Title

Heterogeneous Integration on Silicon-Interconnect Fabric using fine-pitch interconnects (≤10 �m)

Permalink
https://escholarship.org/uc/item/4t08v7cs

Author

Jangam, SivaChandra

Publication Date
2020

UNIVERSITY OF CALIFORNIA

Los Angeles

Heterogeneous Integration on Silicon-Interconnect Fabric using fine-pitch interconnects ($\leq$10 μm)

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Electrical and Computer Engineering

by

SivaChandra Jangam

2020

ABSTRACT OF THE DISSERTATION

Heterogeneous Integration on Silicon-Interconnect Fabric
using fine-pitch interconnects ($\leq$10 μm)

by

SivaChandra Jangam
Doctor of Philosophy in Electrical and Computer Engineering
University of California, Los Angeles, 2020
Professor Subramanian Srikanteswara Iyer, Chair

Today, the ever-growing data-bandwidth demand is pushing the boundaries of the traditional printed circuit board (PCB) based integration schemes. Moreover, with the apparent saturation of semiconductor scaling, commonly called Moore's law, system scaling warrants a paradigm shift in packaging technologies, assembly techniques, and integration methodologies. In this work, a superior alternative to PCBs called the Silicon-Interconnect Fabric (Si-IF) is investigated. The Si-IF is a silicon-based, package-less, fine-pitch, highly scalable, heterogeneous integration platform for wafer-scale systems. In this technology, bare dielets are assembled on the Si-IF at small inter-dielet spacings ($\leq$100 μm) using fine-pitch ($\leq$10 μm) die-to-substrate interconnects. A novel assembly process using a solder-less direct metal-metal (gold-gold and copper-copper) thermal compression bonding was developed. Using this process, sub-10 μm pitch interconnects with a low specific contact resistance of $\leq$0.7 Ω-μm^2 were successfully demonstrated. Because of the tightly packed Si-IF assembly, the communication links between the neighboring dies are short ($\leq$500 μm) with low loss ($\leq$2 dB), comparable to on-chip connections. Consequently, simple buffers can transfer data between dies using a Simple Universal Parallel intERface for chips (SuperCHIPS) protocol at low latency (<30 ps), low energy per bit ($\leq$0.03 pJ/b), and high data-rates (up to 10 Gbps/link), corresponding to an aggregate bandwidth up to 8 Tbps/mm. The benefits of the SuperCHIPS protocol were experimentally demonstrated to provide 4-23X higher data-bandwidth, 3-65X lower latency, and 5-40X lower energy per bit compared to existing integration schemes. This dissertation addresses the assembly technology and communication protocols of the Si-IF technology.

The dissertation of SivaChandra Jangam is approved.

Mark S. Goorsky

Sudhakar Pamarti

Dejan Markovic

Subramanian Srikanteswara Iyer, Committee Chair

University of California, Los Angeles

2020

To my parents . . .

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am extremely indebted to my advisor, Prof. Subramanian Iyer for his support, and guidance throughout this wonderful journey. It was a true priviledge to work under a mentor of highest calibre, technical brilliance, and vast experience. He has always inspired me with his vision, critical thinking, and innovative solutions to challenging problems. I was truly fortunate to explore a wide range of research topics under his guidance and I thank him for giving me that freedom. Apart from technical knowledge, he also taught me many valuable life lessons that helped me become a better person. I am truly grateful to him for being a great guru and hope to continue to learn from him throughout the rest of my career.

I would like to thank Prof. Dejan Markovic for serving on my dissertation committee and guiding me in the design of the compute hardware. I thank him for the insightful discussions and feedback that helped me tremendously.

I also thank Prof. Sudhakar Pamarti for serving on my dissertation committee and his guidance in understanding and developing the models for simulations. I always enjoyed the meaningful discussions I had with him and learned a lot from his deep insights. He always provided me with a fresh perspective on solving problems.

I would also like to thank Prof. Mark Goorsky for being on my dissertation committee and supporting me over the years. His expertise and advice helped me better understand the bonding process.

I was blessed to have many incredible colleagues who were a great pleasure to work with. I thank Dr. Adeel Bajwa for being an excellent mentor and a friend with whom I continued to collaborate on assembly and process development from the start. I would also like to thank Prof. Puneet Gupta, Dr. Boris Vaisband, Dr. Amir Hanna, Dr. Umesh Mogera for their guidance and support. I would like to thank all my colleagues: Saptadeep Pal, Uneeb Rathore, Sumeet Nagi, Kannan K. Thankappan, Niloofar Sharookzadeh, Pranav Ambhore, Yu-tao Yang, Krutikesh Sahoo, Arpan Dasgupta, Meng-Hsiang Liu, Eric Sorensen, Steven Moran, Premsagar Kittur, Dr. Zhe Wan, Jonathan Cox, Randall Irwin, Arsalan Alam, Goutham Ezhilarasu, Aidan Wolk, and others for many successful collaborations over the years and meaningful friendships that I will cherish for life. I am indebted to all my co-authors for all their hard work and enthusiasm without which none of my research would have been possible. I would also like to thank the staff of UCLA Nanolab Research Facility, Integrated Systems Nanofabrication Cleanroom, Center for High-Frequency Electronics for their help and support.

I thank the Electronics Packaging Society and Guru Krupa Institute for providing me with generous fellowships for my graduate education. I thank all the government, consortia, and industry sponsors that funded my research. I thank all our industry collaborators who provided hardware, equipment, technical support, and service required for my research.

Parts of this dissertation were adapted from several publications during my doctoral studies.

VITA

2011–2015	Bachelor of Technology in Electrical Engineering, Indian Institute of Technology, Kanpur
2015–2017	Master of Science in Electrical Engineering, University of California, Los Angeles
2017–2020	Doctor of Philosophy in Electrical and Computer Engineering, University of California, Los Angeles

SELECTED PUBLICATIONS

S. Jangam, U. Rathore, S. Nagi, D. Markovic, and S. S. Iyer, "Demonstration of a Low latency ($<$20 ps) Fine-pitch ($\leq$10 μm) Assembly on the Silicon Interconnect Fabric", Proc. of 70th IEEE ECTC 2020 (Accepted)

S. S. Iyer, **S. Jangam** and B. Vaisband, "Silicon interconnect fabric: A versatile heterogeneous integration platform for AI systems," in IBM Journal of Research and Development, vol. 63, no. 6, pp. 5:1-5:16, 1 Nov.-Dec. 2019

S. Jangam, A. Bajwa, P. Ambhore and S. S. Iyer, "Fine Pitch ($\leq$10 μm) Direct Cu-Cu Interconnects using In-situ Formic Acid Vapor Treatment", Proc. of 69th IEEE ECTC 2019

S. Jangam, A. Bajwa, K. K. Thankappan, P. Kittur and S. S. Iyer, "Electrical Characterization of High Performance Fine Pitch Interconnects in Silicon-Interconnect Fabric", Proc. of 68th IEEE ECTC 2018, San Diego

A. A. Bajwa, **S. Jangam**, S. Pal, B. Vaisband, R. Irwin, M. Goorsky, S. S. Iyer, “Demonstration of a Heterogeneously Integrated System-on-Wafer (SoW) Assembly,” 2018 IEEE 68th Electronic Components and Technology Conference (ECTC), San Diego, CA, 2018

S. Jangam, A. Bajwa, K. K. Thankappan, and S. S. Iyer, “Characterization of Fine Pitch Interconnections (≤10μm) on Silicon Interconnect Fabric for Heterogeneous Integration”, Proc. of 51st IMAPS 2018, Pasadena

S. Jangam, S. Pal, A. Bajwa, S. Pamarti, P. Gupta and S. S. Iyer, “Latency, Bandwidth and Power Benefits of the SuperCHIPS Integration Scheme”, Proc. of 67th ECTC 2017, Orlando, FL, pp. 86-94

A. A. Bajwa, **S. Jangam**, S. Pal, N. Marathe, T. Bai, T. Fukushima, M. Goorsky, and S. S. Iyer. “Heterogeneous Integration at Fine Pitch (≤ 10 m) using Thermal Compression Bonding”, Proc. of 67th IEEE ECTC 2017, Orlando, FL, pp. 1276

CHAPTER 1

Introduction

1.1 Conventional System Integration

Today, mainstream system integration relies on the assembly of individually packaged dies and components on a conventional printed circuit board (PCB). The package is expected to mechanically support the die and protect it from the harsh environment. It also provides for a stable and controlled test environment. In addition, it acts as an intermediate layer to connect the die to the PCB. The PCB acts as a platform to attach multiple packaged dies along with other components and interconnect them to form a system. A schematic of a conventional assembly is shown in Fig. 1.1. This integration methodology has been successfully implemented over the past several decades but this strategy cannot sustain the performance demands of systems today [Iye16].

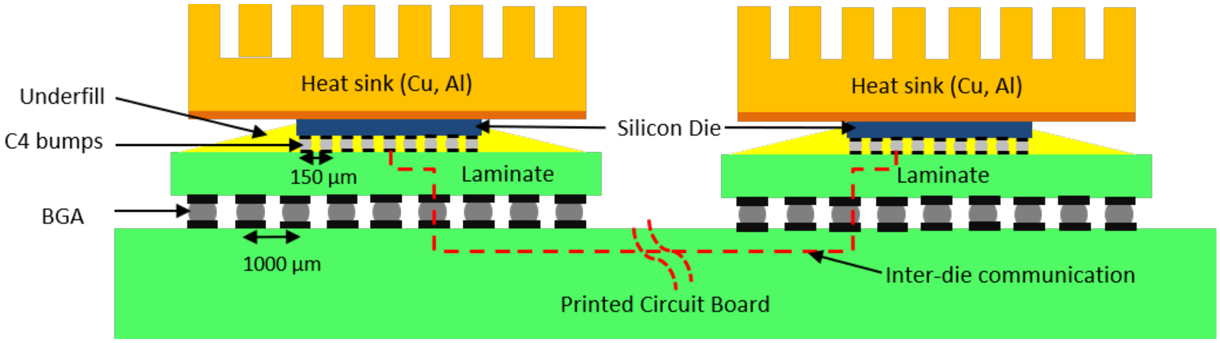


Figure 1.1: Schematic of a conventional assembly and inter-die communication.

Until recently, the demand for system performance has been met by incorporating more and more functionality into a single die, thanks to the aggressive Moore's law scaling of the semiconductor technologies. But with the apparent slow-down of Moore's law, it is no longer cost-efficient nor technologically feasible to scale down the devices further [RS11, Pow08]. Therefore, packaging technologies have to be scaled to ensure improvement in system performance. However, traditional packages and the PCBs are made of organic substrates that cannot use the same fabrication techniques used in semiconductor manufacturing. As a result, in the past four decades, the packaging dimensions have only scaled by 4-5X compared to the 1000X scaling in semiconductor technologies [Iye16]. In addition, the packages and boards use solder-based interconnects such as Controlled Collapsed Chip Connection (C4) bumps on the package at 100-150 μm pitch and Ball Grid Arrays (BGA) at 0.4-1 mm pitch [PS13, Int]. Solder extrusion, bridging, warpage of substrate, and so on limit the scaling of these solder-based interconnects. Consequently, the fine-pitch interconnects on the die (few microns) must be space-transformed to match the interconnect pitch on boards and then scaled back in the next chip leading to long links and inefficiencies in inter-dielet communication as shown in Fig. 1.1. Also, the disparity between the silicon and package dimensions constrains the number of input/output (I/O) connections for a die which in turn restricts the data-bandwidth.

On the other hand, according to Rent's rule [LFR05], as the functionality in a die increases, the number of I/Os increase according to (1.1), where T is the number of I/Os, g corresponds to the number of gates or functional blocks, and t, p are technology dependent constants. Using the transistor count for g, the number of I/Os required by today's processors are in the order of over 10,000. Fig. 1.2 shows the trend of minimum I/O pitch needed for processors [Conc, YW18, Qua19, Nvi18, Lea17] assuming simple peripheral I/Os, according to [IJV19]. From the plot, we observe that today's systems require $\leq$10 μm pitch interconnects. These interconnect pitches cannot be accommodated in traditional packages and therefore, the packages today are 5-18X larger than the dies [PPB+18]. The aforementioned limitations define the inter-die communication link lengths to be at least several millimeters with significant channel losses. Moreover, to accommodate all the I/Os and meet the data-bandwidth requirements, serialization-deserialization circuits (SERDES) are implemented. They have complex transceivers circuits that occupy substantial real estate on the die (up to 30% [Conb]) and consume significant power which can be 30-50% of the total system power [Iye16].

$$T = t * g^p \tag{1.1}$$

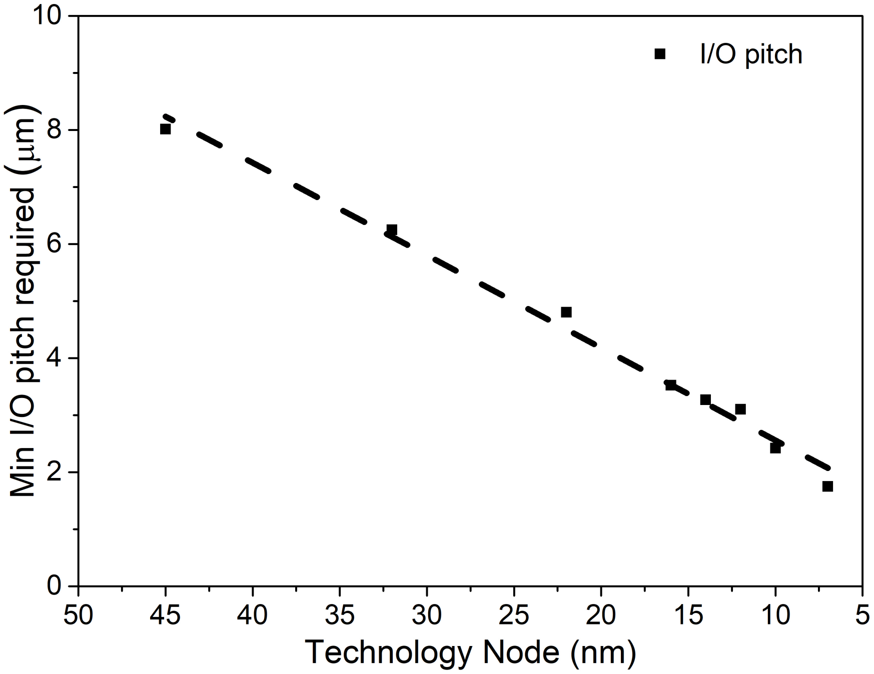


Figure 1.2: Trend of minimum I/O pitch required if no SERDES are used with the scaling of technology nodes for commercial processors [Conc, YW18, Qua19, Nvi18, Lea17].

Therefore, a paradigm shift in packaging technologies is necessary as it plays a more decisive role in determining system performance. As we move closer to data-centric computational systems, packaging technologies need to accommodate the ever-increasing data-bandwidth while simultaneously reducing latencies and communication power. It is no longer just a way to protect the die but rather a way to efficiently interconnect them and add value to the system.

1.2 Advances in System Integration

Over the past several decades, there have been significant advancements in integration methodologies to meet the growing demands of system performance. The integration methodologies can be divided into two categories namely monolithic integration and heterogeneous integration. Monolithic integration refers to the integration of functional blocks or subsystems on a single die. Heterogeneous integration refers to the integration of multiple dies from different technologies on a single platform. Some of the advances in both the monolithic and heterogeneous system integration technologies are described below.

1.2.1 Monolithic Integration

The first breakthrough in on-chip integration was the invention of monolithic integrated circuits where previously discrete transistors were fabricated and connected on a single chip. With the evolution of silicon (Si) fabrication techniques combined with Dennard's scaling theory [DGY$^+$74], made aggressive scaling realizable to achieve a transistor density of ≥100 million transistors/mm^2 in the recent technology nodes (5-7 nm) [Cona]. Besides the devices, the interconnect technology was also scaled to reduce delays using copper wiring levels. With the increase in wiring levels and a wiring hierarchy of fine-pitch wires (<100 nm) connecting neighboring nodes and larger pitch wires (few microns) connecting distant nodes, more functional blocks could be incorporated in a single die. These developments led to a proliferation of larger systems on a single die which is discussed below.

1.2.1.1 System-on-Chip

In the System-on-Chip (SoC) approach, several different functional or intellectual property (IP) blocks required for a system are integrated and fabricated on a single die as illustrated in Fig. 1.3. Availability of fine-pitch wiring and short inter-block spacings on a single chip provides opportunities for high bandwidth and energy efficiency. At first glance, it looks beneficial to integrate more functionality into a single chip and the recent trends of high-performance processors confirm this. The die size has increased to near reticle limits (≈830 mm^2) even though the technology nodes have scaled to have denser transistors. Further, the die size has been a significant factor (20%) apart from process technology (40%) in achieving performance scaling of 2X every 2.5 years in the past decade [Su19]. However, SoCs are extremely complex in design, require IP hardening, Si validation, and so on for every tape-out. This contributes to a high non-recurring engineering (NRE) cost and time to market. In addition, SoCs require large die size which reduces the yield of a die significantly, increasing the cost [SAB$^+$16]. Further, the reticle limit and the slow-down of Moore's law restrict the scalability of this approach. In addition, SoCs are inherently homogeneous in technology and cannot truly integrate different heterogeneous components of a system. As a result, SoCs are limited by the packaging technologies to communicate with other components such as memory.

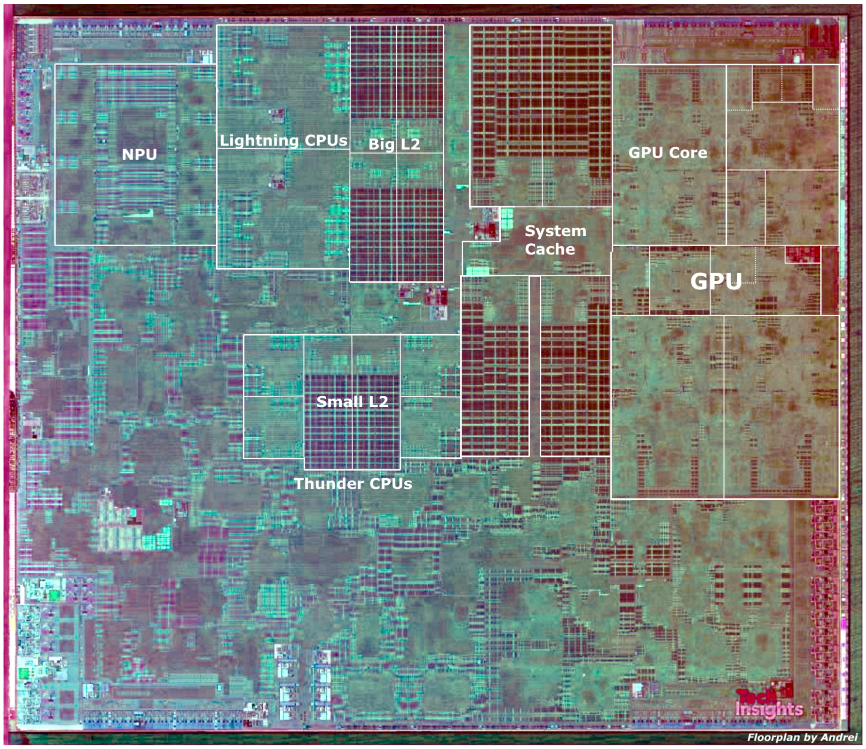


Figure 1.3: Floorplan of Apple A13 processor (SoC) showing different functional blocks integrated on a single monolithic die. (Picture source: [Fru19]).

1.2.1.2 Wafer-scale integration

Wafer-scale integration (WSI) takes the SoC approach to the next level by integrating a massive system on a single wafer. Early efforts were made in the 1980s by Gene Amdahl [MRRS84] to fabricate a fully functional wafer that achieves high communication bandwidth with reduced latency and power. However, technological limitations resulted in a low yield of the system, a situation SoCs are facing today. To improve yield, many redundancy schemes were implemented that increased the length of signal paths and consequently reduced the system speed, making the approach impractical. However, recently a 300 mm wafer-scale functional system for machine learning applications was demonstrated in [Sys]. Novel architectural schemes that use very tiny cores with abundant redundancies were implemented to ensure functionality by re-routing around defective cores. In addition, advanced fabrication and assembly techniques such as reticle stitching, novel connectors, heat extraction solutions were developed to build the system. Although this system shows a lot of promise, it is still a homogeneous system and probably limited by memory capacity. Also, such an architecture may not be suitable for all applications.

1.2.2 Heterogeneous Integration

Heterogeneous integration refers to using packaging technologies to integrate heterogeneous dies or components on a common substrate to overcome the bandwidth challenges on traditional PCBs [Soc]. Recently, there has been a lot of traction in heterogeneous integration because of several promising substrate technologies [MSP+16, HSF+16, CHT+17, OOS+14], and reduced design & cost overhead compared to SoCs. By dividing a large SoC into small chiplets or dielets, the yield of individual dies is improved [PPKG20], corresponding to lower costs. Further, since most of the SoCs have up to 80% of reused IP, the design complexity is simplified, saving time [Gre16]. If these dies are integrated on a heterogeneous platform hopefully with a minimal performance overhead, system scaling can be ensured. Several works [PPT+19, PPKG20, PPB+18, SIL+15, Soc, Su19] have discussed and demonstrated the benefits of heterogeneous integration on different aspects of system performance, architectures, and overall scaling. Some of the heterogeneous integration technologies are listed below.

1.2.2.1 Multi-Chip Modules

Multi-chip-modules (MCMs) were initially developed in the 1980s to integrate a few dies on ceramic-based substrates for high-performance mainframe systems. In MCMs, multiple dies are packaged laterally on a common substrate such as laminate and integrated at finer pitch (<100 μm) than boards [Lau17]. A schematic of the MCM structure is shown in Fig. 1.4. Today, several commercial products are available with multiple dielets in different technologies integrated on organic boards [ABC$^+$17, Su19]. Using the concepts of heterogeneous integration and IP reuse, design and manufacturing costs are significantly reduced (up to 41%) using MCMs [Su19]. Also, recently systems-in-package (SiP) technologies were developed to integrate dielets and packages both laterally on a laminate and vertically using three-dimensional (3D) stacking to form a system or a subsystem. However, both these technologies have low interconnect density and are limited to few dies on a package.

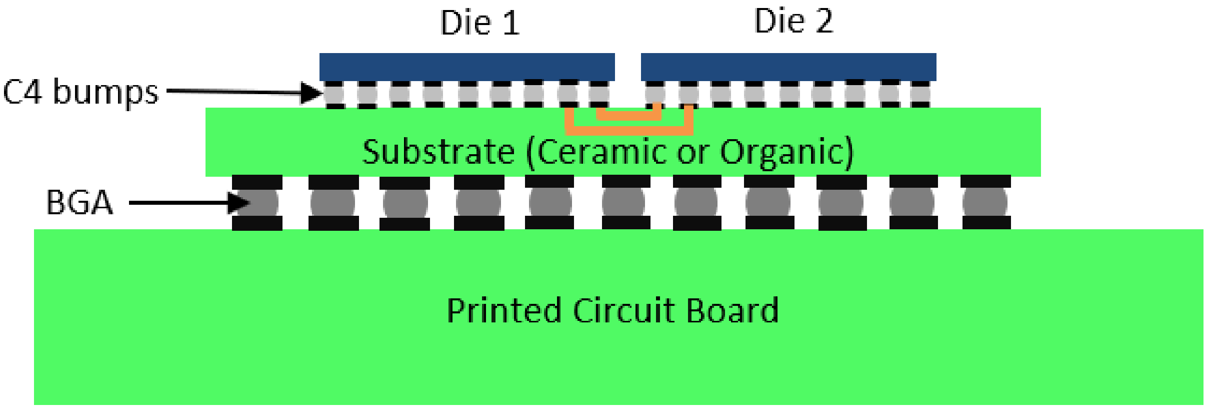


Figure 1.4: Schematic of an MCM package with two dies integrated on a common substrate.

1.2.2.2 Interposers

Interposer technology has been presented as a high interconnect density redistribution layer (RDL) between dielets [LLKK18, CHT+17]. An interposer adds an additional hierarchy level in packaging, where few dies are interconnected using on-chip like wires ($\leq$4 μm pitch) and the interposer assembly is packaged and assembled on PCBs. A schematic of an interposer is shown in Fig. 1.5. Silicon is typically used as the substrate, although, other substrates like glass and organic interposers were also proposed [HSF+16, OOS+14]. Although the wiring pitch is small in interposers, the die-to-substrate interconnect pitch is 40-55 μm, because of the solder-capped Cu pillar interconnects, also called μ-bumps [MSP+16, CHT+17]. However, today's bandwidth requirements demand an interconnect pitch of $\leq$10 μm as shown earlier in Fig. 1.2. Also, the size of the interposer is limited to the reticle size ($\approx$830 mm^2) without stitching. Even though larger interposer of 1700 mm^2 was demonstrated in [Shi20] using reticle stitching, the process becomes extremely complicated and expensive when extended to full wafer which will be discussed in section 2.2.2. Interposers are also thinned to <100 μm to add through-silicon vias (TSV) to connect to the package. The thinned interposers have significant warpage of several microns [MAH+13] limiting the scalability of both the interposer size and the μ-bump pitch. To minimize warpage, Chip-on-Wafer-on-Substrate (CoWoS) technology [CHT+17] uses a thick silicon substrate for assembly of dies. The interposer is back-grinded after dielet assembly which is associated with several reliability concerns. Finally, interposers inflate the overall packaging cost by adding an additional level in the packaging hierarchy [Iye16].

Other approaches use a hybrid of MCM and interposer technology. Intel developed Embedded Multi-die Interconnect Bridge (EMIB) technology [MSP+16] to embed silicon bridges with fine-pitch wires into a package substrate that connects neighboring dies. The proposed advantage is that selective fine-pitch interconnects can be placed at required locations and traditional coarse pitch wiring may be used for the rest of the system to reduce cost. However, the complexity of the bonding process of the die to different interconnect pitches on the silicon bridges and the substrate while ensuring planarity and yield remains

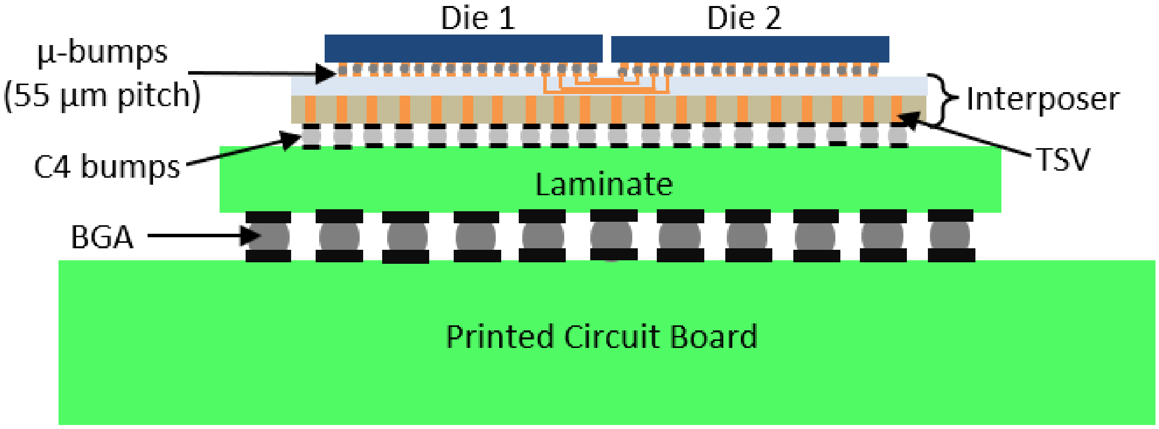


Figure 1.5: Schematic of a dielet assembly on an interposer. The interposer connects the dies at a moderate interconnect pitch (≈55 μm) but adds an additional level in the packaging hierarchy. TSVs are used to transfer the signals to the dies.

quite high. Finally, the use of solder-based μ-bumps in interposers limits the scalability of the interconnect pitch. There is also another proposal to integrate dies with a high density (up to 4 μm pitch) redistribution layer using an approach called fan-out wafer-level packaging (FOWLP) [TLWY16]. In this approach, the dies are first assembled on a handler wafer and the redistribution layer is fabricated on top of the dies creating the package. It is later terminated with C4 bumps for subsequent assembly on boards. Again, the use of organic materials and molding compounds limit the scalability of this approach for wafer-scale systems.

1.2.2.3 3D Integration

Another approach to integration is to vertically stack the dies on top of each other. This solution offers a reduced form-factor and a wide I/O interface between dies using TSVs. A schematic of 3D integration is shown in Fig. 1.6. 3D stacking is done either at wafer-level using wafer-to-wafer bonding or die-level using die-to-die or die-to-wafer bonding. Wafer-to-wafer bonding offers fine interconnect pitches of $\leq$10 μm. The bonded wafers are

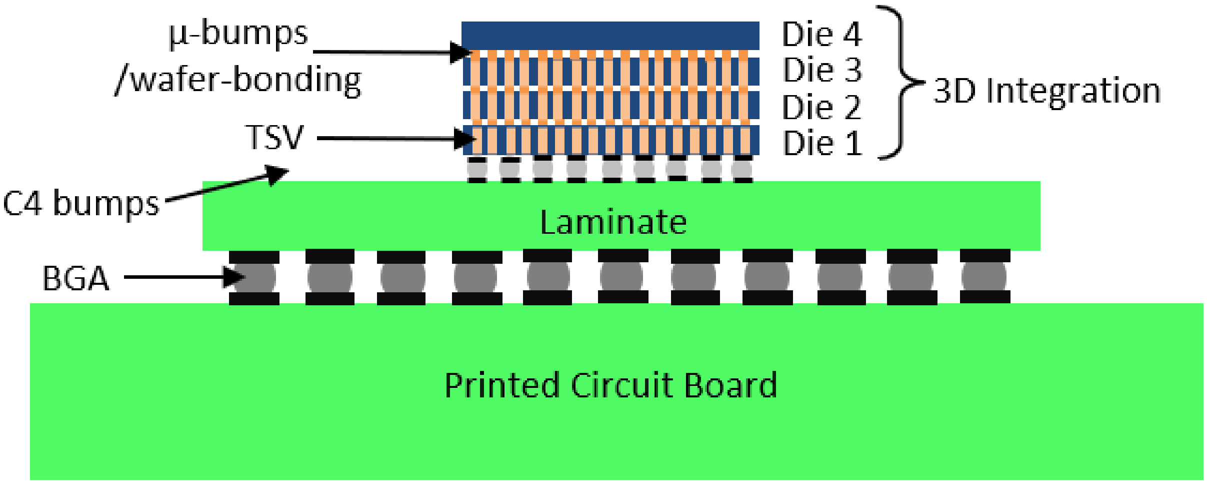


Figure 1.6: Schematic of a 3D stack of 4 dies. The dies are bonded using either μ-bumps or wafer-to-wafer bonding. TSVs are used to transfer signals and power across dielets.

subsequently diced after assembly of the whole stack. But one serious disadvantage of this approach is that a defective area of one wafer can get bonded to functional areas on the second wafer which decreases yield. Therefore, for most applications, including memory stacking, logic-on-memory, and logic-on-logic stacking, die-to-die or die-to-wafer bonding is preferred. 3D stacking of several thinned dies has been successfully implemented especially for memory dies for up to a stack of 12 dies [KAD+08, BAB+06, Loh08, Shi19b]. Intel has demonstrated logic-on-logic stacking called Foveros [IAA+19], and several others have explored logic-on-memory [IK15, LGBS05]. However, logic-die stacking lacks widespread adoption primarily due to thermal and I/O considerations. At the bottom of the stack, the heat from logic-dies cannot be efficiently extracted, limiting the thermal budget. On the other hand, at the top of the stack, logic-dies require a large number of TSVs for I/Os. The corresponding TSV real estate and keep out zone required for proper device functionality inflates the overall die size. Although this technology is not scalable for large-scale systems, it can complement other heterogeneous integration technologies to improve performance.

1.3 Objective of this Work

As discussed earlier, systems today demand on-chip like fine-pitch interconnects ($\leq$10 μm) to meet the ever-growing bandwidth requirements. It has become increasingly clear that a monolithic-style integration is not scalable and cannot sustain the increasingly complex system architectures. Today, there is a need for paradigm shift in packaging technologies to cater to the demands of next-generation systems. A heterogeneous integration approach needs to adapt to provide the same performance and efficiencies supported by the SoCs. Moreover, it must be highly scalable to integrate massive systems consisting of several thousands of dies. In addition, it should provide simple interfaces for inter-die communication to achieve high-performance with minimal overhead. It is an added advantage if the packaging platform is compatible with existing technologies and provides opportunities for the development of novel energy-efficient high-performance architectures.

Considering all the above requirements, in this work, a fine-pitch, highly scalable, package-less, heterogeneous integration platform called the Silicon-Interconnect Fabric (Si-IF) is investigated. The fabrication and assembly processes necessary for such a fine-pitch ($\leq$10 μm) platform are developed. In addition, the benefits of the Si-IF style integration for system performance are demonstrated.

1.4 Organization of this Dissertation

This thesis is organized as follows: Chapter 2 introduces the Si-IF technology and contrasts it with existing packaging technologies. The fabrication process of the Si-IF platform is illustrated in Chapter 3. Chapter 4 describes the fine-pitch dielet assembly process on the Si-IF. The Simple Universal Parallel intERface for Chips (SuperCHIPS) interface protocol is introduced in Chapter 5 along with the experimental characterization, and circuit simulation results. The experimental demonstration of the SuperCHIPS protocol using functional dielet assembly on the Si-IF is presented in Chapter 6. Chapter 7 discusses the benefits of the Si-IF assembly and the SuperCHIPS protocol compared to existing technologies. Finally, the conclusion and future outlook of this work are given in Chapter 8.

CHAPTER 2

Silicon-Interconnect Fabric Technology

The Silicon-Interconnect Fabric (Si-IF) is a novel heterogeneous integration technology that is package-less, fine pitch (≤10 μm), and highly scalable. In the proposed integration scheme, bare dielets are assembled on a silicon substrate at close proximity (≤100 μm) and interconnected at SoC-like wiring pitches. Unlike interposers, the Si-IF technology is developed to replace the PCB and integrate the system on a single packaging hierarchy. Further, the Si-IF leverages the established techniques developed for the mature complementary metal-oxide-semiconductor (CMOS) technology and applies them to the realm of packaging. In this chapter, the Si-IF technology is introduced and the benefits of integration on the Si-IF are discussed.

2.1 Technology Description

The Si-IF consists of a silicon-based substrate with CMOS back-end-of-the-line (BEOL) wiring levels. The number of wiring layers can be up to 4 and there is no fundamental limitation to extend it further. These wiring levels match the top-level fat-wiring layers on-chip. The Si-IF is terminated with ≤10 μm copper (Cu) pillars of diameter ≤5 μm that act as interconnects between the die and the wiring levels on the Si-IF. As a result, the interconnects seamlessly integrate heterogeneous systems to match SoC interconnect density. Bare dies terminated with either Cu or gold (Au) pads are used for direct assembly on the Si-IF substrate. This ensures compatibility with existing CMOS dies that have Cu wiring levels and III-V dies that have Au pads. These bare dies are bonded to the Si-IF using a solder-less metal-metal (Cu-Cu or Au-Au) Thermal Compression Bonding (TCB) process between the metal pillars on the Si-IF and the metal pads on the dies. As a consequence of the elimination of solder, the interconnect pitch is scaled to ≤ 10 μm. Consequently, due to the fine interconnect pitch, more number of I/Os can be incorporated for inter-die communication which significantly increases the data-bandwidth (4-23X). Moreover, by eliminating individual packages, the dies can be assembled at close proximity of ≤ 100 μm. As a result, the near chip communication links are short (50-500 μm) corresponding to low channel losses (≤ 2 dB) and link latencies (≤ 20 ps). Therefore, simple inverters can be used for data-transfer using a Simple Universal Parallel intERface for chips (SuperCHIPS) protocol to reduce communication power tremendously (5-40X). This will be elaborated in Chapter 5. A schematic of the Si-IF assembly is presented in Fig. 2.1.

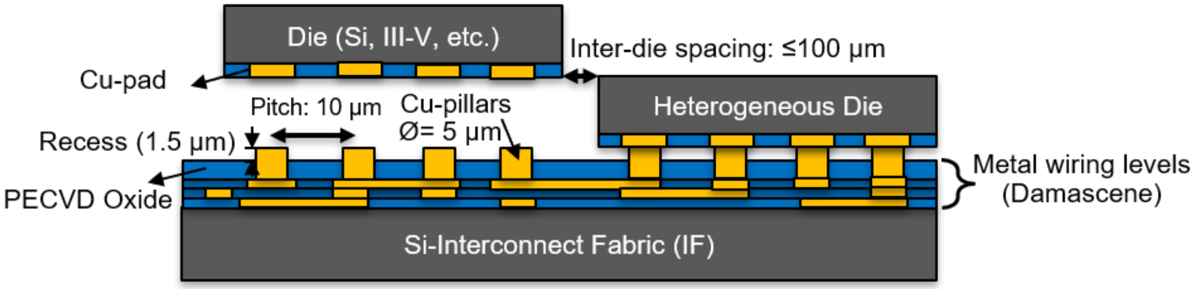
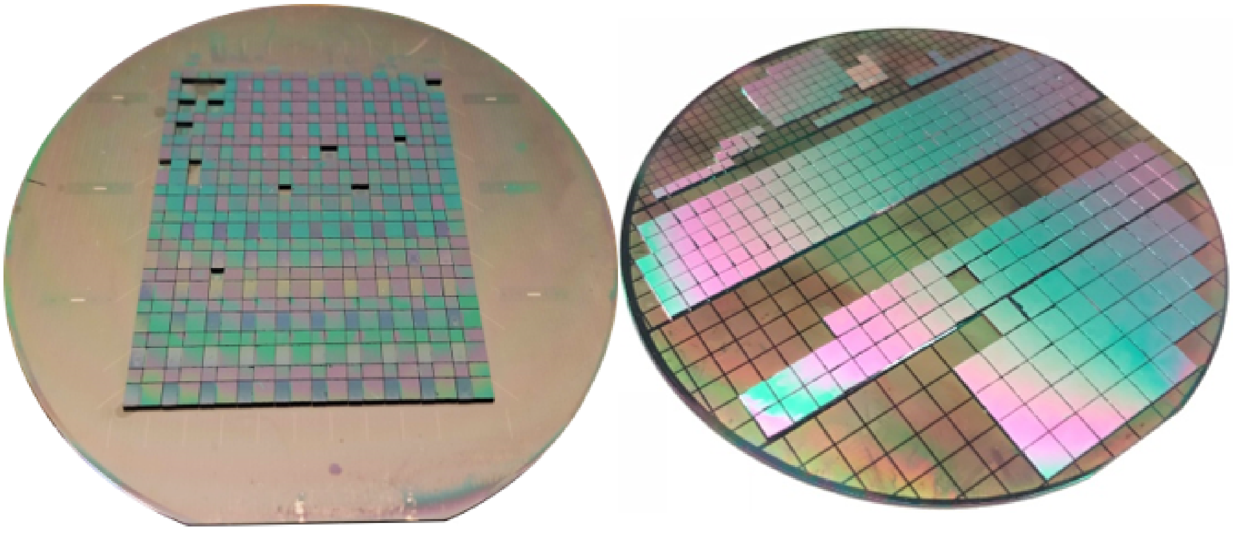


Figure 2.1: Schematic of the fine-pitch assembly on the Silicon-Interconnect Fabric (Si-IF).

Moreover, the Si-IF platform is agnostic to the dielet technology including 3D-stacked dies and passives, therefore allowing for heterogeneous integration. Also, the Si-IF substrate is highly scalable and integrates dies on a silicon wafer up to a diameter of 300 mm. Therefore, the Si-IF provides a platform to integrate a massive wafer-scale system constituting of small heterogeneous dies (2-10 mm [IJV19]). Demonstrations of the wafer-scale assemblies of heterogeneous dies on the Si-IF are presented in Fig. 2.2. Fig. 2.2 (a) shows the integration of 460 heterogeneous dies consisting of 113 3x3 mm^2 dies, 237 3x2 mm^2 dies, and 110 2x2 mm^2 at ≤ 100 μm inter-dielet spacing on a 100 mm Si-IF corresponding to an active area of 2900 mm^2. Fig. 2.2 (b) shows the integration of 371 heterogeneous dies of sizes 1x1 mm^2, 2x2 mm^2, 3x3 mm^2, 4x4 mm^2, and 5x5 mm^2 at ≤ 100 μm inter-dielet spacing on a 100 mm Si-IF, corresponding to an active area of >3100 mm^2.



(a)

(b)

Figure 2.2: Wafer-scale assemblies on the Si-IF. (a) 460 heterogeneous dies (3x3 mm^2, 3x2 mm^2, and 2x2 mm^2) at $\leq$100 μm inter-dielet spacing on a 100 mm Si-IF corresponding to an active area of 2900 mm^2. (b) 371 heterogeneous dies (1x1 mm^2 to 5x5 mm^2) at $\leq$100 μm inter-dielet spacing on a 100 mm Si-IF, corresponding to an active area >3100 mm^2.

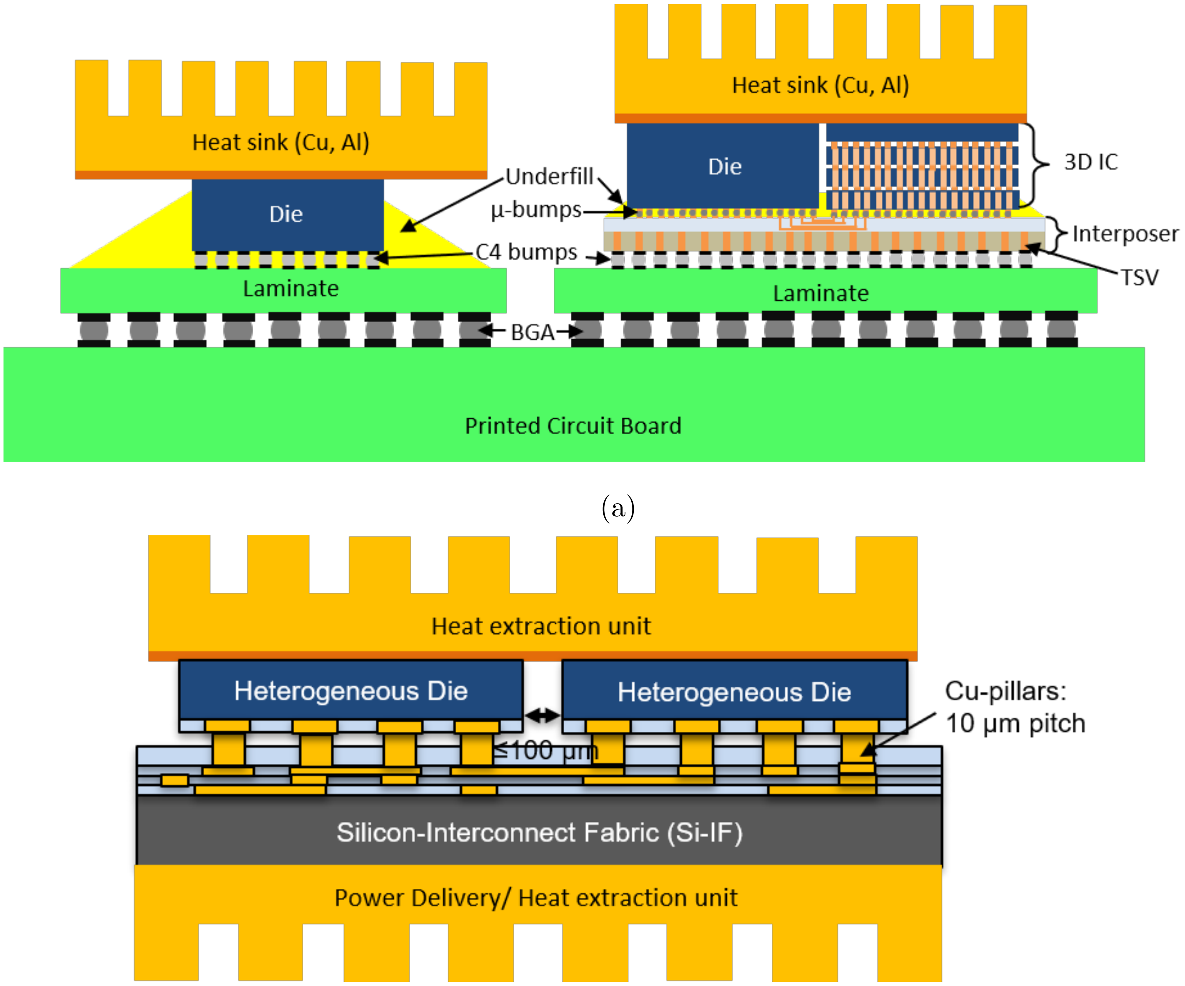
2.2 Comparison with Conventional Technologies

Unlike traditional PCB-based integration where individual chips or sub-systems are packaged and integrated, the Si-IF technology aims to integrate an entire system on a single platform. A schematic of conventional integration schemes and the Si-IF integration is shown in Fig.2.3. In this section, the merits and challenges of the Si-IF technology are discussed and contrasted with conventional technologies.

2.2.1 Merits of the Si-IF Technology

Some of the key merits of the Si-IF technology are listed below.

- Conventional integration schemes have multiple levels in the packaging hierarchy as shown in Fig. 2.3 (a). At each hierarchy level, the interconnect pitch and wiring di-



(a)

(b)

Figure 2.3: Schematic of a comparison of conventional assemblies with the Si-IF assembly. (a) Conventional assembly illustrating traditional die in a package and dies including 3D-stacks assembled on an interposer mounted on board. The schematic shows the different packaging hierarchies. (b) Fine-pitch Si-IF assembly with single packaging hierarchy.

mensions are vastly different (10X). Also note that interposers, although interconnect systems at moderate interconnect densities (40-55 μm), add an additional level in the packaging hierarchy as illustrated in Fig. 2.3 (a). Contrary to these schemes, the Si-IF technology simplifies the packaging hierarchy by integrating the entire system in a single packaging level as shown in Fig. 2.3 (b). Therefore, Si-IF is a wafer-level integration solution to interconnect the entire system at fine wiring pitches ($\leq$2 μm).

