Title

High Throughput Fine-Pitch (≤ 10 µm) Assembly using Cu-Cu Thermal Compression Bonding and Protocol Integration Strategies for Wafer-Scale Systems

ISBN

9798290944838

Author

Sahoo, Krutikesh

Publication Date

2025-08-11

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

High Throughput Fine-Pitch (≤ 10 μm) Assembly using Cu-Cu Thermal Compression Bonding and Protocol Integration Strategies for Wafer-Scale Systems

A dissertation submitted in partial satisfaction

of the requirements for the degree Doctor of Philosophy

in Electrical and Computer Engineering

by

Krutikesh Sahoo

2025

ABSTRACT OF THE DISSERTATION

High Throughput Fine-Pitch (≤ 10 μm) Assembly using Cu-Cu Thermal Compression Bonding and Protocol Integration Strategies for Wafer-Scale Systems

By

Krutikesh Sahoo

Doctor of Philosophy in Electrical and Computer Engineering

University of California, Los Angeles, 2025

Professor Subramanian S. Iyer, Chair

Heterogeneous chiplet-based wafer-scale systems provide a promising path to address the increasing demands for processor-to-memory bandwidth and memory capacity in applications such as artificial intelligence (AI), high-performance computing (HPC), graph processing, and pathfinding. However, the realization of such systems in hardware requires overcoming significant challenges related to materials and processes, high-throughput assembly, yield, inter-die communication, connectors, power delivery, and thermal management.

This dissertation presents a comprehensive study of bonding throughput, first-order assembly yield, and chiplet communication approaches in the context of Silicon

Interconnect Fabric (Si-IF)—a silicon-based platform for wafer-scale integration. We demonstrate a robust, fine-pitch (≤10μm) two-step copper-to-copper (Cu-Cu) thermal compression bonding (TCB) process capable of achieving 330 units per hour (UPH), with an experimentally validated design of experiments (DoE) methodology that outlines a path to exceed 1000 UPH. Furthermore, assembly-related challenges such as bond-head tilt correction, alignment drift, and dicing-induced yield loss are examined, along with design practices and recommendations to improve dielet assembly yield.

Next, we address the challenge of communication protocol interoperability in chiplet-based systems, often hindered due to the proprietary nature of die-to-die signaling standards. Interoperability between signaling protocols can be achieved using translation. We introduce a translator chiplet called “Anuvāda”, which enables cross-protocol communication using a high-bandwidth (2.5 Tbps/mm), low-energy (0.13 pJ/bit, synchronous), low area (~100μm^2/IO) on-wafer interconnect called SuperCHIPS or Simple Universal Parallel intERface for CHIPS. In this proof-of-concept Anuvāda chiplet, we implement protocol translation between PCIe and SuperCHIPS in the 22nm FDX technology node. We discuss the design & simulations of the chiplet, as well as testing results with PCIe loopback. We also discuss the ESD performance of these scaled IOs. The insights from this research offer practical guidelines for chiplet, package, and system designers aiming to enable high-yield, high-throughput wafer-scale integration with interoperable communication protocols.

The dissertation of Krutikesh Sahoo is approved.

Dejan Markovic

Robert N. Candler

Sudhakar Pamarti

Subramanian S. Iyer, Committee Chair

University of California, Los Angeles

2025

To my parents…

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am deeply grateful to my advisor, Prof. Subramanian S. Iyer (Subu), for his invaluable guidance throughout my PhD journey. It has been an honor to work under a mentor of such technical brilliance, experience and extraordinary caliber. His constant encouragement, insightful feedback, and unwavering support pushed me to expand the boundaries of my knowledge and often surpass my own expectations. Our discussions greatly shaped my critical thinking and problem-solving abilities and taught me the importance of presenting results both succinctly and impactfully. Subu gave me the freedom to explore multiple approaches to problem-solving, the space to make mistakes, and the opportunity to learn from them—offering not only technical mentorship but many life lessons along the way. He is everything I could have hoped for in an ideal mentor, both for my PhD and beyond.

I would also like to express my sincere appreciation to the members of my doctoral committee, Prof. Dejan Markovic, Prof. Rob Candler, and Prof. Sudhakar Pamarti, for their guidance and valuable input. I have had the privilege of collaborating with them on package design and learning about chip design, verification, and testing. Prof. Markovic provided exceptional direction on fine-pitch, high-bandwidth interconnect design and effective presentation of results. Prof. Pamarti gave critical feedback on chip testing of Anuvāda, and Prof. Candler offered helpful comments during my work on high-throughput bonding, testing, and qualification of metal-metal interconnects.

I am also thankful to Prof. Puneet Gupta for his insights on chip assembly and digital design of Anuvāda, and to Prof. C. K. Ken Yang and Prof. Boris Vaisband for their constructive feedback throughout my PhD progress reviews.

I would like to acknowledge my colleagues for their generous support across different stages of my research. I am especially thankful to Dr. SivaChandra Jangam for mentoring me in silicon package fabrication and the metal-metal thermal compression bonding process. I am grateful to Dr. Adeel Bajwa for his guidance on bonding methodology and equipment usage. I would also like to thank Dr. Saptadeep Pal and Dr. Steven Moran for their advice on chip design and verification.

Special thanks to Dr. Goutham Ezhilarasu, Dr. Kannan Thankappan, Dr. Uneeb Rathore, Dr. Arsalan Alam, Dr. Niloofar Shakoorzadeh, Dr. Yu-Tao Yang, and Dr. Guangqi Ouyang for their constant encouragement and support during the most challenging phases of my PhD. I am also grateful to my colleagues—now like brothers and sisters—Siyun Qiao, Golam Sabbir, Randall Irwin, Haoxiang Ren, Vineeth Harish, Samuel Wang, Arpan Dasgupta, Jui-Han Liu, Ziyi Guo, Narendharan M. Sundaram, Mansi Sheth, Jerry Ding, Tri Nguyen, Zoe Chen, and Abhik Gude—for their help with fabrication, testing, and device measurements. I hope to continue collaborating with and supporting them in the future.

I would like to thank the industrial partners and fabrication facilities that made this work possible. I am grateful to Kulicke and Soffa for providing bonding tools used in Cu-Cu and Au-Cu thermal compression bonding. I thank GlobalFoundries for their support on the Anuvāda project, including fabrication of the 22nm FDX dielet with custom modifications. I am especially thankful to Dr. Luis Cruz and Dr. John Barth from Synopsys for their guidance on PCIe protocol implementation and their support throughout the Anuvāda development. I would also like to thank the staff at UCLA's Nanolab fabrication facilities for their invaluable assistance over the course of my PhD.

Lastly, I acknowledge the support of the funding agencies that enabled this research. I am especially thankful to the NSF JUMP and ACCENT programs, DARPA, SRC, and the member companies of the CHIPS consortium for their financial support, valuable technical feedback, and guidance on future research directions through the UCLA-CHIPS Industry Advisory Board meetings.

Parts of this dissertation were adapted from several publications produced during the course of my PhD.

VITA

2011 – 2015	Bachelor of Technology in Electronics and Instrumentation engineering
	Biju Patnaik University of Technology, India
2016 – 2018	Master of Technology in Nano Science and Engineering
	Indian Institute of Science, Bangalore, India
2019 – 2025	Doctor of Philosophy in Electrical and Computer Engineering
(expected)	University of California, Los Angeles, USA

SELECTED PUBLICATIONS

Krutikesh Sahoo, V. Harish, H. Ren and S. S. Iyer, "A Review of Die-to-Die, Die-to-Substrate and Die-to-Wafer Heterogeneous Integration," in *IEEE Electron Devices Reviews*, doi: 10.1109/EDR.2024.3501214.

Krutikesh Sahoo *et al.*, “Aluminum to copper thermal compression bonding for heterogeneous integration of legacy dielets,” in *2024 IEEE 74th Electronic Components and Technology Conference (ECTC)*, 2024, pp. 1940–1946. doi: 10.1109/ECTC51529.2024.00329.

Krutikesh Sahoo, M. Zhai and S. Iyer, "Trade-off Between Chiplet Dimensions and Packaging Parameters for Optimal Performance to Cost for Chiplet Based Heterogeneous Integration," *2024 IEEE 26th Electronics Packaging Technology Conference (EPTC)*, Singapore, 2024, pp. 232-237, doi: 10.1109/EPTC62800.2024.10909747.

Krutikesh Sahoo, H. Ren and S. S. Iyer, "A High Throughput Two-Stage Die-to-Wafer Thermal Compression Bonding Scheme for Heterogeneous Integration," 2023 IEEE *73rd Electronic Components and Technology Conference (ECTC)*, Orlando, FL, USA, 2023, pp. 362-366, doi: 10.1109/ECTC51909.2023.00067.

S. S. Nagi, U. Rathore, **K. Sahoo**, T. Ling, S. S. Iyer and D. Marković, "A 16-nm 784-Core Digital Signal Processor Array, Assembled as a 2 × 2 Dielet With 10-µm Pitch Interdielet I/O for Runtime Multiprogram Reconfiguration," in *IEEE Journal of Solid-State Circuits*, vol. 58, no. 1, pp. 111-123, Jan. 2023, doi: 10.1109/JSSC.2022.3212685.

Krutikesh Sahoo, U. Rathore, S. Jangam, T. Nguyen, D. Markovic, and S. S. Iyer, "Functional Demonstration of < 0.4-pJ/bit, 9.8 µm Fine-Pitch Dielet-to-Dielet Links for Advanced Packaging using Silicon Interconnect Fabric," 2022 IEEE 72nd *Electronic Components and Technology Conference (ECTC)*, 2022 doi: 10.1109/ECTC51906.2022.00332.

Krutikesh Sahoo, S. Pal, N. Shakoorzadeh, Y.-T. Yang, and S. S. Iyer, “Copper to gold thermal compression bonding in heterogeneous wafer-scale systems,” in 2021 IEEE 71st *Electronic Components and Technology Conference (ECTC)*, 2021. doi: 10.1109/ECTC32696.2021.00088.

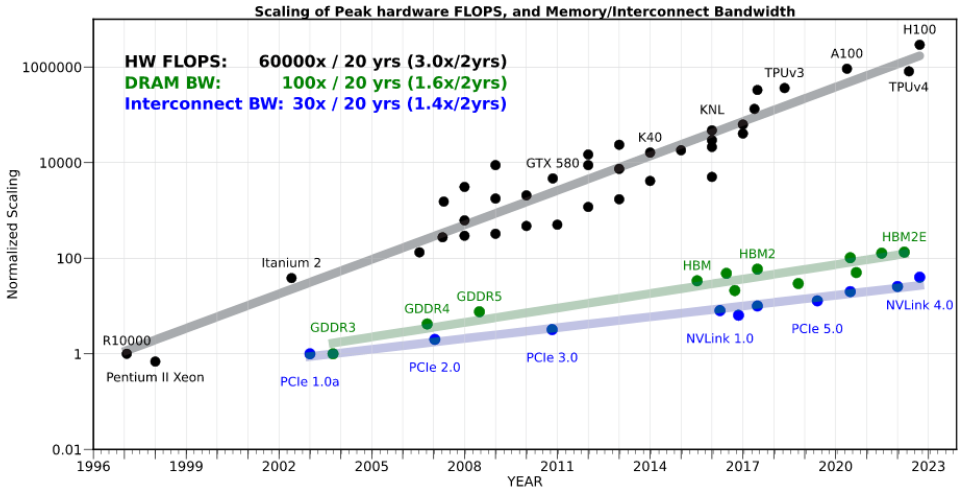
Krutikesh Sahoo, B. Vinnakota, S. Ardalan, S.S. Iyer, “Co-optimizing signaling protocol with semiconductor and packaging technology,” 2021 IEEE 30th *Electrical Performance and Electronics Packaging Systems (IEEE EPEPS)*. doi: 10.1109/EPEPS51341.2021.9609161.

Chapter 1 Wafer-Scale Integration (WSI)—a promising approach for continued package scaling

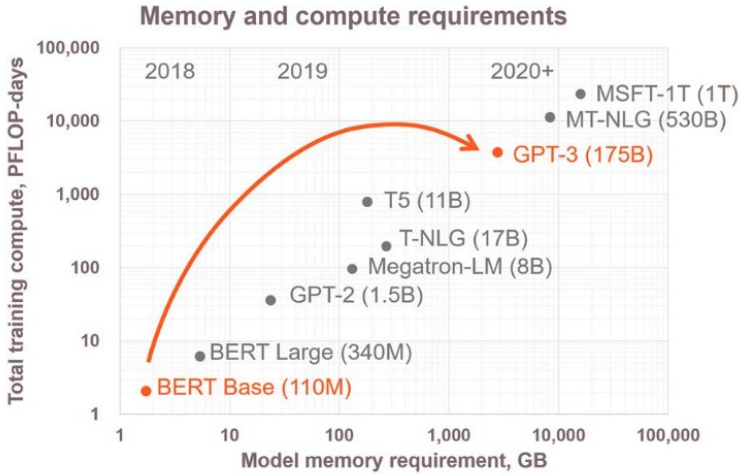
1.1 Introduction

The demand for parallelized hardware to achieve more computing power has never been higher—owing to the proliferation of artificial intelligence (AI) models, scientific simulations, graph processing, and other data-intensive applications. In recent work, Gholami et al. [Ghol24], indicated the amount of computing power needed to train large language models (LLMs) has been steadily growing at a rate of 750x every two years. As a result, the focus of today's architectural development is on increasing computing power of the hardware, often at the expense of simplifying other aspects such as memory hierarchy and using conventional strategies for chip-to-chip communication such as SerDes [Syn1].

However, as highlighted by Iyer et al. [Iyer16], Gholami et al. [Ghol24], and Pal et al. [Spal19], this compute-centric approach often overlooks growing bottlenecks in memory capacity and interconnect bandwidth. In fact, in many AI workloads, intra- and inter-chip communication as well as data transfer to/from AI accelerators are becoming a bigger bottleneck than computational throughput itself. These challenges are commonly referred to as the 'memory wall' problem, a term originally coined by William Wulf and Sally McKee in 1995 [WuMK95]. Despite many innovations in memory technology including HMC [WikiHMC, HMCSpec], HBM [WikiHBM] as well as interconnect scaling [Brain16], this trend still exists today and can be seen in Fig. 1.1, taken from [Ghol24]. While the hardware TOPS (Tera Operations Per Second) has increased by over 60000x over last 20 years, in contrast, DRAM bandwidth has only scaled by ~ 100x over the last 20 years.



(a)



(b)

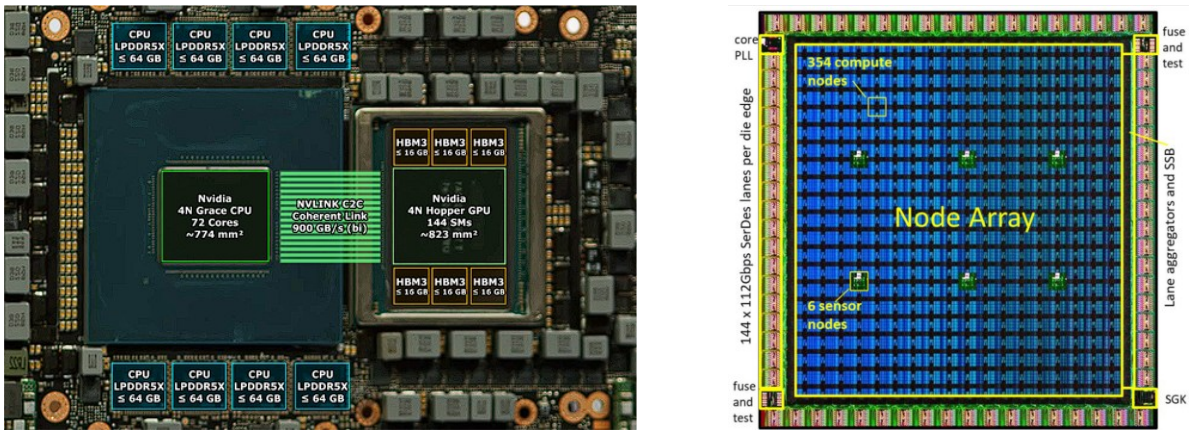
Figure 11.1 The scaling of the bandwidth of different generations of interconnections and memory as well as the peak floating-point operations per second (FLOPS) taken from [Ghol24]. (b) Memory and compute requirements for various state-of-the-art neural networks on log-log scale reported by [Slie23]. PFLOP: Peta Floating Operations per Second. Both these charts indicate the need for innovations in increasing processor to memory bandwidth i.e. interconnect bandwidth as the newer workloads require more compute, memory and inter-communication.

What is more, Interconnect bandwidth scaling suffers even more, with only ~ 30x improvement over the last 20 years. This is indicated by the blue line in the Fig. 1.1(a), where all the links shown are commercially used serialized links. Figure 1.1(b) further

shows the need for compute and memory in many state-of-the-art neural networks, highlighting the need for higher compute-to-memory bandwidth.

To solve the increasing interconnect bandwidth and communication overhead requirements, several academia and industry groups have pursued different approaches based on advanced packaging. In the recent H100 product by NVIDIA [NVH100], they have used chip-on-wafer-on-substrate (CoWoS) technology [TSMCC] from Taiwan Semiconductor Manufacturing Company (TSMC) to achieve a die-to-die bandwidth of about 900GB/s bidirectional between two large chips – grace CPU and hopper GPU. The signaling protocol used is NVIDIA NVLINK C2C [Wei23]. The fundamental idea is to use a silicon interposer for very high bandwidth connections, as fabrication of wires on silicon interposer [Intposr] is dictated by stepper lithography limits which allows for highly dense interconnections when compared to organic laminates and printed circuit boards (PCBs). Each link runs at 40 Gbps [Wei23]. The overall bandwidth per shoreline achieved is about 260 Gbps/mm, one of the highest reported interconnect bandwidths in production using silicon interposers. The communication across interposer still requires use of SerDes, as shown by use of 40Gbps NVIDIA C2C links in H100, thereby requiring IO designs with large drivers and embedded error detection and correction, equalization schemes. The IO energy transfer increases due to these features. The reported energy per bit for communication on NVIDIA C2C is approximately 1.3 pJ/bit.

Another approach to increasing overall communication bandwidth between dies was pursued by the TESLA DOJO team in their D1 Training Processor [TFis23]. It uses a highly parallelized computing architecture with 25 large dies connected using high soeed SerDes for communication. Between two edges of large computing tiles of area 645 mm^2 each, there are 144 SerDes running at 112 Gbps per link, which gives an overall die-to-die bandwidth of 2 TB/s. The approximate bandwidth per shoreline is ~640 Gbps/mm. The packaging technology used by Tesla DOJO is integrated fan out (InFO) [TSMCI] from TSMC, where large dies are embedded first in molding compound before fanout interconnection are made. The IO power can be estimated by considering approximately



(a) **(b)**

Figure 11.2 Commercially available high interconnect bandwidth solutions: (a) NVIDIA H100 using NVIDIA C2C [Wei23] with 260 Gbps/mm bandwidth/shoreline (b) Tesla Dojo D1 training processor [Tfis23] with 640 Gbps/mm bandwidth/shoreline.

10kW of power needed by 25 D1 tiles [TDojo], of which ~20% is assumed to be IO power. This gives us an energy per bit of approximately 1.5 pJ/bit assuming SerDes on all four edges of the die are communicating simultaneously.

In both the approaches discussed, we observe the use of advanced packaging technology such as chip-last CoWoS process [TSMCC] or chip-first InFO process [TSMCI] for die integration and use of SerDes for communication. Advanced packaging refers to integration of multiple dielets at fine bump pitch and short inter-dielet spacing with heterogeneity and scale. Advanced packaging approaches are central to improving bump and interconnect integration density [Brain16], increasing inter-die bandwidth and reducing communication overheads. While the adoption of advanced packaging approaches in industry has increased over the past decade, there are a few alarming trends which threaten to undermine the effectiveness of advanced packaging in increasing the compute, memory and interconnect throughput of a high-performance system. In the next section, we will discuss some of these trends and how our approach has the potential to resolve these challenges.

1.2 Current trends in inter-dielet communication bandwidth scaling and associated challenges

As the demand for inter-die/dielet communication bandwidth increases, today's advanced packaging approaches based on interposer and/or fanout technology and using SerDes for communication will face increasingly greater challenges. We elaborate on some of these challenges below.

- **Use of SerDes for inter-dielet communication:** While implementation of parallelized hardware will benefit greatly from reduced communication overhead between different computing nodes, instead, we observe an opposite trend. Nearly all communication between processors and memory is done through SerDes which are area inefficient and power hungry. The SerDes IO circuitry in today's processors can exceed upwards of 25 % [TFis23], [Sbri] of the chip design area. Power overheads of IOs today can exceed a third (> 30%) of the power budget [Iyer16]. One also needs to consider the efficiency of voltage conversion when delivering power from an external socket in addition to IO power. In fact, any power which is not used for computation is undesired. To address the growing communication needs for AI hardware, the SerDes designs are becoming increasingly faster, more complex with each generation with extensive equalization and error-correction schemes. This is not a sustainable trend. These overheads are only going to grow worse in future as communication energy, latency and IO bandwidth continue scale poorly compared to computation.

- **Use of interposers in packaging:** inclusion of silicon interposers introduces several challenges. Firstly, interposers are an additional level in the packaging hierarchy between dies/dielets and laminates and make the packaging process complicated. The fabrication of interposers requires wafer-grinding [AyodWB], polishing and planarization [CMPov], through silicon via (TSV) fabrication [Bruk20], use of carrier substrate for fabrication, temporary bonding and debonding etc. which makes fabrication challenging and reduces overall manufacturing yield. There are additional challenges due to stress associated with building multiple layers on thinned silicon, which can cause the interposer to delaminate or crack

unless significant process optimization is done. This thinning of silicon interposers is also one of the main reasons why scaling interposer sizes to multiple reticles is difficult, though it can be achieved [Hua21] through direct write and mask stitching. So, in conclusion, while interposers can improve interconnect bandwidth to an extent by introducing silicon processes in their fabrication, they do so at the cost of introducing complexity to IO design, adding multiple fabrication steps and complicating the assembly processes. They also require an extensive supply chain with multiple materials for fabrication, putting limits on time-to-market and product security. Today, we observe that interposer technology is slowly phasing out.

- **Fanout packaging and limitations:** Fanout approaches have high potential for scaling to larger (up to panel) assemblies. Fanout technology can be chip-first or chip-last as shown in Fig. 1.3. Chip-first fanout involves die/dielet assembly on a carrier substrate, followed by wafer/panel level molding, carrier release and wiring layer fabrication to connect the dies/dielets [Sou05], [Tak18]. This technology is affected by die-shift [Kho09] during molding and limitations on wiring pitch when compared to silicon-based packaging. However, with the development of various die-shift identification and correction algorithms [Sha11], [Ouy22] and innovative adaptive pattering approaches such as TrueAdapt™ [Sab23] it is possible to connect the dielets after die-shift. Chip-last fanout involves fabrication of wiring layers prior to dielet placement and bonding. These techniques are not affected by die-shift, however the die-placement accuracy depends on the accuracy of camera vision system of the die placement tool.

- **Architectural limitations to scale-out in current packages**: As the researchers have pointed out in [Slie23], traditional scale-out parallelism techniques have challenges. The fundamental reason is because, in traditional scale-out, memory and compute are tied to each other. Trying to run a single machine learning/artificial intelligence model on thousands of devices turns the scaling of both memory and compute into distributed constraint problems that are interdependent. Conventional approach may just replicate the model in each individual device, but

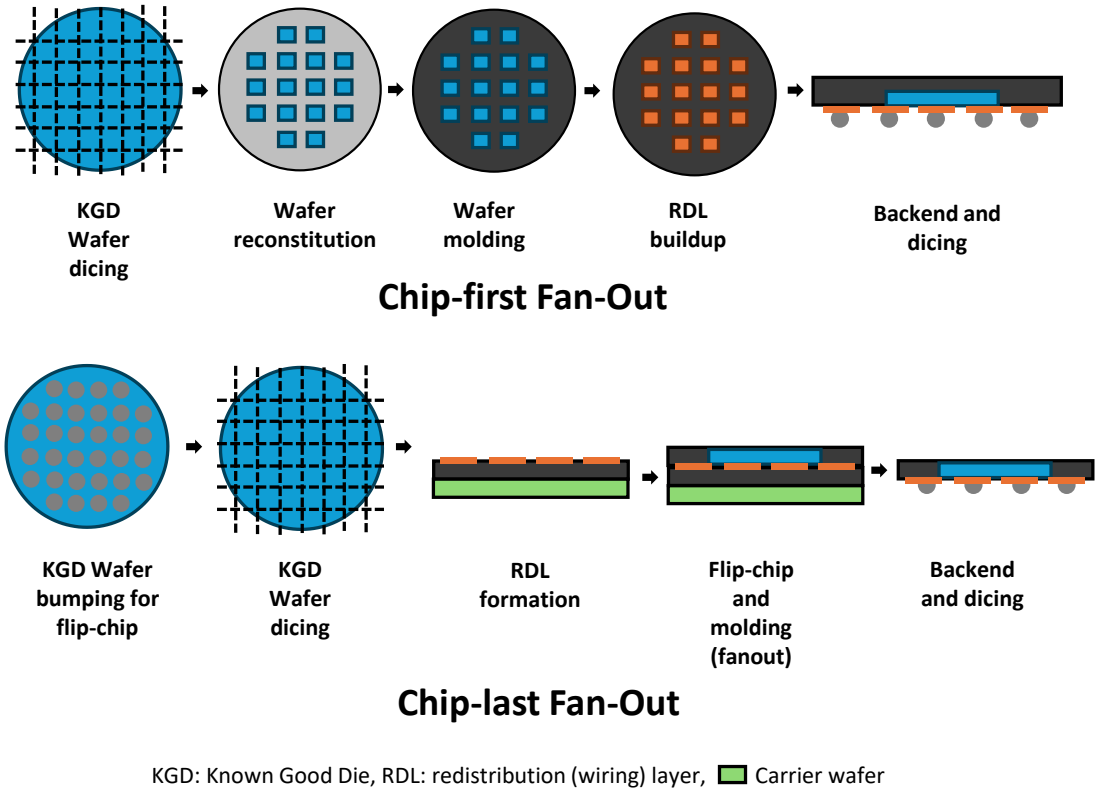


Figure 11.3 Chip first and Chip last Fan-Out processes.

it does not work well when very large models are involved which cannot be contained into a single chip or few chips.

1.3 Scale-down and Scale-out using wafer-scale integration (WSI)

To address the growing interconnect bandwidth and memory requirements, we have proposed a different approach in our previous works [Sjan21], [Spal19]– a wafer-scale computing system. In our approach, a wafer scale system refers to a system of tightly integrated processor and memory dielets assembled using advanced packaging approaches which communicate with one another using an area and energy efficient on-wafer interconnect. Our wafer scale integration platform is called silicon interconnect fabric (Si-IF) which consists of copper (Cu) damascene wiring layers terminating with protruded Cu pillars on a Silicon (Si) substrate, to which bare dielets with Cu pads are bonded using thermal compression bonding (TCB). We will elaborate further on Si-IF platform and integration process in section 1.4. In this section we would like to focus on

the key advantages of building wafer-scale systems. These can be summarized as follows:

- **A multi-functional heterogeneous platform:** A wafer scale integrated (WSI) system can include heterogeneous dielets of varying functional types and physical dimensions such as processor, memory, power etc. on the same silicon wafer. These dielets may be further designed and fabricated using different manufacturing technologies and yet be assembled on the same wafer. Hence WSI represents a highly flexible and versatile packaging platform. Furthermore, each dielet to be bonded to a wafer-scale silicon package can be a known-good die (KGD) ensuring risk reduction in assembly.
- **An ideal platform for increasing dielet-to-dielet bandwidth via "scale down":** Scale-down refers to a steady reduction in all packaging dimensions (including bonding pitch, line and space on package) and increase in number of wiring layers per package. This approach enables the designer to achieve aggressive interconnect bandwidths using lower data-rates (< 10 GHz per wire) for communication, which makes data transfer feasible at bit error rates of 10^{-15} or below, without the need for complex error correction, equalization schemes and such. It would rely on using simple input/output (I/O) designs such as a chain of buffers/inverters [Ksah22, Rfar20] resulting in a low energy-per-bit, low I/O area and low latency link for communication. A wafer-scale system is an ideal platform to achieve fine sub-μm line/space using stepper or direct write lithography approaches. Today 4μm interconnect pitches are routinely achieved on the Si-IF platform [Sjan21] and it will be scaled to sub-μm values in the coming years as the lithographic technology for scale-down already exists. Dielet integration pitches, i.e. bonding pitches today are limited by alignment accuracy of die-to-wafer/die-to-substrate bonders and not due to limitations of manufacturing/fabrication equipment. As these bonding tools improve over time, the bonding/bump pitches will easily scale down. However, die-to-die communication circuits need to change for wafer-scale systems to take advantage of finer bonding and interconnect pitches.

- **An ideal platform for "scale out" packaging:** Scale-out refers to architecting a system with intimately connected semiconductors (Si, III-Vs etc.) and other functional components (passives, sensors, energy storage) through a dense network of chiplets/dielets instead of large chips. Scale-out therefore involves building systems which will evolve over time using chiplets/dielets to provide additional functionality and system-level resiliency. Wafer-scale systems are designed to replace multiple packaging hierarchy levels with a single packaging level which extends to a full 300 mm silicon wafer. Thousands of dielets can be assembled on a single wafer to achieve diverse functionality as well as large memory capacities. In fact, the challenges of memory capacity in AI can be completely resolved by wafer-scale systems which can store several petabytes of data per wafer by use of thousands of stacked memory dielets such as high bandwidth memory (HBM) dielets.
- **Significant potential for programmability:** Compared to GPU clusters or processor-to-memory communication using backplane, wafer-scale systems have considerably lower overhead when it comes to inter-die and intra-die communication. These advantages have been elaborated upon by work done by Iyer et al. [Iyer16] and Pal et al. [Spal19]. This means programmers do not have to be overly concerned about data access across multiple dielets and compilers can generate more flexible and fine-grained hardware resource partitioning and task mapping to effectively utilize the computing resources available.

These reasons should prove the importance of wafer-scale systems in addressing the bandwidth and memory bottlenecks which are increasingly becoming apparent as today's packaging approaches are being scaled to meet the ever-growing AI, graph processing and scientific demands.

While we primarily think of WSI as a platform to integrate thousands of heterogeneous dielets at dense interconnect pitches, an alternative approach to WSI exists in the form of a single full-wafer monolithic processor. This is the case of Cerebras systems wafer scale engine (WSE 1 and 2) [Glau21], [Slie23], where an entire wafer is fabricated to act as a processor. In WSE-2, they have reported 850,000 cores on a single wafer, and a

fabric bandwidth of 4.3 Tbps/mm [Slie23]. However, we believe our approach has several advantages. First, as we know every fabricated wafer has a yield, and manufacturing defects in the wafer can lead to a percentage of devices on any wafer to be non-functional. As a result, using a full wafer as a processor would require devoting either software or hardware resources to the wafer to re-route signals or reorganize hardware partitioning in the event some of the core or memory elements are found to be defective. A great deal of redundancy must be built-in into the system before it is deployed. On a heterogeneous approach to wafer-scale however, the user can select known good dielets (KGDs) to be bonded, bypassing this issue. It should be noted though, the process of assembling a dielet to the substrate also has a yield to be considered. Assembly yield is improved using better tools, better processes and has been steadily improving with each new generation of die-to-wafer and wafer-to-wafer bonding tool and novel advanced packaging techniques. Manufacturing yield on the other hand, requires significant time and resources to improve with each advancement of technology node due to rising complexity and cost. It is not surprising to have a below 50 % yield on a 2nm manufacturing process even at a commercial foundry. As a result, we expect the KGD based wafer-scale approach to eventually be a better integration platform.

Second, wafer-scale system populated with dielets provide heterogeneity both in terms of functionality of the dielets e.g. compute, memory, power etc, as well as manufacturing technology e.g. dielets of different technology nodes can be assembled on the same wafer. This allows a fully populated wafer scale system to integrate the most advanced processor dielets (2nm or below fabrication process) with most advanced stacked memory dielets including high bandwidth memories (HBM4, HBM4E etc.). This is not possible with a full wafer-scale processor. In fact, in case of Cerebras, the on-wafer memory is primarily static random-access memory (SRAM) [Slie23] which limits the amount of on-wafer memory when compared to integrating dynamic RAM (DRAM) HBM dielets on a wafer-scale platform such as Si-IF.

Either implementation of wafer-scale therefore has advantages and disadvantages. For this thesis and our group at UCLA, we have taken the heterogeneous chiplet/dielet based

wafer-scale integration approach. Henceforth wafer-scale in this thesis only refers to the approach taken at UCLA-CHIPS. The platform is discussed in the next section.

1.4 Silicon interconnect fabric (Si-IF) as a wafer-scale integration platform

Silicon interconnect fabric (Si-IF) is our wafer-scale platform for integrating heterogeneous dielets at sub-10 μm pitch and sub-100 μm inter-dielet spacing. The cross-section of Si-IF is shown in Fig. 1.4 (a). Fig 1.4 (b) and 1.4 (c) show wafer level passive die implementations with 460 dielets and 240 dielets respectively. For 1.4 (c) we were able to verify connectivity across the entire assembly [Ksah21]. In general, the Si-IF substrate has the following features:

- It consists of multiple wiring layers fabricated using single or dual copper (Cu) damascene fabrication process described in work done by Jangam et al. [Sjan20t]. We do not go into the details of the fabrication process in this work as it is not a contribution to this thesis. However, the process flow has been already published in Siva Jangam's thesis work [Sjan20t].
- The wiring layers terminate with Cu pillars at 10 μm or below bonding pitch and bare dielets with exposed copper or gold pads are integrated directly onto the copper pillars using direct metal-metal thermal compression bonding (TCB). We have already shown Au-Au [Abaj17] and Cu-Cu TCB [Sjan19] in our previous works. This thesis focuses on scaling up the Cu-Cu TCB process as well as the development of a novel Au-Cu [Ksah21] and Aluminum (Al)-Cu [Ksah24] TCB process. Our goal is to make Si-IF an integration platform which can accept any kind of metal termination and integrate them at ≤ 10 μm bonding pitch.
- The Si-IF can further include passive components such as metal-insulator-metal capacitors (MIM-CAPS) [CTY24], deep-trench capacitors (DECAPS) [Jliu25], and inductors for power delivery purposes. In some cases, Si-IF can also be active [RenVLSI]

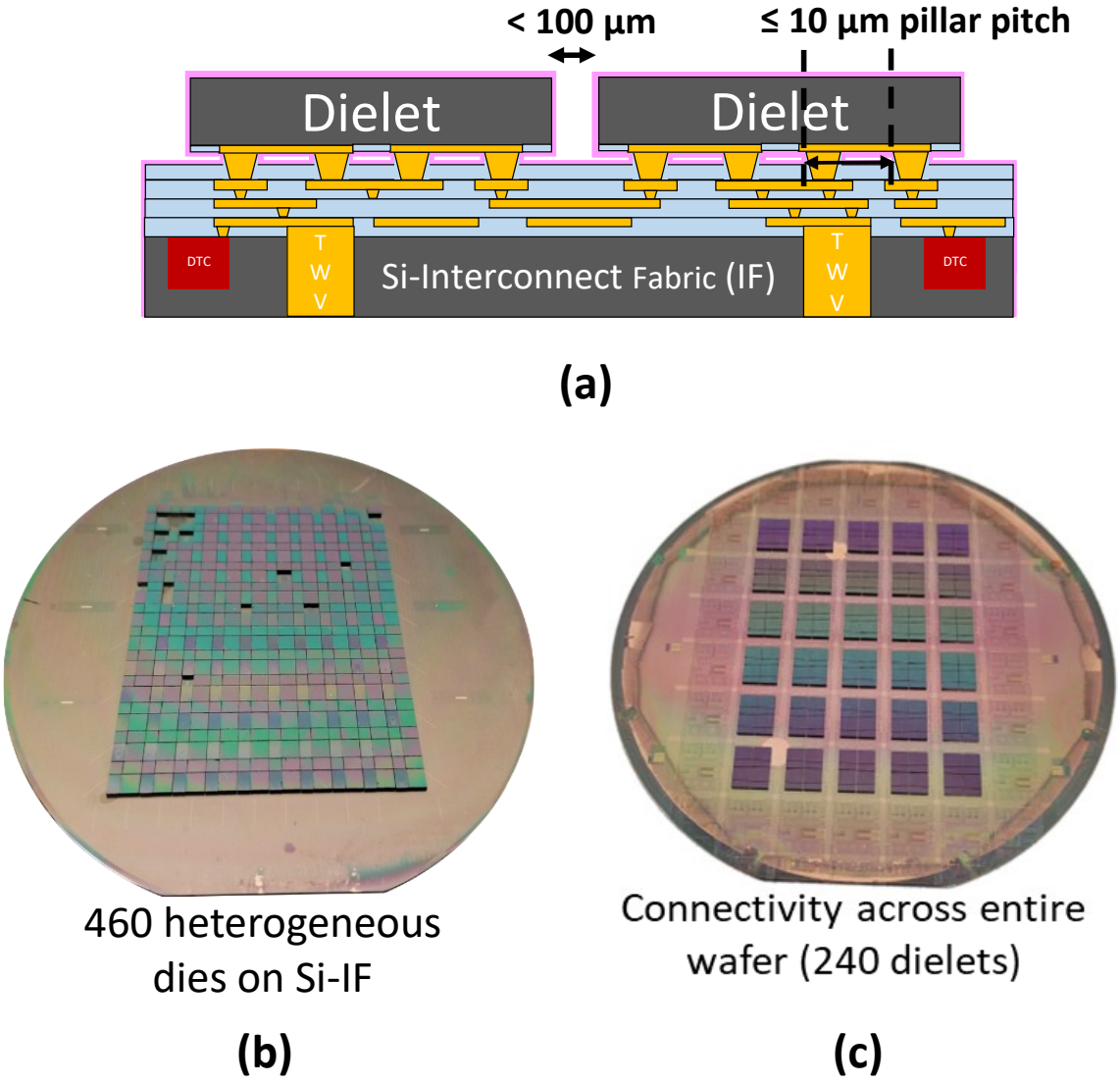


Figure 11.4 (a) Schematic cross-section of Silicon Interconnect Fabric (Si-IF) assembled with dielets. (b) implementation of 460 dielets using Cu-Cu thermal compression bonding (TCB) (c) implementation of 240 dielets using gold-copper TCB with connectivity verification.

- It may also include wafer vias (TWVs) for power and signal delivery, but this process flow is currently being developed for double sided assembly
- The Si-IF platform can also be used to stack multiple dielets.

- It is further possible to extend Si-IF to include laminated organic polymer films on silicon wafer at coarse wiring pitches, followed by low temperature oxide deposition for finer-pitch wiring layers. More information regarding this approach is described in [Vhar24].

All these components are included in the schematic presented in Fig. 1.4 (a). The dielets can be assembled at sub-100μm inter dielet spacing.

Si-IF design follows rules from a design manual which is available upon requirement. The current specifications of a Si-IF wafer with six Cu wiring layers is presented in Table 1.1 below along with future scaling:

Table 1.1 Silicon Interconnect Fabric (Si-IF) design rule values with future scale-up

Parameter	Current Value	Future roadmap value by 2035 as per MRHIEP report [Mrhiep]
Line/Space	2 μm / 2 μm	0.7 μm / 0.7 μm
Bump pitch	10 μm	1 μm
Number of wiring layers	2	6 (4 signal + 2 power)
Layer thickness	2 μm (nominal)	Can vary

Beyond the Si-IF substrate with assembled dielets, an entire system of power delivery, thermal dissipation and communication scheme works to ensure reliable functionality of the system. The scale-out scheme is shown on Fig. 1.5. The space between dielets as well as the top of the dielets is filled with copper which is then planarized. A thermal dissipation unit (TDU) is then attached to the top of the planarized Cu surface as seen in the schematic. Due to a wafer-scale system generating heat in the range of 1 W/mm^2 to 4 W/mm^2 [Usah25], conventional forced air cooling or liquid cooling approaches are inadequate in cooling down the system. A novel two-phase cooling approach [Usah25] has been proposed as a thermal solution for Si-IF.

On the backside of the Si-IF, voltage conversion and power delivery dielets are assembled. It is expected the wafer-scale system receives voltage at 48V from an inlet and then converts it into core and IO voltage at point of load (POL) as shown in Fig 1.5. The conversion can be a direct 48V to POL voltage conversion or a step-based conversion i.e. 48V to 12 V followed by 12V to POL voltage conversion. Details of the conversion along with dielet design for voltage conversion and power delivery are presented in the work done by Ren et al. [Hren23].

Communication within dielets on a wafer-scale system is done via a low-cost on-wafer interconnect called SuperCHIPS. SuperCHIPS [Sjan17] or Simple Universal Parallel intERface for CHIPS is a fine-IO pitch communication protocol which uses simple inverters for data transfer between dielets. Due to its simple architecture, each IO can be integrated to within 10 × 10 μm^2 design area and link energy of communication can be as small as 0.03 pJ/bit [Sjan21]. However, this value is for asynchronous communication using SuperCHIPS which does not involve any flip-flops and ESD for IOs. With the integration of clock and custom ESD which is discussed in Chapter 5 of this thesis, a synchronous SuperCHIPS link with ESD consumes about 0.13 pJ/bit which is very small energy for data transfer even for a synchronous link. It should be noted that the SuperCHIPS links cannot drive very far distances. SuperCHIPS is designed to work at 4Gbps/link for 0.5 mm long link lengths as discussed in chapter 5. This is ideal for a wafer-scale system. For long-haul communication such as multi-wafer communication and protocol translation for communicating with external hardware which uses conventional PCIe protocols is done using a translator chiplet called Anuvada. Details of how wafer level communication is implemented on Si-IF is discussed in Chapter 5.

A fully scaled out wafer-scale system is presented in Fig. 1.5.

1.5 Existing Challenges with implementation of wafer-scale system

Wafer-scale is a transformative “scale down and scale out” solution. Not surprisingly, this approach also brings several unprecedented challenges which must be addressed. A summary of these challenges for a heterogeneous wafer-scale system is as follows:

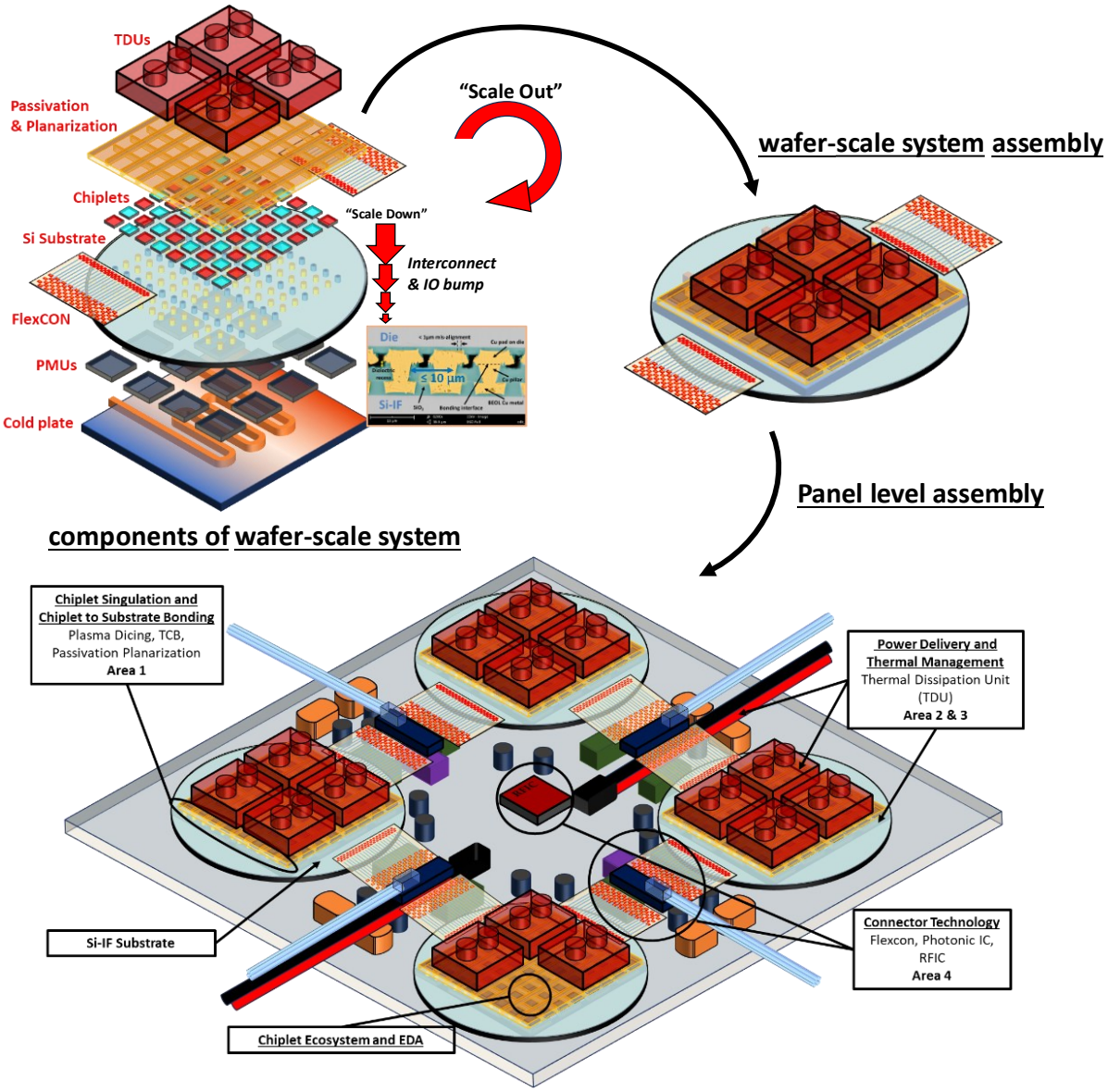


Figure 11.5 Scale-down and Scale-out wafer-scale system on Si-IF showing dielets (compute, memory etc. assembled on an Si-IF with necessary passives, power delivery (power management units), thermal dissipation and connectors for wafer-scale communication [ApexA].

- **Need for novel packaging techniques:** Wafer-scale systems can integrate several hundreds to thousands of dielets into the same wafer depending on dielet size. Furthermore, the dielets must be integrated at fine pitch to take advantage of the low-cost on-wafer links for communication. This necessitates novel assembly

approaches such as metal-metal thermal compression bonding or metal/dielectric hybrid bonding approaches for achieving a fine ≤ 10μm bonding pitches at relatively high throughput. In this research work, we overcome the bonding throughput challenge by proposing and demonstrating a high-throughput two-step Cu-Cu thermal compression bonding process. The potential throughput of this process is > 1000 units per hour (UPH). The methodology is discussed in detail in Chapter 2.

- **Need for novel architecture for wafer-scale:** Many of the current hardware models today exhibit limited scalability. When building a wafer-scale system it is important to consider architectural questions on
    - Clock distribution
    - Power delivery
    - Testing methodology for the assembled dielets
    - Redundancy during wafer-scale operation

  Such questions need to be answered. Fortunately, the research done by Saptadeep Pal [Spal21t], has taken into consideration these questions and proposed feasible solutions which can help when building wafer-scale systems.

- **Wafer-scale communication:** We have discussed that one of the selling points of a wafer-scale integration is that thousands of dielets can communicate via an area and power efficient signaling protocol. We expect simple inverters to be able to drive these links. A demonstration of inverter-based communication was done by Jangam et al. in 2020 [Sjan20t]. This link is termed as SuperCHIPS or Simple Universal Parallel intERface for CHIPS. In this work we have improved SuperCHIPS with inclusion of dedicated custom ESD protection circuitry which will be further elaborated in Chapter 5. However, SuperCHIPS is designed for short-haul communication (up to 1mm). In a wafer-scale system, sometimes it may be required to communicate across multiple dielets or bypass certain dielets – which can be dubbed as medium-haul communication. Furthermore, some data signals

may need to bypass an entire wafer in a multi-wafer system. It is also required for the wafer-scale system to communicate with external hardware which is usually designed to use PCIe or CXL protocols for communication. We term these types of communication as long-haul communication. In short, a wafer-scale communication solution must address short-haul (dielet-to-dielet), medium-haul (between separated dielets on wafer) and long-haul (between wafers or between wafer and external hardware). This research addresses long-haul communication by using protocol translation. We have designed a protocol translation chiplet called Anuvāda which can accept PCIe and convert it into SuperCHIPS and vice versa. More details are given in Chapter 5.

- **Need for novel power architecture:** A wafer scale system will need power delivery exceeding 50kW for all the dielets to operate. Delivering such a large amount of power at low supply noise and with sufficient decoupling capacitors to account for droop, current demand etc. is no trivial task. Ren et al [Hren23] have discussed the implementation of a segmented power delivery architecture including power conversion from a 48V wall outlet supply to point of load (POL) voltage and the use of small passives to support high switching frequency (50MHz and above) to aid in power delivery for wafer scale systems. The reader is requested to read their work to learn more.

- **Need for novel thermal solution:** Given a power delivery of 50kW or more, we expect the wafer-scale system to generate anywhere between 1 W/mm^2 to 4W/mm^2 of heat, with transients rising to 6 W/mm^2. Not only is this a significant amount of heat generated, but also, the heat must be extracted from within the shadow of the wafer without the use of a heat spreader or large heat sink to ensure compatibility with rack mounted systems. This only leaves a two-phase cooling approach on the table to extract heat from a wafer-scale assembly. Shah et al. [Usah25] have proposed a novel two-phase cooling strategy using water and methanol mixture to extract 1 W/mm^2 or greater heat from wafer-scale system and currently work is being done to demonstrate capability up to 4 W/mm^2 [NMSu25].

- **Novel software for wafer-scale:** From a software perspective, today's software has not been introduced to the large computing power and memory capacity available on a wafer-scale system. New compilers are required to make use of the raw power and low inter-dielet communication latency provided by wafer-scale systems. There needs to be an efficient mechanism to fully leverage the power of such systems.

1.6 Contributions of the thesis

In this thesis work, we specifically handle two of the above challenges:

- A novel packaging technology with manufacturable (1000+ UPH) throughput using metal-metal solderless thermal compression bonding for wafer-scale systems.
- Account for protocol interoperability on the wafer-scale system to enable communication with external hardware and other wafers.

In the process of understanding the above challenges and overcoming them, this thesis generates several contributions. Some notable contributions are:

- Demonstration of a novel two-step copper-copper (Cu-Cu) thermal compression bonding technique for high throughput assembly required by wafer-scale systems. At UCLA facility, we have demonstrated a throughput of 330 units-per-hour (UPH) a > 3x throughput improvement over conventional single step Cu-Cu TCB.
- A detailed design of experiments (DOE) for achieving the above throughput and potential to increase the throughput to 1100 UPH by improving the die tacking process.
- A novel Gold-to-Copper (Au-Cu) thermal compression bonding process with gold passivated Cu pillars to increase the duration the Si-IF wafer on a heated chuck for the cases of bonding which can take over several hours even at high throughput.

- A detailed account of additional challenges besides throughput which affect the assembly yield. Specifically, the effect of parallelism on bonding, alignment drifts and dicing on TCB yield will be discussed.
- A novel chip partitioning figure of merit (p-FOM) accounting for the power and area overhead when partitioning a monolithic chip into chiplets. This figure of merit provides guidance to the system designer for ideal chiplet size in addition to the guidance provided by the work done by Iyer et al. [Iyer16].
- An improvement of signaling figure of merit (s-FOM) first reported by Jangam et al. [Sjan20s] which considers additional factors like bit-error-rate and packaging complexity in choice of signaling and packaging technology. An example of choice between two systems with different memory types—hybrid memory cube (HMC) and high bandwidth memory (HBM)— is done to determine the better of the two in terms of signaling and package choice. This cost factor (CF) introduced in this work provides the signaling and package designers with information on deciding the ideal package and signaling protocol given a particular set of inter-dielet bandwidth, signal integrity, memory etc. requirements.
- A novel protocol translator chiplet design for enabling protocol interoperability on wafer-scale systems. This protocol translator chiplet is called “Anuvāda” (Sanskrit for translation) and as a proof-of-concept is designed to translate between PCIe signals and SuperCHIPS signals and vice versa. The goal with translator chiplets is to achieve translations between common protocols such as PCIe, CXL, UCIe etc. using SuperCHIPS as an intermediate signaling scheme. SuperCHIPS interface is designed to operate at 10μm pitches to minimize the area of design and energy of communication. Once functionally verified, the translator chiplet will belong to a marketplace of chiplets used by different wafer-scale systems. It should be noted that this work focuses on proof-of-concept PCIe 3.1 to SuperCHIPS and vice versa conversion.
- A novel electrostatic discharge protection circuitry for SuperCHIPS fine pitch (10μm bonding pitch) IOs including testing the transmission line pulse (TLP)

performance of the ESD diodes implemented in the SuperCHIPS block in Anuvāda chiplet.

This is a summary of the key contributions. The process flows developed during these tasks are detailed in the publications as well as standard operating procedures (SOP) available upon request. The codes used for designing of the Anuvāda chiplet are available on GitHub here.

1.7 Chapter breakdown and Conclusion

In this introductory chapter, we have introduced the importance of wafer-scale systems and their current challenges with implementation. A chapter wise breakdown is presented:

- Chapter 2 discusses the design of experiments used for developing a high throughput Cu-Cu thermal compression bonding process to be used in wafer-scale assemblies. It also discusses the key mechanical and electrical reliability results as per established industry standards like MIL-SPEC 883 [MILS883] and Joint Electron Device Engineering Council (JEDEC) reliability standards [JEDECWIKI].
- Chapter 3 discusses the effect of various factors such as bond-head parallelism, alignment drift and dicing edge roughness of dielets on the overall assembly yield. This is crucial to understand in the process of building a wafer-scale system.
- Chapter 4 introduces partitioning figure of merit (p-FOM) with for partitioning a large monolithic chip into smaller chiplets, which takes into account the power overhead, area overhead of the terminals added as well as the improvement in yield of the chiplets compared to the monolithic chip. An improvement of signaling figure of merit (s-FOM) is also presented in this chapter.
- Chapter 5 gives details regarding the concept of translator chiplet in wafer-scale packaging. It includes the design highlights of the translator chiplet Anuvāda, including test results from the dielet, fabricated in GlobalFoundries 22 nm FDSOI technology. It also discusses a novel electrostatic discharge protection for 10 μm fine pitch IOs and its ESD performance.

The contributions outlined in this work mark significant progress toward the realization of practical, heterogeneous wafer-scale systems. By addressing key implementation barriers—specifically assembly throughput and protocol interoperability—this research lays the foundation for future systems that can overcome today's bandwidth and memory limitations in AI and high-performance computing. These innovations represent critical milestones on the path to scalable, energy-efficient computing architectures of the future.

Chapter 1 References

[Ghol24] A. Gholami, Z. Yao, S. Kim, C. Hooper, M. W. Mahoney, and K. Keutzer, "AI and Memory Wall," *IEEE Micro*, vol. 44, no. 3, pp. 33–39, May 2024, doi: 10.1109/MM.2024.3373763.

[Iyer16] S. S. Iyer, "Heterogeneous Integration for Performance and Scaling," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 6, no. 7, pp. 973–982, Jul. 2016, doi: 10.1109/TCPMT.2015.2511626.

[Spal19] 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)*, Feb. 2019, pp. 250–263. doi: 10.1109/HPCA.2019.00042.

[Syn1] "What is SerDes (Serializer/Deserializer) ?", Synopsys glossary [online], available: https://www.synopsys.com/glossary/what-is-serdes.html

[WuMK95] Wm. A. Wulf and S. A. McKee, "Hitting the memory wall: implications of the obvious," *SIGARCH Comput. Archit. News*, vol. 23, no. 1, pp. 20–24, Mar. 1995, doi: 10.1145/216585.216588.

[WikiHMC] Hybrid Memory Cube [online], available: https://en.wikipedia.org/wiki/Hybrid_Memory_Cube

[HMCSpec] Hybrid Memory Cube Specification V1.0, Wayback Machine [online], available: https://web.archive.org/web/20130513053443/http://www.hybridmemorycube.org/files/SiteDownloads/HMC_Specification%201_0.pdf

[WikiHBM] High Bandwidth Memory [online], available: https://en.wikipedia.org/wiki/High_Bandwidth_Memory

[Brain16] R. Brain, “Interconnect scaling: Challenges and opportunities,” in *2016 IEEE International Electron Devices Meeting (IEDM)*, Dec. 2016, p. 9.3.1-9.3.4. doi: 10.1109/IEDM.2016.7838381.

[NVH100] “NVIDIA H100 Tensor Core GPU Datasheet,” NVIDIA. Accessed: Apr. 22, 2025. [Online]. Available: https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet

[Wei23] Y. Wei *et al.*, “9.3 NVLink-C2C: A Coherent Off Package Chip-to-Chip Interconnect with 40Gbps/pin Single-ended Signaling,” in *2023 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2023, pp. 160–162. doi: 10.1109/ISSCC42615.2023.10067395.

[TSMCC] Chip-on-Wafer-on-Substrate, TSMC [online], available: https://3dfabric.tsmc.com/english/dedicatedFoundry/technology/cowos.htm

[Intposr] “Interposers,” Semiconductor Engineering. Accessed: Apr. 22, 2025. [Online]. Available: https://semiengineering.com/knowledge_centers/packaging/advanced-packaging/2-5d-ic/interposers/

[TFis23] T. C. Fischer *et al.*, “9.1 D1: A 7nm ML Training Processor with Wave Clock Distribution,” in *2023 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2023, pp. 8–10. doi: 10.1109/ISSCC42615.2023.10067658.

[TSMCI] InFO (Integrated Fan-Out) Wafer Level Packaging, TSMC [online]. Available: https://www.tsmc.com/english/dedicatedFoundry/technology/InFO

[TDojo] “Tesla Dojo – Unique Packaging and Chip Design Allow An Order Magnitude Advantage Over Competing AI Hardware,” SemiAnalysis. Accessed: Apr. 22, 2025. [Online]. Available: https://semianalysis.com/2021/08/20/tesla-dojo-unique-packaging-and-chip/

[Sbri] “Sandy Bridge (client) - Microarchitectures - Intel - WikiChip.” Accessed: Apr. 28, 2025. [Online]. Available: https://en.wikichip.org/wiki/intel/microarchitectures/sandy_bridge_(client)

[AyodWB] A. Ayodele, “Wafer Backgrinding: An In-Depth Guide to Semiconductor Manufacturing.” Accessed: Apr. 22, 2025. [Online]. Available: https://www.wevolver.com/article/wafer-backgrinding-an-in-depth-guide-to-semiconductor-manufacturing, https://www.wevolver.com/article/wafer-backgrinding-an-in-depth-guide-to-semiconductor-manufacturing

[CMPov] “Chemical Mechanical Planarization - an overview | ScienceDirect Topics.” Accessed: Apr. 22, 2025. [Online]. Available: https://www.sciencedirect.com/topics/materials-science/chemical-mechanical-planarization

[Mrhiep] MRHIEP final report. [Online]. Available: https://www.chips.ucla.edu/page/MRHIEP/MRHIEP%20Final%20Report

[Bruk20] S. L. Burkett, M. B. Jordan, R. P. Schmitt, L. A. Menk, and A. E. Hollowell, “Tutorial on forming through-silicon vias,” *Journal of Vacuum Science & Technology A*, vol. 38, no. 3, p. 031202, Apr. 2020, doi: 10.1116/6.0000026.

