multiplication has 1.7% difference between these two situations because of its slightly higher memory access rate. There is a much larger difference (16.9%) for Median Filter because it is the most memory-intensive benchmark.

Finally, memory resource pooling is applied to the Median Filter benchmark. Since this benchmark also needs a lot of private memory accesses, four cores running this benchmark does not stress the shared memory sufficiently to reach the memory bottleneck, limiting the performance improvement to 5%. The available benefit from memory resource pooling is proportional to the memory access rate. The memory access rate becomes higher with a larger number of cores on 2D layer and/or by running more memory-intensive applications. When applying resource pooling to more than 2 layers in a 3D system, the benefits are expected to increase as the cores can utilize a larger number of shared memory blocks across different layers.

6.11 Summary

In this chapter, MIRACLE, a test vehicle based on the 3D-MMC architecture has been introduced.

A 36:8 serialization is implemented, vertical data packets are transmitted over multiple cycles, resulting in a vertical data bandwidth of 3.2 Gb/s. Even with such low data rate and the additional latency, the 2-layered 3D-MMC achieves 61% performance improvement compared to a single layered 4-core chip for a set of parallel workloads. This demonstrates the benefit of the proposed 3D serial interconnection in a complete CMP.

The homogeneous integration approach can offer a significant reduction of the Non Recurring

Engineering cost. Stacking identical, fully testable multi-processor dies with 4 processing elements and memory units on each die, leads to an increased yield for the final 3D system, built out of KGD. Coherent design and testing strategies are proposed and demonstrated to ensure robust operation.

A test vehicle, consisting of two layers, has been fabricated using standard UMC 90nm CMOS process. Single dies have been tested to be functional, and then post-processed for the in-house TSV fabrication and stacking. The proposed 3D system can operate at 400MHz, with a vertical bandwidth of 3.2Gbps.

6.12 Acknowledgements

Several people contributed to the development of the MIRACLE project. The author would like to thank Yuksel Temiz and Michael Zervas for their work on the TSV fabrication process, Tiansheng Zhang and Prof. Ayse Coskun for their contribution on the software development and the thermal model, Paolo Giovannini for his work on the testing setup during his master thesis under the author's supervision and Alessandro Cevrero for the successful collaboration that led to the conception of MIRACLE.

7 Summary and Conclusions

Planar on-chip interconnects have become a major concern as their impact on IC performance has been progressively increasing with each technology node, becoming a dominant source of circuit delay and power consumption. As today's semiconductor technology has started facing this "interconnection bottleneck", the ever increasing demand for higher speed, lower power and more functionality has lead to a paradigm shift towards 3D integration. Stacking multiple dies and interconnecting them through the silicon substrate with TSVs is a promising solution to continue the "More-than-Moore" trend. Nevertheless, TSV dimensions are limited by the chip assembly process, and will hardly overcome the sub-micron scale. Hence, the area occupied by TSVs is far from negligible compared to nanometer BEOL structures.

The main objective of this thesis is to explore the speed/power/area trade-offs for cross-chip data communication through state-of-the-art TSV channels. More specifically, the aim of this thesis work consists of proposing design solutions to leverage the high bandwidth and low-delay connection provided by TSV links minimizing the cost in terms of silicon area and capacitance. The summary of the thesis work and the main contributions are presented in the following sections. The dissertation is then concluded with a discussion on the possible directions for future work.

7.1 Summary

The dissertation starts presenting a compact TSV model in Chapter 2. The model is then used for fast and accurate exploration of the TSV performance and impact on electronic systems' performance throughout the remainder of the thesis. Chapter 3 presents a configurable 3D network architecture interconnecting a system composed of a cluster of processing elements, placed on a logic layer, and multiple layers of SRAM modules constituting a single shared L1 memory. TSVs, despite being short and fast, still occupy a significant area compared to the CMOS logic. To address this issue, in Chapter 4 we propose a TSV-based 3D serial link carrying out a design space exploration to identify the serialization level that efficiently balances the area occupied by the TSVs and the power consumption of the vertical link. The effect of the

TSV serialization on the clock distribution network is also discussed. A 3D modular multi-core processor platform integrating the proposed serial TSV link is then presented in Chapter 5. The effect of the serialization on the TSV count and the chip routing are presented. Finally, the capabilities of the proposed serial approach are validated using the test vehicle, MIRACLE, described in Chapter 6. The test vehicle consists of fully functional multi-processor dies based on standard CMOS technology and stacked using an in-house TSV fabrication process to obtain the 3D prototype.

7.2 Main contributions

- We present 3D-LIN, a configurable logarithmic network that can be integrated in a 3D stacked CMP. The architecture of the 3D network has been optimized targeting tightly coupled processor clusters, for which performance critically depends by the the interconnect between the processors and the memory banks. The 3D implementation and the reconfigurability of the 3D network expand the storage capability of the system, still guaranteeing single cycle, low-latency communication. We demonstrate that in the case of memory occupation of 60% of the planar chip, by moving to a system that integrates two memory layers on top of a logic layer, the form factor is improved more than 60%. In terms of latency, the 16x128 configuration of the network can be improved up to around 30%. Latency and area improvements come without a worsening of power consumption.

- A serial vertical high speed TSV link has been developed in order to minimize the area footprint occupied by the TSV channels still guaranteeing the desired bandwidth. We explored the effect of different levels of serialization on the area and energy efficiency for a range of available TSV technologies. We demonstrate that an 8-bit serialization guarantees a good balance between area consumption and energy efficiency across all the explored TSVs.