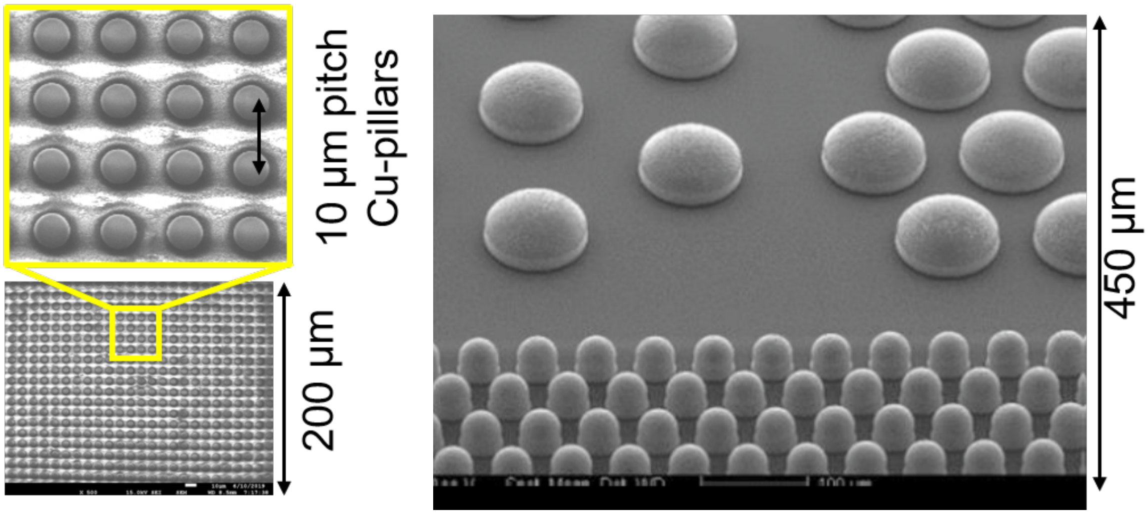
- Today, packaging technologies use many disparate materials such as Si, Cu, FR-4, solder, molding compound, underfill, and so on. These organic and inorganic materials have different thermo-mechanical properties, especially the coefficient of thermal expansion (CTE), thermal conductivity, young's modulus, etc. The mismatch of these material properties impacts the device performance which may lead to failures, commonly known as Chip-Package-Interaction (CPI) related failures. The Si-IF technology, on the other hand, uses a simple material system of Si, Cu, and silicon oxide (SiO_2) that matches the material constituents on the die. Therefore, the mismatch between the dies and the integrating Si-IF platform is significantly low, reducing the CPI-related failures.

- The Si-IF technology uses standard full-thick (500-770 μm) silicon wafer which is mechanically rigid when compared to organic substrates or thinned-interposers. The CTE of silicon is 2.6 ppm/K while that of organic laminates (FR-4) is 14-70 ppm/K. As a result, the warpage of a die-on-wafer assembly is only a few microns while that of a die on thinned-interposer is >33 μm and a die on organic substrate is >200 μm [MAH+13]. The reduction in the warpage helps in reducing the interconnect pitch and dimensions. Moreover, Si is extremely robust with a higher Young's modulus (140 GPa) than steel, although it is brittle.

- The silicon substrate is also an excellent heat spreader with a thermal conductivity of 149 W/mK which is just 3X lower than Cu. This is 600X higher than typical organic substrates like FR-4 with thermal conductivity of 0.25 W/mK [AG96]. This helps in heat sinking and spreading allowing for a higher thermal budget in designing systems for high-performance [PPB+18].

- Using silicon as the packaging material allows us to apply the mature fabrication techniques developed for CMOS processing to easily achieve fine wiring dimensions of $\leq$2 μm. Compared to the dimensions of an organic substrate, this is 50-100X smaller and comparable to on-chip wiring dimensions on the top metal wiring levels.

- Because of all the reasons mentioned above, dies can be assembled on the Si-IF at $\leq$10 μm interconnect pitch using direct metal-metal TCB. As a result, the I/O pad real estate on the die is significantly reduced by 37X and 240X compared to interposer and PCBs respectively, which is discussed in detail in chapter 7. Scanning electron microscope (SEM) images of the Si-IF Cu-pillar interconnects, and μ-bumps and C4-bumps on a die on the same scale are presented in Fig. 2.4. One can observe the difference in the dimensions and pitch between the different interconnects. Moreover, the Si-IF Cu pillars are 5 μm thick with only 1.5-2 μm protruding above the surrounding dielectric. In contrast, μ-bumps are typically 20-30 μm tall with 10-20 μm solder cap and C-4 bumps have >50 μm thick solder balls. In addition, the Si-IF has Cu pillars on the substrate instead of the die which simplifies the die processing, warpage, and improves die yield.



(a)

(b)

Figure 2.4: Comparison of the Cu-pillar interconnects on the Si-IF with the μ-bumps and C4-bumps on a die depicted to scale. (a) Si-IF with 475 Cu pillar interconnects at 10 μm pitch. (b) Intel Stratix die with 44 solder-capped Cu pillar μ-bumps, and 10 C4-bumps for assembly on the EMIB package (Picture source: [MSP$^+$16]).

- Direct Cu-Cu and Au-Au TCB allow for the integration of almost all dielet technologies such as Si and III-V dies with almost no change in the die manufacturing process. III-V dies typically use Au pads and can directly be assembled on the Si-IF using Au-Au TCB. Si-dies use Cu wiring levels but are terminated with aluminum (Al) pads for solder bumping. Instead, Si-die processing can be stopped at the last Cu wiring level before the Al layer and assembled on the Si-IF using direct Cu-Cu TCB. Moreover, Si-IF technology eliminates the need for under bump metallurgy (UBM), reducing fabrication overhead and costs [BL07].

- The Si-IF technology is also legacy compatible with traditional solder-based dies and surface mount components. Fig. 2.5 shows a system of conventional dies and passive components with solder pads assembled on the Si-IF. This was achieved using a thin solder capping layer of 150 nm nickel (Ni) and 350 nm tin (Sn) on the Cu pillars. Note that although this assembly achieves better performance than PCB integration, it cannot exploit all the performance benefits offered by the Si-IF technology due to pitch limitations on the die.

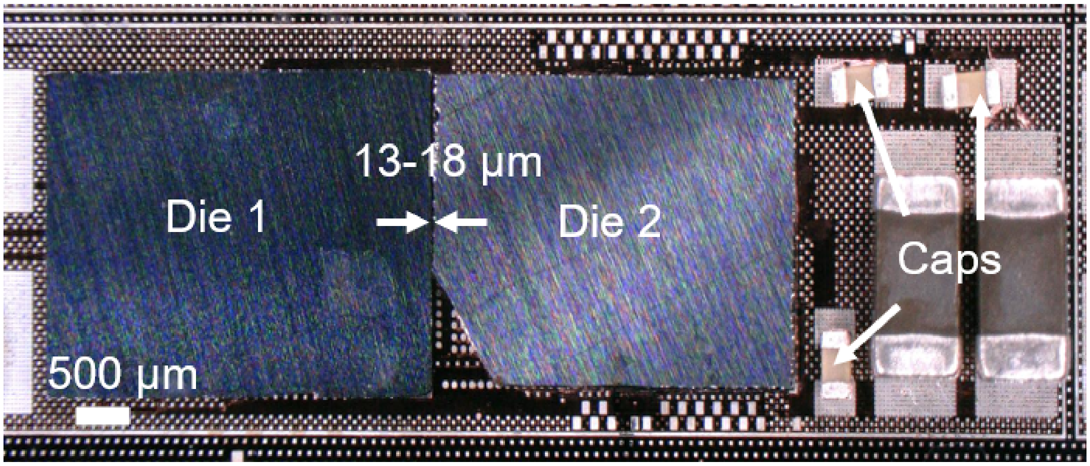


Figure 2.5: Micrograph of an assembly of two test chips with solder termination and discrete surface mount capacitors on the Si-IF.

- The Si-IF uses a minimum wiring pitch of 2-4 μm which is considered coarse for any mature Si technology such as 90 nm, or 65 nm. Therefore, the yield of the Si-IF is very high $\geq$90% [PPT+19]. To estimate the yield, consider the yield formula in (2.1) [ITR07].

$$Yield = (1 + \frac{D_0 * F_{crit} * A}{\alpha})^{-\alpha} \tag{2.1}$$

where D_0 is the defect density, F_{crit} is the fraction of the critical area, A is the total area, and α is the defect clustering factor.

The D_0 for mature CMOS technologies is 2x10^{-3} cm^{-2} [ITR07] which includes all the different layers. The individual layer defect density may be estimated by dividing the D_0 with the number of layers. Moreover, the vast majority of these defects occur in the transistor layer or the first few metal layers with fine pitch wiring (<200 nm). Therefore, the coarse pitch on the Si-IF should be accounted for by reducing the defect density in an appropriate proportion of minimum dimensions as given in [SXSL17]. Therefore, a conservative estimate of the defect density per layer in Si-IF is $\approx$1x10^{-5} mm^{-2}. Further, because the critical interconnects are between tightly spaced dies, the F_{crit} is very small, typically 1%-10%. The value of α is typically between 1 and 3 [ITR07]. So assuming an α of 2, and an effective area of 50,000 mm^2 for 300 mm wafer and 5000 mm^2 for a 100 mm wafer, the yield of the Si-IF is estimated and presented in Table.2.1. The yield drops dramatically if active devices are fabricated on the silicon for the case of wafer-scale SoC integration [KJL15]. Apart from fabrication yield, assembly yield is more important. Fine-pitch die-to-substrate assemblies have been demonstrated with high yield (>99.99%) [MSP+16, CHT+17, Shi19b] and the Cu-pillar bonding is also expected to be >99% with the limited data from the experiments in this work. With the use of pre-tested known-good-dies (KGD) [Lau10], the overall system yield is improved compared to monolithic SoCs. In addition, simple redundancy strategies can be implemented to ensure functionality for the entire wafer-scale system on Si-IF.

- The concept of using Si as a substrate appears costly at first. However, one should note that the majority of the cost comes from the processing of devices and fine features on the Si rather than the substrate itself. For a passive Si substrate with coarse

Si-IF diameter (mm)	**Critical Area (%)**	**Yield per layer (%)**	**Yield for 4 layers (%)**
100	1	99.95	99.80
	10	99.5	98.02
300	1	99.5	98.02
	10	95.18	82.07

Table 2.1: Estimated yield of passive Si-IF wafers.

pitch wiring, the costs are significantly low. Typical passive interposers cost \$500-\$650 per 300 mm wafer of which only 22% comes from damascene processing [Cad07]. Therefore, 300 mm Si-IF should cost <\$250 per wafer. This is also considerably lower compared to the high-performance PCBs which typically cost a few hundred to thousands of dollars for much smaller systems [BL07]. In addition, integrating smaller dies with SoC-like wiring on the Si-IF will tremendously improve yield with little or no penalty on performance. Some of the cost-benefit arguments presented in [SXSL17, SAB^{+}16] for chiplet based designs compared to SoCs are also valid for systems on the Si-IF. However, the assembly cost of the Si-IF can be a considerable amount because of the fine-pitch bonding. Although the fine-pitch bonding process described in chapter 4 has a low bonding cycle time of ≤30 s, it is relatively high (>10X) compared to coarse-pitch solder-based assembly processes. But eliminating the package, UBM, and other processes in the Si-IF technology reduces the cost of many processors by 30-50% [BL07]. Therefore, the impact of Si-IF technology on cost is more pronounced for high-performance wafer-scale assemblies and is arguably competitive even for low-performance or low-cost systems.

2.2.2 Limitations and Challenges

Every technology comes with several challenges and certain limitations. Some of the major challenges and limitations of the Si-IF technology are listed below. Although this list is not exhaustive, it presents an overview of the different types of challenges and strategies to solve them. A conceptual model of the overall system on the Si-IF is illustrated in Fig. 2.6. The figure shows some of the key enablers in realizing a wafer-scale assembly on the Si-IF.

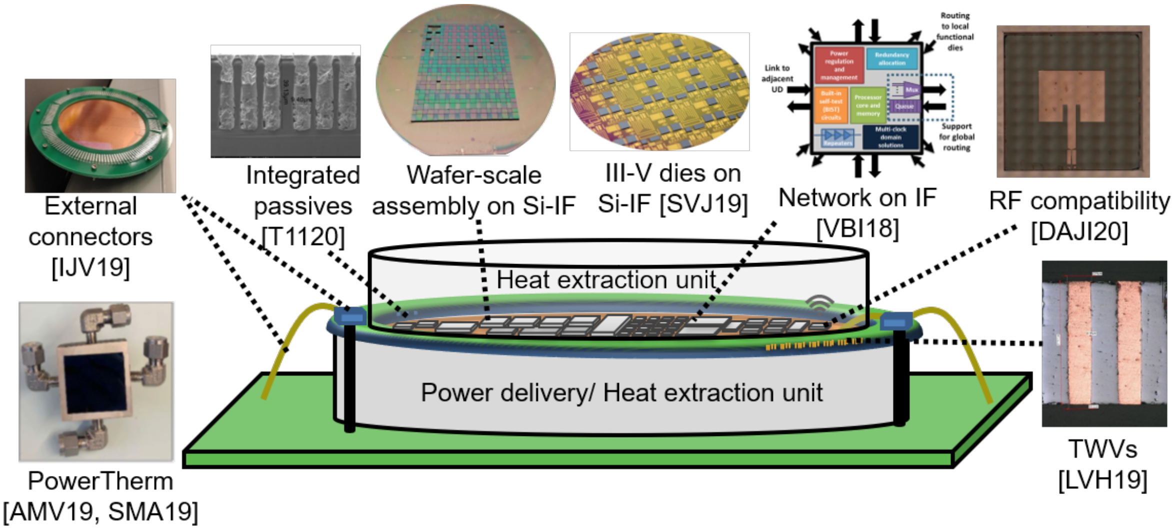

External connectors [IJV19]
Integrated passives [T1120]
Wafer-scale assembly on Si-IF
III-V dies on Si-IF [SVJ19]
Link to adjacent UD
Power regulation and management
Redundancy allocation
Built-in self-test (BIST) circuits
Processor core and memory
Mux
Queue
Repeaters
Multi-clock domain solutions
Routing to local functional dies
Support for global routing
Network on IF [VBI18]
RF compatibility [DAJI20]
Heat extraction unit
Power delivery/ Heat extraction unit
PowerTherm [AMV19, SMA19]
TWVs [LVH19]

Figure 2.6: Schematic representing an overview of the Si-IF technology. Some of the key enablers are highlighted including wafer-scale assembly, heterogeneous dies (III-V) on Si-IF [SVJ$^+$19], global communication network on Si-IF (NoIF) [VBI18], power delivery and heat extraction using PowerTherm [AMV$^+$19, SMA$^+$19], external connectors [IJV19, DAJI20], through wafer vias (TWVs) [LVH$^+$19], and integrated passives [TI20].

- A wafer-scale system requires the underlying routing layers on the Si-IF to accommodate any design across the wafer. This is not possible by using a step-and-repeat of reticle masks done today for 300 mm wafers. Stitching of different adjacent reticle masks has been successfully implemented [Shi20], however, it cannot be scaled to a wafer-scale because of the sheer number of masks required. Recently, maskless lithography techniques using a laser with digital micromirror device (DMD) are gaining traction for fabricating coarse features of >1 μm line/space [ZLW09]. These techniques provide adaptable routing at wafer-scale which is required for Si-IF technology.

- The assembly process on the Si-IF requires a cleanroom environment which will be elaborated in Chapter 4. This is not a typical requirement for package assemblies, though increasingly common. In addition, the dielet and substrate handling must be carefully monitored and pre-cleaning may be implemented to ensure good assembly yield. Other challenges of fine-pitch dielet assembly will be discussed in section 4.5.4. Moreover, unlike solder-based interconnects, direct metal-metal TCB interconnects are not reworkable. If one dielet fails either during assembly or during runtime, it cannot be physically replaced. This severely limits the repairability and serviceability of the Si-IF system. It also tightens the assembly yield requirements to ensure functionality. Therefore, redundancy schemes are essential that can re-route not only around faulty links but also around faulty dies. Note that this limitation also exists in interposer and 3D integration technologies that use solder-cap μ-bumps.

- Testing and probing of the ≤ 10 μm pitch die pads, to isolate the KGD, is challenging because of both the pad dimensions ($\leq$7x7 μm^2) and the damage of pad morphology [KAB+05]. Dedicated larger sacrificial pads may be used on the die for probing to test limited functionality. Moreover, testing of the system after assembly is also difficult because of the enormous number of interconnects. Therefore, novel built-in-self-test (BIST) strategies must be implemented to significantly reduce the testing time.

- Assembly of conventional passive components on the Si-IF cuts into the compute area as shown in Fig. 2.5, reducing the compute density. Integrated passives on the Si-IF [TI20] will minimize if not eliminate the need for passive components. In addition, these passive components should be also incorporated in supporting platforms instead of just the Si-IF wafer.

- Apart from assembly and near-chip communication, long-reach communication, and communication with external systems are required for any technology. The use of a lossy silicon substrate presents several challenges to long-reach communication that will be discussed in chapter 7. Authors in [VBI18] discuss some of the novel protocols and integration strategies that need to be adopted for global communication on the Si-IF. Methodologies and processes for radio frequency (RF) communication compatibility on the Si-IF were discussed in [DAJI20]. In addition, an external connector interface to the Si-IF systems must be developed that is compatible with conventional I/O connectors. These connectors should not only serve as electrical links but also sustain thermo-mechanical stresses when interfacing with the silicon substrate.

- Power delivery and heat extraction are arguably one of the major challenges of wafer-scale systems. With the increase in compute density, the power density also increases to >1 W/mm^2, reaching total power values exceeding 50 kW for a 300 mm wafer [AMV$^+$19,SMA$^+$19]. This amount of power cannot be delivered by just peripherical pads due to high voltage (IR) drop and the corresponding I^{2}R losses. Therefore, power must be delivered from the backside using through-wafer-vias (TWVs) as demonstrated in [LVH$^+$19]. Unlike TSVs that are used for both signal and power transfer, TWVs are primarily used for power delivery. Accordingly, they are at much coarser pitch (100 μm diameter and 200 μm pitch) and traverse full wafer thickness (500-700 μm) [LVH$^+$19]. Moreover, novel power delivery structures, similar to [AMV$^+$19], need to be developed to deliver such large amounts of power. In traditional packages, a large heat sink is installed on the chip that helps in spreading the heat and reducing the heat flux density. However, for a compact system on the Si-IF, the heat flux densities are much higher and conventional forced-air or liquid-based cooling techniques are insufficient. As a result, novel heat extraction strategies, such as two-phase cooling in [SMA$^+$19] must be implemented to extract such enormous heat flux densities.

- Since the dies are package-less, and no underfill or molding compounds are used during assembly, a system-level passivation scheme is vital. This passivation should not only protect the dies and the Si-IF, but also passivate the fine-pitch metal-metal interconnects. Some of these passivation techniques were discussed in [SHI19a, SSYI20]. In addition, several other reliability concerns, particularly at the transition of the Si-IF to other structures, must be studied carefully. Further, although the silicon substrate has high yield strength, it is brittle and requires mechanical support for practical use. Some of the above-mentioned structures could also serve as mechanical support for the assembly.

2.3 Scope of this Work

As mentioned earlier, there are several key enablers for successful wafer-scale assembly of a high-performance system on the Si-IF including the substrate technology, assembly process, power delivery, heat extraction, near-chip communication on the Si-IF, long reach communication strategies on the Si-IF, and novel wafer-scale systems and architectures. Each of these enablers has its unique solutions and challenges. Addressing all the issues would require considerable human resources, facilities, and time which is beyond the scope of this dissertation. Therefore, this dissertation focuses on certain aspects of the Si-IF technology that are considered fundamental. These include developing the fabrication techniques for the Si-IF substrate, developing the die-to-substrate assembly process, characterization of near-chip communication i.e. the SuperCHIPS protocol, and finally, demonstration of a functional system on the Si-IF platform integrated using the SuperCHIPS interface. This work is a first step in demonstrating the technological viability and the performance advantages of the Si-IF technology.

CHAPTER 3

Si-IF Fabrication

The fundamental requirement for an advanced packaging technology is the availability of a highly interconnected fine-pitch substrate. In this chapter, the fabrication process of the Si-IF substrate is described and the results of the fabricated Si-IF samples are presented.

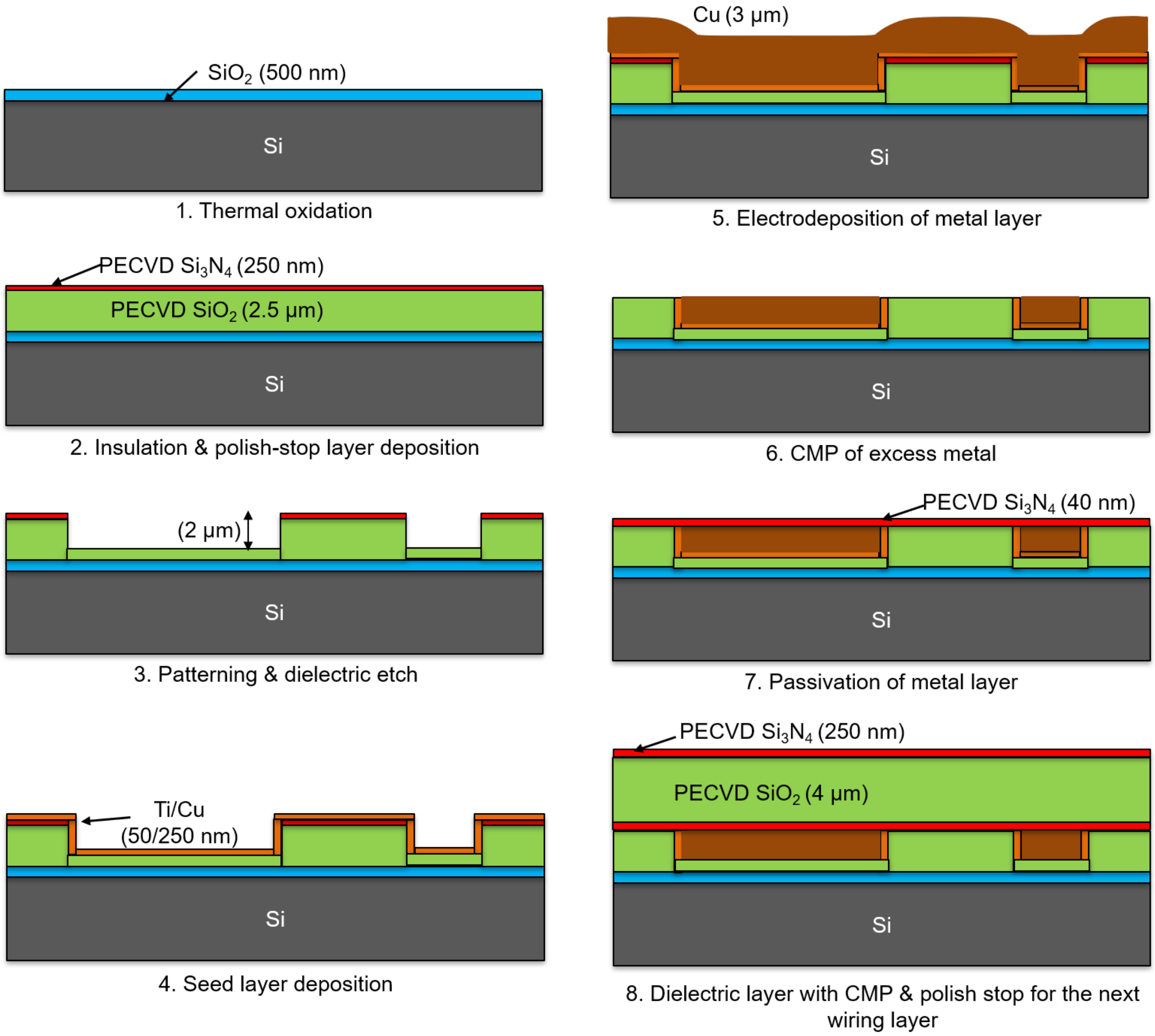
3.1 Fabrication Process Flow

The Si-IF substrate fabrication adopts the already established CMOS BEOL fabrication techniques with Cu wiring levels in SiO_2 dielectric layers. The wiring levels are fabricated using a dual damascene process [Gup09]. The wiring levels on the Si-IF are comparable to the fat-wiring levels on the die that have relatively larger features for CMOS processing. This simplifies the fabrication process of the Si-IF. Also, the Si-IF is completely passive with no transistors, significantly reducing the fabrication steps. The fabrication process flow of the wiring levels in Si-IF is shown in Fig. 3.1. The fabrication process of the wiring levels is described below.

Step 1: The silicon substrate is deposited with a 500 nm of SiO_2 using thermal oxidation in a furnace.

Step 2: SiO_2 of thickness 2.5 μm is deposited using a plasma-enhanced chemical vapor deposition (PECVD) process. Subsequently, a 250 nm silicon nitride (Si_3N_4) is deposited on top that acts as a polish stop layer for consequent steps.

Step 3: The Si-IF is lithographically patterned with the mask of the wiring level using a photoresist. Later the SiO_2 dielectric layer is etched (2 μm) using a dry etch process



SiO$_2$ (500 nm)
Si
1. Thermal oxidation
PECVD Si$_3$N$_4$ (250 nm)
PECVD SiO$_2$ (2.5 μm)
Si
2. Insulation & polish-stop layer deposition
(2 μm)
Si
3. Patterning & dielectric etch
Ti/Cu
(50/250 nm)
Si
4. Seed layer deposition
Cu (3 μm)
Si
5. Electrodeposition of metal layer
Si
6. CMP of excess metal
PECVD Si$_3$N$_4$ (40 nm)
Si
7. Passivation of metal layer
PECVD Si$_3$N$_4$ (250 nm)
PECVD SiO$_2$ (4 μm)
Si
8. Dielectric layer with CMP & polish stop for the next wiring layer

Figure 3.1: BEOL fabrication process flow of wiring layers on the Si-IF.

to form trenches for the wiring level. Consequently, the photoresist is stripped. For the subsequent wiring levels after the first layer, vias should also be etched following the trench (dual damascene process). The vias are similarly patterned and etched through the trenches to reach the metal layer below.

Step 4: A blanket titanium/copper (Ti/Cu: 50 nm/250 nm) layer is sputtered to act as a seed for electroplating. A barrier layer such as tantalum nitride or titanium nitride may also be added to reduce Cu diffusion into the dielectric.

Step 5: Cu metal is deposited using electroplating to fill the trenches.

Step 6: The plated Cu metal is polished using a chemical mechanical polishing (CMP) process to remove the excess metal and planarize the surface. The density of the wiring pattern influences the planarity, dishing, and erosion of the CMP process.

Step 7: A thin (40 nm) layer of Si_3N_4 is deposited on top to protect the Cu wires from oxidation in subsequent steps.

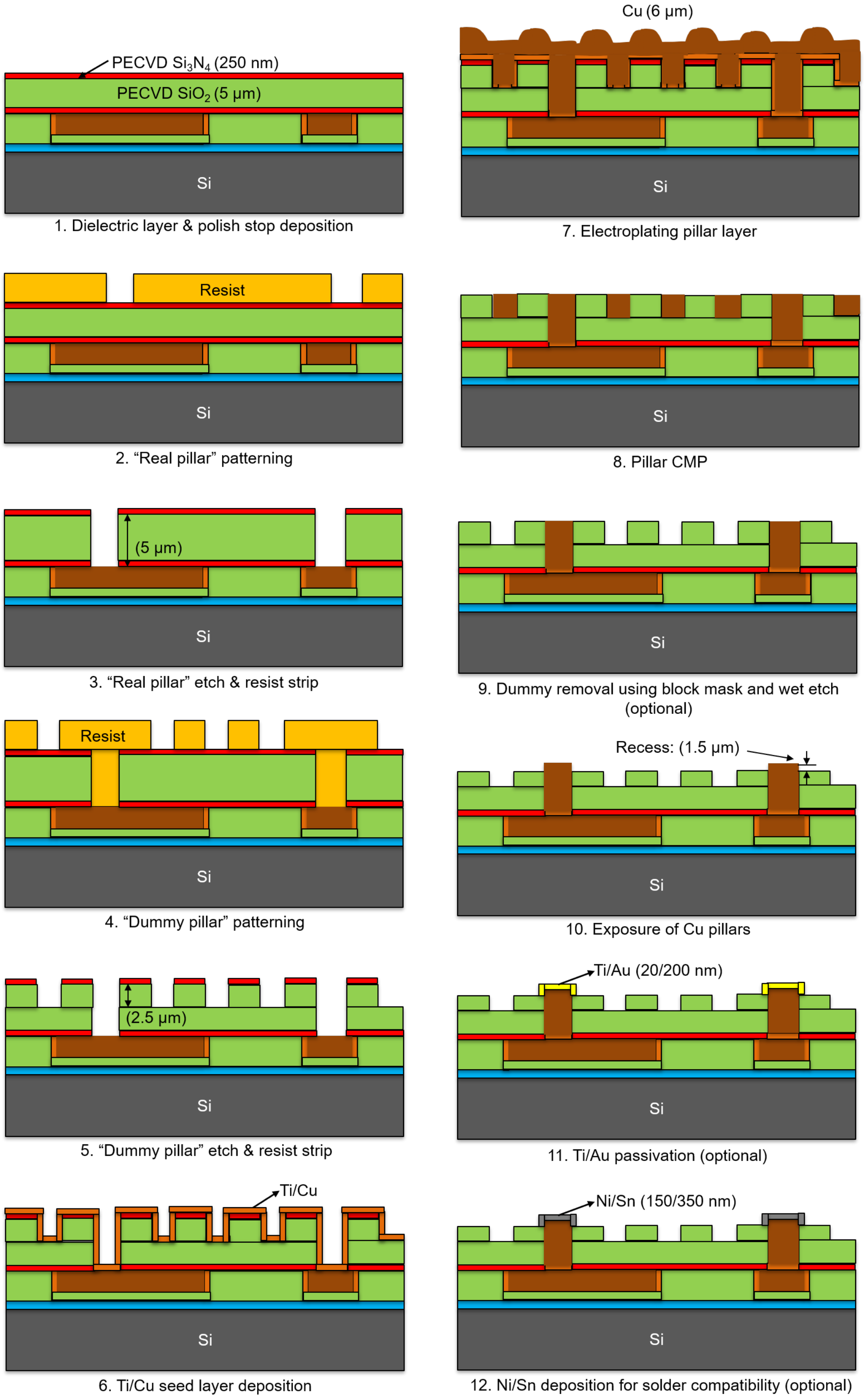
Step 8: A SiO_2 dielectric layer of 4 μm is deposited using PECVD for the next wiring level. The dielectric layer can be planarized using CMP to minimize the variations introduced by the wiring layer below. Finally, similar to step 2, a 250 nm Si_3N_4 polish stop layer is deposited.

Step 9: The process can be repeated for subsequent metal layers for up to the maximum number of layers that the technology allows. In the work, a maximum of two wiring levels were demonstrated.

3.1.1 Pillar Fabrication

The top-most metal layer is terminated with Cu pillars that act as interconnects for bonding to dies. The planarity of the Cu pillars is extremely crucial for the TCB process which will be elaborated in chapter 4. However, since the pillars connect to the I/O and power pads on dies which are spatially sporadic, the pillars do not conform to the density requirements of a typical CMP process. This results in significant non-uniformity and dishing that leads to some pillars not contacting the pads during bonding. Ideally for the CMP process, if the pillars are uniformly populated across the wafer, the best planarity is achieved. Therefore, the traditional damascene process [Gup09] was modified to include “dummy pillars“ that do not contact the metal layer below. The process flow for pillar fabrication is shown in Fig. 3.2 and described below.

Step 1: A 5 μm thick SiO_2 and 250 nm Si_3N_4 layer is deposited on the top of last wiring



1. Dielectric layer & polish stop deposition

2. “Real pillar” patterning

3. “Real pillar” etch & resist strip

4. “Dummy pillar” patterning

5. “Dummy pillar” etch & resist strip

6. Ti/Cu seed layer deposition

7. Electroplating pillar layer

8. Pillar CMP

9. Dummy removal using block mask and wet etch (optional)

10. Exposure of Cu pillars

11. Ti/Au passivation (optional)

12. Ni/Sn deposition for solder compatibility (optional)

Figure 3.2: Fabrication process flow of fine-pitch Cu-pillars on the Si-IF.

layer.

Step 2: The dielectric is lithographically patterned using photoresist with the mask of the pillars that need to connect the die to the wiring layer below, called “real pillars”.

Step 3: The dielectric layer is etched to reach the wiring below and the photoresist is stripped.

Step 4: Photoresist is again spin-coated on top of the Si-IF and it is lithographically patterned with the mask of “dummy pillars”.

Step 5: Subsequently, the dielectric layer is etched only half-way, i.e. 2.5 μm in order to prevent contacting the wiring layer below. Therefore, the dummy pillars do not connect electrically to wiring layer below. Consequently, the photoresist is stripped.

Step 6: A blanket Ti/Cu (50 nm/250 nm) layer is sputtered to act as a seed for electroplating.

Step 7: Cu metal is deposited using electroplating to fill the trenches and form the pillars.

Step 8: The plated Cu metal is polished using the CMP process to remove the excess metal and planarize the surface. In this step, the density requirements for CMP are satisfied because of the dummy pillars.

Step 9: At this point, there is an option to remove the dummy pillars. Sometimes, the dummy pillars can cause unwanted shorts on the dies if not designed appropriately. Moreover, the dummy pillars contribute to the effective bonding area of the die on the Si-IF which in turn restricts the appliable bonding pressure. Therefore, a block mask is used to block the “real pillars” using photoresist, and the dummy pillars are removed using an ammonium per sulfate-based Cu etchant. Later, the photoresist is stripped to expose the “real pillars”.

Step 10: The dielectric layer is then recessed by 1.5 μm using a dry etch process to raise the Cu pillars above the Si-IF surface. This helps in bonding and accommodates for any non-uniformity, and warpage of the dies. Moreover, the recess is tapered as shown in Fig. 3.3 & Fig. 4.9 to ensure that the Cu pillar is enclosed with a dielectric layer for passivation.

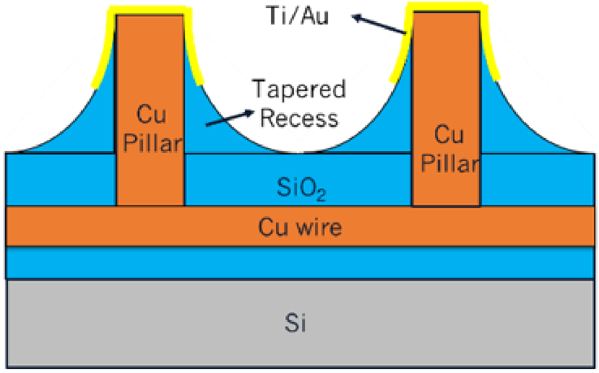


Figure 3.3: Schematic of the Cu-pillars exposed using a tapered recess of the surrounding dielectric and capped with a Ti/Au layer.

Step 11: Further, at this step, there is an option to passivate or cap the Cu pillars with Ti/Au (20 nm/200 nm) layer. This is achieved with a lift-off process using the same mask used for the “real pillars”. During exposure, bias is added to ensure that the Ti/Au pattern is larger than the Cu pillar diameter for overlay compensation as seen in Fig. 3.3 & Fig. 4.9. The Ti/Au layer is essential for direct Au-Au bonding, which will be elaborated in chapter 4, particularly for assembling III-V dies on the Si-IF.

Step 12: Also, the previous step can be replaced with Ni/Sn (150 nm/350 nm) layer for compatibility with legacy dies and surface mount passives with solder bumps.

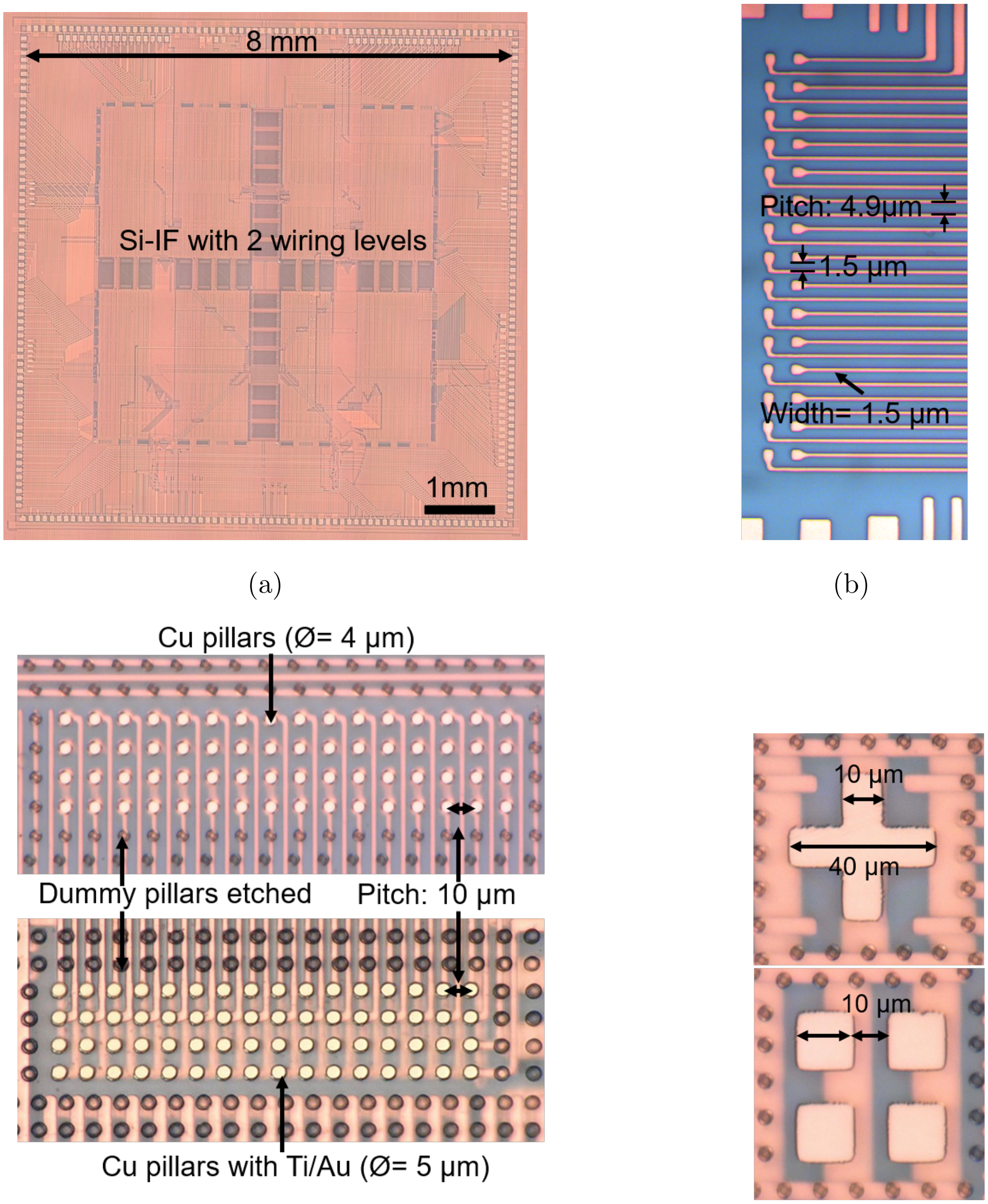
3.2 Results

Each of the fabrication steps mentioned above has been carefully optimized and the processes were controlled to ensure repeatability and robustness. The entire Si-IF fabrication has been developed at UCLA using the cleanroom research facilities and all the Si-IF samples presented in this dissertation were fabricated at UCLA. The micrographs of a fabricated test site on the Si-IF is shown in Fig. 3.4 (a). Fig. 3.4 (b), (c) show the fabricated 1st Si-IF wiring layer, and the Cu pillar layer respectively. Also, the Cu pillars capped with Ti/Au are shown in Fig. 3.4 (c). In addition, alignment marks are required to precisely align the dies to the Si-IF which will be elaborated in chapter 4. The fabricated alignment marks are shown in Fig. 3.4 (d).

The maximum number of layers demonstrated was two wiring layers and a pillar layer. The minimum dimensions achieved within the limitations of UCLA facilities were a wire width of 1.5 μm and a spacing of 1.5 μm. The pillars are 4-5 μm in diameter at 10 μm pitch. The surface roughness of the Cu pillars, which is a critical parameter for the TCB assembly, was measured using an atomic force microscope (AFM). The average root mean square (RMS) roughness of the Cu surface was 3.0 nm (±1.9 nm) [BJP+17].

3.2.1 Design Manual

A design manual for the Si-IF technology was developed with various metallization options. It describes the physical design information and the necessary design layers for the Si-IF fabrication. It contains the physical design rules for different metal layers that are manufacturable using the UCLA fabrication facilities. Also, it includes the recommended alignment marks for both the dies and the Si-IF. The key features of the design manual include six options of metallization, availability of two metal thicknesses, diagonal routing on all the layers, and compatibility with traditional CMOS technologies like 65-90 nm nodes. The important specifications in the design manual include minimum wire width of 1.5 μm, a wire spacing of 1.5 μm, maximum wire width of 24 μm, via dimension of 2x2 μm^2, and overlay tolerance of <0.5 μm. A couple of metallization options are illustrated in Fig. 3.5 (a) & (b) showing four wiring levels with 2 μm thickness, and two wiring levels with 4 μm thickness respectively. Further, the design manual also includes the electrical characteristics of the wiring levels based on both simulated and experimentally measured data. A corresponding Design Rule Check (DRC) deck was developed for the verification of the design layout.



(a)

(b)

(c)

(d)

Figure 3.4: Micrographs of fabricated test sites on the Si-IF. (a) Micrograph of a fabricated test site on the Si-IF wafer with 2 wiring levels and a pillar layer. (b) Fabricated 1st wiring level showing minimum wiring line/space of 1.5 μm. (c) Micrograph of 10 μm pitch Cu-pillars (top), and Cu-pillars terminated with Ti/Au passivation (bottom). The dummy pillars were etched and removed. (d) Complementary alignment marks.

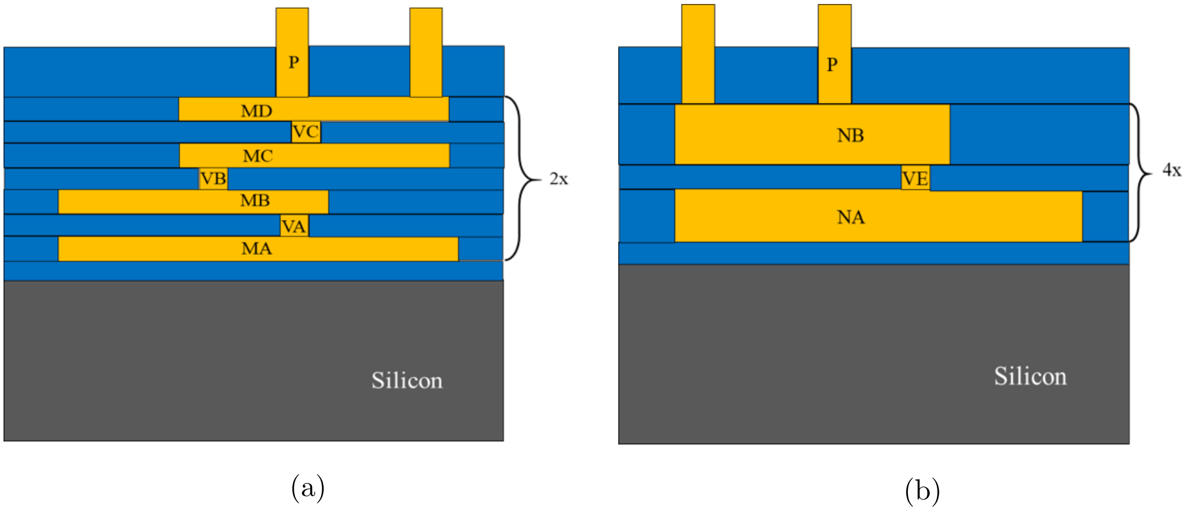
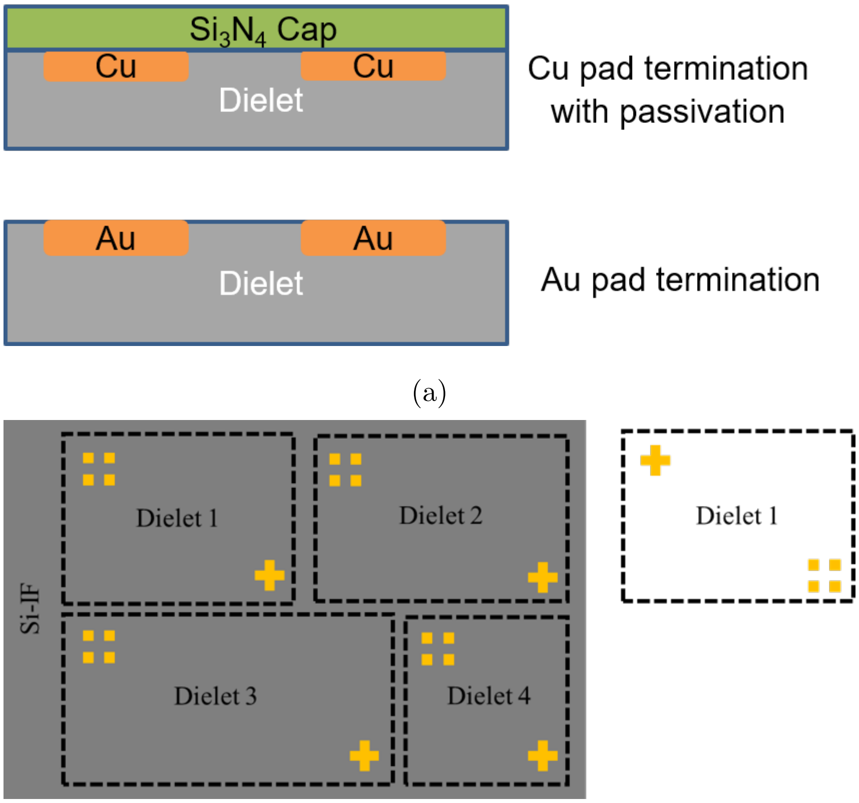


Figure 3.5: Schematics of metallization options on the Si-IF. (a) Four 2 μm thick wiring levels and a pillar level. (b) Two 4 μm thick wiring levels and a pillar level.

A dielet termination standard is also proposed for compatibility with the Si-IF integration. The pads on the dielets must be planar (flat) with the top surface of the dielet and they must be either Au or Cu terminated as shown in Fig. 3.6 (a). In the case of Cu termination, the pads must be passivated with a Si_3N_4 thin film (50-200 nm). Note that the wiring metal stack in CMOS dies is made of Cu and is typically terminated with aluminum (Al) pads. The Al pad layer is used for traditional wire-bonding or solder bumping but is not a fundamental necessity. Therefore, the CMOS dies can easily be terminated with the last Cu wiring layer instead of the Al layer without changing the process flow. Dies and passives with solder bumps can also be integrated on the Si-IF but require more than 10 μm interconnect pitch. Therefore, the solder bumps are bonded to multiple fine-pitch Cu-pillars on the Si-IF. The dies should also include alignment marks for precise assembly on the Si-IF. The alignment schemes are illustrated in Fig. 3.6 (b), where complementary alignment marks are placed on the diagonally opposite corners of the die and the corresponding marks are designed on the Si-IF. These alignment marks are used during assembly to precisely align the die to the substrate within $\leq$1 μm accuracy.



(a)

(b)

Figure 3.6: (a) Schematics of the dielet termination standards. The dies may be terminated with Cu or Au pads. (b) Die alignment scheme for the assembly on Si-IF.

CHAPTER 4

Fine-pitch ($\leq$10 μm) Assembly

The need for a fine-pitch assembly was introduced in chapter 1. In this chapter, we will take a closer look at the challenges of achieving $\leq$10 μm pitch interconnects in traditional packaging using solder-based interconnects. Further, the solder-less direct metal-metal TCB processes developed in this work are presented and sub-10 μm interconnect pitch assemblies are demonstrated.