[Hua21] P. K. Huang *et al.*, “Wafer Level System Integration of the Fifth Generation CoWoS®-S with High Performance Si Interposer at 2500 mm2,” in *2021 IEEE 71st Electronic Components and Technology Conference (ECTC)*, Jun. 2021, pp. 101–104. doi: 10.1109/ECTC32696.2021.00028.

[Sou05] J.-C. Souriau, O. Lignier, M. Charrier, and G. Poupon, “Wafer level processing of 3D system in package for RF and data application,” in *Proceedings Electronic Components and Technology, 2005. ECTC ’05.*, May 2005, pp. 356-361 Vol. 1. doi: 10.1109/ECTC.2005.1441291.

[Tak18] T. Fukushima, A. Alam, A. Hanna, S. C. Jangam, A. A. Bajwa, and S. S. Iyer, “Flexible Hybrid Electronics Technology Using Die-First FOWLP for High-Performance and Scalable Heterogeneous System Integration,” *IEEE Transactions on Components,*

Packaging and Manufacturing Technology, vol. 8, no. 10, pp. 1738–1746, Oct. 2018, doi: 10.1109/TCPMT.2018.2871603.

[Kho09] C. H. Khong *et al.*, "A novel method to predict die shift during compression molding in embedded wafer level package," in *2009 59th Electronic Components and Technology Conference*, May 2009, pp. 535–541. doi: 10.1109/ECTC.2009.5074066.

[Sha11] G. Sharma, A. Kumar, V. S. Rao, S. W. Ho, and V. Kripesh, "Solutions Strategies for Die Shift Problem in Wafer Level Compression Molding," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 1, no. 4, pp. 502–509, Apr. 2011, doi: 10.1109/TCPMT.2010.2100431.

[Ouy22] G. Ouyang *et al.*, "Comprehensive Investigation of In-Plane and Out-of-Plane Die Shift in Flexible Fan-Out Wafer-Level Packaging Using Polydimethylsiloxane," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 12, no. 10, pp. 1692–1701, Oct. 2022, doi: 10.1109/TCPMT.2022.3207031.

[Sab23] G. Sabbir, K. Sahoo, H. Sun, S. Iyer, G. Ouyang, and Y. Wu, "TrueAdaptTM- AI Based Maskless Patterning to Compensate for Die-Shift in Fan-Out Wafer Level Packaging," in *2023 IEEE 73rd Electronic Components and Technology Conference (ECTC)*, May 2023, pp. 2240–2246. doi: 10.1109/ECTC51909.2023.00388.

[Sjan21] S. Jangam and S. S. Iyer, "Silicon-Interconnect Fabric for Fine-Pitch (≤10 μm) Heterogeneous Integration," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 11, no. 5, pp. 727–738, 2021, doi: 10.1109/TCPMT.2021.3075219.

[Ksah22] K. Sahoo, U. Rathore, S. Chandra Jangam, T. Nguyen, D. Markovic, and S. S. Iyer, "Functional Demonstration of < 0.4-pJ/bit, 9.8 μm Fine-Pitch Dielet-to-Dielet Links for Advanced Packaging using Silicon Interconnect Fabric," in *2022 IEEE 72nd Electronic Components and Technology Conference (ECTC)*, May 2022, pp. 2104–2110. doi: 10.1109/ECTC51906.2022.00332.

[Rfar20] R. Farjadrad, M. Kuemerle, and B. Vinnakota, “A Bunch-of-Wires (BoW) Interface for Interchiplet Communication,” *IEEE Micro*, vol. 40, no. 1, pp. 15–24, Jan. 2020, doi: 10.1109/MM.2019.2950352.

[Glau21] G. Lauterbach, “The Path to Successful Wafer-Scale Integration: The Cerebras Story,” *IEEE Micro*, vol. 41, no. 6, pp. 52–57, Nov. 2021, doi: 10.1109/MM.2021.3112025.

[Slie23] S. Lie, “Cerebras Architecture Deep Dive: First Look Inside the Hardware/Software Co-Design for Deep Learning,” *IEEE Micro*, vol. 43, no. 3, pp. 18–30, May 2023, doi: 10.1109/MM.2023.3256384.

[Sjan20t] Siva Jangam PhD thesis, “Heterogeneous Integration on Silicon-Interconnect Fabric using fine-pitch interconnects (≤10 μm)” [online]. Available: https://s3.us-west-1.amazonaws.com/chips.user.media/page_file/Jangam_dissertation_v6+Printed.pdf

[Abaj17] A. A. Bajwa *et al.*, “Heterogeneous Integration at Fine Pitch (≤ 10 μm) Using Thermal Compression Bonding,” in *2017 IEEE 67th Electronic Components and Technology Conference (ECTC)*, 2017, pp. 1276–1284. doi: 10.1109/ECTC.2017.240.

[Sjan19] S. Jangam *et al.*, “Fine-Pitch (≤10 μm) Direct Cu-Cu Interconnects Using In-Situ Formic Acid Vapor Treatment,” in *2019 IEEE 69th Electronic Components and Technology Conference (ECTC)*, 2019, pp. 620–627. doi: 10.1109/ECTC.2019.00099.

[Ksah21] K. Sahoo, S. Pal, N. Shakoorzadeh, Y.-T. Yang, and S. S. Iyer, “Copper to gold thermal compression bonding in heterogenous wafer-scale systems,” in *2021 IEEE 71st Electronic Components and Technology Conference (ECTC)*, 2021, pp. 487–493. doi: 10.1109/ECTC32696.2021.00088.

[Ksah24] K. Sahoo *et al.*, “Aluminum to copper thermal compression bonding for heterogeneous integration of legacy dielets,” in *2024 IEEE 74th Electronic Components and Technology Conference (ECTC)*, 2024, pp. 1940–1946. doi: 10.1109/ECTC51529.2024.00329.

[CTY24] C.-T. Yang, “Structure Design, Process Development, and Integration of MIM Capacitor in Si-IF for Wafer-Scale System,” M.S., University of California, Los Angeles,

United States -- California, 2024. Accessed: Apr. 24, 2025. [Online]. Available: https://www.proquest.com/docview/3103707951/abstract/EE4B451BB51D46DEPQ/1

[Jliu25] J. Liu et al., "Through GaN Via for 3D Heterogeneous Strata Integration," *2025 IEEE 75th Electronic Components and Technology Conference (ECTC)*, Dallas, TX, USA, 2025, pp. TBD, (accepted).

[Vhar24] V. Harish *et al.*, “Fine Pitch (≤ 10µm) Die to Wafer Cu-Cu TCB on Organic Polymer Build Up Films,” in *2024 IEEE 74th Electronic Components and Technology Conference (ECTC)*, May 2024, pp. 2178–2183. doi: 10.1109/ECTC51529.2024.00392.

[Usah25] U. D. Shah, “Transient Analysis of Flash Cooling for Dynamic Thermal Management for High-Performance Electronic Systems,” Ph.D., University of California, Los Angeles, United States -- California, 2023. Accessed: Apr. 24, 2025. [Online]. Available: https://www.proquest.com/docview/2835785749/abstract/1B23C4D66C8E4079PQ/1

[NMSu25] N. Meenakshi Sundaram, R. Pugazhendhi, S. Iyer, and T. Fisher, “Flash boiling of methanol/water mixtures in a microchannel cooler,” in 2025 24th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), May 2025. (accepted)

[Hren23] H. Ren *et al.*, “Heterogeneous Power Delivery for Large Chiplet-based Systems using Integrated GaN/Si-Interconnect Fabric with sub-10 µm Bond Pitch,” in *2023 International Electron Devices Meeting (IEDM)*, Dec. 2023, pp. 1–4. doi: 10.1109/IEDM45741.2023.10413759.

[Spal21t] S. Pal, “Scale-Out Packageless Processing,” 2021. [online]. Available: https://s3.us-west-1.amazonaws.com/chips.user.media/page_file/_Saptadeep_Pal.pdf

[ApexA] available online https://www.chips.ucla.edu/

[Sjan20s] S. Jangam and S. S. Iyer, “A Signaling Figure of Merit (s-FoM) for Advanced Packaging,” *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 10, no. 10, pp. 1758–1761, Oct. 2020, doi: 10.1109/TCPMT.2020.3022760.

[MILS883] MIL-SPEC 883 standard. [online] available: https://www.navsea.navy.mil/Portals/103/Documents/NSWC_Crane/SD-18/Test%20Methods/MILSTD883.pdf

[JEDECREL] JEDEC reliability standards. [online] available: https://www.jedec.org/document_search?search_api_views_fulltext=reliability

[RenVLSI] H. Ren *et al.*, “Power Delivery for Scaled-Out Chiplet-Based Wafer-Scale Systems with 8\mu\mathrmm Cu-Cu Bond Pitch on Active SI-Interconnect Fabric Substrate,” in *2025 Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, Jun. 2025, pp. 1–3. doi: 10.23919/VLSITechnologyandCir65189.2025.11074799.

Chapter 2 High throughput Cu-Cu thermal compression bonding for scale out packaging

2.1 Introduction to thermal compression bonding process

Thermal compression bonding utilizes a synergistic application of temperature and pressure for die-attach. Both Solder-based and metal-metal direct thermal compression bonding approaches have been extensively studied in industry and academia. Solder-based thermal compression bonding [EMD64, MDat03] has been the workhorse of die-attach until bump pitches down to 20 μm [Zli21]. However, going to smaller bonding pitches using solder bumps has proven to be expensive and challenging due to solder extrusion and solder bridging [Xli14] between adjacent solder bumps. Furthermore, as the volume of solder decreases per bump, they become prone to electromigration stress [TCHuang17], and intermetallic growth with thermal aging [BIN06] which affects the long-term reliability of solder interconnects.

Solderless metal-metal thermal compression bonding (TCB) is capable of scaling down bonding pitches to 5 μm for dielet-to-dielet (D2D) and dielet-to-substrate/wafer (D2W) heterogeneous integration, by eliminating solder from conventional TCB. A sample Cu-Cu TCB is represented in Fig. 2.1 (d), where bonding is done at 10μm pitch. It can be seen that the contact is between Cu pads on dielet and Cu pillars on substrate, and the dielectric is recessed to ensure metal-to-metal contact without any dielectric bonding. Metal-metal bonding also increases the reliability of interconnections as most pure metals have higher resistance to electromigration and so long as both the mating surfaces constitute the same metal, there is no IMC growth. Therefore, metal-metal thermal compression bonding has been extensively researched at UCLA-CHIPS and furthermore forms the backbone of the bonding done in this research.

2.2 Previous work on Au-Au and Cu-Cu Thermal Compression Bonding

Under the right conditions of temperature and pressure, interdiffusion and grain growth takes place across the mating surfaces to form a metallic bond. Previously Bajwa et al. [Bajwa], have demonstrated for the first time a gold-gold dielet-to-wafer thermal compression bonding process at 10μm pitch. They deposited a thin film coating of 200 nm gold on Cu pads and Cu pillars prior to bonding to demonstrate thermal compression bonding between Au-Au during assembly as shown in Fig. 2.1. Fig. 2.1(a) shows an assembly of 10 dielets bonded using Au-Au TCB at 10μm pillar to pillar pitch connected via daisy chains with reported specific contact resistance ~1 Ω.μm2 [AABaj17]. Fig 2.1(c) shows a bonded cross-section. The first Cu-Cu thermal compression bonding at dielet-to-wafer level was demonstrated at 10 μm bonding pitch by Jangam et al. [SJan19]. This required the development of a novel formic acid vapor delivery technique to remove copper oxide from the copper pads and pillars prior to bonding. Formic acid vapor effectively reduces CuO into copper formates which are then decomposed at a temperature of ~200ºC during bonding to get pristine Cu. A two-dielet Cu-Cu TCB assembly along with cross-section is shown in Fig. 2.2 (b) and Fig. 2.2 (d) respectively. A cross section of the modified bond-head for in-situ delivery of formic acid is presented in Fig. 2.2 (a). The chemical reactions involved are as follows and taken from [SJan19]:

$$\mathrm{CuO_{(s)}} + 2\mathrm{HCOOH_{(g)}} \longrightarrow \mathrm{Cu(HCOO)_{2(s)}} + \mathrm{H_2O_{(g)}} \quad (1)$$

$$\mathrm{Cu(HCOO)_{2(s)}} \longrightarrow \mathrm{Cu_{(s)}} + 2\mathrm{CO_{2(g)}} + \mathrm{H_{2(g)}} \quad (2)$$

Cu-Cu thermal compression bonding has also applied to functional assemblies. The first demonstration of inter-dielet communication using Cu-Cu TCB involved 2x2 dielet-to-substrate assembly at sub-10μm pitch. The design of the chip, also called universal digital signal processor (UDSP) was published at ISSCC 2022 [URat22] and was developed by Prof. Dejan Markovic's Group at UCLA in collaboration with UCLA-CHIPS. Active silicon dielets represent additional challenges compared to passive dielet assemblies. When

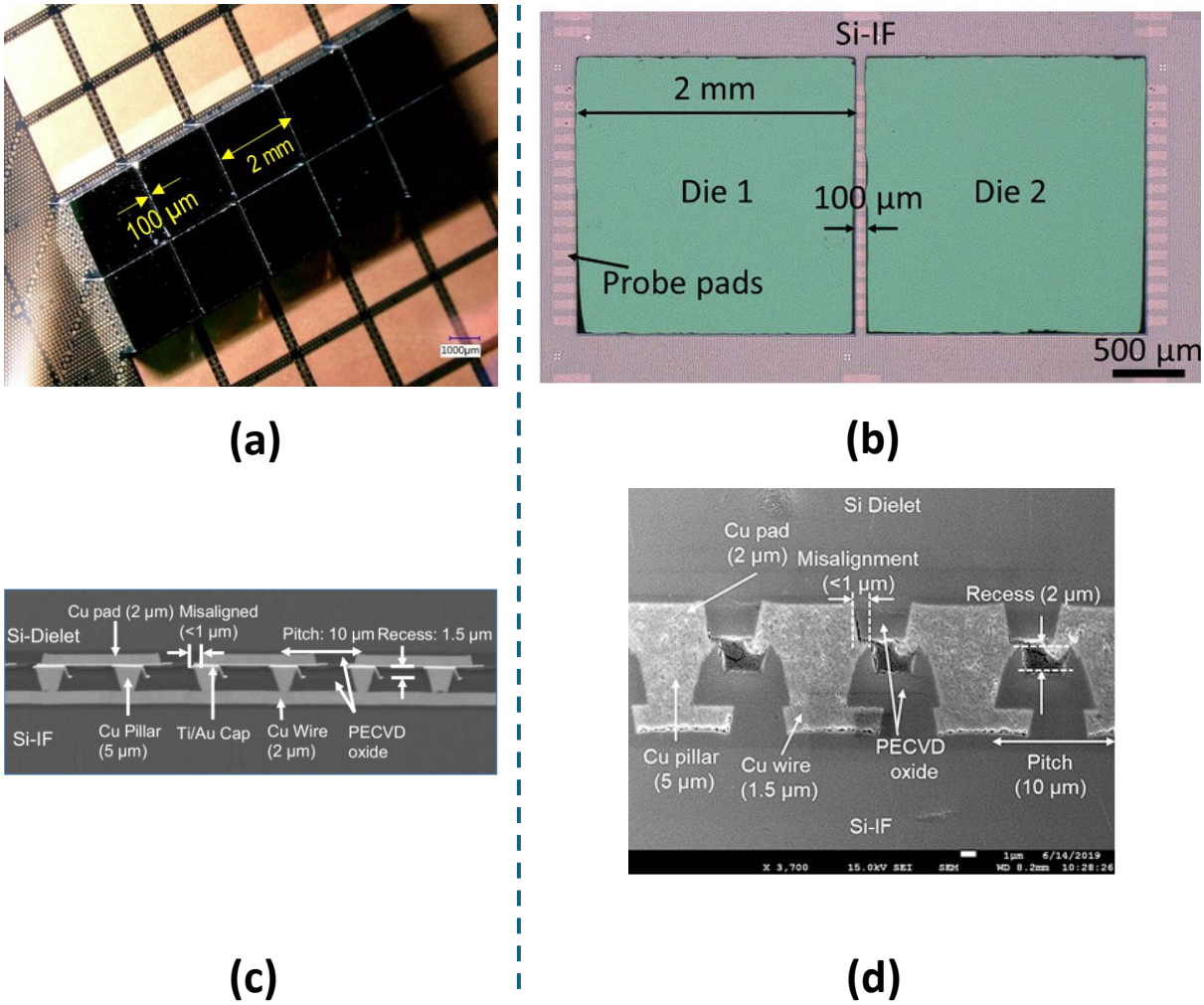


Figure 2.1 (a) Gold-Gold (Au-Au) thermal compression bonding (TCB) on silicon interconnect fabric (Si-IF) published in [AABaj17]. (b) Copper-copper (Cu-Cu) TCB on Si-IF published in ECTC [SJan19]. (c) Au-Au TCB cross-section with Au capped Cu pillars on Si-IF and Au capped Cu pads on dielet (d) Cu-Cu TCB cross-section showing good bonding interface and < 1µm misalignment.

working with active silicon dielets, an ionizer system must be enabled to ensure electro-static discharge (ESD) – safe environment during Cu-Cu TCB. This requires further optimizing both the ionizer flow and formic acid vapor flow to ensure the reducing gas environment necessary to ensure Cu-Cu bonding is unaffected by ionizer flow. The bonding force, bonding temperature and bond-head position along with ionized nitrogen

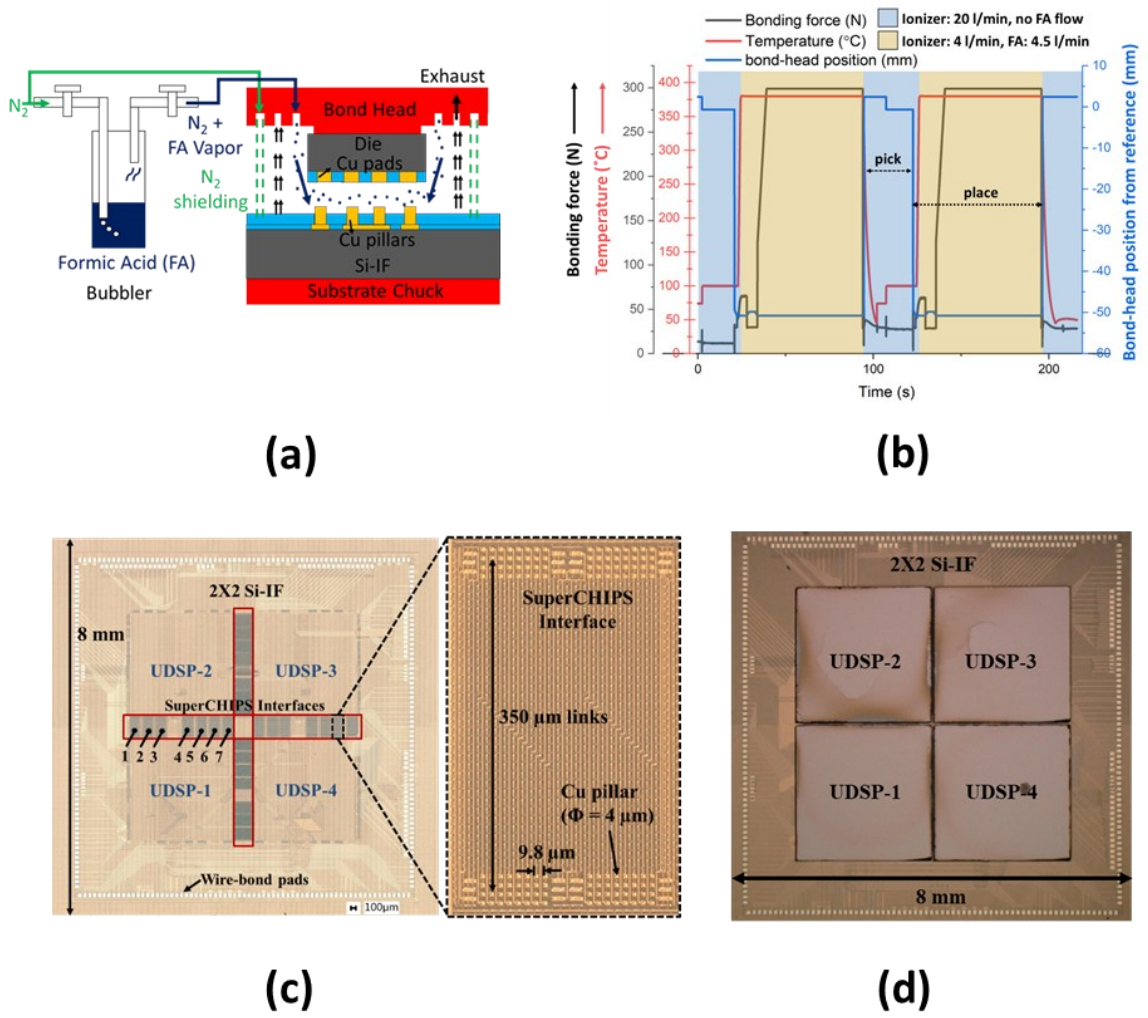


Figure 2.2 (a) schematic cross-section of die-to-wafer bonding setup with formic acid vapor delivery. (b) bonding curves for two consecutively bonded UDSP dielets (c) Si-IF for UDSP assembly showing 7 channels between two UDSP dielets and a fabricated SuperCHIPS channel (d) 2×2 UDSP assembly on Si-IF using Cu-Cu TCB.

(N2) gas flow rate and formic acid vapor flow rate is presented in Fig.1.2(b). This data was published at Journal of Solid Sate Circuits (JSSC) 2023 [SSNagi23] along with dielet communication results. The bonding process was published at the Electronic Components and Technology Conference (ECTC) 2022 [KSah22]. The 2×2 UDSP assembly uses an 8mm X 8mm silicon interconnect fabric (Si-IF) with bump pitch of 9.8µm and trace pitch of 4µm between dielets. The dielets communicate with 7 channels,

achieving an inter-dielet bandwidth of 297 Gbps/mm, shown in Fig. 2.2 (c). In fact, on Si-IF a much higher channel bandwidth up to 2.56 Tbps/mm is possible using high bandwidth links. This channel is referred to as SuperCHIPS, or Simple Universal Interface for CHIPS. The characterization of these channels will be further described in Chapter 5. Overall, there are 8000+ input/output and 22000+ power and ground connections in the assembly. [SJanD20]. Elsewhere researchers have also pursued Cu-Cu thermal compression bonding, albeit at higher bump pitches [CNLi24].

This research builds upon the Cu-Cu TCB process used in UDSP assembly by demonstrating higher throughput necessary for making Cu-Cu TCB competitive for manufacturing, especially in assembly for wafer-scale systems. We start this chapter by justifying the need for using thermal compression bonding in dielet-to-wafer assemblies, followed by proposing a novel two-step bonding approach with potential to increase Cu-Cu TCB throughput to greater than 1000 units per hour. To achieve this throughput, we analyze the Cu-Cu bonding steps and identify the parameters which need to be optimized. Parameter optimization is done through a robust design of experiments (DOE) methodology called Taguchi Method, a powerful tool for doing design and process optimization for quality, performance and cost. Experimental results are provided to confirm the effectiveness of this approach by doing a three-phase, two factor DOE with 20 experiments and 200 dielets. From this DOE, two approaches to increasing Cu-Cu throughput are developed, as described in section 2.7, and either one can be used depending on process/product requirements. Next a mechanism of two-step TCB is proposed. This is followed by electrical and mechanical results of the Cu-Cu bonds including extensive reliability data using various relevant JEDEC reliability tests. Finally, a few approaches to further increase throughput are discussed. The chapter is then concluded.

2.3 Justification for using thermal compression bonding for die-to-wafer assemblies

We strongly advocate that Cu-Cu die-to-wafer (D2W) thermal compression bonding is the ideal approach to demonstrating fine-pitch (≤ 10 μm bonding pitch) assemblies in high volume manufacturing (HVM) in spite of competitive die-to-wafer bonding approaches

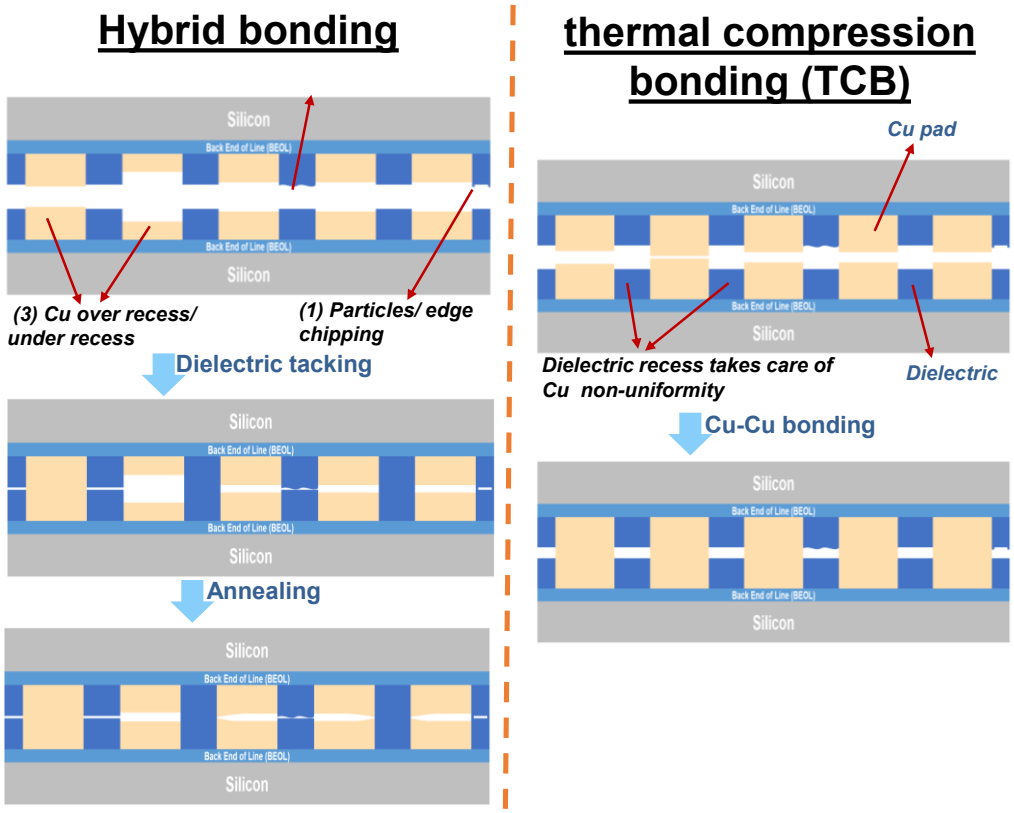


Figure 2.3 Schematic comparison between hybrid bonding (HB) and thermal compression bonding (TCB)

such as Cu-dielectric hybrid bonding, which also have been extensively studied by industry and academia [AElsh21, JJOng22, TWork21]. In previous work done by Ren *et al.* [HRen22], it was discussed that die-to-wafer (D2W) hybrid bonding (HB) offers the potential benefit of higher throughput and room temperature die tacking (placement), but at the cost of lower process tolerance and stringent processing requirements as shown in Fig. 1. These requirements involve: (1) extremely flat (roughness ≤ 1 nm) dielectric surfaces for strong attraction forces during initial dielectric attach, (2) meticulous control of metal recess to ensure successful stress-free metal-to-metal connections during annealing, and (3) complete particle control of bonding environment to avoid failed connections. A previous work by Intel [AElsh21] on Cu/dielectric hybrid bonding, these challenges of hybrid bonding were highlighted. Particle size plays a critical role in hybrid bonding as indicated as the number of failed bonded contacts increases as particle size

increases. Plasma dicing [YLin24] is critical to ensure the lowest particle count for hybrid bonding.

However, a D2W-TCB process is significantly simpler. In TCB, bondable pads/pillars are recessed to achieve metal-to-metal contact. Since there is no dielectric bonding, dielectric roughness requirements are not critical. Surface asperities on bonding pads are flattened by temperature and pressure applied during thermal compression bonding. D2W-TCB is generally independent of the type of dicing used, so blade dicing is applicable. However, a cleaner dicing process like plasma dicing should be preferred to reduce pre-treatment requirements of the dielet and substrate samples prior to bonding. In case of samples diced using blade dicing, ultrasonication based wet treatment in acetone, isopropanol followed by deionized water rinse can be used to remove any particles form the sample prior to bonding. One of the challenges is the air gap between dielet and substrate in Cu-Cu bonding, which can lead to corrosion of exposed Cu surfaces overtime leading to reliability issues. This is overcome by using passivation techniques developed by Sakoorzadeh et al. [NShak20], which will be elaborated upon in section 2.8. The Fig 2.3 shows the schematic comparison of a hybrid bonding and thermal compression bonding process. The key differences between D2W thermal compression bonding and D2W hybrid bonding can be summarized in the table 2.1 below, taken from a previous work [KSah24]:

Table 2.1 Comparison of solderless thermal compression bonding (TCB) and Copper/dielectric hybrid bonding (HB) approaches

	Solderless thermal compression bonding	**Hybrid Bonding**
D2D and D2W Bonding pitch	Up-to 10µm demonstrated [SJan19], Up-to 5µm possible.	Up-to 2µm demonstrated [TWork21] in D2W process, less than 2µm possible.
Bump Requirements	Only metal-metal (metal = Cu, Au, Al) bonding and no dielectric bonding. CMP requirements	Necessitates meticulous control over dielectric flatness (6σ roughness ≤ 1 nm) and

	(~10nm roughness and ~50nm dishing for Cu-Cu TCB [SJan21, KSah23])	metal recess. Requires extensive CMP optimization
Bonding mechanism	Dielet tacking under simultaneous application of temperature and pressure in a suitable bonding environment. May be followed by annealing.	Room temperature hydrophilic bond for dielectric attach followed by annealing for Cu-Cu connections.
Bonding environment	ISO 6 (Class 1000) and above	ISO 3 (Class 1) and below
Die sizes allowed	Limited by maximum bond-head pressure during tacking.	Almost any size due to less tacking pressure requirements.
Die dicing	Dielectric is recessed so, blade dicing with standard wet cleaning is feasible.	Mandatory particle-free dicing. Particles generated by blade dicing at the dielet edge cannot be tolerated.
Passivation post assembly	Atomic layer deposition for passivation [NShak20].	Hermetically sealed by dielectric bonding [AElsh21].
Throughput	1100+ UPH with optimized tack and anneal process [KSah23].	2000+ UPH due to fast dielectric bonding during tacking phase [Besi].
Key challenges	High tacking temperature limits alignment accuracy. Bond-head pressure capability put an upper limit on dielet size.	No tolerance to particles during bonding. If metal recess requirements are not met, bonding can fail.

2.4 Need for high-throughput in Cu-Cu TCB

In Cu-Cu TCB, under the simultaneous application of temperature and pressure at the bonding interface, grain growth takes place between the two Cu mating interfaces. The

kinetics of this growth is further explained in section 2.9. In case of single step TCB, where bonding needs to happen in a single Cu-Cu contact step, the throughput is limited by the time taken for grain boundaries to merge and grow. The time required can be in order of tens of seconds. In the dissertation work done by Siva Jangam [SJanD20], a bonding time of > 30 seconds per dielet in an in-situ formic acid environment was reported to achieve a die-shear strength better than the military standard (MIL-SPEC 883 [MILSPEC883]) shear strength for a given dielet size. This results in a bonding throughput of approximately 100 units per hour (UPH), including time delays due to die pick and transfer from dicing tape or waffle tray to bond-head in a die-to-wafer bonding equipment, as well as die-alignment.

Let us consider a full 300 mm wafer scale assembly to be populated with dielets. Assuming 35 seconds of total bonding time per dielet, the amount of time required to bond 2000 dielets is about 20 hours of bonding time. A variation of total assembly time versus number of dielets is shown in Fig. 2.4. The dielet sizes are chosen based on recommendation of dielet sizes in the National Advanced Packaging Manufacturing Program (NAPMP) [NOFO2]. Throughout this bonding time, the Cu terminated pillars on the Silicon Interconnect Fabric (Si-IF) are exposed to high chuck temperatures of 100^{o}C and above in an ambient air or nitrogen-rich bonding environment. As expected, the oxygen present in the bonding chamber will cause continuous oxidation of exposed Cu into copper oxides. While a localized in-situ cleaning technique in the form of formic acid vapor cleaning described in [SJan19] is always present in-situ in the bonding chamber, such a cleaning technique would not be able to remove several hundred nanometers of copper oxide layers [AABaj17] formed due to 20 hours of exposure to high chuck temperatures. To use single step TCB, one needs to periodically clean the wafer during bonding to remove excess copper oxides. This cleaning could be done via argon plasma treatment or acetic acid treatment [AABaj17]. Such a cleaning process requires pausing the bonding process, taking the wafer from the bonding chamber and then doing wet or dry etching process to remove excess copper oxides, putting the wafer back in the bonding chamber, recalibrating the bonding process and continuing from the site where bonding was paused. As expected, this is a time-consuming process with a high risk of

failure. It is unreliable to scale-up Cu-Cu bonding to achieve 12-inch (300mm) wafer-scale heterogeneous integration.

Thus, there is an urgent need for a higher-throughput bonding process which would reduce the amount of time required to fully populated up to a 300 mm wafer, making wafer-scale integration feasible. At a 1000 UPH, the process of populating 2000 dielets on a 300 mm wafer (active bonded area: 66000 mm^2) would only require about 2 – 3 hours. In this case, local in-situ formic acid cleaning is shown to be adequate [KSah21] for keeping copper oxidation under control during assembly. One of the key contributions

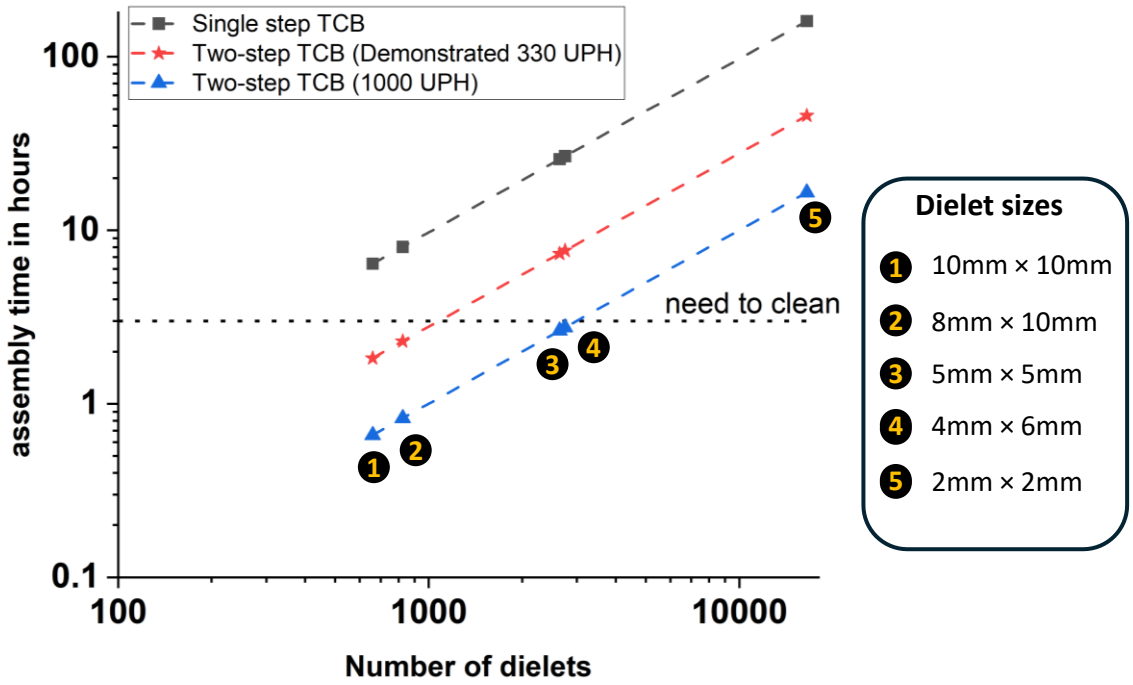


Figure 2.4 assembly time (in hours) versus number of dielets for a 300 mm fully populated wafer (active bonded area: 66000 mm^2) when compared at different bonding throughputs (as shown in legend). The black plot shows single step TCB. The two-step TCB shown by bule and red plots will be discussed from section 2.5. The wafer must be cleaned either by dry or wet cleaning approaches if bonding time exceeds 3 hours. The dielet sizes for this plot are chosen with recommendation of NAPMP [nofo2].

of this dissertation is the development of a high throughput Cu-Cu thermal compression bonding process to address this need. In Fig. 1.4 we also show the assembly time improvement using a 330 UPH and 1000 UPH high throughput Cu-Cu TCB process which can greatly reduce dependence on multiple wafer-level cleaning steps as required by single step TCB. The methodology of high throughput Cu-Cu TCB process is described in the next section.

2.5 High-throughput Cu-Cu TCB using two-step approach

The reliance on a single long duration die bonding step (≥ 30 s/die in our earlier approach) for both mechanical strength and electrical connectivity is the prime challenge which needs to be addressed to achieve manufacturable TCB throughput. To overcome this limitation, we propose a two-stage approach: die tacking to Si-IF or Si-interposer substrate, followed by an annealing of the fully populated assembly. The two steps can be summarized as follows:

Dielet-to-substrate tacking: During the die tacking stage, all the dies are aligned at relatively low temperature of 120 °C and placed within a total time of ≤ 10 seconds/dielet. This step does not ensure final bonding, but it does guarantee a firm-enough attach with a shear strength > 10 N. As expected, this is below the MIL-SPEC shear strength value for a given dielet size. The goal is to ensure no die/dielet movement as the sample is prepared for the next assembly step.

Fully populated annealing (gang-anneal): Once populated, the assembly is batch annealed (batch size depends on furnace capacity) in vacuum for 1 hour. This step ensures Cu grain growth across the mating surfaces needed for successful bonding. By batch annealing several wafers at a time, annealing time is no longer an issue for throughput, and further optimization is still possible at the tacking stage. For the work demonstrated in this dissertation, we can get an aggregate bonding time of ~10 seconds/die, thereby giving a throughput of > 320 units per hour (UPH).

To increase the throughput to 1000+ units per hour, the tacking time needs to be further reduced to 6 seconds/dielet, which can be achieved by optimizing the tacking parameters

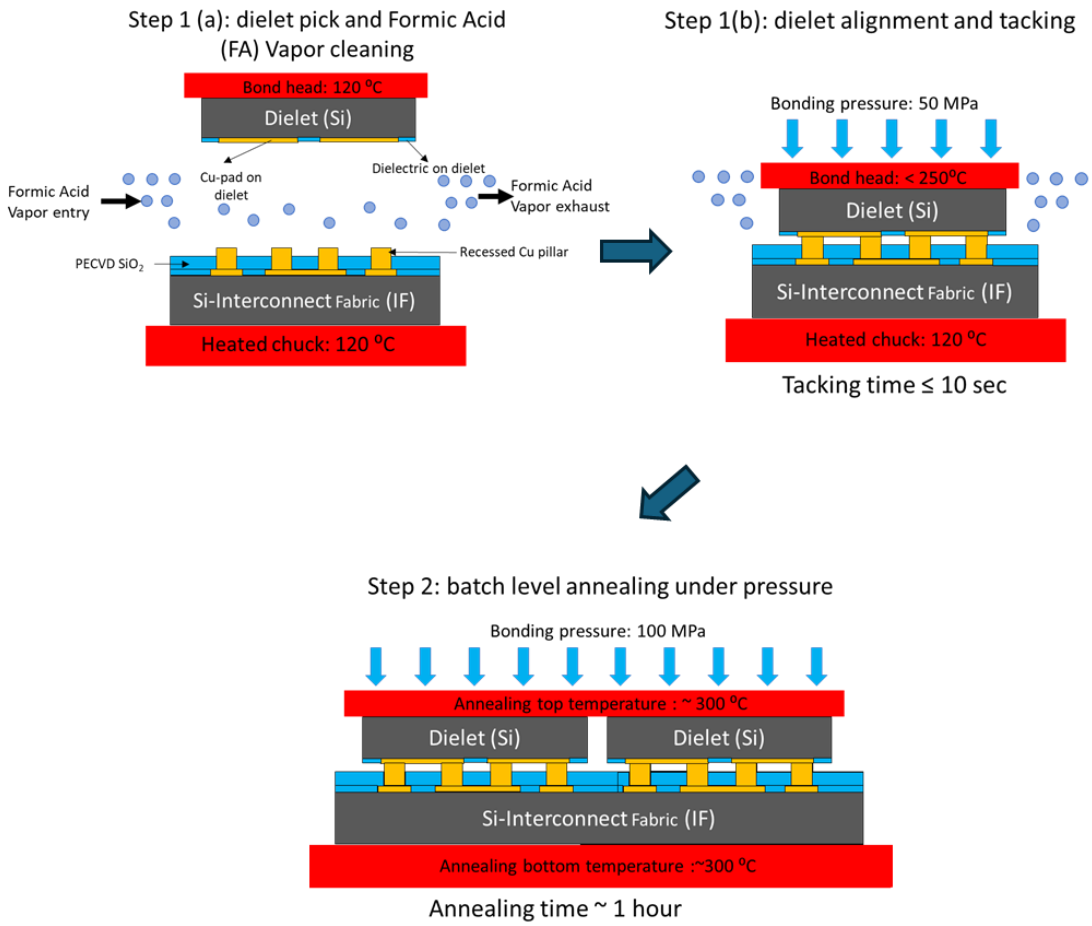


Figure 2.5 Two-step Cu-Cu TCB process. The first step involves dielet transfer to bond-head, dielet cleaning, alignment and tacking. The second step involves batch anneal of the fully populated assembly.

and reducing the dielet-cleaning and dielet-alignment as described in the detailed design of experiments done in this dissertation work.

The two steps are described in detail in Fig. 2.5. The Fig.2.5 step 1 shows the tacking step, followed by step 2 which shows the batch annealing step.

2.5.1 Optimization for increasing throughput for Cu-Cu TCB using Taguchi's Method

Taguchi method was developed by Dr. Genichi Taguchi [NGho17] during 1950s. Taguchi design of experiments provides an efficient and systematic to design and optimize manufacturing processes at a reduce cost. It is based on orthogonal array experiments,

and emphasizes balanced design with equal weightage to all influencing factors with less number of experimental runs. Taguchi has defined a statistical measure of performance called signal to noise ratio (S/N) which is used to evaluate the significance and sensitivity of a process parameter to optimize any manufacturing process [NGho17]. The S/N ratio takes both mean and variability into account. It is defined as the ratio of mean (signal) to the standard deviation (noise). The ratio depends on the characteristics and objective of the process to be optimized. Depending on the objective of process optimization, there are three commonly used S/N ratios in Taguchi design of experiments (DOE):

Higher the better – Typically used when the goal of the optimization is to increase the performance of a specific parameter. The S/N ratio for higher the better is defined as:

$$\eta = -10 \log_{10} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{y_i^2} \qquad (1)$$

Where:

y_i=response variable (e.g. shear force) of n^{th} experiment

n = number of similar experiments

η = signal to noise ratio

Lower the better – Typically used when the goal of optimization is to reduce the cost of a process, when reducing the value of a parameter is desirable. The S/N ratio in case of lower is better is defined as:

$$\eta = -10 \log_{10} \frac{1}{n} \sum_{i=1}^{n} y_i^2 \qquad (2)$$

Nominal the best – Typically used when the goal of optimization is to achieve a particular target. This is most helpful in process optimization when considering both advantages and trade-offs of changing different process parameter to ensure the process or product meets certain target specifications. The S/N ratio for nominal the best is given by:

$$\eta = 10 \log_{10} \frac{\bar{y}^2}{\sigma^2} \qquad (3)$$

Where:

$\bar{y}$= average response (e.g. average shear force) variable

σ= variance of the experimental data., σ^2 is the standard deviation.

For two-step TCB, we use nominal-the-best condition for process optimization because the throughput is being optimized while ensuring a target shear force of the die. In our case, we want to optimize for a minimum shear force of 2x MIL-SPEC to a maximum shear force of 4x MIL-SPEC [MILSPEC883] for the given die size. The other key advantage of using nominal-the-best condition is it gives process knowledge of a given manufacturing process by highlighting the parameters that the process or product is most sensitive to. Therefore, in the design of experiments for Cu-Cu throughput improvement and final analysis, nominal-the-best S/N ratio is used for optimization.

A Design of Experiment (DOE) using more than 200 dies of size 2×2 mm^2, bonded to multiple Si-IF substrates at 10 µm Cu bump/pillar pitch was performed to improve throughput. The goal of the DOE was to reduce the overall die tacking time to ≤ 10 seconds per die and achieve a die shear strength value greater than 2X specified by MIL-STD 883G [MILSPEC883] for the corresponding die size after annealing. For a 2x2 mm^2 die, shear force must exceed 50 N. To minimize the overall bonding time while retaining the required bond strength, all the parameters which affect bond integrity were identified. These parameters include – 1) formic acid vapor dielet cleaning time 2) die placement time 3) bond-chuck temperature 4) tacking pressure 5) annealing temperature 6) annealing pressure. These parameters and the values used for optimization are listed in Table 2.2.

To reduce the number of experiments, but gain useful knowledge of how each parameter affects thermal compression bonding, the DOE was performed in three phases, with two parameters being optimized per phase. The Fig. 2.6 (a) shows the flow-diagram of parameter optimization including the parameters optimized per phase. At the same time in each phase, the parameters which were not tested for were set to the largest value from Table 2.2, unless already optimized.

Table 2.2 Design of experiment parameters for two-step Cu-Cu TCB.

Phase	Parameter name	Value	Unit
1	Die cleaning time	3,5	s
1	Die placement time	3,5	s
2	Chuck temperature	80, 100, 120	°C
2	Tacking pressure	50, 150, 300	MPa
3	Annealing temperature	200, 300, 400	°C
3	Annealing pressure	20, 50, 100	MPa

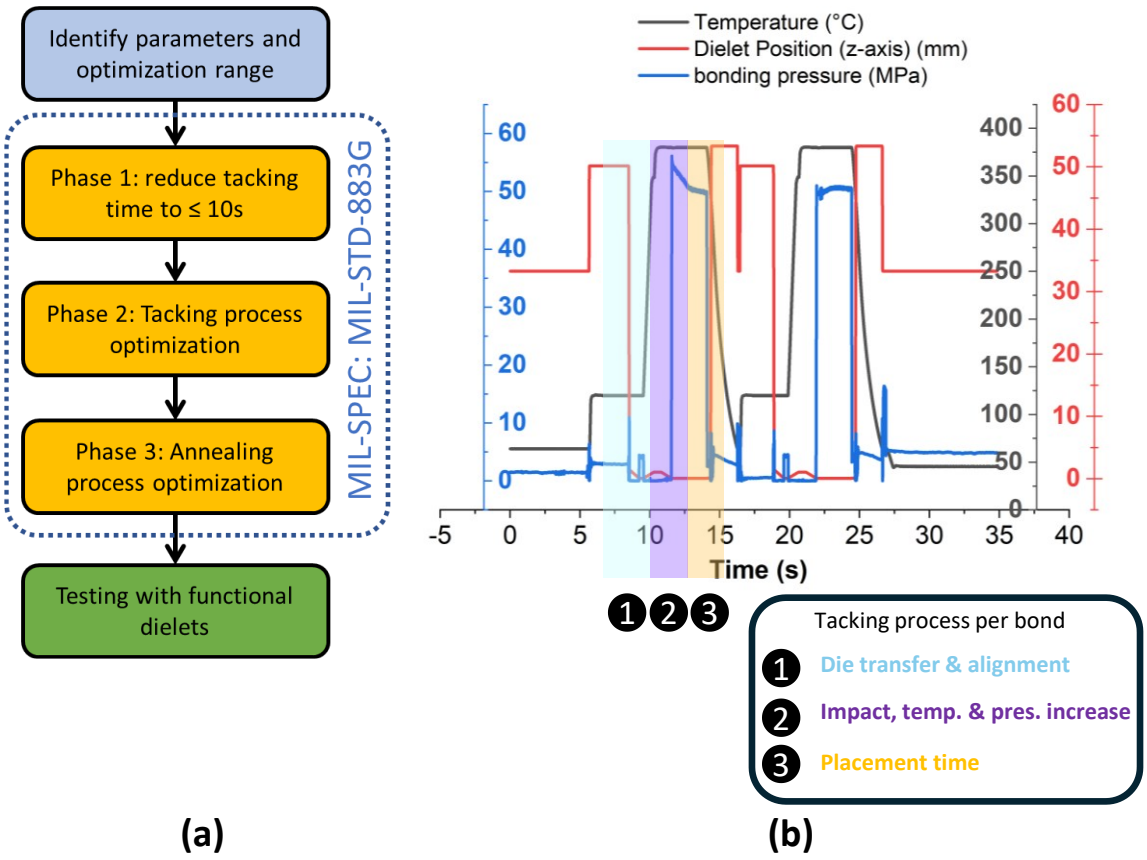


Figure 2.6 (a) Two-step thermal compression bonding design of experiments (DOE) showing optimization being done in three phases with two parameters optimized per phase. (b) The temperature, vertical dielet position and bonding pressure profiles of tacking phase of bonding. The values of temperature and pressure are changed per experiment.

2.5.2 Equipment for dielet to substrate bonding

All the bonding is done using, the chip-to-wafer bonder from Kulicke and Soffa also referred to as APAMA [APAMAC2W]. The dielets to be bonded are placed on dicing tape or a waffle tray, and a picker tool picks up the dielets to transfer them to the bond-head. The schematic of the chip-to-wafer bonder APAMA is shown in Fig. 2.7, including the picker tool and bond-head. The picker is used to pick dies from the dicing tape or waffle tray and transfer them to bond-head placer tool. The placer tool is capable of applying

pressure and temperature to bond the transferred dielet to the site to which it needs to be bonded. On APAMA, the dielet-to-wafer alignment is done using a machine vision system, where a camera comes between the dielet and the substrate prior to bonding. The dielet and substrate site to which it is to be bonded have pre-defined fiducials. The camera vision system comes between the dielet and substrate prior to bonding and determines the location of the fiducials with respect to the camera coordinate system based on prior dielet and substrate teaching steps. Once the relative positions of dielet fiducials and substrate fiducials are determined, the camera vision system retracts, and the dielet moves to make contact with the substrate. This is schematically shown in Fig. 2.2 (a). Then temperature and pressure can be applied for bonding. In summary, the vision system used on K&S APAMA results in a second-order blind alignment between dielet and substrate.

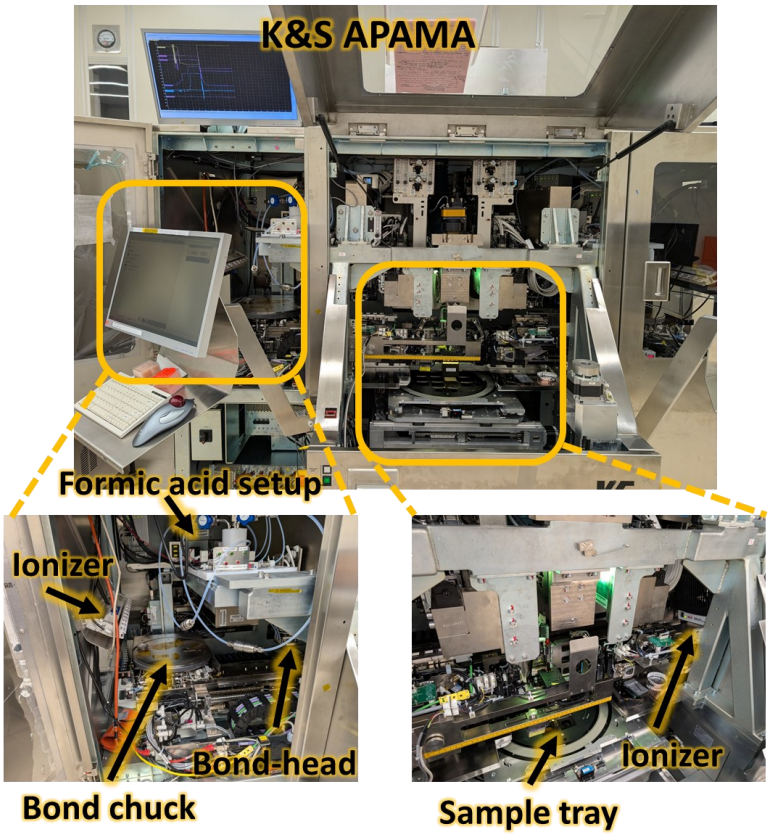


Figure 2.7 Dielet-to-wafer bonding tool APAMA by Kulicke & Soffa [APAMAC2W] which is used for all the bonding experiments.

2.5.3 Details of the samples used for bonding

As described, the dielets to be bonded are of the size 2x2 mm2. These dielets are bonded to Si-IF substrate with 10 µm Cu bump/pillar pitch. An optical microscope (OM) image of the dielet along with a close-up view of the Cu pads is shown in Fig. 2.8 (a), including the Si-IF substrate with copper pillars. The Cu-pillars have a diameter of 5 µm and a pitch of 10 µm. Prior to bonding any dielets, they are cleaned using Acetone + Isopropanol rinse for 1 minute followed by deionized water rinse for 30 seconds. This is to remove any particles sticking to the samples. Then the samples are pre-treated with acetic acid for 10 seconds for each dielet and substrate to remove any excess oxide on the Cu pads. This part of the cleaning process can also be replaced with argon (Ar) plasma treatment for 30 seconds as described in [AABaj17]. Finally, for each split in a phase i.e. for a given set of variables in each phase 10 dielets are bonded.

The number of experiments and number of dielets bonded are as follows:

Table 2.3 Three phase optimization with experiments performed.

Phase number	Total experiments	Total Dielets bonded
1	2	20
2	9	90
3	9	90
Total	20	200

Therefore, a total of 20 splits were performed, and a total of 200 dielets of 2x2 mm^2 were used in the experiment. Each Si-IF had 4 dielets so each split had 2 Si-IFs which were then sheared to determine the shear strength. The determination of shear strength is done using shear tool DAGE 4000 plus [NDAGE] from Nordson DAGE. This equipment has a cartridge equipped to do lateral shear test as shown in Fig. 2.9. The sample is placed on a clamp as shown in Fig. 2.9 (a) and then the chisel is brought before the dielet The die shear was then executed.

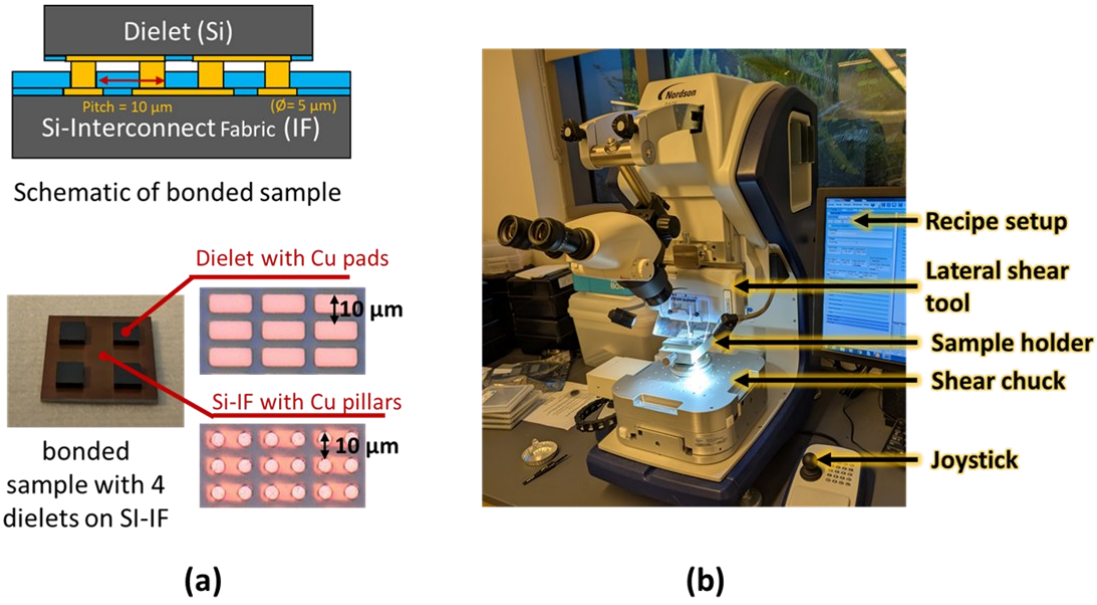


Figure 2.8 (a) Close up of Cu pads and Cu pillars on Si-IF prior to bonding and post bonded sample (b) Nordson DAGE 4000 Plus die-shear tester tool.

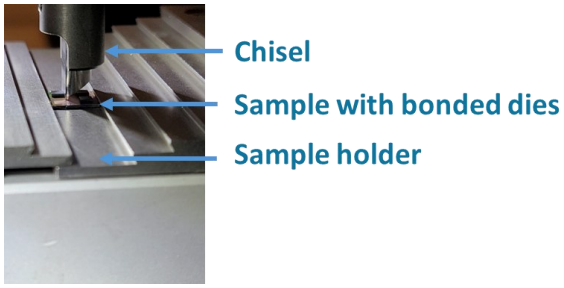
The parameters of the shear were as follows:

Table 2.4 Shear parameters used in Nordson DAGE tool to characterize dielet shear strength of two-step Cu-Cu TCB.

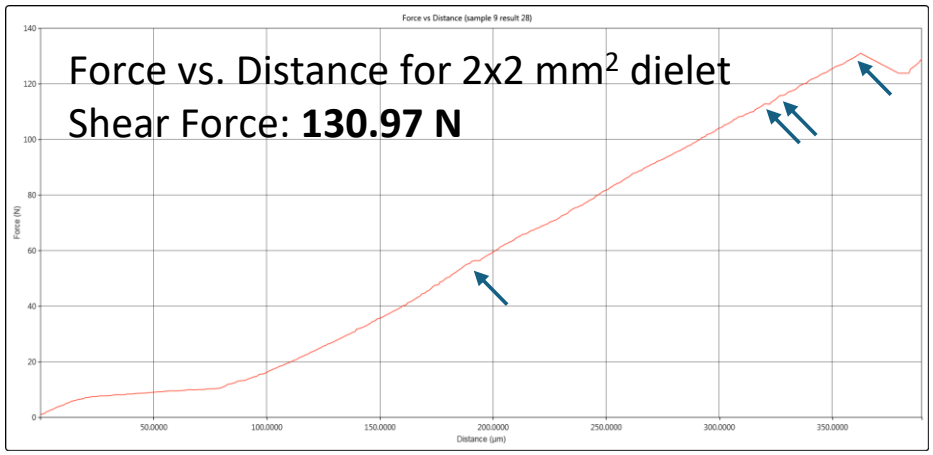
Parameter	Value	Unit
Shear test range	10 – 300	N
Chisel height from substrate surface	11	µm
Initial shear force	2	N
Shear speed	15	µm/s
Maximum distance travelled after shear	0.5	mm

Die shear is considered to be successful if the die is completely ripped from the bonded location or substrate breaks down under the large shear force due to creation of cracks and imperfections. If the die shear was successful, then the result is shown as PASSED. This value is then used for further analysis, such as in tabulation and determination of average shear force and variability in shear force. An sample on the sample holder stage

of DAGE tool is shown in Fig. 2.9 (a). Also, an example of a successful shear force vs distance plot is presented in Fig 2.9 (b). It is important to note that shear is not an instant breakdown of contact between dielet and substrate. Multiple kinks can be seen in the shear plot which indicate the shear happens in steps as few rows of Cu pad to pillar contacts gradually give up when the shear force is increased. When several rows of Cu pad to pillar contacts have failed, the dielet gets detached. However, there are instances of dielet bonded so strongly to the substrate that either the point of delamination is the Titanium seed layer used to attach the pillar to the substrate, or the substrate itself may break.