- A modular 3D stacked multi-processor platform, 3D-MMC, consisting of identical dies has been proposed as an alternative to the memory-on-logic stacking approach. The homogeneous integration approach can offer a significant reduction of the Non Recurring Engineering cost. Stacking identical, fully testable multi-processor dies with four processing elements and memory units on each die, leads to an increased yield for the final 3D system, built out of KGD.

- The serial vertical link has been integrated in the 3D-MMC platform. Results show that the serial approach reaches up to 12.4% wirelength improvement compared to the fully parallel counterpart when using 10μm TSVs. Even for high end TSV technologies such as 5μm TSVs, the wirelength undergoes an average reduction of 5.3%.

- MIRACLE, a test vehicle based on the 3D-MMC architecture has been implemented. The test vehicle includes a serial inter-layer connection and the vertical data packets

are transmitted over multiple cycles, resulting in a vertical data bandwidth of 3.2 Gb/s at 400MHz. Even with such low data rate and the additional latency, the 2-layered prototype achieves 60% performance improvement compared to a single layered 4-core chip for a set of parallel workloads, demonstrating the benefit of the proposed 3D serial interconnection in a complete CMP.

7.3 Future work

Although a significant amount of research has been focused on various aspects of 3D IC technology in the last years, there are still several aspects that need to be addressed before this technology can be widely adopted. Focusing on the topics tackled in this thesis work, we can highlight future research directions for improving 3D IC design.

First of all, the proposed TSV model should be extended to include the crosstalk effect and validated against measurements for the different sizes/process parameters. Fabricated TSV channels integrated with CMOS logic for the signal transmission need to be characterized to extract the actual maximum TSV bandwidth.

In Chapter 3 we have proposed a logarithmic parametric combinatorial network connecting multiple processing elements to on-chip shared multi-banks SRAM. Although this structure provides a convenient shared memory abstraction while avoiding cache coherence overheads, it still has drawbacks in terms of miss latency when the requested data is not in the shared memory, but should be fetched from the main memory. This problem can be a serious performance bottleneck for several applications. Future research efforts should be focused on replacing the shared SRAM with a cache memory.

At the circuit design level, it would be interesting to explore different SERDES topologies to optimize different systems. In this dissertation, we have proposed a high speed 3D link design that maximizes the transmission bandwidth in Chapter 4, while a low-speed, multi-cycle transmission vertical link has been included in the MIRACLE prototype presented in Chapter 6. Focusing on energy efficiency as design priority, a ultra low-power SERDES design for low speed inter-layer data transmission should be investigated.

As the CMOS technology nodes reach the deep sub-micron range, such as the 14nm technology node, the effect of TSV insertion on the system performance is expected to face major challenges. For this reason, future research should be focused on investigating the impact of the proposed serial configuration on 3D ICs built in future CMOS technology nodes.

Bibliography

[1] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, April 1965.

[2] ITRS, "Interconnect," 2011, accessed February 6, 2014. [Online]. Available: http://www.itrs.net/Links/2011itrs/2011Chapters/2011Interconnect.pdf

[3] P. Garrou, B. C., and P. Ramm, Eds., *Handbook of 3D Integration: Technology and Applications of 3D Integrated Circuits*. WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, 2012.

[4] W. Arden, M. Brilluet, P. Cogez, M. Graef, B. Huizing, and R. Mahnkopf, "More than moore white paper," 2010.

[5] I. Bolsens, "2.5d ics: just a stepping stone or a long term alternative to 3d?" accessed February 12th, 2014. [Online]. Available: http://www.xilinx.com/innovation/research-labs/keynotes/3-D_Architectures.pdf

[6] D. Nelson, C. Webb, D. McCauley, K. Raol, J. Rupley, J. DeVale, and B. Black, "A 3d interconnect methodology applied to ia32-class architectures for performance improvements through rc mitigation," in *Proceedings of the 21st Intl. VLSI Multilevel Interconnection Conf.*, 2004.

[7] V. F. Pavlidis and E. G. Friedman, *Three-dimensional Integrated Circuit Design*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2009.

[8] www.tezzaron.com/OtherICs.

[9] Cadence, "Comprehensive infrastructure for 3d ic design, ip, implementation, test, analysis and verification," accessed February 11th, 2014. [Online]. Available: http://www.cadence.com/solutions/3dic/pages/default.aspx

[10] Synopsys, "Accelerating 3d-ic innovation," accessed February 11th, 2014. [Online]. Available: http://www.synopsys.com/Solutions/EndSolutions/3d-ic-solutions/Pages/default.aspx

[11] B. Shi and A. Srivastava, "Liquid cooling for 3d-ics," in *Green Computing Conference and Workshops (IGCC), 2011 International*, 2011, pp. 1–6.

[12] A. Coskun, J. Ayala, D. Atienza, and T. Rosing, "Modeling and dynamic management of 3d multicore systems with liquid cooling," in *Very Large Scale Integration (VLSI-SoC), 2009 17th IFIP International Conference on*, Oct 2009, pp. 35–40.

[13] J. Li and H. Miyashita, "Efficient thermal via planning for placement of 3d integrated circuits," in *Circuits and Systems, 2007. ISCAS 2007. IEEE International Symposium on*, May 2007, pp. 145–148.

[14] L. Zhang, H. Li, G. Lo, and C. Tan, "Thermal characterization of tsv array as heat removal element in 3d ic stacking," in *Electronics Packaging Technology Conference (EPTC), 2012 IEEE 14th*, Dec 2012, pp. 153–156.

[15] J. Cong, G. Luo, J. Wei, and Y. Zhang, "Thermal-aware 3d ic placement via transformation," in *Design Automation Conference, 2007. ASP-DAC '07. Asia and South Pacific*, 2007, pp. 780–785.