4.1 Challenges of Fine-pitch Assembly

Over the past decade, there has been a lot of progress in reducing the die-to-substrate interconnect pitch. Today, conventional packages use C4 bumps at 130 μm pitch [MSP$^+$16] to attach a die and boards use BGA bumps at 400-1000 μm pitch [Int] for package assembly. The major limitation in reducing this pitch is due to the use of solder. Reducing the pitch corresponds to a reduction of the solder volume. This results in the complete consumption of solder by the UBM which leads to intermetallic compounds (IMCs) [KSB13]. These IMCs are extremely brittle and cause major reliability concerns. On the contrary, if the solder volume is left unchanged, reducing the pitch leads to shorting of adjacent bumps [KSB13]. Moreover, the organic laminate warpage can be several tens of microns which is also a major limitation in reducing the solder volume and consequently, the bump pitch [MAH$^+$13]. Rigid silicon interposer technology was able to reduce the bump pitch to 40-55 μm by using solder-capped Cu pillars [LLKK18, CHT$^+$17]. The Cu pillar is tall ($\geq$20 μm) and acts as the UBM to limit solder consumption.

In addition, different assembly techniques were developed to achieve fine-pitch bonding.

Solid Liquid Inter-Diffusion (SLID) process between metal (Cu) and solder (Sn) has been proposed in [HHK+14]. However, as mentioned earlier, IMCs are formed in this process. During cyclic loading, these joints are subjected to high thermo-mechanical stresses, which cause joint failures due to fatigue cracking. TCB of solder-capped Cu pillars has been widely adopted to bond the die to a substrate coated with non-conductive paste (NCP) [GBO+11]. Temperature and pressure are applied on the die which breaks the NCP and forms the bonds with the pads on the substrate. However, this process is hard to control, because, if the bonding pressure is low, no contact is established, while high bonding pressure leads to shorts [GBO+11]. Using these techniques, the interconnect can be reduced to sub-40 μm, however, as argued in chapter 1, systems today require on-chip like interconnect pitch (1-10 μm) which is extremely hard to achieve using solder. To circumvent the problems of solder-based interconnects, solder-less direct metal-metal bonding was proposed which is discussed in the next section.

In addition to the assembly process, fine-pitch assembly also requires precision alignment of a die to the substrate. Today, state-of-the-art tools achieve ≤ 2 μm accuracy (3 σ). However, reducing the interconnect pitch to ≤ 10 μm requires sub-micron accuracies which is challenging given the mechanical vibrations, temperature profile, and pressure profile of the bond-heads. Testing of bonded assembly is also a major challenge because of the large number of connections [KAB+05]. Also, KGD testing without damaging the fine-pitch pads becomes extremely challenging and testing may need dedicated sacrificial pads.

4.2 Solder-less Thermal Compression Bonding

In solder-less TCB, two nominally flat metal surfaces are joined together using a solid-solid diffusion process instead of a molten solder attach. The bonding is performed at elevated temperatures, typically a homologous temperature of 0.3-0.5, with applied pressure on the interface. The solid-solid diffusion process has been extensively studied in the past [DW82, DW84, MGY+12]. Several mechanisms of diffusion were proposed including plastic deformation of surface asperities, grain boundary diffusion, surface diffusion, and creep. Different mechanisms dominate depending on the material properties such as surface roughness, yield strength, and bonding parameters such as temperature, pressure, and time. To summarize the TCB process, when the two mating metal surfaces are brought in contact, initially, the asperities on the surfaces touch. By applying force, these asperities plastically deform because the effective pressure is higher than the yield strength due to the low effective contact area. This forms the initial bond between the metal surfaces. As this process continues, the percentage of bonded area increases, and the asperities are flattened. Accordingly, the effective pressure is reduced below the yield strength. At this point, depending on the temperature and pressure, the surface and grain boundary diffusion mechanisms continue to close the voids as time progresses. Power-law creep deformation typically occurs at higher temperatures and longer bonding times [DW82]. In the TCB process developed in this work, the bonding conditions compel the dominant mechanism to be plastic deformation, followed by a combination of surface and grain boundary diffusion [GSHB$^+$17]. The essential requirements for successful TCB are listed below.

1. Extremely flat mating surfaces with low surface roughness.
2. Pristine surfaces with no surface oxidation or surface contamination.
3. High global planarity of the samples, die or wafer, is also essential.

These surface properties affect the bonding parameters such as temperature, pressure, and bonding time. Moreover, these bonding parameters are also correlated providing a process design space where one parameter can be traded for another. In the Si-IF technology, flat mating surfaces are achieved using the established CMP process to planarize Cu with a surface roughness of 3.0 (±1.9) nm rms [BJP$^+$17]. In addition, global planarity is achieved using uniform pillar density with dummy pillars as discussed in section 3.1.1. Moreover, the global planarity requirements are slightly relaxed for die-to-wafer assembly since planarity has to be ensured only across the die area. However, achieving pristine mating surfaces is challenging.

Ideally, for die-to-wafer assembly, direct Cu-Cu bonding is desirable as Cu is the facto metal in a die metal stack and has excellent electrical and thermal properties. Consequently, using fine-pitch Cu-Cu interconnects would seamlessly attach the dies to the Si-IF like vias in a metal stack. However, the Cu surface is highly prone to the formation of surface oxides (e.g. Cu_2O, CuO, etc.) even under normal atmospheric conditions, making direct Cu-Cu bonding extremely challenging. Furthermore, the rate of oxidation increases with the increase of temperature and time that can be empirically modeled as shown in (4.1) [BJP+17]. It indicates that at bonding temperatures, the Cu surface oxide thickness can be several tens of nanometers, while even at room temperature ≈1 nm thick oxide is formed within 1 hr.

$$Oxide\ thickness\ (\text{Å}) = 0.0076 * exp(0.022 * T) * log(t) \tag{4.1}$$

where T is temperature in Kelvin, and t is time in minutes.

Therefore, today, Cu-Cu bonding is reliable only in wafer-to-wafer TCB processes in a controlled environment such as vacuum or forming gas, with relatively high interface temperatures (300-400 °C), and large bonding times (15-60 min) [KC12, TLA+12, TWB+16, CCLT15]. These approaches, however, are not appropriate for die-to-substrate attachment in practice, primarily because, creating a vacuum in a large machine is difficult, and maintaining an inert environment requires extremely high flow rates (1000-1500 L/min) of gases. Further, the throughput of these processes is extremely low for dielet assembly, inflating the assembly costs. Other approaches such as hybrid bonding [GMF+18] are also used for wafer-to-wafer bonding that are very difficult to extend to die-to-wafer bonding which is discussed later.

4.2.1 Previous Work on Cu-Cu Bonding

Recently, significant research has been directed towards achieving direct Cu-Cu bonding driven primarily by wafer stacking and 3D integration. Several different passivation, pre-treatment, and in-situ treatment techniques were investigated for direct Cu-Cu bonding [KC12, TLA+12, TWB+16, CCLT15]. For most wafer-wafer bonding applications, the

Cu surfaces must be pre-treated and transferred to a vacuum chamber for bonding. Alternatively, the Cu surfaces can be passivated to prevent oxide formation using self-assembled monolayers (SAM) like hexaethiol [TLA+12]. These monolayers can be desorbed at elevated temperatures and bonding was demonstrated in an inert environment. Further, Cu-Cu bonding was also demonstrated using argon (Ar) plasma to clean the Cu surface, called surface activated bonding (SAB). In addition, Ar/Hydrogen (H_2) and Ar/Nitrogen (N_2) plasmas were also shown to clean the surface and form copper hydrides and copper nitrides respectively which passivate the surface, preventing further oxidation [TWB+16, CCLT15]. A 6 μm pitch Cu-Cu wafer-to-wafer bonding was demonstrated using a formic acid in-situ treatment in an enclosed chamber in [XWC+16]. However, these techniques with a vacuum or controlled environment work only for wafer-wafer bonding and cannot be easily extended to die-to-substrate attachment where each die must be sequentially aligned and bonded. Also, as mentioned earlier, having such a bonding system entirely in a controlled environment is not practical.

Ultrasonic bonding was proposed for die-to-wafer bonding that relies on the vibration energy to break the Cu oxide and clean the Cu surface during the bonding process [ANT15]. Temperature may also be applied simulateneously for improving bond quality (thermosonic bonding). However, achieving fine pitch using this process is difficult because it requires tall Cu pillars >20 μm. Moreover, the bonding yield is low, and the bonding interfaces were shown to consist of microscale voids. Authors in [RRSST20] have proposed a thermosonic bonding with low pressure of <6 MPa and process times of <0.5 s as a tacking process for dielet assemblies on a substrate followed by a gang TCB. But the demonstrated Cu pillars were 100 μm in diameter and scaling of this process needs further investigation.

Hybrid bonding was also successfully demonstrated by [GMF+18]. In this process, two dielectric surfaces are first treated with a plasma activation process and subsequently bonded. The plasma activation process is used to leave dangling hydroxyl groups on the surface that form strong covalent bonds with the corresponding dangling bonds on the mating surface. After dielectric bonding, the assembly is annealed for the Cu pads to expand and form bonds by a solid-solid diffusion process. Unlike previous techniques, this process does not require pressure and therefore, can potentially have higher throughput [GMF$^+$18]. But it requires extremely tight process control to ensure extremely flat mating dielectric surfaces, and the control of the plasma treatment and attachment processes is hard for practical implementation of die to wafer assembly.

Authors in [YAS14], have demonstrated Cu-Cu bonding using a formic acid pre-treatment method to clean the Cu surface. The samples were pre-aligned and placed in an N_2 inert chamber and the formic acid was purged just prior to bonding. This approach showed good Cu-Cu bond quality. However, the cleaning time was 10 min which is substantial for die-to-substrate attach, and therefore is detrimental to the process throughput. Furthermore, the loading/unloading of the die and substrate from the chamber for alignment is tedious and not practical adding to assembly time.

4.3 Au-capped Cu Thermal Compression Bonding

In this work, the first approach to prevent Cu surface oxidation was to passivate both the Si-IF Cu pillars and the die Cu pads with a thin film of Ti/Au (20 nm/200 nm). The Ti layer acts as an adhesion layer and the Au acts as the mating surface for TCB. A lift-off process was used to deposit the Ti/Au layer as described in chapter 3. Since Au is an inert metal and free of native oxides, it does not oxidize during the bonding process. As a result, direct Au-Au TCB is successfully achieved in ambient conditions. Moreover, Au has good electrical conductivity next to Cu and has a lower yield strength compared to Cu which aides the TCB process.

The samples are first sputter cleaned with low power (<40 W) Ar-plasma for 3 min to remove any surface contamination. A state-of-the-art die-to-wafer bonder by Kulicke & Soffa (K&S), APAMA was used to bond the dies to the Si-IF. The bonder consists of a bond-head that aligns and bonds the die by applying temperature and pressure. It also consists of a chuck to hold the Si-IF, and a double-sided camera to align the die to the Si-IF.

The alignment scheme was illustrated earlier in Fig. 3.6 (b). The bond-head can be rapidly heated (up to 380 °C) and cooled while the chuck is maintained at a steady temperature of 150 °C because of the large thermal mass. The bonding process is described below.

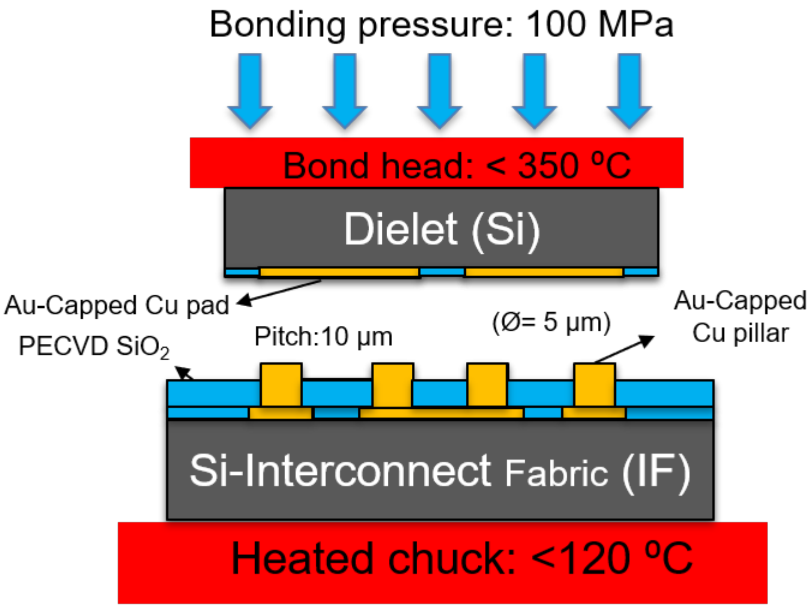
Step 1: Both the die and the Si-IF temperatures are maintained at 150 °C. The double-sided camera detects the alignment marks shown in Fig. 3.4 (d) and precision aligns the die to the Si-IF.

Step 2: The bond-head is lowered to contact the die with the Si-IF. Concurrently, a force is applied that is equivalent to 100 MPa of pressure and the die temperature is raised to 350 °C. This corresponds to an interfacial bonding temperature of 220-250 °C.

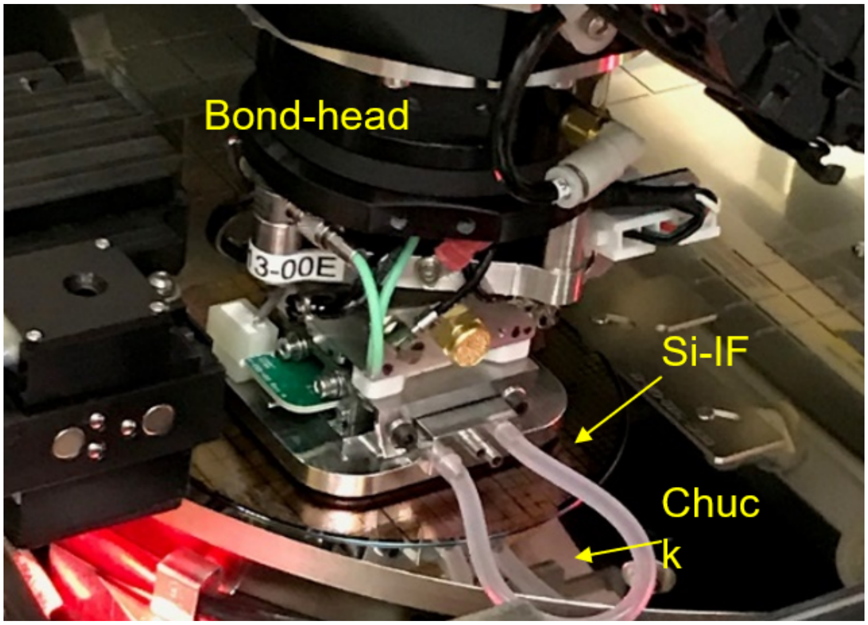
Step 3: Finally, after the bonding process, the bond-head is removed and cooled while the next die is transferred for subsequent bonding.

The schematic of the process and the tool setup are shown in Fig. 4.1 (a) & (b) respectively. The process parameters are presented in Table 4.1. Using this process, direct Au-Au bonding in ambient environment was demonstrated with actual bonding time of 3 s. This corresponds to a bonding cycle time of $\approx$6 s. Note that the process parameters depend on the properties of the samples as mentioned earlier and must be optimized accordingly. After sequential bonding of individual dies, the Si-IF can be annealed at 200-300 °C for a few hours with a slight pressure applied (<1 MPa) to improve the bond quality by allowing diffusion. However, the experimental results presented here do not include this anneal step.

This process allowed for TCB under ambient conditions for fine-pitch ($\leq$10 μm) interconnects. In addition, this method of passivation is effective, and the interface contact resistance of Cu/Ti/Au layers is insignificant (section 4.5.3). Moreover, this process is effective to assemble III-V dies (e.g. indium phosphite, gallium arsenite, etc) which typically have Au pads [SVJ+19].



(a)



(b)

Figure 4.1: (a) Schematic of the Au-Au TCB process. (b) TCB using Kulicke & Soffa (K&S) APAMA bonder tool.

Process parameters	**Value**
Substrate temperature	150 °C
Bond-head temperature	350 °C
Bonding pressure	100 MPa
Bonding time	3 s
Bonding cycle	6 s
Chamber environment	Ambient (Air)

Table 4.1: Process parameters for direct Au-Au thermal compression bonding.

4.3.1 Limitations

Although this assembly process has several advantages stated before, it requires additional processing of the dies which do not have Au pads (especially CMOS). CMOS foundries do not use Au finishing and therefore, the Au layer must be added after dicing which is difficult and impractical. Alternatively, die wafers may be received and processed which is logistically cumbersome. Moreover, the shear tests of the bonded dies revealed failures at the interface of Ti and Cu instead of the Au-Au bonding interface as shown in Fig. 4.7 (a), illustrating poor adhesion of the thin films to Cu.

4.4 Direct Cu-Cu Thermal Compression Bonding

Earlier, we established the need for direct Cu-Cu bonding for $\leq$10 μm fine-pitch assembly of dies on a wafer. It simplifies the CMOS die handling and assembly process compared to the Au-Au bonding process in the last section. However, as described before, Cu-Cu requires a reducing environment to ensure reliable bonding. To address this challenge, a novel approach of local in-situ treatment of the Cu bonding surfaces using formic acid vapor was developed [JBM^{+}19]. The formic acid vapor reduces the Cu-oxides and cleans the Cu surfaces locally below the bond-head during the bonding process. Accordingly, a die-to-wafer assembly was achieved without any vacuum or controlled environment. The details of the mechanism, tool setup, and the bonding process are presented below.

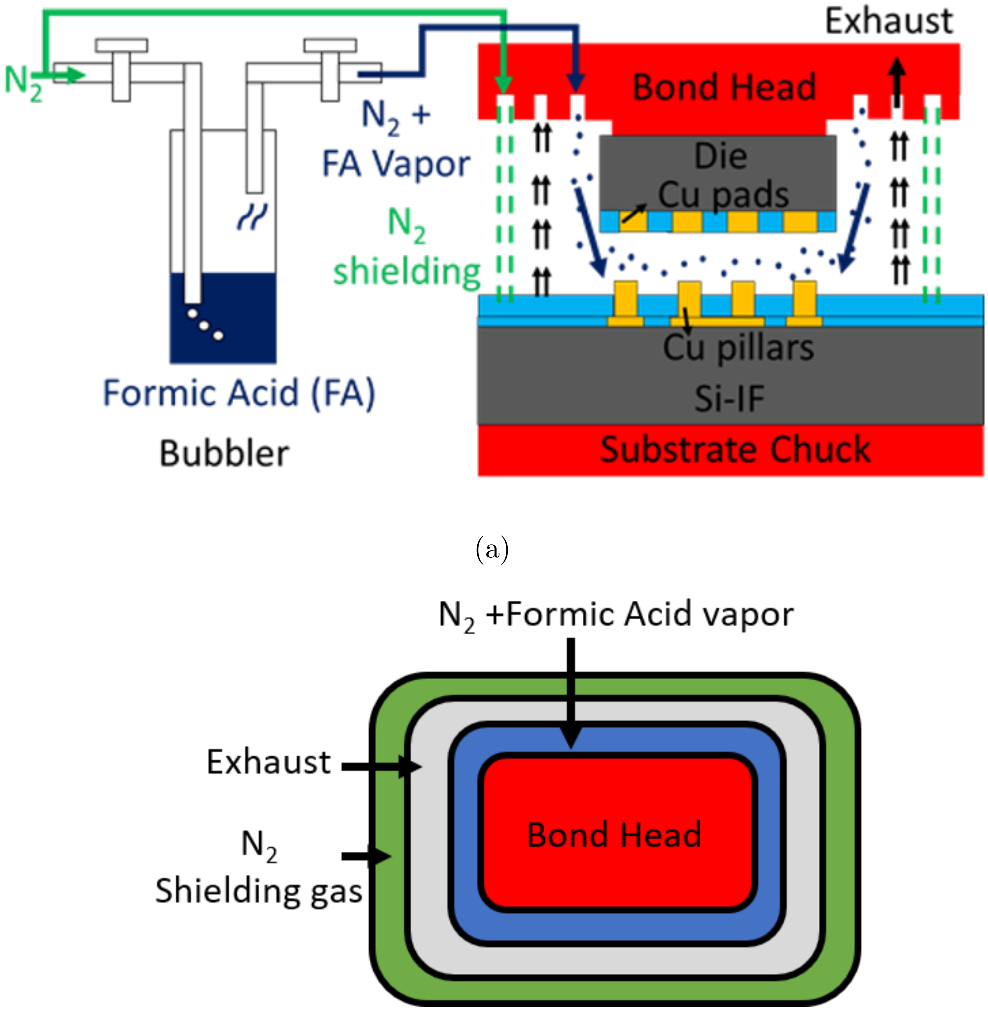
4.4.1 Tool Setup

The APAMA tool was modified in collaboration with K&S to include the formic acid treatment system. The tool setup is shown in Fig. 4.2 (a). The formic acid vapor is obtained by passing a carrier gas (N_2) through a bubbler containing formic acid (HCOOH 95%) solution. As a result, a saturated formic acid vapor is obtained at the output of the bubbler which is then transferred to the bond-head. The percentage of formic acid in the carrier gas can be altered by diluting with N_2 gas. The bond-head was modified to include a shroud consisting of three channels as shown in Fig. 4.2 (b). The innermost channel is used to purge the formic acid vapor that cleans the Cu surfaces locally just prior to bonding. The middle channel provides vacuum and acts as an exhaust for the formic acid vapor, and other reaction products during the bonding process. The outermost channel delivers N_2 as a shielding gas around the shroud. This helps contain the formic acid vapor and other products inside the target area, eliminating the need for any controlled environment in the bonding chamber.

The flow rates in these channels depend on the geometric properties of the samples as well as the process setup. The flow rates were optimized to lower the bonding cycle time and are adjusted such that the shielding gas has a higher flow than the exhaust, which in turn has a higher flow than the formic acid vapor. This ensures that the formic acid vapor and reactant products are exhausted without dispersing into the surrounding chamber. The flow rates of different gases are presented in Table.4.2. These parameters depend on the die and substrate morphology, and other assembly parameters.

4.4.2 Mechanism

The reaction of formic acid (HCOOH) with Cu was extensively investigated by several researchers [Sch12, YHB08, WL99]. A list of chemical reactions of formic acid vapor with oxidized Cu surface is given in (4.2)-(4.5). This is not an exhaustive list but represents the



(a)

(b)

Figure 4.2: (a) Schematic of the tool setup where the N_2 gas is passed through a bubbler containing formic acid to provide the formic acid vapor. (b) Top view of the bond-head shroud showing the three channels for shielding N_2 gas, exhaust, and formic acid vapor.

Gas	**Flow-rate (L/min)**
N_2 through bubbler containing formic acid	2.5
Exhaust	4
Shielding N_2	7

Table 4.2: Flow-rates of various gases in the formic acid vapor treatment setup.

major mechanisms through which the formic acid vapor reduces the surface oxides.

$$2\,HCOOH_{(g)} + CuO \longrightarrow Cu(HCOO)_2 + H_2O_{(g)} \tag{4.2}$$

$$2\,Cu(HCOO)_2 \longrightarrow Cu + 2\,CO_2 + H_{2(g)} \tag{4.3}$$

$$HCOOH_{(abs)} \longrightarrow HCOO_{(abs)} + H_{(abs)} \tag{4.4}$$

$$CuO + 2\,H_{(abs)} \longrightarrow Cu + H_2O_{(g)} \tag{4.5}$$

In gaseous form, the formic acid vapor reacts with the Cu-oxide (CuO) layer and forms Cu-formate ($Cu(COOH)_2$) and water vapor according to (4.2) at temperatures between 100-150 °C. This forms a thin Cu-formate layer that covers the bare Cu surface. When the temperature of the surface is raised above 200 °C, the Cu-formate layer dissociates into carbon dioxide and hydrogen gas, while leaving pure Cu metal on the surface (4.3). Further, the formic acid vapor can be absorbed on the Cu surface and dissociated into formates and hydrogen radicals as shown in (4.4), which is further accelerated by the presence of oxygen activation sites on Cu. These hydrogen radicals also reduce the Cu-oxides as shown in (4.5). These processes together clean the Cu surface of any native oxides and help the Cu-Cu TCB process in ambient conditions. Although most of these reactions are exothermic, they need high activation energy and therefore require higher temperatures (>200 °C) to be effective [GJB74]. At lower temperatures, the Cu-formates on the surface do not dissociate and hinder the bonding process.

4.4.3 In-situ Formic Acid Treatment Process

Similar to the Au-Au TCB process, the samples were pre-treated with low power (<40 W) Ar-plasma for 3 min to remove any surface contamination. As previously mentioned, the formic acid vapor reacts with the Cu-oxide layer and forms Cu-formate which dissociates at elevated temperatures (>200 °C) rapidly. Heating the die on the bond-head to reach these temperatures is easily feasible. However, raising and maintaining the temperature of the entire substrate region (diameter >300 mm) is technologically challenging. Further, since the chamber is at atmospheric pressures, the Cu pillars on the substrate (Si-IF) will oxidize considerably if the chuck is held at high temperatures. Therefore, a novel approach was implemented where the substrate is held at lower temperatures ($\approx$100 °C) and the die is used to transfer heat conductively to the substrate during the bonding process [JBM+19]. This helps dissociate the formates locally under the die. The steps involved in the bonding process are listed below.

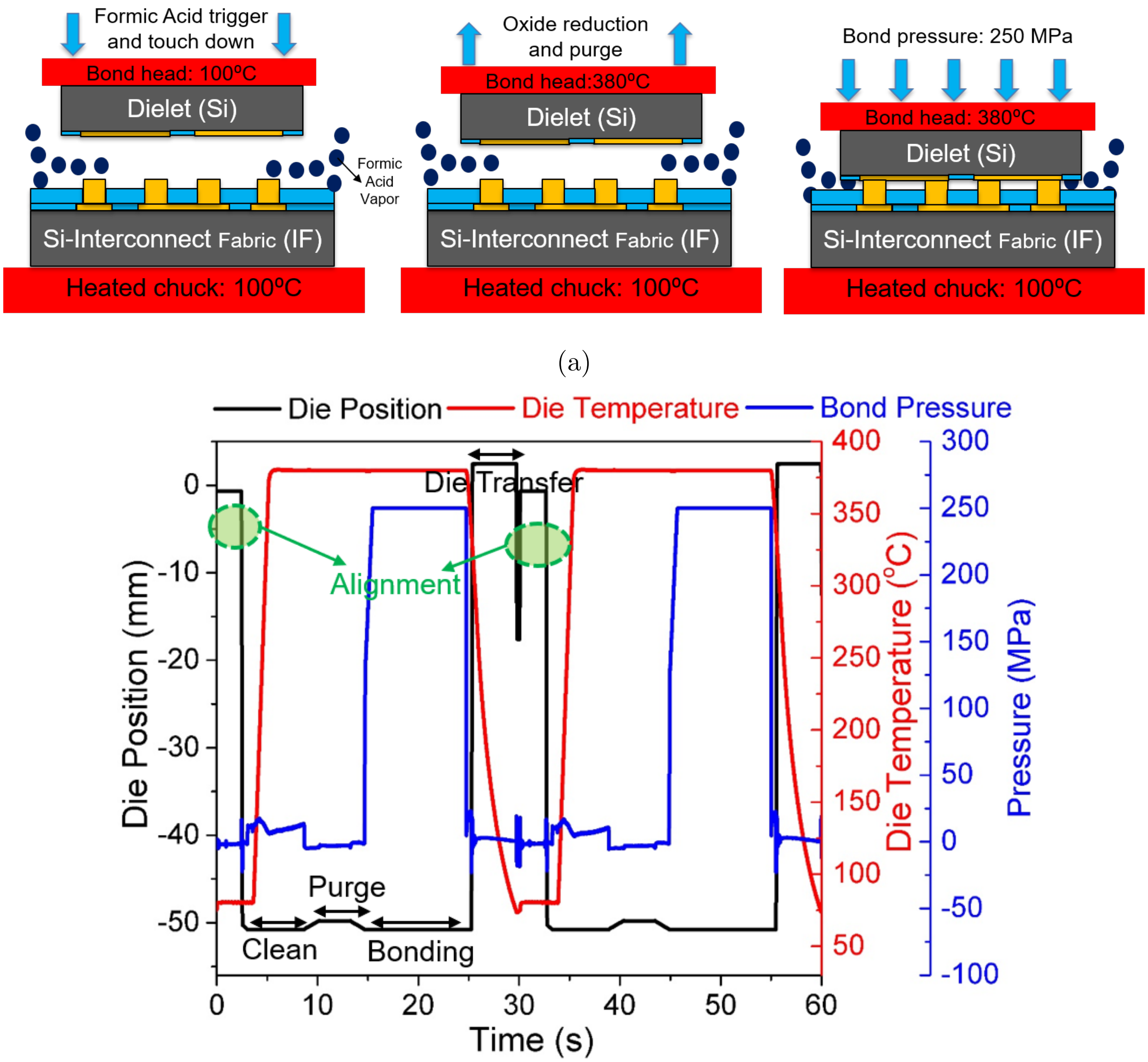
Step 1: Both the die and the Si-IF temperature are maintained at 100 °C. The double-sided camera detects the alignment marks shown in Fig. 3.4 (d) and precisely aligns the die to the Si-IF.

Step 2: As the camera retracts, the formic acid vapor valve is triggered. The die is lowered to contact the substrate with a low force (<1 MPa) and consequently, the bond-head temperature is raised to 380 °C. This establishes pad-to-pillar contact and the heat is transferred from the die to the Si-IF. The interface temperature reaches 200-240 °C which dissociates the Cu-formates that are being formed. This step lasts for 5 s.

Step 3: The bond-head is lifted up for 3 s to allow for the residual formic acid vapors and other reaction products to be sucked up the exhaust. This leaves pristine Cu surfaces both on the die and the Si-IF for the subsequent TCB.

Step 4: Once the oxides are reduced locally from the die pads and Si-IF pillars, the die is lowered to contact the Si-IF and conventional metal-metal TCB is implemented by applying a bonding pressure of 100-250 MPa for up to 10 s.

Step 5: Finally, after the bonding process, the bond-head is removed and cooled for the placement of the next die.



(a)

(b)

Figure 4.3: (a) Schematic of the direct Cu-Cu TCB assembly process illustrating the steps: Formic acid trigger, Oxide reduction, and TCB. (b) Die position, temperature, and pressure profile during the assembly process.

The schematic of the assembly process illustrating the steps is shown in Fig. 4.3 (a) and the profile of the process parameters during bonding are presented in Table 4.3 and illustrated in Fig. 4.3 (b). Using this process, direct Cu-Cu bonding at ≤10 μm pitch was demonstrated with optimized bonding times of <10 s and corresponding cleaning times of <10 s. This corresponds to a total bonding cycle time of <30 s which is a significant improvement in throughput for a TCB process, although considerably longer than solder-based attachment processes. Note that these times by themselves are unacceptably high and efforts are needed to reduce these times which is discussed in section 4.5.4. The selection of these process parameters is highly dependent on surface morphology, i.e. roughness and planarity, and material related factors such as rigidity, surface oxidation, etc. Furthermore, the force and thermal budgets for the TCB process are dictated by the underlying applications.

Process parameters	**Value**
Substrate temperature	100 °C
Bond-head temperature	380 °C
Touch-down pressure	<1 MPa
Bonding pressure	100-250 MPa
Cleaning time	10 s
Bonding time	10 s
Bonding cycle	<30 s
Chamber environment	Ambient (in-situ formic acid vapor treament)

Table 4.3: Process parameters for direct Cu-Cu thermal compression bonding.

4.5 Results

The solder-less metal-metal (Au-Au and Cu-Cu) TCB process is essential for fine-pitch (≤10 μm) integration on the Si-IF. The results of both the Au-Au TCB and direct Cu-Cu TCB processes are presented in this section.

4.5.1 Test Vehicles

Both the TCB processes were developed using daisy chain samples that form a continuous electrical link when the die is attached to the Si-IF as illustrated in Fig. 4.4. The Si-IF consists of 10 μm pitch Cu-pillars with 5 μm diameter that are alternatingly connected using a single wiring level below. Similarly, the die consists of Cu pads that form a daisy chain along the horizontal direction when attached to the Si-IF. The Si-IF also consists of probe pads (80x80 μm^2) between the dies for testing the electrical connectivity. The Si-IF and dies were fabricated according to the process in chapter 3 and are shown in Fig. 4.5 (a) & (b) respectively. In addition, for the initial Au-Au TCB process development, the Cu-pillars on the Si-IF and the Cu-pads on the die were capped with Ti/Au thin layer as shown in Fig. 4.5 (c). Also, the daisy chains can be extended to include multiple dies in series as shown in Fig. 4.4. The inter-dielet spacing between adjacent dies is $\leq$100 μm. Furthermore, various Si-IFs were designed to include dies of different sizes including 1x1 mm^2, 2x2 mm^2, 2x3 mm^2, 3x3 mm^2, 4x4 mm^2, 5x5 mm^2, and 10x6 mm^2 as shown earlier in Fig. 2.2. However, most of the electrical and mechanical results presented in this section correspond to 2x2 mm^2 dies on appropriate Si-IFs.

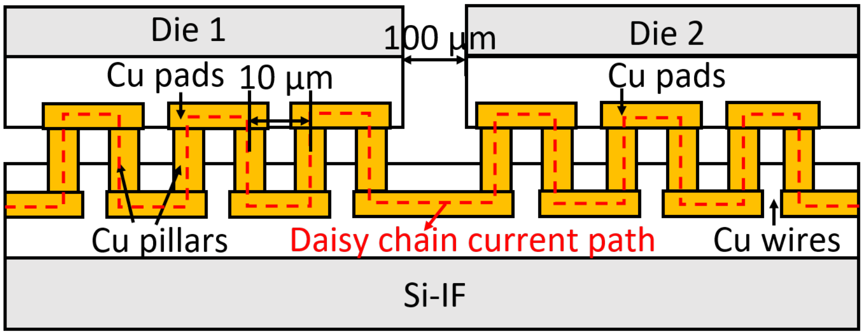
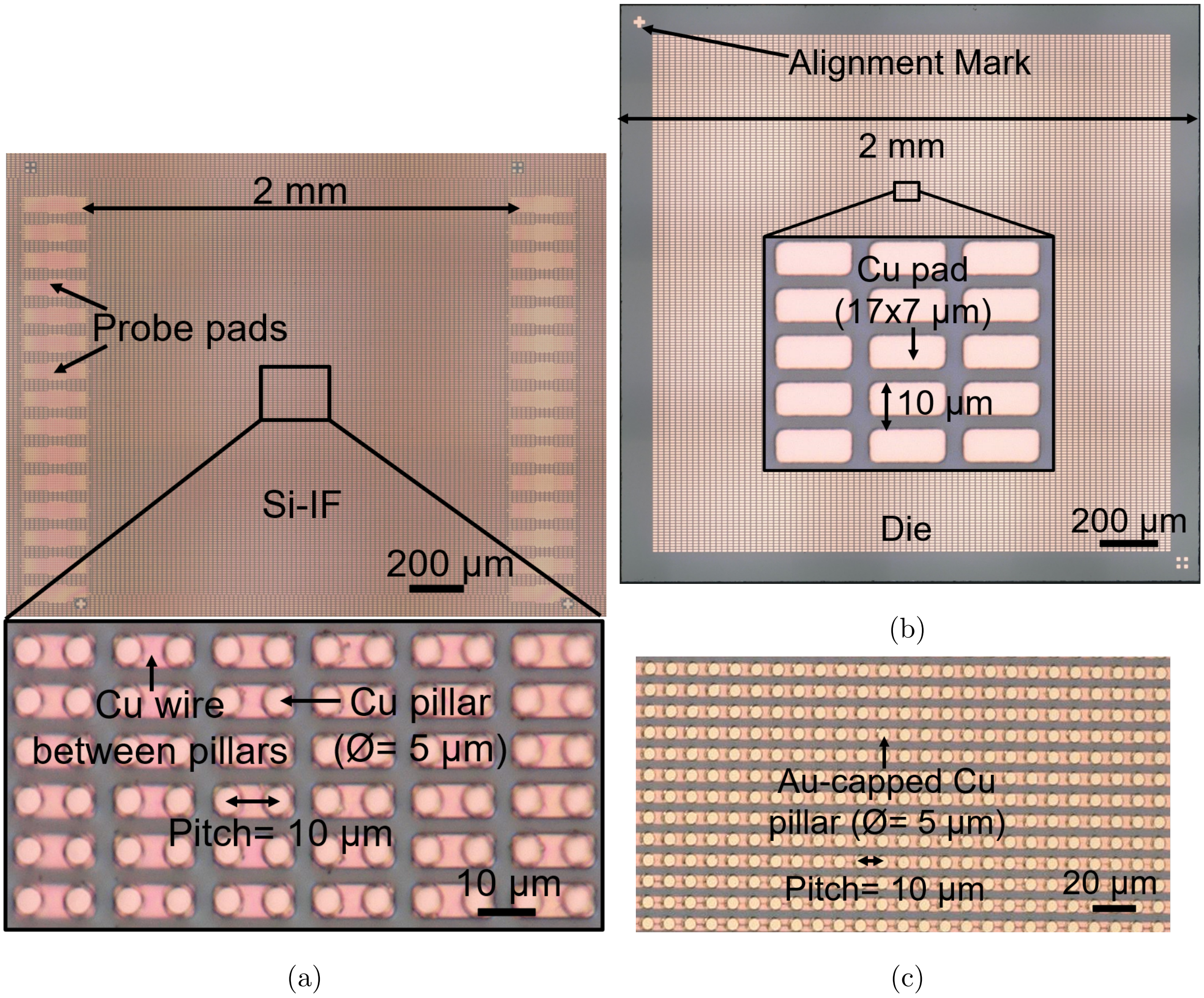


Figure 4.4: Schematic of the daisy chain test structures consisting of Cu wires on the Si-IF and pads on the dies that are attached using 10 μm pitch Cu pillars. Multiple dies can be assembled at 100 μm spacing to extend the daisy chain in series.

The 2x2 mm^2 dies bond to 32,400 ten micrometer pitch Cu-pillars on the Si-IF. This corresponds to a pillar-interconnect density of $\geq$1x10^4 mm^{-2}. For comparison, the intercon-



(a)

(b)

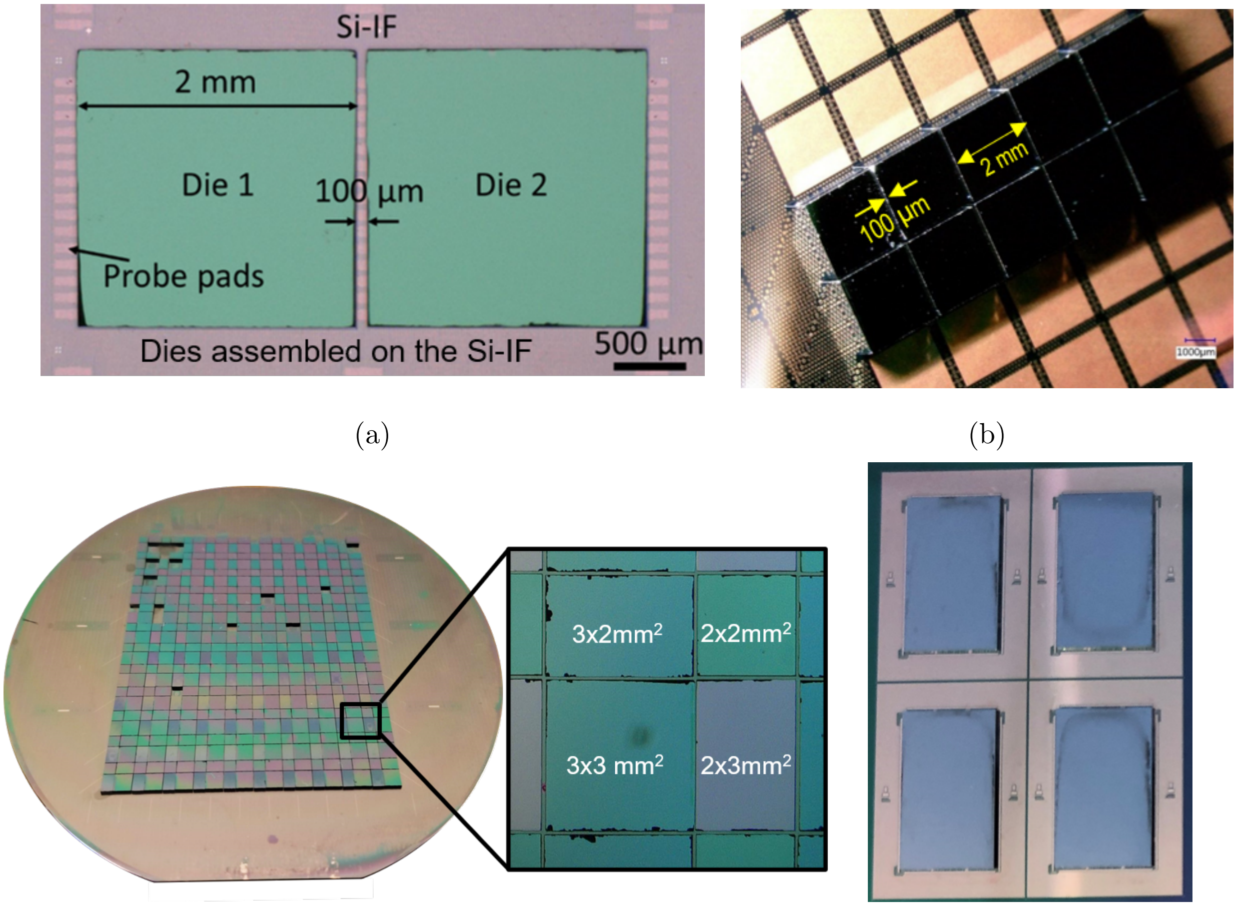
(c)

Figure 4.5: Micrograph of the fabricated test vehicles. (a) A fabricated Si-IF with inset showing the 10 μm fine-pitch Cu pillars on the Cu wires. (b) A fabricated 2x2 mm^2 die consisting of Cu pads. (c) Fabricated Si-IF Cu pillars capped with Ti/Au layer for Au-Au TCB.

nect density of C4 connections is 60-100 mm^{-2}, and BGA connections is 1-6.25 mm^{-2}. Each of these assembled dies consists of 180 horizontal daisy chains and each chain consists of 180 Cu-pillars. However, due to the limitation of the probe pad size, only 15 of these daisy chains can be tested. The testable chains are distributed evenly across the die. Further, note that every Cu-pillar in a daisy chain must be bonded for electrical continuity.

Some of the assemblies of dies bonded to the Si-IF using TCB are shown in Fig. 4.6.

Fig. 4.6 (a) shows two dies bonded at close proximity using Cu-Cu TCB. Fig. 4.6 (b) shows 10 dielets bonded using Au-Au TCB forming a continous daisy chain in series. A wafer-scale assembly of heterogeneous dielets of different die sizes are bonded at close proximity using Cu-Cu, illustrated in Fig. 4.6 (c). Assembly of a larger die 10x6 mm^2 on a corresponding Si-IF using Cu-Cu TCB is shown in Fig. 4.6 (d).



(a)

(b)

(c)

(d)

Figure 4.6: (a) Two dies assembled on the Si-IF at 100 μm inter-dielet spacing using Cu-Cu TCB. (b) An array of 10 dies on the Si-IF assembled using Au-Au TCB. (c) A wafer-scale assembly of heterogeneous dies of 2x2 mm^2, 2x3 mm^2, and 3x3 mm^2 on the Si-IF at 20-100 μm inter-dielet spacing using Cu-Cu TCB. (d) Assembly of larger dies (10x6 mm^2) on the Si-IF.

4.5.2 Mechanical Characterization

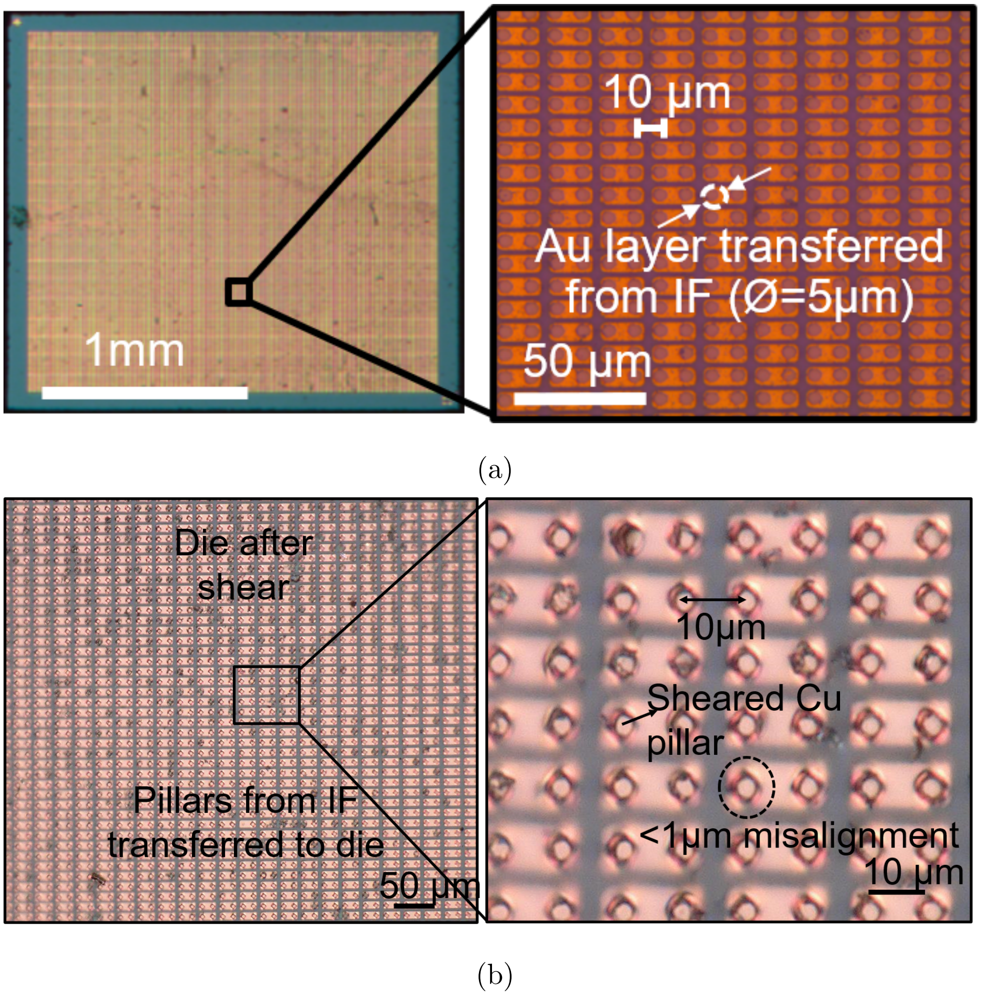
4.5.2.1 Alignment Accuracy

To scale the interconnect pitch to $\leq 10\ \mu$m, it is crucial to achieve a die-to-substrate alignment of $\leq 2\ \mu$m. As mentioned earlier, the APAMA tool has a camera that looks at the fiducials shown in Fig. 3.4 (d) to align the die to the Si-IF. The alignment is done through software control before the die contacts the Si-IF and therefore, is a second-order alignment. Since the alignment is software-controlled, the camera has to be trained to recognize the alignment marks and the offsets have to be corrected. The mechanical stability, optics, and temperature gradients during the operation affect the alignment accuracy.