(a)



(b)

Figure 2.9 (a) Bonded die to be sheared, placed on sample stage of Nordson DAGE tool. (b) A shear curve for a 2mm X 2mm die with shear failure at 130.97 N (> 2x MIL-SPEC 883). The arrows represent kinks where multiple rows of pillars fail. When enough rows fail, the die gets detached from the substrate.

2.6 Experimental results

In this section, the experimental details of the three-phase optimization are discussed.

2.6.1. Phase 1: Dielet tacking time optimization

In phase 1, die tacking time was optimized. The total tacking time is the sum of die cleaning time, die to substrate alignment time and die placement time. The dielet to substrate alignment time is a constant of 2 seconds (s) which depends on the bonding tool used. We use formic acid vapor (FA) to perform an in-situ cleaning of the die pads right before bonding. This represents the die cleaning time. As seen in Fig. 2.10, two cases were tested (first two columns), and both satisfied the 10 s/die tacking constraint, but the one with 3 s cleaning time and 5 s die placement time yielded better results. A sample bonding curve plot from APAMA highlighting different zones during bonding can

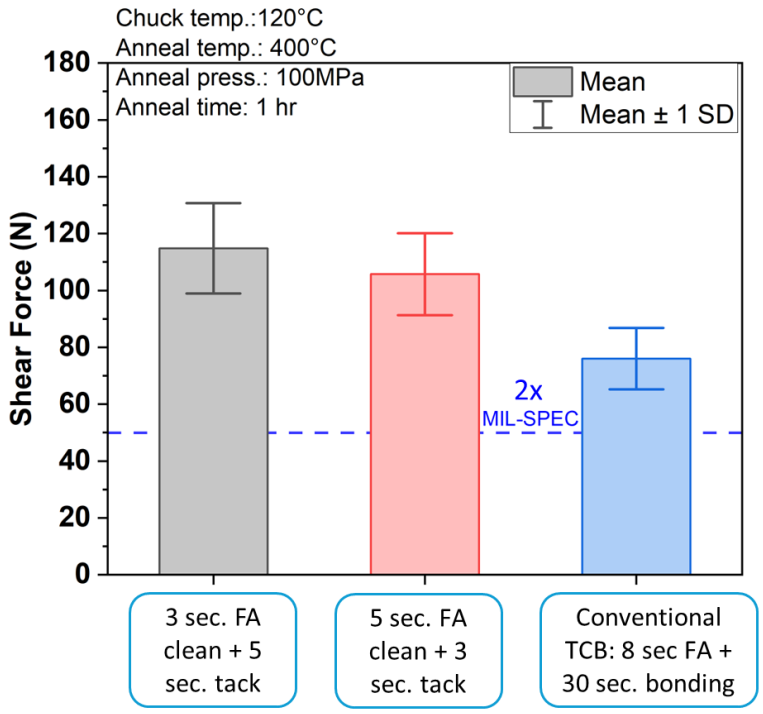


Figure 2.10 Formic acid vapor (FA) cleaning and die placement optimization for phase 1 of two stage direct Cu-Cu TCB process. Two-stage TCB results are compared to conventional > 30s/die TCB and the former shows superior shear strength as compared to conventional TCB process

be seen in Fig. 2.6 (b). Within the same figure, bonding force and tacking temperature at the bond-head is also represented. The total dielet tacking occurs in a time period of ten seconds. This time includes dielet transfer, dielet alignment as well as dielet to substrate placement time.

2.6.2. Phase 2: Dielet tacking condition optimization

Phase 2 addressed the substrate chuck temperature and tacking pressures optimizations. The results are summarized in Fig. 2.11. Variability of shear strength decreased with increase of chuck temperature. This can be explained by considering the fact that in-situ formic acid (FA) vapor reaction is more effective at higher temperatures. Hence for a chuck temperature of 120 °C, the copper oxide reduction to Cu is most effective, and hence the bonding quality is higher. Increasing temperature further is not recommended as it affects alignment accuracy. As we know, the dielet to wafer alignment is done using

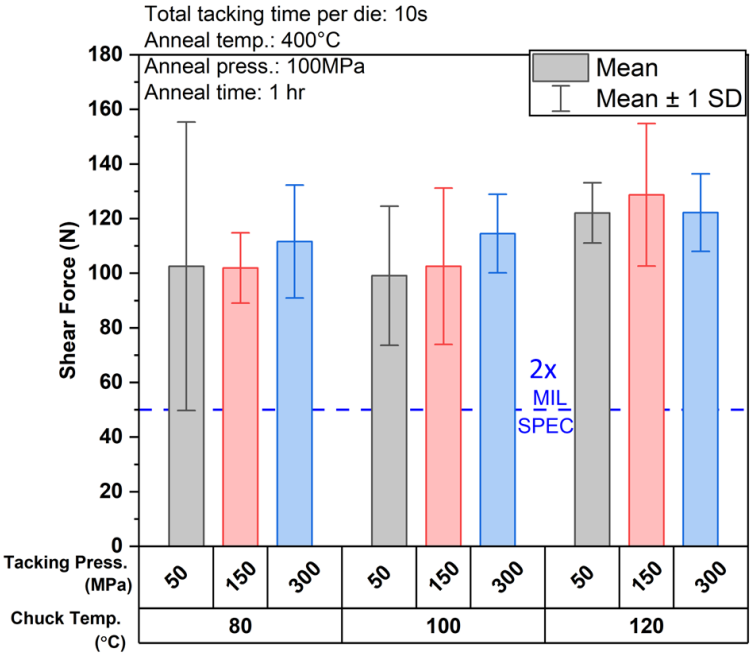


Figure 2.11 Tacking pressure and chuck temperature optimization results from phase 2. Least variability in shear strength occurs at chuck temperature of 120 ℃. Shear strength variability is high after annealing only if both chuck temperature and tacking pressure are low.

a camera vision system, the dielet alignment accuracy depends on refractive index of the air around the camera, dielet and substrate. Use of higher chuck temperatures will result in a larger thermal gradient near the camera, ultimately hampering alignment correction of dielet to substrate. A chuck temperature of 120 °C is found to be ideal. The bonding process is found to be insensitive to tacking pressures, except at low chuck temperatures. At the end of phase 2, a chuck temperature of 120 °C and tacking pressure of 50 MPa was finalized.

2.6.3. Phase 3: Batch annealing parameter optimization

Annealing temperature and pressure were optimized in phase 3. Annealing temperature was found to have the most profound impact on bonding yields. As seen from Fig. 2.12, going from an anneal temperature of 400 °C to 300 °C, the bonding quality generally degrades, whereas bonding catastrophically fails at 200 °C. Pressure has a smaller impact on shear strength for an anneal temperature > 200 °C. The degradation of bonding

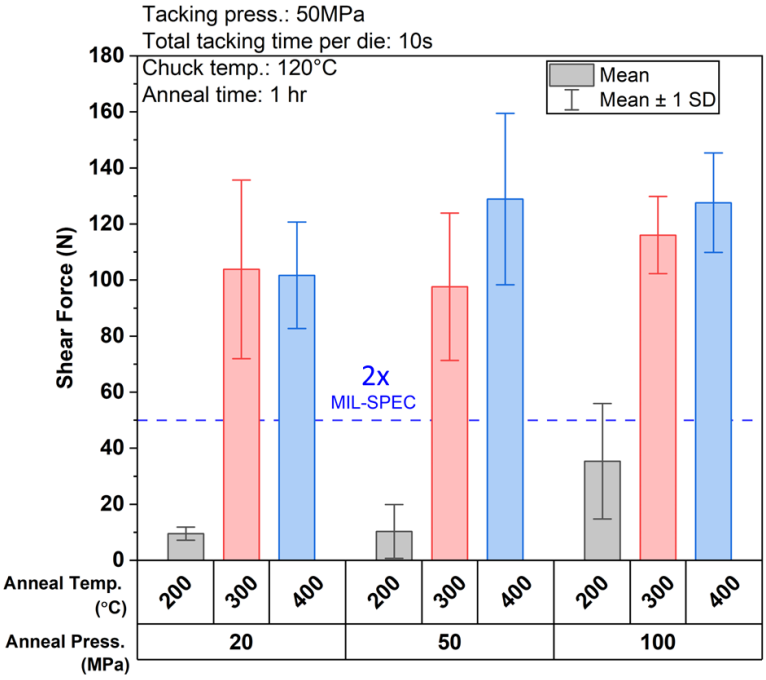


Figure 2.12 Anneal temperature and anneal pressure optimization. Cluster direct bonding is found to fail at 200 °C anneal temperatures which indicates that there is insufficient Cu diffusion across bonding interface.

shear strength with reduction of anneal temperature can be directly linked to grain growth kinetics which will be discussed extensively in section 2.8.

2.7 Optimized die-to-wafer Cu-Cu high throughput TCB process

2.7.1 Optimized Cu-Cu TCB process

Table 2.5 shows two possible variants for reliable two stage D2W TCB after optimization — a low temperature (LT) variant with an anneal temperature of 300 °C but anneal pressure of 100 MPa, and a high temperature (HT) variant with anneal temperature of 400 °C and anneal pressure of 50 MPa. In either variant, the average die shear strength of 2x2 mm^2 dies is found to be > 110 N which is ≥ 2x the military specification (50 N). Depending on the thermal budget requirements of the bonding process and types of dielets used, one of the bonding variants can be chosen.

Table 2.5 Optimized two step Cu-Cu TCB parameters

Parameter name	Low Temp. (LT) anneal	High Temp. (HT) anneal
Die cleaning time (s)	3	3
Die placement time (s)	5	5
Chuck temperature (°C)	120	120
Tacking pressure (MPa)	50	50
Annealing temp. (°C)	300	400
Annealing press. (MPa)	100	50

Several samples were prepared both using high temperature (HT) and low temperature (LT) variant for electrical measurements and further reliability analysis of the bonds. An example of such a sample is shown in Fig. 2.17, which was also published in [KSah23]. In Fig. 2.13, we show the focused ion beam (FIB) cross-section of the Cu-Cu bonds at 300°C and 400°C respectively. At 400°C image shows no voids at the bonding interface, and in fact, we can see grain growth across the bonding interface which represents a

strong metallurgical bond. It should be noted that the sample in the cross-section is further passivated using 10 nm atomic layer deposited alumina (Al_2O_3). As the Al_2O_3 when compared to dimension of Cu pads/pillar and SiO_2 dielectric is very thin, it cannot be identified in the cross-section image.

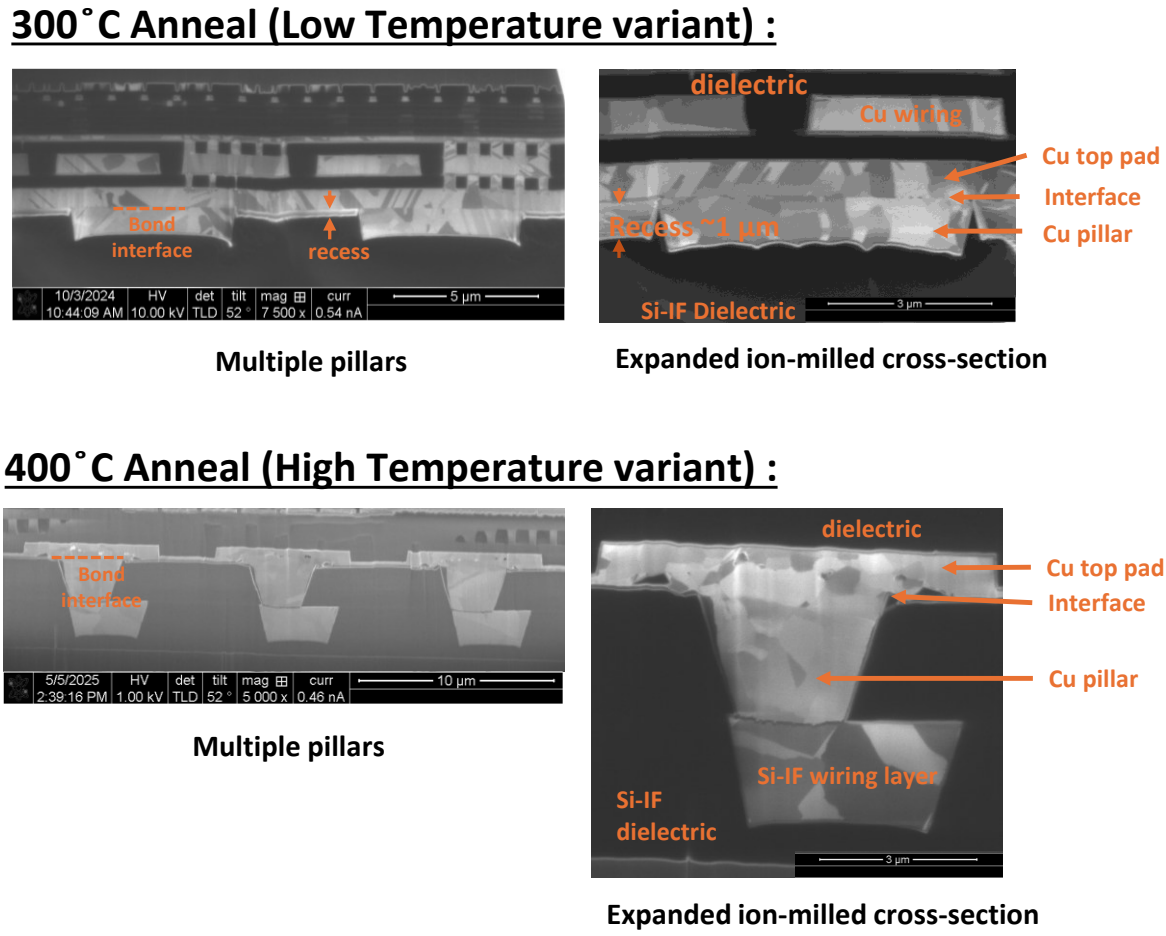


Figure 2.13 Scanning electron micrograph cross-section of two-step Cu-Cu thermal compression bonding (TCB) at 300°C and 400°C. At 400°C anneal temperature, we can observe grain growth across the bonding interface. No visible interface is observed showing strong metallurgical bond.

2.7.2 Parameter sensitivity analysis using Taguchi's S/N

After obtaining all the data required, it can be further processed using Taguchi's analysis methods to determine the parameter sensitivity. From section 2.5.1, a nominal-the-best S/N analysis is used for parameter optimization and for determining parameter sensitivity to bonding. The signal to noise ratio for nominal-the-best approach to optimization from equation (3) is:

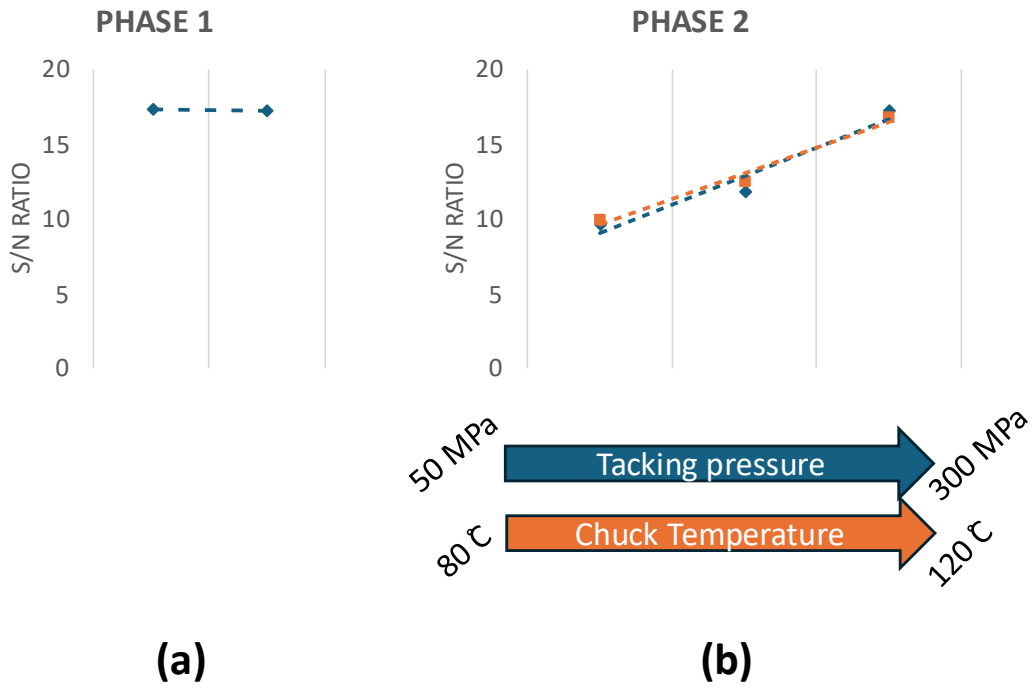
$$\eta = 10 \log_{10} \frac{\bar{y}^2}{\sigma^2} \qquad \text{from (3)}$$

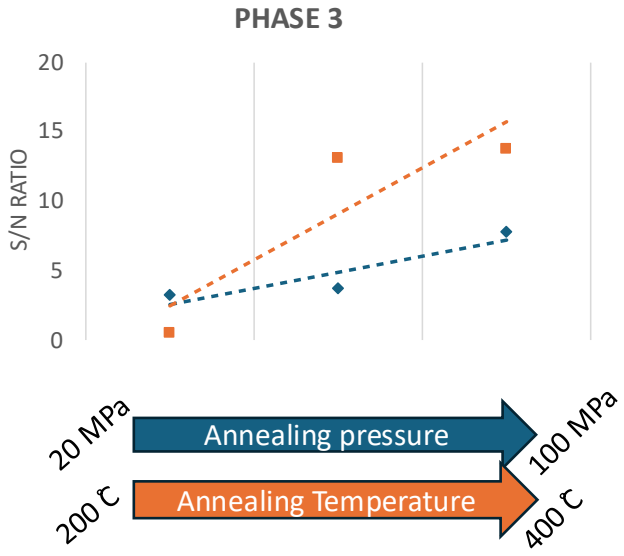
Where:

$\bar{y}$= average response i.e. mean of response variable, in this case average shear force.

σ= variance of the experimental data., σ^2 is the standard deviation.

Figure 2.14 (a), (b) and (c) represent parameter sensitivity at the end of each phase.





(c)

Figure 2.14 Taguchi sensitivity for two-step Cu-Cu TCB process for (a) phase 1 (b) phase 2 and (c) phase 3. In phase 2 change of tacking temperature and tacking pressure has almost similar sensitivity on bonding process. However, in phase 3, annealing temperature has higher sensitivity as compared to annealing pressure due to larger slope.

The slope of the S/N ratio curves indicate sensitivity of a given optimization parameter to optimize the response variable, which in this case is the Cu-Cu shear strength. It can be seen in fig. 2.14 (b) that tacking temperature and chuck temperature have similar slope, indicating similar effectiveness in affecting the bonding quality. However, in fig 2.14 (c) we can observe the annealing temperature has higher slope than annealing pressure, which indicates it is a sensitive parameter for optimization compared to annealing pressure. This is corroborated by the fact that changing the temperature from 300ºC to 200ºC can result in significant bonding failures as seen from Fig. 2.12.

2.8 Proposed mechanism for Cu-Cu two-step TCB

Two step TCB is an extension of single step TCB, which not only increases the throughput, but also the metallurgical bond strength of Cu-Cu interconnection. At its core it is a metal-to-metal diffusion bond, however there is enhanced grain boundary diffusion at the bonding interface due to the additional annealing step which is absent in a single step TCB bond.

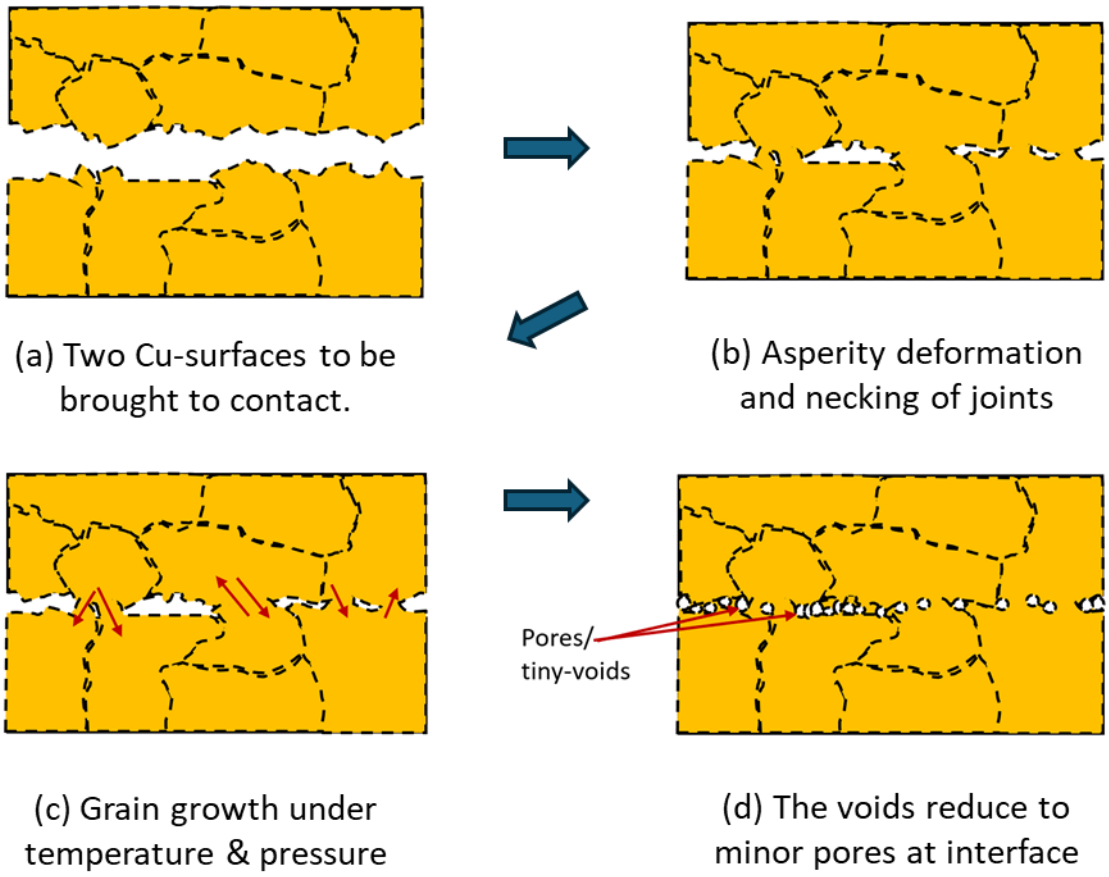
In this section, we propose the mechanism of two-step Cu-Cu TCB. The first step is the dielet tacking step which forms an initial die attach, and the next step is the anneal step where most of the grain boundary growth takes place.

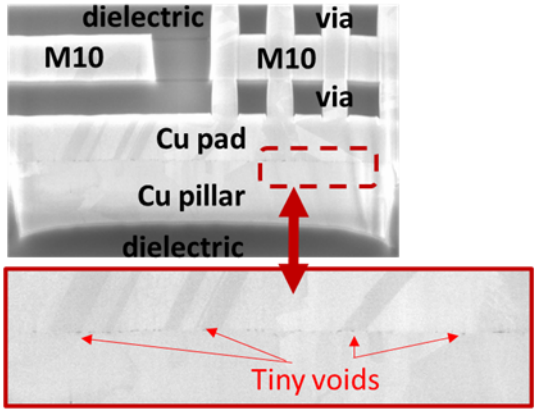
2.8.1 Initial Cu-Cu bond during dielet tacking

Fig. 2.15 (a) shows two Cu surfaces facing each other in a reducing formic acid vapor environment. After reducing the copper oxide, the two Cu surfaces are pristine and ready to bond. At a nanoscopic level, there will be asperities on the Cu surface which are facing each other. As the surfaces come closer to bond, the asperities are the first to make contact on either side. Locally, the pressure on these asperities will cause them to plastically deform leading the asperity linkage as shown in Fig. 2.15 (b). This is a surface diffusion dominated process with necking of asperity joints, but no densification takes place.

At this point, the two surfaces are under both temperature and pressure stress, due to thermal compression which is applied by the bonding tool. Under temperature and pressure, the interface enters a diffusion-controlled mass-transport regime. At such, metal begins to migrate along the connecting grain boundaries, making grain boundary diffusion the dominant diffusion mechanism at the interface. This is schematically shown in Fig. 2.15 (c). Grain boundary diffusion can cause densification, and therefore, the large gaps between the asperities are eventually reduced to small, isolated pores as the bonding continues. This is schematically shown in Fig. 2.15 (d). Post tacking, we have experimentally verified this condition through focused ion beam of the Cu-Cu bonding interface which is shown in 2.15 (e). As can be seen from the micrograph, there are tiny

voids or pores at the interface. At this step, an initial bonding has been achieved, but the shear strength of the bond is below the military accepted shear strength numbers.





The size & number of the pores depends on how long the sample is kept at high temperature and pressure.

Figure 2.15 Mechanism of single step Cu-Cu TCB and dielet tacking.

2.8.2 Increase in Cu-Cu bond strength through annealing

During annealing, the primary bonding process is grain boundary diffusion. The grain boundary diffusion kinetics can be determined by considering the following equations taken from physical metallurgy principles book by Reed Hill [RHillPM]:

$$D_m - D_m^0 = K \cdot t \qquad (4)$$

$$K = K_0 \exp\left(-\frac{Q}{R \cdot T}\right) \qquad (5)$$

Where:

D_m=grain boundary size, K = Kinetics coefficient

As expected, annealing temperature and anneal time have significant impact on the growth of grain boundaries. Keeping annealing time constant, annealing temperature should have significant impact on the bonding. This is further corroborated by the design of experiments, phase 3. In Fig. 2.12, we observe annealing failing at temperatures of 200°C, requiring greater temperatures for appropriate grain growth. In Fig. 2.14, we show large slope of anneal temperature as a parameter, showing the bonding process has high sensitivity to annealing temperature. Anneal pressure also has an effect on bond quality, but not as much as anneal temperature.

From a mechanism perspective, we expect grain growth to occur by moving grain boundaries and the grains can grow into one another at the expense of its volume. Higher anneal temperature can also cause the micro-voids to coalesce to form larger voids, which can move along the grain boundaries. However, these voids are few and do not affect bonding quality as suggested by large shear strength of the dielets post annealing throughout the two-step bonding experiments. This grain growth under annealing conditions is schematically represented in Fig. 2.16(b). We have experimentally verified it again through FIB images of the Cu pad to pillar bonds as seen in Fig. 2.16. Here we do not see a bonding interface, rather a large grain across the bonding interface. This indicates a good metallurgical bond. Also, as shown in section 2.6.3, post bonding when the dielets are shared, the shear force exceeds nearly 4X MIL-SPEC value, i.e. > 110 N for 2x2 mm^2 dielets.

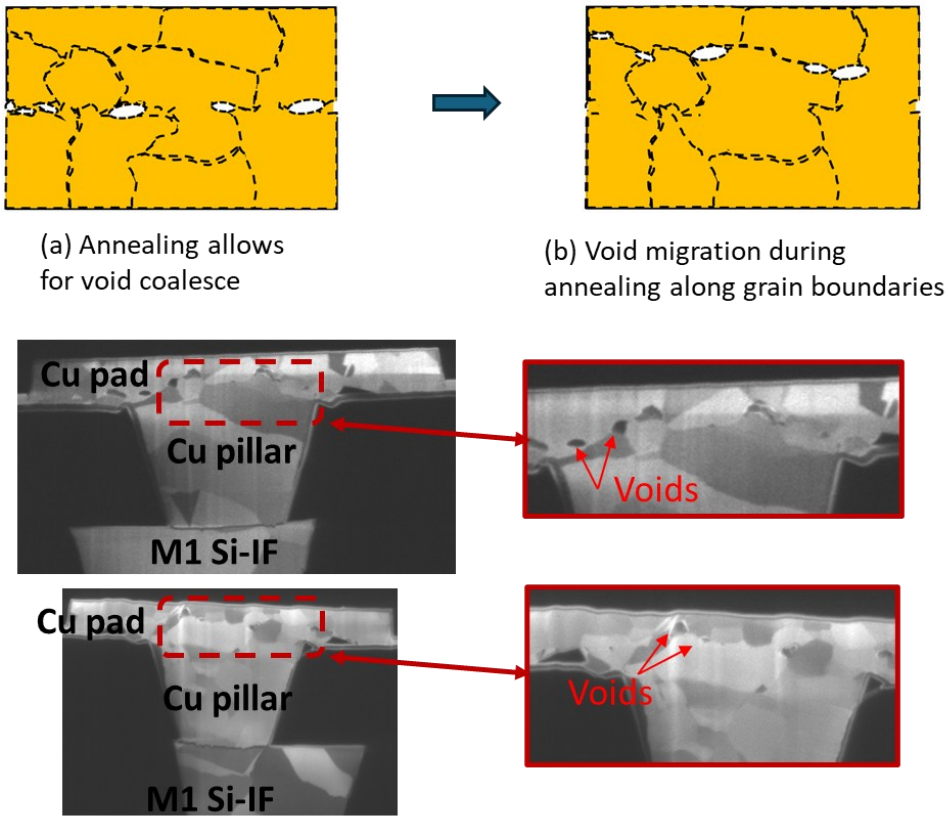


Figure 2.16 Proposed mechanism of the annealing process in two-step Cu-Cu TCB.

2.9 Passivation process for bonded assemblies

Cu-Cu TCB process involves bonding Cu pillars on the Si-IF side to the Cu pads on the substrate side. The dielectric is recessed by about 1 – 1.5 µm, which means there is a gap between the dielet and substrate side. Therefore, there are regions of which are exposed due to dielet-to-substrate misalignment (< 1µm) as well as design rules which require Cu pads on the dielet side to be larger than Cu pillars on the Si-IF side. This exposed Cu is prone to oxidation due to environmental factors such as moisture ingress. This can lead to reliability issues and eventual Cu-Cu bond failure.

Therefore, it is critical to encapsulate or passivate exposed Cu regions. This encapsulation layer would have the following requirements:

- It can effectively protect all exposed Cu layers against environmental factors such as moisture ingress.
- It will have a negligible effect on the thermal dissipation system deployed to remove heat from the dielets.
- It should be a conformal coating which is essential to coat the exposed Cu on the underside of the dielet.

At UCLA-CHIPS Shakoorzadeh et al., [NshakD20] have developed a robust passivation technique using thin film Al_2O_3 deposited using atomic layer deposited (ALD) technique to conformally coat all the exposed surfaces. The deposited thickness varies between 10mm to 17mm, which has negligible added thermal resistance, and does not have any effect on thermal dissipation. Furthermore, atomic layer deposition can conformally coat surfaces and gaps having very high aspect ratio, which makes it the best candidate for encapsulating dielets after bonding. Also, the CTE of Al_2O_3 is about 2.1 – 4 ppm/ºC, which is ideal for coating dielets which are primarily silicon with CTE of 2.6 to 3.3 ppm/ºC.

Therefore, all the bonded assemblies in this work have been passivated with 10 nm ALD Al_2O_3 to maximize resistance against moisture ingress. The parameters used for ALD passivation were obtained from Niloofar Shakoorzadeh's thesis [NShakD21]. The reliability of the Cu-Cu bonds with applied passivation is studied in the next section.

2.10 Electrical and Reliability studies for two-step Cu-Cu TCB

Once the bonding process is optimized, and the assembly is passivated, the next step is to determine the specific contact resistance of the Cu-Cu bonds, as well as perform tests to assess the reliability of the Cu-Cu bonds. An example of a bonded assembly of 16 dielets on Si-IF is presented in Fig. 2.17. Each dielet has several daisy chains and a total of 40 randomly chosen daisy chains are selected to analyze the Cu-Cu specific contact resistance both before and after reliability experiments as described below.

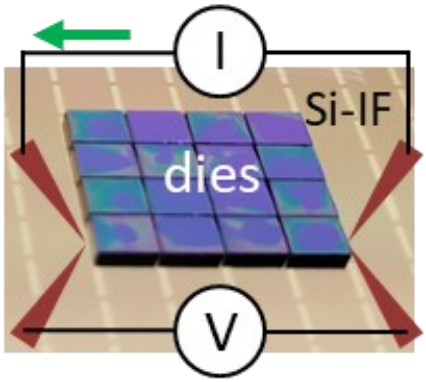


Figure 2.17 Example of 16 dielet assembly bonded using two-step TCB for electrical reliability analysis.

Specific contact resistance of a bond or a contact quantifies the resistance to electrical current flow at the interface between two materials, often a metal and a semiconductor or a metal and another metal [Chang]. It is a key parameter in semiconductor device design and manufacturing as it describes how well electrical signals can be transferred across a contact. It has units of ohm-centimeters squared, though it can also be given as ohm-micrometer squared.

Consider a contact region of uniform area A and length l. In an ideal condition, it is replaced by bulk material – there is no additional resistance to electrical current flow than provided by the bulk of material itself, so specific contact resistance of ideal contact

Ideal specific contact resistance = bulk resistivity X length of contact region

$$= \rho \times l$$

But a non-ideal contact region of uniform area A and length l has a resistance R, which signifies the increased resistance to electrical current flowing across it. In this case, the specific contact resistance is

Measured specific contact resistance = Measured resistance of contact X Area of contact

$$= R_{meas} \times A$$

Specific contact resistance is measured in two ways:

Transfer Length Method/Transmission Line Method (TLM) [WiKiTLM]: It uses a test structure with a series of parallel lines to measure the resistance between them and extract the specific contact resistance. It is often used in the determination of specific contact resistivity of a metal-semiconductor junction

Four-point probe method [Wiki4prb]: Applies a forcing current through the contacts and measures or senses a voltage drop across the contact to determine the contact resistance. Then specific contact resistance can be extracted from the measured resistance and the area of the contact region.

All the measurements done in this work are measured using four-point probe method. The sample is mounted on a manual probe station with four probes. The structure of four-point probe is also shown in Fig. 2.17. Two probes, called forcing probes provide forcing current through the daisy chains, while the other two probes, called sensing probes, are used for measuring voltage across the daisy chains. As no current flows through the sensing probes, there is no voltage drop across the contacts of the probe with the pads. Only the voltage drop across the daisy chain is measured.

2.10.1 Resistance measurements for daisy chain links bonded using two-step thermal compression bonding.

Forty randomly chosen daisy chains, bonded using optimized two step thermal compression bonding approach were chosen for low temperature and high temperature anneal cases for resistance measurements. The four-point probe resistance was measured directly using a digital multimeter model 34401A from Agilent and the distribution for each of the case was plotted. This is shown in Fig. 2.19 where the triangular points show cumulative distribution of 40 four-point resistance measurements before performing any reliability tests in both cases of annealing done at 300°C and 400°C respectively.

2.10.2 reliability tests on daisy chain links bonded using two-step thermal compression bonding.

Electrical reliability tests were carried out in a temperature and humidity test chamber by ESPEC [ESPECSU], shown in Fig. 2.18. This chamber is capable of temperature cycling from -40°C to 125°C (it has a refrigeration unit and heater unit) and also has a humidity control chamber.

For reliability, the passivated samples were subjected to JEDEC unbiased highly accelerated test (UHAST) JESD22-A118, test condition A [JUHAST]. The exact conditions include a temperature of 130 ± 2 C, relative humidity of 85 ± 5 % and a time duration of 96 Hours for test condition A. The resistance of the links post JEDEC UHAST test are presented in Fig. 2.19 which includes resistance measurements both before and after UHAST. After UHAST, the change in resistance is < 8 % for low temperature (LT) anneal case and < 2 % for high temperature anneal (HT) case. We expect this outcome because grain growth is more prominent in the case of high temperature (HT) anneal case, which can be explained by equation [4]. However, in both cases the Cu-Cu interconnections pass JEDEC UHAST test with resistance change < 10 %. A second batch of samples was subjected to steady state humidity bias life test. A long daisy chain consisting of 72000 Cu-Cu contacts was tested in this experiment. The JEDEC steady state humidity test follows the conditions documented in JESD22-A101D [JSSTHB]. In summary the

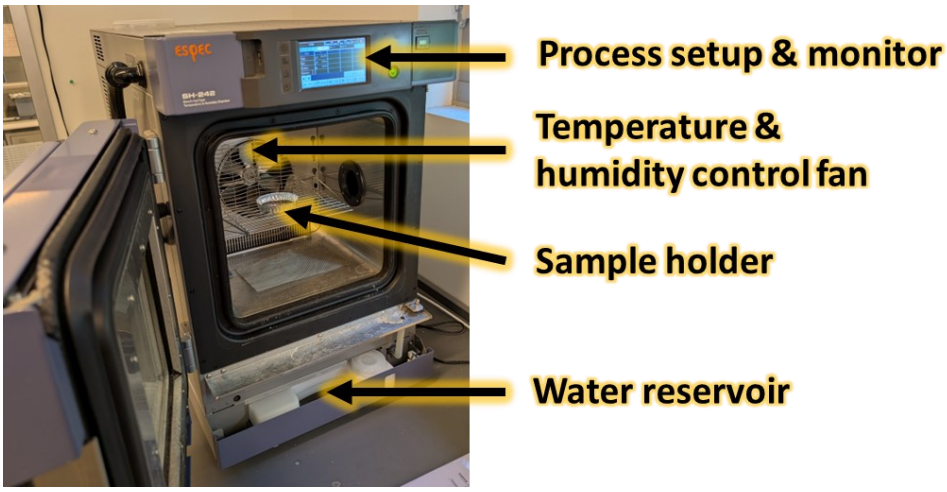


Figure 2.18 ESPEC temperature and humidity control chamber for doing reliability tests.

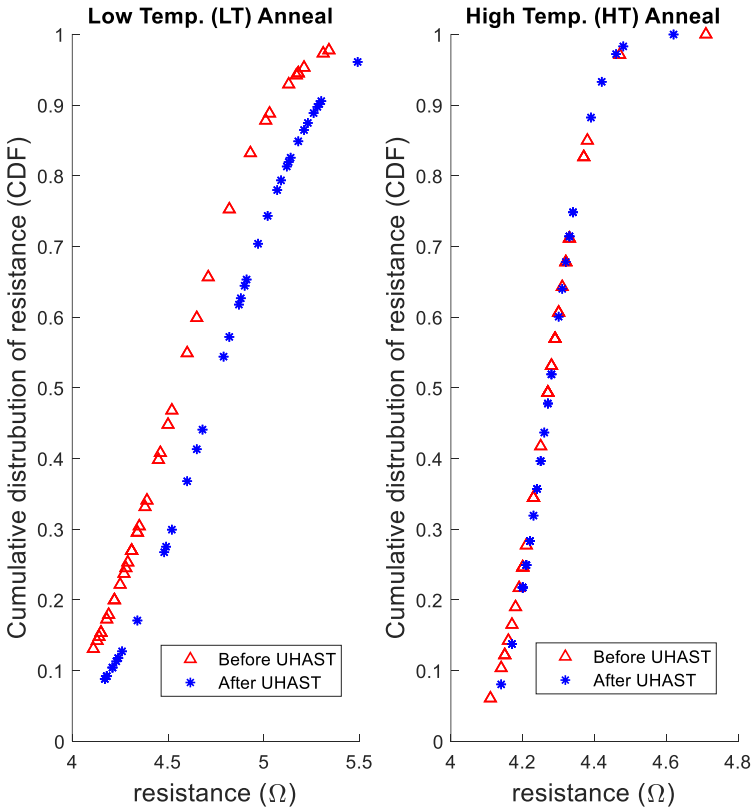
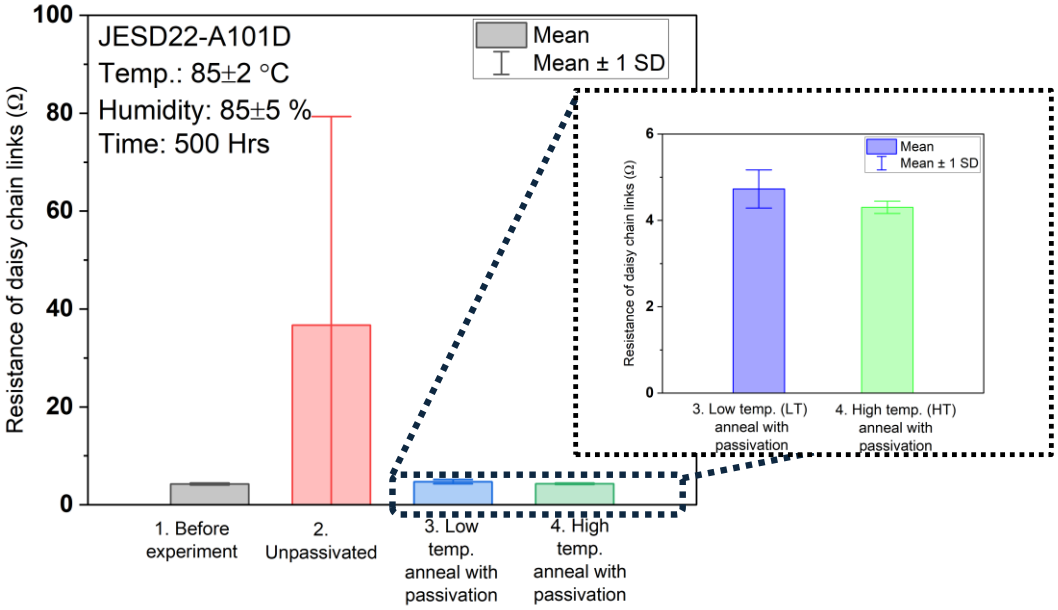


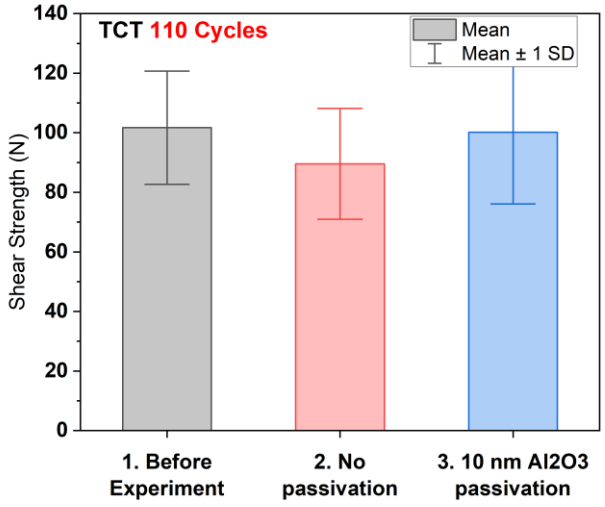
Figure 2.19 Cumulative resistance distribution of 40 randomly chosen daisy chains before and after 96 hours of unbiased highly accelerated stress test (UHAST). In case of both low temperature and high temperature anneal process, the change in the resistance of link after reliability measurements is < 10 %.

samples were subjected to 500 hours of steady state humidity condition of 85 ± 5 % relative humidity and at 85 ± 2 C chamber temperature. The measured resistance results are shown in Fig. 1.20 (a)., Three cases are compared, the case without any passivation but annealed at 400C, the case with passivation and annealed at 300C and the case with passivation and annealed at 400C. The distribution of resistance after steady state humidity test is also shown. From the figure, it is clear that without passivation, there is significant degradation in the specific contact resistance of the daisy chain links.

The third batch of samples was subjected to temperature cycling test (TCT) from JEDEC, using the condition JESD22-A104F [JTC], using test condition K. Again, the samples were subjected to temperature cycling from 0 C to 125 C for 110 cycles. In this instance, only



(a)



(b)

Figure 2.20 (a) Steady state Humidity tests (JESD22-A101D) for passivated and unpassivated samples, showing passivated samples have < 10 % resistance change in both low temperature and high temperature annealing case. (b) Temperature cycling test (TCT) for 110 Cycles showing no change in shear strength in case of Al_2O_3 passivated samples.

the shear stress was investigated, as we expected to test the quality of bonds after continuous temperature cycling. The Fig. 2.20 (b) shows the results of shear when exposed to temperature cycling in the presence and absence of passivation. We can observe the mean shear strength drops without passivation whereas, the change is negligible (< 5%) in presence of 10 nm ALD passivation. Therefore, ALD passivation plays a key role in improving Cu-Cu bonding reliability.

Finally, one of the samples after being exposed to 96 hours of UHAST followed by 1000 hours of steady state humidity test was cross-sectioned using focused ion beam (FIB). The cross-section is shown in Fig. 2.21. As can be seen from the cross-section, there is negligible change to the bonding interface even after significant reliability stress has been applied to the sample. ALD Alumina passivation plays a key role in protecting the Cu-Cu interconnections against moisture ingression. Also, the mating surfaces are both copper. As a result, there is no Kirkendall voiding. Furthermore, due to high electromigration current density of copper (10^6 – 10^7 A/cm^2) when compared to solder (10^4 A/cm^2) we also do not observe electromigration despite electrical testing.

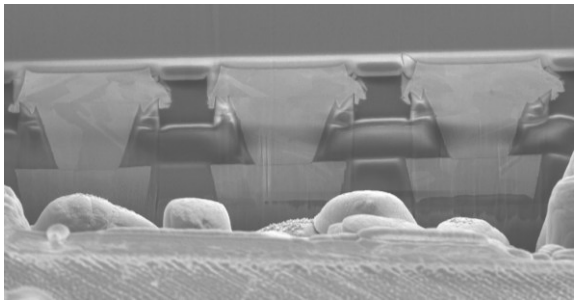


Figure 2.21 Focused ion beam (FIB) cross-section of passivated Cu-Cu TCB showing effectively no change to the bonding interface after extensive reliability stress.

2.11 Application to functional assemblies

Functional dies with active buffered links, designed in 22 nm Global Foundry process, were bonded to corresponding Si-IF using the proposed two-stage bonding scheme. The active links need to be powered for daisy chain testing. After powering up they were tested and verified to be working post bonding as shown in Fig. 1.22. A square wave of 590mV amplitude is applied at the input which results in an approximately 580 mV at the output. This drop can be attributed to IR drop in the daisy chain links as well as probe contact resistance. We can therefore verify that the functionality of the circuit on the dielet is unaffected after two-step Cu-Cu TCB process. This is an important step to demonstrate that the two-step bonding approach does not affect the circuit performance of functional dielets and can be used as a natural improvement over single step TCB.

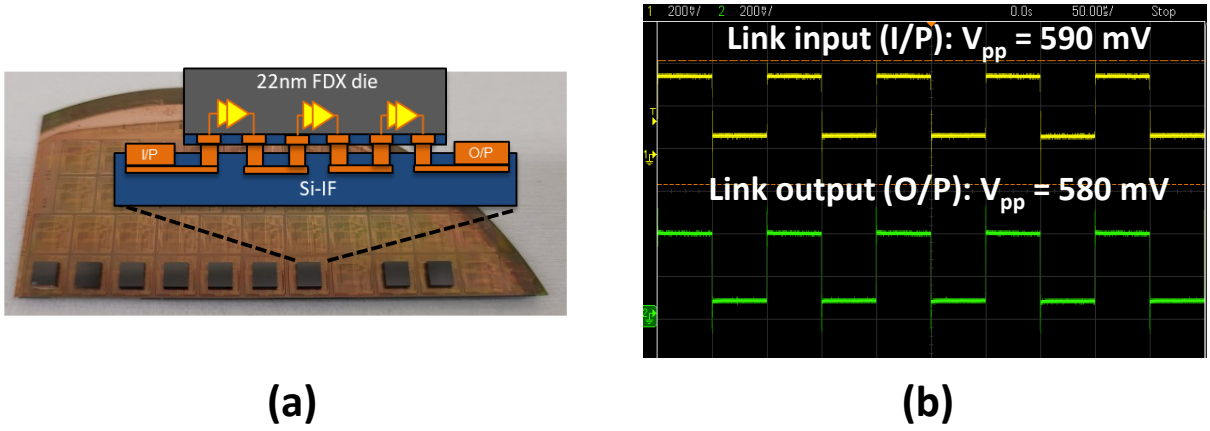


Figure 2.22 Bonding of functional dielets using two-step Cu-Cu TCB. (a) represents dielets with 2880 Cu-Cu interconnections in the active daisy chain link, connected as shown in the schematic. (b) shows applied input to the daisy chain as well as the measured output.

2.12 Methods to further increase Cu-Cu TCB throughput

In this work, we have been able to demonstrate a throughput of 330 units-per-hour (UPH) using two-step Cu-Cu TCB – on a single bond-head. Most bonding tools come with two bond-heads which can work in parallel and hence two dielets can be placed at the same time. This reduces assembly time by half as two wafers can be processed at once. Increasing the number of bond-heads is always beneficial to throughput but it is important to consider the overhead of adding all the additional hardware necessary for the new bond-head work and synchronize operations.

Now, let's focus increasing throughput on a single bond-head. The bonding involves tracking and annealing. Annealing is a parallelized process, and hence annealing time can be eliminated as a throughput concern by annealing several wafers at the same time. However, dielet tacking is a serial process. To improve two-step TCB throughput further, we need to reduce tacking time.

We need to ask the question – what is currently the rate-limiting step in tacking? In general, in standard die-to-wafer bonding, dielet tacking has 5 steps:

- Dielet pick and transfer to bond-head
- Dielet alignment
- Dielet formic acid cleaning
- Dielet contact & raising temperature and pressure to desired TCB values
- Maintaining temperature and pressure for tacking and dielet release

Let us look at each step:

- **Dielet pick and transfer to bond-head:** Dielet is generally picked up from a waffle tray, dicing tape or tape & reel feeder using a picker arm. The picker arm then transfers the dielet to the bond-head. This process may take 2 – 3 seconds depending on camera and X- and Y- axes motor calibrations. However, it is important to note pick and transfer takes place in parallel to the dielet alignment

and formic acid vapor cleaning. While one die is being cleaned and tacked, another dielet is being picked up and ready for transfer. So, from a throughput perspective, this is not the rate limiting step.

- **Dielet alignment:** During dielet alignment the up-and-down place vision camera identifies alignment fiducials on dielet and substrate side and then applies fine-alignment corrections as required. This happens in approximately 1-2 seconds after the dielet and substrate alignment marks are taught to the tool for the bonding process. Unless alignment check is done for more than one iteration, (which may be done to increase alignment correction redundancy,) this step does not become the rate limiting step.

- **Dielet formic acid cleaning:** After alignment, formic acid vapor is used to clean any residual oxides from the Cu mating surfaces prior to contact. This is already explained in section 2.2. In our design of experiments in section 2.6.2, we studied the impact of changing dielet cleaning time from 5 seconds to 3 seconds. From Fig. 2.10, we have not observed a significant difference in bonding strength of the dielets as cleaning time is changed. If bonding is done in an inert environment such as a nitrogen environment or argon environment, we can expect to reduce the dielet cleaning time to less than a second.

- **Dielet contact & raising temperature and pressure to desired TCB values:** The dielet contact takes place when Cu pads on the dielet make physical contact with Cu pillars on the substrate side. When contact is made, the resulting normal force is detected by the bond-head using a load-cell (can be seen by a short force pulse at the beginning of each tacking cycle in Fig. 2.6 (b).) After this step, the desired pressure and temperature are applied by the bond-head. We know that application of pressure is instantaneous, i.e. the bond-head applies a force which is immediately transferred to the Cu pad and Cu pillar in the bonding interface region. However, temperature change is not instantaneous due to finite thermal resistance and thermal capacitance of bond-head, dielet and substrate. As the bond-head heater temperature increases, we need to wait for a few thermal time constants (thermal resistance × thermal capacitance) for this temperature increase

to propagate to the dielet and hence to the Cu-Cu mating surfaces. Furthermore, in steady-state condition, the temperature at the interface is less than temperature of the bond-head.

Thermal compression bonding requires higher temperatures > 240ºC [SJan21] for interdiffusion and bond formation and cannot be reliably done at room temperature. Also, higher temperatures are required for the formic acid vapor cleaning method to work. Analytical measurements of thermal resistance and thermal capacitance can be done to find the time required to raise temperature at interface to ~240°C.

Assuming the dielets are primarily silicon simplifies the calculation. For silicon,

Thermal conductivity of Silicon = K =150 W/m·K

Density of silicon = ρ = 2330 kg/m^3

specific heat capacity of silicon = c_p = 700 J/kg·K

Consider a silicon die of dimensions 2 mm × 2mm and thickness = T = 0.5 mm

Area of die = A = 4 mm^2

Thermal resistance of dielet = R_t = T/K·A = 0.833 K/W

Thermal capacitance of dielet = C_t= $\rho \cdot c_p \cdot A \cdot T = 3.26 \times 10^{-3}$ J/K

Thermal time constant = $\tau = R_t \cdot C_t$ = 2.72 milliseconds

This number is very small. Consider the chuck at steady state temperature (T_o) = 120ºC and the bond-head reaches a temperature T_∞ = 380ºC during bonding.

For the interface to reach T(t) = 240ºC in time t, the equation to solve is given as

$$T(t)=T_\infty+(T_o-T_\infty)e^{-t/\tau}$$

Solving this equation, t = 1.7 milliseconds

Clearly the interface temperature reaches 240ºC in about 1.7 milliseconds as soon as the bond-head temperature is 380ºC. The bond-head takes ~1 sec to go from 120ºC during die alignment to 380ºC at heating rate of 250ºC per second. Hence total time is (1+0.0017) = ~1 second.

- **Maintaining temperature and pressure for tacking and release:** This step of the dielet tacking process is achieved by the bond-head as per user requirements. The user can decide the amount of time required to devote to steady-state bonding

pressure and temperature application. A high tacking time allows sufficient interdiffusion to take place between any metal-metal mating surfaces for a strong bond. Reducing the tacking time would lead to weak bonds, however in two-step bonding, tacking step does not determine the final bond-strength, and hence tacking time can be reduced. Though, it should be noted the tacking time should be just enough for some degree of inter-diffusion to take place to ensure the dielet will not shift from its position while the populated substrate is transferred to an annealing chamber to continue with anneal process. In case of single step TCB, this time can be > 20 seconds or else the bonding can fail. For two-step bonding, we have done design of experiments in section 2.6.1 showing a tacking time of 3 seconds is sufficient for final bonding with anneal. It is possible to fit an anneal chamber into the die-to-wafer bonder to further reduce the steady state time as the populated assembly can be sent for annealing immediately.

The table 2.6 shows the contribution of different aspects of tacking process and how they compare for single step and two step bonding. A proposed tacking step for two step

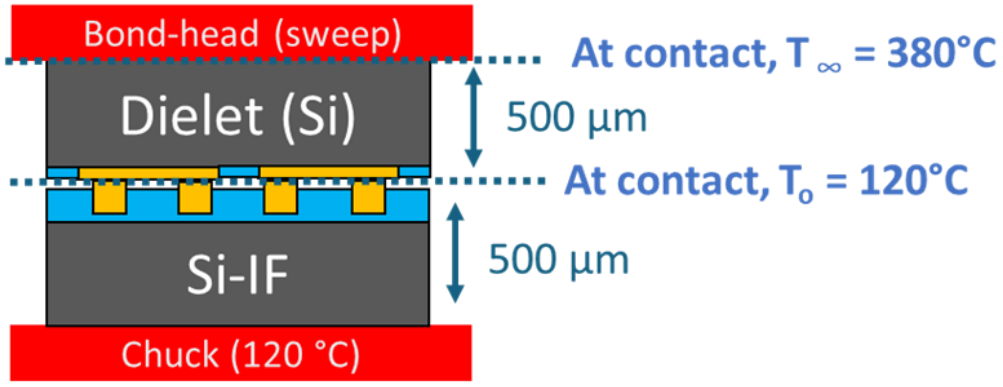


Figure 2.23 Standard schematic to measure the transient time required for the interface to reach a temperature ~240°C at bonding interface during tacking process. Prior to making contact, the substrate is assumed to be at steady state temperature of 120°C. The bond-head after making contact raises the temperature from 120°C to 380°C at the rate of 250°C per second, so it takes ~ 1 second to reach desired TCB temperature and pressure.

bonding to increase the throughput to 1000+ using a single bond-head is also included in the table.

Table 2.6 Throughput roadmap for two-step Cu-Cu TCB done using single bond-head

Parameter (all units in seconds)	**Single -step TCB**	**Two-step TCB (proposed & demonstrated)**	**Two-step TCB step (future)**
Dielet pick and transfer to bond-head	2	5	1.5
Dielet alignment	2		
Dielet formic acid cleaning	8		
Temperature and pressure rise after contact	~ 1	~ 1	~ 1
Steady state application of temperature and pressure	20	3 – 4	1
Total bonding time	**32 – 33**	**9 – 10**	**3.5**
TCB Throughput per bond-head	**~ 100**	**330**	**~ 1000**

2.13 Ti/Au interlayer passivation for Cu pillar using Au-Cu TCB

Using a two-step TCB process, we can potentially achieve > 1000 units per hour (< 6 seconds/dielet) or higher throughput necessary to bond 300 mm wafers without Cu oxidation during bonding becoming a major concern. However, we can further add to the

amount of time the wafer can spend at high chuck temperatures of ≥ 120 °C during bonding by passivating the Cu pillars prior to bonding.

Many Researchers have proposed using metal passivation of Cu surface for direct Cu-Cu bonding. Primarily ultrathin Ti [AKPani16], Ag [TCChou21], and other metals layers are deposited as a passivation layer on Cu to protect Cu from oxidizing. Panigrahi et al. [AKPani16] have deposited Ti layer (3 nm) as a passivation on Cu and by bonding the samples at 2.5 Bar (0.25 MPa) and 160°C.

2.13.1 Development of Au-Cu single step TCB process for passivation

To extend the duration which a Si-IF wafer with exposed Cu pillars can be kept on a hot chuck ≥ 120 °C, we have also pursued use of gold (Au) film passivation on Cu pillars. Au is an inert element, and therefore, does not oxidize at high chuck temperatures unlike Cu. On the substrate side, Cu pillars or pads are capped with 200 nm Au with a 20 nm titanium (Ti) being used as an adhesion layer. On the dielet side, no changes are made to imitate the fact dielets will be from a foundry and technology node which does not process Au. Dielet and substrate are schematically shown in Fig. 2.24.

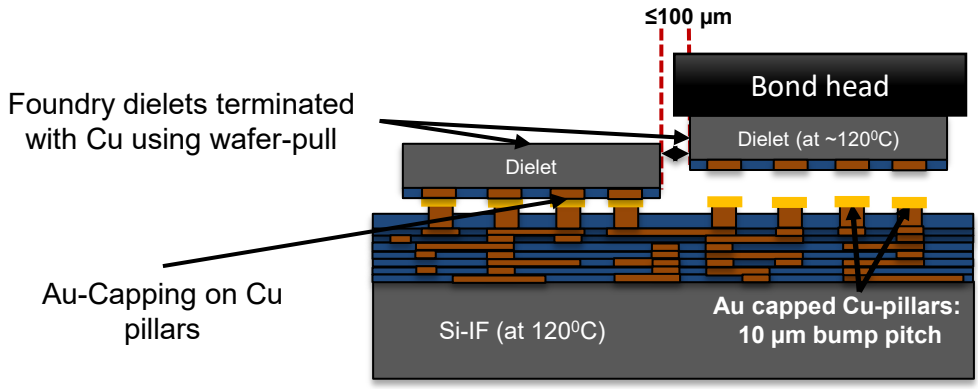


Figure 2.24 Schematic representation of bonding copper (Cu) pad terminated dielets on gold (Au) capped Cu pillars on Si-IF.