[16] X. Zhou, J. Yang, Y. Xu, Y. Zhang, and J. Zhao, "Thermal-aware task scheduling for 3d multicore processors," *IEEE Trans. Parallel Distrib. Syst.*, vol. 21, no. 1, pp. 60–71, Jan. 2010.

[17] JEDEC, "Wide i/o single data rate (wide i/o sdr), jesd229," JEDEC, Tech. Rep., 2011.

[18] G. Beanato, P. Giovannini, A. Cevrero, P. Athanasopoulos, M. Zervas, Y. Temiz, and Y. Leblebici, "Design and testing strategies for modular 3-d-multiprocessor systems using die-level through silicon via technology," *Emerging and Selected Topics in Circuits and Systems, IEEE Journal on*, vol. 2, no. 2, pp. 295 –306, june 2012.

[19] I. Loi, S. Mitra, T. Lee, S. Fujita, and L. Benini, "A low-overhead fault tolerance scheme for tsv-based 3d network on chip links," in *Computer-Aided Design, 2008. ICCAD 2008. IEEE/ACM International Conference on*, 2008, pp. 598–602.

[20] A. Rahimi, I. Loi, M. Kakoee, and L. Benini, "A fully-synthesizable single-cycle interconnection network for shared-l1 processor clusters," in *Design, Automation Test in Europe Conference Exhibition (DATE), 2011*, march 2011, pp. 1 –6.

[21] E. Culurciello and A. Andreou, "Capacitive inter-chip data and power transfer for 3-d vlsi," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 53, no. 12, pp. 1348–1352, Dec 2006.

[22] D. Mizoguchi, Y. Yusof, N. Miura, T. Sakura, and T. Kuroda, "A 1.2gb/s/pin wireless superconnect based on inductive inter-chip signaling (iis)," in *Solid-State Circuits Conference, 2004. Digest of Technical Papers. ISSCC. 2004 IEEE International*, Feb 2004, pp. 142–517 Vol.1.

[23] Y. Temiz, "3d Integration Technology for Lab-on-a-Chip Applications," Ph.D. dissertation, STI, Lausanne, 2012.

[24] R. Ferrant. [Online]. Available: http://www.d43d.com

[25] "Iftle 48 sematech addresses the reliability impact of stress on 3dic," accessed February 20th, 2014. [Online]. Available: http://electroiq.com/insights-from-leading-edge/2011/04/iftle-48-sematech-addresses-the-reliability-impact-of-stress-on-3dic/

[26] K. Athikulwongse, A. Chakraborty, J. seok Yang, D. Pan, and S.-K. Lim, "Stress-driven 3d-ic placement with tsv keep-out zone and regularity study," in *Computer-Aided Design (ICCAD), 2010 IEEE/ACM International Conference on*, 2010, pp. 669–674.

[27] J. Cho, J. Shim, E. Song, J. S. Pak, J. Lee, H. Lee, K. Park, and J. Kim, "Active circuit to through silicon via (TSV) noise coupling," in *Electrical Performance of Electronic Packaging and Systems, 2009. EPEPS '09. IEEE 18th Conference on*, 2009, pp. 97–100.

[28] H. Chaabouni, M. Rousseau, P. LeDuc, A. Farcy, R. El Farhane, A. Thuaire, G. Haury, A. Valentian, G. Billiot, M. Assous, F. De Crecy, J. Cluzel, A. Toffoli, D. Bouchu, L. Cadix, T. Lacrevaz, P. Ancey, N. Sillon, and B. Flechet, "Investigation on tsv impact on 65nm cmos devices and circuits," in *Electron Devices Meeting (IEDM), 2010 IEEE International*, dec. 2010, pp. 35.1.1 –35.1.4.

[29] G. Katti, A. Mercha, J. Van Olmen, C. Huyghebaert, A. Jourdain, M. Stucchi, M. Rakowski, I. Debusschere, P. Soussan, W. Dehaene, K. De Meyer, Y. Travaly, E. Beyne, S. Biesemans, and B. Swinnen, "3d stacked ics using cu tsvs and die to wafer hybrid collective bonding," in *Electron Devices Meeting (IEDM), 2009 IEEE International*, dec. 2009, pp. 1 –4.

[30] J.-S. Kim, C. S. Oh, H. Lee, D. Lee, H.-R. Hwang, S. Hwang, B. Na, J. Moon, J.-G. Kim, H. Park, J.-W. Ryu, K. Park, S. K. Kang, S.-Y. Kim, H. Kim, J.-M. Bang, H. Cho, M. Jang, C. Han, J.-B. Lee, J.-S. Choi, and Y.-H. Jun, "A 1.2 v 12.8 gb/s 2 gb mobile wide-i/o dram with 4x128 i/os using tsv based stacking," *Solid-State Circuits, IEEE Journal of*, vol. 47, no. 1, pp. 107–116, 2012.

[31] M. Farooq, T. L. Graves-Abe, W. F. Landers, C. Kothandaraman, B. Himmel, P. Andry, C. K. Tsang, E. Sprogis, R. Volant, K. Petrarca, K. R. Winstel, J. Safran, T. D. Sullivan, F. Chen, M. J. Shapiro, R. Hannon, R. Liptak, D. Berger, and S. S. Iyer, "3d copper tsv integration, testing and reliability," in *Electron Devices Meeting (IEDM), 2011 IEEE International*, 2011, pp. 7.1.1–7.1.4.

[32] M. Wordeman, J. Silberman, G. Maier, and M. Scheuermann, "A 3d system prototype of an edram cache stacked over processor-like logic using through-silicon vias," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, 2012, pp. 186–187.