To characterize the alignment accuracy, the dies were bonded to the Si-IF using the TCB processes in previous sections and then sheared to observe the interface. The micrograph of a sheared die bonded using Au-Au TCB process is shown in Fig. 4.7 (a). The 2-sigma translational alignment overlay accuracy was $\leq \pm 1\ \mu$m and a rotational accuracy was ≤ 6 mdeg. The micrograph of a sheared die bonded using direct Cu-Cu TCB process is shown in Fig. 4.7 (b). The misalignment for direct Cu-Cu TCB process is higher because of the two touch-down steps involved in the bonding process as discussed in section 4.4. The 2-sigma translational alignment overlay accuracy for direct Cu-Cu TCB is $\leq \pm 2\ \mu$m and the rotational accuracy is ≤ 10 mdeg.

4.5.2.2 Inter-dielet Spacing

Elimination of the individual die packages allows for integrating the dies at $\leq 100\ \mu$m on the Si-IF. Inter-dielet spacings of $\leq 100\ \mu$m up to a minimum spacing of 15 μm were successfully demonstrated as shown in Fig. 4.8. This corresponds to a spacing of 100-200 μm between the actual I/O circuits of the neighboring dies. The inter-dielet spacing is limited by the tolerances of the dicing process of the dies including the roughness and the dicing street variations. Moreover, the physical die size is larger than the actual design size to include some overlay that effects the inter-die I/O spacing. Therefore, state-of-the-art



(a)

(b)

Figure 4.7: Micrograph of the dies sheared after bonding to observe misalignment. (a) Die sheared after Au-Au bonding showing Au-cap of the Cu pillars transferred from the Si-IF to the die. (b) Die sheared after Cu-Cu bonding showing Cu pillars transferred from the Si-IF to the die. Both the dies show misalignment of $\leq\pm1$ μm and ≤10 mdeg.

dicing processes including stealth dicing and plasma dicing [MWMA12] help in reducing the inter-dielet spacing to a few microns.

4.5.2.3 Cross-section

The cross-section of the bonded dies on the Si-IF is inspected using the scanning electron microscopy (SEM) to observe the bonding interface. The SEM cross-section of the dies

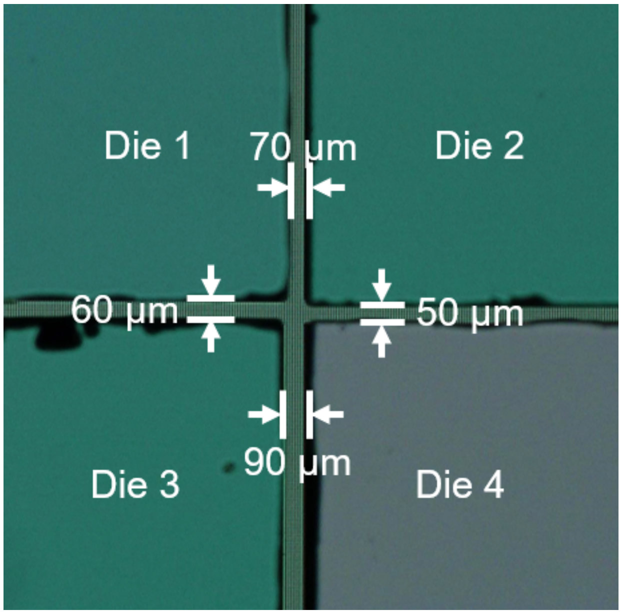


Figure 4.8: Micrograph of the ≤100 μm inter-dielet spacing between adjacent dielets on the Si-IF.

bonded using the Au-Au TCB process is shown in Fig. 4.9. At the bottom is the Si-IF with Cu trace and Cu pillars. The die with Cu-pad is on the top. The thin Ti/Au layers, both on the Si-IF pillars and the die pads, are shown at the bonding interface. As shown, no voids can be observed at the interface. The observed overhang of the Au layer is because of the lift-off pattern to account for alignment overlay and to enclose the Cu-pillars on all sides. The assembly misalignment of ≤1 μm can also be observed.

The cross-section of the dies bonded to the Si-IF using direct Cu-Cu TCB is shown in Fig. 4.10. The structure of the assembly is similar to the Au-Au bonded samples except for the Ti/Au layer. Again, there are no observable voids at the bonding interface, and in fact, there is an extrusion of Cu into the recess indicating that the bonding pressure is high and may be reduced. Moreover, the misalignment of the assembly is observed to be ≤1 μm.

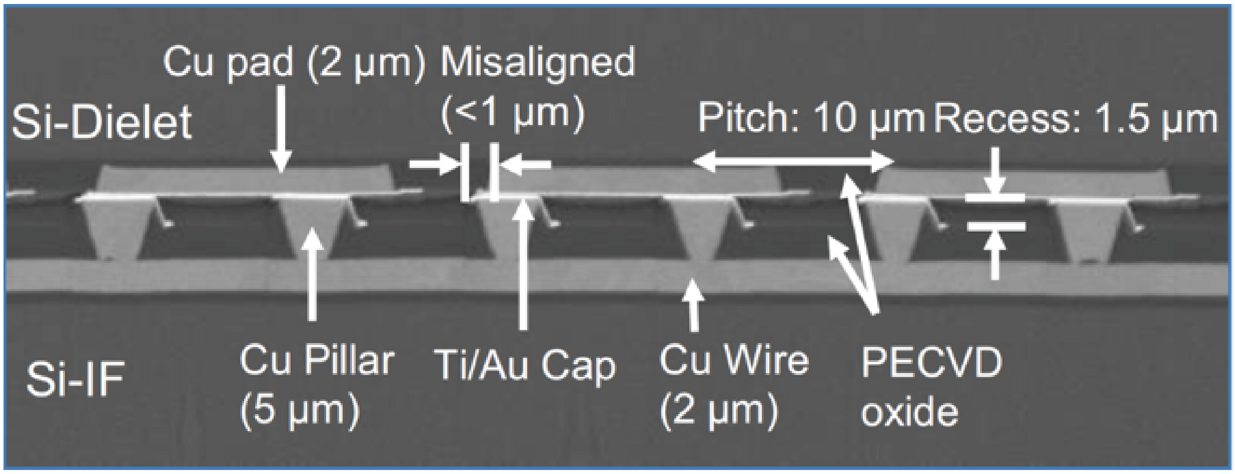


Figure 4.9: SEM cross-section of a dielet assembly on the Si-IF showing the fine-pitch interconnects bonded using Au-Au TCB. The bonding interface is void-free. (Picture Courtesy: Global Foundries)

4.5.2.4 Shear Strength

In the case of soldered interconnects, intermetallic compounds are formed at the interface that are brittle and undergo fatigue cracks which fail during thermal cycling. Direct metal-metal bonding, however, eliminates these intermetallics and forms strong bonds. Shear tests of the dies bonded to the Si-IF were performed to characterize the mechanical strength of the bonds. The dies were sheared using a standard shear tester according to the MIL-STD 883G, method 2019.9. The micrograph of the sheared dies after Au-Au TCB is shown in Fig. 4.7 (a). The average shear strength of the bonded dies is found by dividing the shear force with the effective contact area of the die. The average shear strength of the Au-Au bonded samples for a sample set of 20 was >105 MPa. The average shear strength was improved by the pre-treatment of the samples using the Ar plasma. In addition, observing the sheared dies shows that the failure is not at the Au-Au bonding interface but instead at the Ti to Cu-pillar interface demonstrating that the bonding is strong.

Similar shear tests were performed on the dies bonded to the Si-IF using direct Cu-Cu TCB. The micrograph of the sheared die is shown in Fig. 4.7 (b). The average shear strength of these samples was >127 MPa for a sample set of 12 dies. Moreover, the sheared

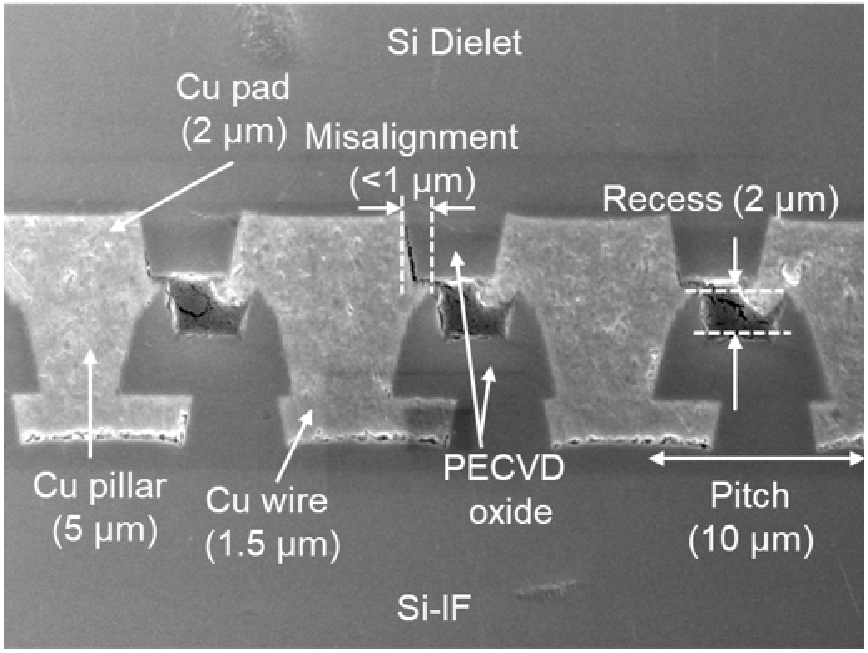


Figure 4.10: SEM cross-section of a dielet assembly on the Si-IF showing the fine-pitch interconnects bonded using direct Cu-Cu TCB. No distinct boundary is observed between the Cu pads on the die and the Cu pillars on the Si-IF demonstrating a high-quality void-free bonding. Note that the morphology at the bottom of the Cu pads is an artifact of the sample preparation. (Picture Courtesy: Pranav Ambhore).

dies show that the failure is not at the Cu-Cu bonding interface but the Cu-pillar on the Si-IF broke and transferred to the die, demonstrating excellent bond quality. Although the die shear values give an overall strength of the assembly, they may not represent the individual pillar shear strength because the applied shear pressure is not uniform across the die. In order to quantify the individual pillar strength, test dies were designed with large Cu pillars (30 μm) on a sacrificial layer that was used to remove the Si substrate of the die after bonding. The average shear strength of these individual Cu pillars was observed to be >200 MPa [JBM$^+$19].

A comparison of the average shear strength of the Si-IF assembly using both Au-Au TCB and Cu-Cu TCB with conventional solder-based interconnects is shown in Fig. 4.11. As demonstrated, the direct Cu-Cu bonded assemblies offer 2X better shear strength compared to μ-bumps. According to MIL-STD 883G method 2019.9, the dies larger than 4 mm^2 should withstand a force of at least 50 N. In a typical assembly, this is only partially supported by the solder bump, and the rest of the shear strength comes from the underfill. However, direct Cu-Cu or Au-Au interconnects can easily withstand these forces without any underfill as shown in Fig. 4.11.

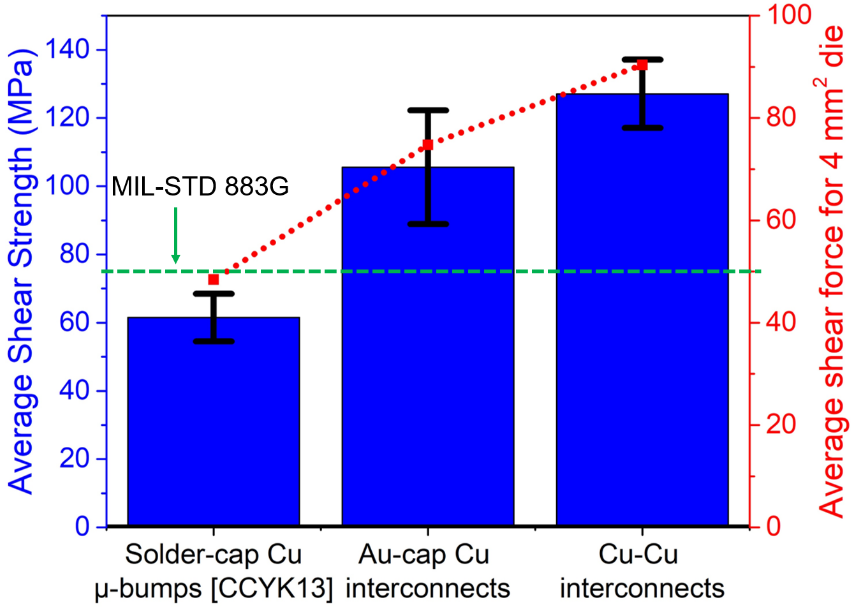


Figure 4.11: Comparison of the shear strength (blue) of solder μ-bumps [CCYK13], and the direct metal-metal interconnects in this work. The shear force for a 4 mm^2 die using these interconnects is also presented (red) and compared with the MIL-STD 883G requirement.

4.5.3 Electrical Characterization

Electrical continuity tests were performed after assembling the dies on the Si-IF to form a daisy chain. As mentioned earlier, the 2x2 mm^2 dies consist of 32,400 fine-pitch Cu-pillars on the Si-IF, or equivalently 180 daisy chains with 180 Cu-pillar per chain. Of these, 15 daisy chains can be tested using probes. The dies were sequentially placed and the resistance of the daisy chains was measured using a 4-point contact after each consecutive die attach. The assembly exhibited a 100% contact continuity yield across all the dies for

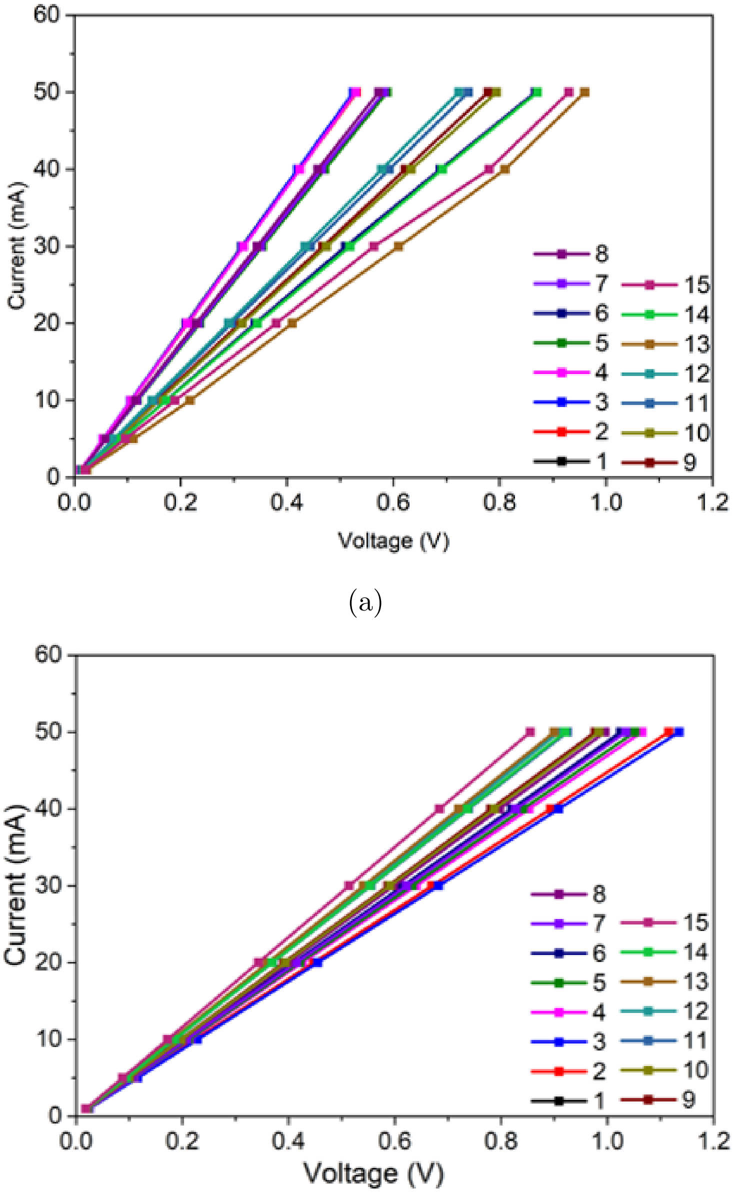
Cu-Cu and Au-Au TCB process. The assembled dies on the Si-IF using direct Cu-Cu TCB are shown in Fig. 4.6 (a). All the 15 testable daisy chains were connected for both the single die and two dies case. The current-voltage (I-V) plots of the daisy chains of a single die and two dies in series are shown in Fig. 4.12 (a) & (b) respectively. For the two dies case, the daisy chains pass through both the dies with a total of 360 interconnects per chain. The different colors represent different daisy chains tested. Both the measurements show well-behaved resistance of the Cu-Cu interconnects. The variation in the measurements between chains can be due to the misalignment during bonding process and measurement errors. The contact resistance of the individual Cu pillar was extracted from the daisy chain resistance by de-embedding the fan-out wire, Cu trace, and Cu pad resistances. The average resistance per pillar was ≈35 mΩ. This corresponds to an effective specific contact resistance of ≈0.685 Ω-μm^2. However, the pillar resistance was observed to vary from 28 mΩ to 50 mΩ which can be attributed to the misalignment during the bonding process.

Similar electrical continuity measurements were performed for dies bonded to the Si-IF using Au-Au TCB. Figure. 4.6 (b) shows the assembly of 10 dielets connected in series with 18,000 fine-pitch interconnects per daisy chain. Once again, a 100% continuity yield was achieved across the dies. The average contact resistance of the interconnects was found to be 42 mΩ which corresponds to effective specific contact resistance of ≈0.82 Ω-μm^2 [BJP+18]. This is 20% higher than interconnects bonded using direct Cu-Cu TCB because of the interfacial resistance between the Ti/Au layer and the Cu-pillar, and Cu-pad. A comparison of the specific contact resistance of various interconnects and geometries is presented in Table 4.4 and a plot of the same is illustrated in Fig. 4.13.

4.5.4 Challenges

Although fine-pitch interconnects (≤10 μm) using direct metal-metal TCB were successfully demonstrated, several challenges still remain that need to be addressed for translating to high volume manufacturing. Some of the challenges are listed below.

1. Alignment accuracy is a major consideration to ensure the repeatability of the process. For solder-based interconnects, molten solder provides self-alignment due to surface tension. However, by eliminating solder, accurate placement becomes criti-



(a)

(b)

Figure 4.12: Current vs voltage plots for (a) Daisy chains of a single die on the Si-IF, (b) Daisy chains of two dies assembled in series on the Si-IF. The different colors represent different daisy chains tested. The average contact resistance was 35 mΩ.

Interconnect	**Diameter** **(μm)**	**Contact pad area** **(μm^2)**	**Contact resistance** **(mΩ)**	**Effective specific contact resistance** **(Ω-μm^2)**
C4 bump [WPG+06]	100	7800	10	78
C4 bump [WPG+06]	50	1950	25	48.7
μ-bump [DWA+07]	23	415	47	19.5
μ-bump [DWA+07]	16	201	43	8.64
Cu pillar [DGT+09]	11.2	100	12	1.2
Au-capped Cu pillar (This work) [BJP+17, BJP+18]	5	19.6	42	0.82
Cu pillar (This work) [JBM+19]	5	19.6	35	0.685

Table 4.4: Geometric and electrical comparison of different interconnect technologies.

cal. Improving the optics, reducing the mechanical disturbances such as vibrations, and improving the software control could improve the alignment. In addition, the multi-touch process in the direct Cu-Cu TCB degrades the native misalignment and therefore, should be eliminated. Alternative methods to locally heat the Si-IF during the bonding process should be explored.

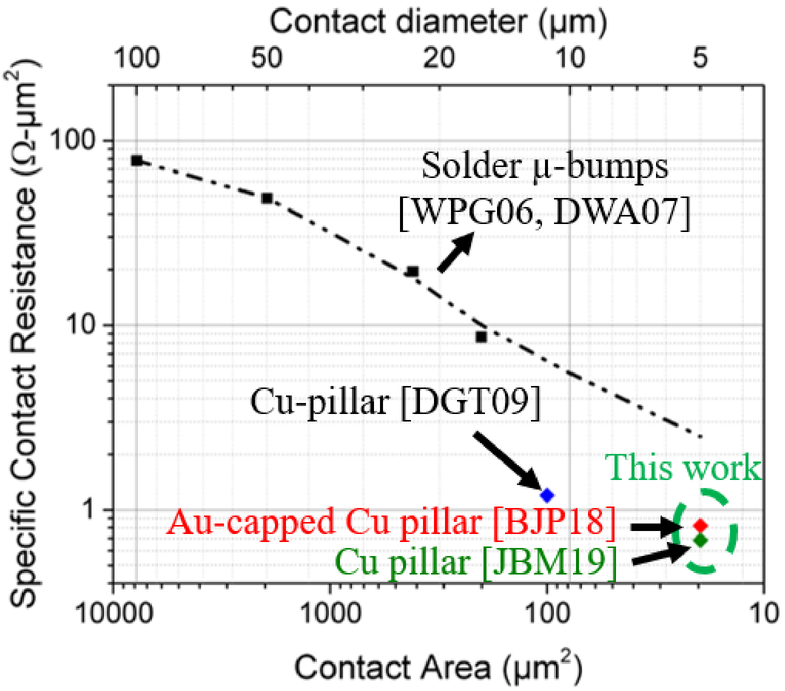
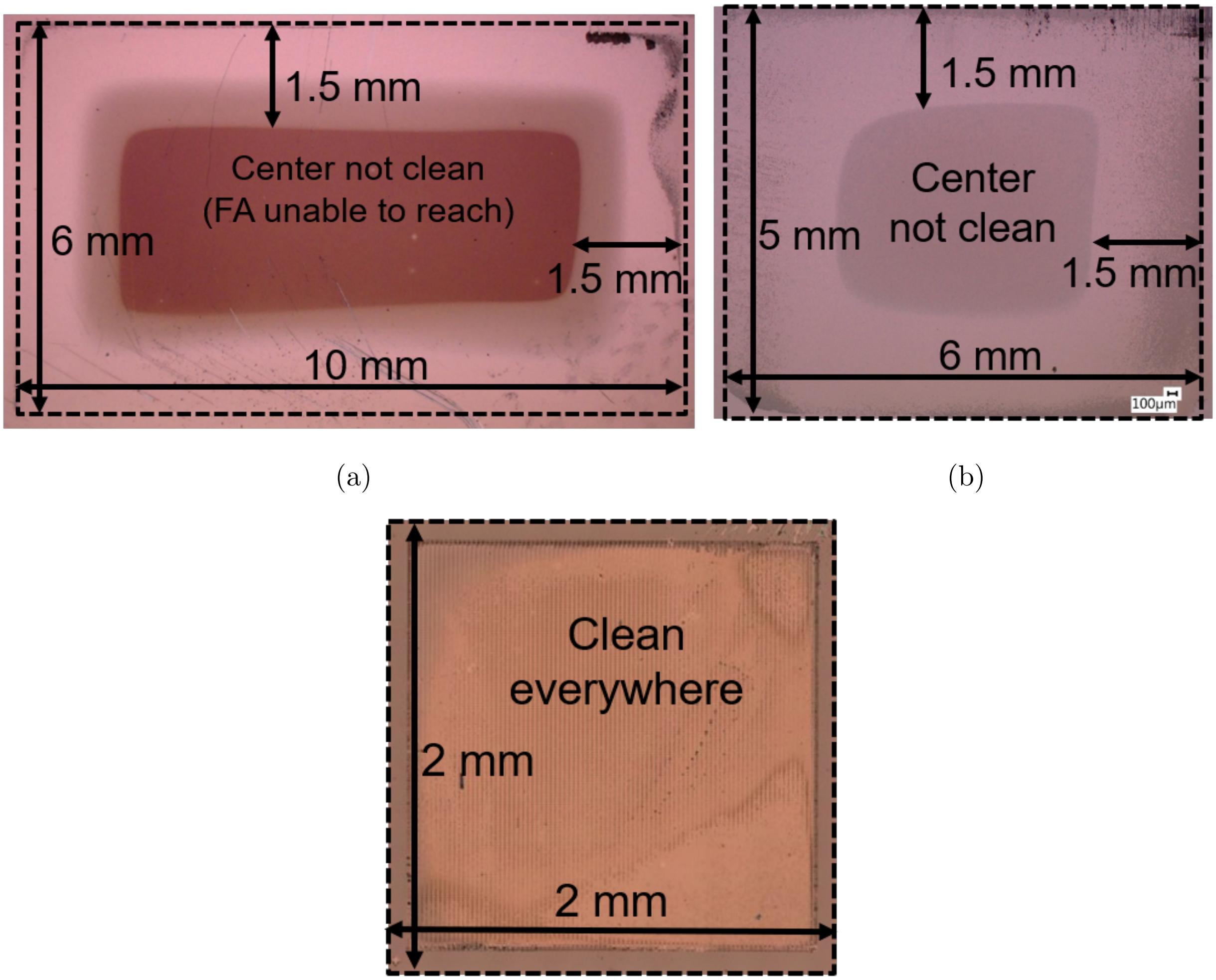


Figure 4.13: Comparison of the specific contact resistance of solder-based interconnects, and the direct metal-metal interconnects in this work.

2. Assembly of larger dies (>20 mm^2) using the direct metal-metal TCB is challenging because of the warpage and non-planarity across the die. There is a trade-off between the die thickness and planarity since full-thick dies have lower warpage while thinned dies are flexible and can be easily flattened during bonding. This trade-off should be explored to achieve successful bonds. In addition, during Cu-Cu TCB, it was observed that the formic acid vapor was not effectively cleaning the center of the die which is shown in Fig. 4.14. By using computational fluid dynamic simulations, the bond-head shroud design should be modified to improve the formic acid vapor flow from turbulent to laminar.

3. Wafer-scale systems present unique challenges in assembly which require the integration of multiple dies with different die sizes and specifications. Wafer-scale integration of heterogeneous dies on the Si-IF was demonstrated using both the direct Cu-Cu TCB earlier in Fig. 2.2 (a) and Au-Au TCB in Fig. 2.2 (b). The current tool



(a)

(b)

(c)

Figure 4.14: Micrograph showing oxidation in the center of larger dies (a) 10x6 mm^2, (b) 5x6 mm^2 because of inadequent formic acid flow. (c) No oxidation is observed on smaller dies 2x2 mm^2.

is not capable to simultaneously handle multiple die sizes. Therefore, all the dies of a particular size are bonded first on the entire wafer. Subsequently, the machine tools are changed to handle the next die size and the process is repeated to assemble all the required dies. Automatic tool changing should significantly improve the flexibility of handling multiple dies.

4. Further, for wafer-scale integration using the Cu-Cu TCB process, the substrate

needs to be held at elevated temperatures (>80 °C) for an extended period of time. This leads to extensive oxidation of the Cu-pillars on the target sites that bond at the end which cannot be cleaned by the formic acid vapor. In this work, the approach to address this problem was to periodically clean the Si-IF after every hour with Ar plasma treatment to sputter the Cu-oxide layer. Another approach was to progressively increase the formic acid vapor cleaning time as more dies are bonded. Other solutions including temporary passivation of the Cu-pillar surface should be explored. Ar/N_2 plasma treatment was shown to form a thin copper nitride that protects the Cu surface from oxidation in [CCLT15] which can be implemented before bonding.

5. Integration of conventional dies with solder bumps along with passives on the Si-IF was demonstrated in Fig. 2.5. A traditional solder-reflow process on the Cu-pillars capped with Ni/Sn layer was used for bonding. However, this no longer has the benefit of 10 μm fine-pitch interconnects and the solder-bump must be bonded to multiple Cu-pillars. Also, the integration of dies with wire-bond pads below the die surface is extremely challenging. The dies have to be bumped to ensure contact with the Si-IF pillars. Moreover, the passive components are difficult to handle because of their surface topology and they occupy a significant area on the Si-IF increasing the inter-dielet spacing. Therefore, it is best to avoid passive components and use built-in deep trench capacitors in the Si-IF [TI20] or mount the passives in a platform below.

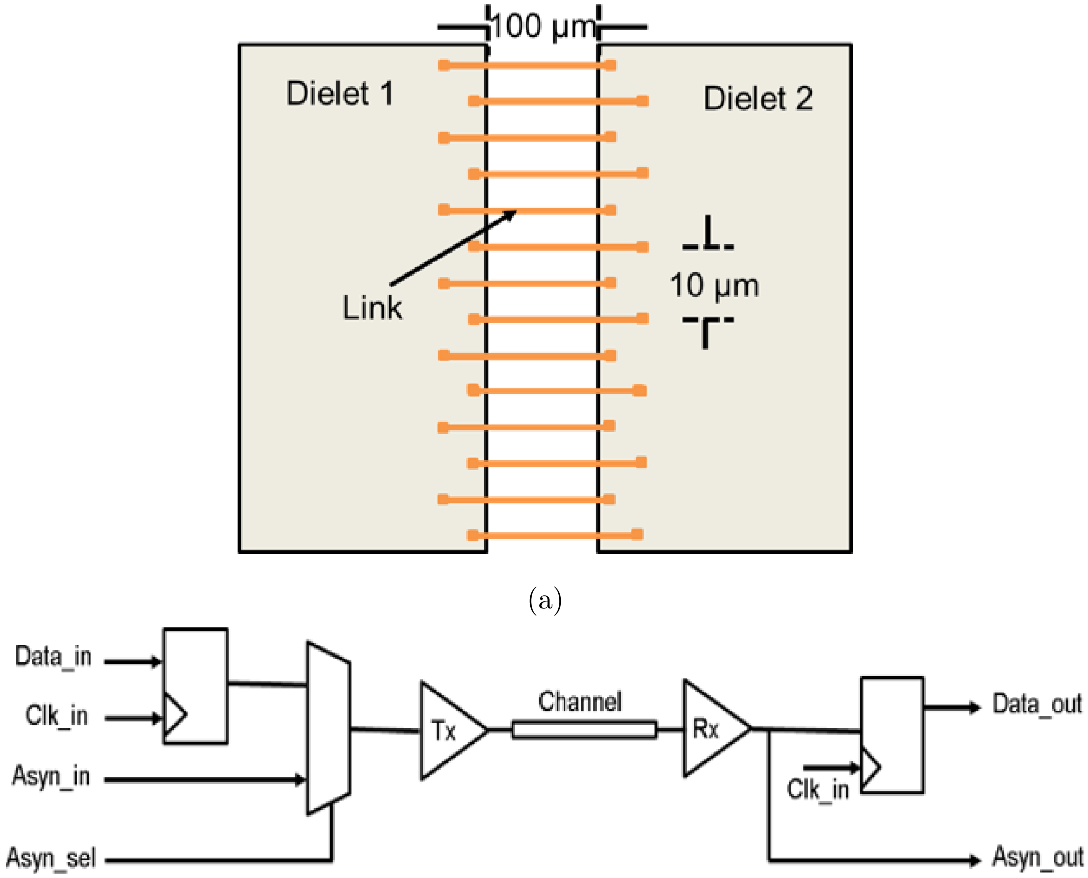
6. Finally, the throughput of the demonstrated TCB process, although, is much higher than other competing TCB technologies, is still lower than the traditional solder-based assembly. However, one should consider the bigger picture here. The Si-IF integration eliminates other traditional processes such as under-bump metallurgy, solder bumping, and so on, reducing the overall assembly time, and cost. But, the TCB process, especially the Cu-Cu TCB, must be optimized to reduce the bonding time to a few seconds (<10 s) to have a competitive advantage. This can be achieved

by improving machine hardware with better temperature ramp-up and cool-down, and having dual bond-heads working simultaneously. In addition, eliminating the multiple touchdowns also reduces the bonding cycle time.

CHAPTER 5

Simple Universal Parallel intERface for Chips

In the previous chapters, we have established the technologies required for fine-pitch integration of wafer-scale systems on the Si-IF platform. In this chapter, we will explore how to translate the benefits of fine-pitch integration to achieve performances comparable to SoCs for heterogeneous systems. A communication interface protocol called the Simple Universal Parallel intERface for Chips (SuperCHIPS) was developed to leverage the Si-IF technology. It efficiently interconnects heterogeneous dies on the Si-IF with simple I/Os for optimal system performance. According to the SuperCHIPS protocol, two adjacent dies assembled on the Si-IF at close proximity are connected using data-links that are only 50-500 μm long. This is possible because of the sub-10 μm die-to-wafer interconnect pitch and the short inter-dielet spacing of $\leq$100 μm. These links are significantly shorter than links on PCBs which are several centimeters long. As a result, the channel loss and link latency overheads are greatly reduced, thus, eliminating the need for complex transceiver circuitry. This significantly reduces the power consumption and the real estate for I/O circuitry by 5-40X and 9-25X respectively. Moreover, the wiring between the dielets are at on-chip like pitches (2-10 μm) providing a greater number of data links compared to existing technologies. With the availability of a large number of data-links, each link can be operated at a relatively lower frequency (<10 Gbps) and at the same time achieve a higher bandwidth density (up to 8 Tbps per millimeter of the die edge). Thus, the need for serialization and deserialization of data is eliminated by parallelizing data transfer. Therefore, using the SuperCHIPS protocol, simple inverter drivers transfer data across a highly parallel interface consisting of short links ($\leq$500 μm). The schematic of the SuperCHIPS interface on the Si-IF and the I/O circuit is shown in Fig. 5.1.



(a)

(b)

Figure 5.1: (a) Schematic of the fine-pitch, short-reach SuperCHIPS interface between two neighboring dies. (b) Schematic of the simple SuperCHIPS I/O.

Some of the key features of the SuperCHIPS protocol are listed below-

- SuperCHIPS is a hard interface protocol to interconnect neighboring dies with high-density wiring and simple buffer I/Os. Any logical or soft protocol can be implemented on the SuperCHIPS interface for communication.

- The SuperCHIPS protocol uses simple buffer I/Os and short links for communication and is therefore efficient only for near-chip communication, especially for neighboring or next to neighboring dies (<5 mm long). It cannot be easily extended for long-haul communication on the Si-IF and would require some modifications described in chapter 7.

- SuperCHIPS relies on key technological enablers such as the development of a superior substrate that allows for fine-pitch interconnects and tight dielet integration. As demonstrated earlier, the Si-IF technology achieves these objectives and is crucial for the implementation of SuperCHIPS.

- In this protocol, all the peripheral links are configured as single-ended unidirectional signals to achieve the maximum data-bandwidth.

5.1 Electrical Characterization

5.1.1 Test Vehicles

To experimentally demonstrate the electrical performance of the short SuperCHIPS links, test structures were designed and fabricated. These structures were designed to emulate the signal transfer between dielets communicating using the SuperCHIPS interface. The dielets have metal pads that are connected to the Si-IF links using the fine-pitch pillar interconnects (10 μm). Daisy chain structures were designed to imitate signal flow between dielets when attached to Si-IF. In a real implementation, the links will be less than 500 μm, (typically 100 μm). However, measuring the signal transfer characteristics of these short links is challenging due to the low channel losses, the physical constraints on the proximity of probes, and the subsequent de-embedding of the fan-out and probe parasitics. Therefore, to get measurable link characteristics, the short link segments on the Si-IF were cascaded in series in a daisy chain fashion using the pads on dies. This forms a long link between the two probing ports with measurable losses. The schematic of the cascaded structure and the cross-section of a link segment are shown in Fig. 5.2. The characteristics of the actual device under test (DUT), which is the short link segment, were later extracted using de-embedding techniques [FCM08, Fri94]. Additionally, this ensures that the parasitics introduced by the bonded interconnects and the assembly process (TCB) are also included in the measurements.

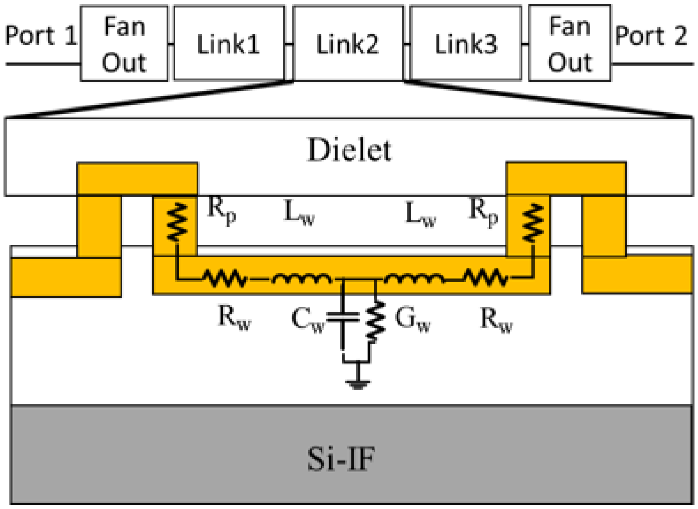


Figure 5.2: Schematic of link segments (DUT) cascaded between the two measuring ports and the cross-section of a DUT.

The DUTs consist of Si-IF links that are configured as coplanar Ground-Signal-Ground (GSG) for the insertion loss measurements and Ground-Signal-Signal-Ground (GSSG) for the cross-talk measurements. Three main parameters of the links were varied, namely the length of the link, the width of the link, and the wiring pitch as depicted in the Table. 5.1. This helps in understanding the effect of each parameter on the link characteristics. The height of the links was 2 μm conforming to the design manual specifications. In all the cases, the link segments were terminated with 10 μm pitch Au-capped Cu pillars with a diameter of 5 μm. The corresponding dies were also designed with Au-capped Cu pads that are 17 μm long, 7 μm wide, 2 μm thick to connect two link segments in series. These test vehicles were fabricated using the process described earlier in chapter 3 and the micrographs of the fabricated Si-IF are shown in Fig. 5.3 & Fig. 5.4. The dies were precisely aligned ($\leq$1 μm) and bonded to the Si-IF using the TCB process in section 4.3 and the bonded assembly is shown in Fig. 5.5. As mentioned earlier, the dielets were assembled to ensure the loss of the bonded interconnects is also included in the measurements. This would give us the actual link behavior when the dielets are in operation. In addition, de-embedding structures with shorted fan-out wires were designed to measure the fan-out wire losses and the probe parasitics. The de-embedding structures are shown in Fig. 5.4 (e).

Length of the link (μm)	**Wire width (μm)**	**Wiring pitch (μm)**
125	2	4
		10
	5	10
585	2	4
		10
	5	10
1750	2	4
		10
	5	10

Table 5.1: Different link parameters used for electrical characterization. In all cases, the link segments were terminated with 10 μm pitch pillar interconnects.

5.1.2 Insertion Loss

Two-port S-parameter measurements were performed on the bonded Si-IF structure in Fig. 5.5, using a 67 GHz Vector Network Analyzer (VNA). The S-parameters were measured for frequencies from 50 MHz to 30 GHz. To calibrate the GSG RF probes, the Line-Reflect-Reflect-Match (LRRM) standard was used. The key challenges for the insertion loss measurements were (1) De-embedding the parasitics introduced by the probes and fan-out wires; (2) Extracting the characteristics of a single link segment from the cascaded structure. To overcome these problems, first, the S-parameters of the de-embedding structures were measured and using the S-to-T-parameter conversion techniques in [FCM08, Fri94], the probe, and fan-out wire parasitics were de-embedded. Finally, using a similar tech-

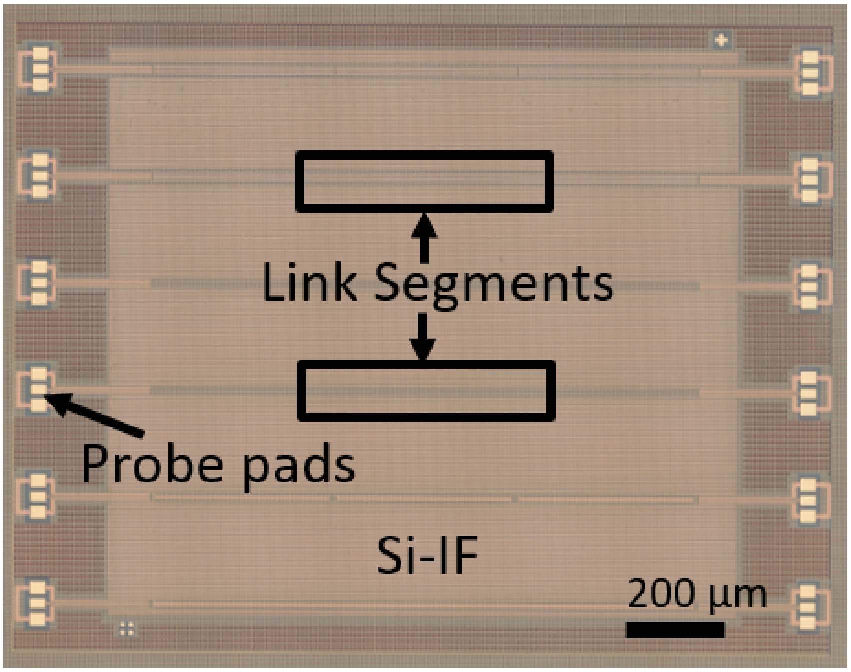
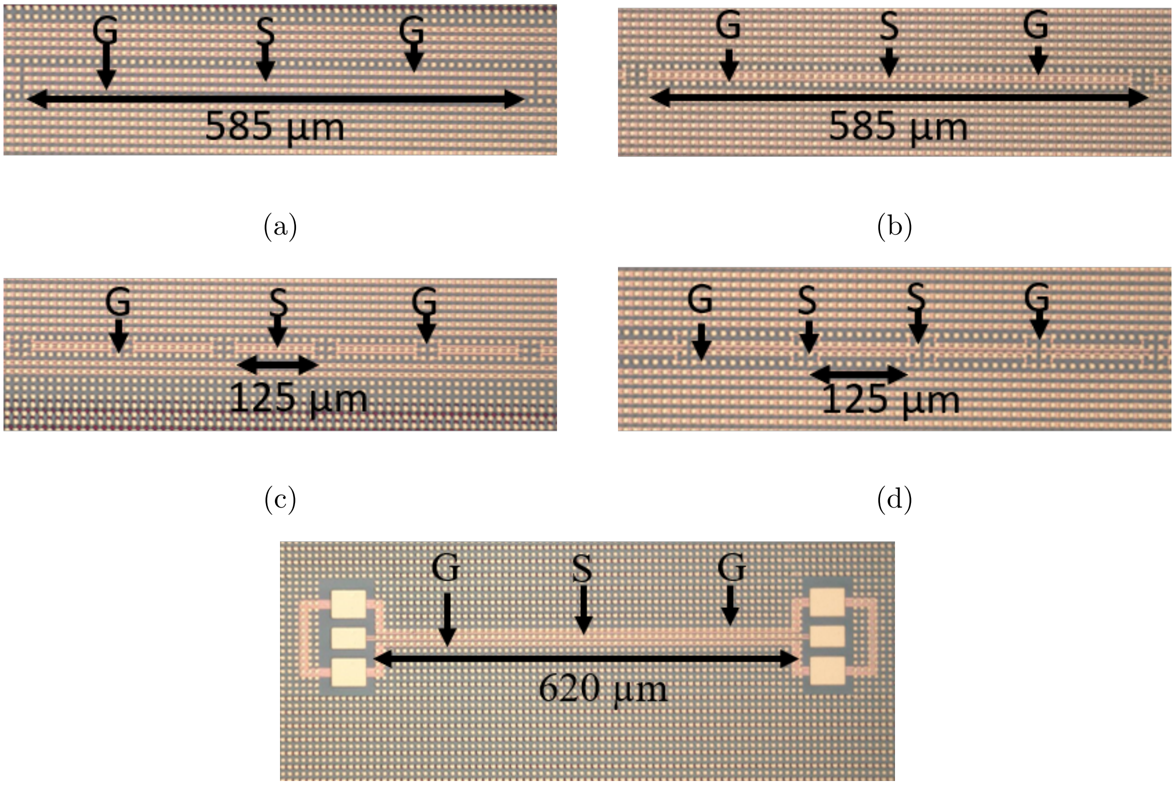


Figure 5.3: Micrograph of the fabricated Si-IF test site consisting of different link segments.

nique, the S-parameters of each link segment is extracted from the cascaded structure. The measured insertion losses (S_{21}) for 585 μm link segments of different wire widths and wiring pitches are shown in Fig. 5.6 (a). The insertion loss of these 585 μm links was found to be <2 dB for frequencies up to 30 GHz. In addition, the measured insertion losses of 2 μm wide, 4 μm pitch links of varying lengths are shown in Fig. 5.6 (b). Accordingly, the measured insertion loss for the 125 μm links is <0.7 dB and the loss for 1.75 mm links is >3 dB for the same frequency range. Furthermore, there is a very good agreement between all the measured characteristics (solid lines) and the simulated values (dashed lines), validating the experimental results. In addition, the insertion loss of SuperCHIPS links is significantly lower than existing interposer technologies [CKL+18, KP14] because of the reduction in link length. Moreover, it is observed that the transfer characteristics of these short SuperCHIPS links have only a single pole. This establishes the RC-like behavior of short links on Si-IF (<500 μm) contrary to the long links on a conventional PCBs (>50 mm) and interposers (3-5 mm) that show an RLC-like resonance at higher frequencies. Further, this re-emphasizes that the inductance of SuperCHIPS interface is



(a)

(b)

(c)

(d)

(e)

Figure 5.4: Optical micrographs of the fabricated link segments on Si-IF. (a) 585 μm GSG link with width: 5 μm, wiring pitch: 10 μm. (b) 585 μm GSG link with width: 2 μm, wiring pitch: 4 μm, (c) 125 μm GSG configured link with width: 2 μm, wiring pitch: 4 μm. (d) 125 μm GSSG configured link width: 2 μm, wiring pitch: 4 μm. (e) De-embedding structure of shorted fan-out wires. In all cases, the link segments were terminated with 10 μm pitch pillar interconnects.

not significant because the link lengths are smaller than the wavelength ($< \lambda/10$) of the propagating EM wave [AN01]. Therefore, like on-chip wires, the short SuperCHIPS links do not have signal reflections during data transfer and consequently, eliminate the need for matching circuitry and complex I/O drivers with equalizers. This permits the use of simple buffers as drivers to significantly reduce the energy/bit to $\leq$0.03 pJ/b.