Next, we briefly explore the bonding mechanism. From the phase diagram of Au-Cu presented in Fig. 2.25, we see several intermetallic phases Au_3Cu, AuCu and $AuCu_3$ to exist near bonding temperatures of 300 – 400 ºC. Exactly at the bonding interface, we expect a mole fraction 0.5 for Cu and 0.5 for Au due to dielet side pads being Cu

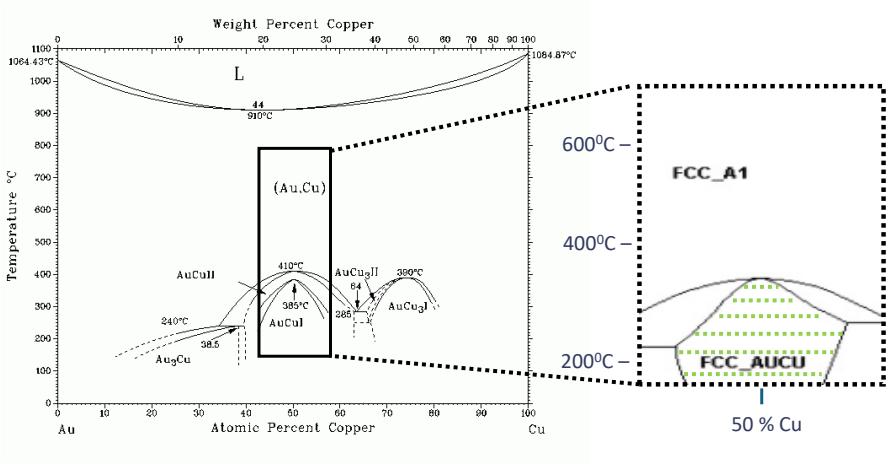


Figure 2.25 Au-Cu phase diagram

terminated and substrate side pads or pillars being Au terminated. Therefore, we expect at 300 – 400 ºC, we will likely have phase AuCu. Then Cu and Au will inter-diffuse to form a dual-phase metallic bond. In summary, under appropriate conditions of temperature and pressure we can form an Au-Cu thermal compression bond. While optimizing the bonding process is necessary, it is not a fundamental challenge. Rather the key challenge with using Au passivation is the cost of deposition as well as the added steps of treatment of depositing Au on Cu pillars/pads on Si-IF or similar substrate.

2.13.2 Experimental observations and results for Au-Cu TCB

In our case, the key research and findings were published in [KSah21]. Note that this work was done prior to two-step Cu-Cu TCB and focused on a proof-of-concept demonstration of Au-Cu TCB. For developing the Au-Cu TCB process, Cu pad to Au-capped-Cu pad bonding on 2x2mm^2 samples was used. The reason for doing pad-to-pad bonding is to reduce the process development time while being able to make samples which can be used for electrical and reliability measurements. On dielet, side, each 2x2 mm^2 dielet was filled with 17x7 um^2 Cu pads. The pads were fabricated using single Cu damascene process as discussed in [SJanD20]. On the substrate side, again single Cu damascene process was used to make Cu pads. This was however followed by deposition of 20 nm titanium (Ti) and 200 nm Gold on Cu using resist lift-off process. The

resist used here is AZ chemical 5214E [AZ5214], which is then patterned. Au is deposited using e-beam deposition for line-of-sight non-conformal deposition. The role of titanium (Ti) is to act as the adhesion layer, as Au is a noble metal and does not adhere well to other metals and dielectrics. Next samples were submerged overnight in Acetone solution for dissolving the photoresist. Ultrasonication process is then used for lift-off of 200nm Au on Cu. Post liftoff, the pads are recessed using fluorine chemistry (CHF3+*CF4*+Ar) for ~1 µm. The sample is then ready for bonding.

A simplified design of experiments was conducted in case of Au-Cu bonding. For Au-Cu bonding process, one step bonding was used as these experiments were designed before the two-step Cu-Cu bonding process was finalized. The parameters used in Au-Cu TCB are presented in the Table 2.7, taken from results published at ECTC.

Table 2.7 Parameter values for Au-Cu TCB.

Parameter name	Value	unit
Bond-head temperature during dielet alignment	100	℃
Substrate temperature during dielet alignment	100	℃
Bonding temperature at interface	~250	℃
Bonding ambient	air	
Bonding force	150, 200, 270	N
Bonding pressure (contact area ~ 1.2 mm^2)	125, 166, 225	MPa
Bonding time	20, 30, 40, 50, 60	s

Prior to bonding, the mating surfaces i.e. Cu surface on the dielet side and Au surface on the substrate side are treated with Ar/H_2 plasma for 1 minute. Ar in Ar/H_2 plasma as described in [AABaj17], helps to clean the mating surfaces by a process called surface activated bonding (SAB) while H_2 helps to provide a reducing environment during the

cleaning process. The bonding was done using dielet-to-wafer bonding tool APAMA, in the same manner as discussed in section 2.5, which includes in-situ formic acid treatment.

To analyze bonding quality, we measure the shear force of the dielets bonded using Au-Cu TCB as they are sheared by the die shear equipment Nordson DAGE 4000Plus. All the shear The Fig. 2.26 (a) shows the results of variation of shear force with bonding force. To exceed MIL-STD 883 [MILSPEC883] shear values, a bonding force of 166 MPa on 2x2 mm^2 dielet was sufficient. Increasing the bonding force leads to increase in shear force, which is expected in thermal compression bonding. At 225 MPa, smaller variability

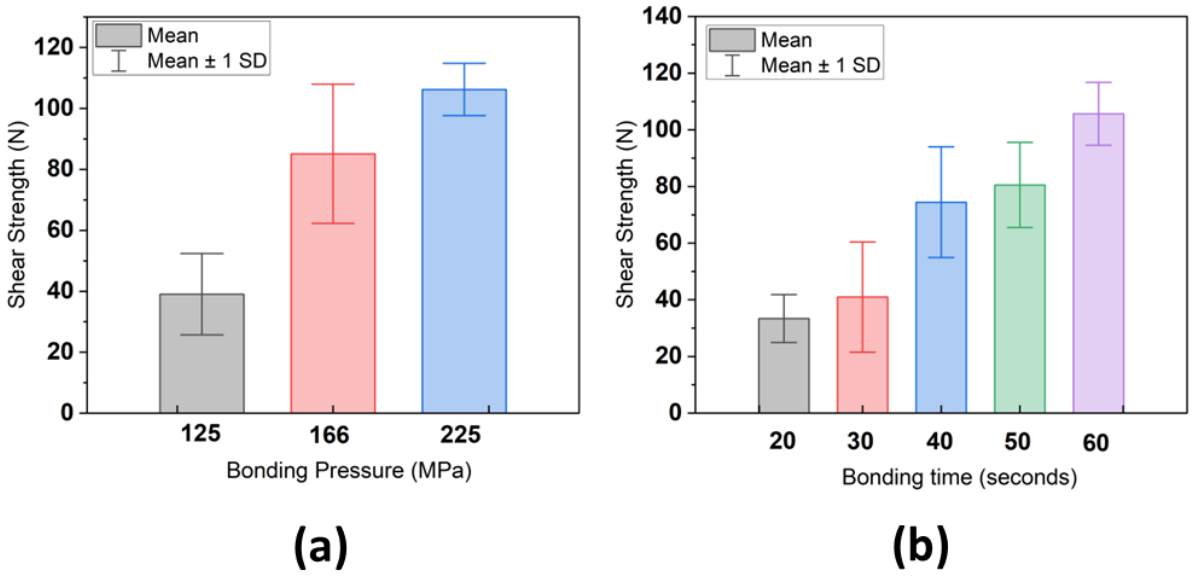


Figure 2.26 (a) Bonding pressure and (b) bonding time optimization of Au-Cu TCB. Note single step TCB was used in this study as it was done before development of high-throughput TCB.

or standard deviation in shear force value was observed. There was no visible damage to the pads even at 225 MPa. Fig. 1.26 (b) shows variation of shear strength with bonding time. As expected, greater bonding time leads to better shear strength, which indicates interdiffusion between Cu and Au increases with bonding time.

For the final assembly, a bonding pressure of 166 MPa and bonding time of 60 seconds was used. As discussed in section 2.13.1, pad to pad bonding was done for testing Au-

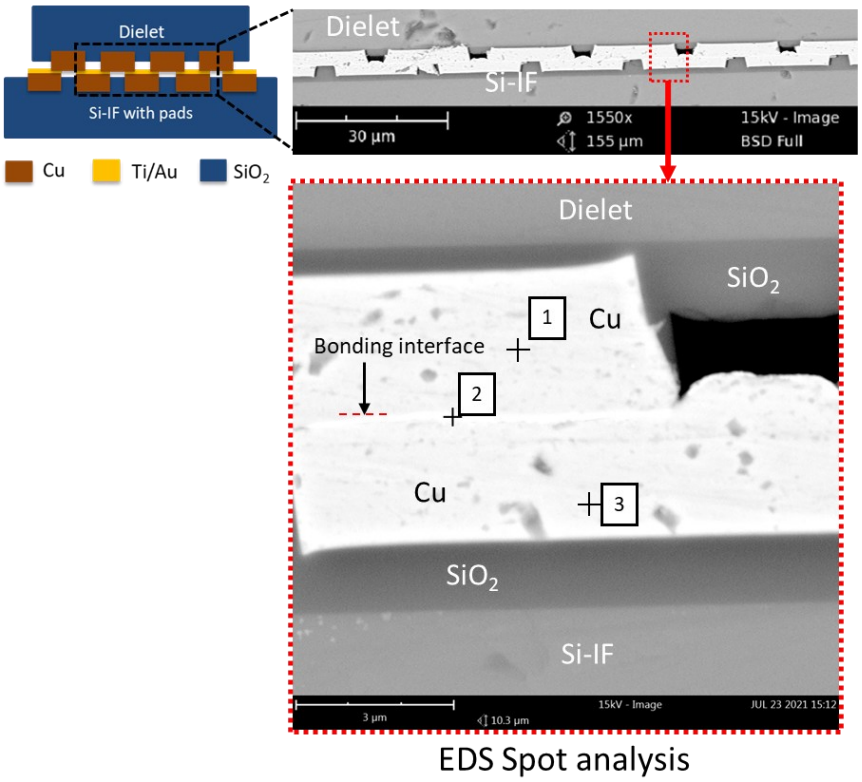


Figure 2.27 Scanning Electron Micrograph (SEM) cross-section of Au-Cu bonding with 3 spot elemental analysis. Analysis of elements is presented in table 2.7.

Cu TCB. The cross-section of pad-to-pad Al-Cu bonding can be seen in Fig. 2.27. At the bonding interface, we do not observe any voids, indicating a good bond. Electron Dispersion Spectroscopy (EDS) is used to identify the Thin film deposited Al layer Cu pads. The elemental identification of metals at the cross-section is shown in Table 2.8. Gold signal is clearly observed at spot 2, with negligible oxygen concentration.

Table 2.8 Electron dispersion spectroscopy (EDS) elemental results for Au-Cu bonding interface.

		Spot 1	Spot 2	Spot 3
Element Number	Element Symbol	Weight Conc.	Weight Conc.	Weight Conc.
79	**Au**	--	**93.29**	--
29	**Cu**	**71.58**	4.89	**73.44**
6	C	9.92	0.72	9.72
14	Si	16.98	1.10	16.83
8	O	1.52	--	--

2.13.3 Reliability of Au-Cu bonds

One of the major concerns in TCB involving multiple metal candidates is the difference between temperature coefficient of thermal expansion (TCE) between the metals. The TCE of copper is about 17.6 ppm per K and that of gold is about 14.2 ppm per K. The difference is less than 5 ppm per K and hence we don't consider any significant effect on the bonding reliability. Nevertheless, the samples were stored at high temperature of 125 C for 2 weeks. These high temperature storage tests followed JEDEC standard JESD22-A103D, Condition A [JHTS]. The results of the shear force measurements are presented in Fig. 1.28 (a). The difference is found to be < 10 % which shows good bonding even after 300 hours of high temperature storage testing.

One of the important goals of this work was to demonstrate the reliability of Cu-Au bonding over extended bonding times of several hours due to the passivation of exposed Cu pads on Si-IF by Ti/Au capping layer. To demonstrate this, the Au-capped-Cu substrate samples were heated to 100 ºC and kept at this elevated temperature for several hours in atmospheric ambient. Dielets were bonded to these heated substrates at specific time

intervals, starting from a fresh sample to the final bonding done on a sample stored at 100 ºC for four hours. The results of the shear test are presented in Fig. 2.28 (b). It is clear that up to 4 hours of testing, we do not see any degradation of bond strength.

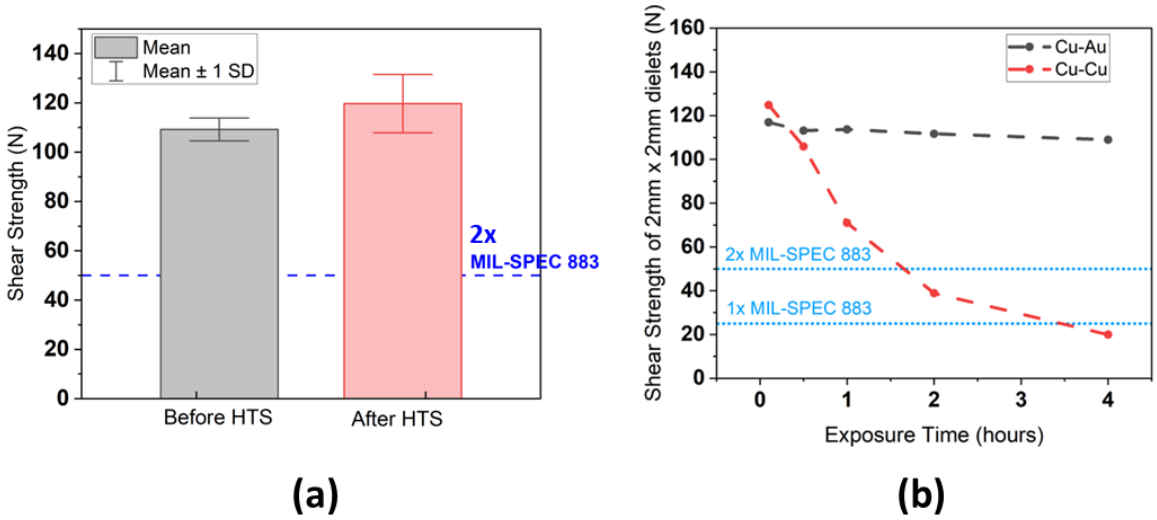


Figure 2.28 Reliability tests for Au-Cu TCB bonds. (a) high temperature storage (HTS) shows < 10 % change in shear strength after 300 hours of test time. (b) Shear force of samples bonded at hourly time intervals demonstrating capability of Au capping layer in preventing oxidation to Cu pillar surface and hence preventing degradation of bonding strength.

2.14 Chapter conclusion

Chapter 2 deals with the process of scaling Cu-Cu thermal compression bonding (TCB) to full 300 mm wafer scale assemblies. To summarize:

- We have demonstrated a novel two-step Cu-Cu thermal compression bonding process with demonstrated throughput of 330 units per hour extendable to 1000 units per hour using extensive design of experiments described in the chapter.
- We have verified the mechanical reliability of the bonded samples by comparing it to military specified (MIL-SPEC) bond strength for samples of a given size. 2mm×2mm dielets require a minimum MIL-SPEC bonding strength of 25 N. We

consider 2x MIL-SPEC i.e. 50 N to be the requirement for Cu-Cu TCB. The average shear strength for two-step Cu-Cu TCB is > 110 N for 2mm×2mm dielets.

- We have experimentally demonstrated the reliability of the samples after passivation with 10nm ALD Al_2O_3. The daisy chains in the samples under test show < 10 % change in resistance of links after JEDEC steady state humidity test for 500 hours and also < 10 % change in resistance of the links after unbiased highly accelerated stress tests (UHAST) for 96 hours of extreme testing (130ºC temperature, 85 % humidity). The extracted specific contact resistance of Cu-Cu bonds is 0.124 $\Omega \cdot \mu m^2$ which is comparable to lowest reported specific contact resistance of bonds in Cu/dielectric hybrid bonding.
- We have demonstrated a novel Au-Cu TCB process to extend assembly times to > 4 hours in air by using the passivation property of Au capping on Cu pillars even if the substrate is present on a chuck at high temperature 100 ºC – 120 ºC. This is especially important when we have to populate a wafer with > 10000 dielets.

Chapter 2 references

[EMD64] E. M. Davis, W. E. Harding, R. S. Schwartz, and J. J. Corning, “Solid Logic Technology: Versatile, High-Performance Microelectronics,” *IBM Journal of Research and Development*, vol. 8, no. 2, pp. 102–114, Apr. 1964, doi: 10.1147/rd.82.0102.

[MDat03] M. Datta, “Electrochemical processing technologies in chip fabrication: challenges and opportunities,” *Electrochimica Acta*, vol. 48, no. 20, pp. 2975–2985, Sep. 2003, doi: 10.1016/S0013-4686(03)00363-3.

[Zli21] Z. Li *et al*., "Scaling Solder Micro-Bump Interconnect Down to 10 μm Pitch for Advanced 3D IC Packages," *2021 IEEE 71st Electronic Components and Technology Conference (ECTC)*, San Diego, CA, USA, 2021, pp. 451-456, doi: 10.1109/ECTC32696.2021.00082.

[Xli14] X. Li, Y. Wang, J. Zhou and T. S. Chang, "Fine pitch BGA device solder bridging mechanism investigation through solder shape modeling," *2014 15th International*

Conference on Electronic Packaging Technology, Chengdu, China, 2014, pp. 796-800, doi: 10.1109/ICEPT.2014.6922769.

[TCHuang17] T.-C. Huang et al., “Scaling Cu pillars to 20um pitch and below: Critical role of surface finish and barrier layers,” in Proc. IEEE 67th Electron. Compon. Technol. Conf. (ECTC), Orlando, FL, USA. Piscataway, NJ, USA: IEEE Press, May 2017, pp. 384–391, doi: 10.1109/ECTC.2017. 324.

[BIN06] B. I. Noh *et al.*, “Effects of number of reflows on the mechanical and electrical properties of BGA package,” *Intermetallics*, vol. 14, no. 10, pp. 1375–1378, Oct. 2006, doi: 10.1016/j.intermet.2005.11.036.

[AABaj17] A. A. Bajwa *et al.*, “Heterogeneous Integration at Fine Pitch (≤ 10 μm) Using Thermal Compression Bonding,” in *2017 IEEE 67th Electronic Components and Technology Conference (ECTC)*, 2017, pp. 1276–1284. doi: 10.1109/ECTC.2017.240.

[SJan19] S. Jangam *et al.*, “Fine-Pitch (≤10 μm) Direct Cu-Cu Interconnects Using In-Situ Formic Acid Vapor Treatment,” in *2019 IEEE 69th Electronic Components and Technology Conference (ECTC)*, 2019, pp. 620–627. doi: 10.1109/ECTC.2019.00099.

[URat22] U. Rathore, S. S. Nagi, S. Iyer, and D. Marković, “A 16nm 785GMACs/J 784-Core Digital Signal Processor Array With a Multilayer Switch Box Interconnect, Assembled as a 2×2 Dielet with 10μm-Pitch Inter-Dielet I/O for Runtime Multi-Program Reconfiguration,” in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2022, pp. 52–54. doi: 10.1109/ISSCC42614.2022.9731582.

[SSNagi23] S. S. Nagi, U. Rathore, K. Sahoo, T. Ling, S. S. Iyer, and D. Marković, “A 16-nm 784-Core Digital Signal Processor Array, Assembled as a 2 × 2 Dielet With 10-μm Pitch Interdielet I/O for Runtime Multiprogram Reconfiguration,” *IEEE Journal of Solid-State Circuits*, vol. 58, no. 1, pp. 111–123, Jan. 2023, doi: 10.1109/JSSC.2022.3212685.

[KSah22] K. Sahoo, U. Rathore, S. Chandra Jangam, T. Nguyen, D. Markovic, and S. S. Iyer, “Functional Demonstration of < 0.4-pJ/bit, 9.8 μm Fine-Pitch Dielet-to-Dielet Links for Advanced Packaging using Silicon Interconnect Fabric,” in *2022 IEEE 72nd Electronic*

Components and Technology Conference (ECTC), May 2022, pp. 2104–2110. doi: 10.1109/ECTC51906.2022.00332.

[SJanD20] SivaChandra Jangam's thesis, [online] available: https://s3.us-west-1.amazonaws.com/chips.user.media/page_file/Jangam_dissertation_v6+Printed.pdf

[CNLi24] C.-N. Li, J.-J. Ong, W.-L. Chiu, S.-C. Yang, H.-H. Chang, and C. Chen, "Low-Temperature Cu-Cu Bonding Using <111>-oriented and Nanocrystalline Hybrid Surface Grains," in *2024 IEEE 74th Electronic Components and Technology Conference (ECTC)*, May 2024, pp. 1312–1316. doi: 10.1109/ECTC51529.2024.00213.

[AElsh21] A. Elsherbini *et al.*, "Enabling Hybrid Bonding on Intel Process," in *2021 IEEE International Electron Devices Meeting (IEDM)*, Dec. 2021, p. 34.3.1-34.3.4. doi: 10.1109/IEDM19574.2021.9720586.

[YLin24] Y. Lin *et al.*, "Direct Die-to-Wafer Hybrid Bonding Using Plasma Diced Dies and Bond Pad Pitch Scaling Down to 2 μm," in *2024 IEEE 74th Electronic Components and Technology Conference (ECTC)*, May 2024, pp. 40–44. doi: 10.1109/ECTC51529.2024.00015.

[JJOng22] J.-J. Ong *et al.*, "Low-Temperature Cu/SiO2 Hybrid Bonding with Low Contact Resistance Using (111)-Oriented Cu Surfaces," *Materials*, vol. 15, no. 5, 2022, doi: 10.3390/ma15051888.

[TWork21] T. Workman *et al.*, "Die to Wafer Hybrid Bonding and Fine Pitch Considerations," in *2021 IEEE 71st Electronic Components and Technology Conference (ECTC)*, Jun. 2021, pp. 2071–2077. doi: 10.1109/ECTC32696.2021.00326.

[HRen22] H. Ren, Y.-T. Yang, and S. S. Iyer, "Recess Effect Study and Process Optimization of Sub-10 μm Pitch Die-to-wafer Hybrid Bonding," in *2022 IEEE 72nd Electronic Components and Technology Conference (ECTC)*, May 2022, pp. 149–156. doi: 10.1109/ECTC51906.2022.00034.

[NShak20] N. Shakoorzadeh Chase, K. Sahoo, Y. Tao Yang, and S. S. Iyer, "Atomic Layer Deposited Al2O3 Encapsulation for the Silicon Interconnect Fabric," in *2020 IEEE 70th*

Electronic Components and Technology Conference (ECTC), Jun. 2020, pp. 1241–1246. doi: 10.1109/ECTC32862.2020.00198.

[KSah24] K. Sahoo, V. Harish, H. Ren, and S. S. Iyer, “A Review of Die-to-Die, Die-to-Substrate and Die-to-Wafer Heterogeneous Integration,” *IEEE Electron Devices Reviews*, pp. 1–28, 2024, doi: 10.1109/EDR.2024.3501214.

[SJan21] S. Jangam and S. S. Iyer, “Silicon-Interconnect Fabric for Fine-Pitch (≤10 μm) Heterogeneous Integration,” *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 11, no. 5, pp. 727–738, 2021, doi: 10.1109/TCPMT.2021.3075219.

[KSah23] K. Sahoo, H. Ren, and S. S. Iyer, “A High Throughput Two-Stage Die-to-Wafer Thermal Compression Bonding Scheme for Heterogeneous Integration,” in *2023 IEEE 73rd Electronic Components and Technology Conference (ECTC)*, 2023, pp. 362–366. doi: 10.1109/ECTC51909.2023.00067.

[Besi] BESI Bonding. Accessed: Sep. 5, 2024. [Online]. Available: https://www.besi.com/products-technology/productgroup/die-attach/

[MILSPEC883] “MIL-STANDARD 883,” Accessed: Sep. 6, 2024. [Online]. Available: http://everyspec.com/MIL-STD/MIL-STD-0800-0899/MIL-STD-883H_21092/

[NOFO2] NAPMP NOFO2, [online] Available: https://www.nist.gov/system/files/documents/2024/10/18/CHIPS%20NAPMP%20NOFO%202.pdf

[KSah21] K. Sahoo, S. Pal, N. Shakoorzadeh, Y.-T. Yang, and S. S. Iyer, “Copper to gold thermal compression bonding in heterogenous wafer-scale systems,” in *2021 IEEE 71st Electronic Components and Technology Conference (ECTC)*, 2021, pp. 487–493. doi: 10.1109/ECTC32696.2021.00088.

[NGho17] N. Ghosh, P. K. Pal, and G. Nandi, “GMAW dissimilar welding of AISI 409 ferritic stainless steel to AISI 316L austenitic stainless steel by using AISI 308 filler wire,” *Engineering Science and Technology, an International Journal*, vol. 20, no. 4, pp. 1334–1341, Aug. 2017, doi: 10.1016/j.jestch.2017.08.002.

[APAMAC2W] Kulicke and Soffa, Apama C2W tool, [online]. Available: https://www.kns.com/Products/Equipment/Advanced-Solutions-(TCB)/APAMA-C2W

[NDAGE] Nordson Dage 4000 plus bond-tester, [online]. Available: https://www.nordson.com/en/products/test-and-inspection-products/dage-4000plus-bondtester

[RHillPM] Physical metallurgy Principles by Reed Hill, accessed: 4/20/2025, Available: https://books.google.com/books/about/Physical_Metallurgy_Principles.html?id=_plTAAAAMAAJ

[NShakD21] Nelly Shakoorzadeh Chase's dissertation. Available: https://s3.us-west-1.amazonaws.com/chips.user.media/page_file/Niloofar_Shakoorzadeh.pdf

[Chang] Metal/Semiconductor ohmic contacts, available: https://web.stanford.edu/class/ee311/NOTES/Ohmic_Contacts.pdf

[WikiTLM] Transfer length method, Wikipedia, Available: https://en.wikipedia.org/wiki/Transfer_length_method

[Wiki4prb] Four terminal sensing, Wikipedia, Available: https://en.wikipedia.org/wiki/Four-terminal_sensing

[ESPECSU] ESPEC temperature and humidity chamber, [online], available: https://espec.com/na/products/family/sh_humidity

[JUHAST] JEDEC Unbiased Highly Accelerated Stress Test (UHAST), available: https://www.jedec.org/standards-documents/docs/jesd-22-a118

[JSSTHB] JEDEC steady state temperature humidity bias life test, available: https://www.jedec.org/document_search?search_api_views_fulltext=A101D

[JTC] JEDEC temperature cycling test, available: https://www.jedec.org/document_search?search_api_views_fulltext=A104F

[AKPani16] A. K. Panigrahi, S. Bonam, T. Ghosh, S. G. Singh, and S. R. K. Vanjari, "Ultra-thin Ti passivation mediated breakthrough in high quality Cu-Cu bonding at low

temperature and pressure," *Materials Letters*, vol. 169, pp. 269–272, 2016, doi: https://doi.org/10.1016/j.matlet.2016.01.126.

[TCChou21] T.-C. Chou *et al.*, "Electrical and Reliability Investigation of Cu-to-Cu Bonding With Silver Passivation Layer in 3-D Integration," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 11, no. 1, pp. 36–42, 2021, doi: 10.1109/TCPMT.2020.3037365.

[AZ5214] AZ series 5214E resist, documentation available: https://www.microchemicals.com/dokumente/datenblaetter/tds/merck/en/tds_az_5214e_photoresist.pdf

[JHTS] JEDEC high temperature storage life test [online], available: https://www.jedec.org/document_search?search_api_views_fulltext=A103D

Chapter 3 Important considerations for improving Cu-Cu assembly yield for scale-out systems

3.1 Understanding factors beyond thermal compression bonding which affect the assembly yield

In Chapter 2, we demonstrated a high-throughput Cu-Cu thermal compression bonding (TCB) process capable of delivering over 1000 units per hour (UPH), a key milestone toward making Cu-Cu TCB viable for high-volume manufacturing. We examined in detail how various factors affected bonding quality in the dielet tacking and annealing steps—including tacking pressure, chuck temperature, annealing pressure, and annealing temperature. Additionally, we presented a design of experiments (DOE) framework aimed at optimizing the Cu-Cu TCB process for maximum throughput.

However, Cu-Cu bonding is only one part of the larger assembly workflow. The assembly process begins once dielet and substrate wafers are released from fabrication and encompasses several critical stages: inspection, testing, and verification of dielets and substrates; dielet singulation and preparation prior to bonding; bonding tool capability assessment; execution of the Cu-Cu bonding process; and post-bonding inspection, testing, and verification of the assembled substrate. For a product to be considered manufacturable, the assembly yield must meet or exceed 99.9997%, which corresponds to a 6-sigma yield—equivalent to just 3.4 defects per million assemblies.

In this chapter, we shift our focus to the broader set of factors that impact the die-to-substrate (D2S) or die-to-wafer (D2W) assembly yield beyond Cu-Cu bonding. Section 3.2 describes the factors affecting assembly yield which are studied in this chapter. The

subsections of 3.3 delves into experiments performed including types of dielets used, the experimental procedure and results. Section 3.4 discusses dielet dicing requirements for Cu-Cu TCB, 3.5 discusses the effect of dishing, 3.6 discusses design related consideration for achieving higher assembly yield, section 3.7 discusses known good substrate approaches to improve Si-IF wafer fabrication yield.

3.2 Parameter drift vs. wafer-scale assembly yield

The factors which affect the assembly yield beyond Cu-Cu bonding are described in Table 3.1. While temperature and pressure are key bonding enablers for thermal compression bonding, the yield depends on dielet and substrate design choices, process assumptions for dielet and substrate fabrication, dielet and substrate conditioning prior to bonding and most importantly, the specifications of the bonding tool used for doing the dielet-to-dielet (D2D), dielet-to-wafer or dielet-to-substrate (D2S/D2W) bonding. These parameters also depend on the type of bonding done such as solder flip-chip, Cu/dielectric Hybrid Bonding (HB), Cu-Cu Thermal Compression Bonding (TCB) etc. We have discussed the different types of D2W and D2S bonding schemes in our review work in IEEE Electron Device Reviews [KSah24]. It is crucial to understand that no process related parameter and tool related parameter stays absolutely within specifications as multiple dielets are tacked or bonded. Every process in real life is subject to a process drift, and this drift needs to be characterized and accounted for as we scale up the bonding process to wafer-scale assemblies.

In this chapter, we will elaborate on the design, processing and bonding tool related factors in the context of Cu-Cu TCB, including experimental observations of how process and bonding tool drifts can affect assembly process multiple wafers are tacked. The bonding tool used in this work is APAMA C2W bonder as in the previous chapter, however the lessons learnt are applicable to any bonder and bonding process. For testing the bonding tool tilt and bonding tool alignment, passive dielets with Cu pads were used for analysis. The details are described in section 3.3.1. For analyzing process related and design related factors which affect assembly process, dielets fabricated in-house as well as from commercial foundry were used.

Table 3.1 Factors affecting assembly yield in die-to-wafer Cu-Cu TCB

Design Related	Processing Related	Bonding tool related
Bonding pad distribution (determines bonding pressure and pressure uniformity)	Dielet edge condition after dielet singulation	Bond-head tilt (tilt < 1μm/10mm, as described by manufacturer)
Known Good Dielet (KGD) and Known Good Substrate (KGS)	Cu pad roughness on dielet and substrate (Rq ~10 nm RMS)	Alignment accuracy (3σ = ± 2μm) and accuracy drift
Detailed documentation of the dielet and substrate	Cu dishing on bonding pillars/pads (< 50 nm RMS)	
	Cu recess pillar height (~1 μm)	

In developing the high throughput bonding process discussed in the previous chapter, we have taken precautions to ensure these factors are always within target specifications (described in the brackets in the table 3.1). This is done via bonding process re-teaching, dielet and substrate conditioning as required and so on. In this chapter, we discuss how the assembly yield gets affected when the factors drift from the target specifications. This ensures the results obtained from Chapter 2 are independent of the extended analysis done in this chapter.

3.3 Effect of bonding-tool related parameters on assembly yield

In a pick and place dielet-to-substrate/wafer bonding tool, the dielets as well as substrate to be bonded are first treated either with argon plasma, acetic acid treatment or both to reduce copper oxide contaminants from the Cu pads to be bonded. Either approach is experimentally found to be equally effective in cleaning Cu pads. They are then assembled on a waffle chip tray or a tape and reel feed to be picked up by the bonding

tool pick arm for tacking. During tacking process, the tool picker transfers the dielet from the waffle tray to the place tool mounted on a bond-head. Then the dielet is aligned to the corresponding site on the substrate and tacked in an in-situ formic acid environment [SJan19] as described in chapter 2. We have studied two parameters of bonding tool which have a direct effect on assembly yield –

- Bond-head parallelism or bond-head tilt
- Alignment drift across wafer-scale dielet bonding

3.3.1 Effect of bond-head tilt on assembly yield

The bond-head surface must be maintained parallel to the substrate to ensure the dielet makes uniform contact with all the Cu pillars on the substrate. We assume the pillar heights are uniform and within specifications. Any tilt in the bond-head directly impacts the assembly quality: if the bond-head is significantly tilted, one edge of the dielet will contact the substrate before the other. As a result, the pillars on the early-contacting side will experience higher-than-expected pressure, leading to potential deformation. This issue becomes particularly significant in case of functional dielets, as IO circuitry is often placed below the Cu pads which are to be bonded to the Cu pillars. Furthermore, often IO pads are not uniformly distributed across the dielet to be bonded. An example of a dielet with non-uniform pillar density is shown in Fig. 3.3 (a). Under tilt conditions, IO pads that contact the Cu pillars first are subjected to elevated bonding pressure, which can damage underlying circuits. This phenomenon is illustrated in Fig. 3.1 and presents a significant challenge for wafer-scale integration.

To access the bond-head tilt, tests were done with passive as well as functional dielets. The passive dielets used in bond-head tilt study were 2 mm X 2 mm dielets. In this experiment, the bond-head tilt was first reduced to specified value of < 1μm/10mm in both x-direction and y-direction of bond-head prior to starting any dielet tacking. This was done using a standard procedure already established by the equipment manufacturer. Next 5 dielets were bonded and the bond-head tilt evolution was measured. This was followed by 5 more dies before measuring tilt, and by 5 more dies for final measurement. The evolution of bond-head tilt at end of 5, 10 and 15 dielets is shown in Fig. 3.2.

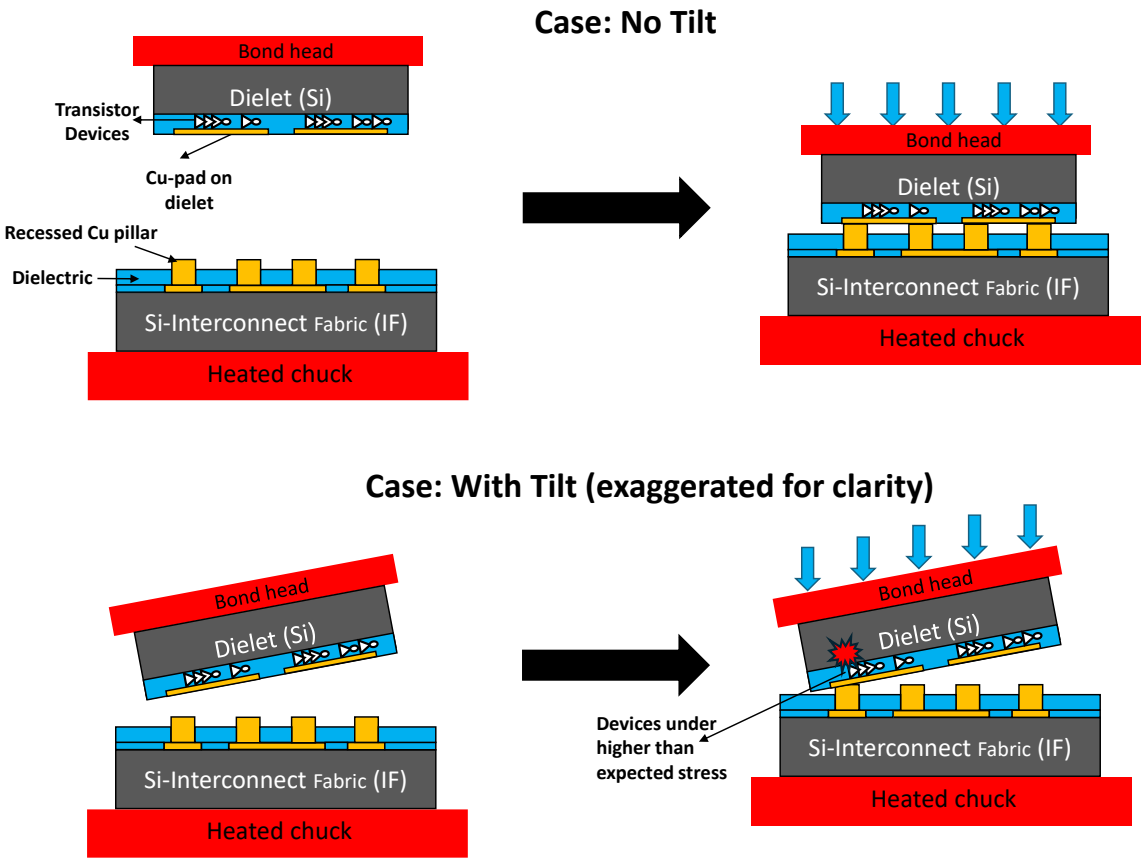


Figure 3.1 Comparison of bonding in the absence and presence of tilt. As can be seen from the figure, when tilt is present, the devices under Cu pads on the tilted side are under higher stress than expected during dielet tacking.

From the figure 3.2, we can observe by the time the 10th dielet is bonded after 1st bond-head tilt correction, the tilt is out of specification. In dielet designs where Cu bonding pads are not uniformly distributed, bond-head tilt can cause the circuitry below a Cu-pad to be damaged during tacking to a Cu pillar and it has the potential to reduce assembly yield.

To address the challenges introduced by bond-head tilt, two approaches can be considered. The first approach involves performing bond-head tilt correction at each individual bonding site. This method relies on confocal parallelism measurements at every bond location [DMay22], and several equipment manufacturers have begun implementing per-site parallelism correction. However, not all bonding tools support this capability, and integrating confocal measurement systems can significantly increase tool cost and complexity.

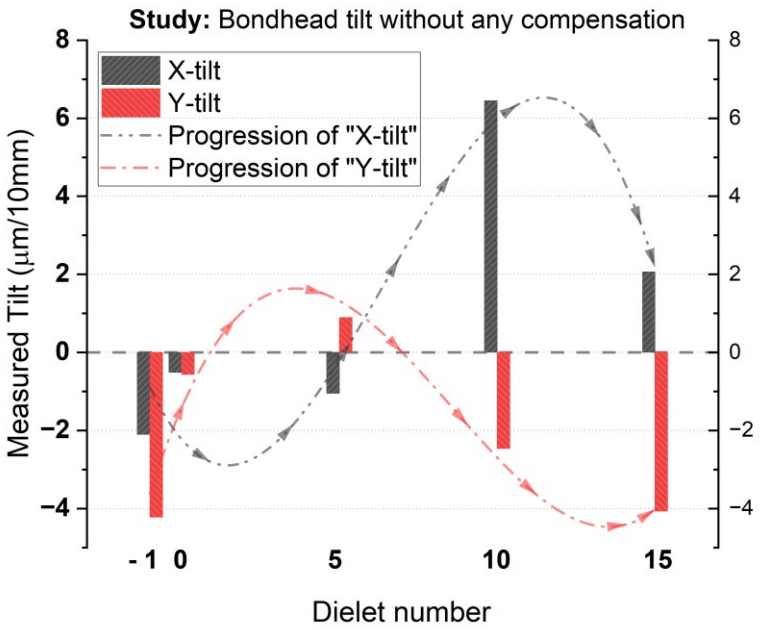


Figure 3.2 Bond-head tilt behavior with continuous bonding.

The second approach which is adopted in this work aims to use a uniform dummy pillar density across the substrate, to reduce the pressure per pillar during dielet tacking. To demonstrate the effect of adding supporting pillars, let us consider an example of the dielet shown in Fig. 3.3 (a) which has Cu pads non-uniformly distributed at top and bottom rows of the dielet as shown in the Figure. This memory dielet was designed by Pal et al. [SpalD21] and fabricated by TSMC in 40 nm low power (40LP) technology and we do not take any credit for the dielet design and fabrication. The dielet has only 1600 Cu dielet pads or input/output (IO) in four rows, two at the top of the dielet and two at the bottom. As per the Si-IF design manual, each Cu pad has one Cu pillar and hence we only have 1600 Cu pillars (each is an IO pillar) to bond to the Cu pads on the dielet. In this situation, bonding tilt is especially a concern for dielet, as if there is a tilt before dielet attach, significant pressure can be felt by the side of the dielet which contacts Cu pillars first on the Si-IF and can lead to damage as described by Fig. 3.1 earlier.

The memory dielet is bonded and sheared using Nordson DAGE tool described in details in Chapter 2. When the applied bonding force is 8N for this dielet, which corresponds to a bonding pressure ~ 250 MPa, imprints are observed only on right side of the dielet, but none occur the left side of the die, as seen from Fig 3.4 (a). Clearly, this is a direct effect of bond-head tilt, and this dielet is found to be non-functional after bonding. To get imprints on both the left side and right side, bonding force was increased. At 60N bonding force on the dielet (with only 1600 pillars) we observe pillar imprints on both left and right side. The dielet is well bonded, however the pressure on the dielet during bonding is > 1700 MPa. It was found that at such high bonding pressure more than 50 % of the bonded dies exhibited high leakage, indicating damage to internal circuitry under the IO pads. The high pressure was originally applied to overcome bond-head tilt.

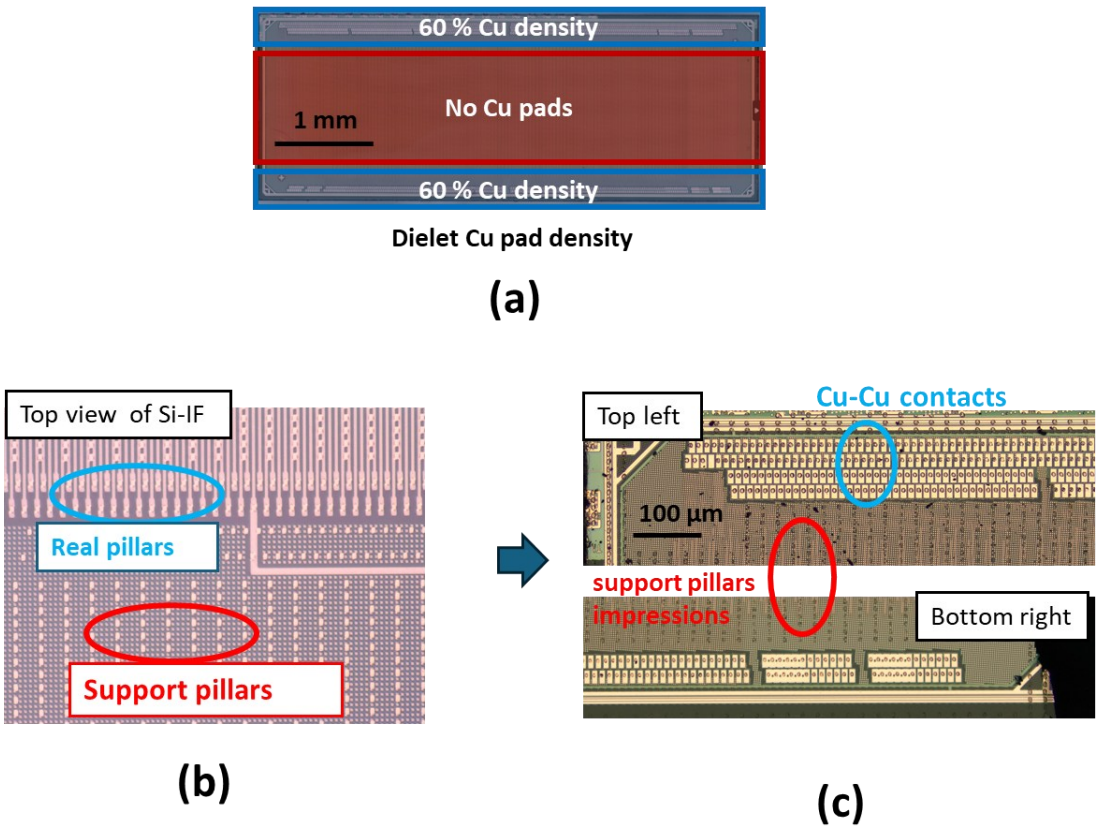


Figure 3.3 (a) Dielet with uneven Cu pad density. (b) Si-IF with real pillars connected to wiring layers and support pillars etched halfway into the dielectric (c) pillar imprints after bonding to an Si-IF with support pillars.

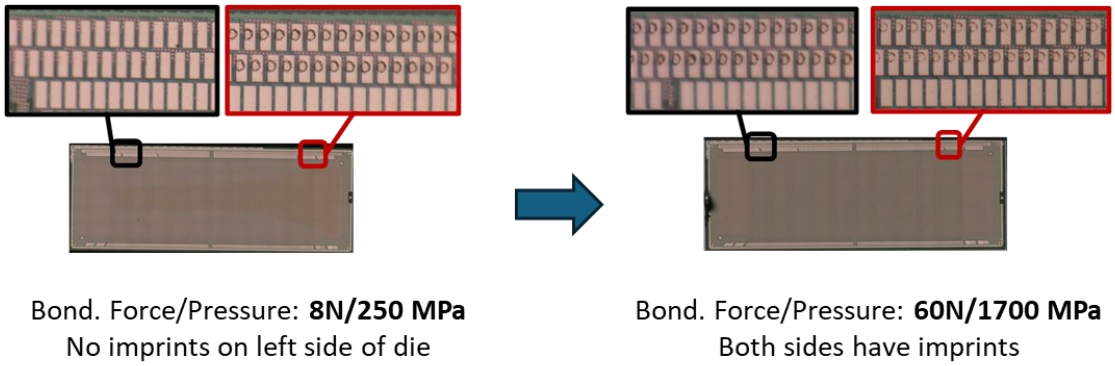


Figure 3.4 Effect of bonding tilt on assembly. Due to bonding tilt, the entire dielet does not make contact with the Cu pillars on the substrate. Bonding pressure can be increased to make contact on both sides, but it can lead to damage to the circuitry on the dielet

To remove tilt, active tilt corrections are necessary per site. This can be done via confocal measurement of tilt at each site as described earlier, but our bonding equipment does not have this facility. However, we sought to address the tilt challenges from a design perspective. Design changes were made to the SI-IF side to reduce the impact of non-uniform bonding pad/pillar distribution. The key approach was to add a Cu supporting pillars in those regions where no Cu pads are present. This is shown in Fig. 3.3 (b). These supporting pillars are only etched halfway into the Si-IF pillar dielectric, ensuring they do not connect to the wiring layers below. The purpose of these pillars is purely mechanical i.e. to support the dielet during bonding and uniformly distribute the bonding force to ensure a uniform pressure distribution. Fig. 3.3 (c) shows the Si-IF with real pillars (blue box) connected to wires in the topmost wiring layer and dummy pillars/support pillars (red box) just for support. Furthermore, supporting pillars are added to the periphery of the active pillars particularly on the shielding ring of the dielet.

To capture the effect of adding support pillars, ANSYS mechanical simulation were done. The simulation was created with two models. Model 3.A has only Cu IO pillars and no support pillars, and Model 3.B has Cu IO pillars + support pillars with a pillar density ~ 10 %. This is shown in Fig. 3.5. Fig 3.5(a) shows the side view of a die bonded to Si-IF and a blue rectangle inset shows the region of interest where the two rows IO pillars are located. The thickness of the die is 300μm to match the thickness of the memory dielet

discussed above. For clarity we hide the Silicon die and back end of line (BEOL) wiring layers in Fig. 3.5(b) and show model 3.A with only Cu IO pillars which protrude 1μm from the dielectric surface and model 3.B with both Cu IO pillars and Cu support pillars with same 1μm protrusion as previous case.

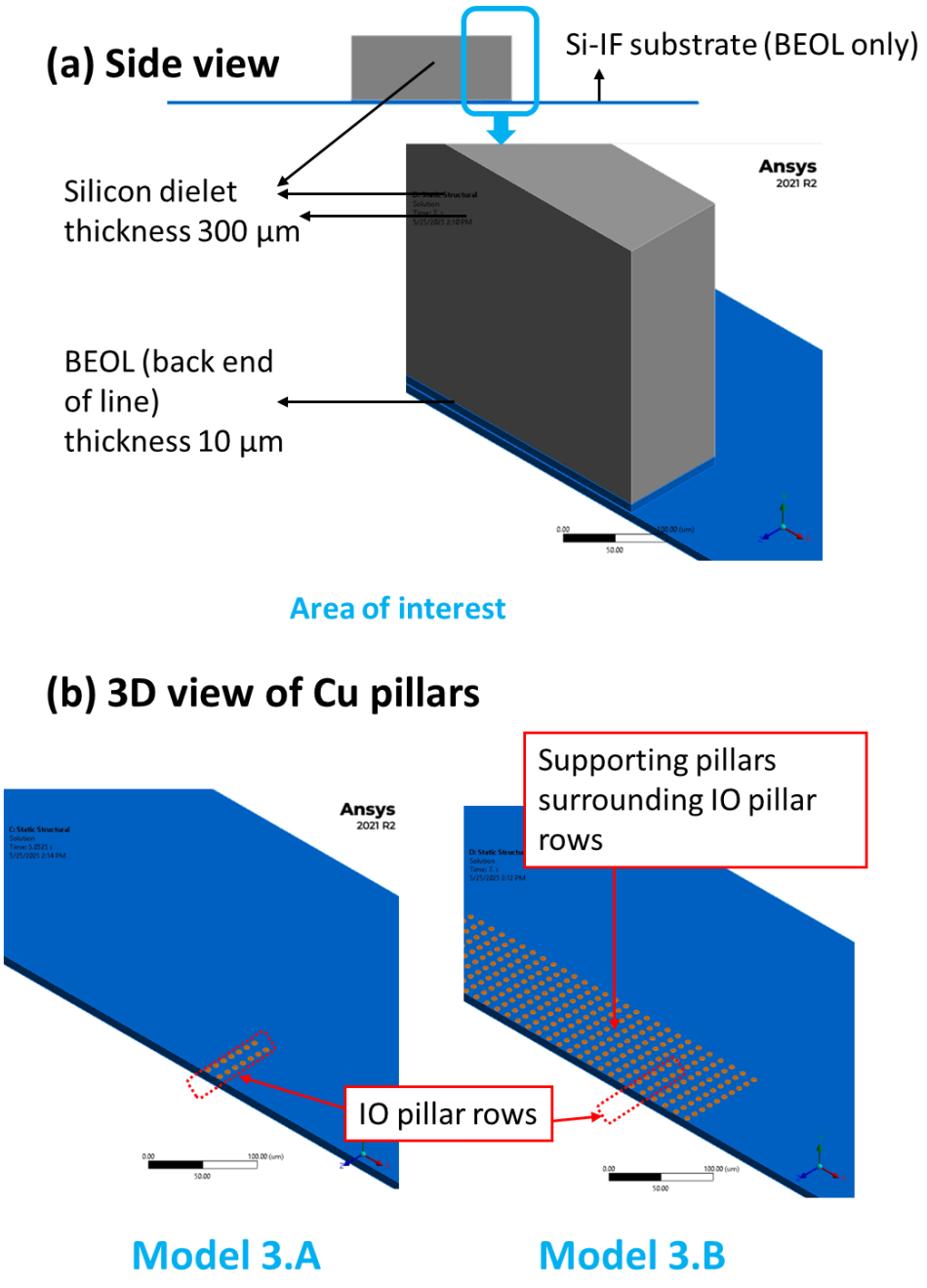
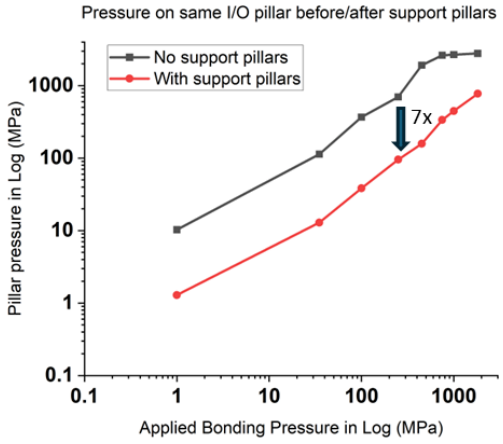
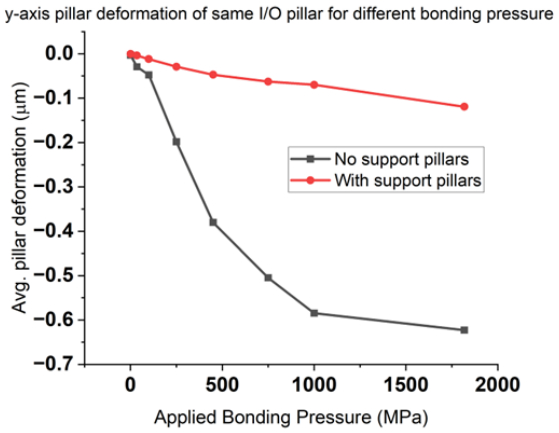


Figure 3.5 (a) ANSYS mechanical simulation to determine the effect of adding support pillars to the Si-IF. (b) model after removing the top die. Model 3.A has only real Cu IO pillars and model 3.B has both real and support pilalrs. Pilars protrude 1μm from the dielectric surface.



(a)



(b)

Figure 3.6 (a) Normal Stress (MPa) on the same IO copper pillar before and after addition of support pillars. (b) reduction in deformation of the IO copper pillar before and after addition of support pillars.

Fig. 3.6 (a) shows the maximum stress on the same Cu pillar after application of bonding pressure. It can be seen at high bonding pressures, the individual IO pillar experiences a maximum local pressure > 2 GPa when the overall bonding pressure on the memory dielet is about 1700 MPa. On the other hand, when supporting pillars are added, the

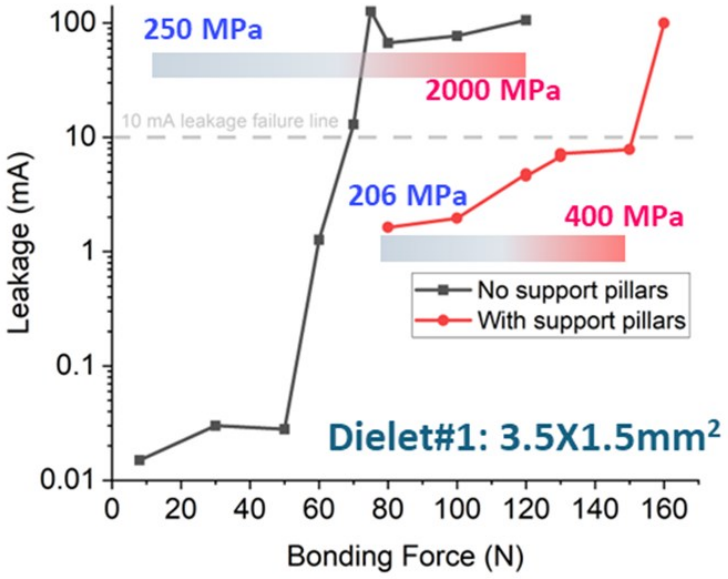


Figure 3.7 Memory die leakage (mA) versus bonding force (N) before and after addition of support pillars. After addition of support pillars, we have higher control of bonding pressure.

individual IO pillar pressure reduces to about 550 MPa which is a 4 times reduction. Also pillar deformation due to extremely high pressure has reduced. At a bonding pressure of 250 MPa, the pillar pressure reduces by ~ 7 times. This gives a much larger design window of bonding pressure without incurring damage to the IO circuitry present underneath IO pads and IO pillars. The dielet is found to warp under application of uniform force from the bond-head around the Cu pillars which act as pincers. The bending of dielet in case of only two rows of protruded Cu pillars is much significant compared to the bending of dielet in case of uniformly distributed Cu pillar density. The amount of deformation of the IO pillar without and with surrounding support pillars is shown in Fig. 3.6 (b).

A direct effect of addition of support pillars can be seen in the leakage measurement of the memory dies in Fig. 3.7. Before adding supporting pillars, the optimization of memory die bonding was extremely difficult, with requirement of high bonding forces (and consequently high bonding pressures) to account for tilt. This would cause the memory dielet to fail frequently, resulting in high leakage between VDD (power) and VSS (ground).

However, after the addition of supporting pillars, a much wider window in terms of bonding force is available, which helps account for tilt, without increasing the bonding pressure significantly during the bonding process. The acceptable bonding pressures to overcome tilt and get pillar imprints on both left and right side of the die are now below 300 MPa.

A rigorous implementation of uniform supporting pillar density in Si-IF packaging is done by inclusion of a new design rule check (DRC) in the Si-IF process manual which requires at least 10 % supporting pillar density in all regions where IO pillars are not present.

3.3.2 Alignment drift across wafer scale assembly

Scaling the Cu-Cu bonding process to assemble an entire 12-inch wafer demands not only high throughput but also precise alignment of each dielet with its corresponding bond site on the silicon substrate. Any misalignment can result in critical failures—either opens, where bonds fail to form, or shorts, which can lead to excessive leakage or device failure.

To achieve a bonding pitch of 10 µm, a stringent misalignment tolerance of $3\sigma \leq \pm 2$ µm is required. Scaling the pitch further down to 5 µm tightens this requirement to $3\sigma \leq \pm 1$ µm. Moreover, this level of alignment accuracy must be consistently maintained across the entire wafer to ensure successful assembly. At UCLA-CHIPS, using the high-throughput Cu-Cu thermocompression bonding (TCB) process described in Chapter 2, we can initially achieve sub-micron misalignment—often ≤1 µm—at the start of the dielet tacking and bonding process. However, alignment drift becomes a concern over time, gradually increasing misalignment and eventually leading to failures such as opens or shorts.

This drift is primarily caused by mechanical shifts in bonding tool components, process-induced variations, and thermal gradients, which can interfere with the vision system's ability to accurately detect dielet and substrate fiducials. An additional challenge stems from the bonding pitch itself. While our work routinely achieves 10 µm pitch bonding, the bonding tool in use—K&S APAMA—was originally designed for a 30 µm pitch with an alignment accuracy limit of ±2 µm. Operating at tighter tolerances pushes the tool beyond its intended specifications, contributing to drift, particularly when bonding hundreds to thousands of dielets across a full wafer.

Given these challenges, we conducted a detailed study to characterize alignment drift across the wafer when targeting a 10 µm bonding pitch and a misalignment tolerance of <2 µm. This analysis is essential for guiding future improvements in tool calibration, process stability, and system-level alignment strategies.

For this study, we designed 3.5 mm X 2.5 mm glass dielets to be assembled on a 4-inch silicon wafer with alignment marks/fiducials. The glass dielets had alignment fiducials fabricated using titanium/gold (Ti/Au) 20nm/200nm liftoff process. Gold was chosen for its resistance to oxidation, ensuring reliable fiducial detection during alignment. The substrate is a 4-inch wafer with alignment fiducials again lifted using Ti/Au 20nm/200nm process to ensure oxidation does not affect the alignment drift process. The glass used to prepare the dielets was transparent borosilicate glass.