[33] G. Van der Plas, P. Limaye, I. Loi, A. Mercha, H. Oprins, C. Torregiani, S. Thijs, D. Linten, M. Stucchi, G. Katti, D. Velenis, V. Cherman, B. Vandevelde, V. Simons, I. De Wolf, R. Labie, D. Perry, S. Bronckers, N. Minas, M. Cupac, W. Ruythooren, J. Van Olmen, A. Phommahaxay, M. de Potter de ten Broeck, A. Opdebeeck, M. Rakowski, B. De Wachter,

M. Dehan, M. Nelis, R. Agarwal, A. Pullini, F. Angiolini, L. Benini, W. Dehaene, Y. Travaly, E. Beyne, and P. Marchal, "Design issues and considerations for low-cost 3-d tsv ic technology," *Solid-State Circuits, IEEE Journal of*, vol. 46, no. 1, pp. 293 –307, jan. 2011.

[34] G. Katti, M. Stucchi, K. De Meyer, and W. Dehaene, "Electrical modeling and characterization of through silicon via for three-dimensional ics," *Electron Devices, IEEE Transactions on*, vol. 57, no. 1, pp. 256 –262, jan. 2010.

[35] M. Goldfarb and R. Pucel, "Modeling via hole grounds in microstrip," *Microwave and Guided Wave Letters, IEEE*, vol. 1, no. 6, pp. 135 –137, june 1991.

[36] F. Leferink, "Inductance calculations; methods and equations," in *Electromagnetic Compatibility, 1995. Symposium Record., 1995 IEEE International Symposium on*, 1995, pp. 16–22.

[37] M. Zervas, "High throughput screening platform for ultra large scale integration based on high-k silicon nanowire ISFET and 3D integration," Ph.D. dissertation, STI, Lausanne, 2013.

[38] F. Sun, "Self-Aligned 3D Chip Integration Technology and Through-Silicon Serial Data Transmission," Ph.D. dissertation, STI, Lausanne, 2011.

[39] C. Ryu, J. Lee, H. Lee, K. Lee, T. Oh, and J. Kim, "High frequency electrical model of through wafer via for 3-D stacked chip packaging," in *Electronics Systemintegration Technology Conference, 2006. 1st*, vol. 1, 2006, pp. 215–220.

[40] S. Kannan, S. Evana, A. Gupta, B. Kim, and L. Li, "3-D Copper based TSV for 60-GHz applications," in *Electronic Components and Technology Conference (ECTC), 2011 IEEE 61st*, 2011, pp. 1168–1175.

[41] S. Hu, Y.-Z. Xiong, J. Shi, L. Wang, B. Zhang, D. Zhao, T.-G. Lim, and X. Yuan, "THz-wave propagation characteristics of TSV-based transmission lines and interconnects," in *Electronic Components and Technology Conference (ECTC), 2010 Proceedings 60th*, June 2010, pp. 46–50.

[42] K. Yoon, G. Kim, W. Lee, T. Song, J. Lee, H. Lee, K. Park, and J. Kim, "Modeling and analysis of coupling between tsvs, metal, and rdl interconnects in tsv-based 3d ic with silicon interposer," in *Electronics Packaging Technology Conference, 2009. EPTC '09. 11th*, dec. 2009, pp. 702 –706.

[43] Y. Xie, "Processor architecture design using 3d integration technology," in *VLSI Design, 2010. VLSID '10. 23rd International Conference on*, jan. 2010, pp. 446 –451.

[44] J. Owens, W. Dally, R. Ho, D. Jayasimha, S. Keckler, and L.-S. Peh, "Research challenges for on-chip interconnection networks," *Micro, IEEE*, vol. 27, no. 5, pp. 96 –108, sept.-oct. 2007.

[45] S. Borkar and A. Chien, "The future of microprocessors," *Commun. ACM*, vol. 54, pp. 67–77, May 2011.

[46] R. Banakar, S. Steinke, B.-S. Lee, M. Balakrishnan, and P. Marwedel, "Scratchpad memory: a design alternative for cache on-chip memory in embedded systems," in *Hardware/Software Codesign, 2002. CODES 2002. Proceedings of the Tenth International Symposium on*, 2002, pp. 73–78.

[47] L. Benini and G. De Micheli, "Networks on chips: a new soc paradigm," *Computer*, vol. 35, no. 1, pp. 70 –78, jan 2002.

[48] A. Balkan, G. Qu, and U. Vishkin, "A mesh-of-trees interconnection network for single-chip parallel processing," in *Application-specific Systems, Architectures and Processors, 2006. ASAP '06. International Conference on*, sept. 2006, pp. 73 –80.

[49] L. Plurality, "The hypercore architecture," in *White Paper*, January 2010.

[50] F. Li, C. Nicopoulos, T. Richardson, Y. Xie, V. Narayanan, and M. Kandemir, "Design and management of 3d chip multiprocessors using network-in-memory," *SIGARCH Comput. Archit. News*, vol. 34, no. 2, pp. 130–141, May 2006.

[51] G. Loh, "3d-stacked memory architectures for multi-core processors," in *Proceedings of the 35th Annual International Symposium on Computer Architecture*, ser. ISCA '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 453–464.

[52] D. H. Woo, N. H. Seong, D. Lewis, and H.-H. Lee, "An optimized 3d-stacked memory architecture by exploiting excessive, high-density tsv bandwidth," in *High Performance Computer Architecture (HPCA), 2010 IEEE 16th International Symposium on*, jan. 2010, pp. 1 –12.