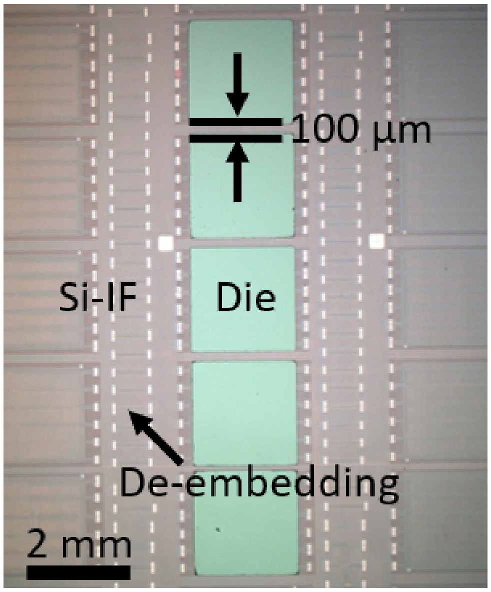
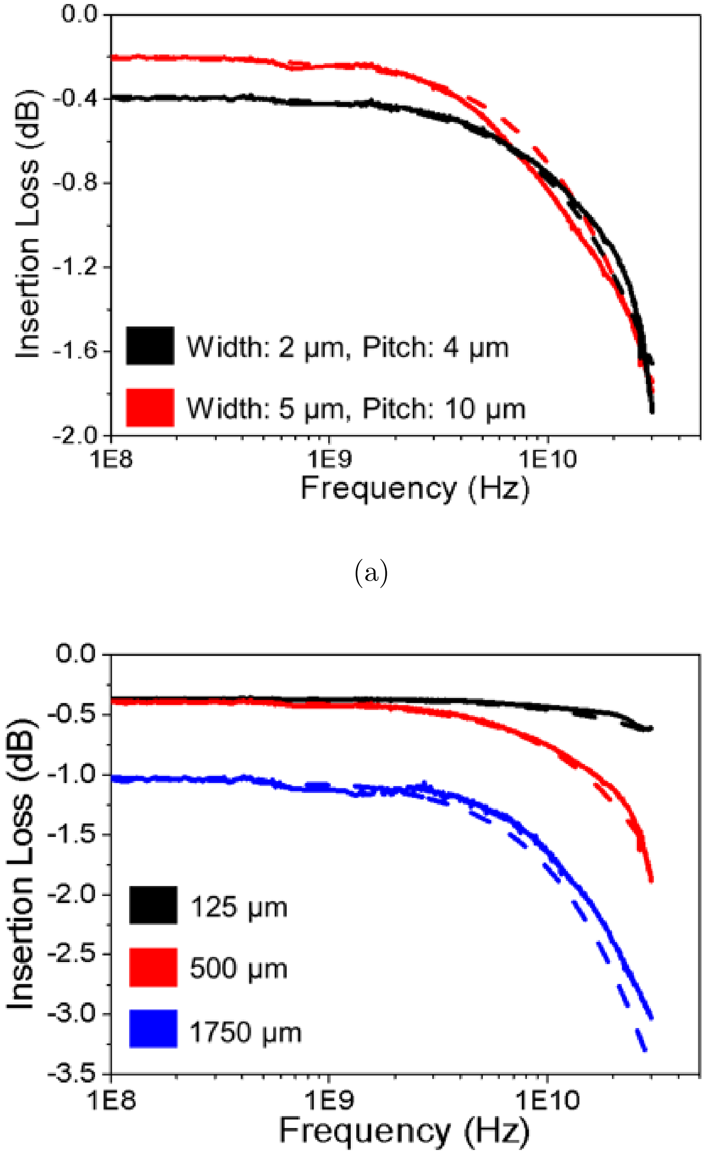


Figure 5.5: Optical micrographs of the test vehicle with 2x2 mm^2 dies bonded to the Si-IF at 100 μm spacing.

5.1.3 Cross-talk

The cross-talk between the SuperCHIPS links is predominantly due to the capacitive coupling between the signal traces rather than the fine-pitch pillars. However, some resistive and inductive coupling also exists due to the shared ground for single-ended signals. Therefore, short link lengths are essential to guarantee low cross-talk. To characterize the cross-talk in the SuperCHIPS links, four-port S-parameter measurements of the GSSG configured links were performed for frequencies from 50 MHz to 20 GHz. Both the ground traces of the GSSG links were shorted, establishing a shared ground for the signals. Using similar methods as described earlier, a four-port de-embedding with T-parameters [FCM08] was used to extract the cross-talk of the short link segments from the cascaded structure. The variation of the near-end cross-talk (NEXT) for 585 μm links with different wire widths and pitches is shown in Fig. 5.7 (a). The NEXT in these links is <-15 dB for the measured frequency range. The cross-talk is relatively higher due to the ground bounce effects because of the shared grounds. Besides, the NEXT for 2 μm wide, 4 μm pitch links of different link lengths are shown in Fig. 5.7 (b). The NEXT for the 125 μm link is <-30 dB,



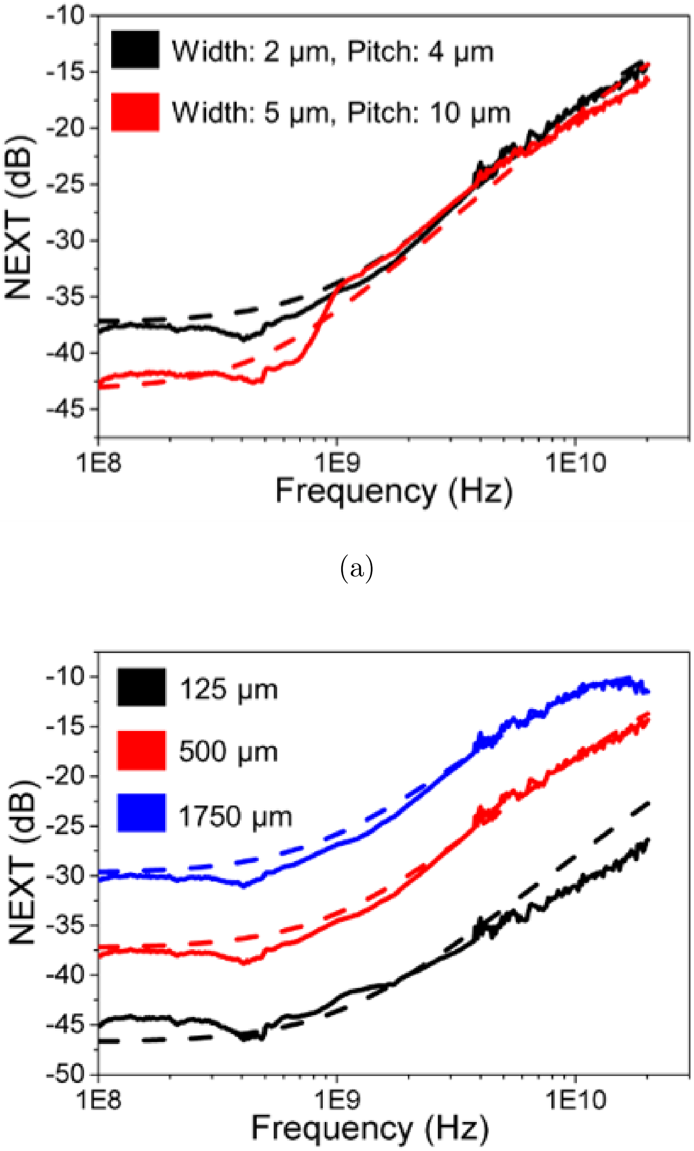
(a)

(b)

Figure 5.6: Plots of insertion loss vs frequency for GSG configured links (solid: measured, dashed: simulated). (a) 585 μm links with different wire widths and pitches showing <2 dB insertion loss. (b) 2 μm width and 4 μm pitch links with different lengths.

and the NEXT for the 1.75 mm link is <-10 dB for the measured frequency range.

Additionally, the far end cross-talk (FEXT) for the 585 μm links with various wire widths and pitches is shown in Fig. 5.8 (a). In both the cases, the FEXT is <-30 dB for



(a)

(b)

Figure 5.7: Plots of NEXT vs frequency for GSSG configured links (solid: measured, dashed: simulated). (a) 585 μm links with different wire widths and pitches showing NEXT of <-15 dB. (b) 2 μm width and 4 μm pitch links with different lengths.

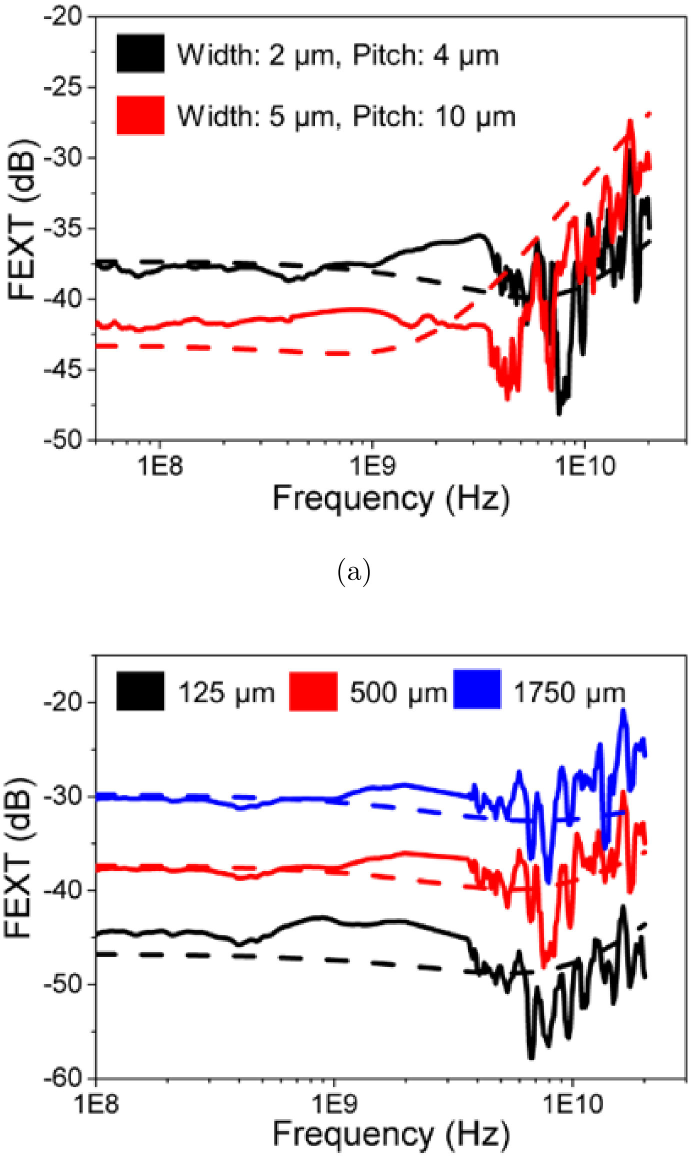
frequencies up to 20 GHz. Similarly, the variation of FEXT for 2 μm wide, 4 μm pitch links of different lengths is shown in Fig. 5.8 (b). The FEXT for the 125 μm link is <-45 dB, and the FEXT for 1.75 mm links is <-20 dB for the same frequency range. The noise in the measurement at higher frequencies can be attributed to noise of the measurement setup that did not have adequent shielding. As shown, all the NEXT and FEXT measurements (solid lines) agree well with the simulations (dashed lines) and are less than a typical acceptable value of -12 dB in all cases.

The insertion loss-to-cross-talk ratio (ICR) which corresponds to the signal-to-noise ratio (SNR) is presented in Fig. 5.9 for varying SuperCHIPS link lengths. When compared to interposer links (3-5 mm) that have an ICR $<$15 dB [BJH^{+}17] at 4 GHz, the short SuperCHIPS links of 500 μm and 125 μm have ICR $>$23 dB, and $>$35 dB respectively. This underlines the necessity for short links to minimize signal degradation and associated driver overhead.

5.1.4 Parasitics Extraction

The measured S-parameters were used to extract the parasitics in the Si-IF links. An RLGC transmission line model of the link was used for the parasitics extraction. However, we established that the short Si-IF links cannot be modeled as transmission lines. Therefore, the S-parameters of the long 1.75 mm links were used for the parasitic extraction. The extracted parasitics are shown in Fig. 5.10. The extracted values include the parasitics of the interconnects and die pads amortized across the length of the wires which is negligible. The extracted resistance, inductance, capacitance, and conductance per unit length are presented in Table. 5.2. As shown, all the measured values (solid lines) concur with the simulation results (dashed lines) and are identical to the on-chip top wiring level parasitics in a 65-90 nm CMOS technology node. Moreover, the variation in the extracted resistance is like the trend observed in [EE92]. Besides, the decrease in measured inductance is consistent with the previous studies [EE92, JAG^{+}17].

Further, the difference in the number of interconnects (Cu-pillars) among different measurements was used to extract the resistance and capacitance of a single pillar, shown in Fig. 5.11. The resistance/pillar is 50-70 mΩ and the capacitance/pillar is 3-4 fF. The parasitics of these interconnects are negligible when compared to the link parasitics and



(a)

(b)

Figure 5.8: Plots of FEXT vs frequency for GSSG configured links (solid: measured, dashed: simulated). (a) 585 μm links with different wire widths and pitches showing FEXT of <-30 dB. (b) 2 μm width and 4 μm pitch links with different lengths.

therefore, can be ignored. Furthermore, this highlights the efficiency of the Si-IF platform to integrate heterogeneous dielets in the same way as functional blocks on a monolithic SoC.

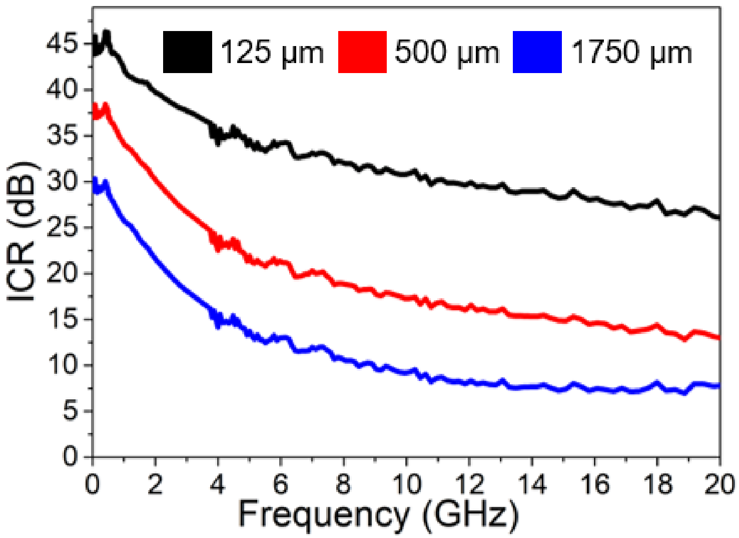
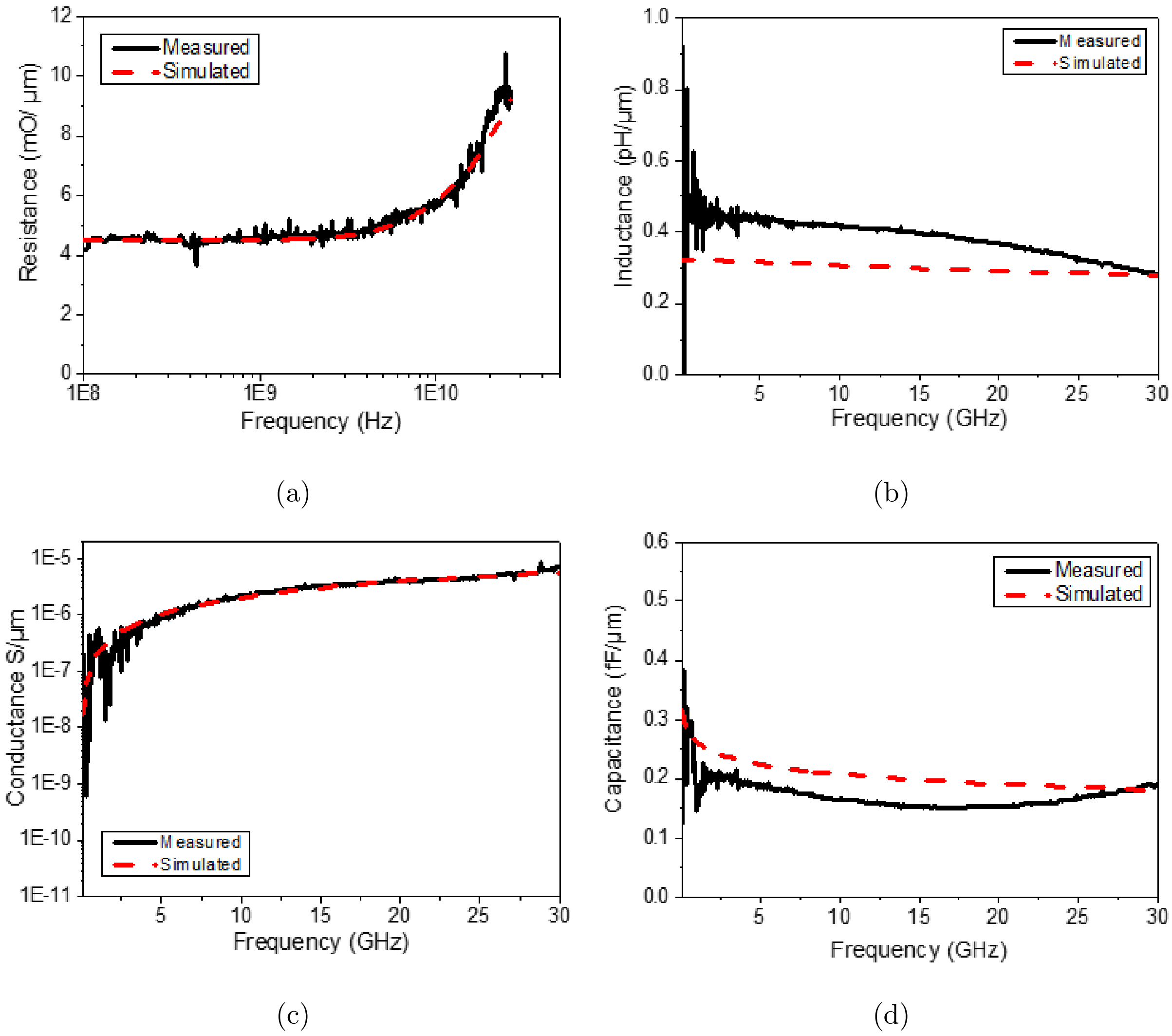


Figure 5.9: Plot of insertion loss to cross-talk ratio (ICR) for 2 μm width and 4 μm pitch Si-IF links with different lengths.

Parasitic per unit length	**Value**	
Resistance (DC)	4.6	mΩ/μm
Capacitance	0.2	fF/μm
Inductance	0.42	pH/μm
Conductance	10^{-6}	Ω^{-1}/μm

Table 5.2: Extracted parasitics of the Si-IF links per unit length.

A comparison of the total parasitic load on the driver using the SuperCHIPS interface on the Si-IF, interposers, and PCB-based assemblies is shown in Table. 5.3. The values presented include the total parasitics of the traces, interconnects, and packages, which is the total load on the driver. The package parasitics are applicable only to the PCB substrates. Besides, the major difference between the interposer and SuperCHIPS interface is the length of the traces. Moreover, the capacitance due to Electro-Static-Discharge (ESD) protection is not included for Si-IF assemblies that can add significant (>0.1 pF) parasitic capacitance. Overall, compared to PCB, the Si-IF has 40-200X lower parasitic inductance and 30-150X lower parasitic capacitance. Compared to interposers, the Si-IF has 20X lower parasitic inductance and capacitance.



(a)

(b)

(c)

(d)

Figure 5.10: The plots of (a) Resistance, (b) Inductance, (c) Conductance, and (d) Capacitance per unit length of the links, extracted from the S-parameters (measured: solid, simulated: dashed) using RLGC line model. The plots show good agreement of the measured data with simulations.

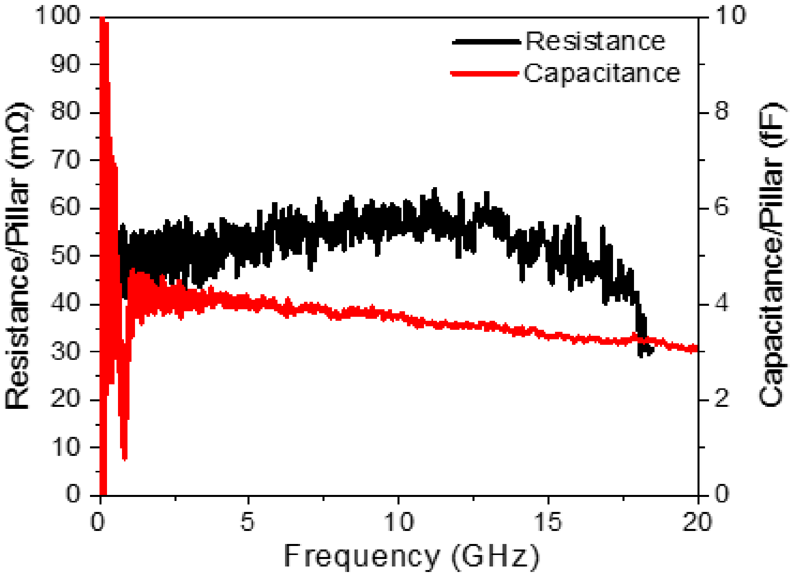


Figure 5.11: The plots of resistance (50-70 mΩ), and capacitance (3-4 fF) per pillar extracted from the measurements.

Technology	**Si-IF**	**Interposer**	**Package**	**PCB**
Interconnect pitch (μm)	10	55	150	1000
Typical link length (mm)	0.25	5	10	50
Inductance (nH) (Normalized to interposer)	0.1 (0.05X)	1.97[a] (1X)	4.25 [KE19] (2.16X)	19.25 [DWC19] (9.77X)
Capacitance (pF) (Normalized to interposer)	0.05 (0.05X)	1.04[a] (1X)	1.68 [KE19] (1.62X)	8.1 [DWC19] (7.79X)

[a] [DWC19, KFK13]

Table 5.3: Comparison of the typical parasitic load on the driver due to the links in different packaging technologies.

5.2 Circuit-level Simulations

The previous section established that the SuperCHIPS interface has low channel loss and cross-talk because of which, simple inverters can be used as transceivers to efficiently stream data between dielets. To validate this theory, circuit simulations were performed with tapered I/O buffer drivers designed using standard cell inverters in TSMC 16 nm technology. Note that the data transfer depends on the driver strength, and the voltage swing and therefore, would change with the die technology. The equivalent driver on-resistance for the buffers used is 250 Ω and the voltage swing is 0.8 V (core voltage). The measured S-parameters and extracted RLGC parameters were used to model the links for a circuit-level simulation study. These RLGC parameters were adjusted to resemble single-ended links and worst-case parasitics. They also include the values corresponding to cross-talk. A practical implementation of the SuperCHIPS interface was considered with 8 fine-pitch single-ended links with a wire width of 2 μm and wire pitch of 5 μm corresponding to two rows of pads per wiring layer as shown in Fig. 5.12. The length of the links was varied from 100 μm to 5 mm to observe the change in characteristics. The input of the driver was presented with a pseudo-random bit stream (PRBS) sequence at various frequencies and the output of the SuperCHIPS link and the receiver were analyzed. The rise and fall time of the input was assumed to be 20 ps which is typical in this technology. Two scenarios were evaluated, (1) without ESD protection circuitry; (2) with ESD protection circuitry shown in Fig. 5.13 (a) & (b) respectively. The ESD protection adds significant load (assumed to be 50 fF per terminal) on the drivers, that is comparable to the parasitics of short links ($\leq$500 μm), increasing the delay and power by almost 2X. Note, due to the low contact area per interconnect and the minimal die handling in the assembly process, the required ESD protection is expected to be lower ($\leq$50 fF) for integration on Si-IF, when compared to a PCB or interposer style integration ($\geq$100 fF) [KVI19, TBV$^+$19]. The results of the simulations are listed below.

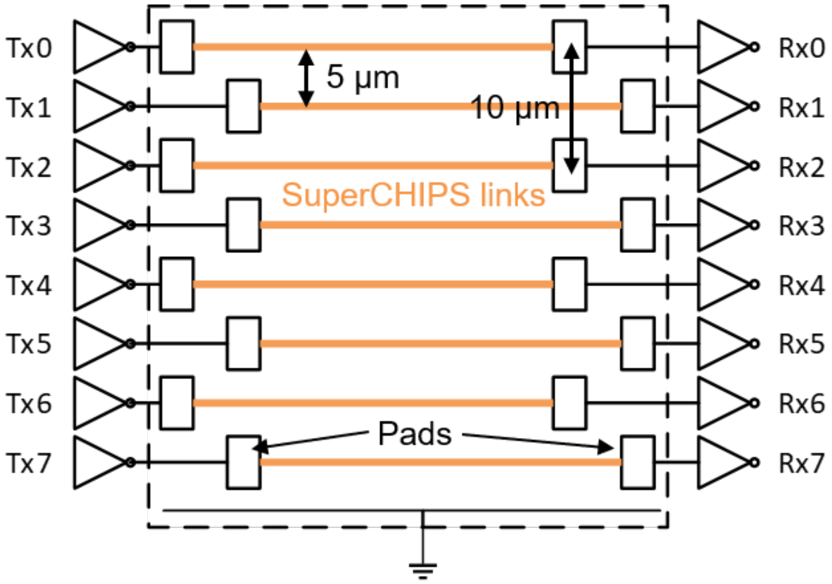
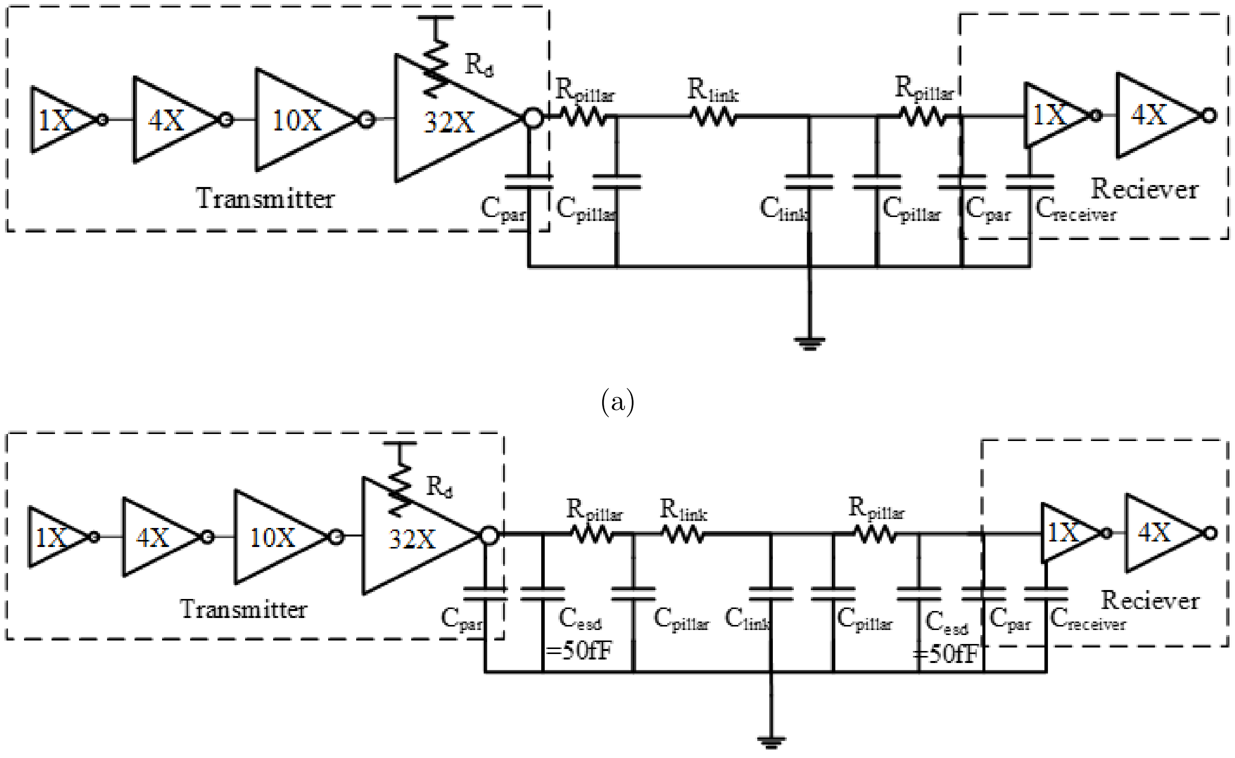


Figure 5.12: Schematic of a SuperCHIPS interface consisting of 8 links used for simulations.

5.2.1 Data-rate

The maximum data-rate achievable per link depends on the driver load and the SNR of the links. The maximum frequency for single-ended links is approximately dictated by the equation shown in (5.1) where t_r is the rise time of the pulse, R_{driver} is the on-resistance of the driver, and C_{link} is the link capacitance including parasitics such as ESD. This directly correlates to the length of the links. The advantage of the Si-IF technology is the availability of fine-pitch ($\leq$10 μm) die-to-substrate interconnects which result in short communication links ($\leq$500 μm). This is not feasible in interposers or PCBs which will be discussed in chapter 7. Therefore, high data-rates (>10 Gbps) can be easily achieved using the short SuperCHIPS links. The simulated eye diagrams of the signal at the output of a 100 μm SuperCHIPS link with and without ESD protection circuitry for 10 Gbps data-rate are shown in Fig. 5.14. As shown, for the case without ESD protection, the eye is completely open with an eye width of 97.5 ps and eye height of $\approx$800 mV. Also, as shown, an ESD protection capacitance of 50 fF changes the transfer characteristics of short links even though the eye width and eye height are similar. Note that the effect of the ESD capacitance diminishes for longer link lengths ($\geq$1 mm) because the link parasitics



(a)

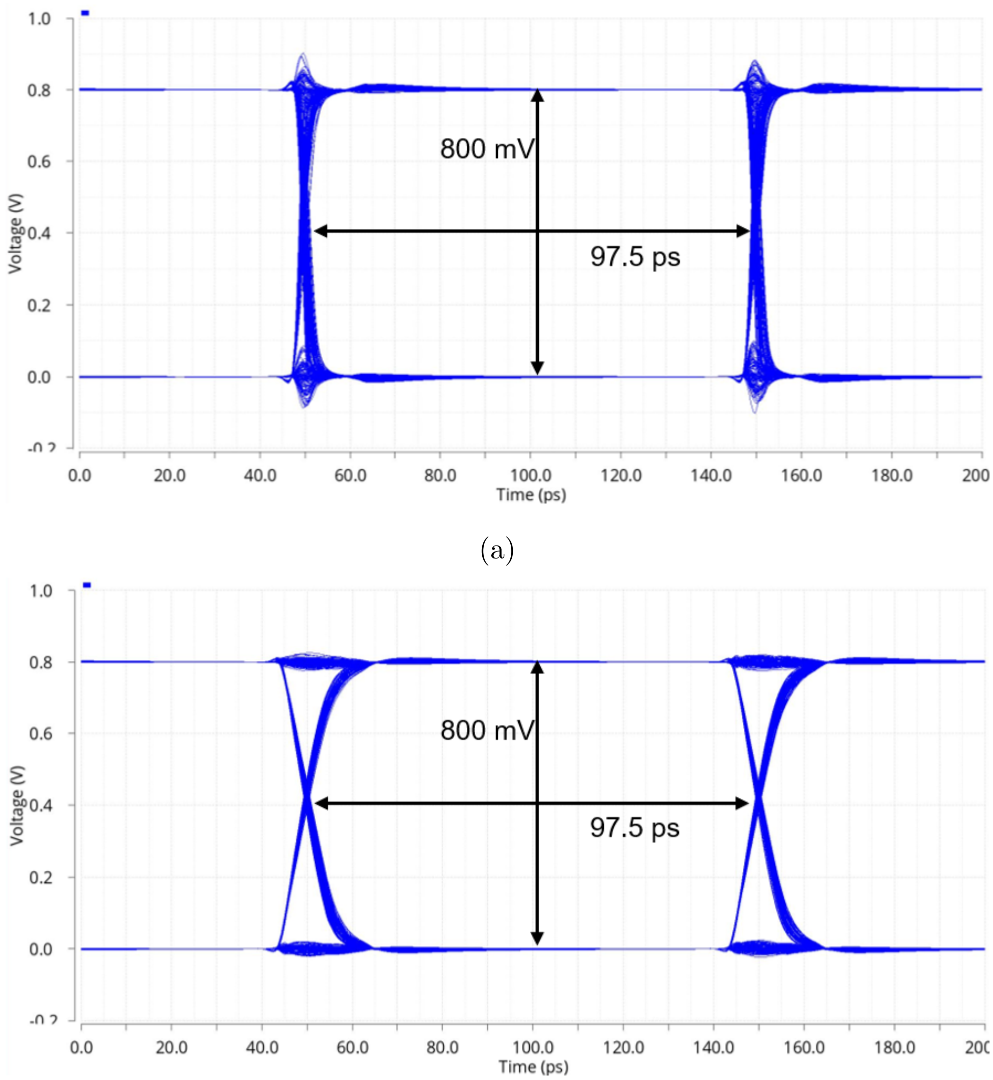
(b)

Figure 5.13: Schematic of simulated transceiver circuits: (a) without ESD protection, and (b) with ESD protection.

are relatively higher. The eye diagram of a 10 Gbps signal at the output of a 500 μm SuperCHIPS link with ESD protection is shown in Fig. 5.15. The eye width is 88.5 ps and the eye height is 780 mV. It can be observed that the eye-opening deteriorates compared to the 100 μm link as expected.

$$Maximum\ Frequency = \frac{0.35}{t_r} = \frac{0.16}{R_{driver} * C_{link}} \tag{5.1}$$

In the simulations, no input jitter was added and the jitter observed is purely due to the SuperCHIPS links only. The plot of the jitter induced by the SuperCHIPS links vs the frequency of data transfer for different link lengths is shown in Fig. 5.16. It can be observed that the induced jitter increases with link length and frequency proportionately. Note that these are simulated values and a real implementation would have higher (20%-30%) jitter.



(a)

(b)

Figure 5.14: Simulated eye diagram for a 100 μm SuperCHIPS link at 10 Gbps data-rate: (a) without ESD protection, and (b) with ESD protection.

Using these simulations, and using the jitter and eye-opening values, the maximum data-rate for a SuperCHIPS link was estimated. The maximum data-rate vs link length and is shown in Fig. 5.17. For asynchronous transfer, the data-rate was estimated using the ICR plot in Fig. 5.9, and a jitter tolerance of 20% unit interval (UI). As shown, data-rates of >10 Gbps can easily be achieved for links <1 mm. The data-rate can be further improved for longer links using shielded or differential signaling identical to other interfaces in typical

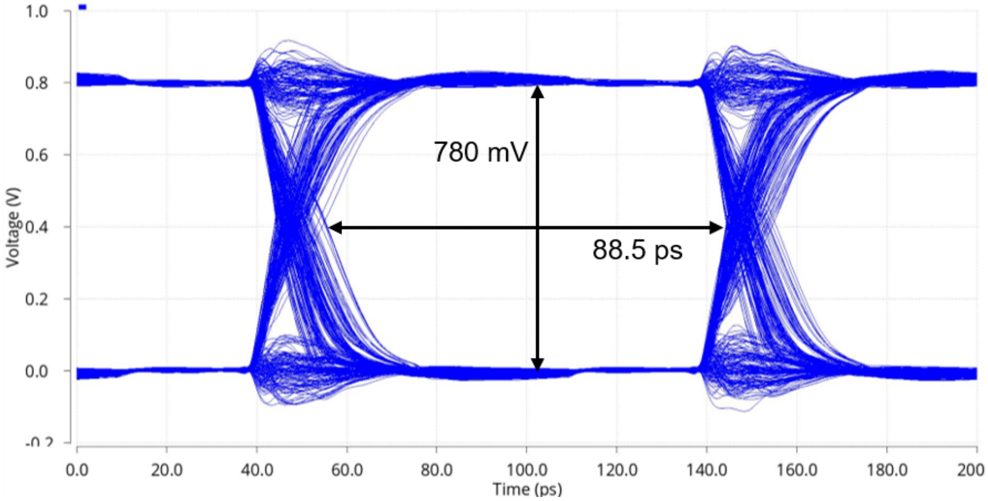


Figure 5.15: Simulated eye diagram for a 500 μm SuperCHIPS link at 10 Gbps data-rate with ESD protection.

packaging substrates. While the data-rate is limited by the driver strength for asynchronous transfer, the clock jitter and uncertainty dominates for synchronous transfer. Generating and distributing a high-speed clock (>2 GHz) with low jitter and distortion is extremely difficult and energy-intensive. Therefore, for synchronous transfer, the data-rates are much lower to account for uncertainty in the clock. Therefore, as shown, short SuperCHIPS links (<500 μm) can support 10 Gbps asynchronous data-transfer and 4 Gbps synchronous double data-rate (DDR) data-transfer.

5.2.2 Bandwidth

As mentioned earlier, because of the fine interconnect pitch, there are a large number of parallel links using SuperCHIPS which contribute to improvement in inter-dielet communication bandwidth. The bandwidth of the SuperCHIPS interface is found by multiplying the data-rate per link with the number of links. To standardize the SuperCHIPS protocol, four data-rates are considered for the synchronous mode of data-transfer including single data-rate (SDR) and double data-rate (DDR) modes that are listed in Table 5.4. The corresponding maximum data-bandwidths are also presented. The values presented assume

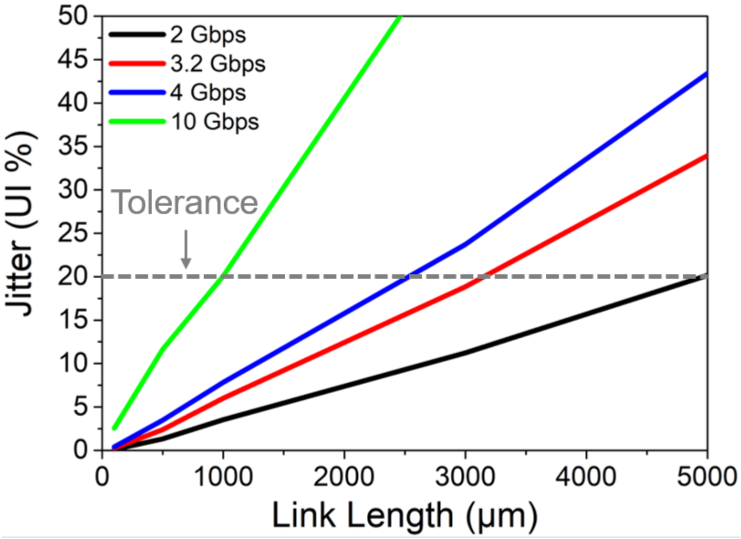


Figure 5.16: Comparison of simulated jitter vs SuperCHIPS link length for different data-rates.

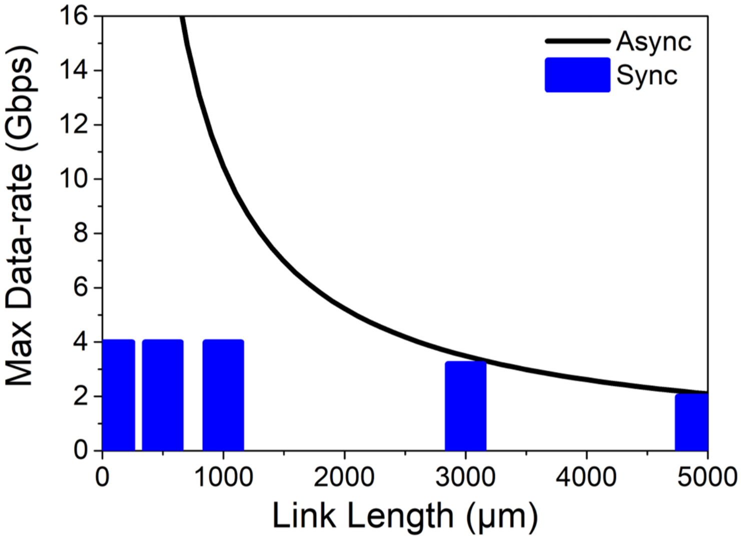


Figure 5.17: Estimated maximum data-rate vs SuperCHIPS link length for both synchronous and asynchronous data transfer. In synchronous case, the clock frequency determines the data-rate and could be increased significantly without penalty for short links (≤500 μm).

four layers of wiring on the Si-IF and single-ended wires at a wiring pitch of 5 μm. Also, it is assumed that all the wires are used for signaling (maximum bandwidth). However, note that this is not true and typically 20% of the edge is allocated for control signals like clock and power wires for a real implementation described in section 6.1.

SuperCHIPS mode	**Clock frequency**	**Data-rate per link (Gbps)**	**Maximum data-bandwidth per die edge (Gbps/mm)**
Synchronous	1	1 (SDR)	800
		2 (DDR)	1600
	2	2 (SDR)	1600
		4 (DDR)	3200
Asynchronous	N.A	up to 10	8000

Table 5.4: Data-rate and maximum bandwidth of the SuperCHIPS protocol.

5.2.3 Latency

The latency introduced by the short SuperCHIPS links is shown in Fig. 5.18, that can be found using standard Elmore delay formulation [KM97] given in 5.2, where $R_{d_{eff}}$ is the effective driver resistance, C_p is the pillar capcitance, C_{link} is the link capacitance, C_{par} is the parasitic capacitance on the driver including the ESD capacitance, and C_r is the receiver capacitance. Note that the pillar resistance ($\mathrm{R_p}$), and the link resitance ($\mathrm{R_{link}}$) are ignored compared to the on-resistance of the driver ($\mathrm{R_d}$) because of the reasons mentioned above in section 5.1. The overall latency of the SuperCHIPS I/O can be found using 5.3, where $t_{overall}$ is the overall latency, t_{Tx} & t_{Rx} are the latencies of transmitter and receiver respectively, and t_{link} is the SuperCHIPS link latency. The simulated waveforms of the input and output of the SuperCHIPS I/O with 500 μm link at 10 Gbps data-rate with ESD protection is shown in Fig. 5.18. The overall latency from the input of the transmitter to the output of the receiver is <26.5 ps and <31.5 ps for the scenarios without and with ESD respectively. The latency added by the SuperCHIPS link compared to just on-chip wire is <14 ps, which is very close to the theoretical value. For synchronous communication, the data can be transfered within 1 clock cycle.

$$t_{link} = R_{d_{eff}} * (2C_p + C_{link} + C_{par} + C_r) \tag{5.2}$$

$$t_{overall} = t_{Tx} + t_{link} + t_{Rx} \tag{5.3}$$

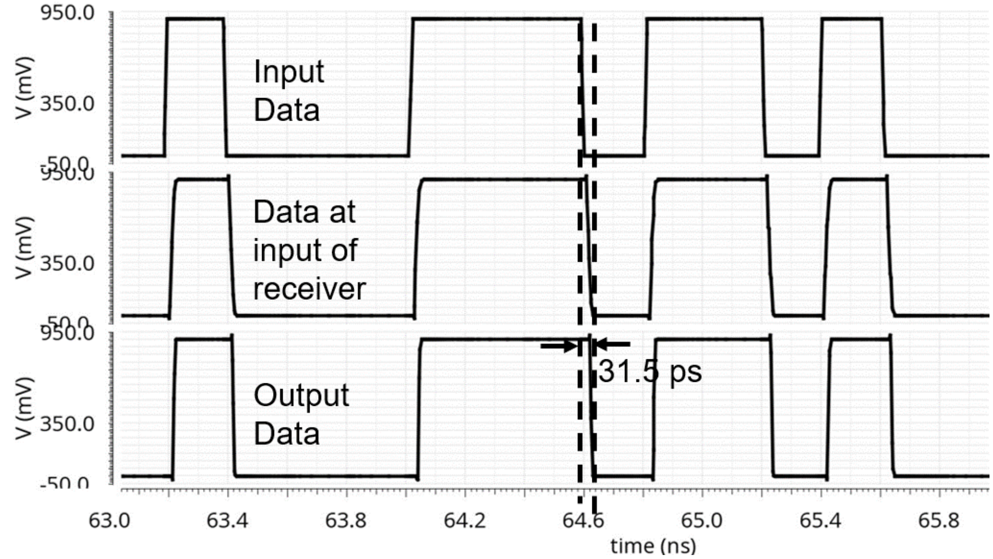


Figure 5.18: Simulated waveforms: 10 Gbps PRBS input; the transmitted data across the link before receiver; the receiver output.

5.2.4 Energy per bit

The use of low loss channels in the SuperCHIPS interface allows for the use of simple buffer I/Os and simple control logic to significantly reduce the energy per bit. The energy per bit variation of the SuperCHIPS protocol with link length is shown in Fig. 5.19. The plot shows the contribution of the I/O without ESD, the contribution of the ESD capacitance, and the contribution of logic and clock for synchronous transfer. The contribution of the I/O and ESD do not change considerably with frequency and the values are presented for a 10 Gbps asynchronous transfer. For the I/O control logic, a modified "lite" version of the Advanced Interface Bus (AIB) [Keh19] soft protocol that is suitable for 10 μm I/O pitch called the Short Near Range-10 (SNR-10) is assumed. This protocol uses simple single-ended unidirectional SuperCHIPS I/Os for data-transfer. The logic energy values presented in Fig. 5.19 are estimated for a 2 GHz clock frequency with a very high activity factor of 50%. As shown, the energy per bit for asynchronous transfer using short SuperCHIPS links ($\leq$500 μm) is $\leq$0.03 pJ/b for the case without ESD protection and $\leq$0.06 pJ/b for the case with ESD protection. For equivalent SuperCHIPS interface with synchronous transfer, the energy per bit is $\leq$0.15 pJ/b that includes the logic, and ESD contributions.

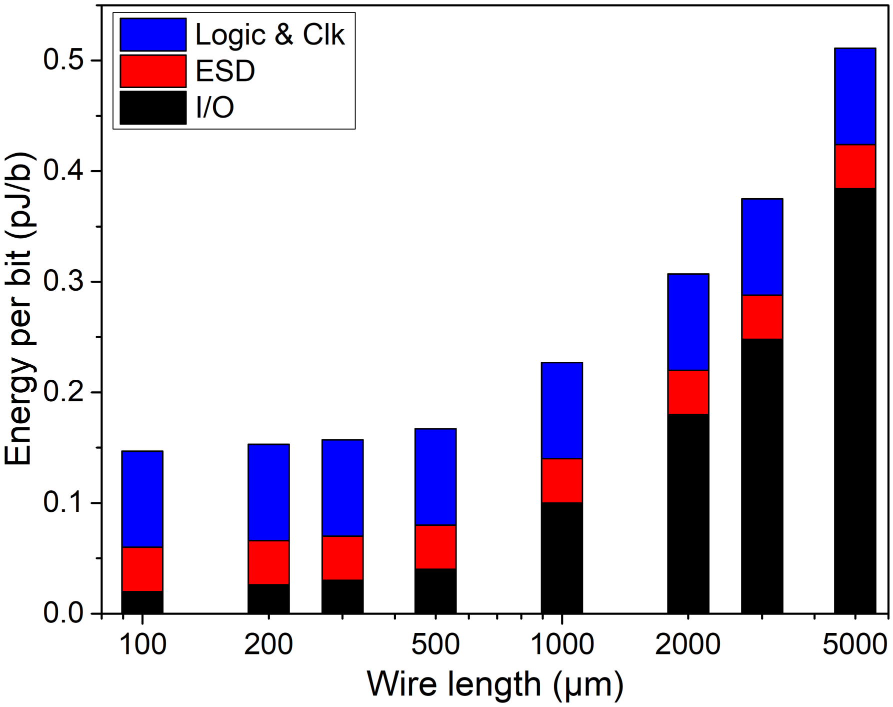


Figure 5.19: Plot of energy per bit vs link length for SuperCHIPS communication for an activity factor of 50%. The energy contributions of the link, ESD, and the logic are shown separately.