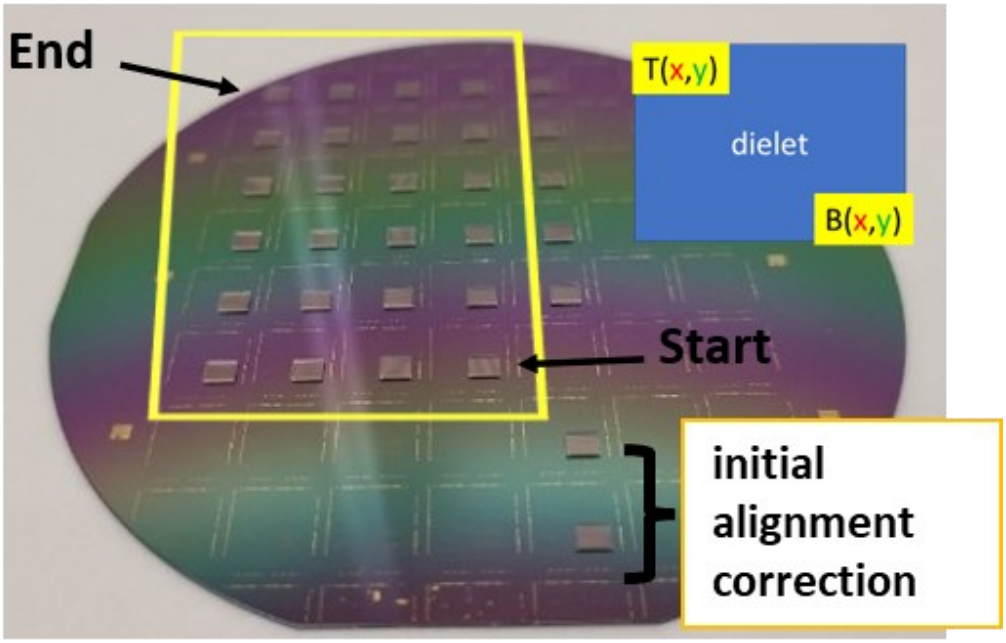


Figure 3.8 Au-Au glass die bonding experiment for measurement of alignment drifts when bonding across a 4-inch wafer.

A total of 24 dielets were bonded onto the wafer with an x- and y-spacing of 10 mm between each dielet. The bonding strategy is illustrated in Fig. 3.8. We used the standard Au-Au thermocompression bonding (TCB) process previously demonstrated by Bajwa et al. [AABaj17]. Bonding was performed from the bottom-right to the top-left corner of the

wafer in a column-by-column manner. During bonding, the bond-head moves only in the y-direction, while x-direction motion is achieved by moving the substrate via the chuck. This motion strategy significantly impacts the alignment behavior, as we discuss below.

Because the dielets are made of transparent glass, we were able to optically measure x- and y-misalignments by observing the alignment marks post-bonding, as shown in Fig. 3.9. We analyzed alignment drift at the top-left and bottom-right fiducials of each dielet to track cumulative drift effects as bonding progressed across the wafer.

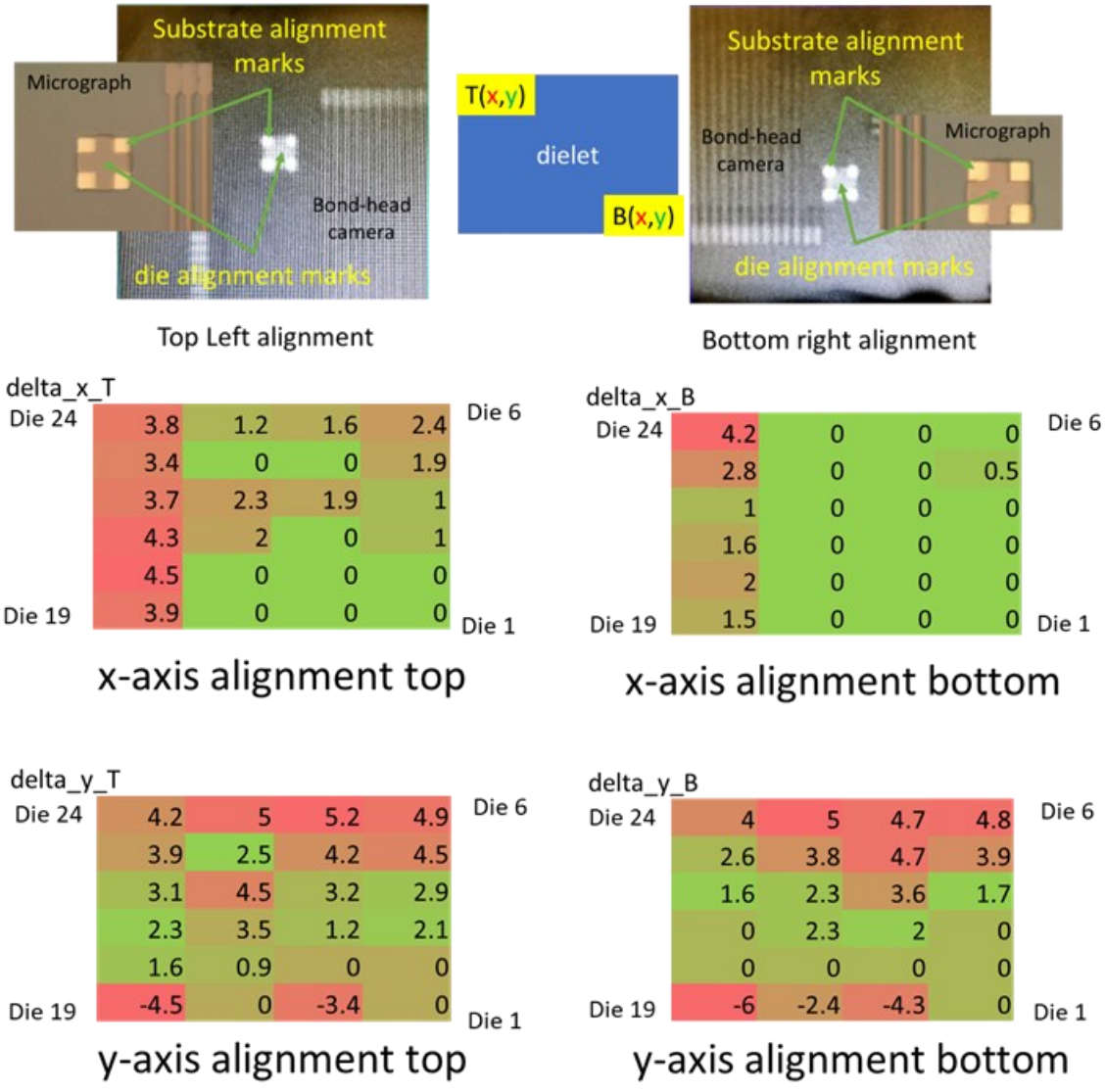


Figure 3.9 Details on calculation of measurement drifts at top left and bottom right alignment marks including values obtained for alignment drift in x- and y- directions.

The results are summarized in Fig. 3.9 tables, which shows heat maps of x- and y-drift at the fiducial locations. A clear trend of y-axis drift was observed, which correlates strongly with the bond-head's motion direction. When the bond-head reverses direction (after bonding six dielets per column), the direction of y-drift also reverses—further confirming this relationship. The maximum drift of approximately 6 μm is observed which is enough to completely misalign the dielet with respect to the substrate and cause unwanted shorts between dielet and substrate when bonding is done at 10μm pitch. Therefore, under current operating conditions, the alignment accuracy of the bonding tool is inadequate for wafer-scale assembly at a 10 μm pitch.

It is important to note, however, that the 6 μm drift is acceptable in a 30 μm solder-based bonding process, which aligns with the original specifications of the K&S APAMA bonding tool.

3.3.3 Improving Alignment Accuracy via. Bond-Head Initialization

To explore potential improvements, we investigated whether periodic reinitialization of the bond-head could mitigate drift. By initializing the bonding system after each dielet is tacked and bonded, we observed a significant improvement in alignment accuracy. Heat maps illustrating the drift with this approach are shown in Fig. 3.10.

The initialization process takes approximately 30 seconds and ideally only needs to be performed once before continuous bonding. However, for this study, initialization was performed before bonding each dielet, which completely reset alignment drift between bonding steps.

While this method successfully restored alignment accuracy, it comes with a significant throughput penalty. To illustrate this -

- Manufacturer-specified bonding process throughput for 30 dielets:
  Initial setup: 30 s
  Bonding 24 dielets: 24 × 10 s = 240 s
  Total time: 270 s

- Modified bonding process with reinitialization before each dielet:

  24 dielets × (30 s initialization + 10 s bonding) = 960 s

  This is a **~3×** increase in total bonding time.

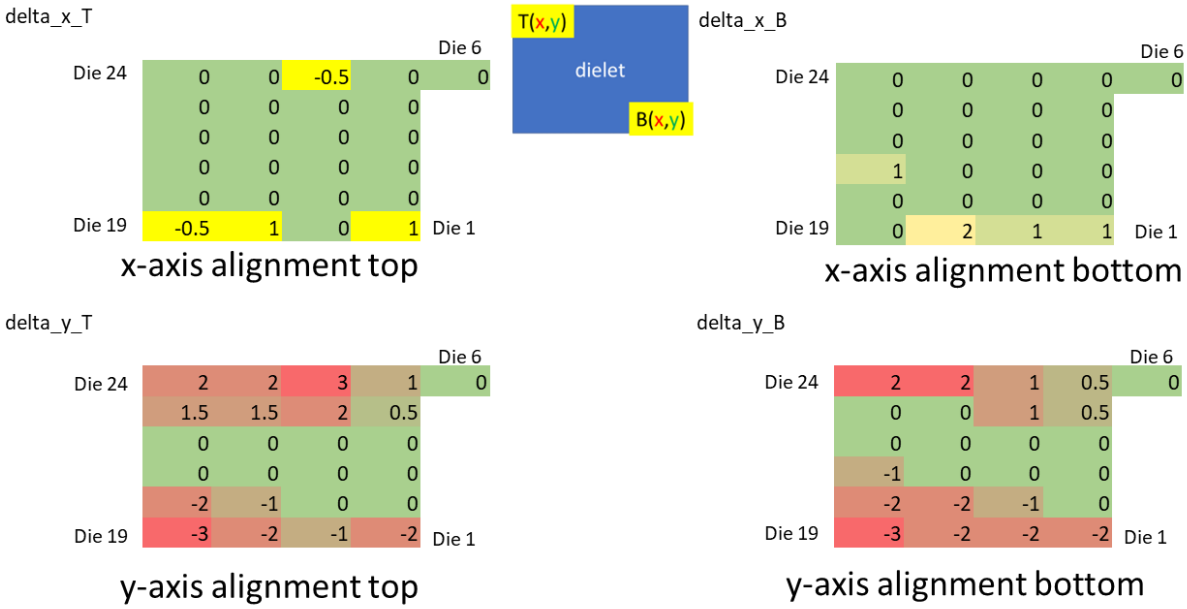


Figure 3.10 Improvement in alignment drift after addition of bond-head initialization step prior to each bond. This improvemnt in alignment accuracy comes at the cost of bonding throughput.

In summary, while bond-head reinitialization offers a practical path to achieving the alignment accuracy needed for 10 µm pitch bonding on a 4-inch wafer, it imposes a significant trade-off in throughput. Future work could explore optimized scheduling strategies, tool upgrades, or partial reinitialization protocols that balance alignment precision with bonding speed to support full wafer-scale assembly under tight pitch constraints.

3.3.4 Effect of mis-alignment on signaling between dielets

The APAMA has a rated misalignment accuracy of ± 2µm and 30 µm bonding pitch [APAMAC2W]. However, it can routinely achieve ~ 1µm alignment accuracy which is reasonable for achieving 10µm bonding pitch in the beginning of the bonding process and for next few dielets, after which we need to recheck misalignment. Nevertheless, we have

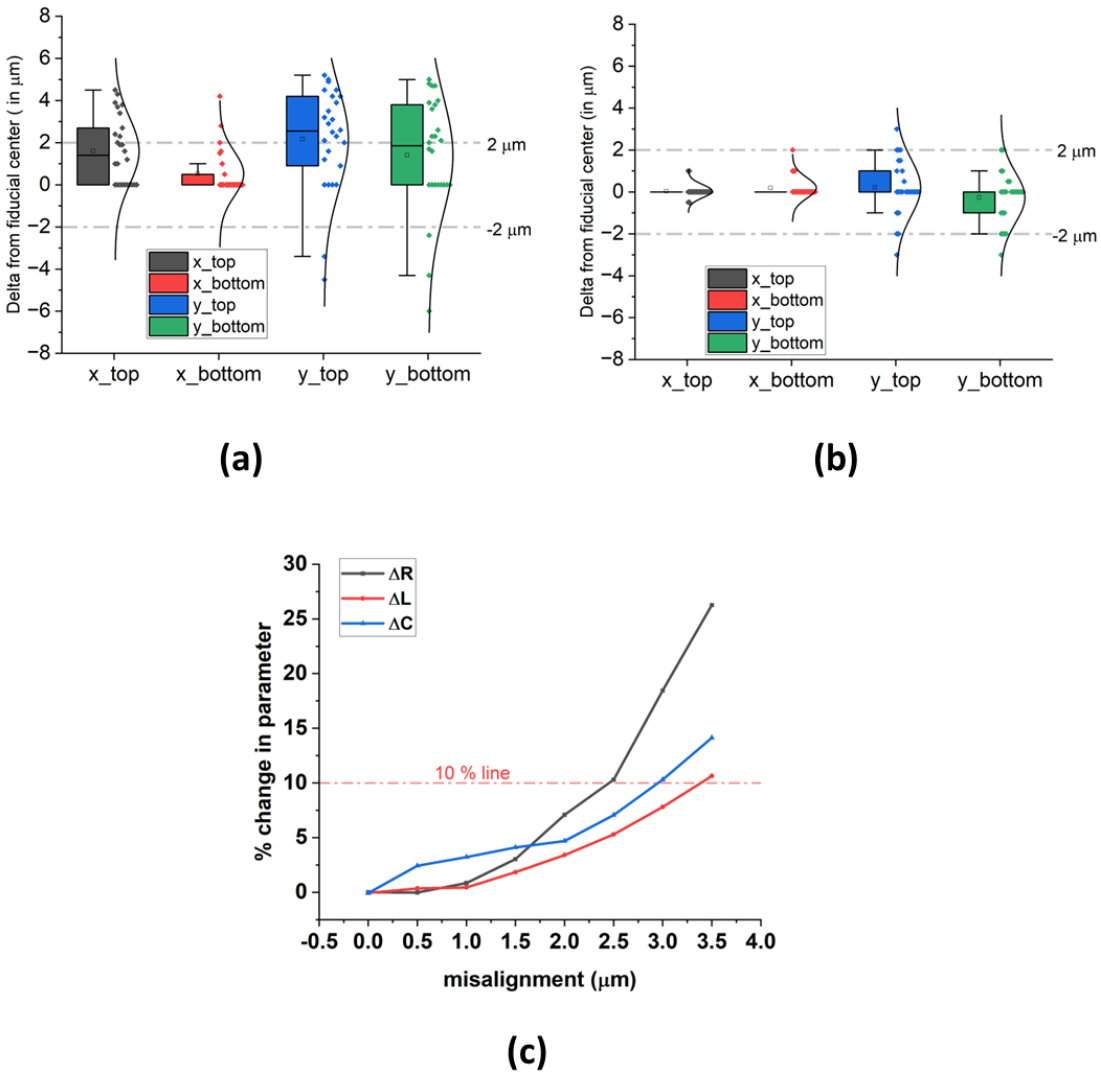


Figure 3.11 Comparison of distribution of alignment drifts for (a) continuous Au-Au bonding (b) Au-Au bonding with bond-head initialization. (c)) variation of resistance, inductance and capacitance of the bonds with misalignment for a signal of 1.1 GHz passing throuh a Cu pillar of diameter 4µm.

explored the change in the pillar parasitics that results from misalignment. Fig. 3.11(c) shows the extracted RLC parameters at 1.1 GHz for a perfectly aligned contact compared with a contact where the Cu pad on the dielet is 2 µm misaligned with respect to the Cu pillar (Φ = 4 µm). Resistance increases with misalignment due to reduction in contact area and current spreading. However, we found that the parasitic inductance and capacitance do not change significantly (<5%). Going beyond a misalignment of 2.4 µm,

however, the changes in RLC parameters of the Si-IF joint are no longer acceptable. Nevertheless, it should not be an issue, as the typical misalignment in our process is ≤ 1 μm.

3.4 Dielet singulation requirements

Singulation is the process of separating individual dielets or chiplets from a larger wafer or panel. It is a critical step in advanced packaging and heterogeneous integration, especially in chiplet-based designs. Dielet singulation/dicing method can have significant effect on assembly yield. The key goal of dicing singulation process is to ensure the dielets can be separated from one another without damage /crack to the active silicon region on the die and without leaving much residue behind due to the singulation process. Several approaches to dielet singulation exist in the form of blade dicing, laser-assisted blade dicing, stealth dicing, plasma dicing. The key differences are presented in the table below:

Table 3.2 Comparison of different dicing approaches.

Dicing approach	Methodology	Pros	Cons
Blade dicing	Uses a diamond tipped circular saw to cut the wafer	1. Mature, widely used technology 2. Low cost of processing	1. Mechanical stress can cause the dielets to crack. Also causes chipping at dielet edges. 2. Throughput can be low if there are a lot of cuts to be made.
Laser dicing	Uses a laser beam to ablate silicon and separate dielets on a wafer	1. Higher precision	1. Ablated silicon can re-deposit ton the dielet edges causing significant

			silicon debris at edge. 2. Has higher equipment cost.
Stealth dicing	Uses a laser to create internal damage along silicon planes along which then the wafer can be split	1. No visible Kerf 2. Minimizes debris at the edge 3. Good for brittle materials	1. Slower process 2. Higher process complexity
Plasma dicing	Uses plasma etching to separate dielets without mechanical stress	1. Low stress on the dielets during cut 2. Minimal to no chipping 3. Helpful for ultra-thin wafers 4. Very high throughput of cutting (single cut can separate dielets across wafer.)	1. Requires expensive equipment. 2. Less mature though industry adoption is accelerated due to its requirement in Hybrid bonding.

As anticipated, significant chipping or silicon debris at the dielet edge can critically compromise the quality of Cu-Cu thermal compression bonding (TCB). Silicon debris not only introduces particulate contamination during the bonding process but can also obstruct the physical contact between Cu pads on the dielet and those on the substrate. This interference can lead to poor or open electrical and mechanical connections leading to assembly failure.

Several techniques can be employed to mitigate dicing-related damage and Si-debris at die edge. For example, ultra-sonication cleaning helps to remove loosely attached

particles on the dielet edge, especially in case of blade dicing and stealth dicing. However, plasma dicing stands out as the most favorable method for singulation. It produces minimal mechanical stress during dicing and virtually eliminates edge chipping, resulting in the cleanest dielet edges thereby significantly improving the assembly yield. In fact, in case of Cu/dielectric hybrid bonding, plasma dicing is considered mandatory [YLin24] to ensure proper dielectric-dielectric bonding as even a single particle at the dielectric bonding interface can cause hybrid bonding to fail. In contrast, some dicing approaches, especially laser dicing which work on the principle of ablation, are notorious for leaving un-removable silicon debris at the dielet edge. These are very difficult to bond as the debris can often protrude along the silicon edge and prevent Cu-Cu contact during bonding.

A clear example illustrating the unsuitability of laser dicing for fine-pitch Cu-Cu TCB is presented in this work. The Fig. 3.12 shows a dicing edge from one of the samples being bonded at UCLA-CHIPS which was found to have silicon debris sticking out of the dielet edges. Despite implementing a pillar recess of 1 μm to 2 μm to compensate, a high assembly yield was not achieved. A cross-sectional image reveals how the debris contacts the bond site, exerting mechanical force against the Cu pad and adjacent dielectric, thereby preventing proper Cu-Cu interface formation. Notably, as we move

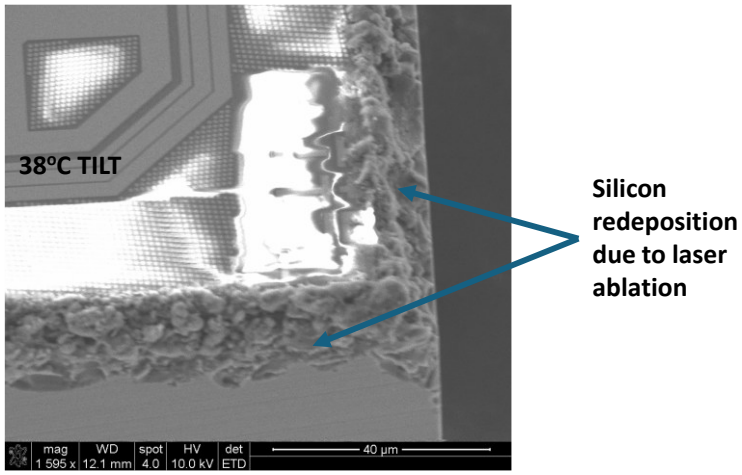


Figure 3.12 A foundry dielet with laser assisted dicing. Due to laser ablation process, silicon at dicing street has redeposited on the edges creating challenge for thermal compression bonding.

inward from the dielet edge, full Cu-Cu contact is restored, indicating that the issue is localized to the dicing edge region. These results make it evident that, with laser dicing, we cannot guarantee a proper Cu-Cu bond at the dielet edges, and the assembly yield suffers.

While the preferable solution to mitigate edge debris/residue is to switch to a cleaner dicing method, this is not feasible for dielets that have already been processed using laser dicing. Fortunately, we have developed an innovative post-processing technique to address this challenge. Since the debris present at the dielet edge is predominantly silicon, we proposed the use of isotropic silicon wet etching to selectively remove it. A good example of isotropic silicon etchant is HNA solution i.e. hydrofluoric acid – nitric acid – acetic acid solution. As discussed in chapter 1 and chapter 2, all dielets to be bonded to Silicon interconnect fabric (Si-IF) are terminated with metal pads but then passivated with 150 nm of Silicon Nitride (Si_3N_4). Therefore, we exposed the dielets with nitride passivation to 1:3:5 HNA solution i.e. a 1 parts hydrofluoric acid (HF) to 3 parts Nitric acid (HNO_3) to 5 parts acetic acid (CH_3COOH) solution. A small-scale design of experiments for different etch times of 20 seconds and 1 minute was done to observe the removal of Si debris at edge and determine if the passivation is failing at some point in the etch.

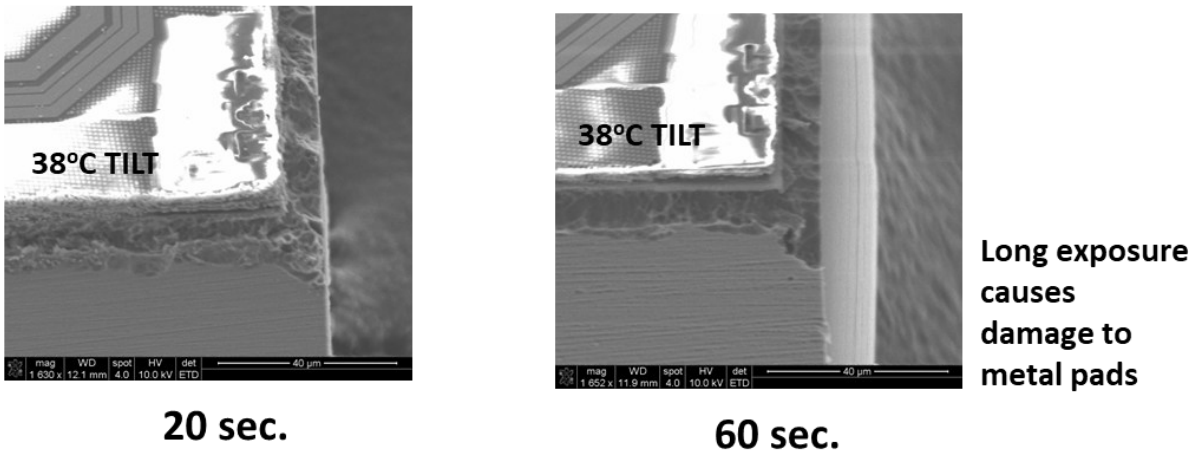


Figure 3.13 Use of HNA solution ($HF + HNO_3 + CH_3COOH$) for isotropic etching of silicon at the dielet edge. The dielet is passivated with thin (150nm) silicon nitride passivation. 20 sec of HNA etch is found to be adequate in removing silicon damage at dielet edge.

The results are shown in Fig. 3.13. We observe that even with 20 seconds of etch time, we can effectively remove a majority silicon debris from the die edge. Increasing the etch time to 1 minute further improved the dicing edge, however, the Cu pads beneath the nitride passivation begin to oxidize, (which as visually observed) and therefore, the passivation can fail with prolonged exposure to HNA.

Based on these observations, we conclude that for laser-diced, nitride-passivated dielets, a 20-second HNA etch is a critical post-processing step. It significantly reduces edge debris and thereby enhances the Cu-Cu bonding quality and overall assembly yield, while avoiding damage to the underlying metallization.

3.5 Effect of Cu pad dishing on bonding

The Cu pillars/pads for Cu-Cu TCB are usually fabricated using Cu damascene process. The roughness and dishing on these pad/pillars are crucial for bonding. High roughness reduces the contact area between the bonding interfaces, and causes uneven pressure distribution during bonding, which can lead to stress concentrations and potentially affect the quality of bond. A high dishing value on the Cu pads/pillar surfaces can lead to creation of voids during bonding, degrading the quality of bond. An example of a high dishing

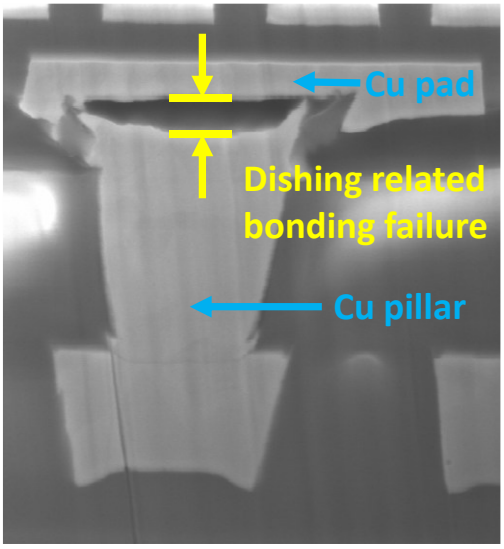


Figure 3.14 Bonding failure due to significant dishing damage during chemical mechanical polishing (CMP). We suggest a dishing < 50nm for Cu-Cu TCB.

leading to bonding failure in a Cu-Cu TCB can be seen in Fig. 3.14. While these issues can be mitigated by applying high bonding pressure and temperature as shown by Ambhore et al. [Pranav] which can flatten the surfaces enough to get a good Cu-Cu bond, the applied pressure must be moderated to avoid any potential damage of the dielectrics below the bonding pad. This is a process related challenge and requires optimization of chemical mechanical polishing (CMP) process to reduce both roughness and Cu dishing. Cu dishing also depends on the density of copper structures in each damascene layer. In Si-IF fabrication, a mandatory density rule is imposed which requires the polished Cu pad/pillar density to meet a 20% to 60% requirement for polishing. Pillars/pads may be selectively etched out to reduce the contact area with the dielet to be bonded to modify the bonding pressure. This is done if required by the dielet designer/manufacturer.

3.6 Die/dielet design-related considerations prior to tape-out which can improve Cu-Cu assembly and testing post-assembly

So far, we have examined how bonding tool configurations and bonding process parameters can significantly influence assembly quality. However, dielet design decisions made prior to tape-out also play a crucial role in determining the overall assembly yield. Based on our experience with die-to-wafer Cu-Cu thermal compression bonding (TCB) of functional dielets, we offer the following design recommendations to assist future designers—particularly those targeting wafer-scale integration—in achieving higher bonding success rates:

- Incorporate active and passive test structures: Include active and passive test structures which can be electrically characterized after bonding has been completed. Simple passive daisy chains are extremely valuable for assessing post-bonding alignment and connectivity. Meanwhile, active silicon structures such as ring oscillators or buffered daisy chain links enable functional verification of logic and signal paths. This has become especially important as dielet 3D-stacking such as memory-on-logic or logic-on-memory stacking is rapidly becoming a preferred

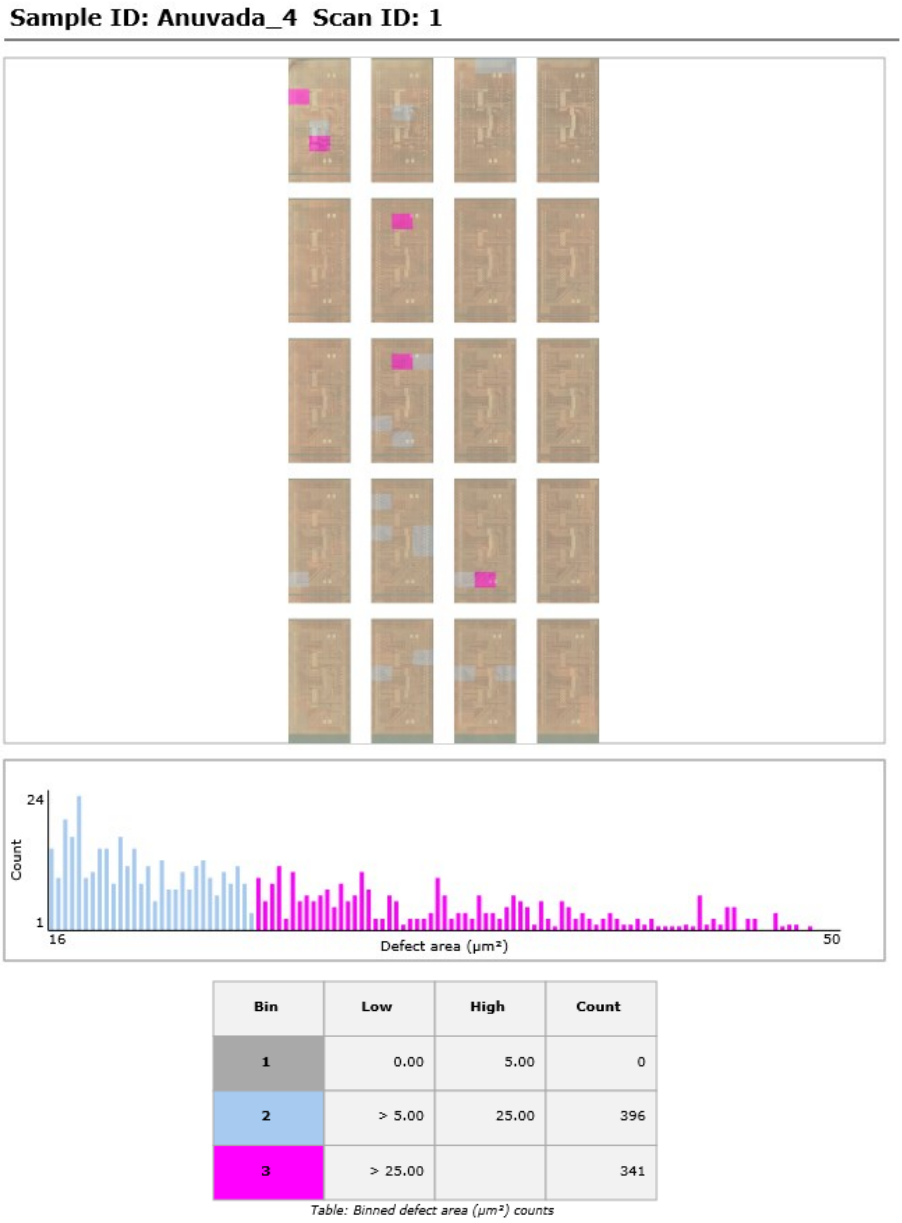
approach for reducing processor to memory latency, processor-to-memory bandwidth etc.

- Maintain comprehensive dielet-level and test documentation: Proper dielet level and dielet-test documentation is crucial irrespective of bonding process used. Lack of documentation can often make debugging difficult in case of assembly related issues.
- Ensure uniform Cu bonding pad distribution: Dielet designs are strongly recommended to have a uniform Cu bonding pad distribution. These Cu pads may be floating or preferably connect to a power mesh if possible. This can significantly reduce damage in case there is a tilt between bond-head and chuck in the die-to-wafer bonding tool.

3.7 Known Good Si-IF Substrate (KGS) fabrication considerations for improving substrate yield for assembly

In addition to selection of known good dies (KGD), we have also developed a known good Si-IF approach to ensure the best Si-IFs were used for the assembly. As described in reference [SJanD20], Cu wires are fabricated on Si-IF substrate using BEOL Cu damascene or dual damascene process flow. Post fabrication of each metal layer, Nanotronics defect identification equipment [NANOTRONICS] was used to identify defects, if any, produced during fabrication. The camera on the Nanotronics scans the entire wafer to take multiple images of the Si-IF sites being fabricated. The focusing algorithm was optimized to attain clarity in the images taken while maintaining an acceptable throughput. The tool combines all the image data acquired to create a golden template of the Si-IF site. Using information from the mask file (GDS) as well as the golden template, defects produced during fabrication can be identified with high precision enhanced by Artificial Intelligence/Machine Learning (AI/ML) methods.

To verify that the defect detection procedure was able to detect any defects that would be catastrophic to functionality of the assembly, Si-IF samples with two metal layers and



Bin	Low	High	Count
1	0.00	5.00	0
2	> 5.00	25.00	396
3	> 25.00		341

Table: Binned defect area (µm²) counts

Figure 3.15 Identification of defects in Si-IF sample using Nanotronics Microscope.

intentionally added defects were tested. These samples would mimic the Si-IF process flow but would never be used for real assembly. During the preparation of this Si-IF some of the wafer cleaning steps were intentionally omitted to add defects. The Fig. 2.15 shows the defect detection micrograph when the sample is scanned by Nanotronics using modified defect detection parameters. This sample has twenty spanning over an effective area of 1100 mm^2. During sample scan, the large scan area is segmented into smaller

areas in which the number of defects of a given size counted. A color is applied to each area depending on defect count binning parameters. A defect histogram is also shown in Fig. 3.15. The analysis was able to detect defects as expected. Hence, for all the Si-IF samples, Nanotronics defect detection is now an intrinsic part of the Si-IF inspection process.

3.7 Conclusion

In this chapter, we discuss some of the factors beyond Cu-Cu TCB which can affect the assembly yield. These include bonding tool related parameters such as bond-head tilt and alignment drift; process parameters such as dielet singulation strategy, Cu pillar distribution across bond-site, Cu pillar dishing and roughness; and dielet design factors such as availability of proper test structures, maintaining proper documentation and having uniform Cu pad density even for non-IO pads. The objective is to provide practical guidance to chiplet/dielet designers and package engineers in developing robust pad layouts and design strategies that enhance post-bonding assembly yields. While the experiments done and observations presented here are in context of die-to-wafer/substrate Cu-Cu TCB, the insights are expected to be broadly applicable to other flip-chip bonding technologies, including hybrid bonding and solder-based flip-chip assembly.

Chapter 3 references

[KSah24] K. Sahoo, V. Harish, H. Ren, and S. S. Iyer, “A Review of Die-to-Die, Die-to-Substrate and Die-to-Wafer Heterogeneous Integration,” *IEEE Electron Devices Reviews*, pp. 1–28, 2024, doi: 10.1109/EDR.2024.3501214.

[SJan19] S. Jangam *et al.*, “Fine-Pitch (≤10 μm) Direct Cu-Cu Interconnects Using In-Situ Formic Acid Vapor Treatment,” in *2019 IEEE 69th Electronic Components and Technology Conference (ECTC)*, 2019, pp. 620–627. doi: 10.1109/ECTC.2019.00099.

[DMay22] D. May, J. Heilmann, E. Boschman, M. Abo Ras, and B. Wunderle, “Tilt- and warpage measurement as inline quality assessment tool,” *Microelectronics Reliability*, vol. 135, p. 114601, Aug. 2022, doi: 10.1016/j.microrel.2022.114601.

[SpalD21] Saptadeep Pal's thesis [online] available: https://s3.us-west-1.amazonaws.com/chips.user.media/page_file/_Saptadeep_Pal.pdf

[AABaj17] A. A. Bajwa *et al.*, "Heterogeneous Integration at Fine Pitch (≤ 10 μm) Using Thermal Compression Bonding," in *2017 IEEE 67th Electronic Components and Technology Conference (ECTC)*, 2017, pp. 1276–1284. doi: 10.1109/ECTC.2017.240.

[APAMAC2W] Kulicke and Soffa, Apama C2W tool, [online]. Available: https://www.kns.com/Products/Equipment/Advanced-Solutions-(TCB)/APAMA-C2W

[YLin24] Y. Lin *et al.*, "Direct Die-to-Wafer Hybrid Bonding Using Plasma Diced Dies and Bond Pad Pitch Scaling Down to 2 μm," in *2024 IEEE 74th Electronic Components and Technology Conference (ECTC)*, May 2024, pp. 40–44. doi: 10.1109/ECTC51529.2024.00015.

[NANOTRONICS] Nanotronics nSpec defect analysis equipment [online], available: https://nanotronics.ai/product/nspec-ls

[SJanD20] SivaChandra Jangam's thesis, [online] available: https://s3.us-west-1.amazonaws.com/chips.user.media/page_file/Jangam_dissertation_v6+Printed.pdf

Chapter 4 Chip disaggregation and die-to-die signaling for heterogeneous integration

4.1 The need for chip disaggregation

The exponential growth of data—from petabytes to zettabytes—has driven a corresponding demand for significantly greater computational power and higher compute-to-memory bandwidth. To harness the performance benefits of advanced semiconductor technology nodes at 5 nm and below, designers are striving to integrate as much compute and memory as possible within a single package. This has led to a dramatic increase in monolithic chip sizes, in some cases reaching the maximum reticle dimensions allowed by lithography steppers (reticle limits).

However, building large monolithic chips presents substantial challenges. As technology nodes scale, the complexity, cost, and risk associated with design, verification, and manufacturing grow disproportionately [AnySil]. The design process involves multiple stages—RTL simulation, post-synthesis, post-place-and-route, and parasitic extraction—each requiring extensive resources and time. Formal verification and late-stage engineering change orders (ECOs) further extend development cycles, particularly for large designs, negatively impacting time-to-market. Manufacturing large chips at leading-edge nodes is not only technically demanding but also prohibitively expensive. Yields at nodes like 2 nm are often below 70% [emb25], significantly increasing the cost per functional die. Packaging these large chips introduces additional difficulties, including die handling constraints, die and substrate warpage, and coefficient of thermal expansion (CTE) mismatches between different materials used in the package stack.

As illustrated in Fig. 4.1, adapted from Iyer et al. [SIyer19], increasing chip size exacerbates issues across the board—from die handling and IP reuse to testing and cost. Testing monolithic chips post-packaging is particularly challenging due to the sheer

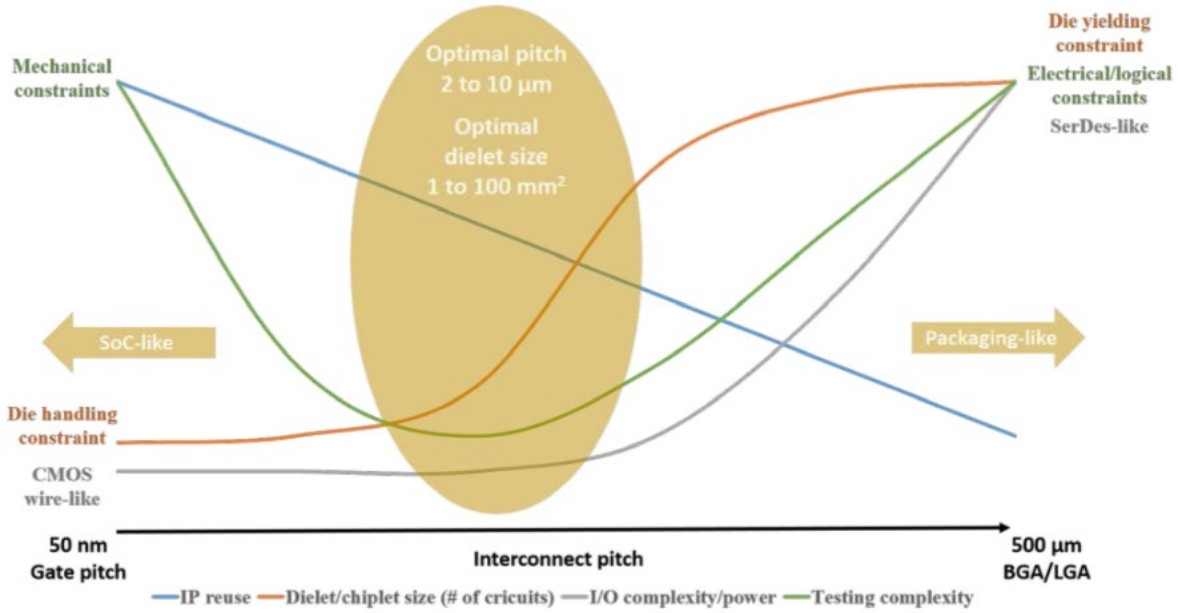


Figure 4.1 Design space for determination of optimal chiplet size and interconnect pitches on silicon interconnect fabric as reported in [Iyer16]. As shown, several factors including die handling constraints, mechanical constraints, yield constraints, electrical testing and IP (chiplet intellectual property) constraints determine optimal chiplet size. This chapter will introduce a partitioning figure of merit to help with determination of ideal size given certain metrics.

number of integrated IP blocks. Furthermore, advanced packaging techniques such as copper-to-copper (Cu-Cu) thermal compression bonding or hybrid bonding are non-reworkable. If the final assembly fails functional testing, the entire package must be scrapped—an outcome that is far more costly when larger dies are involved. In such scenarios, even a single re-spin becomes a critically evaluated and costly decision.

These limitations highlight the growing need for chip disaggregation, where functionality is partitioned across multiple smaller chiplets. This approach enables better yield, cost-efficiency, and flexibility, while addressing the scalability bottlenecks of traditional monolithic chip design.

4.2 Chip disaggregation process into chiplets/dielets

Chiplets offer a new approach to building System-on-Chips (SoCs) which can greatly improve yields and reduce design and manufacturing costs when compared to monolithic chips [GMou16], [Cade], [RPFil], [SIyer19], [PGup19]. Chiplets are optimized to do a specific function extremely well – which includes logic, memory, input/output (I/O), power, signal processing and more. Therefore, they offer heterogeneity in terms of design, and functionality, with high manufacturing yields. Different chiplets can be mixed and matched to create highly customized SoCs, without the need to fit all functions into a monolithic chip.

The chiplet paradigm involves partitioning a monolithic chip into modular, well-defined functional components. These components use standardized interfaces for communicating with other chiplets to form a complete system [KSahEPEPS]. The goal of such interfaces is to achieve on-chip like bandwidths in the order of those between adjacent processing cores or core and memory on the same chip. However, unlike within a chip, where wire pitches can be in tens of nanometers (nm) or below, terminal pitches between chiplets are much higher (few microns to tens of microns) due to I/O design and assembly constraints [SIyer19], [KSahEPEPS], [SJansFOM].

This shift from monolithic to chiplet-based design introduces trade-offs in system-level performance, power efficiency, and area. Understanding these trade-offs is crucial for system designers striving to create optimized chiplet-based architectures. This chapter aims to equip designers with the tools and insights needed to make informed partitioning decisions.

To that end, we introduce a **Partitioning Figure of Merit [EPTC] (p-FOM)**—a metric to guide chip architects in determining optimal chiplet sizes under various design and system constraints. The chapter is structured as follows:

- **Section 4.3** introduces the derivation of the partitioning figure of merit (p-FOM) and outlines its role in chiplet system design.

- **Section 4.4** explores case studies demonstrating how p-FOM can be applied to real-world partitioning problems.
- **Section 4.5** provides an in-depth discussion on p-FOM, including practical considerations and limitations.

From **Section 4.6** onward, the focus shifts to die-to-die communication, where we present the Signaling Figure of Merit (s-FOM$_{UCLA}$) [SJansFOM], [KSah22]. We detail the critical parameters affecting die-to-die signaling and expand on the original s-FOMUCLA metric proposed by Jangam et al. by incorporating additional factors such as bit error rate and relative packaging cost, culminating in the development of a comprehensive cost function [KSahEPEPS].

- **Section 4.6** introduces cost function for selection of D2D PHY with electrical, chip design and packaging constraints and applies this cost function to a processor-to-memory package example to evaluate package choices based on performance and cost-effectiveness.
- **Section 4.8** concludes the chapter by summarizing key insights and takeaways for chiplet-based system design.

4.3 Derivation of partitioning figure of merit (p-FOM) to assess the performance vs. cost relationship for chiplets

In this section we investigate the process of partitioning a monolithic chip into a system of interconnected chiplets mimicking the chip, including the tradeoffs associated with this process. Let us consider the problem of partitioning a square monolithic chip of area ‘A’ into densely interconnected ‘m’ identical square chiplets. In the monolithic chip, we assume we have wires going either in horizontal or vertical direction (Manhattan routing

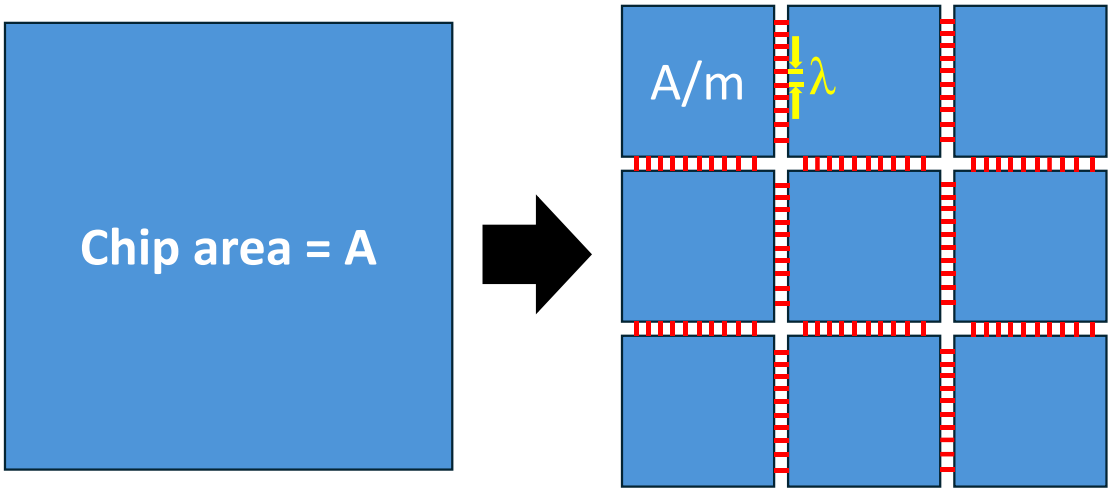


Figure 4.2 Partitioning of a monolithic chiplet of area 'A' into smaller identical 'm' chiplets interconnected with terminals at pitch λ.

scheme [8]) at near infinitesimal wiring pitch. However, when we partition the chip into smaller chiplets, we create terminals at the chiplet boundary to transfer information between chiplets, as seen in Fig. 4.2. These terminals are at a finite pitch of 'λ'. The terminals are driven by transceivers and have a finite area and power associated with inter-chiplet data transfer. We consider the following parameters as inputs for our analysis:

- Bandwidth generated by chip per unit area (bits/sec) = b
- Number of chiplets = m
- Terminal pitch (in m) = λ
- Area occupied by each terminal (in m^2) = λ^2 (i.e.., the I/O is fully contained withing the terminal)

Area of each chiplet $= \frac{A}{m}$

Let us assume as we partition into chiplets, terminals are created on all 4 sides of the chiplet

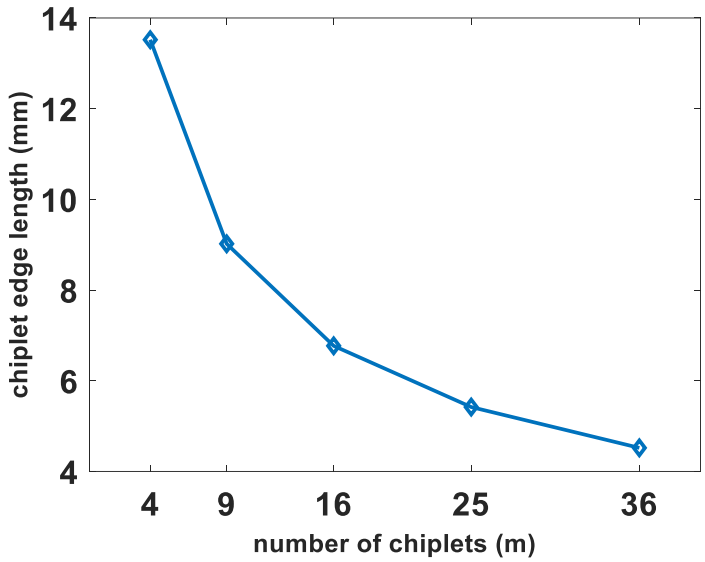


Figure 4.3 The chiplet size relation with increase in the number of chiplets/partitions (m).

Then, number of terminals per chiplet $= \frac{4}{\lambda}\sqrt{\frac{A}{m}}$

Now bandwidth per chiplet = bandwidth per unit area × area of chiplet $= b \times \frac{A}{m}$

As terminals are created on all 4 sides for inter-chiplet communication, and this will lead to an increase in power consumed by the chiplet as well as an increase in area of the chiplet (if we keep the compute performance constant). Our goal is to determine this increase in power and increase in area of the system incurred due to the partitioning process.

First, we look at the increase in power consumed due to addition of terminals for inter-chiplet communication. We assume that as in the original chip some terminals are input, and others are output. To first order we assume that overheads such as clock forwarding, and such are included.

Data-rate per terminal $= f_T = \frac{\text{bandwidth per chiplet}}{\text{terminals per chiplet}} = \frac{\lambda b}{4}\sqrt{\frac{A}{m}}$

that total power consumed by transceivers during switching can be given by the expression:

= dynamic power + static power

= switching power + short-circuit power + leakage power

Where:

Switching power = power dissipated during charging and discharging of the load capacitance ($\mathrm{C_L}$) of the terminal during switching

Short-circuit power = power dissipated during the instantaneous short-circuit connection between supply voltage and ground during gate switching operation (assuming CMOS design).

Leakage power = continuous dissipation due to leakage currents. Sub-threshold leakage is a function of supply voltage $\mathrm{V_{DD}}$, threshold voltage $\mathrm{V_{th}}$, the transistor size (W/L) and temperature T

To simplify the analysis, we assume the power per terminal during communication is given by

power per terminal = $\mathrm{P_T} = \alpha \mathrm{C_L V_{DD}^2 f_T^q}$

where $1 < \mathrm{q} \leq 3$, depending on technology and device architecture.

Therefore, the increase in terminal power due to chiplet disaggregation of large monolithic chip –

$$\Delta \mathrm{P} = \text{number of chiplets} \times \text{terminals per chiplet} \times \text{power per terminal}$$

Which can be further simplified as

$$\Delta \mathrm{P} = \alpha \mathrm{C_L V_{DD}^2} 4^{(1-q)} b^q \times \lambda^{(q-1)} \frac{A^{\left(\frac{q+1}{2}\right)}}{m^{\left(\frac{q-1}{2}\right)}} \qquad (4.A)$$

The increase in area due to added terminals:

$$\Delta A = \text{number of chiplets} \times \text{terminal per chiplet} \times \text{area per terminal}$$

Which is simplified as

$$\Delta A = 4\lambda\sqrt{Am} \qquad (4.B)$$

Another important parameter to consider is the yield of the die manufacturing. Popular yield models used by chip manufacturers are Poisson model, Murphy model, Seeds Model, Bose-Einstein model etc. which are described in detail in [9]. For this work, we consider the Bose-Einstein yield model [9], [10] which is given by

Yield = $\left(\frac{1}{1+d_0 A}\right)^N$

d_0= defect density on critical mask layer

A = area of the chiplet

N = number of critical mask layers

So, for an individual chiplet of total area $\frac{A}{m} + 4\lambda\sqrt{\frac{A}{m}}$ the yield is

$$Yield = \left(\frac{1}{1 + d_0\left(\frac{A}{m} + 4\lambda\sqrt{\frac{A}{m}}\right)}\right)^N \qquad (4.C)$$

4.3.1 Partitioning figure of merit (p-FOM)

To analyze the impact of partitioning a monolithic chip into smaller chiplets, we need to define a performance-to-cost figure of merit. It should consider the performance, power and area (PPA) of adding terminals and account for chiplet yields. This figure of merit can be defined as –

$$\text{p-FOM} = \frac{\text{performance}}{\text{power overhead} \times \text{area overhead}} \times \text{Yield}$$

In our case the performance of the monolithic chip is identical to the performance of the system of interconnected ‘m’ chiplets. So, performance is constant.

$$\text{p-FOM} = \frac{b \times A}{\alpha C_L V_{DD}^2 4^{(1-q)} b^q \times \lambda^{(q-1)} \frac{A^{\left(\frac{q+1}{2}\right)}}{m^{\left(\frac{q-1}{2}\right)}} \times 4\lambda\sqrt{Am}} \times \left(\frac{1}{1+d_0\left(\frac{A}{m}+4\lambda\sqrt{\frac{A}{m}}\right)} \right)^N$$

$$equation\ (4.D)$$

4.4 Effect of partitioning in different cases

We study the effect of partitioning in different cases. We want to point out that minimum number of partitions cannot be 1 i.e. m ≠ 1, because it essentially represents a case of no partitions which is the monolithic chip. If there are no partitions, there are no terminals created. This case represents a singularity and out treatment will be invalid.

4.4.1 p-FOM for a fixed terminal pitch

Let us consider the case where we fix the terminal pitch λ = 10μm. Fig. 4.4 shows the plots obtained for power, area, yield and figure of merit for different number of chiplets. Our inferences are as follows –

- Fig. 4.4 (a) shows the total power added by terminals reduces as number of chiplets/partitions (m) increase. As the monolithic chip is partitioned into greater number of chiplets (m), more terminals are created and the data-rate required per terminal decreases. Given power is proportional to (data-rate)^q, overall this leads to a decrease in total power dissipated by terminals as data-rate per terminal reduces. But eventually the incremental amount of power saved by increasing the number of partitions diminishes at higher number of partitions. It should also be noted that, in the case m = 1, as stated earlier, λ = 0 (no terminals are created, no area or power is added), and therefore m = 1 is not a valid condition for analysis.

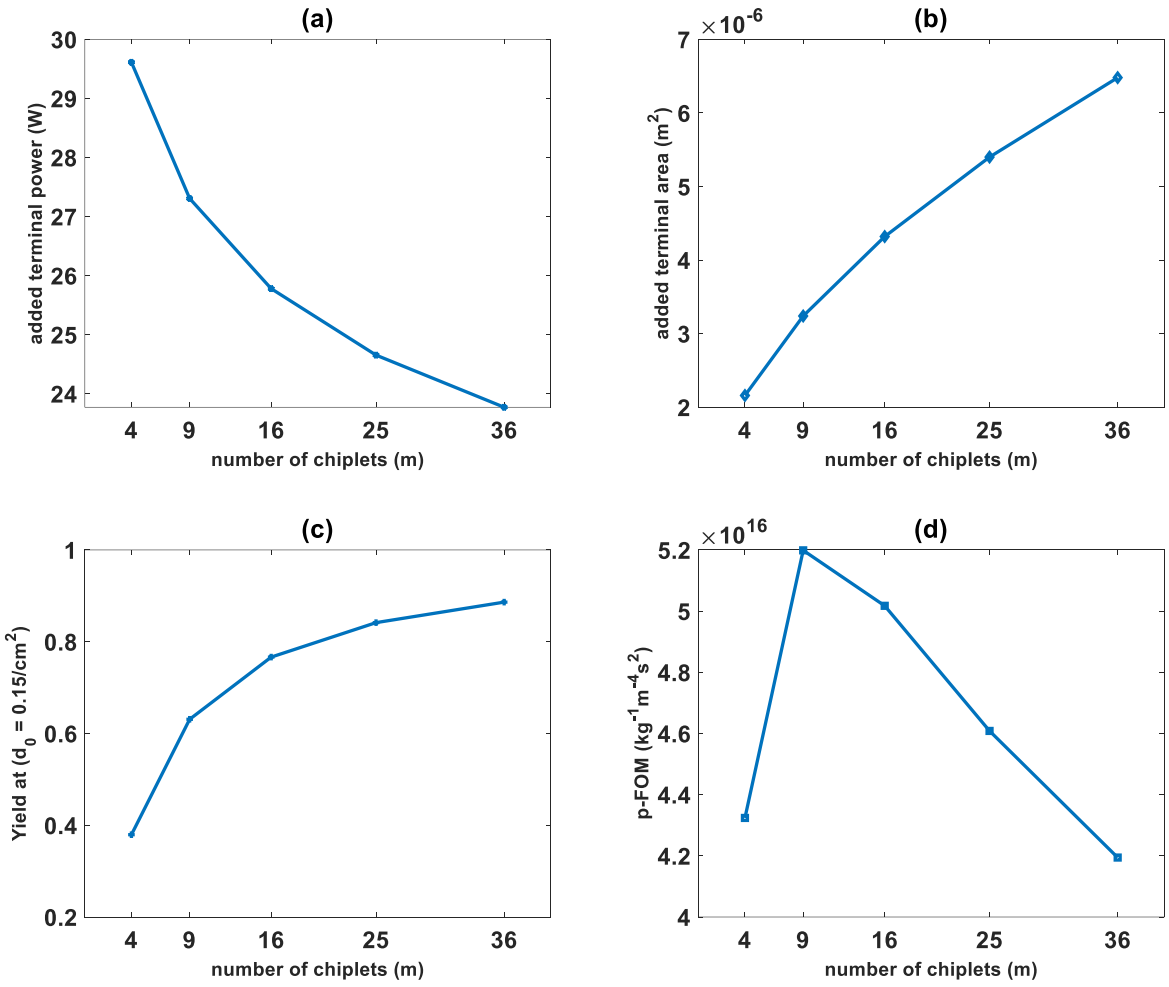


Figure 4.4 Partitioning of a 729 mm^2 monolithic chiplet into 'm' chiplets for terminal pitch λ = 10 µm. (a) shows the total terminal power variation with number of partitions or chiplets (b) shows terminal area variation with number of partitions or chiplets (c) shows yield variation with number of chiplets. Chiplet size reduces as number of chiplets increase. (d) partitioning figure of merit (p-FOM) variation with number of partitions or chiplets. The figure of merit has a maximum when partitioning the chip into 9 chiplets, each of size ~ 9 mm × 9 mm.

- Fig. 4.4 (b) shows the increase in area due to addition of terminals. It is important to note that infinite partitions of the chip are not possible, because in that situation, chiplet edge length < λ, there are no terminals per chiplet. So, a limiting factor of chiplet edge

length > λ is required. However, we do not expect such a situation to arise as leads to poor figure of merit.

- Fig. 4.4 (c) shows increase in yield with number of chiplets/ partitions (m). As m increases, the chiplet edge length reduces per Fig. 2. So, there is an increase in chiplet manufacturing yield.
- Fig. 4.4 (d) shows the partitioning figure of merit (p-FOM) for chiplets. We see that the maximum p-FOM is achieved at a number of chiplets of 9, which shows that a 27 mm × 27 mm chip can be broken into 9 chiplets of size ~9.02 mm × 9.02 mm for optimized chiplet based implementation. It further solidifies the optimal chiplet size lies in 'golden-dielet' regime discussed by Iyer et al. in their work [SIyer19]. Note that memory manufacturers have for several generations settled on chip areas that are between 80 and 100 mm^2 for yield reasons. It is interesting that our analysis of chip partitioning yields similar conclusions.