[53] N. Madan, L. Zhao, N. Muralimanohar, A. Udipi, R. Balasubramonian, R. Iyer, S. Makineni, and D. Newell, "Optimizing communication and capacity in a 3d stacked reconfigurable cache hierarchy," in *High Performance Computer Architecture, 2009. HPCA 2009. IEEE 15th International Symposium on*, feb. 2009, pp. 262 –274.

[54] A. Mishra, X. Dong, G. Sun, Y. Xie, N. Vijaykrishnan, and C. Das, "Architecting on-chip interconnects for stacked 3d stt-ram caches in cmps," *SIGARCH Comput. Archit. News*, vol. 39, no. 3, pp. 69–80, Jun. 2011.

[55] J. Kim, C. Nicopoulos, D. Park, R. Das, Y. Xie, V. Narayanan, M. Yousif, and C. Das, "A novel dimensionally-decomposed router for on-chip communication in 3d architectures," in *Proceedings of the 34th annual international symposium on Computer architecture*, ser. ISCA '07. New York, NY, USA: ACM, 2007, pp. 138–149.

[56] D. Park, S. Eachempati, R. Das, A. Mishra, Y. Xie, N. Vijaykrishnan, and C. Das, "Mira: A multi-layered on-chip interconnect router architecture," in *Proceedings of the 35th Annual International Symposium on Computer Architecture*, ser. ISCA '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 251–261.

[57] Y. Xu, Y. Du, B. Zhao, X. Zhou, Y. Zhang, and J. Yang, "A low-radix and low-diameter 3d interconnection network design," in *High Performance Computer Architecture, 2009. HPCA 2009. IEEE 15th International Symposium on*, feb. 2009, pp. 30 –42.

[58] L. Xue, Y. Gao, and J. Fu, "A high performance 3d interconnection network for many-core processors," in *Computer Engineering and Technology (ICCET), 2010 2nd International Conference on*, vol. 1, april 2010, pp. V1–383 –V1–389.

[59] A. Ben Ahmed, A. Ben Abdallah, and K. Kuroda, "Architecture and design of efficient 3d network-on-chip (3d noc) for custom multicore soc," in *Broadband, Wireless Computing, Communication and Applications (BWCCA), 2010 International Conference on*, nov. 2010, pp. 67 –73.

[60] V. Pavlidis, I. Savidis, and E. Friedman, "Clock distribution networks for 3-d integrated circuits," in *Custom Integrated Circuits Conference, 2008. CICC 2008. IEEE*, sept. 2008, pp. 651 –654.

[61] *Design Compiler®User Guide*, Synopsys, December 2011, version F-2011.09-SP2.

[62] B. Goplen and S. Sapatnekar, "Thermal via placement in 3d ics," in *Proceedings of the 2005 international symposium on Physical design*, ser. ISPD '05. New York, NY, USA: ACM, 2005, pp. 167–174. [Online]. Available: http://doi.acm.org/10.1145/1055137.1055171

[63] H. Yu and L. He, "Dynamic power and thermal integrity in 3d integration," in *Communications, Circuits and Systems, 2009. ICCCAS 2009. International Conference on*, july 2009, pp. 1108 –1112.

[64] ISSCC2014AdvanceProgram.pdf, accessed December 2nd, 2013. [Online]. Available: http://isscc.org/program/

[65] B. Black, M. Annavaram, N. Brekelbaum, J. DeVale, J. Jiang, G. Loh, D. McCaule, P. Morrow, D. Nelson, D. Pantuso, P. Reed, J. Rupley, S. Shankar, J. Shen, and C. Webb, "Die stacking (3d) microarchitecture," in *Microarchitecture, 2006. MICRO-39. 39th Annual IEEE/ACM International Symposium on*, dec. 2006, pp. 469 –479.

[66] P. Reed, G. Yeung, and B. Black, "Design aspects of a microprocessor data cache using 3d die interconnect technology," in *Integrated Circuit Design and Technology, 2005. ICICDT 2005. 2005 International Conference on*, may 2005, pp. 15 – 18.

[67] Z. Li, X. Hong, Q. Zhou, J. Bian, H. H. Yang, and V. Pitchumani, "Efficient thermal-oriented 3d floorplanning and thermal via planning for two-stacked-die integration," *ACM Trans. Des. Autom. Electron. Syst.*, vol. 11, no. 2, pp. 325–345, Apr. 2006. [Online]. Available: http://doi.acm.org/10.1145/1142155.1142159

[68] M. Aoki, F. Furuta, K. Hozawa, Y. Hanaoka, H. Kikuchi, A. Yanagisawa, T. Mitsuhashi, and K. Takeda, "Fabricating 3d integrated cmos devices by using wafer stacking and via-last tsv technologies," in *Electron Devices Meeting (IEDM), 2013 IEEE International*, 2013.

[69] Y. Liu, W. Luk, and D. Friedman, "A compact low-power 3d i/o in 45nm cmos," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2012 IEEE International*, 2012, pp. 142–144.

[70] S. Pasricha, "Exploring serial vertical interconnects for 3d ics," in *Design Automation Conference, 2009. DAC '09. 46th ACM/IEEE*, 2009, pp. 581–586.

[71] Y. Ghidini, M. Moreira, L. Brahm, T. Webber, N. Calazans, and C. Marcon, "Lasio 3d noc vertical links serialization: Evaluation of latency and buffer occupancy," in *Integrated Circuits and Systems Design (SBCCI), 2013 26th Symposium on*, 2013, pp. 1–6.