CHAPTER 6

Experimental Demonstration of SuperCHIPS

6.1 Test Vehicles

To demonstrate the functionality of the Si-IF assembly and the performance of the SuperCHIPS interface, functional dies and macros were designed in both the TSMC 16 nm finfet (16FF), and GF 22 nm FDSOI (22FDX) technologies. The dies were designed in collaboration with Prof. Markovic's research group at UCLA [Mar] as a part of the DARPA Common Heterogeneous Integration and IP Reuse (CHIPS) program [Gre16]. The die in TSMC 16FF is designed by Prof. Markovic's group to function as a Universal Digital Signal Processor (UDSP). It also includes test macros to test the SuperCHIPS protocol. The die in GF 22FDX, which will be referred to as GF die for simplicity, consists of a neural inference engine designed by Prof. Iyer's group [CHI], but it also includes macros to test the UDSP core and the SuperCHIPS protocol. Furthermore, the UDSP and GF dies were designed to be terminated with 9.8 μm and 10 μm pitch Cu pads respectively. The Cu pad size is $\approx$7x7 μm^2 and is compatible with the Si-IF assembly process instead of the typical Al pads with sizes $\geq$25x25 μm^2. This was accommodated by stopping the wafer processing at the last Cu wiring level in the metal stack (wafer-pull) and passivating with a 200 nm Si_3N_4 layer. Moreover, all the SuperCHIPS I/O transceivers use the existing standard cell buffers with the highest drive strength in a given technology. Further, there is no ESD protection circuitry at the terminals except for antenna diodes.

6.1.1 SuperCHIPS Macros

As mentioned above, test macros were designed in both the technologies to independently test the assembly, measure the latency introduced by the Si-IF links, demonstrate high-speed data-transfer, and estimate the bit-error-rate (BER) of the SuperCHIPS interface. Each die consists of two copies of the test macros that are isolated from the rest of the die. The UDSP die has these test macros placed on the east and west edge of the die so that they can be connected to a test macro in an adjacent die as shown in Fig. 6.6 (a). The GF die has the test macros closely placed, separated by 400 μm to be consistent with the worst-case communication distance between two neighboring dies as shown in Fig. 6.6 (b). These test macros are completely isolated on the chip and communicate only using the Si-IF links to emulate two different dies. The overall schematic of the SuperCHIPS macros is shown in Fig. 6.1. Each macro consists of three modules described below.

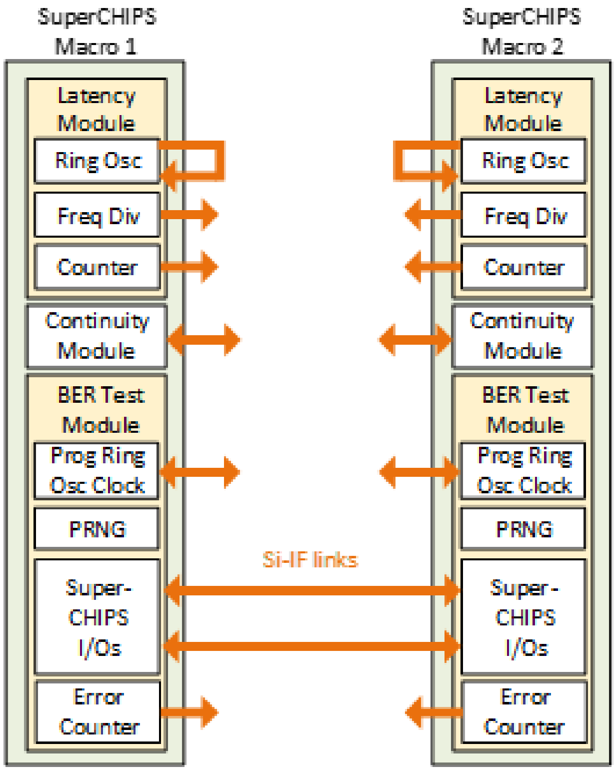


Figure 6.1: Block diagram of the implemented SuperCHIPS test macros showing continuity, latency, and BER characterization modules.

6.1.1.1 Continuity Check

Daisy chain structures were implemented to check for continuity of the electrical links after assembly on the Si-IF. There were two types of daisy chains, namely passive and active chains. The passive chains are simple wires on the die, that when attached to the Si-IF, form a continuous link similar to the ones demonstrated earlier in section 4.5.3. These structures help to debug any misalignment and bonding related failures. The active chains consist of buffers instead of just wires between the pads on the die. Continuity of these chains would ensure no device failures after the assembly process. The schematic of both the daisy chain structures is shown in Fig. 6.2. Multiple of these chains were connected in series to check continuity and verify the assembly process.

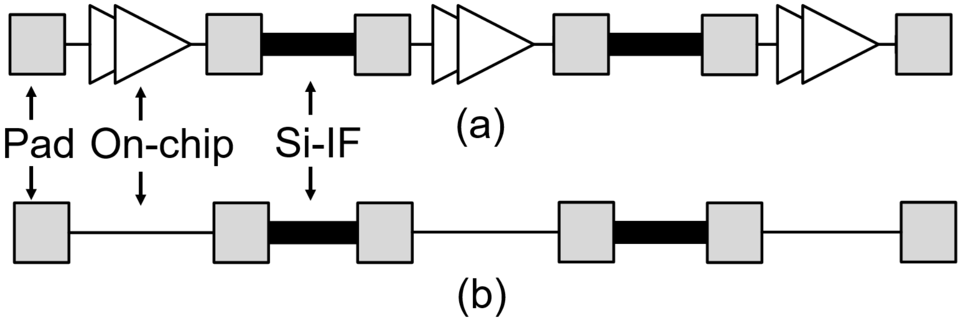


Figure 6.2: (a) Schematic of an active daisy chain consisting of buffers on the die and links on the Si-IF connected alternatively. (b) Schematic of a passive daisy chain consisting of wires on the die and links on the Si-IF connected alternatively.

6.1.1.2 Latency Characterization

To characterize the latency introduced by the Si-IF links, two identical sets of buffer delay and inverting delay blocks were designed. One of these sets was internally connected on-chip to form a ring oscillator which acts as a reference. The second set was terminated with Cu pads that are connected externally using the Si-IF links to form a ring oscillator. The schematic of the testing circuit is shown in Fig. 6.3. The I/Os of both these blocks are consistent with the SuperCHIPS buffers and the schematic shown earlier in Fig. 5.1 (b) & Fig. 5.13 (a).

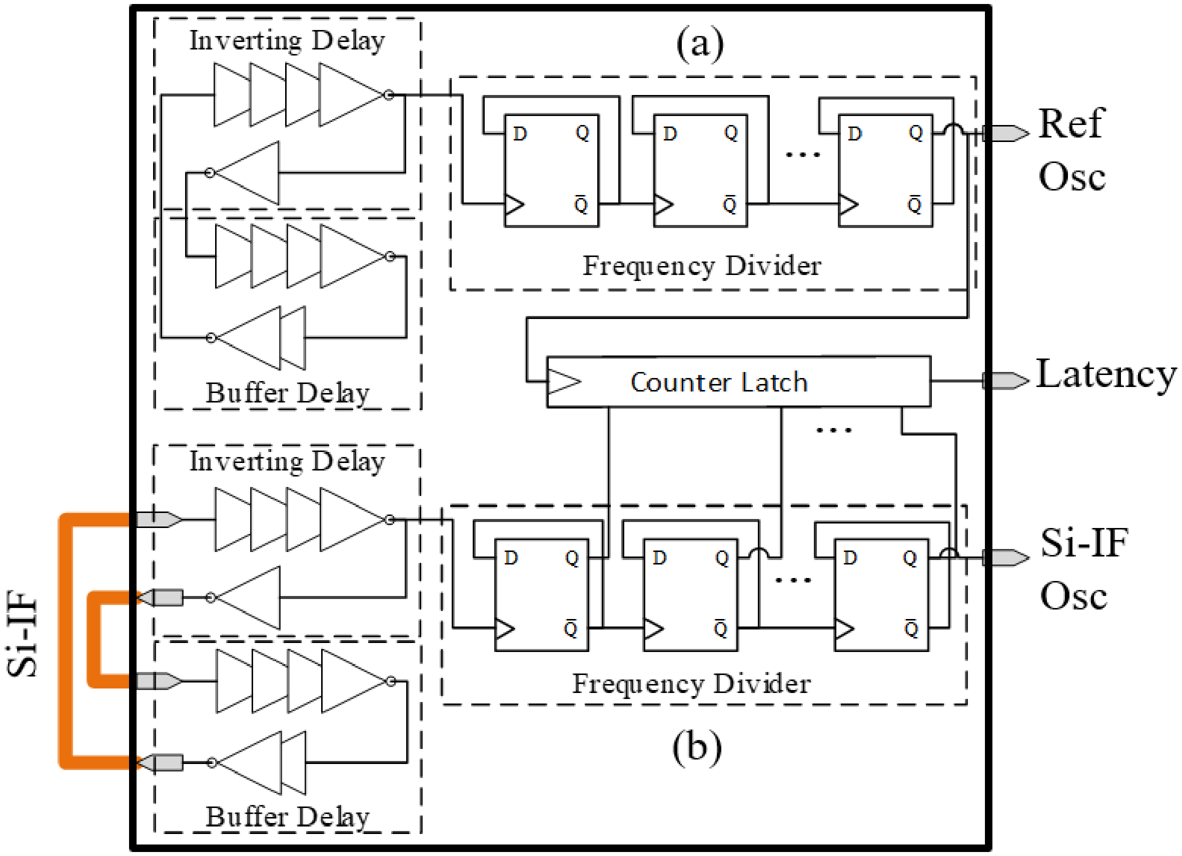


Figure 6.3: Schematic of the latency characterization module. (a) On-chip reference ring oscillator. (b) Ring oscillator formed by connecting the buffer delay and inverting delay blocks using Si-IF links. Also shown are the frequency divider and the on-chip counter latch to measure latency.

The delay due to the link can be found by (5.2). For on-chip oscillator, the t_{link} is small and can be ignored. Also, the C_r and C_{par} in the die technologies are small compared to the Si-IF link capacitance. Therefore, the time period of the two oscillators is shown below.

$$T_{Si-IF} = 2(t_{inv} + t_{buf} + 2t_{link}) \tag{6.1}$$

$$T_{ref} = 2(t_{inv} + t_{buf}) \tag{6.2}$$

where $T_{\textit{Si-IF}}$ is the time period of the oscillator with Si-IF links, T_{ref} is the time period of the reference oscillator, t_{inv} is the delay of the inverting block, and the t_{buf} is the delay of the buffer block. Using (6.1) and (6.2), the latency corresponding to Si-IF links can be found by (6.3).

$$t_{link} = (T_{Siif} - T_{ref})/4 \tag{6.3}$$

The reference oscillator will resonate at a higher frequency compared to the oscillator with Si-IF links. Moreover, the frequency of the oscillator through Si-IF depends on the length of the links. The reference oscillator was designed to resonate at 3-4 GHz for both the dies. Measuring these high-speed signals directly is challenging, therefore, a 12-stage on-chip frequency divider is implemented to reduce the output frequency by 2^{12} to get a measurable waveform, Fig. 6.16. Consequently, the latency of the Si-IF links can be found by dividing the measured delay in (6.3) by 2^{12}. Additionally, there is an on-chip counter latch designed to quantify the exact difference between the number of cycles of the oscillators within a given time.

6.1.1.3 High-speed Data Transfer & Bit Error Rate (BER) Estimation

For high-speed data transfer and BER measurement, two identical copies of the test macros connected using the SuperCHIPS interface are used. Each macro consists of 8 transmitters and 8 receivers to send and receive data using 16 SuperCHIPS links. As mentioned earlier, the I/O circuits use the available standard-cell buffers, and registers for data transfer. Further, we designed an 8-bit Pseudo-Random Number Generator (PRNG) to generate an 8-bit input data for the SuperCHIPS links. The data is transmitted from one macro to the other using the SuperCHIPS interface consisting of ≈450 μm links. Subsequently, the received data is compared with the generated data internally at the receiver and the bit errors are counted using an on-chip error counter. For the GF die, both the macros on a single die can be connected to do the measurements. However, for the UDSP die, the test macros of two neighboring dies have to be connected. As mentioned earlier, the SuperCHIPS links have very low latency and can support up to 10 Gbps data-rate per link.

But the clock jitter and synchronization limit this data-rate. Therefore, we incorporated a programmable ring oscillator to generate the clock at variable frequencies from 500 MHz to 3 GHz. Consequently, we can measure the BER vs the clock frequency. In addition, the programmable ring oscillator clock frequency is divided by 2^8 and sent as an output for observation. The schematic of the test macro is shown in Fig. 6.4.

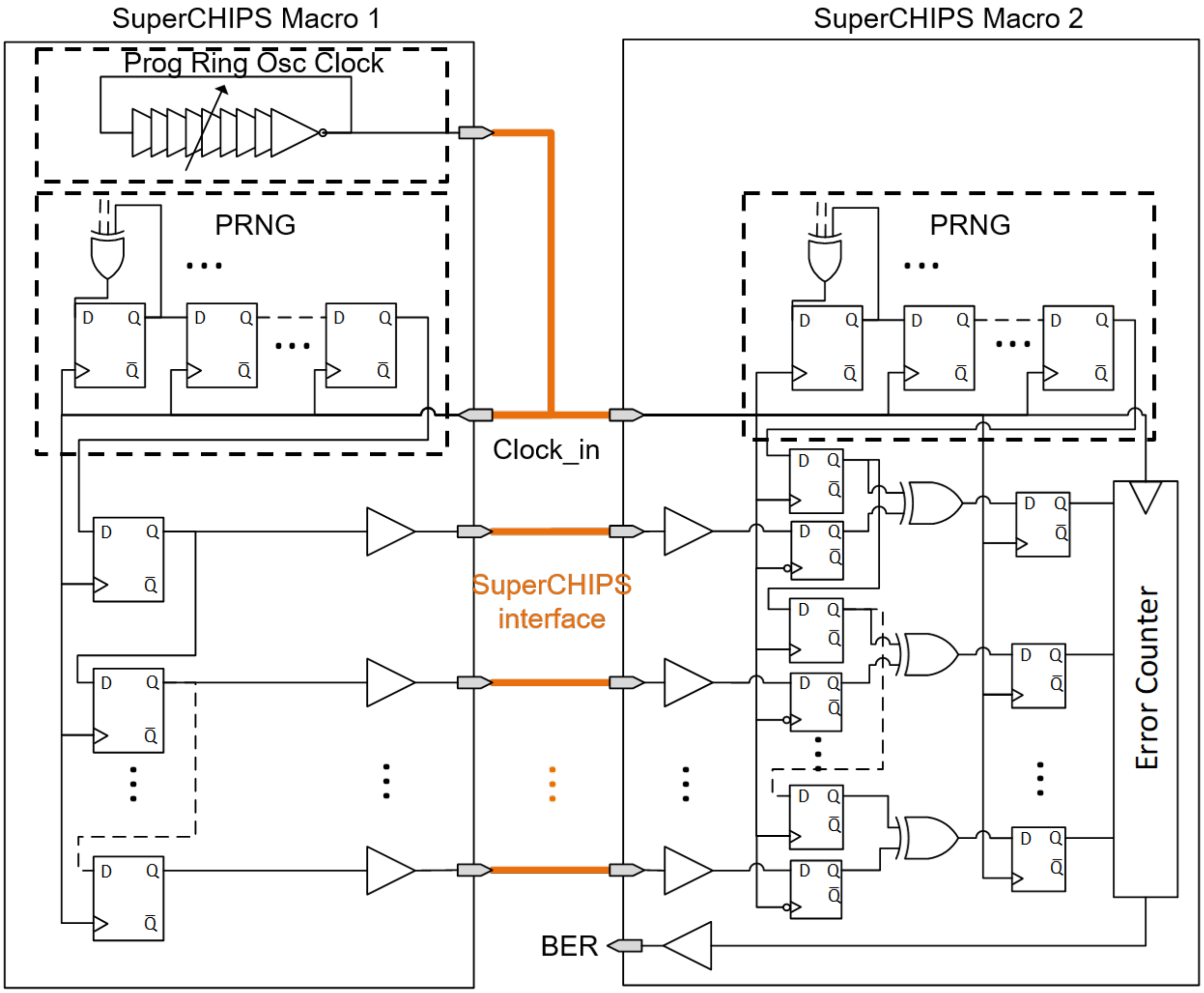


Figure 6.4: Schematic of the high-speed data transfer and BER measurement circuit showing the two macros connected with the SuperCHIPS interface. Both macros are identical although a simplified transmitter schematic is shown on the left. The macros consist of a programmable ring oscillator as the clock, PRNG, comparator, and error counter.

6.1.2 Universal Digital Signal Processor

The UDSP die is designed to be a digital signal processor to perform signal processing computations on inputs. It consists of many small cores that are interconnected at different logical hierarchies. What makes the UDSP universal is that these connections between the cores can be re-programmed to perform many different functions like a field-programmable gate array (FPGA) [WYYM14]. As mentioned earlier, the design and implementation of the UDSP die is by Prof. Markovic's group and is beyond the scope of this thesis. From a system point of view, the UDSP design can be scaled to incorporate a vast number of cores to improve both functionality and performance. However, practical limitations of the die size restrict the number of these cores. But, if multiple dies can be efficiently integrated such that the inter-die core-to-core communication latency, and bandwidth are comparable to the values within a single die, then the UDSP system can be extended to include a massive number of cores. This is where the Si-IF platform provides the fine-pitch integration that satisfies these requirements. Further, the SuperCHIPS protocol provides a simple, low latency, low energy, and high bandwidth interface for core-to-core communication comparable to on-chip metrics.

As a result, the UDSP is designed in TSMC 16FF with a die size of 2.5x2.5 mm^2, consisting of 196 cores that communicate to neighboring dies using the SuperCHIPS interface. Moreover, as a part of DARPA's effort to standardize the communication protocol between dies [Gre16], a modified version of the Advanced Interface Bus (AIB) protocol [Keh19] was adopted as the soft protocol. However, the AIB protocol is designed for interposer/EMIB style integration and has several features for an end-to-end solution that are not needed for simple neighboring die communication. Therefore, a "lite" version of AIB was implemented in the UDSP prototypes, called the Short Near Range-10 (SNR-10) which is suitable for sub-10 μm pitch I/Os that is simple and compatible with AIB. The SuperCHIPS I/Os were designed to conform to this communication standard. The UDSP consists of 8 of these channels on each side and each channel consists of 32 input (Rx) and 32 output (Tx) links. Apart from data-channels, the UDSP also consists of a control channel that is used to program and control the UDSP, and a phase-locked loop (PLL) IP to generate high-speed clock up to 3 GHz on-chip. The UDSP was designed to work at clock frequencies up to 1 GHz. The schematic of the UDSP is shown in Fig. 6.5.

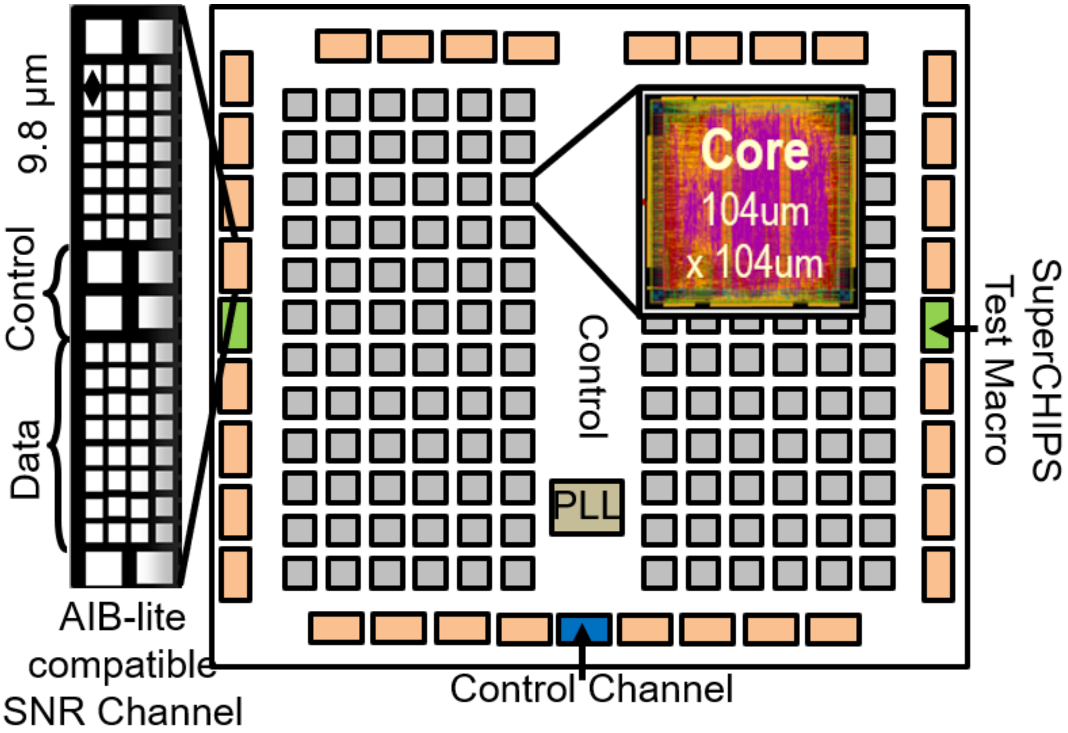
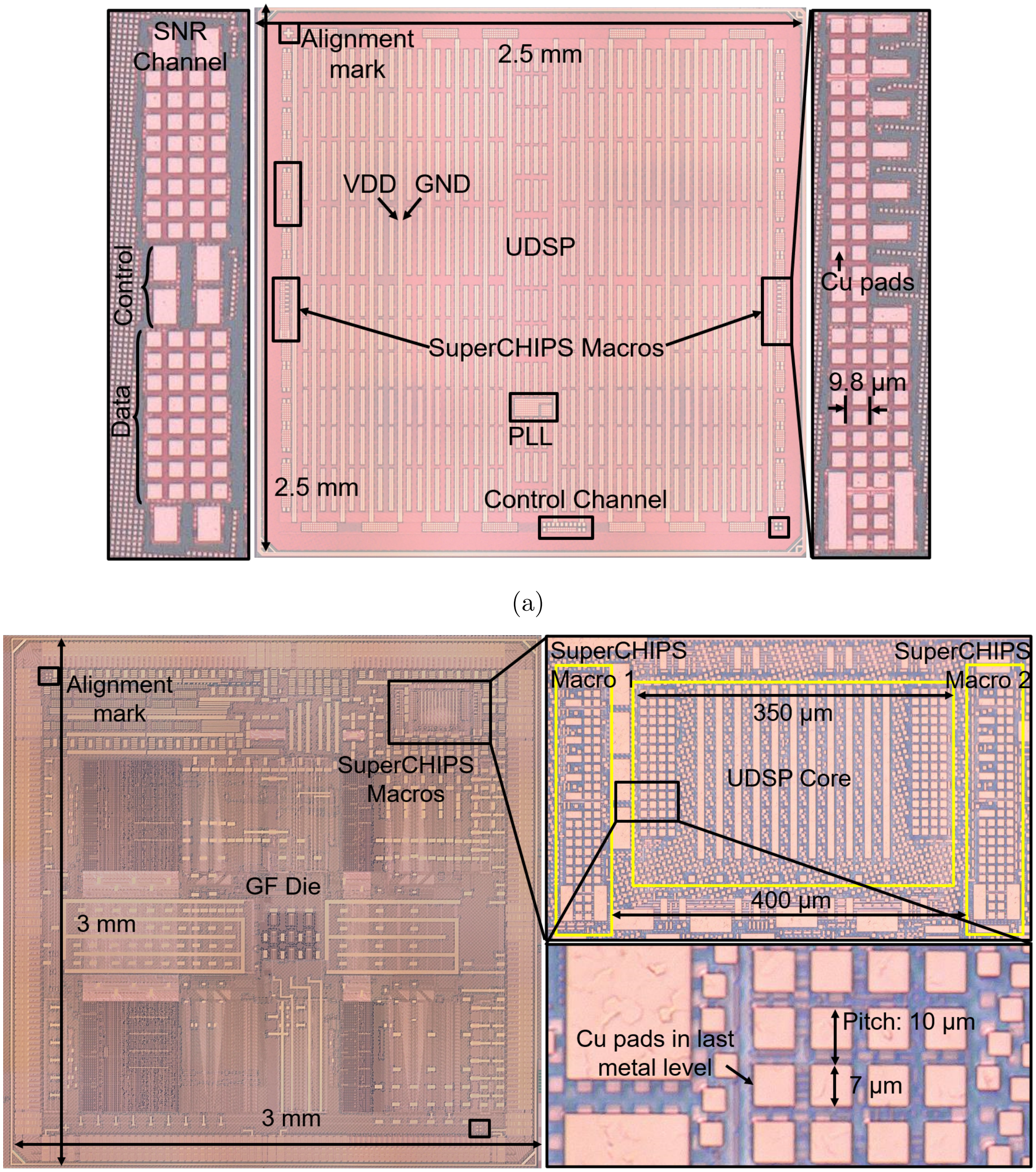


Figure 6.5: Schematic of the UDSP die showing different components of the UDSP (Courtesy: Prof. Markovic's Group [Mar]).

A macro of four UDSP cores with 2 SNR channels is also included in the tape-out of the GF die which is shown in Fig. 6.6 (b). In addition, the UDSP die also includes the SuperCHIPS macros discussed above.

6.1.3 Fabricated Dies

The fabricated UDSP die in TSMC 16FF is presented in Fig. 6.6 (a), showing different components of the die including the Cu metal termination and the 9.8 μm pitch pads. The fabricated GF die is shown in Fig. 6.6 (b) and consists of the SuperCHIPS and UDSP test macros.



(a)

(b)

Figure 6.6: Fabricated dies- (a) UDSP in TSMC 16 nm finfet technology, (b) Die in GF 22 nm FDSOI technology, showing fine-pitch Cu metal termination and the SuperCHIPS channels & macros.

6.2 Design and Fabrication of Si-IF

Three different assemblies were demonstrated on the Si-IF as listed below.

1. Single UDSP on the Si-IF to demonstrate the functionality of the die. The goal is to perform a low-speed test to validate the UDSP cores and the assembly.

2. Single GF die on the Si-IF to characterize the SuperCHIPS macros and verify the functionality of the UDSP core.

3. An array of 2x2 UDSPs integrated on the Si-IF at close spacing ($\leq$55 μm) to demonstrate the high-speed inter-die communication using SuperCHIPS, and the functionality of the system.

Three different Si-IFs were designed to test the three assemblies mentioned above. All the Si-IFs were fabricated with two wiring levels that include both the signal and power wiring. The Si-IFs were terminated with 4 μm diameter Cu pillars at 9.8 μm and 10 μm pitch for the UDSP and GF dies respectively. All the external I/Os were fan-ed out to pads at the periphery with 100 μm pitch for testing and wire-bonding to board. All three Si-IF test sites were placed on a single wafer and processed together. The Si-IFs were fabricated at the UCLA facilities using the process mentioned earlier in chapter 3. The details of individual test site design and fabrication are presented below.

6.2.1 Single UDSP

The Si-IF platform can support numerous I/Os because of the fine-pitch interconnects. However, all the I/Os cannot be fan-ed out to periphery pads for wire-bonding because of the larger pitch. Therefore, for the single UDSP testing, only the control channel, and two SuperCHIPS data-channels, each corresponding to 64 I/Os (32 Tx, 32 Rx) were fan-ed out to be tested. The I/O pads on the die bond to 4 μm diameter Cu-pillars and two wiring levels are used to fan-out the connections to periphery pads. Also, the power is distributed using multiple Cu-pillars across the UDSP die. Moreover, a single row of wire-bonding pads was used that connect to two rows of staggered pads on the testing board. This is limited by the PCB pitch, routing, and wire-bonding complexity. Also, the Si-IF includes test pads for the continuity check of the SuperCHIPS macro daisy chains to validate successful assembly. The Si-IF size is 5.5x5.5 mm^2 and the fabricated Si-IF is shown in Fig. 6.7. The design consists of 7574 pillar interconnects of which 308 are for signal transfer, 5737 are for power transfer, and the rest are dummy pillars.

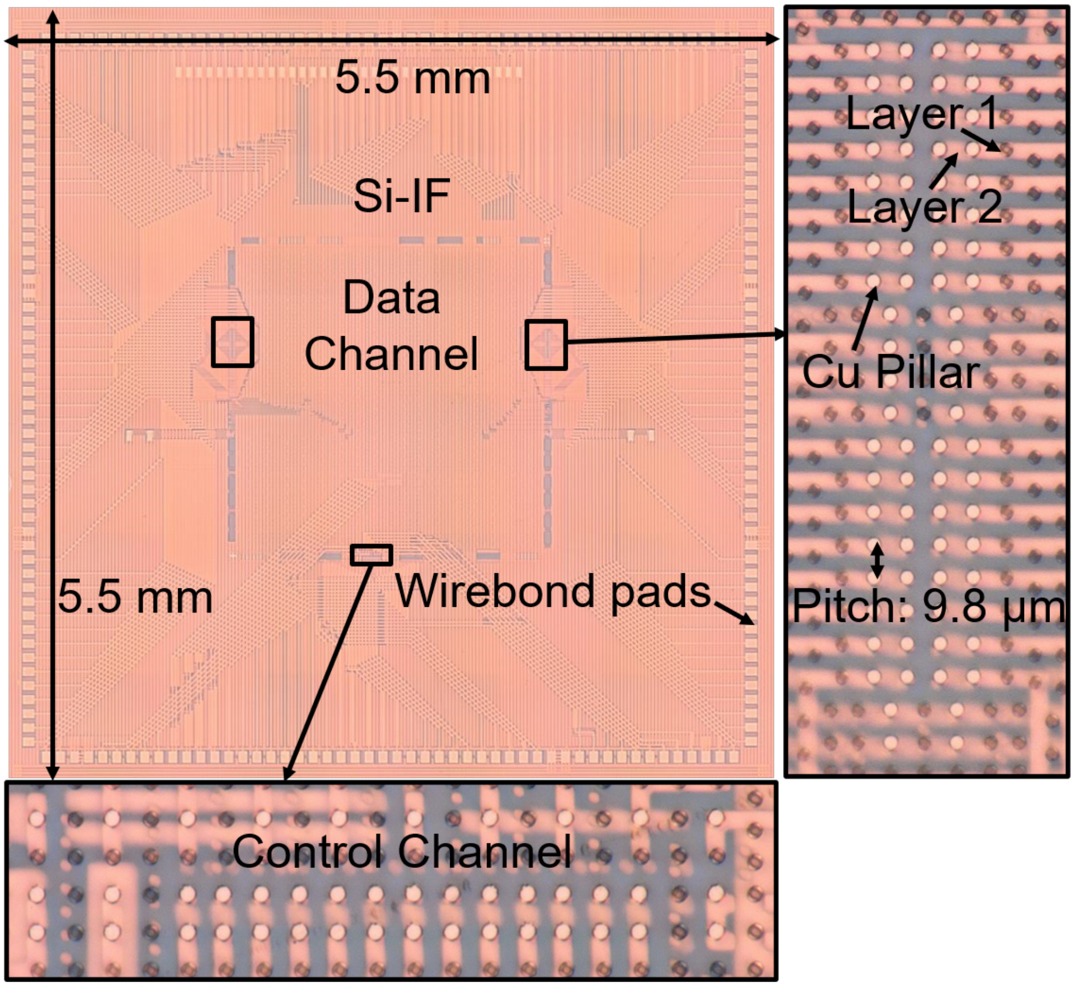


Figure 6.7: Micrograph of the fabricated Si-IF for single UDSP assembly. Inset shows the data and control channels with 9.8 μm Cu pillars that are fan-ed out to pads for wire-bonding.

6.2.2 Single GF Die

The Si-IF for the GF die is designed to be compatible with the single UDSP testing board consisting of a similar layout of the periphery pads and the 5.5x5.5 mm^2 Si-IF size. A second row of periphery pads was also included for additional test points and to serve as alternative wire-bonding pads. The pads of the daisy chain structures were connected in series of three using the Si-IF links and fan-ed out to probing pads. Also, the ring oscillator of the two macros on each die was connected using the Si-IF links of two different lengths: 200 μm, and 500 μm. This allows for measuring the change in latency vs the link length. The SuperCHIPS links connecting the two BER modules are $\approx$450 μm long. The wire width is 1.5 μm and the wiring pitch is 5 μm. The fabricated Si-IF is shown in Fig. 6.8.

Note that only 364 pillar interconnects are required for the test, but additional 15,480 dummy pillar interconnects were included for the mechanical stability and bond strength of the assembly. Moreover, other test macros that are related to the neural engine on the die were also fan-ed out to the probe pads for testing as shown in Fig. 6.8. The testing of these macros is beyond the scope of this work.

6.2.3 2x2 UDSP System

To demonstrate a functional UDSP system, an 8x8 mm^2 Si-IF was designed to integrate four UDSPs in a 2x2 array. The fabricated Si-IF is shown in Fig. 6.9. Two adjacent UDSPs are connected using the short SuperCHIPS channels of length $\approx$350 μm as shown. The wire width is 1.5 μm and the wiring pitch is 4.9 μm. The inter-dielet spacing is crucial in achieving the desired SuperCHIPS link lengths and one has to account for the die fabrication shrinkage, physical die size after dicing which is typically larger than the design, and variation in the die edge due to dicing. The SuperCHIPS channels between the UDSPs can communicate using either synchronous mode at core clock frequency or asynchronous mode. There a total of 8 SuperCHIPS channels between two adjacent UDSPs that correspond to a bandwidth of 512 Gbps between the dies. Few of the periphery SuperCHIPS channels

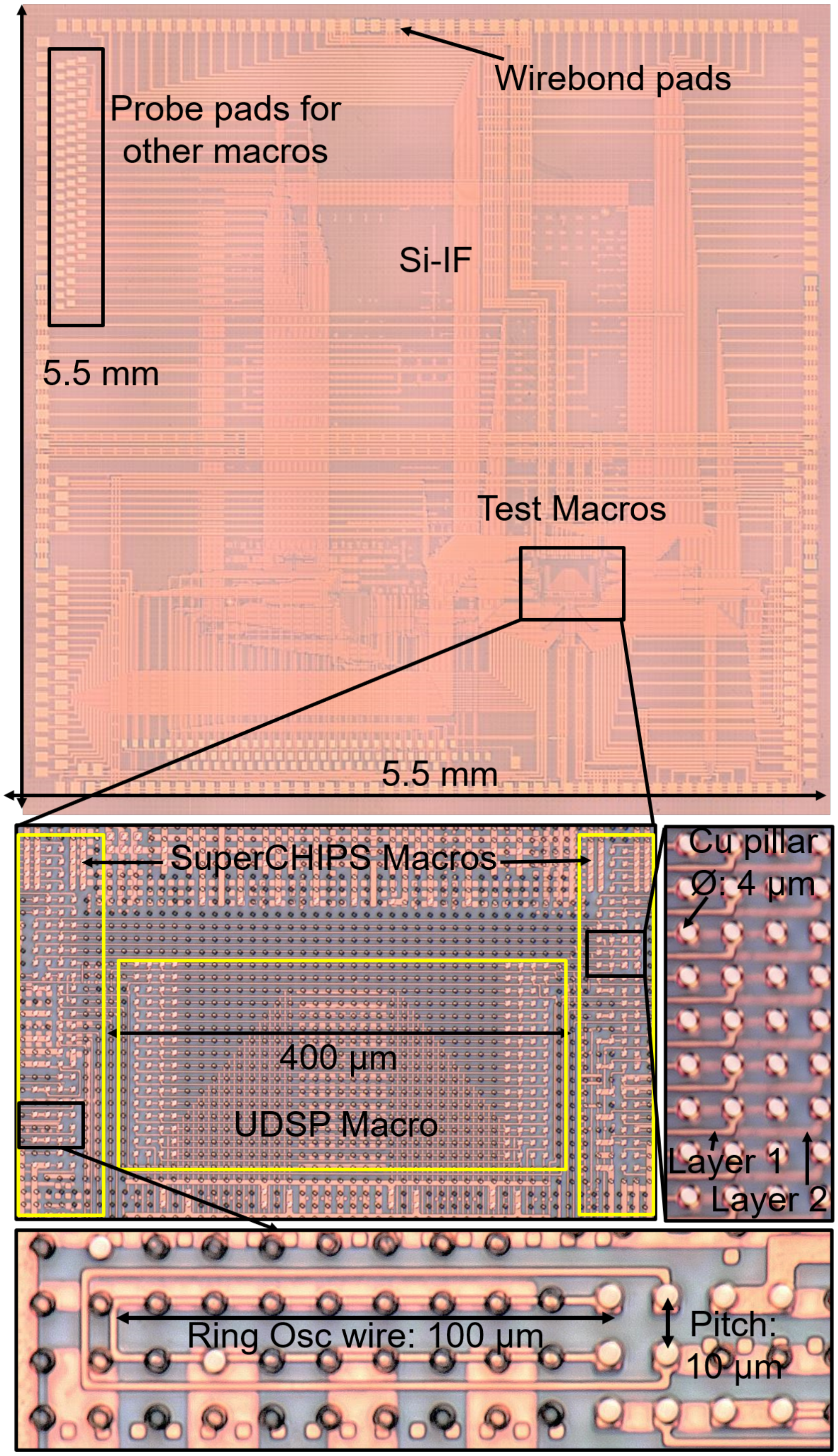


Figure 6.8: Micrograph of the fabricated Si-IF for the GF die assembly. Insets show the test macros, Si-IF wires connecting the ring oscillator, and 10 μm pitch Cu pillars.

are fan-ed out to serve as external data I/O to the system. There are a total of 30,402 pillar interconnects for the system of which 8,111 are for signal transfer and 22,291 are for power transfer.

The 2x2 UDSP system is designed to function at high frequency and a corresponding clock cannot be reliably supplied externally from the board. Therefore, a slow reference clock is given as input which is distributed on the Si-IF using a simple H-tree and given to all the UDSPs. This ensures the same clock delay and skews for all the UDSPs. Moreover, another H-tree is used to distribute a high-speed clock generated by the PLL of one of the UDSPs to all the others. This serves as the core clock which ensures synchronicity across all the UDSPs. Therefore, the 2x2 UDSP accomplishes high-bandwidth (512 Gbps) data-transfer between two adjacent dies. Although this is nowhere close to a wafer-scale system, it is the first step in demonstrating the performance of the Si-IF platform and SuperCHIPS communication.

Apart from the UDSP, the SuperCHIPS test macros on all the dies were also connected using the Si-IF. The continuity daisy chains extend across two dies in series to verify the bonding of both the dies. In addition, similar to the Si-IF for GF die, the ring oscillators were connected using the Si-IF links of two different lengths, 200 μm and 500 μm, to measure the difference in latency. For high-speed data transfer and BER measurement, two SuperCHIPS macros on adjacent dies were interconnected as shown in Fig. 6.9. The ring oscillator and BER modules are routed to peripheral wire-bonding pads on the second row as shown and share some of the bonding pads with the data-channel on the PCB. Therefore, for a sample, either the data-channel or the SuperCHIPS macro can be bonded and tested.

6.3 Assembly

As mentioned earlier, the dies were terminated with pads at the last Cu metal wiring level and passivated with 200 nm of Si_3N_4 layer. Before bonding to the Si-IF, this passivation

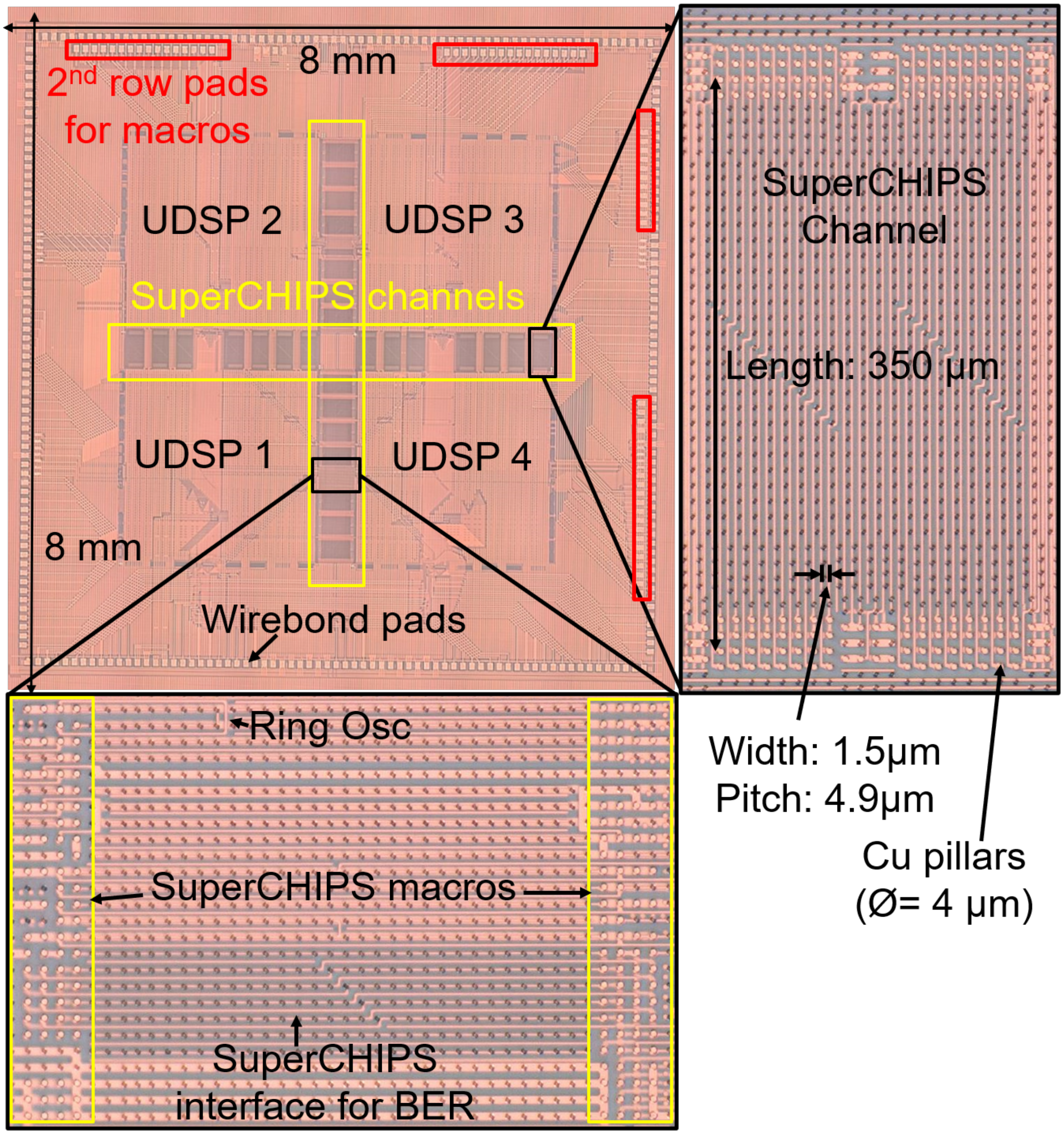
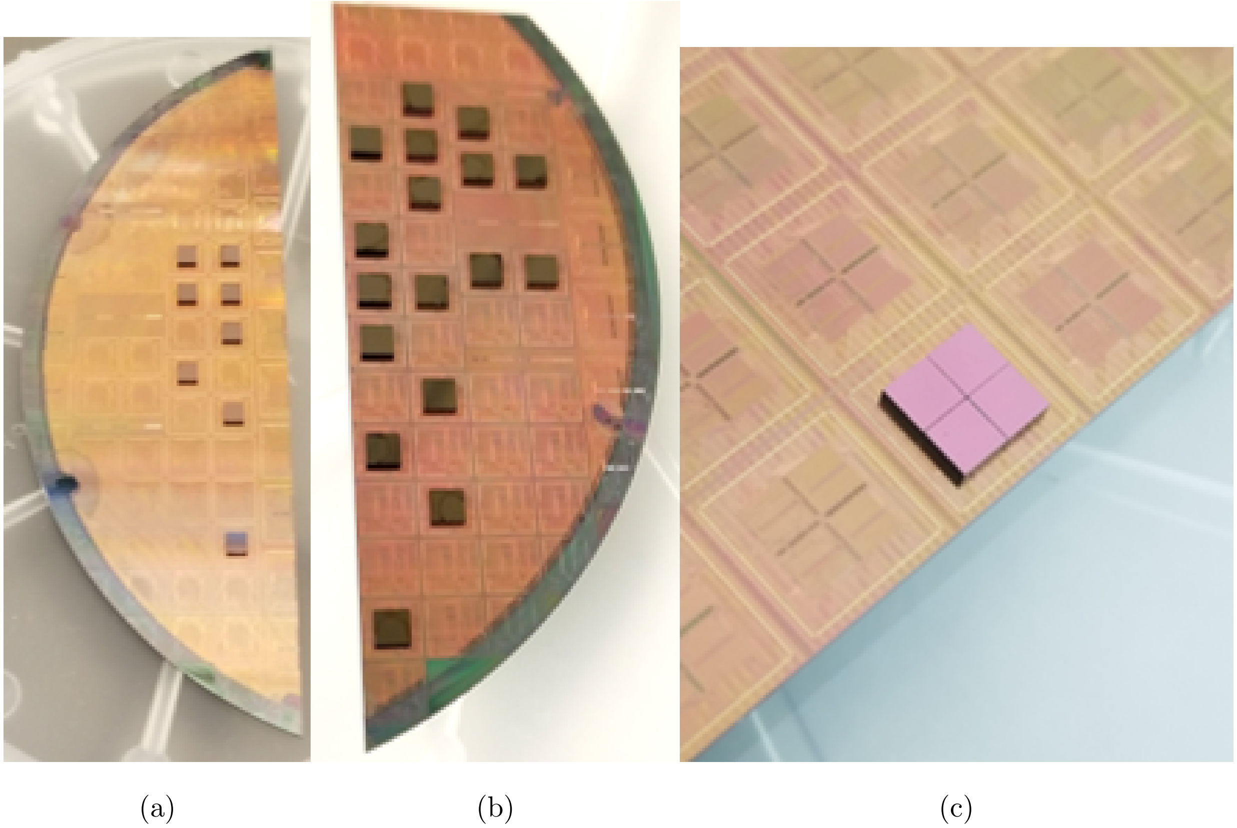


Figure 6.9: Micrograph of the fabricated Si-IF for assembly of 2x2 UDSP dies. Insets show the SuperCHIPS channels between the dies and the connections between the SuperCHIPS macros of two dies.

on the dies was removed using a dry etch of the Si_3N_4 layer, exposing the Cu pads. Subsequently, the dies and the Si-IF were treated with Ar-plasma for 3 min to remove any surface contamination. Using the direct Cu-Cu TCB process described in section 4.4, the dies were precision aligned and bonded to the Si-IF. The process parameters used are given in Table 4.3. All the different sites on the Si-IF and their corresponding die alignment marks were individually taught before assembly. All the different sites were first bonded on the Si-IF wafer as shown in Fig. 6.10. The Si-IF wafer is later diced to separate individual assemblies. The individual assemblies are presented below.