4.4.2 p-FOM for a fixed partitioning (m = 9)

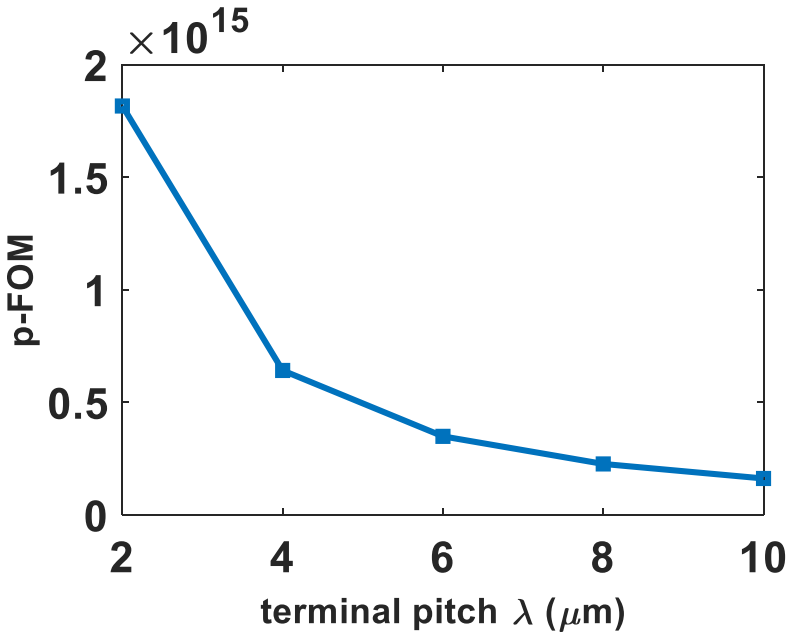


Figure 4.5 Variation in figure of merit for a fixed m with terminal pitch varying. λ = 2 μm to 10 μm.

Considering the case we have 9 chiplets, Fig. 4.5 shows the p-FOM for different terminal pitch ranging from 2µm to 10µm. p-FOM decreases with increase in terminal pitches, which states we always gain by going to lower interconnect/terminal pitches. It should be noted that going to lower terminal pitches will increase the resistance of the wire and the wire delay will increase. This can be detrimental to the performance of the chip and hence a limit to the pitch reduction also exists. We cannot transfer data lower than a rate it is generated at i.e. the lowest possible frequency for data transfer is the clock frequency of the chip.

4.4.3 p-FOM for various defect densities

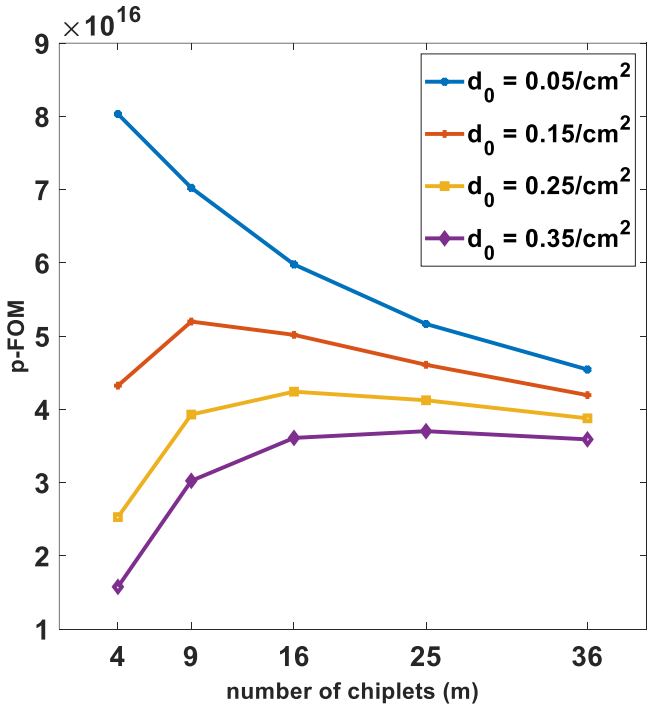


Figure 4.6 Variation in figure of merit for at different values of defect densities. It can be seen use of larger chiplet sizes and fewer number of chiplets becomes more feasible as defect densities decrease.

It is important to consider the effect of process maturity on manufacture of large monolithic chips and need for partitioning. In the beginning when a manufacturing process is still in the maturation phase, the chip has high defect density, but as the process matures (usually a few quarters), there are substantial process improvements, material changes,

manufacturing tool improvements and more, which can significantly reduce both systematic defects, as well as overall defect density. In Fig. 4.6, the effect of reducing overall defect densities is shown. At high defect densities ~0.35 defects/cm^2, higher FOM is achieved at greater number of partitions as chiplet sizes are small and have high yield. But eventually when defect densities reach ~0.05 defects/cm^2, partitioning may no longer result in any gains, because of large yields even for monolithic chips. This is an important consideration. The cost benefits of chiplet portioning will be less important as the process matures.

It should be noted that usually process is said to become manufacturable at defect density of 0.5 /cm^2 [11], but best when defect densities are reduced to below 0.1 /cm^2. For very advanced nodes attaining this level of defect density may take a few years and may not be even achievable. In those cases, partitioning will continue to have lasting value.

4.5 Discussion and limitations of partitioning Figure of Merit (p-FOM)

The takeaways from this work can be summarized in the points below:

- A monolithic chip will always have a performance superior to a system of chiplets that provide identical functionality as the monolithic version. This is due to the added area and power overhead to the chiplets due to inclusion of terminals at the periphery of chiplets.
- Due to rising performance requirements in, Artificial Intelligence (AI), and high-performance computing, today die sizes have increased to occupy nearly the entire reticle, which combined with increasingly complex lithography and processes, leads to low yields and a high cost of manufacturing.
- Designs that partition large chips into chiplets will improve this yield and reduce cost, however the optimal chiplet size depends on careful understanding of relationship of area overhead, power overhead of added terminals as well as yield improvement. A partitioning figure of merit (p-FOM) is developed in this work to demonstrate the performance to cost trade-offs. It is implemented for a design example involving a 729

mm^2 chip. The best performance to cost is obtained by partitioning the 729 mm^2 chip into 9 chiplets of 9.02 mm × 9.02 mm dimensions. This is of course considering the inputs given in Table 1.

- Process maturity of manufacturing has a crucial role to play in determination of performance to cost. As process matures, larger and larger chiplet sizes become feasible. The need for partitioning can be reduced for defect densities below 0.05 /cm^2. The question of whether defect densities at very advanced nodes will reach this level of defect density remains to be seen.

- In this paper we have not broached the topic of heterogeneous integration and limited ourselves to the simpler problem of partitioning a monolithic chip. Heterogenous integration will add additional performance benefits that will be system architecture dependent and validate partitioning even when an if yields mature. That is beyond the scope of this work. Our analysis in the paper has been somewhat simplistic and we have not considered the yield implications of fine pitch assembly. However, given the relative infancy of advanced packaging and the global efforts being devoted to it, we expect rapid strides in process simplification and assembly considerations [12] that will bring packaging yields to the point where they are not a detractor.

In the section 4.2 – 4.5, an analytical approach to determine the partitioning parameters for a monolithic chip are investigated. The optimal chiplet sizes after partitioning depends on bandwidth requirements, terminal pitch, electrical characteristics of the transceivers, defect density and other manufacturing parameters. We believe the equations developed in this analysis would be valuable to the system designer in deciding the optimal chiplet size for their application and understanding the necessary trade-offs.

4.6 Signaling Figure of Merit (s-FOM_{UCLA}) for die-to-die signaling on advanced packages

Once a monolithic chip has been partitioned into chiplets, the next critical step is to integrate these discrete dielets to function collectively as a unified system—essentially replicating the behavior of a monolithic SoC. This integration is enabled by advanced packaging technologies, which offer a myriad options of interconnect platforms for chiplet-

to-chiplet (dielet-to-dielet, or D2D) communication. These platforms include organic substrates [SArd20], glass or silicon interposers, and wafer-level solutions such as Silicon Interconnect Fabric (Si-IF) [SJan21], [Siyer19] and Fan-Out Wafer-Level Packaging (FOWLP) [TFuku17].

Selection of the D2D PHY protocol is a significant design decision because transceivers on both ends of a link need to be compatible. In effect, the D2D PHY selection for a design is a selection for a family of designs. To provide guidance on how the D2D PHY design choice can be made, we need to optimize a cost function parameterized with all the relevant performance and cost parameters. This function incorporates key electrical, chip-level, and packaging-level parameters that influence performance, power, and cost. A summary of these influencing factors is depicted in Fig. 4.7.

- Electrical factors are typically dictated by system requirements and include data rate per I/O pin, Bandwidth per millimeter of die shoreline (BW/mm), Bit Error Rate (BER), Link latency, Signaling protocol characteristics, including channel control, redundancy, and encoding overhead. Also, channel integrity metrics such as insertion loss, crosstalk, and return loss govern the achievable signaling integrity (SI) and bandwidth efficiency of a package.
- Chip design factors influence how well the PHY can be integrated into the dielet. These include technology node for the I/O circuitry, which in turn affects I/O power consumption, physical footprint per I/O and integration density and feasibility within the chiplet layout
- Packaging factors are equally critical and often determine the mechanical and electrical feasibility of implementing a signaling scheme. The factors to be considered include bump pitch, trace pitch, dielet-to-dielet spacing and other design metrics which need to be obtained from an assembly design kit (ADK) for a given package. These parameters are typically standardized for each packaging technology and optimized to achieve high manufacturing yields (>99.999%).

For achieving the best performance from D2D interfaces after dielet/chiplet disaggregation, one has to consider the interrelated nature of electrical, design, and

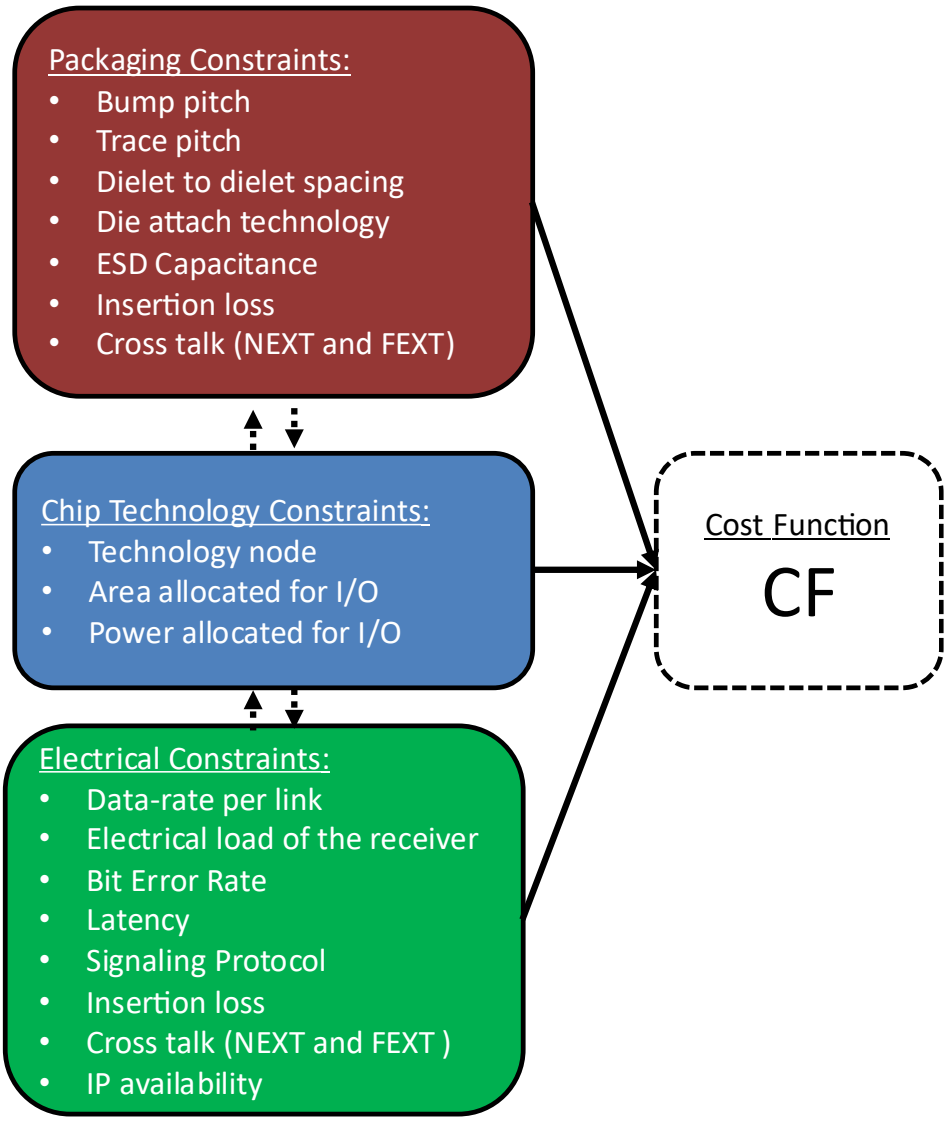


Figure 4.7 Factors important in making I/O decision.

packaging factors. In the past, Green et al. [DAG] introduced a technology figure of merit using energy per bit to evaluate available PHY options. It is represented as:

$$\text{s-FOM}_{\text{I}} = \frac{\text{Bandwidth/Shoreline}}{\text{Energy/bit}} \qquad (4.E)$$

However, this figure of merit does not account for packaging parameters, link latency, transceiver area and link-length between dielets, making it inadequate in providing design insights to the system architect. Later, Jangam and Iyer [SJansFOM] addressed most of these shortcomings using s-FOM$_{\text{UCLA}}$ described in equation (4.F). The s-FOM$_{\text{UCLA}}$ is a

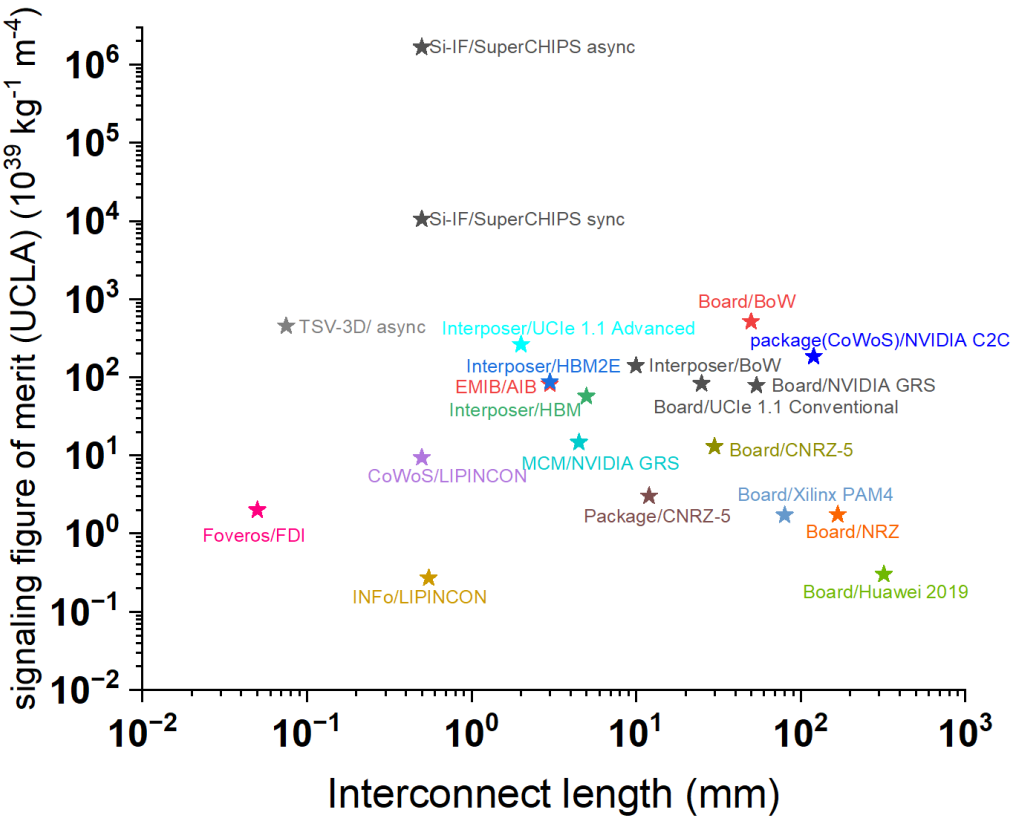


Figure 4.8 Variation of s-FoM$_{UCLA}$ (10^{39} kg^{-1}m^{-4}) versus reported link length (mm)

useful function for comparing commercially available PHYs. This comparison is represented in Fig. 4.8 against link length.

$$\text{s-FOM}_{\text{UCLA}} = \frac{\left(\frac{\text{Bandwidth}}{\text{Shoreline}}\right) * (\text{Length}_{\text{Link}})}{\left(\frac{\text{Energy}}{\text{bit}}\right) * \left(\text{IOcols} * \frac{\text{TrArea}}{\text{link}}\right) * \text{Latency}} \qquad (4.F)$$

While all the above proposed figures of merit are useful in evaluating commercially available PHY and signaling protocols, they do not propose a methodology of selecting an optimal D2D PHY in the context of a particular design problem. In an actual system design scenario, the system designer must conform to specific design requirements, look at available packaging solutions (which have their own requirements) and select the PHY which results in an optimized performance to efficiency ratio. Therefore, in this work, we demonstrate two contributions –

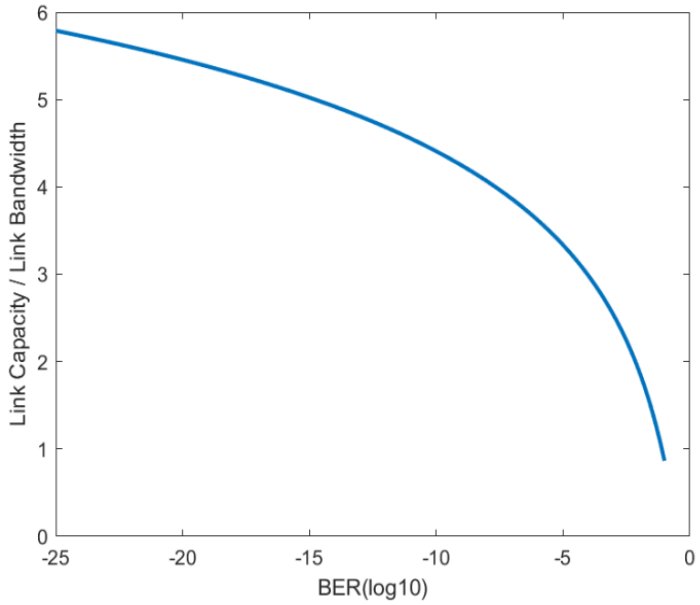


Figure 4.9 Link Capacity/Link Bandwidth (used in CF) vs. BER (in log) scale showing lower BER correspond to higher Link Capacity/ Link Bandwidth.

- we introduce a cost metric – the Cost Function (CF) derived from s-FOM$_{UCLA}$ which improves upon it by adding the effect of bit error rate (BER) of the signaling scheme as well as packaging cost factor to evaluate multiple packaging solutions for certain electrical and chip design requirements to select the optimal D2D PHY.
- We implement CF to evaluate the PHY used in context of a processor-memory design problem.

4.6.1 Development of Cost Function (CF)

In equation (B) The Bandwidth/Shoreline represents the bandwidth per mm of die edge. Length$_{Link}$ represents the maximum drivable link length of the D2D PHY between transmitter and receiver. IOcols is the number of columns of I/O pads used (perpendicular to die edge). TrArea is the actual circuit area (not I/O pad area) of both the transmitter and the receiver and accounts for the depth of the IOs represented by IOcols. Energy per bit and Latency represent the energy consumed and latency in accessing one bit of data. The definitions of the parameters used in s-FoM$_{UCLA}$ have also been validated by Open Domain-Specific Architecture (ODSA) [ODSA] association.

To develop the cost function, we introduce two additional factors - BER as well as the relative cost of the packaging technology used to implement the D2D PHY. We have chosen to represent the BER as the link capacity normalized to link bandwidth. We invoke the Shannon-Hartley theorem [SHT] using normalized signal to noise ratio, Eb/No (energy per bit to noise power spectral density ratio). It is especially useful when comparing the BER performance of different digital modulation schemes independent of bandwidth. We relate BER to Eb/No for additive white gaussian noise model and get to equation (4.G):

$$\mathrm{C/B} = \log_2(1 + \mathrm{erfc}^{-2}(2 * \mathrm{BER})) \qquad (4.G)$$

This expression penalizes high BER and saturates at very low BER as shown in Fig. 3. This approach is justified because error detection and correction schemes have a lower penalty at low BER, but if the BER are very low, that penalty does not improve dramatically. We are now ready to address the cost function that needs to be optimized. We defined this cost function (CF) as in equation:

$$\mathrm{CF} = \frac{\alpha}{\left[\frac{\mathrm{C}}{\mathrm{B}}\right] * \text{s-FOM}_{\mathrm{UCLA}}} \qquad (4.H)$$

In equation (D), *α* represents the normalized cost of the packaging construct being employed.

4.6.2 Implementation of Cost Function for processor-memory scenario and findings

The factors in CF i.e., bandwidth/shoreline, transceiver area, energy per bit, latency, channel characteristics, and cost map directly to the choices which a system designer needs to make during a I/O decision. We demonstrate the application of CF in two scenarios involving a processor-memory design:

1. A high-performance general processing unit (GPU) with High Bandwidth Memory (HBM) [ChaNay, Mike17] dielets implemented on a Si-Interposer.
2. The same GPU as a packaged chip with Hybrid Memory Cube (HMC) [JTPaw11, HMCG2] packages implemented on an organic MCM.

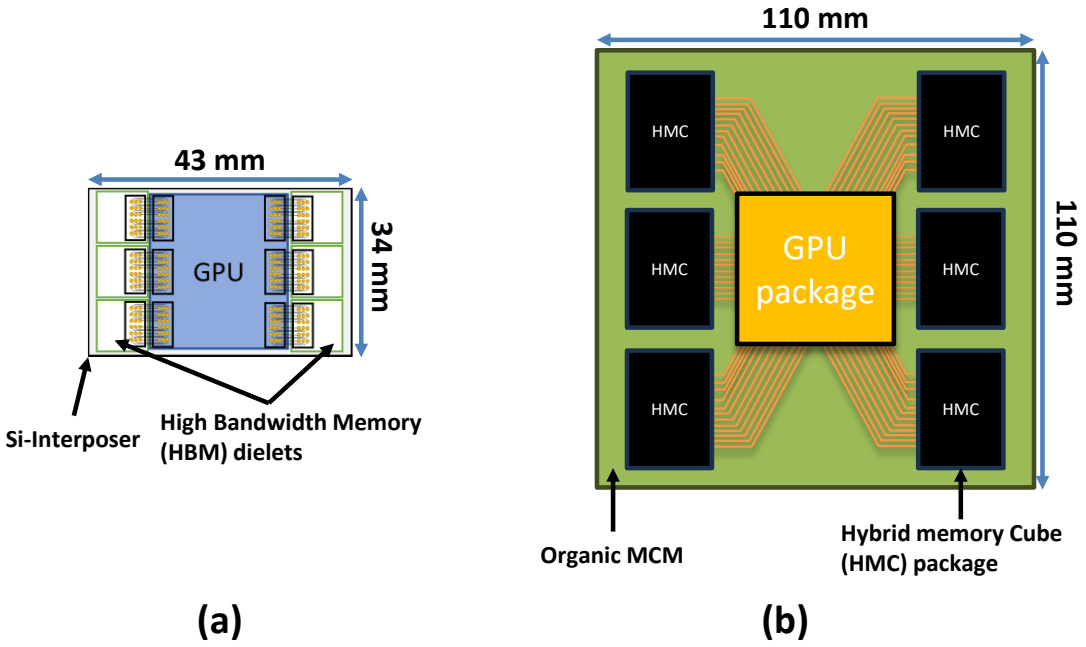


Figure 4.10 (a) An interposer implementation of a GPU dielet and 6 HBMs (b) An organic MCM implementation of the same problem with the same GPU in packaged form and HMC Gen2 packages

The maximum memory capacity per HMC package is 2 GB [HMCG2]. To make the comparison fair, we have chosen to limit the memory capacity of each HBM dielet to 2GB, even if larger memory capacities are possible as seen from literature [Mike17]. Therefore, in both cases, we have fixed the total memory capacity requirement to 12 GB and total memory bandwidth requirement to ~1.5 TB/s. Table 4.1 shows the details of the implementation.

The HBM is designed to be integrated on a silicon interposer. The GPU-HBM on interposer implementation is shown in Fig. 4.10(a). It consists of one GPU of area ~ 800 mm^2 and 6 HBM dies of area ~ 70 mm^2 each with 2GB memory capacity. The data rate per pin is 2 Gb/s. While the HBM dies are commercially available, the GPU is custom designed using 12 nm TSMC process and is approximately 31 mm X 26 mm. The effect of die yield was considered using Murphy's die yield model [TYM]. The GPU and HBM dielets are interconnected using traces at 55 µm bump pitches and 3 µm trace pitch. Considering maximum trace length of 5 mm between GPU and HBM dielets, the size of

the interposer to implement this design is estimated to be 43 mm X 34 mm. the interposer has 4 metal layers – two for signals, one for power and one for ground. The GPU-HMC

Table 4.1 components used in the implementation of the processor-memory package design.

	HBM-Si interposer implementation	**HMC-Organic MCM implementation**
Data I/O pins (memory)	1024 per HBM dielet	128 per HMC package
Data rate/pin (processor)	2 Gb/s	16 Gb/s (modified interface)
Memory Capacity/stack	2 GB (bare stack)	2 GB (packaged stack)
I/O bump pitch	55 μm (μbump)	1 mm (BGA)
Area of Package Substrate	1462 mm^2	12100 mm^2 (8.2x)
Additional Comments	GPU area (~800 mm^2)	GPU die area increased by 10% to accommodate SerDes

implementation is shown in Fig. 4.10(b). The HMC is designed with a conventional packaging hierarchy. Each package has a memory capacity of 2GB. The implementation consists of a packaged GPU and 6 HMC generation 2 (Gen2) packages interconnected via an organic substrate at 1 mm bump pitch. The links are made to run at 16Gbps using a custom SerDes implemented on both HMC and GPU to achieve the required memory bandwidth. The HMC Gen2 package is 31 mm X 31 mm [HMCG2]. We have had to increase the GPU size by ~ 10% to accommodate the SerDes necessary for GPU-HMC communication. The effect of increase in GPU size on die yield was included using Murphy's die yield model. The GPU package size after fanout is estimated to be 36 mm

X 36 mm. The organic substrate has 10 layers (4-2-4 stack) as determined from HMC specifications [HMCG2]. The size of the organic multichip module (MCM) is estimated to be 110 mm X 110 mm. The cost of the individual components were obtained from vendor inputs and the estimated cost of the assembled module was calculated for each implementation. The Table. 4.2. represents estimated costs normalized to HBM-Si-Interposer implementation

Table 4.2 Estimated costs normalized to HBM-Si interposer implementation.

Component	**HBM-Si Interposer Implementation**		**HMC-Organic MCM implementation**	
	Quantity	Normalized cost	Quantity	Normalized cost
Memory	6 (bare die)	1x	6 (packaged)	1.5x
GPU	1 (bare die)	1x	1 (packaged)	1.5x
Package Assembly	1 (4-layer interposer + package substrate)	1x	1 (10-layer organic MCM)	0.6x
Normalized assembled module cost		1x		1.2x

(**Note:** The normalized costs have been estimated based on vendor input for 50 units of each component. However, costs can vary depending on vendor and volume. The numbers are for illustrative purposes only.)

It is found that the GPU-HMC implementation raw package cost is about 1.2x higher than the GPU-HBM on interposer implementation. Basically, cost factor α for GPU-HMC Gen2 organic MCM implementation is 1.2. This is counter intuitive as conventional wisdom is that organic packaging is cheaper, however, to meet the constraints of bandwidth this is not necessarily the case as we have seen here. However, the raw bill of materials and assembly cost are not the whole story. We also need to account for the performance metrics. We do this through the cost function (CF) shown in equation (4.H). While we

have kept the basic functionality consistent through the memory size and the bandwidth between the memory and the processor, we were not able to simultaneously constrain the two solutions to have the same energy per bit and latency. The cost function accounts for this discrepancy. And that analysis is shown in Table. 4.3. As a result, the cost function for the HBM-Si-interposer solution is about 11x better than the HMC-Organic MCM solution. This difference is driven primarily by the performance metrics of the two solutions. Thus, the HBM-Si-interposer solution is more practical than the HMC-organic MCM solution. This has also been borne out in the marketplace and HMCs have been discontinued and the industry has migrated to the HBM-Si interposer solution

Table 4.3 Values of the factors used in calculation of cost factor (CF) for processor memory design scenario. Lower CF is preferred.

Parameter name	HBM-Si-Interposer Implementation	HMC-Organic MCM implementation
Energy per bit (pJ/bit)	3.9 [14]	13 [13]
Latency (ns)	1.5 [2]	~7 [2]
Bandwidth/Shoreline (Gbps/mm)	281	234.5
CF ($x10^{-31}$ Jm^2s^2)	538.8	6131 (11.37x higher)

4.6.3 Inferences from Cost Factor Analysis

cost function (CF) provides guidance for the optimal choice of D2D PHY for a given design. Individual component and individual packaging constructs cannot be optimized independently. The integrated solution needs to be optimized and this can be done via the cost function. As a practical example we have compared two embodiments of a processor-memory packaging construct: one using advanced packaging such as the silicon interposer and the other using more conventional techniques such as organic MCM. We show that the advanced packaging solution has ~ 11x better cost function. We believe that novel methods of scaling the package such as Si-IF and wafer-scale integration will provide even greater benefits to the cost factor.

4.7 Chapter Conclusion

This chapter explored the critical transition from monolithic chip design to chiplet-based architectures, emphasizing the key challenges and optimization strategies required for effective system integration. We examined the implications of chip partitioning, the introduction of die-to-die (D2D) interfaces, and the co-optimization of signaling protocols and packaging technologies to meet performance, power, and cost goals. The individual lessons learnt from partitioning figure of merit (p-FOM) and chip-package cost function (CF) are already provided towards end of sections 4.5 and 4.6 respectively. To conclude this chapter, we present a summary here:

- **System Reconstruction Through Chiplets:** A central objective of the chiplet ecosystem is to partition large monolithic chips into smaller, functional chiplets and reassemble them using high-performance D2D interfaces such that the resulting system closely approximates the functionality and performance of the original monolithic design. It should be noted that the performance of monolithic chip will always be superior to a system of chiplets.
- **Partitioning Trade-offs and the Partitioning Figure of Merit (p-FOM):** Partitioning of large chips into chiplets will improve manufacturing yield and reduce cost, but it also introduces trade-offs in area, power, and design complexity. Determining the optimal chiplet size requires balancing these trade-offs, particularly the overhead introduced by additional I/O terminals against the yield gains. The p-FOM introduced in this chapter provides a quantitative framework to evaluate and guide such partitioning decisions.
- **Die-to-Die Communication and Cost Optimization:** After partitioning, chiplets must communicate using die-to-die interfaces. The choice of an optimized D2D PHY depends on electrical, circuit design and packaging factors. Given system requirements and packaging options available, the cost function can provide valuable guidance to the system architect to select the right D2D PHY for interconnecting chiplets with optimized values of power, area and communication cost.

Together, the partitioning figure of merit (p-FOM) and the chip-package cost function offer a systematic approach to designing high-performance, cost-effective chiplet-based systems. By applying these metrics, system designers can make informed architectural decisions that balance yield, performance, and manufacturing complexity.

4.8. Chapter 4 references

[AnySil] anysilicon, “Design Challenges in Single-Digit Technology Nodes,” AnySilicon. Accessed: May 31, 2025. [Online]. Available: https://anysilicon.com/design-challenges-in-single-digit-technology-nodes/

[emb25] S. Lovati, “TSMC’s 2nm Technology Almost Ready for Mass Production,” Embedded. Accessed: May 31, 2025. [Online]. Available: https://www.embedded.com/tsmcs-2nm-technology-almost-ready-for-mass-production/

[SIyer19] 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*, vol. 63, no. 6, p. 5:1-5:16, Nov. 2019, doi: 10.1147/JRD.2019.2940427.

[GMou16] G. Mounce, J. Lyke, S. Horan, W. Powell, R. Doyle, and R. Some, “Chiplet based approach for heterogeneous processing and packaging architectures,” in *2016 IEEE Aerospace Conference*, Mar. 2016, pp. 1–12. doi: 10.1109/AERO.2016.7500830.

[Cade] “The rise of chiplets” [online] available: The Rise of Chiplets | Advanced PCB Design Blog | Cadence

[RPFil] R. P. Filho, “What Is A Chiplet, And Why Should You Care” [online] available: What Is A Chiplet, And Why Should You Care? (semiengineering.com)

[PGup19] P. Gupta and S. S. Iyer, “Goodbye, motherboard. Bare chiplets bonded to silicon will make computers smaller and more powerful: Hello, silicon-interconnect fabric,” *IEEE Spectrum*, vol. 56, no. 10, pp. 28–33, Oct. 2019, doi: 10.1109/MSPEC.2019.8847587.

[KSahEPEPS] K. Sahoo, B. Vinnakota, S. Ardalan, and S. S. Iyer, “Co-optimizing signaling protocol with semiconductor and packaging technology,” in *2021 IEEE 30th Conference on Electrical Performance of Electronic Packaging and Systems (EPEPS)*, Oct. 2021, pp. 1–3. doi: 10.1109/EPEPS51341.2021.9609161.

[SJansFOM] S. Jangam and S. S. Iyer, “A Signaling Figure of Merit (s-FoM) for Advanced Packaging,” *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 10, no. 10, pp. 1758–1761, Oct. 2020, doi: 10.1109/TCPMT.2020.3022760.

[KSah22] K. Sahoo, U. Rathore, S. Chandra Jangam, T. Nguyen, D. Markovic, and S. S. Iyer, “Functional Demonstration of < 0.4-pJ/bit, 9.8 µm Fine-Pitch Dielet-to-Dielet Links for Advanced Packaging using Silicon Interconnect Fabric,” in *2022 IEEE 72nd Electronic Components and Technology Conference (ECTC)*, May 2022, pp. 2104–2110. doi: 10.1109/ECTC51906.2022.00332.

[NMax87] N. Maxemchuk, “Routing in the Manhattan Street Network,” *IEEE Transactions on Communications*, vol. 35, no. 5, pp. 503–512, May 1987, doi: 10.1109/TCOM.1987.1096802.

[JACun90] J. A. Cunningham, “The use and evaluation of yield models in integrated circuit manufacturing,” *IEEE Transactions on Semiconductor Manufacturing*, vol. 3, no. 2, pp. 60–71, May 1990, doi: 10.1109/66.53188.

[BEYield] “Yield-models – Bose-Einstein” [online], available: https://fog.misty.com/perry/cod/c1/notes_16.html

[DefDen] “Defect Density”, [online], available: https://anysilicon.com/semipedia/defect-density-dd/#:~:text=While%20the%20acceptable%20DD%20can,technology%20advancements%20and%20specific%20applications.

[NAPVIS] NAPMP vision paper available online at: https://www.nist.gov/system/files/documents/2023/11/19/NAPMP-Vision-Paper-20231120.pdf

[SArd20] S. Ardalan *et al.*, “Bunch of Wires: An Open Die-to-Die Interface,” in *2020 IEEE Symposium on High-Performance Interconnects (HOTI)*, Aug. 2020, pp. 9–16. doi: 10.1109/HOTI51249.2020.00017.

[SJan21] S. Jangam and S. S. Iyer, “Silicon-Interconnect Fabric for Fine-Pitch (≤10 μm) Heterogeneous Integration,” *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 11, no. 5, pp. 727–738, 2021, doi: 10.1109/TCPMT.2021.3075219.

[TFuku17] T. Fukushima *et al.*, “‘FlexTrate ^TM’ — Scaled Heterogeneous Integration on Flexible Biocompatible Substrates Using FOWLP,” in *2017 IEEE 67th Electronic Components and Technology Conference (ECTC)*, May 2017, pp. 649–654. doi: 10.1109/ECTC.2017.226.

[DAG] Daniel Green, DARPA CHIPS

[ODSA] Open Domain-Specific Architecture [online]: https://www.opencompute.org/

[SHT] Shannon Hartley Theorem: http://www.cs.man.ac.uk/barry/mydocs /CS3282 /Notes/DC06 7.pdf

[JTPaw11] J. T. Pawlowski, "Hybrid memory cube (HMC)," 2011 IEEE Hot Chips 23 Symposium (HCS), 2011, pp. 1-24

[HMCG2] HMC Gen 2 specification: https://www.micron.com/-/media/client/global/documents/products/data-sheet/hmc/gen2/hmc_gen2.pdf.

[TYM] “Test Yield Models - Poisson, Murphy, Exponential, Seeds.” Accessed: Jun. 01, 2025. [Online]. Available: https://www.eesemi.com/test-yield-models.htm

[ChaNay] Changyong Ahn & Nayoung Lee, “An In-depth study of High Bandwidth Memory”, SK Hynix.

[Mike17] Mike O’ Conner et al., “Fine-Grained DRAM: Energy-Efficient DRAM for Extreme Bandwidth Systems,” MICRO-50 October 14–18, 2017, USA

Chapter 5 Anuvāda: a re-usable translator design for cross-protocol communication in advanced packaging

5.1 Wired Communication in wafer-scale systems

Efficient communication between dielets is a crucial aspect of wafer-scale systems. As highlighted by Pal et al. [Spal19], three distinct signaling architectures are typically employed, depending on drive length, energy per bit, and latency requirements:

- **Short-haul communication** – for adjacent or neighboring dielets;
- **Medium-haul communication** – for dielets separated by moderate distances across the wafer (several dielets in between);
- **Long-haul communication** – for communication across multiple wafers or with external peripheral systems using standardized protocols such as PCIe (peripheral component interconnect express), Ethernet, etc.

Figure 5.1 compares bandwidth, energy efficiency, and latency of on-wafer communication links with those of multi-chip modules (MCMs), board-level interconnects, and PCB backplanes. On-wafer links exhibit performance comparable to on-chip interconnects and significantly outperform conventional packaging solutions.

In the context of the Silicon Interconnect Fabric (Si-IF), short-haul communication is enabled via SuperCHIPS — a low-energy, inverter-based I/O protocol that operates at a fine bump pitch of 10 µm. Originally proposed in [SJan20+, KSah22], and further validated in [URat22, SNagi23], SuperCHIPS enables robust and scalable parallel signaling across neighboring dielets. Its detailed architecture is discussed in Section 5.4, along with key results obtained so far from our prior works. In this work, we retain the core SuperCHIPS

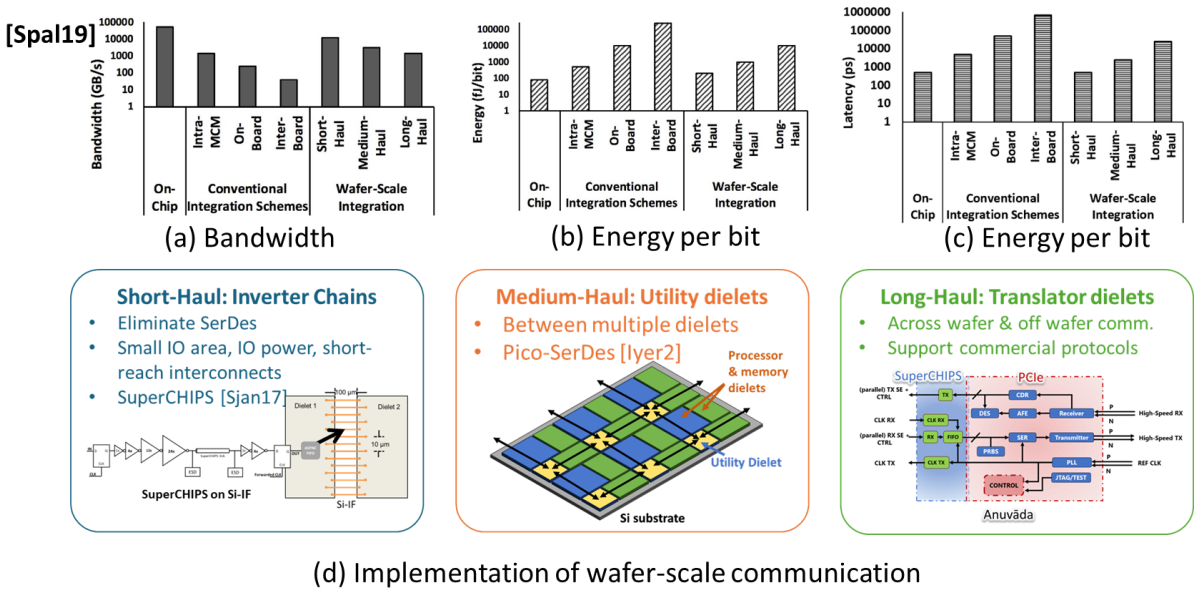


Figure 5.1 Different regimes of communication on a wafer-scale system. (a), (b) and (c) are taken from work done by [Spal19].

PHY design but enhance its resilience against electrostatic discharge (ESD) events, which are critical during dielet transfer, storage, and bonding, by adding custom designed ESD protection diodes to each IO.

For medium-haul communication, we propose a hub-style architecture using a utility dielet equipped with low-power pico-SerDes. This dielet can facilitate longer intra-wafer communication. While the concept of the utility dielet remains under active development, it is not covered in the current scope of this dissertation.

Long-haul communication addresses two key challenges: (1) interfacing with off-wafer peripheral systems, and (2) enabling inter-wafer cross-protocol communication. Peripheral hardware typically uses standardized high-speed protocols such as PCIe [PCIeWiki] (peripheral component interconnect express), Ethernet [ETHE], and QSFP/QSFP+ [FSQSFP] (quad small form-factor pluggable) optical fibers. Therefore, wafer-scale systems like Si-IF must support protocol interoperability—translating between internal SuperCHIPS-based signaling and these external interfaces. Moreover, Si-IF envisions integration of heterogeneous dielets with potentially diverse I/O standards

and architectures. To ensure seamless communication, the system must support protocol translation between SuperCHIPS and other I/O protocols.

To address this, we introduce a specialized **translator chiplet** that bridges SuperCHIPS with standard protocols such as PCIe, CXL [CXLWiki], UCIe [UCIeSYN], and Ethernet. All dielets on Si-IF are expected to use SuperCHIPS PHY for communication. The translator chiplet facilitates two functions: (1) enabling wafer-to-periphery communication, and (2) supporting non-SuperCHIPS protocols within the wafer by using SuperCHIPS as a common intermediary.

Figure 4.2 illustrates the translator chiplet concept. As shown, it allows incoming data in any supported protocol to be converted into SuperCHIPS format and subsequently re-translated as needed—effectively forming a modular and interoperable signaling backbone for the Si-IF platform.

The remainder of this chapter details the design, fabrication, and testing of a proof-of-concept translator chiplet named **Anuvāda**, introduced in the following section.

5.2 Concept of chiplet level protocol translation using Anuvāda

Anuvāda (Sanskrit for "translation") is a protocol translator chiplet designed to facilitate seamless communication between dielets that use different signaling schemes. It can be

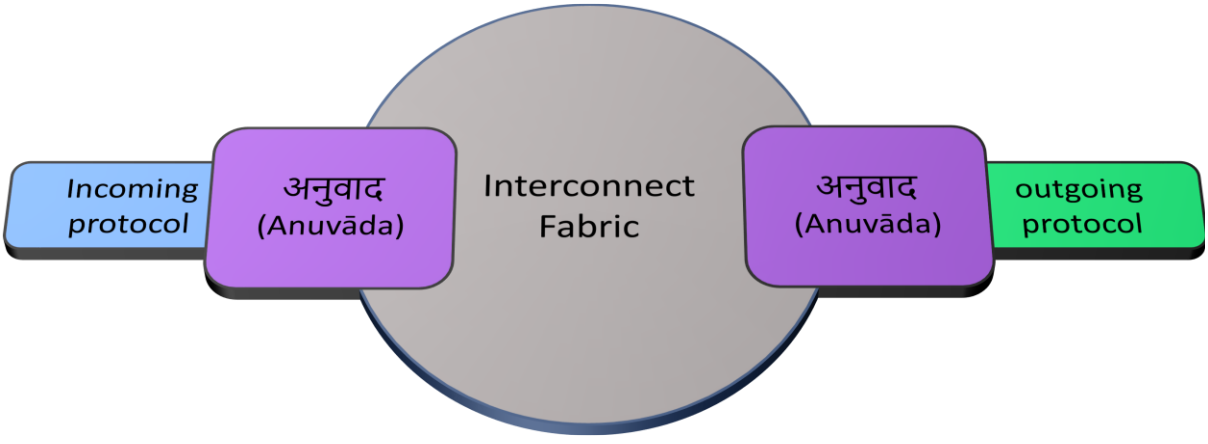


Figure 5.2 Schematic of protocol translation concept.

integrated into wafer-scale, interposer, or flexible packages, acting as an intermediary by converting all incoming signals into a standardized 10µm fine-pitch protocol called SuperCHIPS—and vice versa.

In this work, we utilize SuperCHIPS as the common protocol for inter-dielet communication. The Anuvāda dielet accepts commercial protocols such as PCIe, UCIe, and CXL, translating them into SuperCHIPS signals. Protocol selection is determined via control signals during the handshake process. Additionally, Anuvāda performs bidirectional conversion, translating SuperCHIPS signals back into PCIe, UCIe, or CXL as required. It also incorporates electrostatic discharge (ESD) protection, error detection, and correction mechanisms associated with these commercial protocols.

Beyond intra-package (intra-Si-IF) translation, Anuvāda can be placed at the package periphery to mediate between off-package and on-package I/Os. A high-level system overview featuring processors, memory, and Anuvāda translator dielets is illustrated in Fig. 5.3. In the schematic, peripheral Anuvāda dielets convert incoming PCIe signals into SuperCHIPS for internal on-package use. Most dielets within the package communicate natively via SuperCHIPS, eliminating the need for additional protocol conversion.

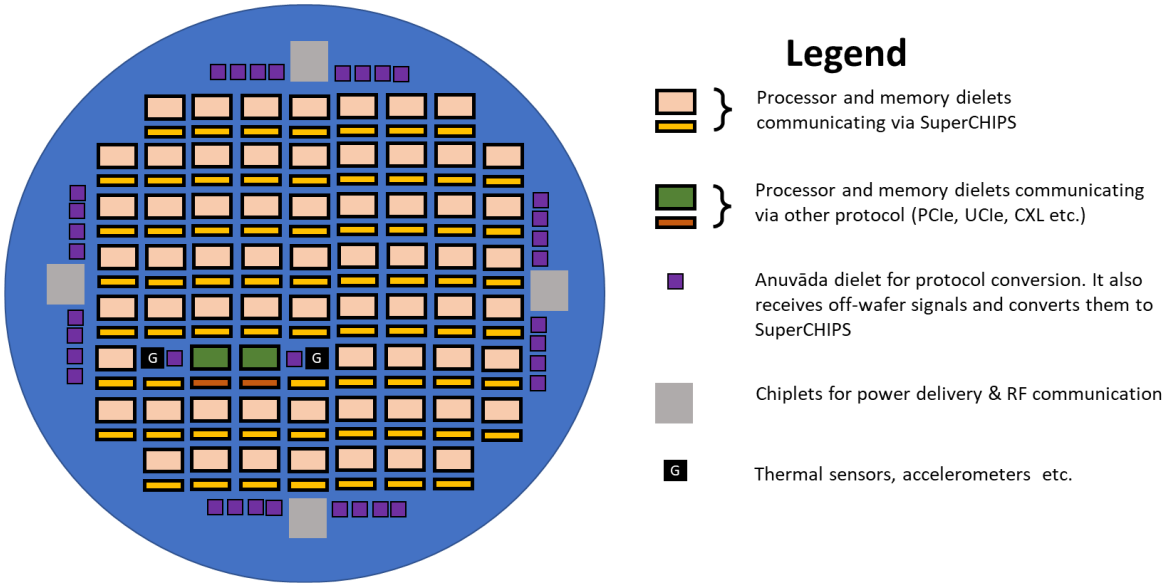


Figure 5.3 A fully populated wafer-scale system with processor, memory, and anuvada dielets. Anuvada dielets are responsible for translation of SuperCHIPS data to PCIe and other widely used protocols.

However, in case there are any chiplets which use another protocol such as UCIe, they require Anuvāda dielets for UCIe-to-SuperCHIPS conversion.

Thus, Anuvāda can support multiple protocols via translation and helps to remove the burden of redesigning signaling schemes, allowing dielet manufacturers to focus on core functionality. Moreover, once its operation is hardware-verified, system architects can reuse the IP and avoid the overhead of integrating various I/O protocols manually, thereby streamlining the overall design process.

The design and functional hardware verification of the Anuvāda dielet is structured into multiple phases. The first phase focuses on a proof-of-concept demonstration at a small scale, showcasing communication between a 234 µm I/O pitch, 8 Gbps PCIe 3.1 serial interface and a 10 µm fine-pitch, 500 Mbps SuperCHIPS interface. While SuperCHIPS supports data rates up to 3 Gbps per link as discussed in section 5.4, this initial phase is constrained by PCIe 3.1 specifications. This phase helps identify and resolve design challenges, optimize functionality, and refine integration strategies for the dielets. The next phase involves increasing data rates to 32 Gbps or higher, enabling compatibility with a broader range of high-speed I/O protocols such as PCIe 4, 5, 6, and UCIe.

For the first phase, the Anuvāda dielet will be assembled on a Silicon Interconnect Fabric (Si-IF), designed and fabricated at UCLA. Modifications were necessary to accommodate the PCIe 3.1 IP received from Synopsys for bonding to copper (Cu) pillars on Si-IF. PCIe 3.1 requires termination pads fabricated with Aluminum when used with GlobalFoundries 22nm FDSOI technology. Aluminum (Al) cannot be reliably bonded to Cu pillars due to the fast oxidation rate of Al to Al_2O_3 or alumina, which does not bond to Cu [KSah24]. As a result, all aluminum pads were overlayed with thin titanium/gold (Ti/Au) overlayer of thickness 20nm/200nm was to enable bonding to Cu pillars on the Si-IF. In this case, the bonding occurred between Au layer (200nm) on the Anuvāda dielet and Cu pillars on the corresponding Si-IF. The parameters used in this bonding have been described in section 2.13, chapter 2.

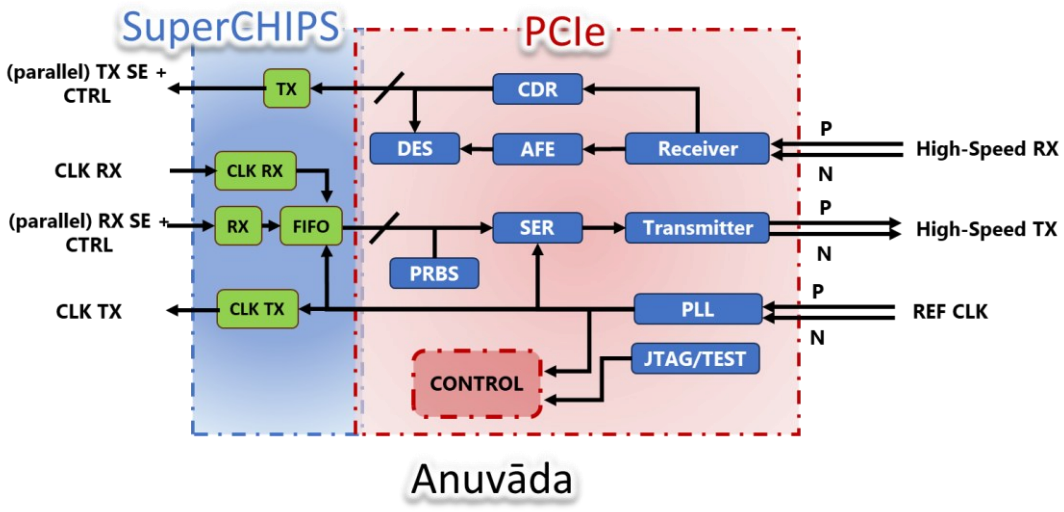


Figure 5.4 Schematic overview of proof-of-concept PCIe3.1 to SuperCHIPS Anuvāda chiplet.

5.3 Anuvāda phase 1: PCIe 3.1 to SuperCHIPS specifications

The simplified block diagram of the implemented Anuvāda chiplet is shown in Fig. 5.4. Anuvāda is designed in 22 nm fully depleted silicon on insulator (22nm FDSOI) [GF22FDX] technology node from Global Foundries. The chiplet receives a high speed 8 Gbps signal through a differential input and converts it into 20 parallel SuperCHIPS outputs depending on 8b/10b encoding scheme which each pin running at 500 Mbps. Also, it can receive 20 SuperCHIPS inputs with 8b/10b encoding and convert it to a high speed 8 Gbps PCIe 3.1 output. In addition, the chip consists of control circuitry, necessary clock domain crossing (CDC) blocks between PCIe and SuperCHIPS to accommodate any phase mismatches etc. The reference clock used for chip operation is 100 MHz differential. The reference clock requirements for Anuvāda are described as follows:

Table 5.1 Reference clock requirements for Anuvāda

Parameter	*Value*	*Unit*
REF CLK frequency	100	MHz
Allowed frequency offset	-300 to 300	ppm

Max. RMS Jitter	1	ps, rms
Max. Deterministic Jitter	2.8	ps, pp
Voltage min. typical	0	V
Voltage max. typical	0.8	V
Voltage swing	0.8 to 0.88	V

As described in Section 5.2, the PCIe 3.1 IP used in the Anuvāda chiplet was sourced from Synopsys [PCISYN] to streamline the chip design process. In contrast, the SuperCHIPS IP was developed in-house at UCLA and represents an updated version of the design originally introduced by Jangam et al. [SJanD20]. A detailed description of the SuperCHIPS architecture is provided in Section 5.4. Anuvāda also integrates a chiplet configuration module based on the JTAG standard, which interfaces with a Control and Status Register (CSR) block via an AMBA AHB channel. The overall chip architecture, including this configuration logic, is discussed in Section 5.5. The majority of the chiplet design was implemented in Verilog, with synthesis performed using the Design Compiler (DC) from Synopsys. Certain specialized blocks—such as a custom-designed Electrostatic Discharge (ESD) protection block for the SuperCHIPS IO—were developed using Synopsys Custom Compiler, with timing models generated via NanoTime. For physical design, Fusion Compiler was primarily used for placement, clock planning, routing, and final wire/via routing optimization. Functional verification at the testbench level was conducted using Synopsys VCS, while formal verification was carried out using Formality. Design Rule Checks (DRC) and Layout vs. Schematic (LVS) signoffs were initially performed in Fusion Compiler and subsequently re-verified using Mentor Graphics Calibre DRC standards. Finally, Static Timing Analysis (STA) was completed using PrimeTime prior to tapeout. The complete design and verification flow is summarized in Sections 5.4 and 5.5.

5.4 Design of SuperCHIPS input/output IP included in Anuvāda

The SuperCHIPS input/output (IO) drivers used in Anuvāda are an updated version of the work done by Siva Jangam [SJanD20]. The key differentiating factor for SuperCHIPS IO used in this work are –

1) Drive-strength has been increased to extend channel reach from 0.4 mm to 2 mm up to 4 Gbps data rates. Our previous results were published in [SJan20+].
2) Addition of ESD protection in the form of back-to-back diodes, to provide additional protection to any charge accumulation during assembly process.

The hard requirement for SuperCHIPS is to amortize the driver and ESD design area to within 10X10 μm^2 per IO, excluding clock domain crossing (CDC) circuits. Therefore, we adopt the simplest driver structure, i.e. even number of inverters for driving the IO link. Diode protection has been added, but as we will see below the back-to-back diodes occupy ~85% of the IO design area. This prevents any additional circuitry for equalization, error correction or impedance matching at source and destination of the SuperCHIPS link. The consequence of this is the limited reach of the channel, maximum up to 2 mm at 4 Gbps data-rate per IO to maintain a bit error rate (BER) of < 10^{-15}. However, considering SuperCHIPS is designed to be a highly parallel interface, even with data-rate per link of 4 Gbps, with a wire-density of 200 wires per layer per mm we can still achieve a high bandwidth given by:

SuperCHIPS synchronous bandwidth =

= 4 Gbps * 200 wires/layer/mm * 0.8 (*20% overhead: clock, control, redundancy*)

= 640 Gbps/layer/mm

For example, in a 6-layer Si-IF, where 4 layers are dedicated to signaling/wiring and 2 layers for power ang ground the bandwidth can be given by:

SuperCHIPS aggregate bandwidth (4 layers) = 640 * 4 Gbps/mm = 2560 Gbps/mm

In Summary – the Table 5.2 provides details on performance of SuperCHIPS against Signaling done on Interposers and PCBs

Table 5.2 Comparison between SuperCHIPS and other commercial protocols.

Interface protocol/Tech	***SuperCHIPS/Si-IF***		***UCIe1.1 /Interposer [U1,U2,U3]***	***BoW/Board [B1,B2]***	***SerDes /Board [S1,S2]***	***Improve-ment***
	Async, No ESD	***Sync, With ESD***				
Reach	Neighbor		Neighbor	Neighbor	Long Reach	
Bump pitch	10 µm		25 – 40 µm	130 µm	1 mm	4 – 100X
I/O depth (µm)	80		750[U3]	~400[B1]	1027	9 – 25X
Data-rate/link (Gbps)	10	4	2 – 16[U1,U2]	32[B1]	56-112	0.1 – 2X
Overall Latency (ps)	30	750	~ 2000[U1]	< 3000	~6000	3 – 65X
Energy/bit (pJ/b)	0.03	< 0.15	0.5 – 1[U2]	0.7[B1]	4 - 6.9[S1, S2]	5 – 40X
Bandwidth/mm (Gbps/mm)	8000[a]	2560[a,b]	1317	1760	149-298	4 – 23X

[a] 4 wiring levels, [b] Assuming 20% overhead

5.4.1 Driver design for SuperCHIPS link

The schematic of SuperCHIPS IO is presented in the Fig. 5.5 below. SuperCHIPS drivers are CMOS drivers with rail-to-rail (VDD-to-VSS) switching. The transmitter has four

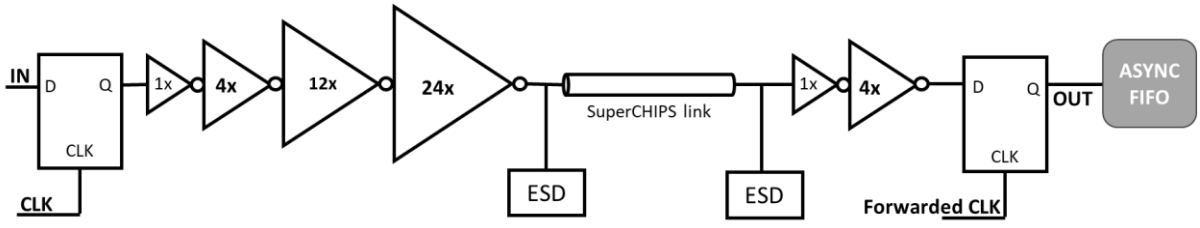


Figure 5.5 Circuit schematic of SuperCHIPS link. SuperCHIPS consist of simple inverter drivers for data transfer.