[72] B. C. Hien, S.-M. Kim, and K. Cho, "Design of a wave-pipelined serializer-deserializer with an asynchronous protocol for high speed interfaces," in *Quality Electronic Design (ASQED), 2012 4th Asia Symposium on*, July 2012, pp. 265–268.

[73] R. Dobkin, M. Moyal, A. Kolodny, and R. Ginosar, "Asynchronous current mode serial communication," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 18, no. 7, pp. 1107–1117, July 2010.

[74] A. Jose, G. Patounakis, and K. L. Shepard, "Pulsed current-mode signaling for nearly speed-of-light intrachip communication," *Solid-State Circuits, IEEE Journal of*, vol. 41, no. 4, pp. 772–780, April 2006.

[75] Y.-H. Hsu, M.-S. Kao, H.-C. Tzeng, C.-T. Chiu, J.-M. Wu, and S.-H. Hsu, "A 20 gbps scalable load balanced birkhoff-von neumann symmetric tdm switch ic with serdes interfaces," in *Design Automation Conference, 2007. ASP-DAC '07. Asia and South Pacific*, Jan 2007, pp. 102–103.

[76] D. Tondo and R. Lopez, "A low-power, high-speed cmos/cml 16:1 serializer," in *Micro-Nanoelectronics, Technology and Applications, 2009. EAMTA 2009. Argentine School of*, Oct 2009, pp. 81–86.

[77] W.-Y. Tsai, C.-T. Chiu, J.-M. Wu, S. Hsu, and Y.-S. Hsu, "A novel low gate-count pipeline topology with multiplexer-flip-flops for serial link," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 59, no. 11, pp. 2600–2610, Nov 2012.

[78] M. Kurisu, M. Kaneko, T. Suzaki, A. Tanabe, M. Togo, A. Furukawa, T. Tamura, K. Nakajima, and K. Yoshida, "2.8 gb/s 176 mw byte-interleaved and 3.0 gb/s 118 mw bit-interleaved 8:1 multiplexers," in *Solid-State Circuits Conference, 1996. Digest of Technical Papers. 42nd ISSCC., 1996 IEEE International*, 1996, pp. 122–123.

[79] M. Yuffe, E. Knoll, M. Mehalel, J. Shor, and T. Kurts, "A fully integrated multi-cpu, gpu and memory controller 32nm processor," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International*, 2011, pp. 264–266.

[80] ITRS, 2011, accessed February 6, 2014. [Online]. Available: http://www.itrs.net/reports.html

[81] NIST, accessed March 18, 2014. [Online]. Available: http://www.nist.gov

[82] Agilent, "Jitter analysis ttechnique for high data rates," Agilent Technologies, Application Note 1432.

[83] H.-H. Lee, "2.5 and 3d ic - the status, the challenges and their prosect," Presentation slides, 2014.

[84] ISSCC, "Isscc 2013 trends," 2013, accessed March 20, 2014. [Online]. Available: http://isscc.org/trends/

[85] D. H. Kim, S. Mukhopadhyay, and S. K. Lim, "Through-silicon-via aware interconnect prediction and optimization for 3d stacked ics," in *Proceedings of the 11th International Workshop on System Level Interconnect Prediction*, ser. SLIP '09. New York, NY, USA: ACM, 2009, pp. 85–92. [Online]. Available: http://doi.acm.org/10.1145/1572471.1572486

[86] T. Kgil, S. D'Souza, A. Saidi, N. Binkert, R. Dreslinski, T. Mudge, S. Reinhardt, and K. Flautner, "Picoserver: using 3d stacking technology to enable a compact energy efficient chip multiprocessor," in *Proceedings of the 12th international conference on Architectural support for programming languages and operating systems*, ser. ASPLOS-XII. New York, NY, USA: ACM, 2006, pp. 117–128.

[87] I. Loi, P. Marchal, A. Pullini, and L. Benini, "3d nocs - unifying inter & intra chip communication," in *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on*, june 2010, pp. 3337 –3340.

[88] W. L. Byung-Gyu Ahn, Jaehwan Kim and J.-W. Chong, "Effective estimation method of routing congestion at floorplan stage for 3d ics," *Journal of semiconductor technology and science*, vol. 11, pp. 344 – 349, 2011.

[89] M.-C. Tsai, T.-C. Wang, and T. Hwang, "Through-silicon via planning in 3-d floorplanning," *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, vol. 19, no. 8, pp. 1448–1457, Aug 2011.

[90] J. Cong and G. Luo, "A multilevel analytical placement for 3d ics," in *Design Automation Conference, 2009. ASP-DAC 2009. Asia and South Pacific*, Jan 2009, pp. 361–366.

[91] B. Lee and T. Kim, "Algorithms for tsv resource sharing and optimization in designing 3d stacked ics," *Integr. VLSI J.*, vol. 47, no. 2, pp. 184–194, Mar. 2014. [Online]. Available: http://dx.doi.org/10.1016/j.vlsi.2013.11.001

[92] L. Benini, E. Flamand, D. Fuin, and D. Melpignano, "P2012: Building an ecosystem for a scalable, modular and high-efficiency embedded computing accelerator," in *Design, Automation Test in Europe Conference Exhibition (DATE), 2012*, march 2012, pp. 983 –987.

[93] N. Magen, A. Kolodny, U. Weiser, and N. Shamir, "Interconnect-power dissipation in a microprocessor," in *Proceedings of the 2004 International Workshop on System Level Interconnect Prediction*, ser. SLIP '04. New York, NY, USA: ACM, 2004, pp. 7–13. [Online]. Available: http://doi.acm.org/10.1145/966747.966750

[94] A. Azevedo, C. Meenderinck, B. Juurlink, A. Terechko, J. Hoogerbrugge, and M. Alvarez, "Parallel h.264 decoding on an embedded multicore processor," in *Proceedings of the 4th International Conference on High Performance and Embedded Architectures and Compilers*, 2009.