(a)

(b)

(c)

Figure 6.10: Multiple dies assembled on the Si-IF wafer before dicing: (a) Single UDSPs, (b) GF dies, (c) 2x2 array of UDSPs.

6.3.1 Assembly of Single UDSP on Si-IF

The micrograph of a single UDSP assembled on the Si-IF is shown in Fig. 6.11. As shown, the die size is 2.5x2.5 mm^2 and the Si-IF size is 5.5x5.5 mm^2. The periphery wire-bond pads were used for preliminary probe testing and later wire-bonded to the testing PCB.

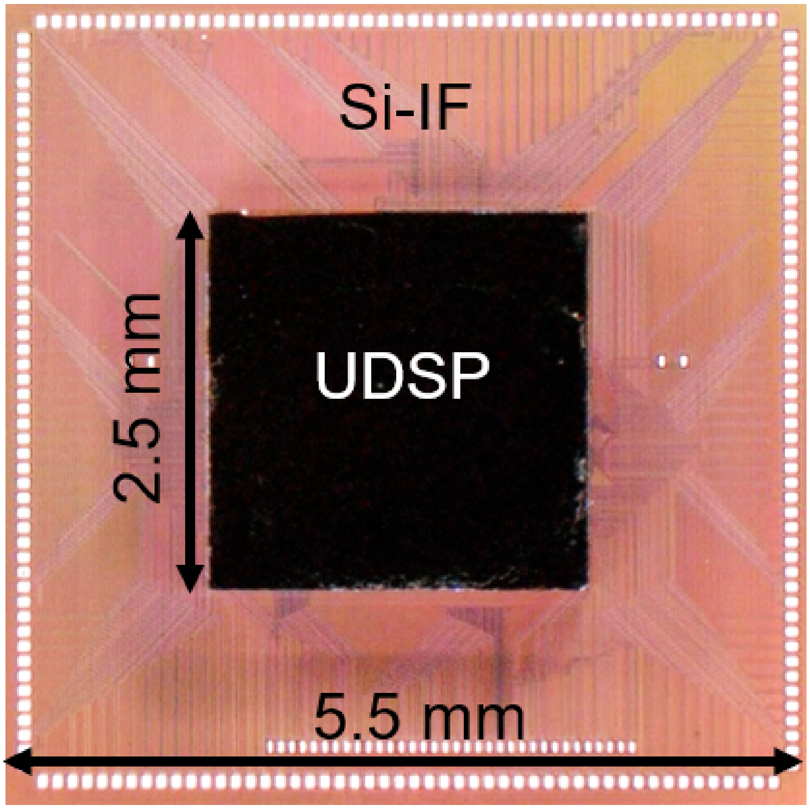


Figure 6.11: Micrograph of a single UDSP die assembled on the Si-IF.

6.3.2 Assembly of Single GF die on Si-IF

The micrograph of a single GF die assembled on the Si-IF is shown in Fig. 6.12. As shown, the die size is 3x3 mm^2 and the Si-IF size is 5.5x5.5 mm^2. The periphery wire-bond pad layout is the same as a single UDSP Si-IF. However, only some of these pads correspond to the SuperCHIPS and UDSP core macros. Preliminary probe testing was performed using this assembled Si-IF before wire-bonding to the PCB.

6.3.3 Assembly of 2x2 UDSPs on Si-IF

The micrograph of a 2x2 UDSP array assembled on the 8x8 mm^2 Si-IF is shown in Fig. 6.13. The inset shows the inter-dielet spacing between two adjacent UDSPs is $\approx$55 μm. As mentioned earlier, the variation in die size after dicing constraints this spacing. For the UDSP die, the die edge left-over after dicing is 30±5 μm larger than the design size on each

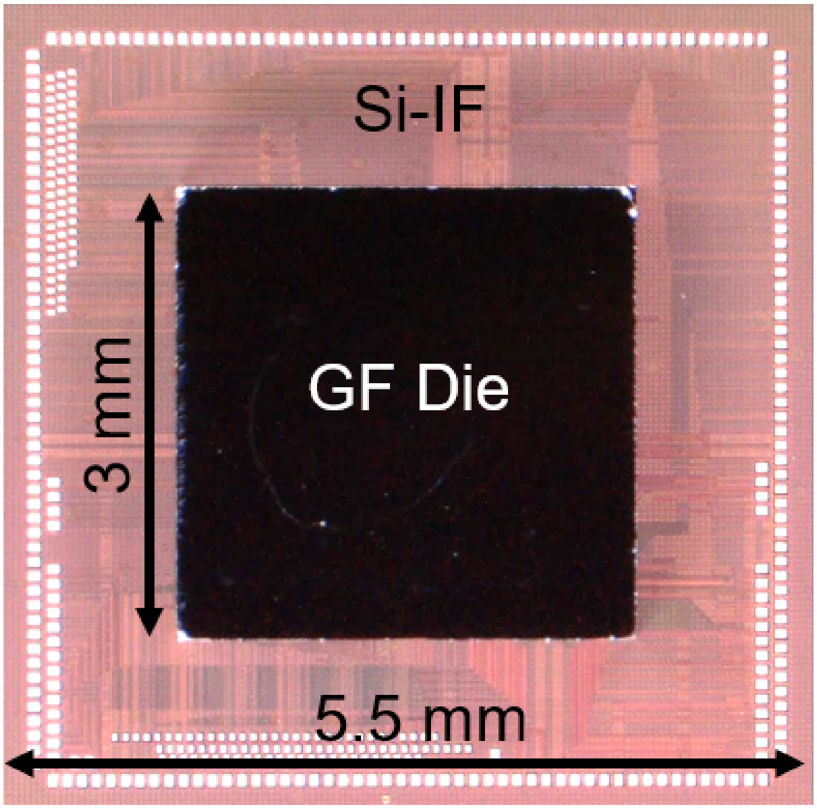


Figure 6.12: Micrograph of the GF die assembled on the Si-IF.

side. The four UDSPs are bonded sequentially on the Si-IF with no intermediate cleaning process other than the in-situ formic acid treatment process. Once again, preliminary probe tests were performed as mentioned above before wire-bonding to the PCB.

6.3.4 Assembly of Si-IFs on PCB

For the complete functional testing of the dielet assemblies, the Si-IFs were mounted on testing PCBs and wire-bonded. Two different PCBs were designed to test the single die assemblies, and the 2x2 UDSP assemblies. Because there was no packaging of the Si-IF, the assemblies had to be directly wire-bonded to the PCBs which presented some challenges. However, the Si-IFs were successfully wire-bonded to the PCBs and the final assemblies of the three different samples is shown in Fig. 6.14. An FPGA was used for programming, and interfacing with the boards to perform tests.

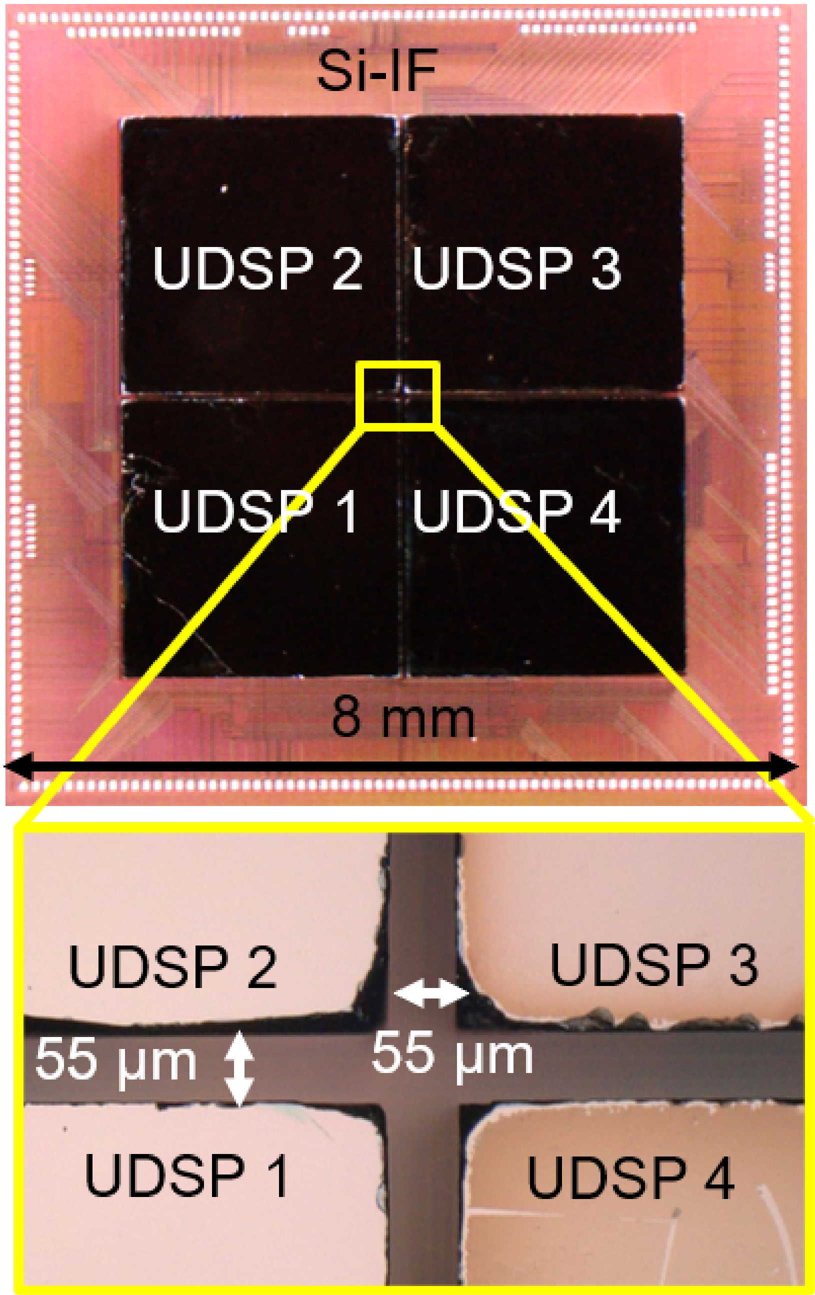
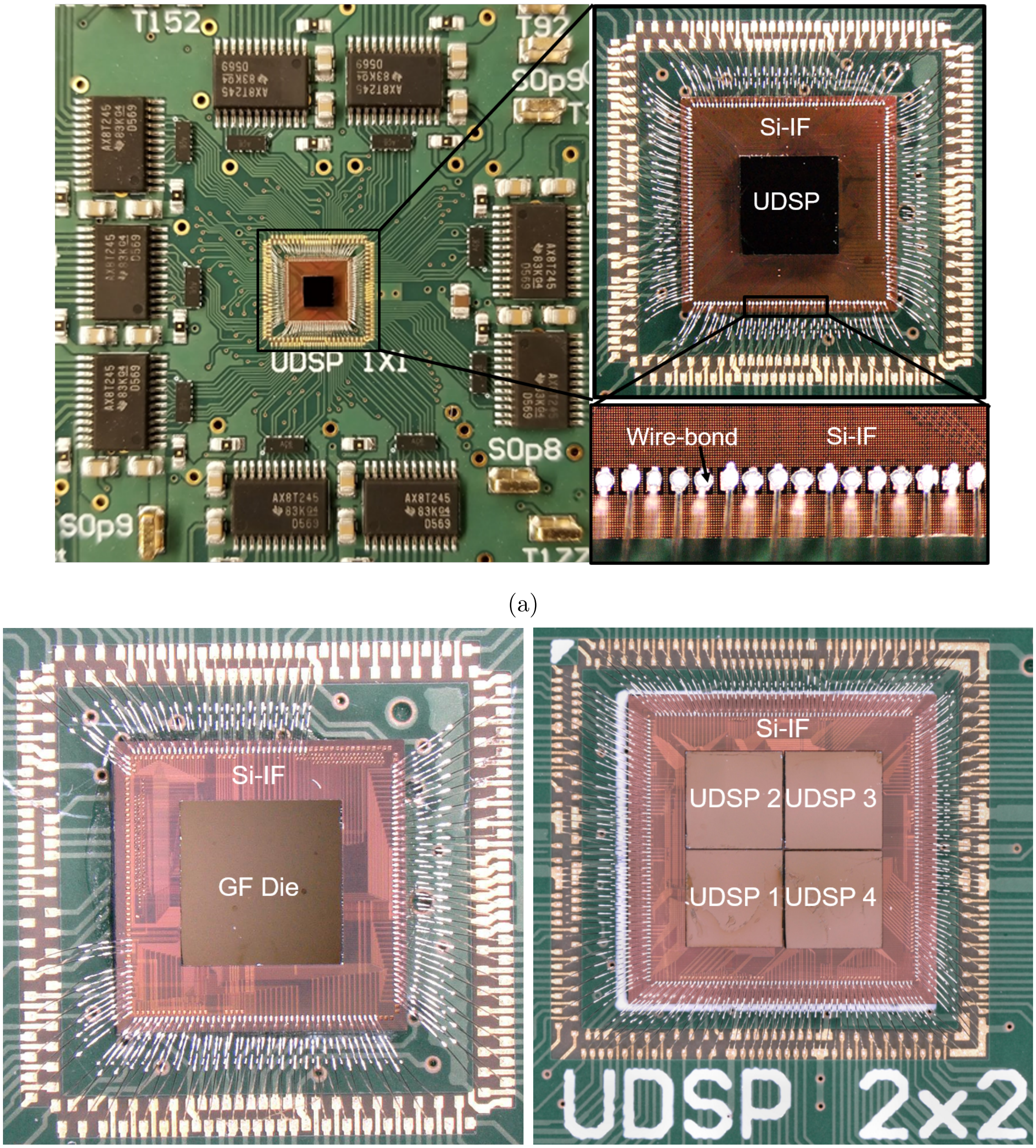


Figure 6.13: Micrograph of a 2x2 array of UDSP dies assembled on the Si-IF. Inset shows the inter-dielet spacing of $\leq$55 μm.

6.4 Results of SuperCHIPS Macros Characterization

The SuperCHIPS macros of the dies in both the technologies were characterized and the results are presented below.



(a)

(b)

(c)

Figure 6.14: Micrographs of wire-bonded samples: (a) Si-IF with single UDSP, (b) Si-IF with GF die, (c) Si-IF with 2x2 array of UDSPs.

6.4.1 Continuity

The continuity check is the first test that was performed to validate the bonding. Both the passive and active daisy chains of all the bonded assemblies were tested using probing of the Si-IF pads. For a single die, the daisy chains on the opposite edges were tested to ensure alignment. Additionally, for 2x2 UDSPs assembly, the daisy chains that pass through two dies in series were also successfully tested that ensured bonding of all the dies. A square wave was applied to the inputs and the outputs of the daisy chains were measured. The passive daisy chains passed the continuity tests for all the bonded samples. The measured waveform is shown in Fig. 6.15 (a). This establishes that the TCB process is successful, reliable, and repeatable. In addition, the output of the active daisy chains was also observed to follow the input waveform as illustrated in Fig. 6.15 (b). This demonstrates that the assembly process does not affect the functionality of the devices. Therefore, no bonding pressure or ESD related failures were observed. The reduction in the output voltage swing of the active daisy chain is because of the voltage drop (100 mV) from the supply to the power and ground pillars of the Si-IF. All the assemblies passed the continuity tests.

6.4.2 Latency Characterization

As previously stated in section 6.1.1, the ring oscillator output was frequency divided and measured by probing on the fan-ed out pads on the Si-IFs. For these measurements, both the GF die assembly, and the 2x2 UDSP dies assembly on Si-IF were used. The average measured frequencies of the reference, and the Si-IF ring oscillators with 200 μm and 500 μm link lengths are listed in Table 6.1. The actual oscillator frequencies are found by multiplying the measured frequencies with 2^{12}. Subsequently, the latencies introduced by the Si-IF links are determined using (6.3) which are also presented in Table 6.1. The measured output waveforms are illustrated in Fig. 6.16.

The latency introduced by the Si-IF links is dependant on the driver strength. The TSMC 16FF library had larger buffers than GF 22FDX and the values presented in Ta-

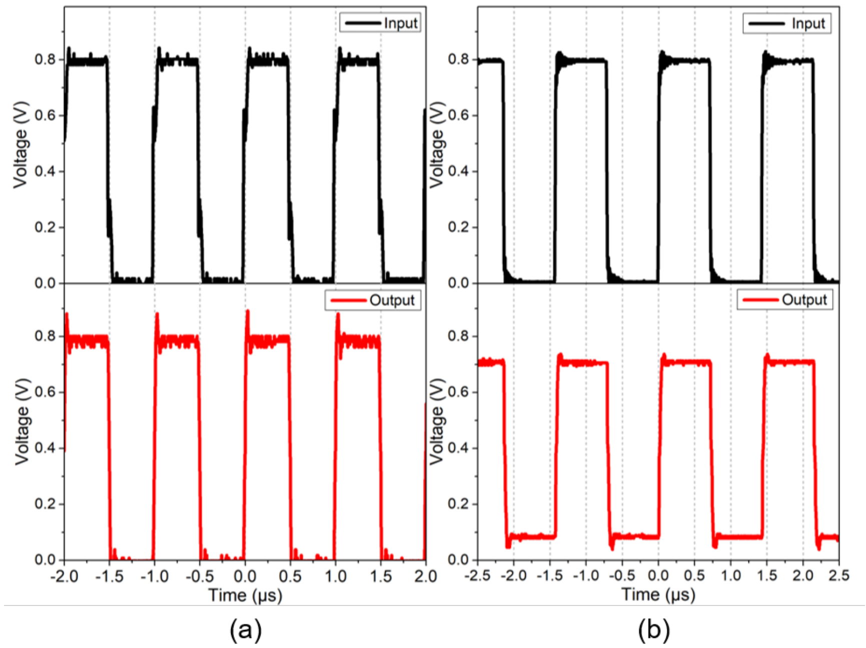
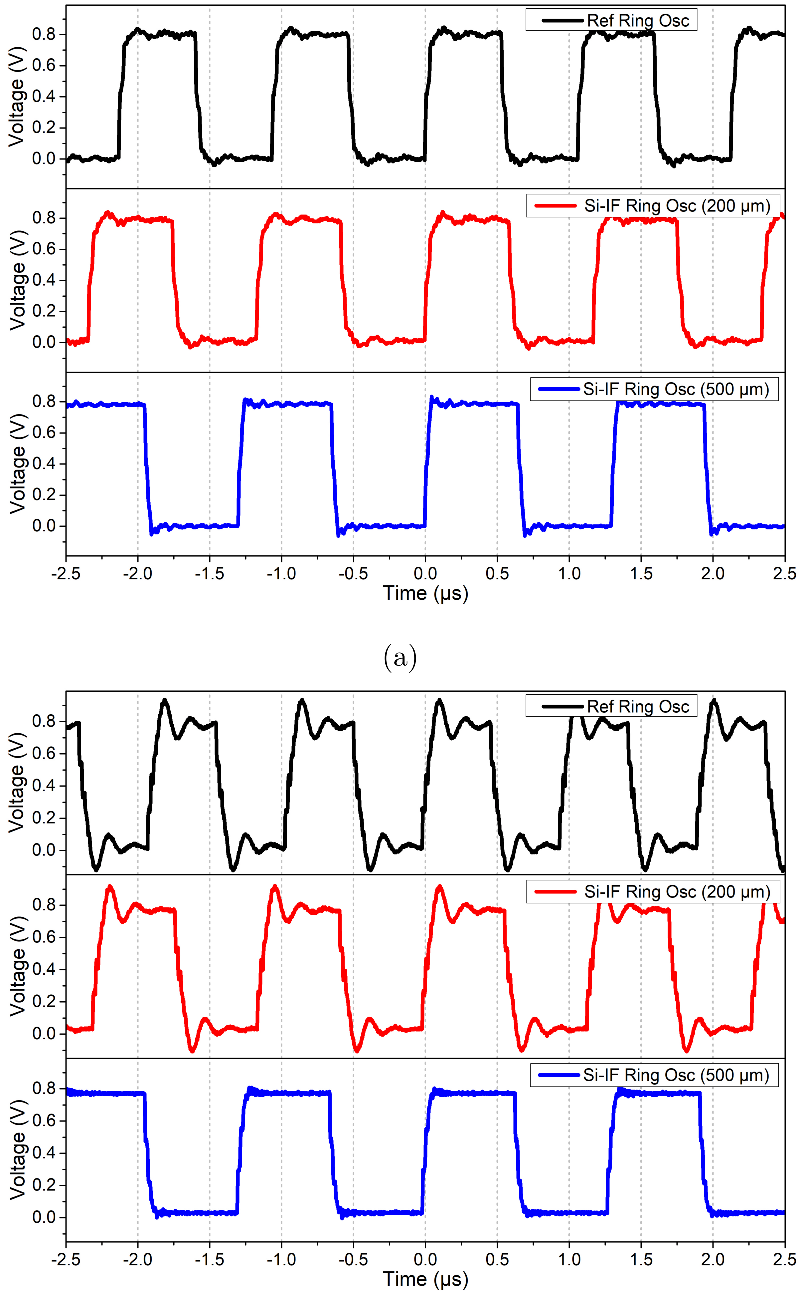


Figure 6.15: Measured waveforms verifying electrical continuity after assembly. (a) Passive daisy chain, (b) Active daisy chain.

Ring oscillator	**Measured frequency (kHz)**	**Actual frequency (before division) (GHz)**	**Measured latency of the Si-IF links (ps)**
2x2 UDSP array on Si-IF			
On-chip reference	921.1	3.77	N.A
With 200 μm Si-IF links	836.8	3.43	6.67
With 500 μm Si-IF links	762.3	3.12	13.80
GF die on Si-IF			
On-chip reference	1033.9	4.23	N.A
With 200 μm Si-IF links	877.6	3.59	10.51
With 500 μm Si-IF links	760.3	3.11	21.26

Table 6.1: Characterization of the latency introduced by the Si-IF links.



(a)

(b)

Figure 6.16: The measured output waveforms of the ring oscillators after frequency division by 2^{12} for the macros in (a) 2x2 UDSP dies assembled on the Si-IF, (b) GF die assembled on the Si-IF. Presented are the waveforms of the reference ring oscillator (black), ring oscillator with 200 μm (red), and 500 μm Si-IF links (blue).

ble 6.1 reflect that and are consistent with the theory (5.2). The latency values were also verified with the on-chip cycle counter measurements and the values match perfectly. Moreover, the latency introduced by Si-IF links is comparable to on-chip buffer delays. As a result, the latency using the SuperCHIPS protocol is 50X lower than the typical SERDES interface on PCBs. Further, the latency is 10X lower compared to interposers. Therefore, by using a fine-pitch ($\leq$10 μm) assembly and small inter-die spacings ($\leq$100 μm), the SuperCHIPS interface achieves superior latency performance. Furthermore, as presented in Table 6.1, the ring oscillators using both the 200 μm and 500 μm Si-IF links have operating frequencies up to 4 GHz. Achieving such high clock speeds ($\geq$4 GHz) on interposers is challenging which will be discussed in chapter 7. On the contrary, the short Si-IF links ($\leq$500 μm) in SuperCHIPS, achieve high data-rates $\geq$10 Gbps/link as shown in section 5.2.

6.4.3 High-speed Data Transfer & BER

For high-speed data transfer and BER characterization, the assembly of the GF die on Si-IF was wire-bonded to the testing PCB. The programmable ring oscillator clock frequency was varied from 500 MHz to 3 GHz which was verified by the 2^8 divided output clock frequency. The module was triggered and the output bits of the error counter were monitored. The testing showed no errors for all the frequencies from 500 MHz to 3 GHz, demonstrating successful data-transfer up to 3 Gbps/link. The aggregate bandwidth for the 16-bits across both the macros is 48 Gbps which corresponds to a maximum data-bandwidth/mm of 1200 Gbps/mm for the two-layer Si-IF. Increasing the frequency from 3 GHz to 6 GHz caused timing closure problems because of the technology limitation and the data-transfer beyond 3 GHz could not be verified. The testing was continued for more than 43 hrs with no errors, corresponding to a BER of <10^{-14} with 99% confidence. The BER testing is limited by the testing time and the actual BER is expected to be much lower. The SNR at the sampling frequency is estimated to be >35 dB from the plot shown in Fig. 5.9. Using this, one can estimate the BER to be much lower (<10^{-25}). Verifying this BER is extremely challenging and would require a variable sampling point at the receiver to plot a bathtub curve. However, this circuitry was not implemented in the macros.

Moreover, the difference in the power between the active and reset state was measured to estimate the energy per bit. The measured difference in power was 1.34 mW for 48 Gbps data-transfer across 450 μm long SuperCHIPS interface. This corresponds to an energy per bit of 0.028 pJ/b which includes the clocking, registers, and I/O buffers. However, note there is no ESD protection implemented. The ESD protection capacitance of 50 fF would add 0.03 pJ/b to the overall energy per bit and would also double the link latency.

6.5 Results of UDSP Characterization

6.5.1 Single UDSP Functionality

The UDSP die functionality was analyzed using the testing PCB. The UDSP was successfully booted up and the idle power was ≈100 mW. First, a clock loop-back test was performed using both an external clock, and the PLL reference clock. In the clock loop-back test, the clock is transferred to the center of the clock-tree within the UDSP and distributed to all the nodes. One of the leaf nodes is given as an output for observation. The clock loop-back test was successful with the output waveform following the input clock. This establishes that the clock tree within UDSP is functional. Second, a programming loop-back test was performed using an FPGA to transfer the program to the UDSP which is subsequently looped-back to the output and observed. The programming loop-back was successful where all the programs transferred were correctly looped-back. Moreover, other control flags were also working as expected and the output can be observed reliably. Next, the UDSP core and the data-plane were programmed and the output was monitored. This test, however, showed inconsistent results, and the exact problems are under scrutiny. The summary of the observations suggests that the control plane of the UDSP is functional without flaws, however, the data-plane had errors. The errors occured internal to the data channels on the UDSP because the errors were sampled with the UDSP internal clock. Programming the cores seemed to trigger faults in some of the data bits, an issue that is being investigated. In addition, some of the input bits have a loss of data in the input path while all the output bits show activity.

The macro of the UDSP cores on the GF die was also tested in the same sequence as above. Once again, the clock and programming loop-back are functional and consistent with the UDSP die results. Programming the cores again resulted in inconsistent results, however, some programs showed complete functionality while others showed errors. These experiments also suggest that the problem is internal to the data-channels on the die, particularly the input path, because all the output bits show expected data.

6.5.2 2x2 UDSP System

The 2x2 UDSP assembly was first tested by probing to verify the SuperCHIPS macros performance. These results are presented above in section 6.4, and the successful functionality of the macros was established. The assembly was mounted on a testing PCB for further tests. Once again, the assembly was successfully booted up and the idle power was ≈400 mW. Because of the design choice in section 6.2, the clock loopback can only be verified using the PLL reference clock since the external clock input of all the UDSPs were tied to the PLL output of one of the UDSP. The PLL clock loop-back gave inconsistent results both in the single UDSP and 2x2 UDSPs system. Further, none of the four PLLs in the 2x2 UDSP system achieved a lock, although the output waveform of the PLL suggests some partial locking was achieved. The PLL is essential to achieve a high-frequency clock to test the high-speed data transfer of the UDSP system. However, the experiments suggest that the failures are in the input data-path similar to the faults observed in single UDSP testing. This may be the reason for the failure of the PLL as well. Further testing couldn't progress until the single UDSP functionality is verified and the reason for the input data path failures is debugged.

6.5.3 Conclusions

From the experiments, it is clear that the problem is not related to the Cu-Cu TCB process, because both the SuperCHIPS macros and the output channels show expected behavior. Also, no passive coupling or short behavior was observed. Further, all the faults observed were sampled according to the internal UDSP clock. As a result, the faults seem to be internal to the UDSP such as manufacturing defects or errors in the input channel logic. Although, having no ESD protection leads to the suspicion that there are ESD related failures in the input path. This, however, couldn't explain all the results of the experiments. Moreover, the data from the SuperCHIPS macro testing suggests that ESD related failures didn't occur, or at least are not common. The problems are under study by Prof. Markovic's research group [Mar].

CHAPTER 7

SuperCHIPS Benefits & Signaling Figure of Merit

From the previous chapters, both experimental and circuit simulations demonstrate that the SuperCHIPS protocol achieves low energy ($\leq$0.03 pJ/b), low latency ($\leq$30 ps), and high bandwidth (8 Tbps/mm) communication between dielets. This protocol is particularly efficient for a streaming interface between dielets that is comparable to on-chip communication on an SoC. In this chapter, the benefits of the SuperCHIPS protocol are contrasted with those of the existing technologies. As mentioned earlier, the SuperCHIPS protocol is a hardware interface protocol with simple buffer I/Os. Any logical protocol can be implemented using the SuperCHIPS including SERDES. However, to achieve the benefits presented in this work, a simple logical protocol is needed and the SNR-10 protocol described in section 6.1 serves this purpose. The values presented in this chapter assume this logic protocol implementation.

7.1 Comparison with Conventional Technologies

Today, the best bandwidth performance is achieved by using SERDES for packages & PCBs, and High Bandwidth Memory (HBM) or AIB protocols for interposers. For a PCB or package, the wiring density is only 2 wires/mm/layer compared to >200 wires/mm/layer on interposers and the Si-IF. Moreover, as mentioned earlier, the interconnect pitch for package and PCB is limited by C4-bump pitch (130 μm) and BGA pitch (0.4-1 mm) respectively. Therefore, for a PCB style integration, SERDES is the only way to achieve high bandwidth. Over the past decade, there has been significant research in increasing the SERDES data-rate and improving the energy efficiency owing to the data-bandwidth demands. The current state-of-the-art SERDES typically operate at data-rates of 56 Gbps with differential wires and typically use PAM-4 signaling [ECH+18, LWL+19]. Recently, higher data-rate SERDES of 112 Gbps/link were also demonstrated [KBD+19, KPL+20]. These SERDES typically have energy efficiencies of 4-7 pJ/b depending on reach and can compensate for signal attenuation of 20-35 dB [KPL+20, KBD+19, LWL+19, DZM+19, NCH+15]. For a neighboring die on MCM package or on a board, the signal attenuation is only 4-10 dB, and therefore, the SERDES energy efficiency can be improved to 1-2 pJ/b [PDC+13, PWT+19, TBC+20, SCF+16] using single-ended links with lower data-rate per link of 10-25 Gbps. Apart from the interconnect pitch, the SERDES circuits are also limited by the real estate on a chip. Typical SERDES circuits that operate at 56 Gbps occupy an area of 2x0.31 mm^2 per link [KZ19]. Therefore, the I/O circuitry extends 2 mm deep into the die which is >25% even for large dies (>625 mm^2). The DARPA CHIPS program targets a bandwidth density of 1 Tbps/mm [Gre16] which would require 12 mm depth along the die perimeter just for I/Os. This is almost the whole die area and therefore, is not practical [KZ19]. Some recent SERDES implementations [KPL+20, KBD+19] have shown smaller area and higher data-rates, but they still require significant die area (>3 mm) to meet the 1 Tbps/mm specification.

Interposers, on the other hand, have moderate interconnect densities and connect neighboring dies using relatively simple I/O cells. Therefore, higher energy efficiencies are achieved (<1 pJ/b [Keh19, OCL+17]) using a lower data-rate of 2-4 Gbps/link [Keh19, Sta20]. At the same time, a higher bandwidth density of >500 Gbps/mm is achieved due to fine wiring pitch of ≤ 4 μm, reduced interconnect pitch of 40-55 μm [CHT+17, MSP+16], and short link lengths of 1-5 mm for neighboring dies. The I/O real estate and reach in interposers are typically limited by the μ-bump pitch rather than the I/O circuitry. Therefore, to achieve the DARPA CHIPS target of 1 Tbps/mm, interposers today require 1.85 mm depth along the die perimeter [KZ19]. By reducing the bump pitch to <35 μm, this can be reduced to 0.92 mm [KZ19]. Authors in [LHT+20], have shown high bandwidth density of 1.6 Tbps/mm^2 (corresponding to 533.33 Gbps/mm per die edge) with high data-rate of 8 Gbps/link on CoWoS platform using a Low-voltage-InPackage-INterCONnect (LIPINCON) interface [MCC+16]. The energy per bit was low (0.56 pJ/b) because of the low swing transfer but the area of their I/O was larger to accommodate the circuit complexity. Moreover, the reach for these links was only 500 μm which was achieved by configuring the 40 μm bumps appropriately, and reducing the inter-dielet spacing to <70 μm. Therefore, the I/O depth is limited to eight columns and may not be easily scalable.

The advantage of the SuperCHIPS protocol in terms of the physical implementation is shown in Table 7.1. The SuperCHIPS protocol has a good balance of fine pillar pitch ($\leq$10 μm) on the Si-IF and much simpler I/Os with a transceiver area of <10x10 μm^2 including the control logic. Therefore, to achieve a 1 Tbps/mm of bandwidth density, only 50 μm of the die perimeter is utilized which in turn allows for the short inter-dielet link lengths of 100-500 μm. This corresponds to a 37X and 240X improvement in area efficiency compared to PCB or interposer-based implementations. Moreover, the SuperCHIPS protocol simplifies the I/Os implementation by designing with standard cells and to work at core frequency and transfer data either at SDR or DDR.

A comparison of different metrics of the SuperCHIPS protocol with the state-of-the-art SERDES protocol on PCBs and HBM or AIB interfaces on interposers is presented in Table 7.2. As mentioned earlier, the short links ($\leq$500 μm) in the SuperCHIPS interface have a latency of <30 ps or 1 clock cycle. This corresponds to an improvement of 4-65X and 3-50X when compared to PCB and interposer-based interfaces respectively. Moreover, this is comparable to 1-2 stage on-chip buffer delays. Also, because of the simple I/O cells, the energy per bit using SuperCHIPS is <0.03 pJ/b for the asynchronous mode, and <0.15 pJ/b for the synchronous mode. For reference, global communication on SoCs typically has an energy efficiency of 0.01 pJ/b/mm [LLS+13]. Therefore, the SuperCHIPS energy/bit is significantly lower (5-40X) compared to traditional systems on PCBs and interposers. At the same time, due to the fine-pitch interconnects, the bandwidth density is up to 8 Tbps/mm for asynchronous mode, and up to 2.56 Tbps/mm for synchronous transfer. Although the data-rate per link of the SuperCHIPS protocol is 4-10 Gbps, which

Parameter		PCB/ SERDES	Interposer/ AIB	Si-IF/ SuperCHIPS
Wire density (lines/mm/layer)		2	250 [MSP+16]	200-250[b]
Data-rate (Gbps)		56-112[a]	2-4 [Keh19]	2-4
Typical I/O size	Height along die edge	310 μm [KZ19]	104 μm [KZ19]	10 μm
	Depth into the die	0.5-2 mm [KZ19]	27.5 μm [KZ19]	10 μm
I/O depth required for 1 Tbps/mm bandwidth density		3-12 mm [KZ19]	0.92-1.85 mm [KZ19]	50 μm

[a] References: [ECH+18, KBD+19, KPL+20, LWL+19].

[b] Assuming UCLA fabrication facilities and corresponding design rules.

Table 7.1: Typical I/O area required to meet 1 Tbps/mm data-bandwidth specification for different implementations.

is 10X slower than the SERDES interface, the bandwidth density is extremely high (7-23X) due to the fine interconnect pitch and wiring density. Moreover, the data-rate per link of SuperCHIPS is comparable if not higher than interposer interfaces and is limited by the core operating frequency for simplicity of implementation. Compared to interposer interfaces, the bandwidth density of SuperCHIPS is 4-11X higher.

7.2 Signaling Figure of Merit

In the Table. 7.2, several different metrics including latency, energy/bit, bandwidth are listed and compared separately. This is the current norm for comparison of signaling for any packaging technology. Although these parameters represent different aspects of perfor-

Technology/ Interface protocol	**Si-IF/ SuperCHIPS**		**Interposer/ HBM2E, AIB**	**PCB/ SERDES**		**Change**
	Async	Sync				
Reach (length)	Neighbor ($\leq$500 μm)		Neighbor (1-5 mm)	Neighbor ($\approx$50 mm)	Long reach ($\approx$300 mm)	
Interconnect pitch (μm)	10		40-55	100-150	400-1000	4-100X
I/O depth (μm)	80		715 [Sta20]-1320 [WAA+20]	686 [PWT+19]	1027[b]	9-25X
Data-rate/link (Gbps)	10	4	2 [Keh19]-3.2 [Sta20]	25 [PWT+19]	56-112[c]	0.1-5X
Overall latency (ps)	30	1 clock cycle (500)	1500 [Keh19]	$\approx$2000	$\approx$6000	3-65X
Energy/bit (pJ/b)	<0.03	<0.15	0.8 [OCL+17]-0.85 [Keh19]	1.17 [PWT+19]	6.9 [LWL+19]	5-40X
Maximum bandwidth/mm (Gbps/mm)	8000	2560[a]	707.7[a]	354	149-298[b]	4-23X

[a] Assuming 20% overhead for power and control signals.

[b] Estimated from data in [KPL+20, KBD+19, LWL+19].

[c] References: [ECH+18, KBD+19, KPL+20, LWL+19].

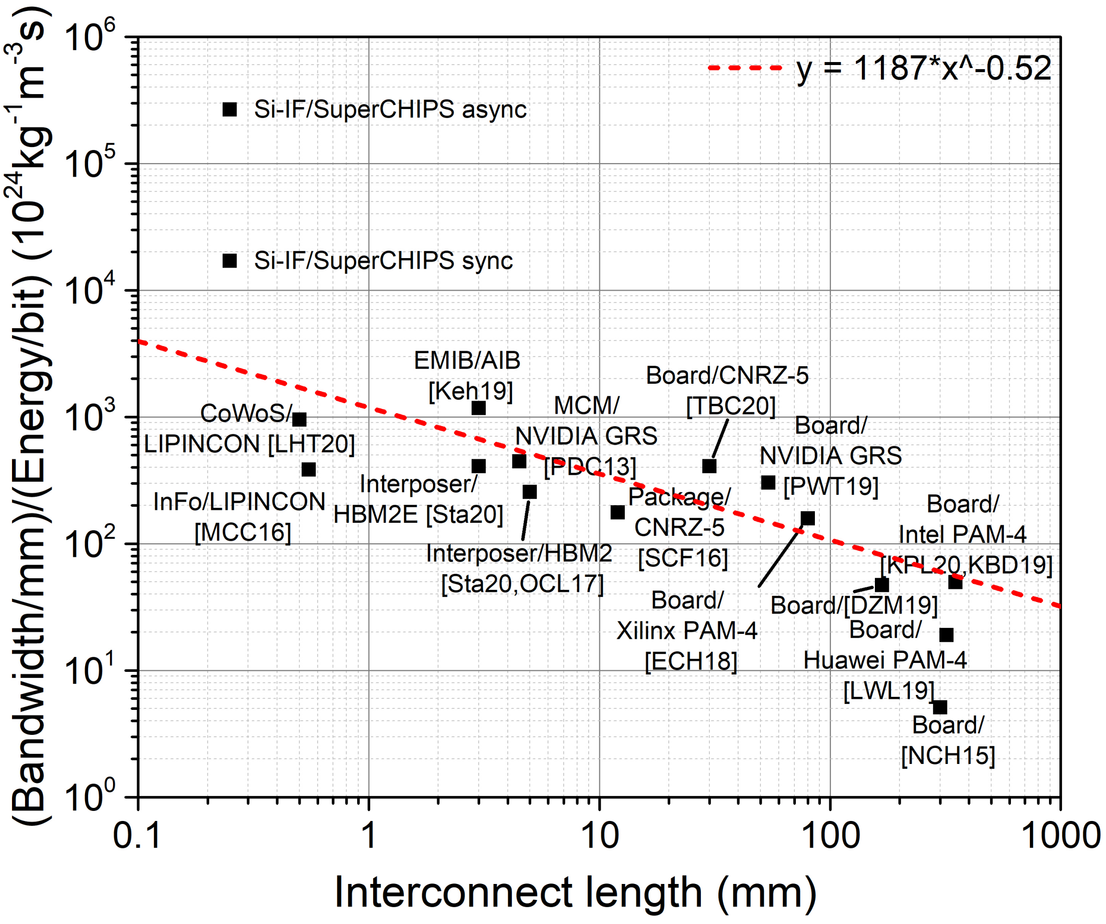
Table 7.2: Comparison of the SuperCHIPS interface protocol with existing technologies.

mance, they are inter-dependent and there exists a trade-off between them. Therefore, it would be good to have a Figure of Merit (FoM) for signaling that captures all the different parameters into a single metric. One FoM was proposed in [WAA+20] by Intel, shown in (7.1), where *Bandwidth/mm* is the bandwidth per millimeter of the die edge and *Energy/bit* is the energy consumed for transferring single bit. This FoM_{Intel} is useful in understanding the system performance benefits because one would naturally desire higher bandwidth/mm and lower energy/bit. A plot of the FoM_{Intel} vs the interconnect length is shown in the Fig. 7.1.

$$FoM_{Intel} = \frac{Bandwidth/mm}{Energy/bit} \tag{7.1}$$

Although this FoM is useful, it does not capture several other aspects that are important in signaling. For example, the interconnect length is not considered which is a packaging technology attribute and depends on how far the chips are. Note that the FoM_{Intel} of technologies decreases with the increase in length. This is expected as interconnect length increases, the interconnect density decreases, and the energy/bit increases to compensate for the increased loss. Therefore, it is not fair to compare an I/O designed to drive long distances with an I/O that only communicates with a neighboring die. Moreover, bandwidth/mm is not well defined because it depends on the number of wiring layers used for routing. For example, one could increase the bandwidth/mm by going deeper into die, i.e. use more columns of I/O pads, area, and additional wiring levels. Once again, a fair comparison between technologies does not exist. The latency and transceiver area are also important metrics to consider.

A good FoM should consider all the different metrics and combine them in a way that is meaningful, easy to quantify, and weigh them appropriately for a fair comparison. Keeping this in mind, a novel FoM, called the $\mathrm{FoM_{UCLA}}$ is proposed in this work and shown in (7.2). Some of the terms are explained as follows- *shoreline* is the length along the die edge through which the I/Os communicate; *IOcols* is the number of columns of the I/O pads used, perpendicular to the die edge; *TransceiverArea* is the actual circuit area of both the transmitter and the receiver (not I/O pad area). The explanation and justification of the terms in FoM are presented below.



Estimated from the references to the best of knowledge.

Figure 7.1: Plot of the FoM_{Intel} (Bandwidth per mm/ Energy per bit ((Gbps/mm)/(pJ/b))) vs interconnect length for different state-of-the-art signaling schemes.

$$FoM_{UCLA} = \frac{(\frac{Bandwidth}{shoreline*IOcols}) * (Length_{link})}{(\frac{Energy}{bit}) * (\frac{TransceiverArea}{Link}) * Latency} \tag{7.2}$$

- *Bandwidth/(shoreline*IOcols)* represents the bandwidth per mm of the die edge but it also includes the number of I/O columns used. It normalizes the number of wiring layers used and provides fair comparison within and across technologies. A higher *Bandwidth/(shoreline*IOcols)* is desired and therefore, the $\mathrm{FoM_{UCLA}}$ is directly proportional to this term.

- *Length$_{link}$* is the interconnect length between the transmitter and receiver. This term represents the load on the driver and justifies designing larger drivers for longer links. Longer reach transceivers should be given higher merit and therefore, it is in the numerator.

- *Energy/bit* is intuitive to be in the denominator as lower energy is desired.

- The *TransceiverArea/link* represents the silicon area occupied by the transceiver to achieve the metrics. Note that this is not the I/O pad area which is already accounted for in the *Bandwidth/(shoreline*IOcols)* term. Transceivers do not contribute to the functionality of a system except they are an inevitable burden for data-transfer through the connecting links. In addition, the transceiver area cuts into the active area of the die that could be used for computation and memory. As a result, a lower transceiver circuit area is desired which should correspond to a higher FoM value. In addition, the directionality of links is also accounted for in this term. For unidirectional links, both the transmitter and receiver area are considered, while for bi-directional links, the entire transceiver area on one terminal is considered.

- *Latency* is also considered and lower latency is desirable. This is included since many of the applications today including real-time processing are latency sensitive.

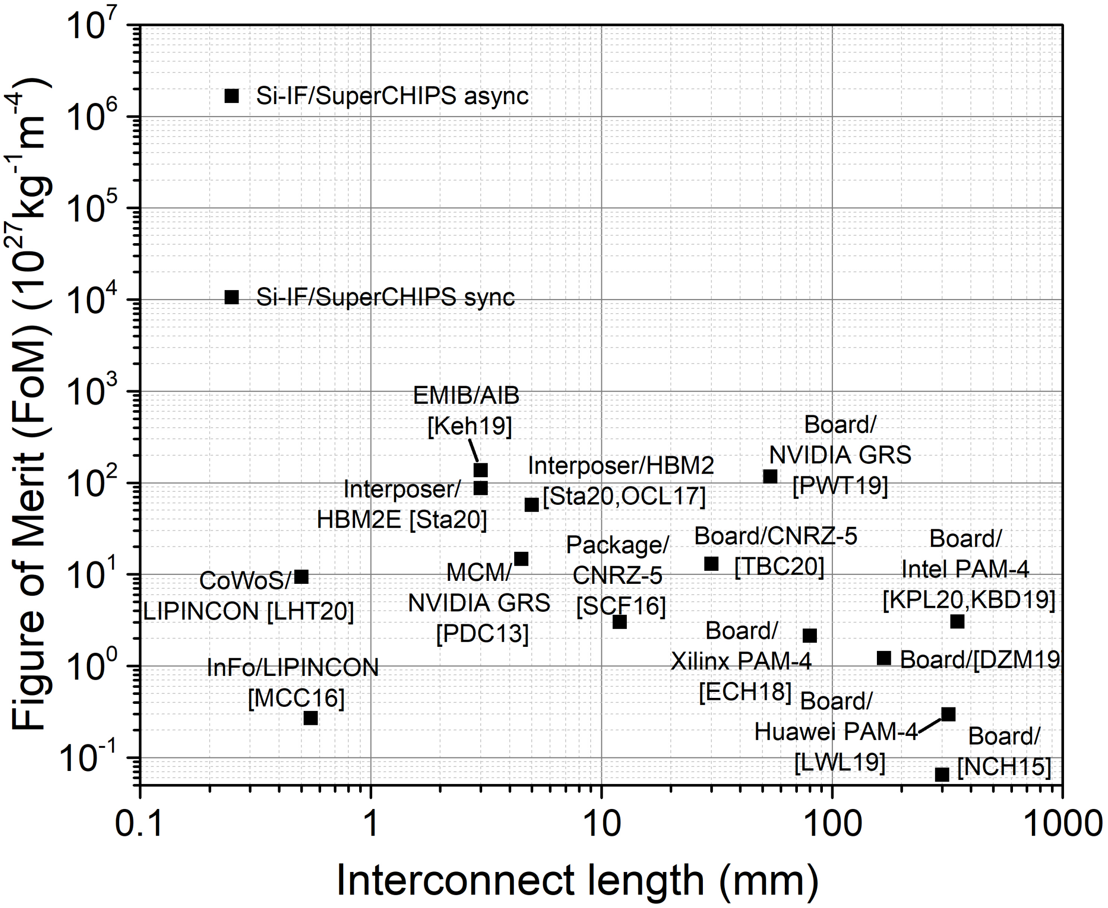
Note that these parameters depend both on the semiconductor and packaging technologies which is true for the performance of any signaling scheme including SuperCHIPS. A smaller Si technology node reduces the I/O circuit area, and energy, while a better packaging technology increases the bandwidth per mm and reduces latency. For example, moving from GF 22FDX to TSMC 16FF die technology provided ≈60% improvement in the FoM$_{\text{UCLA}}$ for the same SuperCHIPS signaling on the Si-IF platform. On the other hand, consider a SuperCHIPS signaling scheme between dies placed at close proximity ($\leq$100 μm) on an interposer. The corresponding FoM$_{\text{UCLA}}$ value decreases by a factor of 2.5X compared to the Si-IF technology due to a lower *Bandwidth/(shoreline*IOcols)* and an increase in *Energy/bit* because of the underlying packaging technology. Moreover, the

FoM_{UCLA} not only represents the performance but also considers the cost for such performance. For example, consider a comparison between MCM and packaged dies on PCB. The performance for MCM may be better than PCB but the cost of integration is also higher which is not beneficial. This is indirectly accounted for in the $Length_{link}$ term. Typical cost is correlated with the feature sizes and smaller links mean higher cost which could nullify the benefit. Therefore, this FoM_{UCLA} represents the overall efficiency of both the die and packaging technologies.