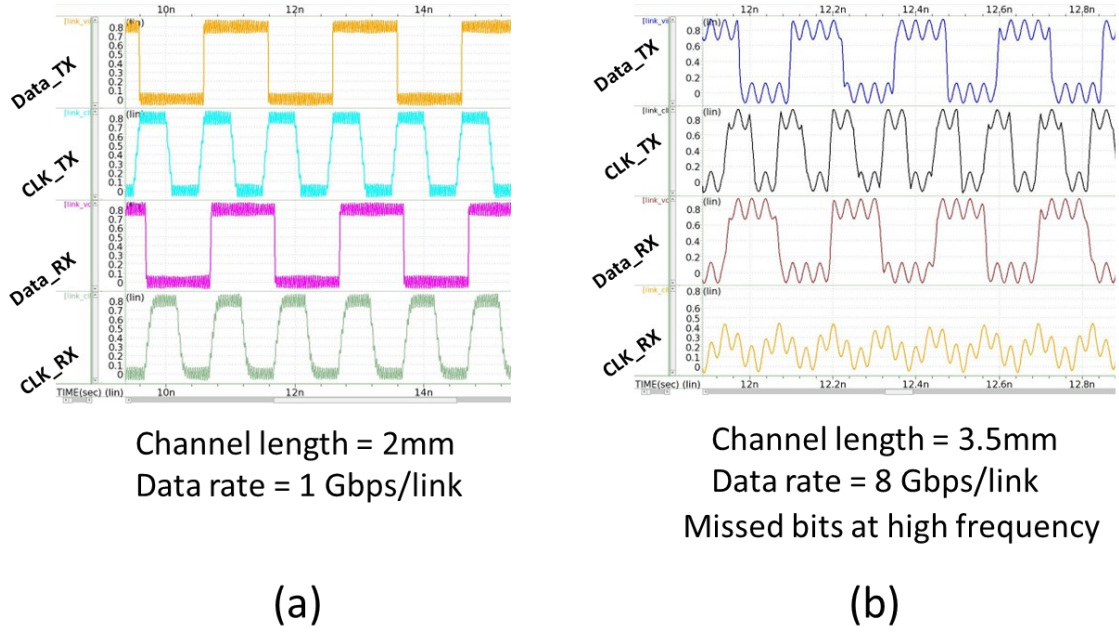


Figure 5.6 SuperCHIPS data and clock transmit and receive waveforms at the receiver for (a) 1Gbps/link, 2mm and (b) 8 Gbps/link, 3.5mm conditions. Note that at high speed, long links received clock is highly deteriorated.

inverters, sized as 1x, 4x, 12x and 24x. The sizing of the inverters was done in accordance with logical effort analysis [ISuth99] of the links. The inverters were then chosen from the 22nm standard cell library. This helped minimize the driver design area with respect to the desired drive strength for driving up to 2 mm long links.

The SuperCHIPS channel is a source synchronous channel, and the clock driver has the same design as data driver. On the transmitter side of SuperCHIPS, transmitted data is sampled at the rising edge of the clock however, on the receive size, received data is sampled at the falling edge of the clock. This ensures the clock edge falls approximately at the center of the data eye. The transmitted and received waveforms for data (DATA) and clock (CLK) at the transmitter and receiver are shown in Fig. 5.6 (a) for a link of channel length of 2mm and at data rate of 1 Gbps. The SPICE model used to determine these waveforms also includes inductive effect of bonds and chip and package decoupling capacitors to ensure simultaneous switching noise (SSN) is taken into consideration during the simulation. This also includes the effect of the ESD diodes, which will be discussed in the next section.

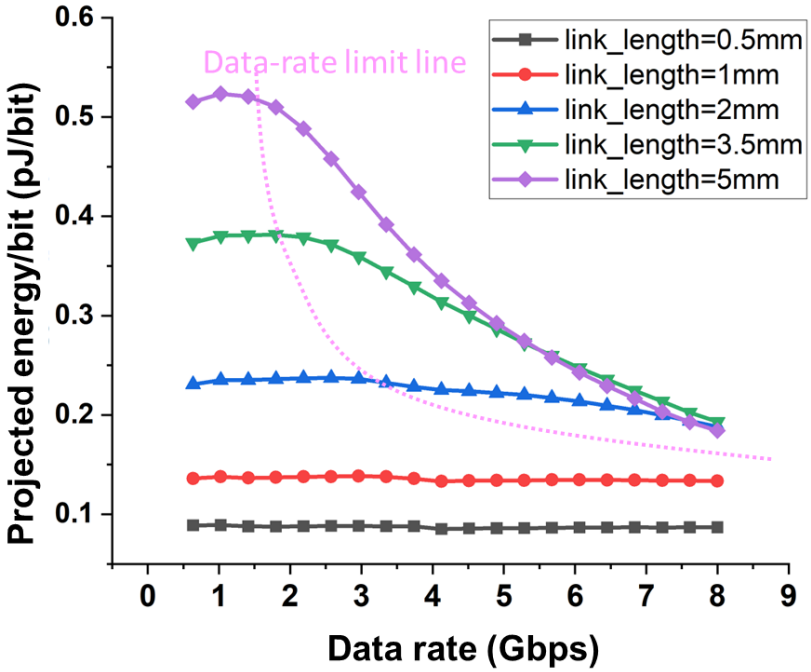


Figure 5.7 A projected energy per bit vs channel data-rate simulation for SuperCHIPS links. At high data-rates, a lots of bits are missed resulting in lower than expected energy consumed in the circuit.

Fig. 5.7 shows the projected energy per bit for different data rates and for different link lengths. In calculating energy per bit values, the energy consumed by all components of the link is first calculated by integrating the product of supply voltage (VDD) and current drawn from the supply over a given simulation time. This energy is then divided by the number of expected bits transferred in the simulation time. For example, for a link running at 2 Gbps for simulation time of 100 ns, the expected number of bits transferred is 100 ns x 2 Gbps = 200 bits. And energy per bit is total energy consumed during simulation time from the supply divided by 200 bits. It should be noted that expected number of bits transferred can be equal or greater than the real number of bits transferred, depending on whether some of the real bits are lost during the communication process. This could happen due to driver & channel properties, jitter, electromagnetic interference (emi) etc. With that in mind, let us interpret the results obtained in the projected energy per bit plot.

At short channel lengths, we see energy per bit is relatively flat with respect to increase in data-rate. Intuitively, energy per bit should be constant with an increase in data rate, as even though more bits are transferred with higher data rates, energy of data transfer also increases in proportion. However, we see energy per bit reducing for long links at high data rates. This seems non-intuitive at first glance. This is explained by considering we are dividing the simulated energy value not by real number of bits transferred, but by an expected number of bits transferred. When certain bits are not captured by the flipflop used in the transmitter design, the SuperCHIPS link is not charged, and no energy is absorbed/dissipated by the link or inverters. As a result, there is a drop in projected energy per bit. This drop simply indicates certain bits are either not sent by the transmitter or received by the receiver.

An important question then is—why is the link failing at longer link lengths and higher frequencies? The answer has to do with signal integrity and effect of simultaneous switching noise (SSN) particularly at high frequencies. All SuperCHIPS links can be treated as low pass filter like in behavior and hence will filter out high frequency components of the square waves transferred through them. For a link, a good rule of thumb for minimum channel bandwidth for a square wave of rise-time/fall-time of tr is given as:

Channel BW = $0.35/t_r$, where t_r = signal rise time

At higher frequencies, rise-time is shorter, and channel bandwidth requirements are more stringent. Hence many of the high frequency components of the square waves will be filtered out by the channel, causing the square wave to widen in time domain at the receiver, causing inter-symbol interference (ISI) and therefore, signal integrity challenges. There is no pre-emphasis, de-emphasis or equalization in SuperCHIPS IO to compensate for ISI and channel loss. Another concern is the lack of termination. Reflections are not a concern at low frequencies, as they get absorbed in the rise-time/fall-time of the signals carried by the SuperCHIPS links. But at higher frequencies, the reflections will directly interfere with the clock or data waveform causing overshoot and ringing. This can lead to misinterpretations of voltage levels at both transmitter and receiver side. A third challenge is Simultaneous switching noise (SSN), which is caused due to inductance in the power

and ground path from the supply source and causes noise fluctuations in the waveforms. This effect is also called ground bounce. This type of noise, however, can be greatly reduced by including significant amounts of decoupling capacitance at different frequencies of operation. The Fig. 5.8 provides the amount of ground bounce in terms of (VDD – VSS) with respect to increase in the decoupling capacitance added to the chip. In this plot, the simulation was done at 1 Gbps data rate, for a 2 mm long link. The expected value of VDD – VSS should be 0.8 V, and maximum accepted ripple is 10% due to ground bounce. Increasing of decoupling capacitance will help reduce the variation from 0.8 V.

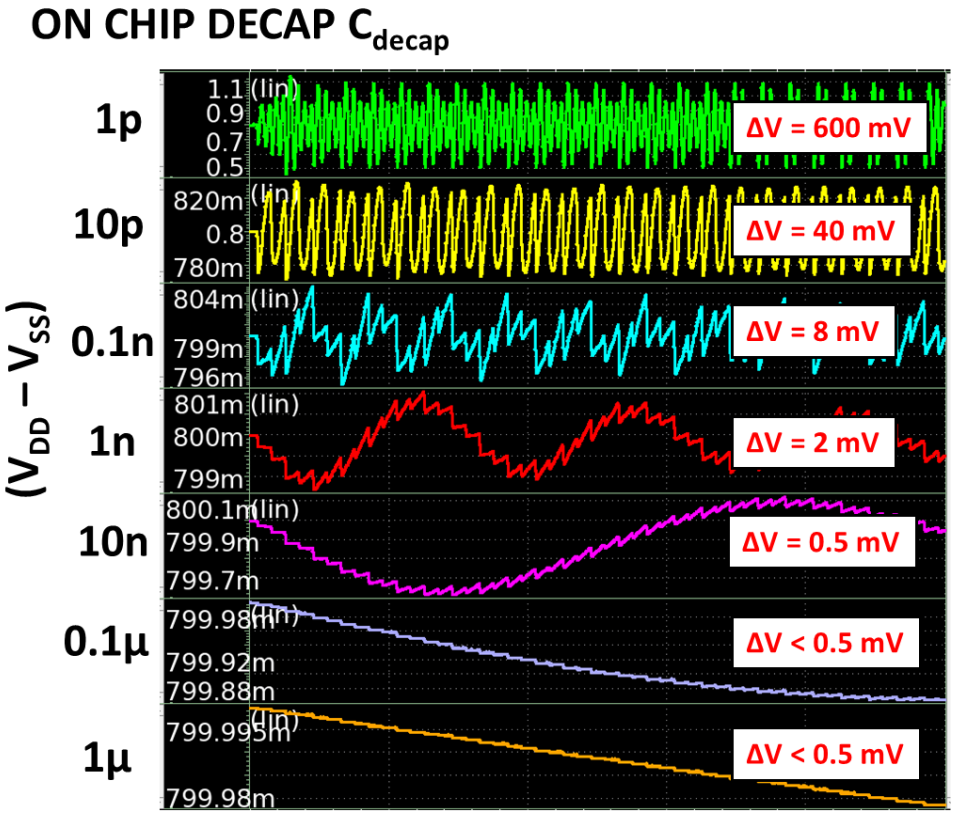


Figure 5.8 Variation of ground bounce ($\Delta V = V_{DD} - V_{SS}$) at different levels of decoupling capacitance.

Decoupling capacitors are introduced whenever possible both at the dielet design level as well as on the Si-IF and on the break-out PCB design. The details of decoupling capacitance insertion are not discussed here; however, the reader is referred to [DHyu25,

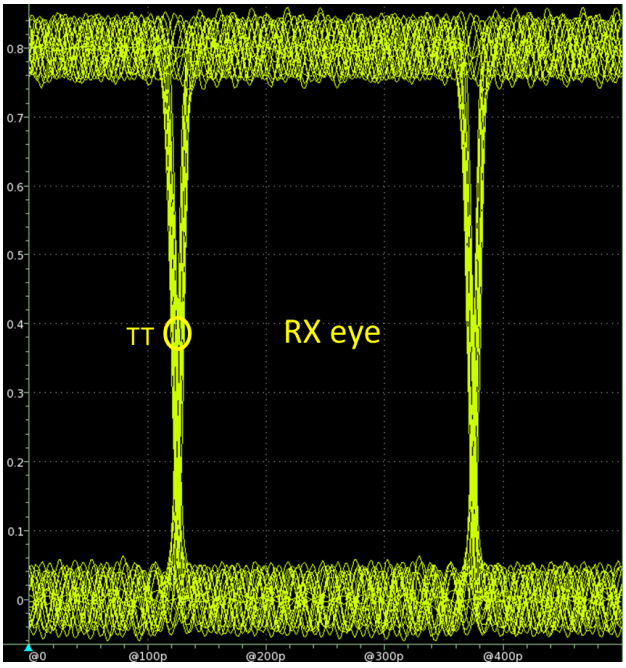


Figure 5.9 Eye diagram of data eye at the SuperCHIPS receiver at typical corner.

DecapWiki, DecapAlt] for understanding the capacitance calculation and insertion procedure. For this work, the summary of decoupling capacitor insertion in Anuvāda is as follows:

- After VRM (voltage regulator module) multiple decoupling capacitors of size 10µF, 4.7µF, 100nF and 10nF are added around the Anuvāda chip over the PCB which is used to test the chip.
- In the case where Anuvāda is further mounted on a Si-IF, there is provision to add an additional 100 nF surface mount (0201) capacitor on Si-IF.
- The extracted (Calibre PEX) on-chip decoupling capacitance is approximately 60nF. This includes gate-to-channel MOSCAPs in 22nm technology.

Fig. 5.6 (b) shows a SuperCHIPS data and clock link running at 8 Gbps, for a 3.5mm long link. We can observe the clock is unable to have a full rail to rail switching due to low pass filter like behavior of the SuperCHIPS link. Further there is significant ringing on the Clock signal, which can be attributed to lack of termination in the SuperCHIPS IOs. As the clock

is not well defined, data cannot be reliably captured, resulting in unacceptable error rates. This ultimately limits the capability of SuperCHIPS link.

In summary, SuperCHIPS is a simple, low power, low area, high bandwidth IO. The energy per bit for a 1 mm SuperCHIPS link is about 0.13 pJ/bit. The high bandwidth advantage of SuperCHIPS comes from its small IO pitch of 10μm and below. We believe on a wafer-scale system, SuperCHIPS is one of the most efficient die-to-die links which will reduce the overall communication power. Fig. 5.9 shows the simulated eye diagram of SuperCHIPS link at the receiver carrying data over 1 mm long link running at 2Gbps data rate.

5.4.2 electrostatic discharge (ESD) protection circuitry for SuperCHIPS IOs

Input/Output (IO) bonding pads can accumulate static charges during transport or even during the assembly process. An Electrostatic Discharge (ESD) event [ESDFu] occurs when a sudden flow of electricity happens between two objects with different electrical potentials. This typically occurs when a charged object (such as a person or tool) comes into contact with or comes close to the dielet, causing a rapid transfer of static charge. ESD events can generate voltage levels in the range of hundreds to thousands of volts, while semiconductor devices typically operate at only a few volts. The sudden surge of current can exceed the device's tolerances, damaging delicate transistor gates and interconnects. The following effects are commonly caused by ESD events:

- Oxide Breakdown: The transistor gates in a semiconductor are insulated by an extremely thin oxide layer (often just a few nanometers thick). An ESD event can apply enough voltage to break down this insulating layer, causing permanent damage or due to a short between gate and channel.
- Thermal Damage: The rapid discharge of static electricity can create a localized heating effect. This can melt or fuse internal connections in the silicon dielet, making the chip unusable.
- Latch-up Effect: Some ESD events can trigger a parasitic conduction path within the die, leading to a condition called latch-up, where the chip gets stuck in a high current state, potentially burning out the device.

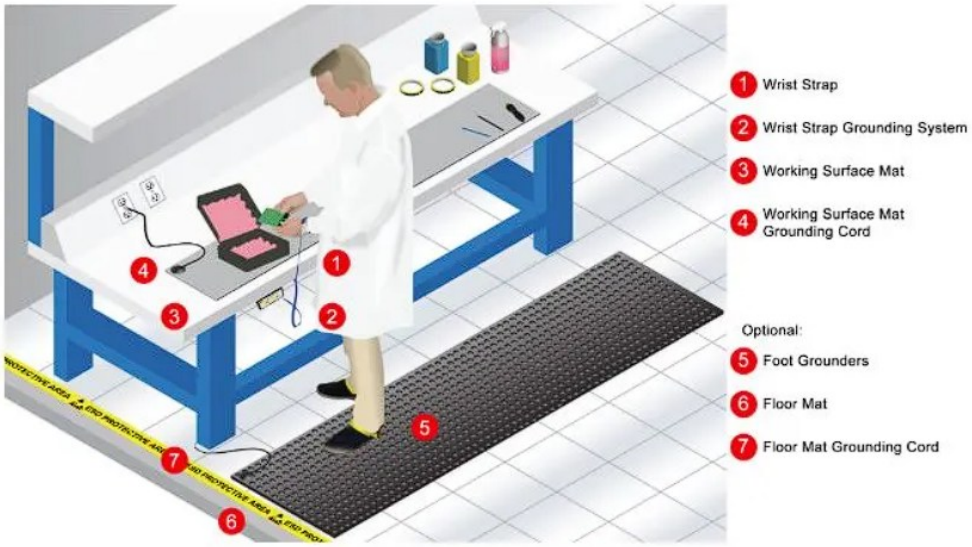


Photo Courtesy: SCS

Figure 5.10 Standard operating procedure when working with ESD sensitive dielets. [ANagu25]

- Degradation and Latent Defects: Not all ESD damage is immediately catastrophic. Some events cause minor degradation, leading to long-term reliability issues. This can cause premature failure of chiplet/dielet in the field.

Preventing damage to internal circuits due to ESD events is therefore crucial to the operation of any chip. We would like the reader to note ESD events are physical events, not within anyone's control. Preventative measures must be taken to reduce damage from ESD. These measures are in the form of wearing ESD straps during testing, using ESD certified boots, gloves and working in an ESD safe environment. During thermal compression bonding, we have always ensured to take the above measures, and in addition, there are three ionizers on the bonding tool, which help remove any buildup of charges during die transfer, alignment and bonding. While testing the chip, we have ensured to wear straps, use ESD safe mats and overall take precautions as shown in the Fig. 5.10 below taken from [ANagu25].

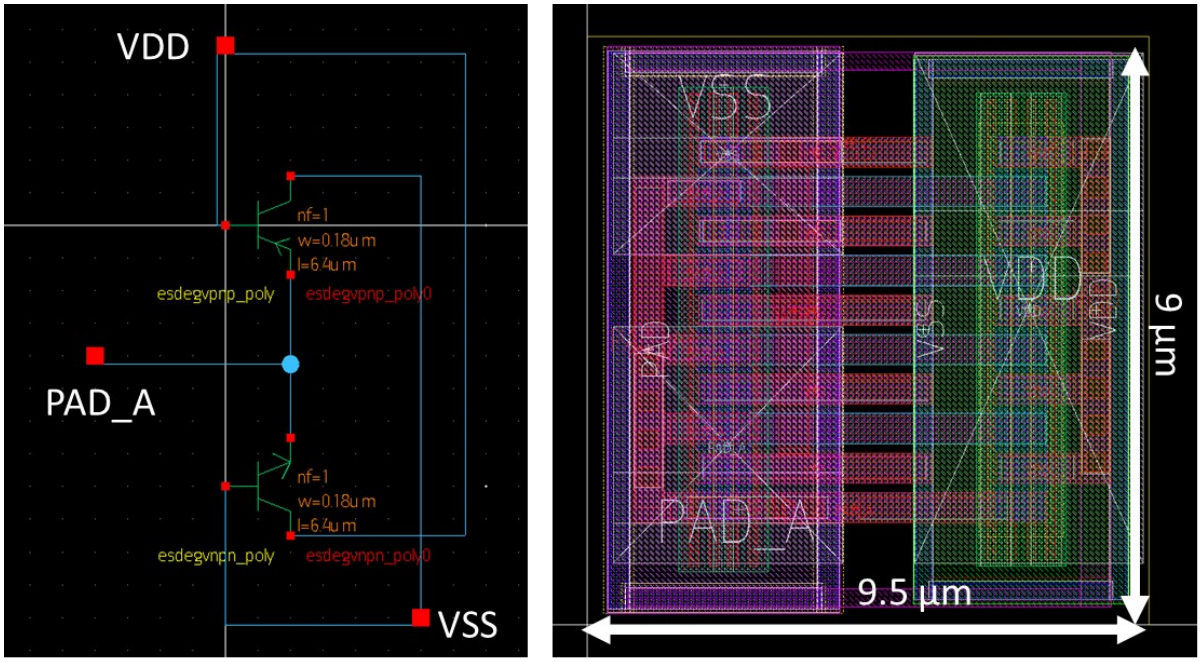


Figure 5.11 Schematic of the ESD cell used in each SuperCHIPS IO. (b) Layout of ESD cell used in SuperCHIPS IO.

Even with all the preventative measures during bonding and testing, ESD protection circuitry has been added to Anuvāda during design phase. SuperCHIPS has custom designed IOs and therefore, care needs to be taken to increase protection against ESD events. As such, two poly-bound diodes are connected to the input/output terminals of the SuperCHIPS IO as shown in the schematic in Fig. 5.11 (a). The size of ESD diode however is limited by the allotted design area of 100 μm^2, and the requirements of ESD parameterized cells (PCELLS) in the 22nm device library from GlobalFoundries. The schematic and layout of the ESD cell used in SuperCHIPS IO is presented in Fig. 5.11.

There are two vertical BJTs being used as base-emitter diodes. The up-diode is a vertical PNP bi-polar device used as a P+/NW/Sub diode, where P+ stands for heavily doped P-type and NW stands for N-type well contact. The substrate used is lightly doped P-type. The down-diode is a vertical NPN bi-polar device used as a N+/P+/NW diode, (where diode is N+/P+). Both the diodes are identical in dimensions with gate length 6.4 µm, and

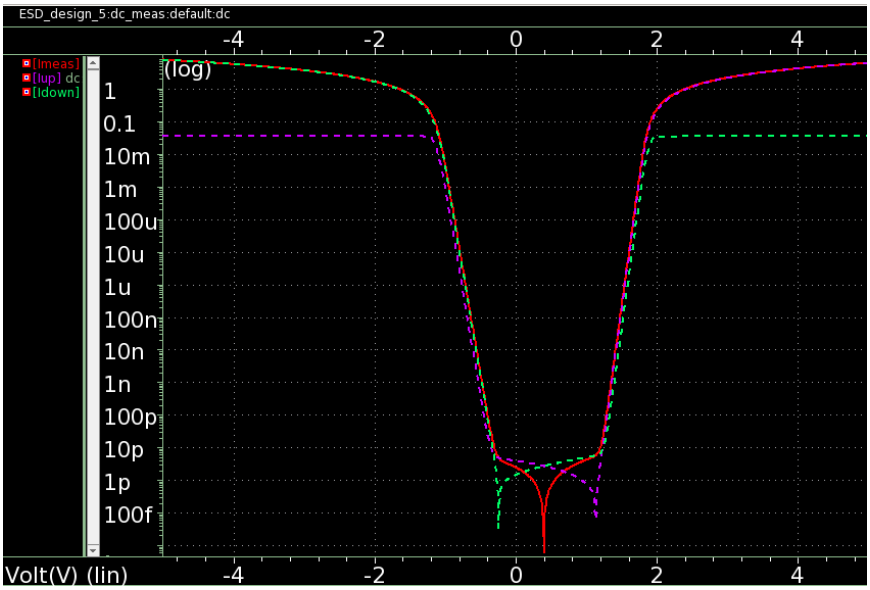


Figure 5.12 DC characteristics of ESD back-to-back diodes used in SuperCHIPS IO. The operation range of the ESD diodes is in range of -0.4 V to 1.3 V.

overall area occupied by the two diode ESD cell is ~ 85 μm^2, which takes up about 85 % overall space for the ESD protection from the 100 μm^2. The layout is represented in Fig. 5.11 (b). The DC characteristics, and hence the turn-on voltages of the diode are represented in plot in Fig. 5.12. The red curve indicates the current through the back-to-back diodes with voltage sweep. The range of diode operation is approximately -0.4 V to 1.3 V, meaning one of the diode (up or down) gets forward biased and sinks the excess current without affecting the internal circuits.

The ESD performance of these diodes however can be estimated through two tests:

- Transmission line pulse (TLP) pulse test: TLP involves applying fast, high current pulses like real electrostatic discharge (ESD) events. It generates a pulsed IV curve that helps designers understand how ESD protection diodes, clamps, and I/O buffers respond to transient state. It provides quantitative data on how circuits behave under ESD-like stress conditions, making it an industry standard.
- Latch up: Latch up is a failure mode in a CMOS circuit caused by parasitic thyristor (PNPN) structures which can cause uncontrolled high current flow, potentially

damaging the device. A latch-up test characterizes the device's robustness to overvoltage and overcurrent conditions. However, it is not a direct or complete measure of the ESD sensitivity of the IO. Even so, it is a very useful test as it can provide insights into the circuit's susceptibility to ESD events. Latch up reveal vulnerabilities in a circuit which can be exacerbated by ESD events.

For Anuvāda testing, the IOs were subjected to TLP testing, and the procedure and results will be discussed in section 5.9.4. We expected TLP tests to be able to accurately demonstrate the capability of the IOs.

5.4.3 SuperCHIPS FIFO and control signals.

In Anuvāda, the SuperCHIPS IO are not driven at their maximum capability. Anuvāda receives high speed PCIe 3.1 inputs at 8 Gbps and translates it into 16 outputs which undergo 8b/10b encoding to result in 20 SuperCHIPS outputs. The clock frequency of SuperCHIPS IO is therefore 500 MHz. Similarly, the chip can receive 20 SuperCHIPS 8b/10b encoded inputs and convert it to a high speed 8 Gbps output.

SuperCHIPS has an asynchronous FIFO (first in first out) at the receiver required for clock domain crossing between signals being received at the receiver inputs and the internal PCIe clock. There are two important control signals related to the FIFO, which also serve as the control signals for SuperCHIPS in Anuvāda. The depth of the FIFO is 16 and its role is to ensure data can be written by received source-synchronous clock to the FIFO and read by internal PCIe CLK asynchronously. Both the received SuperCHIPS clock, and internal PCIe clock (also called PCLK) run at 500 MHz. Therefore, they are identical in frequency, but they have no phase relationship. As frequency is identical for both clocks, we expect FIFO depth of 16 for each IO is sufficient. The overall SuperCHIPS signals are described in Table 5.3 below.

Table 5.3 SuperCHIPS channel data and control signals

Signal Name	Signal Type	Function
SuperCHIPS Transmitter		
Data_Out[19:0]	Output	8b/10b transmitted outputs

CLK_out	Output	SuperCHIPS clock output (500MHz) single ended
	SuperCHIPS Receiver	
Data_In[19:0]	Input	8b/10b received inputs
CLK_in	Input	Received clock
fifo_empty	Output	No received data in FIFO
fifo_full	Output	SuperCHIPS FIFO is full, and further incoming data may be lost

5.5 Overall Anuvāda Chiplet design

In this section, we will summarize the overall design process of building Anuvāda from conception to final tapeout. The tapeout followed Synopsys design flow and Synopsys design tools. For DRC and LVS operations, Calibre was used. The design was tested using test-bench at individual block level as well as full chip level, gate level and post routing at static timing analysis level. The chip architecture is presented in Fig. 5.13. The major components of the chip are:

- Data-loader module
- 6.4 kB SRAM block
- SuperCHIPS block
- PCIe 3.1 block

In the following paragraphs, we will explain these components briefly.

5.5.1 Anuvada data-loader module

The purpose of the data loader module is to configure the PCIe 3.1 IP and SuperCHIPS IP for operation. The schematic of the data-loader module is shown in Fig. 5.13 on left side. It consists of a JTAG IEEE 1149.1 Standard [JTAGS] for loading configuration data

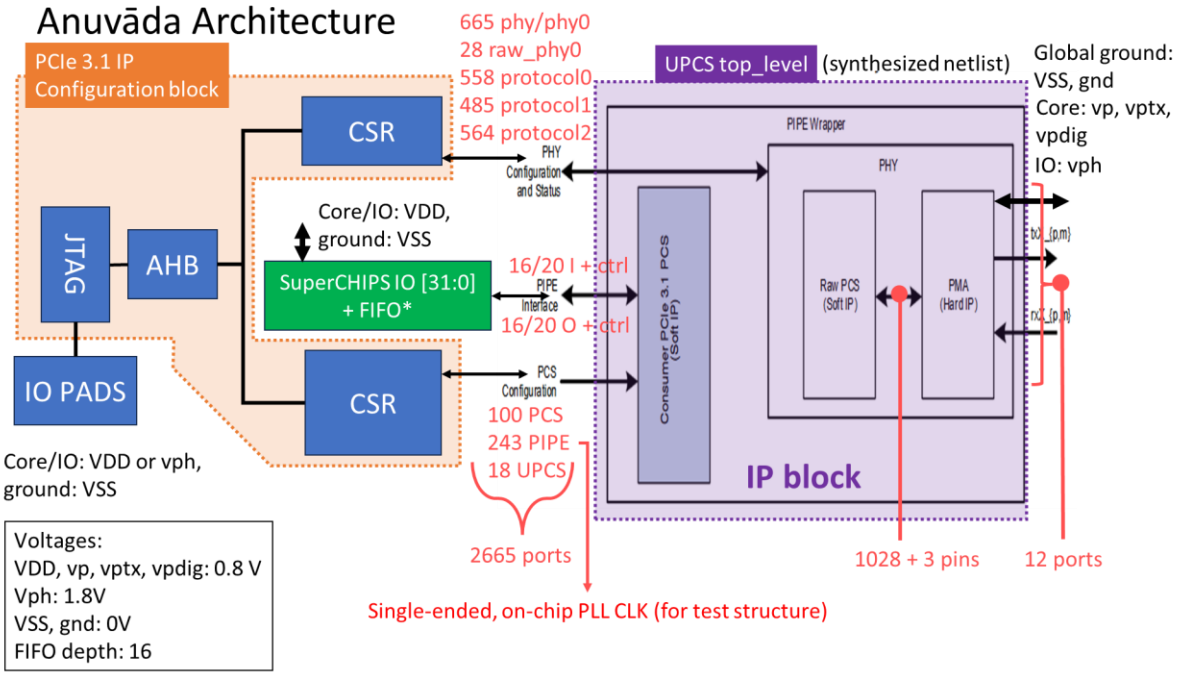


Figure 5.13 Architecture of Anuvāda chiplet.

into the chip or reading data from the chip. The standard signals are included in Table 5.4. The data can be loaded into a set of registers called control and status registers (CSR). The CSR is directly connected to the inputs of the PCIe 3.1 block. The data loaded into the CSR block is used to configure the PCIe inputs such as setting the operation mode, data-rate of operation, setting the phase locked loop (PLL) configuration of the PCIe, setting the data-width, equalization gain, power-state, power gating signals etc. It is also possible to read the status outputs of the PCIe through JTAG. The exact state diagram of the Test Access Port (TAP) controller and operation of JTAG is not discussed here, but it can be found in reference documents [JTAGT].

Table 5.4 Signals used in Anuvāda dataloader block for configuration of PCIe block.

JTAG IO	Purpose
TDI	Test Data Input, used to serially send a 32-bit data word to the CSR block

TDO	Test Data Output, used to serially output 32-bit data stored in the data register
TCK	JTAG clock (1 MHz single ended clock for Anuvāda)
TRST	JTAG reset
TMS	Test Mode Signal, used to control the transitions of the Test Access Port (TAP) controller, a state machine that manages JTAG operations.

The communication between JTAG and CSR is done through AMBA-AHB interface. AMBA AHB is a bus interface suitable for high-performance synthesizable designs. We use a 32-bit parallel AMBA AHB interface to transfer control and data signals between the JTAG and CSR block. All the AHB interface control signals are internal to the data loader module and cannot be accessed externally to the chip.

The Verilog codes for JTAG, AMBA-AHB, CSR block were coded from scratch with instructions from the reference manuals, and no external code was used. These codes are open-source and available at GitHub.

5.5.2 6.4 kB SRAM block

The PCIe IP required data to first be loaded to an internal SRAM block before loading it into PCIe digital block – also referred to as the Physical Coding Sublayer (PCS). This SRAM was implemented using INVECAS SRAM IP available through global foundries. Detailed documentation of the SRAM block will be available on GitHub.

5.5.3 SuperCHIPS IP

There is one SuperCHIPS interface present on Anuvāda. The layout is done such that there is a 600μm gap between SuperCHIPS transmitters and SuperCHIPS receivers. SuperCHIPS links can connect between the row of transmitters and receivers on the same site on Si-IF if required. This channel is represented in Fig. 2.2 (c) of section 2.2. Also a 2 mm long channel between two Anuvāda dielets was fabricated on a similar Si-IF to show communication between two dielets on Si-IF. Details on SuperCHIPS design

process, driver and ESD has already been discussed in Section 5.4.2, and we refer the reader to this section for more details.

5.5.3 PCIe block and implementation

Anuvāda consists of one PIPE 4.3 compliant PCIe 3.1 protocol obtained from Synopsys, and one SuperCHIPS interface. The SuperCHIPS interface acts as a surrogate for MAC (media access control) for the PIPE 4.3 compliant PCIe 3.1. The SuperCHIPS interface uses a repeated/buffered version of PCLK (the synchronized clock signal for parallel data conversion in PCIe 3.1) as a clocking signals. The PCLK frequency is 500MHz. The PCIe 3.1 block has two parts:

- PMA: full duplex transceiver/SerDes (mostly analog)
- PCS: To configure the PMA block (purely digital block)

The PMA block is a hard macro, with no design or modification required. However, the physical coding sublayer i.e. PCS block was provided by Synopsys as a soft IP, which needs to be designed and tested using the technology node used by the designer. In this case, it is GlobalFoundries 22nm FDSOI technology node. This testing was done using

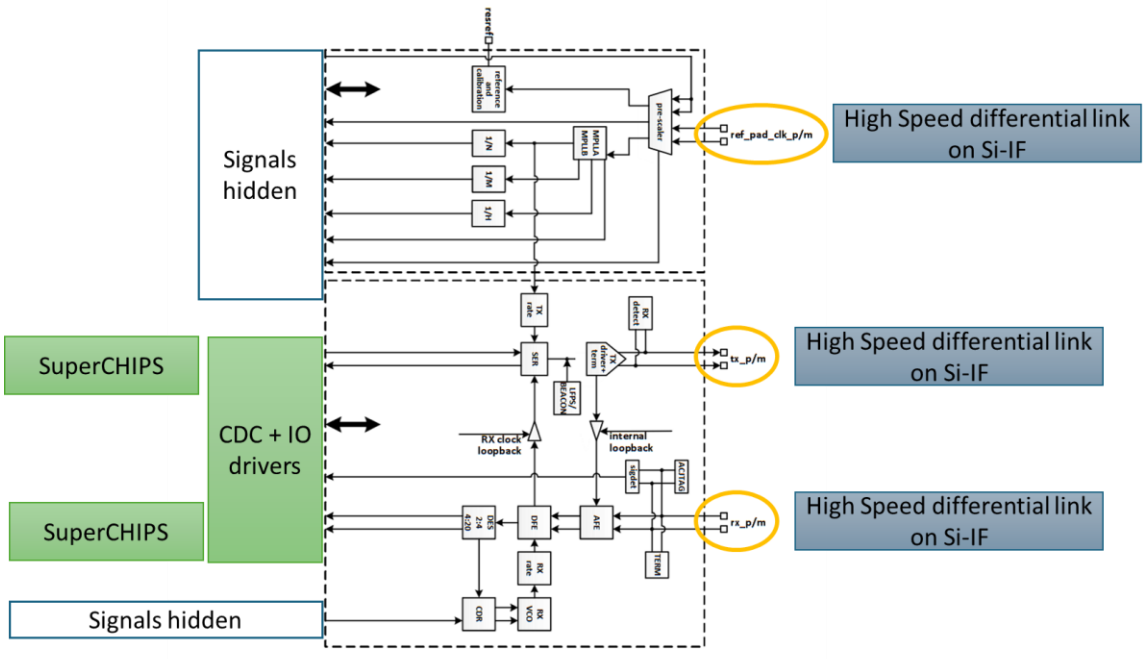


Figure 5.14 Schematic representation of PCIe 3.1 Architecture.

both testbench provided by Synopsys, followed by testbench written in this work. Necessary test benches will be included on GitHub. The PCIe 3.1 block simplified architecture is presented in Fig. 5.14.

There are two important user design changes for the PCIe 3.1 design in Anuvāda. The first one has to do with hand placement / manual placement of Clock Domain Crossing modules (CDC) between some of the high speed and slow speed regions of the PCIe 3.1 IP. These CDC blocks were not pre-placed by Synopsys during the design of the IP, rather the designer must come up with the right cells to be used during synthesis. An example of this is a three flip-flop (FF) synchronizer. While it is possible to use three standard flip-flops for doing this synchronization, there is high risk of metastable state propagating from the first FF which is in the source-synchronous clock domain and third FF which is in the current-chip clock domain. Instead, we selected appropriate 3 FF synchronizers already available in the standard cell technology library to do the CDC. These synchronizers are designed and layout to ensure the lowest MBTF (mean time to failure) for metastability. Many of the RTL documents were modified to now include the selected synchronizers. Logical equivalency was established through test bench tests. One example is shown in Fig. 5.15.

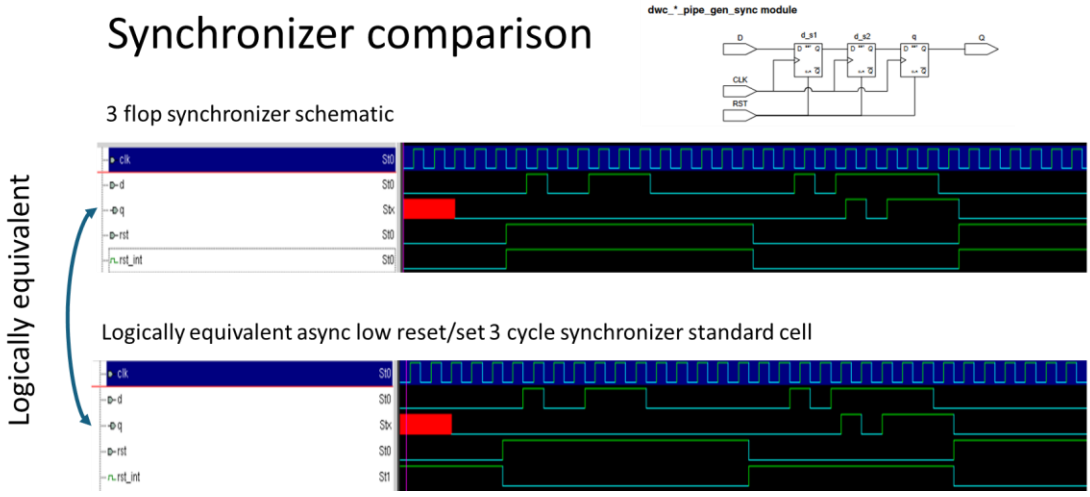


Figure 5.15 Hand placed blocks used in PCIe soft macro and their simulation to verify functionality.

The second design change involves is the CDC including transfer of control signals between PCIe and SuperCHIPS. The PCIe data and receiver valid signals communicate with the read and write enable signals on the SuperCHIPS FIFO, for letting the FIFO know when the PCIe 8b/10b signals will be received at the receiver. The same control signals are also transferred through the channel for another SuperCHIPS receiver to receive them specifically to do a loopback through the SuperCHIPS.

The control signal transfer through SuperCHIPS for loopback in Fig. 5.16 below.

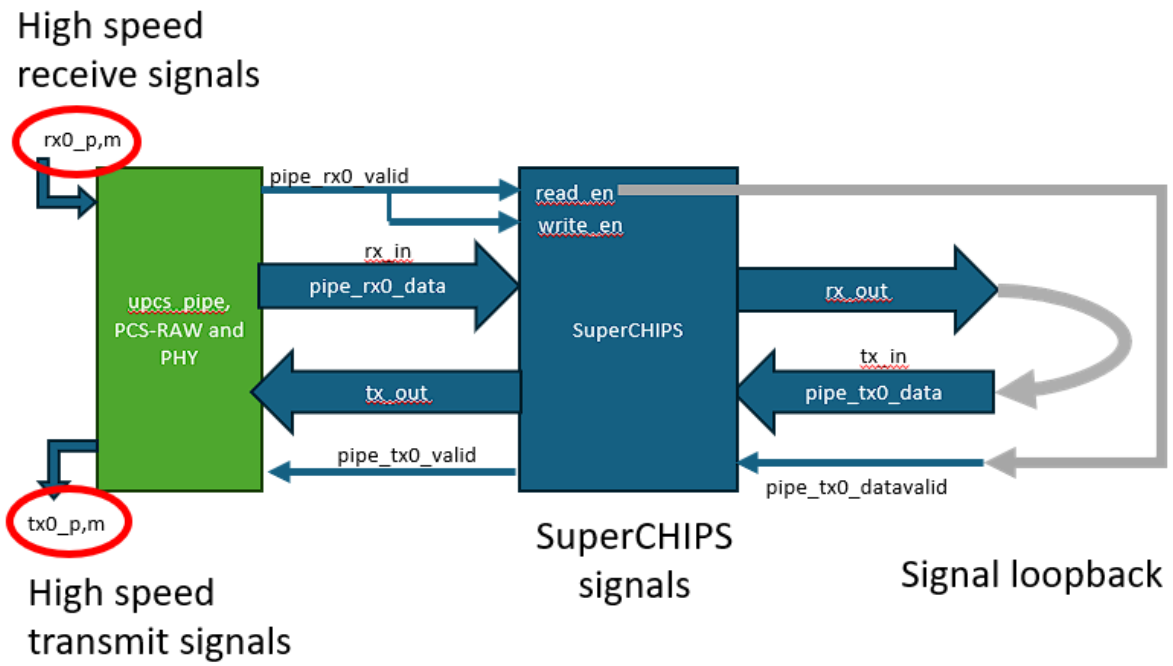


Figure 5.16 Control signals sent from PCIe block to SuperCHIPS block for loopback operation. Note that SuperCHIPS is agnostic to soft protocols, and it forwards the control signals from the PCIe receiver to PCIe transmitter interface.

The PCIe 3.1 block was tested both with Synopsys provided testbench, user written testbench, and after layout and parasitic extraction.

5.6 Anuvāda design, simulation and tape out

Fig. 5.17 shows the final layout from fusion compiler after P&R and Signoff. Notice that the SuperCHIPS IO design area is 100x smaller than the PCIe design area for the same data bandwidth. This will be more apparent in Fig. 5.22 (a), which shows the figure of

Anuvāda after it was fabricated at GlobalFoundries. The table 5.5 represents a summary of Anuvāda chip level signals IO pads necessary to understand the Anuvāda operation. Some of the signals are bunched together to indicate they are not considered important to demonstrate Anuvāda operation and are primarily used for debugging purposes.

Table 5.5 Anuvāda top level pad signals.

Signal Name	Type	Description
TCK	I	JTAG clock (1 MHz single ended clock for Anuvāda)
TMS	I	Test Mode Signal, used to control the transitions of the Test Access Port (TAP) controller, a state machine that manages JTAG operations.
TDI	I	Test Data Input, used to serially send a 32-bit data word to the CSR block
TDO	O	Test Data Output, used to serially output 32-bit data stored in the data register
TRST	I	JTAG reset (active low)
phystatus	O	PCIe status signal. 1 = PCIe not configured and 0 = PCIe is configured
power_present	O	PCIe power present signal
SuperCHIPS_reset	I	To reset the SuperCHIPS block and clear the FIFO
phy_reset	I	PCIe reset signal
phy_resref	I	200 ohm precision resistor is connected to the phy_resref (phy resistor reference)
ref_clk_p/m	I	Reference clock for PCIe 3.1 (100 MHz)
tx_data_p/m	O	8 Gbps data transmitted from the PCIe transmitter
rx_data_p/m	I	8 Gbps data received by the PCIe receiver
VDDIO	I	IO voltage = 1.8 V

VDD, vpdig, vptx0	I	Digital core voltage = 0.8 V
vph	I	Analog IO voltage = 1.8 V
vp	I	Analog core voltage = 0.8 V
VSSIO, VSS, gd	I	Ground
pcie_dc_jtag	I/O	PCIe JTAG
pcie_ac_jtag	I/O	AC boundary scan for PCIe

Simulations of Anuvāda was done at 4 levels –

- **Block level:** RTL and post-synthesis simulations for SuperCHIPS, data loader block, PCIe 3.1 was done prior to full chip integration
- **Full chip RTL:** after integration of PCIe 3.1, Anuvāda data loader block, SRAM block and SuperCHIPS block at RTL level, a full loopback was tested.
- **Full chip Gate Level:** post synthesis, the design passed formal verification by formality to ensure no change to the ports, missing clocks or other terminals and then loopback was tested at gate level. In this case, do delay or parasitic back-annotation was used, to maintain a quick testing and result turn around.
- **Full chip post-Static Timing Analysis (STA) level:** Using extracted netlist from the layout, with addition of path delays in SDF format generated by PrimeTime and parasitic back annotated values in SPEF format, a full chip simulation of loopback was again done to verify the desired paths were operational.

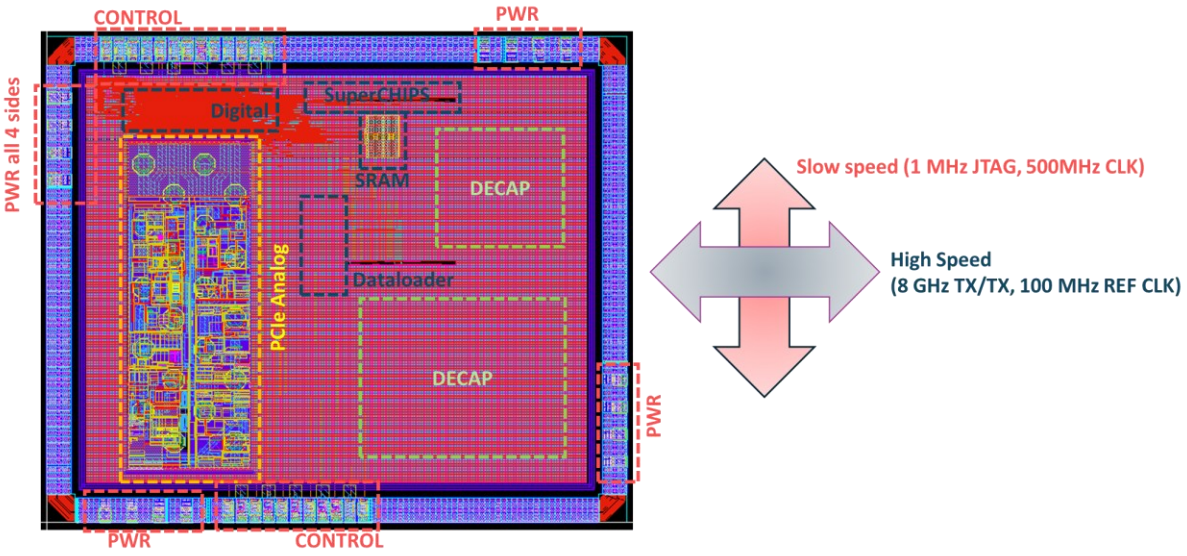
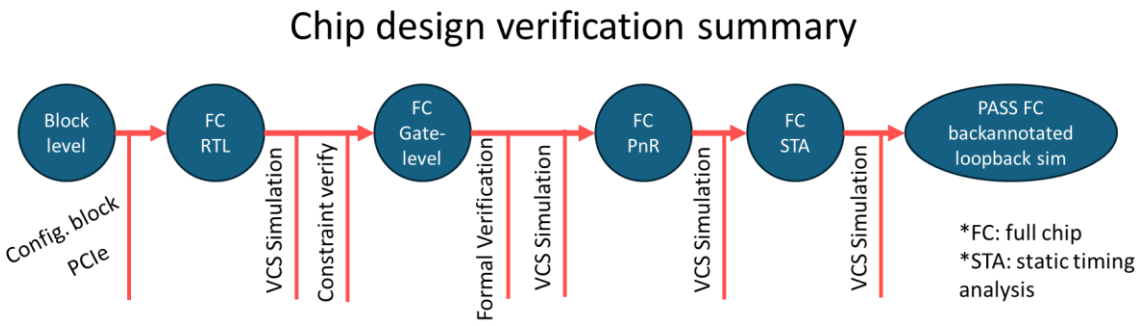


Figure 5.17 The layout of Anuvada from fusion compiler. The slow speed signals are accessed from the top and bottom of the dielet and the high speed signals are accessed from the right and left of the dielet.

A summary of testing is shown in Fig. 5.18.

For full chip level testing, the test bench involves a sequence of activities including IP power up, PCIe 3.1 configuration to loopback test. The sequence is as follows

1. Power up the VDDIO, VSSIO supplies which are 1.8 V digital IO voltage supplies. This will also power up analog voltage supply of the PCIe IP i.e. vph = 1.8 V.
2. Observe phystatus == 1. Assert signal phy_reset = 1 to keep the PCIe IP in reset mode.
3. Power up the VDD, VSS supplies which are 0.8 V digital core voltage supplies. This will also power up analog and digital voltage supplies of the PCIe IP i.e. vp, vptx0 and vpdig.
4. Use the JTAG to load PCIe 3.1 configuration to the chip.
5. Dessert phy_reset.
6. Observe phy_status fall (from 1 → 0) in about 20 ms.



	SuperCHIPS	JTAG+AHB+CSR	SRAM	PCIe	Full-chip Loopback
SPICE	YES	--	--	--	--
RTL	PASS	PASS	PASS	PASS	PASS - with limits of not having a PCIe controller
Gate level	PASS	PASS	--	--	PASS - with limits of not having a PCIe controller
Post-STA with Primetime SDF (typical corner)	--	--	--	--	PASS - with limits of not having a PCIe controller

Figure 5.18 Summary of all testing and verification done on the Anuvāda chiplet.

7. Change the powerdown state of the PCIe to 2'b00. This will cause the PCIe IP to start giving random 8 Gbps signals through the outputs.
8. Send a PCIe 3.1 packet through the high speed PCIe receiver.
9. This signal should be looped back through the SuperCHIPS.
10. Receive PCIe 3.1 packet through the high speed PCIe transceiver.

This loopback function is demonstrated to be working in simulation after RTL level, Gate level and after Static Timing Analysis (STA). For this the standard delay format (SDF) file generated by PrimeTime is used for delay annotation. The post-STA simulation results are shown in Fig. 5.19 which shows loopback. As we can see from the simulation results, about 88,557 words were transmitted by a PCIe signal generator block (not shown) to the Anuvāda DUT (device under test), which were received, translated, sent through SuperCHIPS IP and looped back to the input of PCIe transmitter form where they were serialized and transmitted back to the PCIe signal generator block for comparison. Approximately 20,000 words are used up during the initial calibration phase, hence only 61,216 words were ultimately verified. However, there are no errors in transmission.

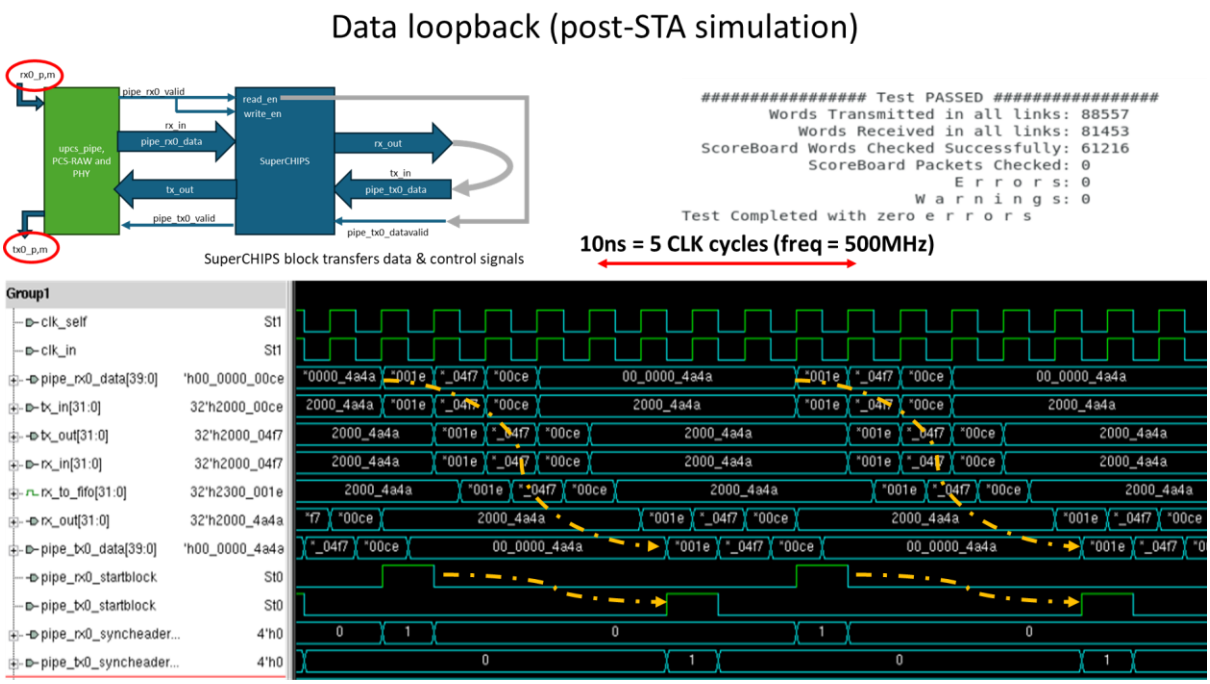


Figure 5.19 Post static timing analysis (STA) data loopback showing translation in action.

The STA reports show no violating paths in worst case (setup), best case (hold) or typical scenarios as ween in the Fig 5.21. Also, the functional verification report shows successful verification.

The final step is clearing the physical design rule checks (DRC) set by the foundry. In case of Anuvāda, due to the use of 10µm IO, the chip did not pass all the ESD level checks. These tests fail due to not having enough emitter area in use of up and down diodes. Since these IOs are intended, we asked GlobalFoundries for waivers on these ESD level checks and accepted the risks with the fabrication.

5.7 Anuvāda post tape out and Si-IF fabrication

The die-shot of the fabricated chiplet in GlobalFoundries 22nm FDX node is presented in Fig. 5.22 (a). As seen in the figure, power is provided from all 4 edges of the dielet, though it is still possible to power up the die by accessing power and ground pads on single edge. The location of the PCIe 3.1 IP from Synopsys and the custom designed SuperCHIPS IP can also be seen. The Size of SuperCHIPS IP and PCIe 3.1 IP is shown in the table below:

```
******************************** Verification Results ********************************
Verification SUCCEEDED
   ATTENTION: RTL interpretation messages were produced during link
              of reference design.
              Verification results may disagree with a logic simulator.
-------------------------------------------------------------------
 Reference design: REF:/WORK/anuvada_chip_top
 Implementation design: IMP:/WORK/anuvada_chip_top
 13427 Passing compare points
-------------------------------------------------------------------------------------
Matched Compare Points     BBPin    Loop    BBNet     Cut     Port     DFF     LAT   TOTAL
-------------------------------------------------------------------------------------
Passing (equivalent)        1111       0        0       0       15   12270      31   13427
Failing (not equivalent)       0       0        0       0        0       0       0       0
Not Compared
  Constant reg                                                          8     601     609
  Unread                       0       0        0       0        0     592      16     608
*************************************************************************************
Removed container 'IMP'
Information: Implementation design is no longer set (FM-071)
Removed container 'REF'
Information: Reference design is no longer set (FM-070)
Cleared container 'i'
Cleared container 'r'

Maximum memory usage for this session: 7959 MB
CPU usage for this session: 3235.56 seconds ( 0.90 hours )
Current time: Tue Jun  4 02:43:40 2024
Elapsed time: 2435 seconds ( 0.68 hours )
Number of cores used: 8

Thank you for using Formality (R)!
touch ver16
[krutikesh_linux@vajra synth]$
```

Figure 5.20 Formal verification report for Anuvāda chiplet.

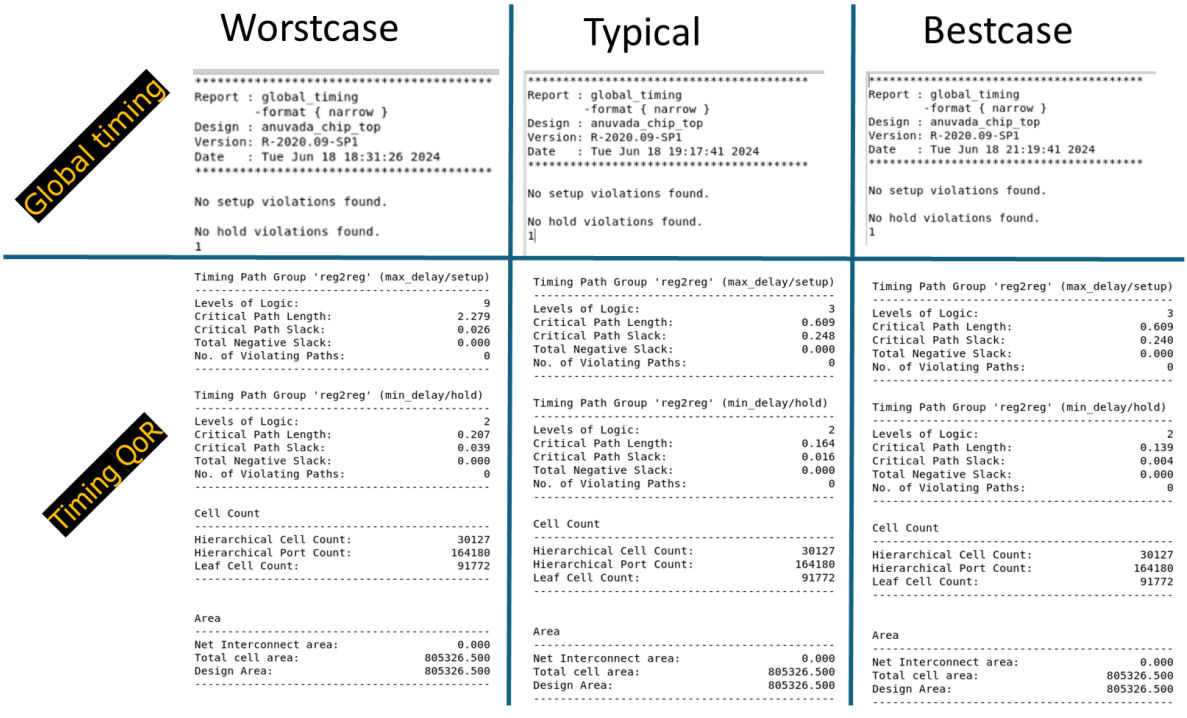


Figure 5.21 STA reports for wrostcase (0.72V, 1.68V, 125°C), typical (0.8V, 1.8V, 25°C) and bestcase (0.88V, 1.98V, -40°C) showing no violating paths in the design.