[95] G. Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities," in *Proceedings of the April 18-20, 1967, spring joint computer conference*, ser. AFIPS '67 (Spring). New York, NY, USA: ACM, 1967, pp. 483–485.

[96] J. Reinders, *Intel threading building blocks*, 1st ed. Sebastopol, CA, USA: O'Reilly & Associates, Inc., 2007.

[97] A. Topol, D. L. Tulipe, L. Shi, D. Frank, K. Bernstein, S. Steen, A. Kumar, G. Singco, A. Young, K. Guarini, and M. Ieong, "Three-dimensional integrated circuits," *IBM Journal of Research and Development*, vol. 50, no. 4.5, pp. 491 –506, july 2006.

[98] U. Kang, H.-J. Chung, S. Heo, D.-H. Park, H. Lee, J. H. Kim, S.-H. Ahn, S.-H. Cha, J. Ahn, D. Kwon, J.-W. Lee, H.-S. Joo, W.-S. Kim, D. H. Jang, N. S. Kim, J.-H. Choi, T.-G. Chung, J.-H. Yoo, J. S. Choi, C. Kim, and Y.-H. Jun, "8 gb 3-d ddr3 dram using through-silicon-via technology," *Solid-State Circuits, IEEE Journal of*, vol. 45, no. 1, pp. 111 –119, jan. 2010.

[99] U. Kang, H.-J. Chung, S. Heo, S.-H. Ahn, H. Lee, S.-H. Cha, J. Ahn, D. Kwon, K. J.H., L. J.-W., H.-S. Joo, K. W.-S., K. H.-K., L. E.-M, S.-R. Kim, K.-H. Ma, D.-H. Jang, N.-S. Kim, M.-S. Choi, S.-J. Oh, J.-B. Lee, T.-K. Jung, J.-H. Yoo, and C. Kim, "8gb 3d ddr3 dram using through-silicon-via technology," in *Solid-State Circuits Conference - Digest of Technical Papers, 2009. ISSCC 2009. IEEE International*, feb. 2009, pp. 130 –131,131a.

[100] F. Clermidy, F. Darve, D. Dutoit, W. Lafi, and P. Vivet, "3d embedded multi-core: Some perspectives," in *Design, Automation Test in Europe Conference Exhibition (DATE), 2011*, march 2011, pp. 1 –6.

[101] M. Healy, K. Athikulwongse, R. Goel, M. Hossain, D. Kim, Y.-J. Lee, D. Lewis, T.-W. Lin, C. Liu, M. Jung, B. Ouellette, M. Pathak, H. Sane, G. Shen, D. H. Woo, X. Zhao, G. Loh, H. Lee, and S. K. Lim, "Design and analysis of 3d-maps: A many-core 3d processor with stacked memory," in *Custom Integrated Circuits Conference (CICC), 2010 IEEE*, sept. 2010, pp. 1 –4.

[102] D. Lewis, M. Healy, M. Hossain, T.-W. Lin, M. Pathak, H. Sane, S. Lim, G. Loh, and H.-H. Lee, "Design and test of 3d-maps, a 3d die-stack many-core processor," in *first IEEE International Workshop on Testing Three-Dimensional Stacked Integrated Circuits (poster)*, Nov. 2010.

[103] P. Garrou, C. Bower, and P. Ramm, *Handbook of 3D Integration: Technology and Applications of 3D Integrated Circuits*. John Wiley & Sons, 2008, no. v. 2. [Online]. Available: http://books.google.com.au/books?id=LtbakwQlNs0C

[104] E. Marinissen, "Testing tsv-based three-dimensional stacked ics," in *Design, Automation Test in Europe Conference Exhibition (DATE), 2010*, march 2010, pp. 1689 –1694.

[105] F. Sun, A. Cevrero, P. Athanasopoulos, and Y. Leblebici, "Design and feasibility of multi-gb/s quasi-serial vertical interconnects based on tsvs for 3d ics," in *VLSI System on Chip Conference (VLSI-SoC), 2010 18th IEEE/IFIP*, sept. 2010, pp. 149 –154.

[106] Y. Temiz, M. Zervas, and Y. Leblebici, "A cmos compatible chip-to-chip 3d integration platform," in *IEEE 62nd Electronic Components and Technology Conference*, 2012, to appear.

[107] C. Bower, D. Malta, D. Temple, J. Robinson, P. Coffinan, M. Skokan, and T. Welch, "High density vertical interconnects for 3-d integration of silicon integrated circuits," in *Electronic Components and Technology Conference, 2006. Proceedings. 56th*, 2006, p. 5 pp.

[108] P. Morrow and S. Muthukumar, "Chapter 34. 3d and microprocessors," in *Handbook of 3D Integration: Technology and Applications of 3D Integrated Circuits*, C. B. . R. P. P. Morrow, S. M. P. Garrou, Ed. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2008, 2008.

[109] M. Zervas, Y. Temiz, and Y. Leblebici, "Fabrication and characterization of wafer-level deep tsv arrays," in *Electronic Components and Technology Conference (ECTC), 2012 IEEE 62nd*, 2012, pp. 1625–1630.

[110] K. Skadron *et al.*, "Temperature-aware microarchitecture," in *ISCA*, June 2003, pp. 2–13.

[111] J. Meng, K. Kawakami, and A. Coskun, "Optimizing energy efficiency of 3-D multicore systems with stacked DRAM under power and thermal constraints," in *DAC*, June 2012, pp. 648–655.