The FoM_{UCLA} for different signaling schemes and technologies is plotted against the length in Fig. 7.2. From the plot, typical SERDES interfaces on PCBs have an FoM_{UCLA} value in the range of 1-10 while packages and interposers have an FoM_{UCLA} value in the range of 10-100 which is a 10X improvement over boards. However, the SuperCHIPS interface in both synchronous and asynchronous exceeds an FoM_{UCLA} value of >10,000. This corresponds to an overall improvement of 100-10,000X in the FoM_{UCLA}. Note that the Si-IF link lengths are only 500 μm, which negatively impacts the FoM_{UCLA}, but the improvement in all the other parameters is so significant that we see such a high improvement. This demonstrates the SuperCHIPS protocol is a highly efficient signaling interface with good balance between all the terms in the FoM_{UCLA}.

7.3 Limitations

As mentioned earlier, the SuperCHIPS protocol is a hardware protocol that is dependant on fine-pitch technologies such as the Si-IF. Therefore, the SuperCHIPS protocol requires the $\leq$10 μm interconnects and cannot be easily implemented on other integration schemes. Moreover, the range of the SuperCHIPS is limited to neighboring die with short links of $\leq$500 μm. This range can be extended to be comparable to interposer link lengths >5 mm with a slight hit in data-rate of $\leq$2 Gbps as shown earlier in Fig. 5.17. Note that although the range is increased from 500 μm to 5 mm, other parameters such as the data-rate, energy per bit, and latency take a hit, reducing the FoM_{UCLA} by about 30%. A range of >5 mm



Estimated from the references to the best of knowledge.

Figure 7.2: Plot of the FoM_{UCLA} vs interconnect length for different state-of-the-art signaling schemes.

allows the SuperCHIPS protocol to communicate with the next-to-neighbor dies if the die sizes are relatively small (≤5 mm). The range can be extended further by using differential signaling, however, it is still limited because of the Si substrate that is lossy compared to organic PCBs.

As a result, it is not a fair comparison to contrast SuperCHIPS with long reach SERDES without the FoM_{UCLA}. In a typical PCB-based integration, the dies are packaged and assembled on boards and several of the boards communicate using a data-backplane. State-of-the-art SERDES are designed to accommodate losses of over 35 dB [KPL+20,LWL+19]. Today, the die-to-package ratio is 5-18X [PPT+19] and the packages have typical losses of about 4-10 dB [PWT+19]. Moreover, the PCB wires have a typical loss of ≈0.1-0.2 dB/mm. Therefore, these long-range SERDES are designed to serve 30-40 cm long backplane connections. Now, integration on Si-IF miniaturizes the overall system by eliminating packages, reducing the inter-dielet spacing, and assembling the entire system on a single wafer. Therefore, the overall form factor of the system is reduced remarkably by a factor of at least 5-10X. Even if we assume a conservative 5X reduction, the equivalent length the SuperCHIPS protocol needs to reach is ≈60 mm. This cannot be achieved by an end to end connection on passive Si-IF and requires intermediate boost on dies. Therefore, the dies in the system need to allocate certain SuperCHIPS channels for feed-through to pass the signals to the neighboring dies. As a result, the signals area transmitted by several hops along the path to the destination. Depending on the system and the size of the dies, the number of hops can vary and latency and energy per bit increase appropriately. The comparison of the communication using this scheme with SERDES is presented in Table 7.3. The underlying assumptions are that the die sizes are 2-10 mm (side) with the signals passed from the first die and repeated through every die with on-chip energy/bit of 0.01 pJ/b/mm [LLS+13]. Also, two link lengths are assumed (1) 300 mm long, and (2) 60 mm long assuming 5X scaling because of technology. The FoM_{UCLA} for the SuperCHIPS interface is still 35-830X better than the long-reach SERDES interface.

Alternatively, a SERDES protocol for global communication on the Si-IF, described in [VBI18], may be implemented. Also, because of the lossy Si substrate, the SuperCHIPS interface cannot be used for radio frequency (RF) communication across the wafer or with external sources. As a result, technologies such as quartz inlays [DAJI20] may be implemented on the Si-IF for RF communication. The implementation and implications of such wafer-level communication schemes are beyond the scope of this thesis.

Technology/ Interface protocol	**Si-IF/ SuperCHIPS**		**PCB/ SERDES**	**Change**
Link length (mm)	300	60	300	5X
Bandwidth/mm (Gbps/mm)	<2560	<2560	298[a]	9X
Energy/bit (pJ/b)	4.55-21.45	1.19-5.7	6.9 [LWL+19]	0.3-6X
Overall latency (ns)	29-142	5.9-28.6	≈6	0.04-1X
UCLA Figure of Merit (FoM)	70-333	345-1667	2	35-830X

[a] References: [KBD+19, KPL+20].

Table 7.3: Comparison of the SuperCHIPS protocol with hops and conventional long-reach SERDES protocols.

CHAPTER 8

Conclusion

8.1 Summary

In this dissertation, a package-less, closely packed, highly scalable, fine-pitch heterogeneous integration technology called the Silicon-Interconnect fabric (Si-IF) was developed. The fundamental aspects of a scalable technology including packaging substrate fabrication, fine-pitch assembly process, and high-bandwidth communication interface protocols were developed. Further, the characteristics and performance benefits of the fine-pitch integration on the Si-IF were investigated.

By repurposing mature semiconductor fabrication techniques, a process flow for fine-pitch silicon-based packaging substrate was developed and demonstrated in chapter 3. Accordingly, a design manual was established for the Si-IF technology.

A solder-less direct metal-metal thermal compression bonding (TCB) process was demonstrated in chapter 4 to achieve sub-10 μm die-to-substrate interconnect pitch. Both direct gold-gold (Au-Au) and copper-copper (Cu-Cu) TCB processes were demonstrated with bonding cycle times of $\leq$6 s and $\leq$30 s respectively. For the direct Cu-Cu TCB, a novel in-situ formic acid vapor treatment process was developed. The fine-pitch pillar interconnects, bonded using these techniques, show an average shear strength of over 127 MPa. Further, electrical continuity was demonstrated across multiple dies with a low specific contact resistance of $<$0.7 Ω-μm^2.

In chapter 5, a Simple Universal Parallel intERface for Chips (SuperCHIPS) protocol was proposed as an interface protocol for near-neighbor communication on the Si-IF.

Experimental characterization of the short SuperCHIPS links of $\leq$500 μm shows a low insertion loss of $\leq$2 dB up to 30 GHz, and a near-end cross-talk of <-15 dB for frequencies up to 20 GHz. Moreover, the parasitics of the Si-IF assembly were found to be 20-50X lower compared to interposer and PCB counterparts. Further, simulation studies of the SuperCHIPS interface were presented.

The performance benefits of the SuperCHIPS protocol were demonstrated using functional hardware assembled on the Si-IF in chapter 6. Dies in two different technologies (TSMC 16FF & GF 22FDX) with sub-10 μm pitch pads were assembled on the Si-IFs using the fabrication and assembly techniques developed in this work. Experimental characterization of the hardware shows a low link latency of <21.56 ps for $\leq$500 μm Si-IF links. Further, 3 Gbps/link data-transfer across the SuperCHIPS interface was also demonstrated corresponding to a data-bandwidth of 1.2 Tbps/mm of die edge. The energy per bit was also measured to be <0.03 pJ/b.

Finally, chapter 7 compares and contrasts the SuperCHIPS protocol with interfaces in interposers and PCB-based assemblies. The SuperCHIPS protocol achieves 4-23X improvement in data-bandwidth, 3-65X reduction in latency, and 5-40X reduction in energy per bit compared to nearest-neighbor communication on interposers and PCBs. Further, a new figure of merit, called the FoM_{UCLA} has been proposed according to which the SuperCHIPS protocol supersedes existing technologies by 100-10,000X.

8.2 Outlook

The Si-IF technology is a superior alternative to PCBs for heterogeneous integration of massive wafer-scale systems. Fine-pitch integration on the Si-IF provides SoC-like performance while ensuring technology heterogeneity. Although the viability and merits of the fine-pitch integration on the Si-IF platform are demonstrated in this work, several challenges remain for wafer-scale assemblies. Three major directions for the future are suggested below:

1. Integration of other enabling technologies in the Si-IF: As mentioned in section 2.2.2,

there are several key enablers for the Si-IF technology which include multiple wiring levels using maskless lithography, TWVs, integrated passives, quartz inlays, and passivation. These technologies were demonstrated independently [LVH$^+$19, TI20, DAJI20, SHI19a, SSYI20] but they must be integrated together for wafer-scale systems. Although, these processes are compatible with conventional semiconductor processing, integrating them may be challenging and must be earnestly pursued. Also, technology transfer to industry is essential to ensure yield and repeatability. Other supplement technologies like connectors for a reliable communication interface with external systems, power delivery, and heat extraction for massive wafer-scale systems are also important. These technologies [AMV$^+$19, SMA$^+$19] present a new domain of challenges and must be integrated with the Si-IF assembly appropriately.

2. Scaling of the assembly process: Assembly of an entire wafer would require handling of multiple dies of different sizes which is limited by tooling availability. The throughput of the TCB process should also be improved further for wafer-scale systems as suggested in section 4.5.4. Also, a contactless substrate heating is essential for the Cu-Cu TCB process because the current approach of plasma-cleaning at regular intervals is not practical. One can consider a laser-based heating or temporary passivation approach for this purpose. Moreover, the reliability of the TCB process must be investigated in great detail in order to ensure high bonding yield across the entire wafer.

3. Novel wafer-scale architectures and systems: To efficiently utilize the benefits of the Si-IF assembly, new architectures have to be explored. The highly parallel SuperCHIPS interface provides high data-bandwidth at low energy and latency which should be capitalized upon. Some of the ideas presented in [PPT$^+$19, Sys] show the tremendous potential for wafer-scale systems.

REFERENCES

[ABC^{+}17] Akhil Arunkumar, Evgeny Bolotin, Benjamin Cho, Ugljesa Milic, Eiman Ebrahimi, Oreste Villa, Aamer Jaleel, Carole-Jean Wu, and David Nellans. Mcm-gpu: Multi-chip-module gpus for continued performance scalability. In *Proceedings of the 44th Annual International Symposium on Computer Architecture*, ISCA '17, page 320–332, New York, NY, USA, 2017. Association for Computing Machinery.

[AG96] K. Azar and J. E. Graebner. Experimental determination of thermal conductivity of printed wiring boards. In *Twelfth Annual IEEE Semiconductor Thermal Measurement and Management Symposium. Proceedings*, pages 169–182, 1996.

[AMV^{+}19] P. Ambhore, U. Mogera, B. Vaisband, U. Shah, T. Fisher, M. Goorsky, and S. S. Iyer. Powertherm attach process for power delivery and heat extraction in the silicon-interconnect fabric using thermocompression bonding. In *2019 IEEE 69th Electronic Components and Technology Conference (ECTC)*, pages 1605–1610, 2019.

[AN01] R. Achar and M. S. Nakhla. Simulation of high-speed interconnects. *Proceedings of the IEEE*, 89(5):693–728, 2001.

[ANT15] Y. Arai, M. Nimura, and H. Tomokage. Cu-cu direct bonding technology using ultrasonic vibration for flip-chip interconnection. In *2015 International Conference on Electronics Packaging and iMAPS All Asia Conference (ICEP-IAAC)*, pages 468–472, 2015.

[BAB^{+}06] B. Black, M. Annavaram, N. Brekelbaum, J. DeVale, L. Jiang, G. H. Loh, D. McCaule, P. Morrow, D. W. Nelson, D. Pantuso, P. Reed, J. Rupley, S. Shankar, J. Shen, and C. Webb. Die stacking (3d) microarchitecture. In *2006 39th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'06)*, pages 469–479, 2006.

[BJH^{+}17] W. Beyene, N. Juneja, Y. Hahm, R. Kollipara, and J. Kim. Signal and power integrity analysis of high-speed links with silicon interposer. In *2017 IEEE 67th Electronic Components and Technology Conference (ECTC)*, pages 1708–1715, 2017.

[BJP^{+}17] A. A. Bajwa, S. Jangam, S. Pal, N. Marathe, T. Bai, T. Fukushima, M. Goorsky, and S. S. Iyer. Heterogeneous integration at fine pitch ($\leq$ 10 μm) using thermal compression bonding. In *2017 IEEE 67th Electronic Components and Technology Conference (ECTC)*, pages 1276–1284, 2017.

[BJP^{+}18] A. A. Bajwa, S. Jangam, S. Pal, B. Vaisband, R. Irwin, M. Goorsky, and S. S. Iyer. Demonstration of a heterogeneously integrated system-on-wafer (sow) assembly. In *2018 IEEE 68th Electronic Components and Technology Conference (ECTC)*, pages 1926–1930, 2018.

[BL07] J. B. Brinton and J. R. Lineback. Packaging is becoming biggest cost in assembly, passing capital equipment. `https://www.eetimes.com/packaging-is-becoming-biggest-cost-in-assembly-passing-capital-equipment/`, 2007.

[Cad07] L. Cadix. Lifting the veil on silicon interposer pricing. `https://sst.semiconductor-digest.com/2012/12/lifting-the-veil-on-silicon-interposer-pricing/`, 2007.

[CCLT15] S. L. Chua, G. Y. Chong, Y. H. Lee, and C. S. Tan. Direct copper-copper wafer bonding with ar/n2 plasma activation. In *2015 IEEE International Conference on Electron Devices and Solid-State Circuits (EDSSC)*, pages 134–137, 2015.

[CCYK13] Y.J. Chen, C.K. Chung, C.R. Yang, and C.R. Kao. Single-joint shear strength of micro cu pillar solder bumps with different amounts of intermetallics. *Microelectronics Reliability*, 53(1):47 – 52, 2013. Reliability of Micro-Interconnects in 3D IC Packages.

[CHI] CHIPS. Neuro ctt. `https://www.chips.ucla.edu/research/project/6`.

[CHT+17] W. C. Chen, C. Hu, K. C. Ting, V. Wei, T. H. Yu, S. Y. Huang, V. C. Y. Chang, C. T. Wang, S. Y. Hou, C. H. Wu, and D. Yu. Wafer level integration of an advanced logic-memory system through 2nd generation cowos® technology. In *2017 Symposium on VLSI Technology*, pages T54–T55, 2017.

[CKL+18] K. Cho, Y. Kim, H. Lee, H. Kim, S. Choi, J. Song, S. Kim, J. Park, S. Lee, and J. Kim. Signal integrity design and analysis of silicon interposer for gpu-memory channels in high-bandwidth memory interface. *IEEE Transactions on Components, Packaging and Manufacturing Technology*, 8(9):1658–1671, 2018.

[Cona] Wikichip Contributors. 7 nm lithography process (wikichip). `https://en.wikichip.org/wiki/7_nm_lithography_process`.

[Conb] Wikichip Contributors. Skylake (server)- microarchitectures - intel (wikichip). `https://en.wikichip.org/wiki/intel/microarchitectures/skylake_(server)`.

[Conc] Wikipedia Contributors. Transistor count. `https://en.wikipedia.org/wiki/Transistor_count`.

[DAJI20] A. Dasgupta, A. Alam, G. Ouyang S. Jangam, and S. Iyer. Antenna on silicon interconnect fabric. In *2020 IEEE 70th Electronic Components and Technology Conference (ECTC)*, 2020.

[DGT$^+$09] L. Di Cioccio, P. Gueguen, R. Taibi, T. Signamarcheix, L. Bally, L. Vandroux, M. Zussy, S. Verrun, J. Dechamp, P. Leduc, M. Assous, D. Bouchu, F. de Crecy, L. Chapelon, and L. Clavelier. An innovative die to wafer 3d integration scheme: Die to wafer oxide or copper direct bonding with planarised oxide inter-die filling. In *2009 IEEE International Conference on 3D System Integration*, pages 1–4, 2009.

[DGY$^+$74] R. H. Dennard, F. H. Gaensslen, H. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc. Design of ion-implanted mosfet's with very small physical dimensions. *IEEE Journal of Solid-State Circuits*, 9(5):256–268, 1974.

[DW82] B. Derby and E. R. Wallach. Theoretical model for diffusion bonding. *Metal Science*, 16(1):49–56, 1982.

[DW84] B. Derby and E. R. Wallach. Diffusion bonding: development of theoretical model. *Metal Science*, 18(9):427–431, 1984.

[DWA$^+$07] B. Dang, S. L. Wright, P. S. Andry, C. K. Tsang, C. Patel, R. Polastre, R. Horton, K. Sakuma, B. C. Webb, E. Sprogis, G. Zhang, A. Sharma, and J. U. Knickerbocker. Assembly, characterization, and reworkability of pb-free ultra-fine pitch c4s for system-on-package. In *2007 Proceedings 57th Electronic Components and Technology Conference*, pages 42–48, 2007.

[DWC19] B. Dehlaghi, N. Wary, and T. C. Carusone. Ultra-short-reach interconnects for die-to-die links: Global bandwidth demands in microcosm. *IEEE Solid-State Circuits Magazine*, 11(2):42–53, 2019.

[DZM$^+$19] E. Depaoli, H. Zhang, M. Mazzini, W. Audoglio, A. A. Rossi, G. Albasini, M. Pozzoni, S. Erba, E. Temporiti, and A. Mazzanti. A 64 gb/s low-power transceiver for short-reach pam-4 electrical links in 28-nm fdsoi cmos. *IEEE Journal of Solid-State Circuits*, 54(1):6–17, 2019.

[ECH$^+$18] M. Erett, D. Carey, J. Hudner, R. Casey, K. Geary, P. Neto, M. Raj, S. McLeod, H. Zhang, A. Roldan, H. Zhao, P. Chiang, H. Zhao, K. Tan, Y. Frans, and K. Chang. A 126mw 56gb/s nrz wireline transceiver for synchronous short-reach applications in 16nm finfet. In *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, pages 274–276, 2018.

[EE92] W. R. Eisenstadt and Y. Eo. S-parameter-based ic interconnect transmission line characterization. *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, 15(4):483–490, 1992.

[FCM08] J. Frei, X. Cai, and S. Muller. Multiport s -parameter and t -parameter conversion with symmetry extension. *IEEE Transactions on Microwave Theory and Techniques*, 56(11):2493–2504, 2008.

[Fri94] D. A. Frickey. Conversions between s, z, y, h, abcd, and t parameters which are valid for complex source and load impedances. *IEEE Transactions on Microwave Theory and Techniques*, 42(2):205–211, 1994.

[Fru19] A. Frumusanu. The apple iphone 11, 11 pro 11 pro max review: Performance, battery, camera elevated. `https://www.anandtech.com/show/14892/the-apple-iphone-11-pro-and-max-review/2`, October 2019.

[GBO$^+$11] M. Gerber, C. Beddingfield, S. O'Connor, M. Yoo, M. Lee, D. Kang, S. Park, C. Zwenger, R. Darveaux, R. Lanzone, and K. Park. Next generation fine pitch cu pillar technology — enabling next generation silicon nodes. In *2011 IEEE 61st Electronic Components and Technology Conference (ECTC)*, pages 612–618, 2011.

[GJB74] A. K. Galwey, D. Jamieson, and M. E. Brown. Thermal decomposition of three crystalline modifications of anhydrous copper(ii) formate. *The Journal of Physical Chemistry*, 78(26):2664–2670, Dec 1974.

[GMF$^+$18] G. Gao, L. Mirkarimi, G. Fountain, L. Wang, C. Uzoh, T. Workman, G. Guevara, C. Mandalapu, B. Lee, and R. Katkar. Scaling package interconnects below 20μm pitch with hybrid bonding. In *2018 IEEE 68th Electronic Components and Technology Conference (ECTC)*, pages 314–322, 2018.

[Gre16] Daniel Green. Common heterogeneous integration and intellectual property (ip) reuse strategies chips). `https://www.darpa.mil/attachments/CHIPSoverview%20Sept212016ProposerDay.pdf`, September 2016.

[GSHB$^+$17] Mark S. Goorsky, Kari Schjølberg-Henriksen, Brett Beekley, Tingyu Bai, Karthick Mani, Pranav Ambhore, Adeel Bajwa, Nishant Malik, and Subramanian S. Iyer. Characterization of interfacial morphology of low temperature, low pressure au–au thermocompression bonding. *Japanese Journal of Applied Physics*, 57(2S1):02BC03, dec 2017.

[Gup09] Tapan Gupta. *The Copper Damascene Process and Chemical Mechanical Polishing*, pages 267–300. Springer New York, New York, NY, 2009.

[HHK$^+$14] C. Honrao, T. Huang, M. Kobayashi, V. Smet, P. M. Raj, and R. Tummala. Accelerated slid bonding using thin multi-layer copper-solder stack for fine-pitch interconnections. In *2014 IEEE 64th Electronic Components and Technology Conference (ECTC)*, pages 1160–1165, 2014.

[HSF$^+$16] B. Hedrick, V. Sukumaran, B. Fasano, C. Tessler, J. Garant, J. Lubguban, S. Knickerbocker, M. Cranmer, K. Ramachandran, I. Melville, D. Berger, M. Angyal, R. Indyk, D. Lewison, C. Arvin, L. Guerin, M. Cournoyer, M. P. L. Ouellet, J. Audet, F. Baez, S. Li, and S. Iyer. End-to-end integration of a multi-die glass interposer for system scaling applications. In *2016 IEEE 66th*

Electronic Components and Technology Conference (ECTC), pages 283–288, 2016.

[IAA$^+$19] D. B. Ingerly, S. Amin, L. Aryasomayajula, A. Balankutty, D. Borst, A. Chandra, K. Cheemalapati, C. S. Cook, R. Criss, K. Enamul, W. Gomes, D. Jones, K. C. Kolluru, A. Kandas, G. . Kim, H. Ma, D. Pantuso, C. F. Petersburg, M. Phen-givoni, A. M. Pillai, A. Sairam, P. Shekhar, P. Sinha, P. Stover, A. Telang, and Z. Zell. Foveros: 3d integration and the use of face-to-face chip stacking for logic devices. In *2019 IEEE International Electron Devices Meeting (IEDM)*, pages 19.6.1–19.6.4, 2019.

[IJV19] S. S. Iyer, S. Jangam, and B. Vaisband. Silicon interconnect fabric: A versatile heterogeneous integration platform for ai systems. *IBM Journal of Research and Development*, 63(6):5:1–5:16, 2019.

[IK15] S. S. Iyer and T. Kirihata. Three-dimensional integration: A tutorial for designers. *IEEE Solid-State Circuits Magazine*, 7(4):63–74, 2015.

[Int] Intel. Ball grid array (bga) packaging. `https://www.intel.com/content/dam/www/public/us/en/documents/packaging-databooks/packaging-chapter-14-databook.pdf`.

[ITR07] ITRS. Yield enhancement. `https://www.semiconductors.org/wp-content/uploads/2018/06/5_2015-ITRS-2.0-Yield-Enhancement.pdf`, 2007.

[Iye16] S. S. Iyer. Heterogeneous integration for performance and scaling. *IEEE Transactions on Components, Packaging and Manufacturing Technology*, 6(7):973–982, 2016.

[JAG$^+$17] H. Jacquinot, L. Arnaud, A. Garnier, F. Bana, J. C. Barbe, and S. Cheramy. Rf characterization and modeling of 10 μm fine-pitch cu-pillar on a high density silicon interposer. In *2017 IEEE 67th Electronic Components and Technology Conference (ECTC)*, pages 266–272, 2017.

[JBM$^+$19] S. Jangam, A. A. Bajwa, U. Mogera, P. Ambhore, T. Colosimo, B. Chylak, and S. Iyer. Fine-pitch ($\leq$10 μm) direct cu-cu interconnects using in-situ formic acid vapor treatment. In *2019 IEEE 69th Electronic Components and Technology Conference (ECTC)*, pages 620–627, 2019.

[KAB$^+$05] J. U. Knickerbocker, P. S. Andry, L. P. Buchwalter, A. Deutsch, R. R. Horton, K. A. Jenkins, Y. H. Kwark, G. McVicker, C. S. Patel, R. J. Polastre, C. D. Schuster, A. Sharma, S. M. Sri-Jayantha, C. W. Surovic, C. K. Tsang, B. C. Webb, S. L. Wright, S. R. McKnight, E. J. Sprogis, and B. Dang. Development of next-generation system-on-package (sop) technology based on silicon carriers with fine-pitch chip interconnection. *IBM Journal of Research and Development*, 49(4.5):725–753, 2005.

[KAD$^+$08] J. U. Knickerbocker, P. S. Andry, B. Dang, R. R. Horton, M. J. Interrante, C. S. Patel, R. J. Polastre, K. Sakuma, R. Sirdeshmukh, E. J. Sprogis, S. M. Sri-Jayantha, A. M. Stephens, A. W. Topol, C. K. Tsang, B. C. Webb, and S. L. Wright. Three-dimensional silicon integration. *IBM Journal of Research and Development*, 52(6):553–569, 2008.

[KBD$^+$19] J. Kim, A. Balankutty, R. K. Dokania, A. Elshazly, H. S. Kim, S. Kundu, D. Shi, S. Weaver, K. Yu, and F. O'Mahony. A 112 gb/s pam-4 56 gb/s nrz reconfigurable transmitter with three-tap ffe in 10-nm finfet. *IEEE Journal of Solid-State Circuits*, 54(1):29–42, 2019.

[KC12] Cheng-Ta Ko and Kuan-Neng Chen. Low temperature bonding technology for 3d integration. *Microelectronics Reliability*, 52(2):302 – 311, 2012. Low Temperature Processing for Microelectronics and Microsystems Packaging.

[KE19] J. Kim and Y. Eo. Ic package interconnect line characterization based on frequency-variant transmission line modeling and experimental s-parameters. *IEEE Transactions on Components, Packaging and Manufacturing Technology*, 9(6):1133–1141, 2019.

[Keh19] D. C. Kehlet. Accelerating innovation through a standard chiplet interface: The advanced interface bus (aib). `https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/accelerating-innovation-through-aib-whitepaper.pdf`, 2019.

[KFK13] M. A. Karim, P. D. Franzon, and A. Kumar. Power comparison of 2d, 3d and 2.5d interconnect solutions and power optimization of interposer interconnects. In *2013 IEEE 63rd Electronic Components and Technology Conference*, pages 860–866, 2013.

[KJL15] A. Kannan, N. E. Jerger, and G. H. Loh. Enabling interposer-based disintegration of multi-core processors. In *2015 48th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pages 546–558, 2015.

[KM97] A. B. Kahng and S. Muddu. An analytical delay model for rlc interconnects. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 16(12):1507–1514, 1997.

[KP14] H. Kalargaris and V. F. Pavlidis. Interconnect design tradeoffs for silicon and glass interposers. In *2014 IEEE 12th International New Circuits and Systems Conference (NEWCAS)*, pages 77–80, 2014.

[KPL$^+$20] Y. Krupnik, Y. Perelman, I. Levin, Y. Sanhedrai, R. Eitan, A. Khairi, Y. Shifman, Y. Landau, U. Virobnik, N. Dolev, A. Meisler, and A. Cohen. 112-gb/s pam4 adc-based serdes receiver with resonant afe for long-reach channels. *IEEE Journal of Solid-State Circuits*, 55(4):1077–1085, 2020.

[KSB13] Sung-Kwon Kang, Da-Yuan Shih, and William E. Bernier. *Flip-Chip Interconnections: Past, Present, and Future*, pages 85–154. Springer US, Boston, MA, 2013.

[KVI19] K. K.T., B. Vaisband, and S. S. Iyer. On-chip esd monitor. In *2019 IEEE 69th Electronic Components and Technology Conference (ECTC)*, pages 2225–2233, 2019.

[KZ19] D. C. Kehlet and J. Zhang. Accelerating innovation through a standard chiplet interface: The advanced interface bus (aib). 2019 IEEE 27th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM) Workshop, 2019.

[Lau10] J. H. Lau. Tsv manufacturing yield and hidden costs for 3d ic integration. In *2010 Proceedings 60th Electronic Components and Technology Conference (ECTC)*, pages 1031–1042, 2010.

[Lau17] John Lau. Mcm, sip, soc, and heterogeneous integration defined and explained. `https://www.3dincites.com/2017/08/mcm-sip-soc-and-heterogeneous-integration-defined-and-explained/`, August 2017.

[Lea17] Richard Leadbetter. Inside the next xbox: Project scorpio tech revealed. `https://www.eurogamer.net/articles/digitalfoundry-2017-project-scorpio-tech-revealed`, April 2017.

[LFR05] M. Y. Lanzerotti, G. Fiorenza, and R. A. Rand. Microminiature packaging and integrated circuitry: The work of e. f. rent, with an application to on-chip interconnection requirements. *IBM Journal of Research and Development*, 49(4.5):777–803, 2005.

[LGBS05] C. C. Liu, I. Ganusov, M. Burtscher, and Sandip Tiwari. Bridging the processor-memory performance gap with 3d ic technology. *IEEE Design Test of Computers*, 22(6):556–564, 2005.

[LHT$^+$20] M. Lin, T. Huang, C. Tsai, K. Tam, K. C. Hsieh, C. Chen, W. Huang, C. Hu, Y. Chen, S. K. Goel, C. Fu, S. Rusu, C. Li, S. Yang, M. Wong, S. Yang, and F. Lee. A 7-nm 4-ghz arm^1-core-based cowos^1 chiplet design for high-performance computing. *IEEE Journal of Solid-State Circuits*, 55(4):956–966, 2020.

[LLKK18] J. Lee, C. Y. Lee, C. Kim, and S. Kalchuri. Micro bump system for 2nd generation silicon interposer with gpu and high bandwidth memory (hbm) concurrent integration. In *2018 IEEE 68th Electronic Components and Technology Conference (ECTC)*, pages 607–612, 2018.

[LLS$^+$13] S. Lee, S. Lee, D. Sylvester, D. Blaauw, and J. Sim. A 95fj/b current-mode transceiver for 10mm on-chip interconnect. In *2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, pages 262–263, 2013.

[Loh08] G. H. Loh. 3d-stacked memory architectures for multi-core processors. In *2008 International Symposium on Computer Architecture*, pages 453–464, 2008.

[LVH+19] M. Liu, B. Vaisband, A. Hanna, Y. Luo, Z. Wan, and S. S. Iyer. Process development of power delivery through wafer vias for silicon interconnect fabric. In *2019 IEEE 69th Electronic Components and Technology Conference (ECTC)*, pages 579–586, 2019.

[LWL+19] M. LaCroix, H. Wong, Y. H. Liu, H. Ho, S. Lebedev, P. Krotnev, D. A. Nicolescu, D. Petrov, C. Carvalho, S. Alie, E. Chong, F. A. Musa, and D. Tonietto. 6.2 a 60gb/s pam-4 adc-dsp transceiver in 7nm cmos with snr-based adaptive power scaling achieving 6.9pj/b at 32db loss. In *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, pages 114–116, 2019.

[MAH+13] K. Murayama, M. Aizawa, K. Hara, M. Sunohara, K. Miyairi, K. Mori, J. Charbonnier, M. Assous, J. Bally, G. Simon, and M. Higashi. Warpage control of silicon interposer for 2.5d package application. In *2013 IEEE 63rd Electronic Components and Technology Conference*, pages 879–884, 2013.

[Mar] Prof. Dejan Markovic. Parallel data architectures group. `http://icslwebs.ee.ucla.edu/dejan/researchwiki/`.

[MCC+16] Mu-Shan Lin, Chien-Chun Tsai, Cheng-Hsiang Hsieh, Wen-Hung Huang, Yu-Chi Chen, Shu-Chun Yang, Chin-Ming Fu, Hao-Jie Zhan, Jinn-Yeh Chien, Shao-Yu Li, Y. . Chen, C. . Kuo, Shih-Peng Tai, and K. Yamada. A 16nm 256-bit wide 89.6gbyte/s total bandwidth in-package interconnect with 0.3v swing and 0.062pj/bit power in info package. In *2016 IEEE Hot Chips 28 Symposium (HCS)*, pages 1–32, 2016.

[MGY+12] Riko I Made, Chee Lip Gan, Liling Yan, Katherine Hwee Boon Kor, Hong Ling Chia, Kin Leong Pey, and Carl V. Thompson. Experimental characterization and modeling of the mechanical properties of cu–cu thermocompression bonds for three-dimensional integrated circuits. *Acta Materialia*, 60(2):578 – 587, 2012.

[MRRS84] J. F. McDonald, E. H. Rogers, K. Rose, and A. J. Steckl. The trials of wafer-scale integration: Although major technical problems have been overcome since wsi was first tried in the 1960s, commercial companies can't yet make it fly. *IEEE Spectrum*, 21(10):32–39, 1984.

[MSP+16] R. Mahajan, R. Sankman, N. Patel, D. Kim, K. Aygun, Z. Qian, Y. Mekonnen, I. Salama, S. Sharan, D. Iyengar, and D. Mallik. Embedded multi-die interconnect bridge (emib) – a high density, high bandwidth packaging interconnect. In *2016 IEEE 66th Electronic Components and Technology Conference (ECTC)*, pages 557–565, 2016.

[MWMA12] N. Matsubara, R. Windemuth, H. Mitsuru, and H. Atsushi. Plasma dicing technology. In *2012 4th Electronic System-Integration Technology Conference*, pages 1–5, 2012.

[NCH+15] R. Navid, E. Chen, M. Hossain, B. Leibowitz, J. Ren, C. A. Chou, B. Daly, M. Aleksić, B. Su, S. Li, M. Shirasgaonkar, F. Heaton, J. Zerbe, and J. Eble. A 40 gb/s serial link transceiver in 28 nm cmos technology. *IEEE Journal of Solid-State Circuits*, 50(4):814–827, 2015.

[Nvi18] Nvidia. Hot chips 30: Nvidia xavier soc. `https://fuse.wikichip.org/news/1618/hot-chips-30-nvidia-xavier-soc/`, December 2018.

[OCL+17] M. O'Connor, N. Chatterjee, D. Lee, J. Wilson, A. Agrawal, S. W. Keckler, and W. J. Dally. Fine-grained dram: Energy-efficient dram for extreme bandwidth systems. In *2017 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*, pages 41–54, 2017.

[OOS+14] K. Oi, S. Otake, N. Shimizu, S. Watanabe, Y. Kunimoto, T. Kurihara, T. Koyama, M. Tanaka, L. Aryasomayajula, and Z. Kutlu. Development of new 2.5d package with novel integrated organic interposer substrate with ultra-fine wiring and high density bumps. In *2014 IEEE 64th Electronic Components and Technology Conference (ECTC)*, pages 348–353, 2014.

[PDC+13] J. W. Poulton, W. J. Dally, X. Chen, J. G. Eyles, T. H. Greer, S. G. Tell, J. M. Wilson, and C. T. Gray. A 0.54 pj/b 20 gb/s ground-referenced single-ended short-reach serial link in 28 nm cmos for advanced packaging applications. *IEEE Journal of Solid-State Circuits*, 48(12):3206–3218, 2013.

[Pow08] J. R. Powell. The quantum limit to moore's law. *Proceedings of the IEEE*, 96(8):1247–1248, 2008.

[PPB+18] S. Pal, D. Petrisko, A. A. Bajwa, P. Gupta, S. S. Iyer, and R. Kumar. A case for packageless processors. In *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 466–479, 2018.

[PPKG20] S. Pal, D. Petrisko, R. Kumar, and P. Gupta. Design space exploration for chiplet-assembly-based processors. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 28(4):1062–1073, 2020.

[PPT+19] S. Pal, D. Petrisko, M. Tomei, P. Gupta, S. S. Iyer, and R. Kumar. Architecting waferscale processors - a gpu case study. In *2019 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pages 250–263, 2019.

[PS13] Eric Perfecto and Kamalesh Srivastava. *Technology Trends: Past, Present, and Future*, pages 23–52. Springer US, Boston, MA, 2013.

[PWT+19] J. W. Poulton, J. M. Wilson, W. J. Turner, B. Zimmer, X. Chen, S. S. Kudva, S. Song, S. G. Tell, N. Nedovic, W. Zhao, S. R. Sudhakaran, C. T. Gray, and W. J. Dally. A 1.17-pj/b, 25-gb/s/pin ground-referenced single-ended serial link for off- and on-package communication using a process- and temperature-adaptive voltage regulator. *IEEE Journal of Solid-State Circuits*, 54(1):43–54, 2019.

[Qua19] Qualcomm. Qualcomm datacenter technologies announces commercial shipment of qualcomm centriq 2400 – the world's first 10nm server processor and highest performance arm-based server processor family ever designed. `https://www.qualcomm.com/news/releases/2017/11/08/qualcomm-datacenter-technologies-announces-commercial-shipment-qualcomm`, November 2019.

[RRSST20] A. Roshanghias, A. Rodrigues, S. Schwarz, and A. Steiger-Thirsfeld. Thermosonic direct cu pillar bonding for 3d die stacking. *SN Applied Sciences*, 2(6):1091, May 2020.

[RS11] K. Rupp and S. Selberherr. The economic limit to moore's law. *IEEE Transactions on Semiconductor Manufacturing*, 24(1):1–4, 2011.

[SAB+16] D. Stow, I. Akgun, R. Barnes, Peng Gu, and Y. Xie. Cost analysis and cost-driven ip reuse methodology for soc design based on 2.5d/3d integration. In *2016 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 1–6, 2016.

[SCF+16] A. Shokrollahi, D. Carnelli, J. Fox, K. Hofstra, B. Holden, A. Hormati, P. Hunt, M. Johnston, J. Keay, S. Pesenti, R. Simpson, D. Stauffer, A. Stewart, G. Surace, A. Tajalli, O. T. Amiri, A. Tschank, R. Ulrich, C. Walter, F. Licciardello, Y. Mogentale, and A. Singh. 10.1 a pin-efficient 20.83gb/s/wire 0.94pj/bit forwarded clock cnrz-5-coded serdes up to 12mm for mcm packages in 28nm cmos. In *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 182–183, 2016.

[Sch12] Martin Schmeißer. Reduction of copper oxide by formic acid an ab-initio study, 2012.

[SHI19a] N. Shakoorzadeh, A. Hanna, and S. Iyer. Bilayer passivation film for cu interconnects on si interconnect fabric. In *2019 IEEE International Reliability Physics Symposium (IRPS)*, pages 1–5, 2019.

[Shi19b] Anton Shilov. Samsung develops 12-layer 3d tsv dram: Up to 24 gb hbm2. `https://www.anandtech.com/show/14952/samsung-develops-12layer-3d-tsv-dram`, October 2019.

[Shi20] Anton Shilov. Tsmc & broadcom develop 1,700 mm2 cowos interposer: 2x larger than reticles. `https://www.anandtech.com/show/15582/tsmc-`

broadcom-develop-1700-mm2-cowos-interposer-2x-larger-than-reticles, March 2020.

[SIL$^+$15] M. J. Schulte, M. Ignatowski, G. H. Loh, B. M. Beckmann, W. C. Brantley, S. Gurumurthi, N. Jayasena, I. Paul, S. K. Reinhardt, and G. Rodgers. Achieving exascale capabilities through heterogeneous computing. *IEEE Micro*, 35(4):26–36, 2015.

[SMA$^+$19] U. Shah, U. Mogera, P. Ambhore, B. Vaisband, S. S. Iyer, and T. S. Fisher. Dynamic thermal management of silicon interconnect fabric using flash cooling. In *2019 18th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, pages 1228–1233, 2019.

[Soc] Electronics Packaking Society. Heterogeneous integration roadmap. https://eps.ieee.org/images/files/Roadmap/Heterogeneous-Integration-Roadmap-Generic-Final.pdf.

[SSYI20] N. Shakoorzadeh, K. Sahoo, Y. Yang, and S. Iyer. Atomic layer deposited al2o3 encapsulation for the silicon interconnect fabric. In *2020 IEEE 70th Electronic Components and Technology Conference (ECTC)*, 2020.

[Sta20] JEDEC Standard. Jesd235c : High bandwidth memory (hbm) dram. https://www.jedec.org/document_search?search_api_views_fulltext=jesd235, January 2020.

[Su19] Lisa Su. Delivering the future of high-performance computing, darpa eri summit. https://eri-summit.darpa.mil/docs/Su_Lisa_AMD_Final.pdf, July 2019.

[SVJ$^+$19] E. Sorensen, B. Vaisband, S. Jangam, T. Shirley, and S. S. Iyer. Integration and characterization of inp die on silicon interconnect fabric. In *2019 IEEE 69th Electronic Components and Technology Conference (ECTC)*, pages 543–549, 2019.

[SXSL17] D. Stow, Y. Xie, T. Siddiqua, and G. H. Loh. Cost-effective design of scalable high-performance systems using active and passive interposers. In *2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pages 728–735, 2017.

[Sys] Cerebras Systems. Building a wafer-scale deep learning chip: Lessons learned. https://s3.us-west-1.amazonaws.com/chips.user.media/page_file/Building%20a%20Wafer%20Scale%20Chip%20Lessons%20Learned%20-%20JP%20Fricker%20Post.pdf.

[TBC$^+$20] A. Tajalli, M. Bastani Parizi, D. A. Carnelli, C. Cao, K. Gharibdoust, D. Gorret, A. Gupta, C. Hall, A. Hassanin, K. L. Hofstra, B. Holden, A. Hormati, J. Keay, Y. Mogentale, V. Perrin, J. Phillips, S. Raparthy, A. Shokrollahi,

D. Stauffer, R. Simpson, A. Stewart, G. Surace, O. Talebi Amiri, E. Truffa, A. Tschank, R. Ulrich, C. Walter, and A. Singh. A 1.02-pj/b 20.83-gb/s/wire usr transceiver using cnrz-5 in 16-nm finfet. *IEEE Journal of Solid-State Circuits*, 55(4):1108–1123, 2020.

[TBV+19] K. K. Thankappan, A. Bajwa, B. Vaisband, S. Jangam, and S. S. Iyer. Reliability evaluation of silicon interconnect fabric technology. In *2019 IEEE International Reliability Physics Symposium (IRPS)*, pages 1–5, 2019.

[TI20] K. K. Thankappan and S. Iyer. Deep trench capacitors in silicon interconnect fabric. In *2020 IEEE 70th Electronic Components and Technology Conference (ECTC)*, 2020.

[TLA+12] C.S. Tan, D.F. Lim, X.F. Ang, J. Wei, and K.C. Leong. Low temperature cucu thermo-compression bonding with temporary passivation of self-assembled monolayer and its bond strength enhancement. *Microelectronics Reliability*, 52(2):321 – 324, 2012. Low Temperature Processing for Microelectronics and Microsystems Packaging.

[TLWY16] C. Tseng, C. Liu, C. Wu, and D. Yu. Info (wafer level integrated fan-out) technology. In *2016 IEEE 66th Electronic Components and Technology Conference (ECTC)*, pages 1–6, 2016.

[TWB+16] Koki Tanaka, Wei-Shan Wang, Mario Baum, Joerg Froemel, Hideki Hirano, Shuji Tanaka, Maik Wiemer, and Thomas Otto. Investigation of surface pretreatment methods for wafer-level cu-cu thermo-compression bonding. *Micromachines*, 7(12):234, Dec 2016.

[VBI18] B. Vaisband, A. Bajwa, and S. S. Iyer. Network on interconnect fabric. In *2018 19th International Symposium on Quality Electronic Design (ISQED)*, pages 138–143, 2018.

[WAA+20] M. Wade, E. Anderson, S. Ardalan, P. Bhargava, S. Buchbinder, M. L. Davenport, J. Fini, H. Lu, C. Li, R. Meade, C. Ramamurthy, M. Rust, F. Sedgwick, V. Stojanovic, D. Van Orden, C. Zhang, C. Sun, S. Y. Shumarayev, C. O'Keeffe, T. T. Hoang, D. Kehlet, R. V. Mahajan, M. T. Guzy, A. Chan, and T. Tran. Teraphy: A chiplet technology for low-power, high-bandwidth in-package optical i/o. *IEEE Micro*, 40(2):63–71, 2020.

[WL99] Wei Lin and Y. C. Lee. Study of fluxless soldering using formic acid vapor. *IEEE Transactions on Advanced Packaging*, 22(4):592–601, 1999.

[WPG+06] S. L. Wright, R. Polastre, H. Gan, L. P. Buchwalter, R. Horton, P. S. Andry, E. Sprogis, C. Patel, C. Tsang, J. Knickerbocker, J. R. Lloyd, A. Sharma, and M. S. Sri-Jayantha. Characterization of micro-bump c4 interconnects for si-carrier sop applications. In *56th Electronic Components and Technology Conference 2006*, pages 8 pp.–, 2006.

[WYYM14] C. C. Wang, F. Yuan, T. Yu, and D. Markovic. 27.5 a multi-granularity fpga with hierarchical interconnects for efficient and flexible mobile computing. In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pages 460–461, 2014.

[XWC+16] L. Xie, S. Wickramanayaka, S. C. Chong, V. N. Sekhar, D. Ismeal, and Y. L. Ye. 6um pitch high density cu-cu bonding for 3d ic stacking. In *2016 IEEE 66th Electronic Components and Technology Conference (ECTC)*, pages 2126–2133, 2016.

[YAS14] W. Yang, M. Akaike, and T. Suga. Effect of formic acid vapor in situ treatment process on cu low-temperature bonding. *IEEE Transactions on Components, Packaging and Manufacturing Technology*, 4(6):951–956, 2014.

[YHB08] T.G.A. Youngs, S. Haq, and M. Bowker. Formic acid adsorption and oxidation on cu(110). *Surface Science*, 602(10):1775 – 1782, 2008.

[YW18] Daniel Yang and Stac Wegner. Apple iphone xs max teardown. `https://www.engadget.com/2018-09-12-apple-a12-bionic-7-nanometer-chip.html`, September 2018.

[ZLW09] Y. Zhang, C. Liu, and D. Whalley. Direct-write techniques for maskless production of microelectronics: A review of current state-of-the-art technologies. In *2009 International Conference on Electronic Packaging Technology High Density Packaging*, pages 497–503, 2009.