[112] B. Noia and K. Chakrabarty, "Pre-bond probing of tsvs in 3d stacked ics," in *Test Conference (ITC), 2011 IEEE International*, sept. 2011, pp. 1 –10.

[113] H. Homayoun, V. Kontorinis, A. Shayan, T.-W. Lin, and D. Tullsen, "Dynamically heterogeneous cores through 3d resource pooling," in *High Performance Computer Architecture (HPCA), 2012 IEEE 18th International Symposium on*, feb. 2012, pp. 1 –12.

GIULIA BEANATO

Av. De la Rochelle 6
1008 Prilly, Switzerland
+41 (0) 79 365 04 15
giulia.beanato@gmail.com

Birth: 11/08/1985, Torino Italy
Female
Italian citizenship
Swiss B permit
Single, no children

EDUCATION

Doctor of philosophy (PhD) in Micro-Systems and Micro-Electronics — March 2010- Expected: June 2014
EPFL - Switzerland

Master in Micro and Nanotechnologies for the Integrated Systems — September 2007- September 2009
EPFL-Switzerland, INP Grenoble-France, Politecnico di Torino - Italy

Bachelor in Mechatronic Engineering — September 2004- July 2007
Politecnico di Torino - Italy

RESEARCH AND WORK EXPERIENCES

Research Assistant - *EPFL, Microelectronic Systems Laboratory – Lausanne, Switzerland* — 2010-on going
Exploration and design of digital circuit solutions for 3D IC technology. Design and testing of a fully functional chip multi-processor taped out in 90nm CMOS.
A list of publications is available at: http://people.epfl.ch/giulia.beanato

Internship - *EPFL, Microelectronic Systems Laboratory– Lausanne, Switzerland* — September 2009 - January 2010
Digital calibration of pipelined ADC. Semi-custom digital design.

Master Project - *EPFL, Microelectronic Systems Laboratory– Lausanne, Switzerland* — January 2009 - September 2009
Design of very low power SAR-ADC in UMC 90nm technology. Modeling, semi-custom digital design, full-custom analog design.

Internship - *ChiLab – research laboratory – Chivasso, Italy* — July 2008 - September 2008
Research on Lab-on-Chip (LOC), microfluidic devices. Clean-room.

Internship - *Promec Elettronica S.r.L. – Ivrea, Italy* — September 2003 - January 2004
FPGA developer for a system of current amplifiers driving the gradient coils of a Magnetic Resonance Imaging (MRI) machine.

COMPUTER SKILLS

- **FPGA tools**: Xilinx ISE and EDK, Altera Quartus II.
- **CAD tools**: Cadence Virtuoso IC, Encounter digital Implementation (EDI), Synopsys IC/Design compiler, Modelsim, Agilent Advanced Design System (ADS).
- **Programming languages**: C, VHDL, Verilog, Verilog-A, Verilog-AMS, Matlab/Simulink, shell scripts.
- **Text editing, framework and tools:** MS Office Suite, Latex, MS Windows, Linux.

LANGUAGES

- **English:** fluent (C1-C2)**.**
- **French:** good (B2)**.**
- **Italian:** native speaker.

INTERESTS

Improvisational theatre, classic theatre. I love travelling and knowing new cultures.

LIST OF PUBBLICATIONS

- **G. Beanato**, A. Cevrero, G. De Micheli and Y. Leblebici. *Low Power 3D Serial TSV Link for High Bandwidth Cross-Chip Communication*. 51st ACM/EDAC/IEEE Design Automation Conference (DAC), San Francisco, California, USA, 2014.

- **G. Beanato**, A. Cevrero, G. De Micheli and Y. Leblebici. *3D serial TSV link for low-power chip-to-chip communication*. 2014 IEEE International Conference on Integrated Circuit Design and Technology (ICICDT), Austin, Texas, USA, 2014.

- **G. Beanato**, I. Loi, G. De Micheli, Y. Leblebici and L. Benini. *Configurable Low-Latency Interconnect for Multi-core Clusters*, in VLSI-SoC: From Algorithms to Circuits and System-on-Chip Design, p. 107-124, 2013.

- T. Zhang, A. Cevrero, **G. Beanato**, P. Athanasopoulos and A. Coskun et al. *3D-MMC: A Modular 3D Multi-Core Architecture with Efficient Resource Pooling*. Design, Automation & Test in Europe Conference (DATE 2013), Grenoble, France, 2013.

- **G. Beanato**, I. Loi, G. De Micheli, Y. Leblebici and L. Benini. *3D-LIN: A Configurable Low-Latency Interconnect for Multi-Core Clusters with 3D Stacked L1 Memory*. IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC), Santa Cruz, California, USA, 2012.

- A. Cevrero, **G. Beanato**, P. Athanasopoulos and Y. Leblebici. *Towards Cost Effective Multi-Core Processor Platforms Using 3-D Stacking Technology*. 49th Design Automation Conference (DAC 2012), San Francisco, 2012.

- P. Giovannini, **G. Beanato**, A. Cevrero, P. Athanasopoulos and Y. Leblebici. *A 3D Stacked Multi-Core Processor Platform with Improved Testability*. 12th Design Automation and Test in Europe Conference (Date 2012), Dresden, 2012.

- **G. Beanato**, P. Giovannini, A. Cevrero, P. Athanasopoulos and M. Zervas et al. *Design and Testing Strategies for Modular 3-D-Multiprocessor Systems Using Die-Level Through Silicon Via Technology*, in IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 2, num. 2, p. 295-306, 2012.