Table 5.6 Comparison of IP design area in Anuvāda Chiplet

IP name	Total Bandwidth	Final Designed size (mm^2)	Improvement
PCIe 3.1 IP	8 Gbps	0.694	
SuperCHIPS channel IP	8 Gbps	0.007	99x

We observe that the size of SuperCHIPS IP is 99 times smaller than the size of PCIe 3.1 IP for the same bandwidth.

Si-IF were also fabricated to integrate the Anuvāda dielet for full assembly and testing. The fabrication of the Si-IF uses Cu dual-damascene process for wiring layers and Cu single-damascene process for pillar layer. The fabrication process is elaborately explained by research done by Jangam et al [SJan21]. The key difference in this Si-IF is the reduction of the pillar height from 5μm to 2μm to reduce wafer-scale warpage. The Cu pillars are then recessed by 1μm to expose them for bonding. In Fig. 5.22 (d), an Si-IF design is shown with 0.6 mm SuperCHIPS channel inter-dielet communication. A cross-section micrograph of the Si-IF which shows both metal layers and pillar is presented in Fig. 5.22(e).

5.8 Anuvāda testing process and integration on Si-IF

Anuvāda chip testing was divided into four parts. The goal was to test the chip a step-by-step manner so that if issues are encountered, it would be easier to understand and resolve them. The three main tests done are as follows:

1. Die-level slow speed data loader testing
2. Full Si-IF implementation and high-speed loopback testing
3. Full Si-IF implementation and ESD transmission line pulse measurements.

The intended tests for loopback operation are visually shown in Fig. 5.26. In the section below, we elaborate on each test and also discuss test results.

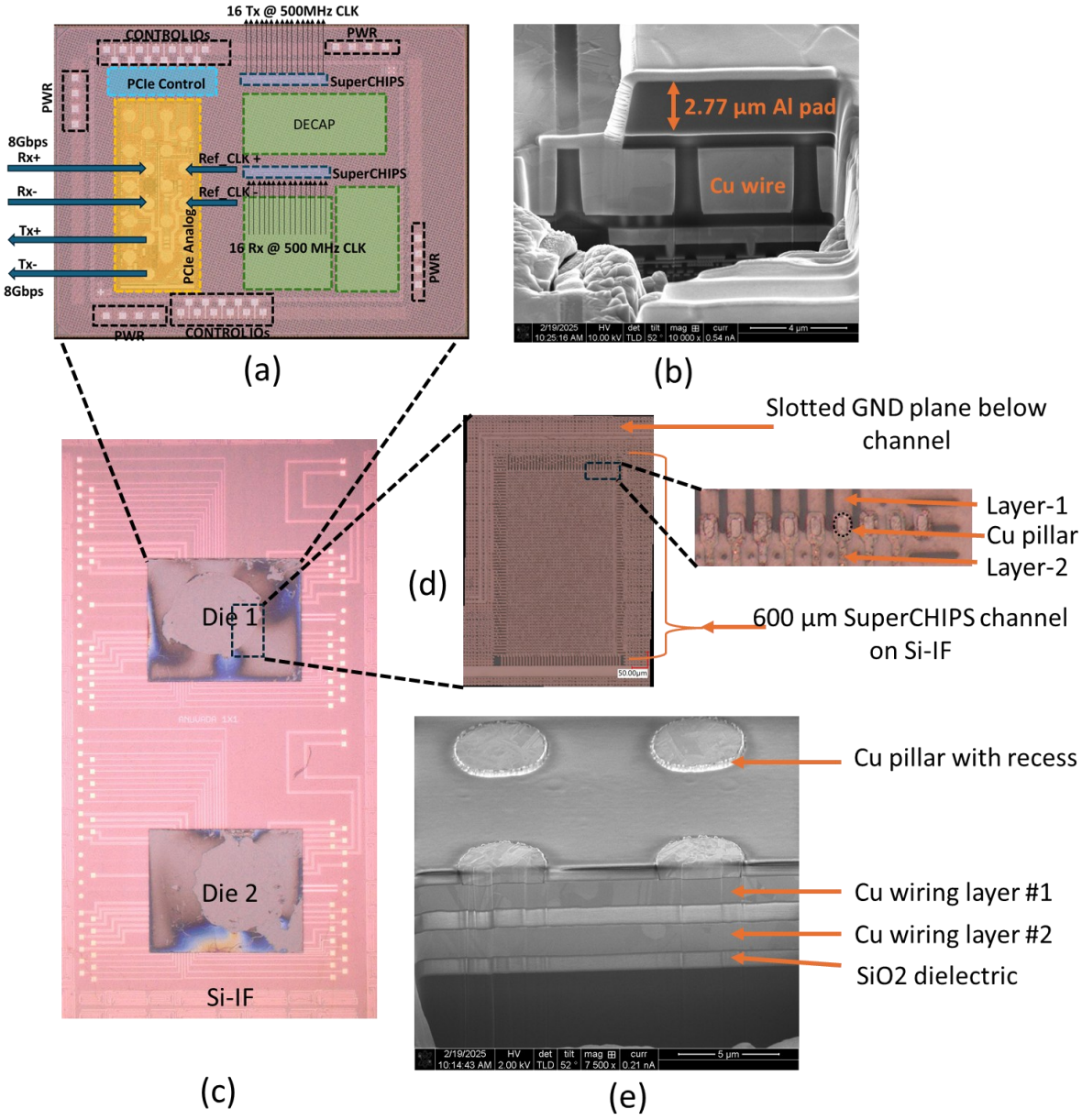


Figure 5.22 (a) micrograph of Anuvāda dielet. SuperCHIPS design region is presented in blue whereas, PCIe in yellow and light blue. (b) cross-section of the 22nm FDX dielet showing Aluminum termination. (c) Si-IF with two Anuvāda dielets integrated. (d) closeup of two-layer Si-IF showing 600 µm SuperCHIPS channel including an inset image of Cu pillar and two wiring layers on Si-IF. (e) a SEM micrograph of two-layer Si-IF used in Anuvāda.

Next the methodology of integration of Anuvāda on Si-IF is discussed. Anuvāda dielets are terminated with Aluminum pads instead of Copper pads. This is due to the PCIe 3.1 hard macro-IP being already terminated in Al pads. This is shown in Fig. 5.21 (b). At UCLA we have always proposed bonding to be between Au-Cu or Cu-Cu or Au-Au which would require the dielet pads to be copper terminated. However, in this case, Cu termination is not possible. While the option of wafer-pull at last Cu metal layer is possible, The final Al pad layer on PCIe 3.1 IP does more than just laying out pads. Within the IP, passive structures were implemented on the Al pad layer including several wiring connections to the previous Cu metal layer. Changing the Al layer to Cu would not only tamper with the IP but also, we do not have a way to ascertain how it will affect the performance of the IP since internal information about the schematic, devices connectivity and layout is protected by IP rules. Therefore, we decided not to tamper with the IP, rather come up with a novel integration scheme for bonding the Anuvāda dielet to Cu pillars on the Si-IF.

However, thermal compression bonding involving Aluminum is very difficult [KSah24], or not reliable, which we found through initial tests. This is due to the rapid oxidation rate of Al to Al_2O_3, requiring only a few milliseconds. The Al_2O_3 then acts as a barrier for thermal compression bonding to Cu pillars on Si-IF and it is not possible to bond the dielet.

The process of making Anuvāda bonding compatible with Cu pillars on Si-IF involves introduction of a metal overlayer. In the case of Anuvāda, we lifted off 20nm/200nm or titanium/gold (Ti/Au) on the PCIe pads. As per our process on record, all Anuvāda dielets were obtained with 100nm Silicon Nitride passivation. This passivation is removed using dry etching and then pads openings are patterned on Anuvāda dielets over the Al pads. Next, using evaporation, we deposit 20nm/200nm Ti/Au on the Al pads. The purpose of Ti layer is to act as an adhesion layer for sticking Au to Al pads.

The bonding of Au capped pads to Cu pillars is done using in the Au-Cu TCB process as exactly described in Chapter 2, section 2.12. The Fig. 5.23 shows pads on Anuvāda dielet with Ti/Au capping. The capping is smaller than the pad dimensions to ensure no overflow on the pad in case of misalignment during lithography process. This is very important as any misalignment can lead to shorts. After bonding Anuvāda dielets using Au-Cu bonding process, a sample of two Anuvāda dielets on Si-IF is shown in Fig. 5.21 (c). Post bonding

both dielets were tested to ensure they powered up properly. In the next section we will discuss all the testing results in detail.

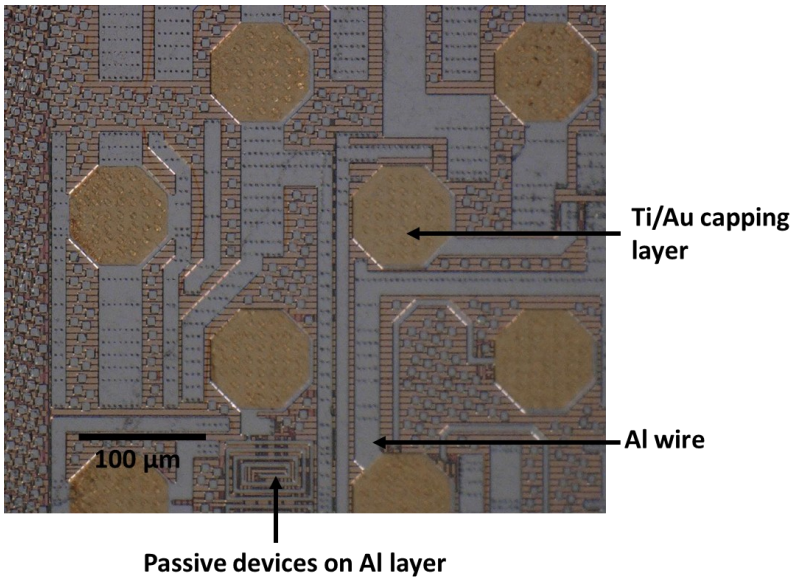


Figure 5.23 Ti/Au capping layer on Al pads to enable Au-Cu TCB of Anuvāda dielets to Cu pillars on Si-IF. Notice that the Ti/Au dimensions are smaller than the corresponding Al pad dimensions to account for any misalignment during lithography.

5.9 Anuvāda testing results

Anuvada die was mounted on a break-out PCB for testing the slow-speed data-loader, which runs at a clock frequency of 1 MHz. The purpose of data-loader is to load the PCIe 3.1 Configuration while PCIe is set to reset mode (phy_reset = 1), so when reset is deasserted, PCIe inputs are already stable.

5.9.1 Die-level show speed data loader testing

As discussed in section 5.8 of this chapter, First, we set VDDIO → 1.8V, VSSIO→ 0V, followed by VDD → 0.8V and VSS→ 0V.

The leakage of the chip is found to be within specification. The “phystatus” which is the output indicating status of the PCIe is driven up to 1.8 V or HIGH, which is the expected initial state of this signal. This shows the PCIe starts up to the expected initial state.

Next, PCIe configuration signals need to be supplied through the TCK, TMS, and TDI inputs. These signals are applied through the Xilinx Artix 7 FPGA [ARTIX7]. The coding necessary to supply these signals was independently done. The Test setup for slow speed testing is shown in Fig. 5.24.

The output of TDO, exactly matches the expected simulated output, as shown in Fig 5.25. This is the first important test result which shows the PCIe configuration data is loaded to the input of PCIe block. Throughout this process, PCIe is in reset mode by keeping phy_reset asserted to 1. Therefore, we demonstrate the process of configuring the PCIe chip through the data-loader module.

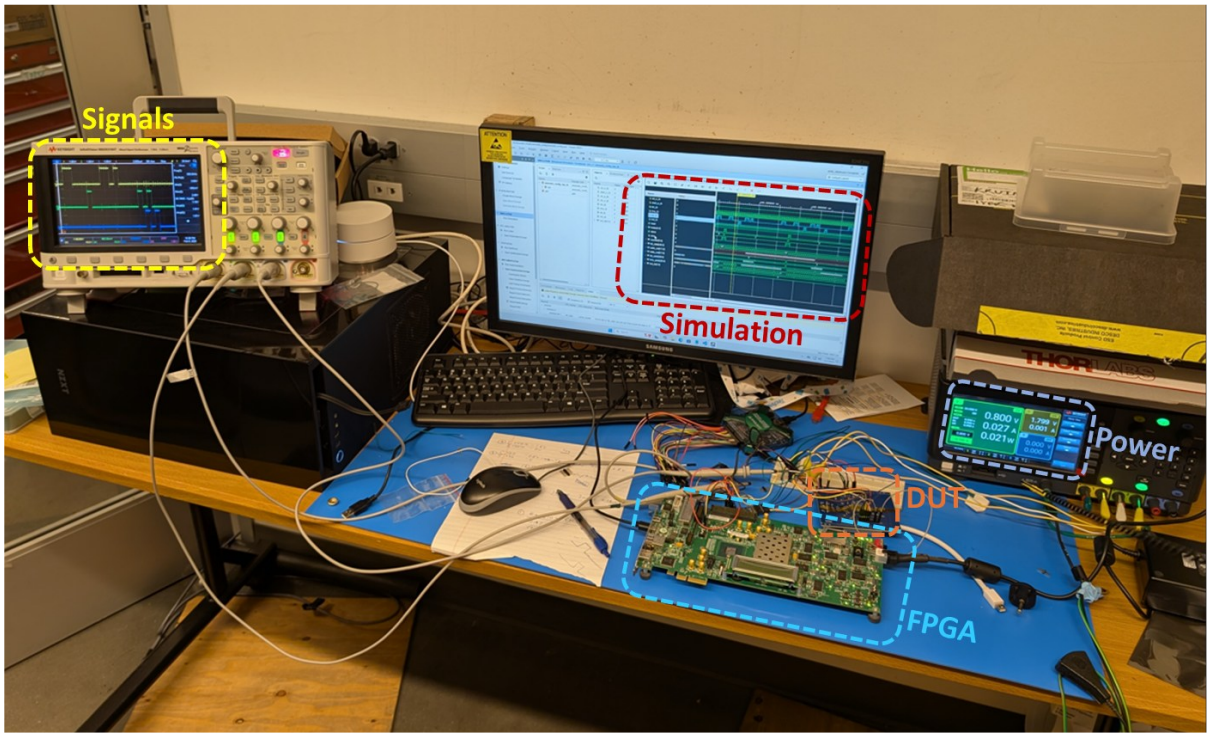


Figure 5.24 Measurement setup for slow speed testing.

Figure 5.25 Comparison of STA waveforms with the actual waveforms obtained from the Anuvāda chip showing a perfect match. These inputs (TDI/TMS) and outputs (TDO, obtained from chip) are used to configure the PCIe block.

5.9.3 Full Si-IF implementation and loopback testing

For complete loopback test, the integrated bit error rate (iBERT) testing tool on a FPGA from Alinx AXKU042 board [AXKU] was used. The iBERT provides an integrated interface to apply various PRBS waveforms in PCIe format and do a loopback to demonstrate PCIe loopback functionality. The transmit and receive signals are handled through the standard PCIe interface.

To verify translation, the goal is to send a PRBS sequence generated by FPGA with necessary PCIe specific header to the first Anuvāda chip, which will be translated into SuperCHIPS signals to be sent on the SuperCHIPS hardware interface on the Si-IF as seen in Fig. 5.21 (d). The loopback signals are then received by a second or even the same Anuvāda chip for translation back into high speed signals to be sent back to FPGA. The FPGA then compares the signals and accounting for time delays, it can calculate the bit error rate in comparing the received PRBS sequence to the originally transmitted PRBS sequence. This is schematically shown in Fig. 1.25. If the bit error rate (BER) is below the BER of PCIe 3.1 block, we consider translation to be successful. Our goal is to

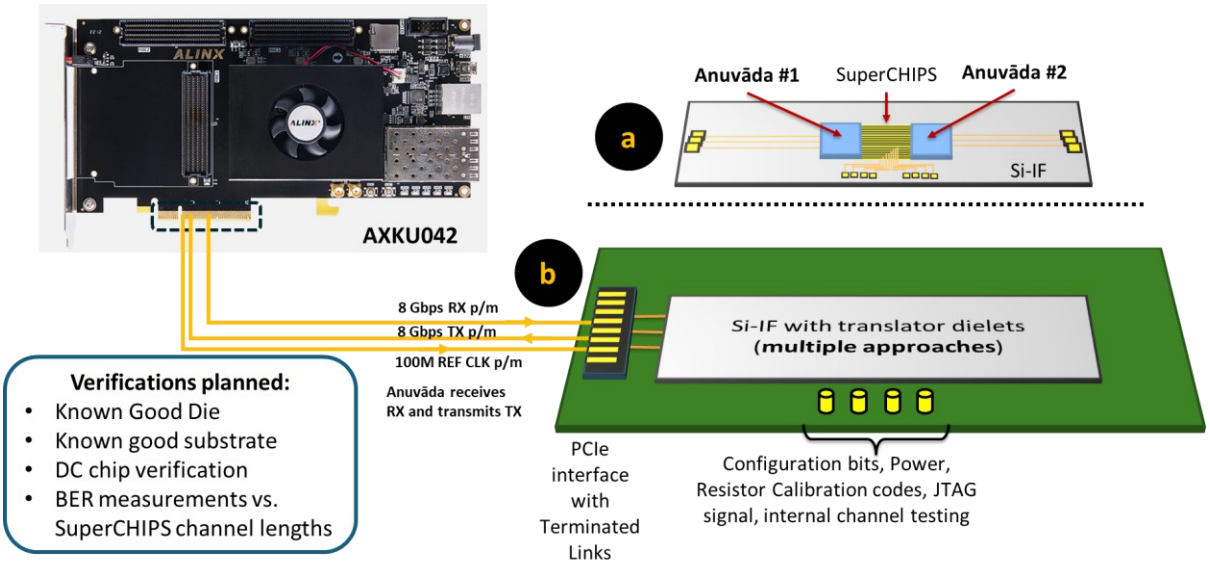
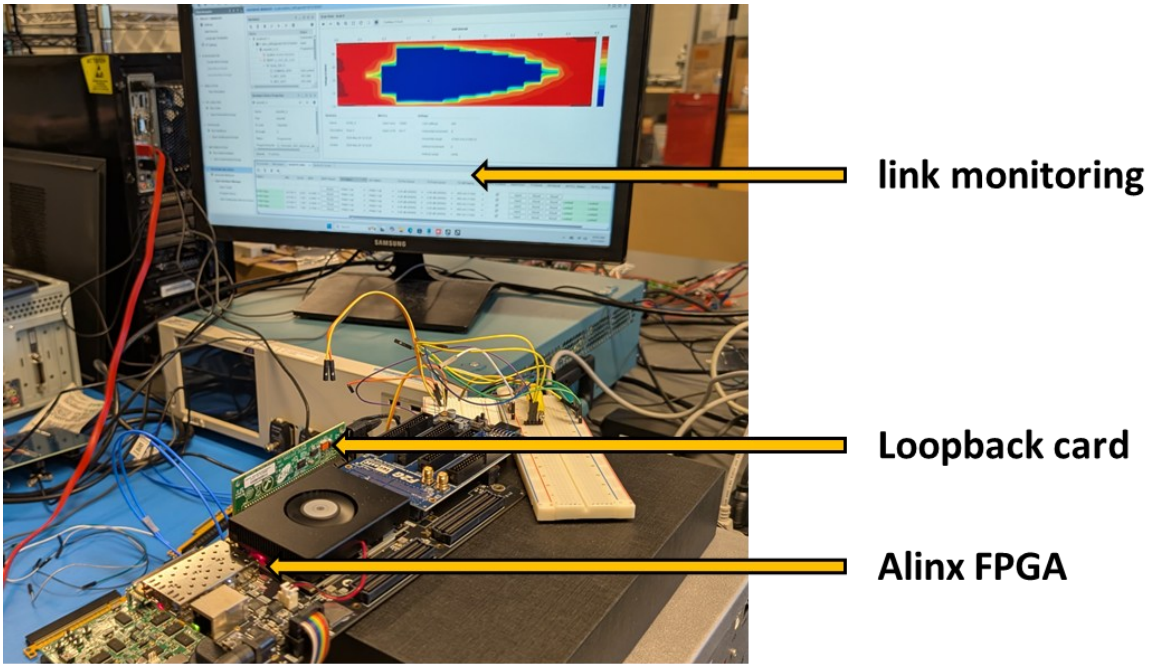


Figure 5.26 PCIe loopback test for translation through ALINX AXKU042 FPGA. (a) shows integration of two Anuvāda chiplets with SuperCHIPS PHY between them. (b) shows integration on a PCB.

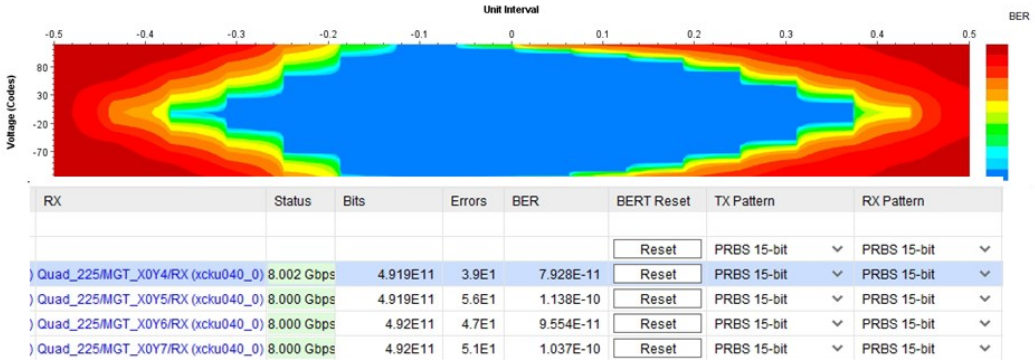
perform the loopback test first with single Anuvāda chiplet and then with two Anuvāda chiplets. In the Fig. 5.25 (a) first we bond the Anuvāda dielets to the corresponding Si-IF using Au-Cu TCB and in Fig. 5.25 (b) we wirebond the sample to a breakout PCB board with terminated links. The wirebond connections were simulated to ensure they were compatible with PCIe speeds.

First, we need to determine the BER of the PCIe transceivers on the FPGA using a standard PCIe loopback card. A standard loopback card from whizz systems [CLOOP] is used to verify the 8 Gbps PCIe transceiver operation on the Ailinx board. Fig. 5.26 (a) shows the setup. As can be seen from the figure, the loopback card is inserted into the PCIe x8 slot. The loopback card has a reference clock generator which is used to generate the reference 100 MHz clock signal for generation of PCIe data.

Fig. 5.26 (b) shows the output for the loopback card. The loopback card has 4 links compared to only 1 link on current version of Anuvāda dielet. All the links on the loopback card run at 8 Gbps data rate and the link achieves a bit error rate (BER) of 8E-11. A standard acceptable BER for PCIe link is ≤ 10^{-12} [S1] which is sufficient for reliable high speed data transfer across the PCIe fabric.



(a)



RX	Status	Bits	Errors	BER	BERT Reset	TX Pattern	RX Pattern
					Reset	PRBS 15-bit	PRBS 15-bit
) Quad_225/MGT_X0Y4/RX (xcku040_0)	8.002 Gbps	4.919E11	3.9E1	7.928E-11	Reset	PRBS 15-bit	PRBS 15-bit
) Quad_225/MGT_X0Y5/RX (xcku040_0)	8.000 Gbps	4.919E11	5.6E1	1.138E-10	Reset	PRBS 15-bit	PRBS 15-bit
) Quad_225/MGT_X0Y6/RX (xcku040_0)	8.000 Gbps	4.92E11	4.7E1	9.554E-11	Reset	PRBS 15-bit	PRBS 15-bit
) Quad_225/MGT_X0Y7/RX (xcku040_0)	8.000 Gbps	4.92E11	5.1E1	1.037E-10	Reset	PRBS 15-bit	PRBS 15-bit

(b)

Figure 5.27 (a) Setup for loopback using loopback card from whizz systems []. The loopback card is inserted into PCIe x8 connector. (b) the measured link rate (8 Gbps) with PRBS 15-bit pattern and eye opening at receiver.

Now we have verified the initial condition using the default loopback card. Next, we replace the loopback card with the PCB on which Anuvāda is integrated. For the Anuvāda loopback test, we use only a single Anuvāda dielet assembled on an Si-IF. The links on Si-IF have been designed to facilitate loopback using only a single chip if necessary. The

Si-IF is wirebonded to the PCB which has high speed links designed for 8 Gbps signals. The new setup is shown in Fig. 5.27(a).

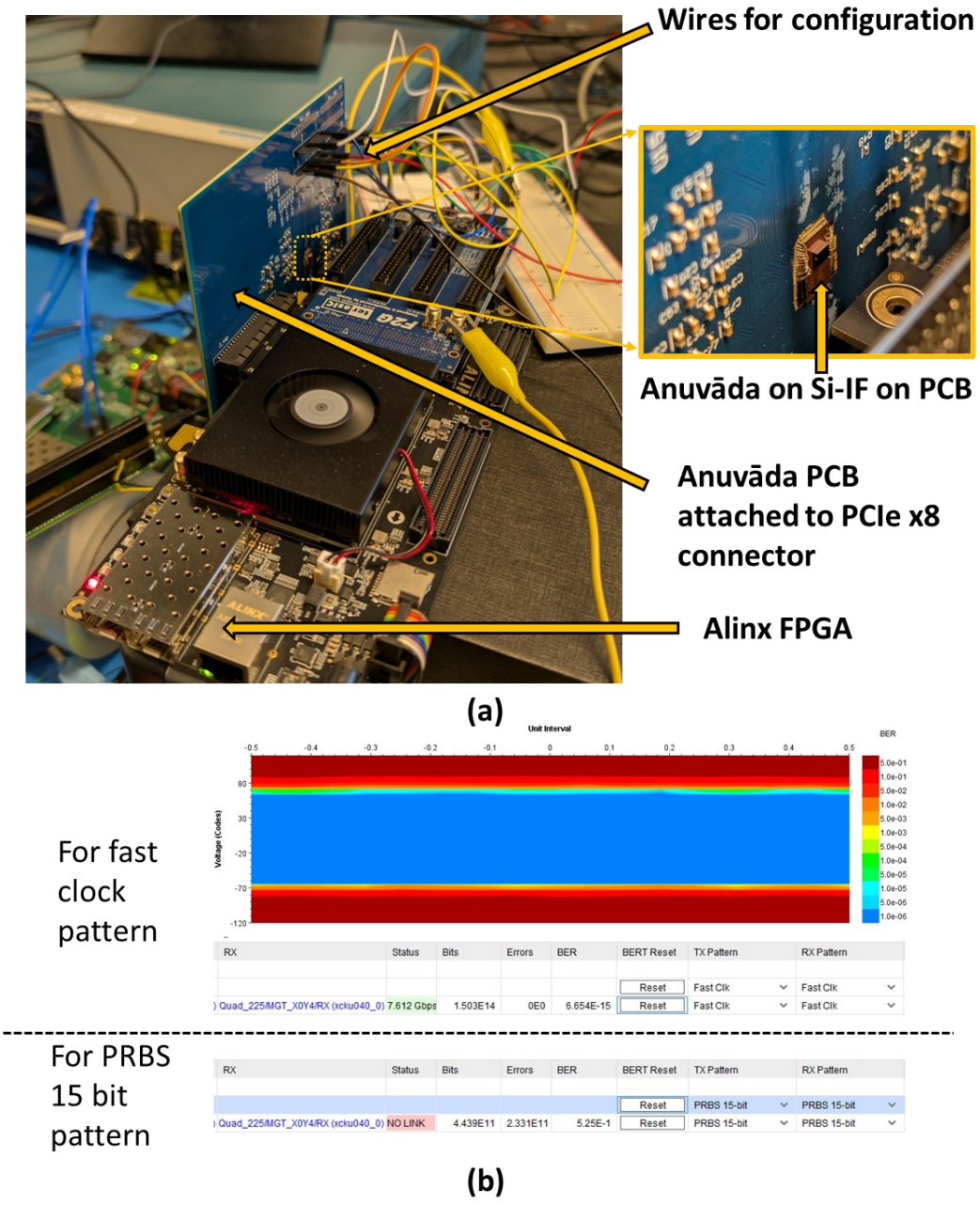


Figure 5.28 The loopback setup using bonded Anuvāda dielet on Si-IF further assembled on a breakout PCB. The PCB is inserted into PCIe x8 connector. (b) the measured link rate (7.612 Gbps) with a clock pattern at transmitter and receiver as well as challenges with PRBS 15-bit signal.

The reference clock for the test is derived from the standard loopback card to maintain the same clocking conditions. First the Anuvāda dielet is powered up and then configured by exactly following the sequence discussed in Section 5.6. First the known good die signals were verified to work on the sample after assembly, indicating proper bonding on Si-IF. Fig. 5.27 (b) shows the current loopback results. As we can see, we were able to lock the PLL of the FPGA for transmitting PCIe PRBS packets. The figure shows when the transmitted signal from the FPGA is a clock pattern or alternating 1's and 0's (100 % toggle rate), the same pattern is obtained at the receiver of FPGA. This means a clock like pattern was received by Anuvāda, translated into SuperCHIPS, and looped back. Once looped back it is translated back to high-speed clock pattern and sent to the FPGA receiver, which is able to detect the signal.

The data rate of transmission is 7.612 Gbps which is lower than 8 Gbps. We expect the parasitics of the link play a role in reduction of data-rate.

However, as we change the signal transmitted by FPGA to PRBS 15-bit, the link fails to work. We expect there may be challenges with alignment of PCIe packet header between the transmitted PRBS signals and received PRBS signals. A new version of PCB will be designed to address this issue. Since the link only worked for clock signal there are no errors in transmission and BER ≤ 10^{-15}. However, this is not the true BER. To measure the true BER, we need to send and receive the identical PRBS signals.

5.9.4 Full Si-IF implementation and ESD transmission line pulse measurements.

Through the Anuvāda design and assembly process, we have asked for wavier on ESD rules applied on SuperCHIPS IOs. Due to small size (~85 μm^2) of total ESD protection area instead of expected ~500 μm^2 protection desired by GlobalFoundries process, it is important for us to identify that no fabrication related damage such as plasma related charging, antenna effects etc. can affect or damage the IO circuits.

Furthermore, while we do not expect bonding to be a cause of ESD due to our generous use of ionizers during bonding, we do extra processing of the die and transfer it to several stations for passivation removal, plasma activation and testing after bonding. ESD diodes will help provide paths for excess charges to flow to ground in event of charge

accumulation while the dies are being transferred. Hence it is important to quantify the protective performance of ESD diodes for SuperCHIPS IOs.

In section 5.4.2 we discussed the transmission line pulse (TLP) test which enables study of behavior of the IO circuits under transient events, such as ESD. The important aspect of TLP is it is a non-destructive test where the capability of IO to protect against ESD events is quantified without damaging the IO as in case of HBM (human body model [ESDAg]) or CDM (Charged Device Model [ESDAg]) tests. TLP involves applying controlled voltage stress with very fast rise time and measuring the current through the IOs after specific time interval. TLP generates controlled pulses and measures the current vs. voltage (I-V) behavior of the device under test (DUT). This gives quantitative insights into how the protection structures behave under stress.

The schematic of a standard TLP tester is shown in Fig. 5.29, taken from [ETLP]. A basic pulse generator consists of a charge line (TL1) of length L, a switch (S1), and a high voltage power supply (V_o). The charge line is first charged to the desired voltage, and then as the S1 changes position, it is discharged through another Transmission line (TL2) as a pulse. The other end of TL2 is connected to the input/output (IO) pad which will

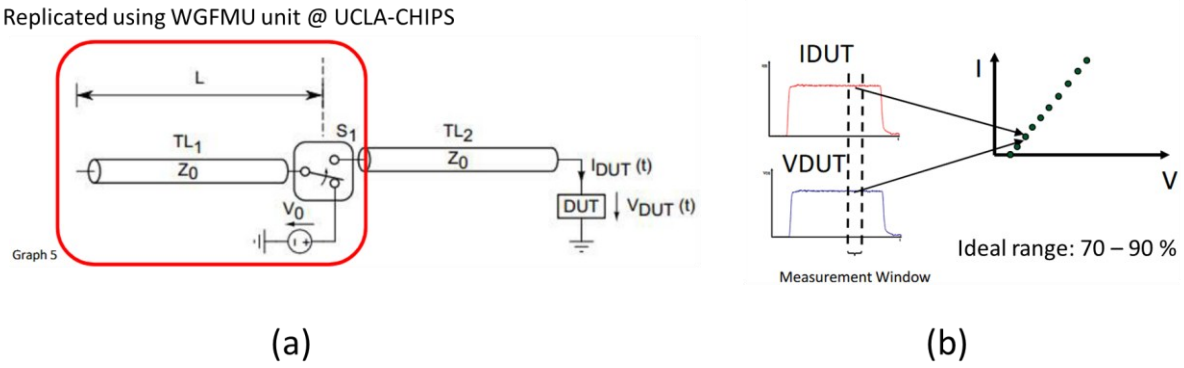


Figure 5.29 (a) Schematic of ESD TLP setup [ETLP] (b) ESD TLP IV measurement approach.

experience the pulse. The voltage (VDUT) and current (IDUT) across the device under test (DUT) are captured by oscilloscope or similar measurement device and a ESD-TLP IV plot is drawn as shown in Fig. 5.29 (b). Some aspects of TLP measurement include:

- The length of the charge line determines the pulse width
- Typically pulse width can be 10ns to ~1μs with a sharp rise time about 10ns. We have used WGFMU and Keysight Analyzer to provide this pulse. Due to equipment capability the sharpest rise time obtainable is about 80 ns.
- Pulses are incrementally increased until the device fails, or maximum voltage on tester is reached.
- Measurement window is typically between 70 – 90 % region to obtain a point on the I-V curve. In our WGFMU equipment, we have measured the current at about 80% of applied voltage stress interval.

Next let us discuss the experimental details of the TLP measurement done in this work. We have used the WGFMU (Waveform Generation Fast Measurement Unit) in conjunction with Keysight Analyzer to provide a sequence of pulses from 0 V to 5 V. Our testing is limited by the rise time limitations of the WGFMU, which is 80ns. The details of the TLP pulse are shown in Table 5.7. Fig. 5.30 shows the sequence of pulses applied to the DUT.

Table 5.7 Details of the ESD TLP pulse generation

	Given (WGFMU)
Pulse rise time	80 ns
Pulse width	200 ns
Measurement duration	80 % of pulse width
Voltage	Pulse train
Voltage range	0 V to 5 V

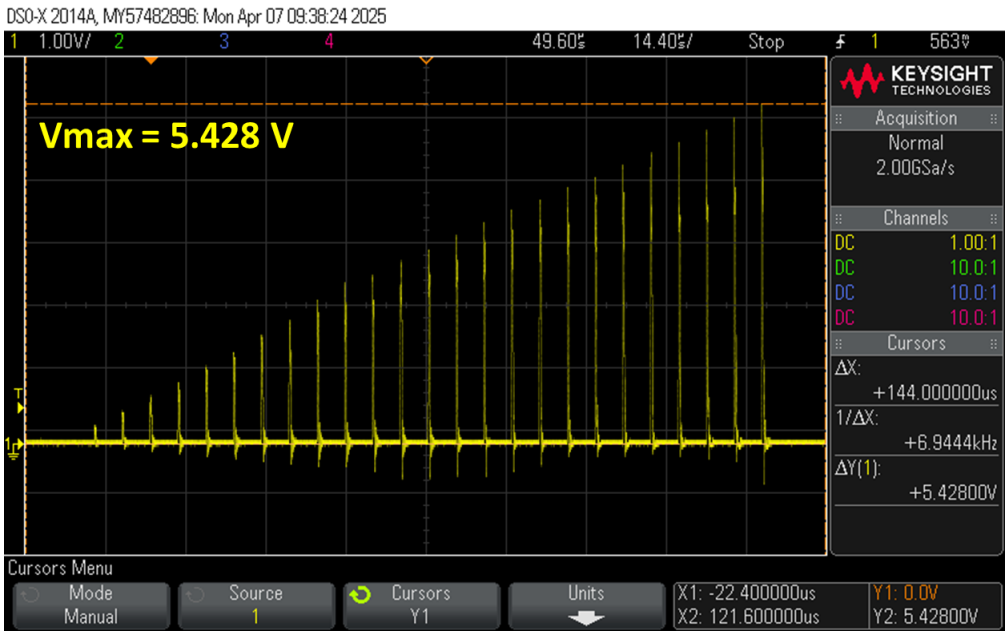


Figure 5.30 Train of pulses from 0V - 5V applied to each IO device under test. notice a gap of ~1µs is provided between each step for device cooldown.

The TLP test involves measuring the I-V characteristics of two types of IO devices (DUTs) used in Anuvāda under ESD stress:

- SuperCHIPS IO: SuperCHIPS IO have 10µm pitch work in the voltage domain of 0.8V. These IOs have ESD cells custom designed in this work as described in previous sections of this chapter.
- Synopsys IO: Synopsys IO consist of standard 22nm GlobalFoundries input/output devices obtained from Synopsys which are designed in accordance with GlobalFoundries ESD design rules.

We look at the current-voltage characteristics of the 10 different IO devices per die. Each IO is measured 4 times to ensure voltage stress is applied to each IO four times. Overall, 40 TLP IV stress curves are measured in each die.

Test1: measurement of TLP IV characteristics of IOs before and after bonding.

In the first test we measure the ESD TLP IV measurements of both SuperCHIPS IOs and Synopsys IOs before and after bonding the same die. In this test, no actual bonding is performed, but the dielet is made to go through all the steps in tacking such as dielet passivation removal, acetic acid cleaning, dielet pick and transfer to bondhead and tacking under temperature and pressure. The same IOs are measured before and after this bonding cycle and results are presented in Fig. 5.31.

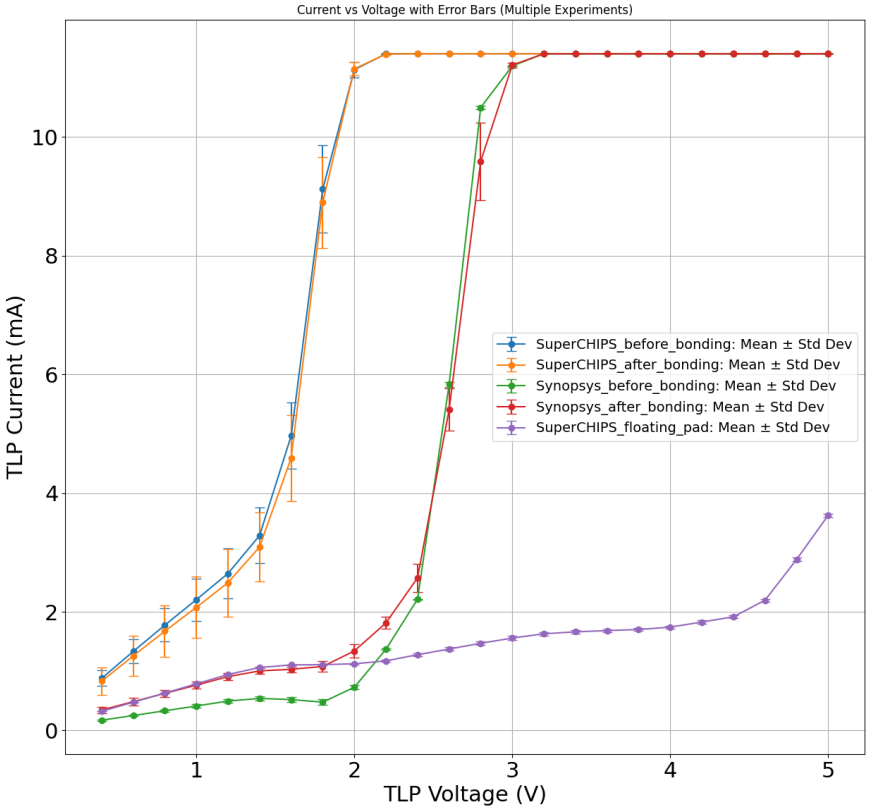


Figure 5.31 TLP IV measurements before and after bonding.

A key point to note is that the WGFMU unit is not designed to measure currents greater than 11mA to protect the internal circuitry. However, from Fig. 5.30 above, we ensure that we are applying voltages up to 5.4 V across the IO device and ground. We expect actual current in the IO to be higher than 11mA, though it cannot be measured. In spite of the limitation, we can report each IO must at least pass 11 mA of current during the test, which

his substantially higher than its current capability. Clearly the ESD diodes spring into action during the applied high voltage fast rise time stress.

From the Fig 5.31 we can observe the following:

- The TLP-IV curve of the device remains identical for both Synopsys IO and SuperCHIPS IO before and after bonding. This shows bonding does not affect the ESD performance of these devices. In fact, the IOs are well protected even after bonding cycle.
- Each IO was subjected to the 0V – 5V pulse train repeatedly for 4 times. The results are therefore repeatable.
- The Synopsys IO and SuperCHIPS IO have different clamping voltages due to operating in different voltage domains. Synopsys IOs operate in 1.8 V voltage domain and SuperCHIPS IOs operate in 0.8 V voltage domain. The average clamping voltage of Synopsys IO is 2.5 V and that of SuperCHIPS IO is 1.6 V.
- A final IV plot is done on a floating pad with no connection to a ESD protection device. As seen clearly from the chart, this floating pad does not show ESD diode IV behavior.

Test 2: Comparing ESD TLP IV measurements in two different bonded dielets to determine variability with different dies.

In the second test, we take two bonded dielets and measure the ESD TLP IV characteristics of both Synopsys and SuperCHIPS IOs. The IV measurements are shown in Fig. 5.32.

We observe the following:

- The IV performance of ESD diodes under stress is identical across different dielets.
- The dynamic ON resistance of the ESD devices for SuperCHIPS IOs is 55.56 Ω while that of Synopsys IOs is 56.49 Ω. This shows both diodes offer similar resistance during an ESD event. We expect this resistance to be related to the

parasitic resistance of the Anuvāda dielet. In short, the resistance of both the ESD diodes when activated is smaller than the parasitic resistance of the dielet.

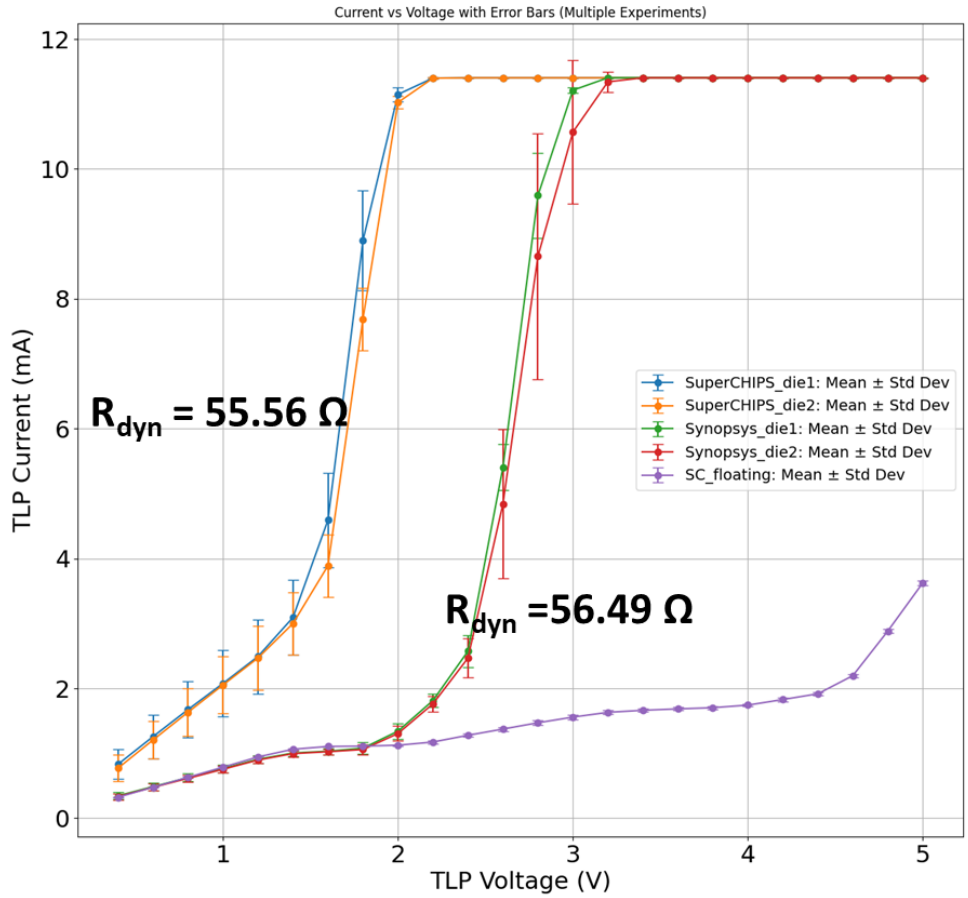


Figure 5.32 TLP IV measurements across different dielets.

This concludes our ESD TLP measurements. In general, we find SuperCHIPS IOs to perform equivalently to Synopsys IOs and therefore, expect similar performance from both the diodes in case of an ESD event.

Chapter conclusion

In this chapter, we have explored the concept of translation and its importance for a chiplet based ecosystem. We expect a library of protocol translator dielets to be available on demand on the chiplet marketplace to help with design as required. Achieving protocol

interoperability is necessary to achieve compatibility with commercial protocols such as PCIe, ethernet as well as emerging protocols like UCIe etc. SuperCHIPS is an ideal PHY to translate between different protocols due to its simple overhead in terms of power, area and high bandwidth > 2Tbps/mm. In this chapter we have discussed a proof of concept translator chiplet called Anuvāda – which can translate between PCIe 3.1 and SuperCHIPS and vice versa. In addition, dedicated ESD protection devices were added to SuperCHIPS IOs for the first time and TLP IV measurements of these devices were performed showing they have performance very similar to foundry recommended Synopsys IOs. The detailed design architecture of Anuvāda, simulations of translation post layout are presented in this chapter. Also, the measurement setup and procedure for performing loopback and verifying translation is discussed. Finally actual loopback measurement show transfer of an alternate 1's and 0's pattern through the high-speed link to Anuvāda and back to the signal source i.e. the FPGA. We expect Anuvāda to be the building block on which the translator chiplet ecosystem is established under the umbrella of chiplet ecosystem.

Chapter 5 References

PCI Express® 5.0 Compliance Testing

[Spal19] 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)*, Feb. 2019, pp. 250–263. doi: 10.1109/HPCA.2019.00042.

[SJan20+] S. Jangam, U. Rathore, S. Nagi, D. Markovic, and S. S. Iyer, "Demonstration of a Low Latency (<20 ps) Fine-pitch (≤10 µm) Assembly on the Silicon Interconnect Fabric," in *2020 IEEE 70th Electronic Components and Technology Conference (ECTC)*, Jun. 2020, pp. 1801–1805. doi: 10.1109/ECTC32862.2020.00281.

[KSah22] K. Sahoo, U. Rathore, S. Chandra Jangam, T. Nguyen, D. Markovic, and S. S. Iyer, "Functional Demonstration of < 0.4-pJ/bit, 9.8 µm Fine-Pitch Dielet-to-Dielet Links for Advanced Packaging using Silicon Interconnect Fabric," in *2022 IEEE 72nd Electronic*

Components and Technology Conference (ECTC), May 2022, pp. 2104–2110. doi: 10.1109/ECTC51906.2022.00332.

[URat22] U. Rathore, S. S. Nagi, S. Iyer, and D. Marković, “A 16nm 785GMACs/J 784-Core Digital Signal Processor Array With a Multilayer Switch Box Interconnect, Assembled as a 2×2 Dielet with 10µm-Pitch Inter-Dielet I/O for Runtime Multi-Program Reconfiguration,” in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, Feb. 2022, pp. 52–54. doi: 10.1109/ISSCC42614.2022.9731582.

[SNagi23] S. S. Nagi, U. Rathore, K. Sahoo, T. Ling, S. S. Iyer, and D. Marković, “A 16-nm 784-Core Digital Signal Processor Array, Assembled as a 2 × 2 Dielet With 10-µm Pitch Interdielet I/O for Runtime Multiprogram Reconfiguration,” *IEEE Journal of Solid-State Circuits*, vol. 58, no. 1, pp. 111–123, Jan. 2023, doi: 10.1109/JSSC.2022.3212685.

[PCIeWiki] PCI Express, [online], Available: https://en.wikipedia.org/wiki/PCI_Express

[ETHE] “What Is Ethernet?,” Cisco. Accessed: Jun. 01, 2025. [Online]. Available: https://www.cisco.com/c/en/us/solutions/enterprise-networks/what-is-ethernet.html

[FSQSFP] FS, “Understanding SFP and QSFP Ports on Switches,” FS.com. Accessed: Jun. 01, 2025. [Online]. Available: https://www.fs.com/blog/understanding-sfp-and-qsfp-ports-on-switches-530.html

[CXLWiki] “Compute Express Link,” *Wikipedia*. May 23, 2025. Accessed: Jun. 01, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Compute_Express_Link&oldid=1291743959

[UCIeSYN] “What is UCIe? – How it Works | Synopsys.” Accessed: Jun. 01, 2025. [Online]. Available: https://www.synopsys.com/glossary/what-is-ucie.html

[KSah24] K. Sahoo *et al.*, “Aluminum to copper thermal compression bonding for heterogeneous integration of legacy dielets,” in *2024 IEEE 74th Electronic Components and Technology Conference (ECTC)*, 2024, pp. 1940–1946. doi: 10.1109/ECTC51529.2024.00329.

[GF22FDX] “FDX™ FD-SOI,” GlobalFoundries. Accessed: Jun. 01, 2025. [Online]. Available: https://gf.com/technology-platforms/fdx-fd-soi/

[PCISYN] “PCI Express IP Solutions | Synopsys.” Accessed: Jun. 01, 2025. [Online]. Available: https://www.synopsys.com/designware-ip/interface-ip/pci-express.html

[SJanD20] SivaChandra Jangam’s thesis, [online] available: https://s3.us-west-1.amazonaws.com/chips.user.media/page_file/Jangam_dissertation_v6+Printed.pdf

[ISuth99] I. Sutherland, R. F. Sproull, and D. Harris, *Logical Effort: Designing Fast CMOS Circuits*. Morgan Kaufmann, 1999.

[DHyu25] D. Hyun, Y. Jung, I. Cho, and Y. Shin, “Decoupling Capacitor Insertion Minimizing IR-Drop Violations and Routing DRVs,” in *2023 28th Asia and South Pacific Design Automation Conference (ASP-DAC)*, Jan. 2023, pp. 271–276. Accessed: Jun. 01, 2025. [Online]. Available: https://ieeexplore.ieee.org/document/10044831

[DecapWiki] “Decoupling capacitor,” *Wikipedia*. Jul. 15, 2024. Accessed: Jun. 01, 2025. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Decoupling_capacitor&oldid=1234659025

[DecapAlt] “Decoupling Capacitor and Bypass Placement Guidelines,” Altium. Accessed: Jun. 01, 2025. [Online]. Available: https://resources.altium.com/p/bypass-and-decoupling-capacitor-placement-guidelines

[ESDFu] “EOS/ESD Fundamentals Part 1 | EOS/ESD Association, Inc.” Accessed: Jun. 01, 2025. [Online]. Available: https://www.esda.org/esd-overview/esd-fundamentals/part-1-an-introduction-to-esd/

[ANagu25] A. Nagus, “How to set up an ESD Workstation | ESD Table Bench Setup | Correct Products,” Correct Products - Leaders in ESD Packaging & Static Control. Accessed: Jun. 01, 2025. [Online]. Available: https://blog.correctproducts.com/2015/12/11/how-to-set-up-an-esd-workstation/

[JTAGT] “The JTAG Test Access Port (TAP) State Machine - Technical Articles.” Accessed: Jun. 01, 2025. [Online]. Available: https://www.allaboutcircuits.com/technical-articles/jtag-test-access-port-tap-state-machine/

[SJanD20] SivaChandra Jangam’s thesis, [online] available: https://s3.us-west-1.amazonaws.com/chips.user.media/page_file/Jangam_dissertation_v6+Printed.pdf

[U1] UCIe V1.1 specification

[U2] X. F. Brun, S. Wong, M. Mota, and E. Udofia, “World’s First UCIe Interoperability Silicon Enabling Open Standards Heterogeneous Integration,” in *2024 IEEE 74th Electronic Components and Technology Conference (ECTC)*, May 2024, pp. 264–268. doi: 10.1109/ECTC51529.2024.00050.

[U3] UCIe Chiplet Summit Presentation, [Online] Available: https://chipletsummit.com/proceeding_files/a0q5f0000044zma/20240206_PreConD_Pasdast.PDF#page=8.00

[B1] R. Farjadrad, M. Kuemerle, and B. Vinnakota, “A Bunch-of-Wires (BoW) Interface for Interchiplet Communication,” *IEEE Micro*, vol. 40, no. 1, pp. 15–24, Jan. 2020, doi: 10.1109/MM.2019.2950352.

[B2] S. Ardalan *et al.*, “Bunch of Wires: An Open Die-to-Die Interface,” in *2020 IEEE Symposium on High-Performance Interconnects (HOTI)*, Aug. 2020, pp. 9–16. doi: 10.1109/HOTI51249.2020.00017.

[S1] PCIe 5.0 specification, PCI-SIG, 2025, www.pcisig.com

[S2] A. Shokrollahi *et al.*, “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)*, Jan. 2016, pp. 182–183. doi: 10.1109/ISSCC.2016.7417967.

[JTAGS] JTAG test standard, [online], Available: https://grouper.ieee.org/groups/1149/1/

[SJan21] S. Jangam and S. S. Iyer, “Silicon-Interconnect Fabric for Fine-Pitch (≤10 µm) Heterogeneous Integration,” *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 11, no. 5, pp. 727–738, 2021, doi: 10.1109/TCPMT.2021.3075219.

[KSah24] K. Sahoo *et al.*, “Aluminum to copper thermal compression bonding for heterogeneous integration of legacy dielets,” in *2024 IEEE 74th Electronic Components and Technology Conference (ECTC)*, 2024, pp. 1940–1946. doi: 10.1109/ECTC51529.2024.00329.

[ARTIX7] Xilinx Artix7 FPGA, product brief, [online], Available: https://www.xilinx.com/support/documents/product-briefs/artix7-product-brief.pdf

[AXKU] Alinx Kintex Ultrascale board [online], Available: https://www.en.alinx.com/Product/FPGA-Development-Boards/Kintex-UltraScale/AXKU042.html

[CLOOP] Standard loopback card for PCIe 3.1, Available: https://bmisurplus.com/product/whizz-systems-3900151-01-pci-loopback-card/?srsltid=AfmBOooX23sTMdEc7VWyRZE6EdjWQ1SPoSZ5yZwXnciFAOTPcYnHw-Qp

[ESDAg] S. Agarwal, “What’s The Difference Between HBM, CDM, And MM Test?,” Electronic Design. Accessed: Jun. 02, 2025. [Online]. Available: https://www.electronicdesign.com/technologies/power/article/21799383/whats-the-difference-between-hbm-cdm-and-mm-test

[ETLP] “Introduction of Transmission Line Pulse (TLP) Testing for ESD Analysis”, [online] Available: https://www.esdemc.com/public/docs/TechnicalSlides/ESDEMC_TS001.pdf#page=1.00

Chapter 6 Conclusion and outlook

6.1 Concluding remarks

Wafer-scale integration is a transformative "scale-down and scale-out" paradigm aimed at meeting the rapidly growing demands for memory capacity and interconnect bandwidth in modern computing systems. While this approach offers unprecedented performance benefits, it also introduces a range of formidable implementation challenges. This thesis addresses several of these challenges, focusing on scalable assembly methods, assembly yield optimization, monolithic chip partitioning strategies, application based cost modeling, and protocol interoperability for heterogeneous wafer-scale systems.

Chapter 2 presents a high-throughput, fine-pitch (≤10 μm) two-step Cu-Cu thermal compression bonding (TCB) process for 300 mm silicon wafers. This process is essential for enabling manufacturable wafer-scale assembly. The first step involves dielet tacking onto a passive silicon interconnect fabric (Si-IF) substrate, followed by a second wafer-level annealing step to establish mechanically and electrically reliable bonds. The proposed TCB process achieves a single bond-head throughput of ~330 units-per-hour (UPH) and demonstrates bond integrity that meets MIL-SPEC mechanical standards and JEDEC electrical reliability criteria. Strategies for further improving throughput, potentially up to 1000 UPH are detailed, including design of experiments (DoE) to shorten tacking time. In practice, parallel bond-heads can already meet the required throughput for full-wafer population, highlighting the scalability of this bonding methodology.

Chapter 3 explores additional factors that influence assembly yield beyond bonding throughput. These include design-related, tool-related, and process-related parameters such as bond-pad distribution, bond-head parallelism, alignment drift during progressive dielet bonding to wafer, and dielet dicing quality. Fabrication yield of the Si-IF substrate is also characterized. These findings are intended to guide dielet designers and system integrators in optimizing assembly processes. Emphasis is placed on the importance of comprehensive documentation and pre-bond testing (known good

die/substrate), as well as a well-structured post-bond testing protocol. Effective yield and risk management strategies are discussed for ensuring high-reliability wafer-scale assemblies.

Chapter 4 introduces key figures of merit (FOMs) relevant to wafer-scale architectures. A *partitioning figure of merit* (p-FOM) is defined to guide the disaggregation of monolithic chips into smaller chiplets. This metric accounts for the area and power overhead introduced by additional die-to-die terminals upon disaggregation, as well as the yield benefits of using smaller dies. The trade-off between area, power, and yield is analyzed, and the study concludes that dielet of ~100 mm² or smaller area provide an optimal balance for most use cases. A cost function (CF) is also introduced, based on a signaling FOM proposed in [SJansFOM], to evaluate different packaging strategies. This framework assists system designers in selecting the most suitable packaging technology by comparing bandwidth, latency, transceiver area, drivable length, and energy per bit trade-offs across architectures.

Chapter 5 addresses *protocol interoperability*, a critical challenge in wafer-scale systems. While intra-wafer communication can leverage custom, high-bandwidth, low-power interconnects, communication with external devices (e.g., storage, memory, networks) must rely on standardized protocols such as PCIe, Ethernet, UCIe, and CXL. To bridge this gap, this work introduces Anuvāda, a protocol translation chiplet designed to interface between on-wafer custom signaling and commercial I/O standards. Anuvāda leverages the SuperCHIPS interface—*Simple Universal Parallel intERface on CHIPS*—which supports fine-pitch (≤10 µm) signaling and was first proposed in [SJan17]. In this work, SuperCHIPS is used in a synchronous mode with a novel ESD protection scheme for robust integration. Chapter 5 details the architectural design, simulation, and silicon implementation of Anuvāda on the Si-IF platform, along with testing results.

In summary, this dissertation presents a comprehensive exploration of the design, assembly, and system integration challenges involved in realizing wafer-scale systems. The innovations and methodologies introduced here significantly advance the feasibility of practical, heterogeneous wafer-scale computing platforms. Looking ahead, wafer-scale integration is poised to become a foundational enabler for next-generation AI and high-

performance computing (HPC) systems—offering unmatched improvements in compute density, memory bandwidth, interconnect performance, and system scalability.

6.2 Outlook

While this research addresses several critical implementation challenges in wafer-scale integration, it does not resolve all the issues necessary for realizing a fully practical heterogeneous wafer-scale system. Below, we outline key future challenges that must be addressed to advance the development of such systems:

- **Low-temperature Cu-Cu bonding for temperature-sensitive devices:** This work presents a scalable two-step Cu-Cu thermal compression bonding (TCB) process for 300 mm wafers, which involves annealing at 300 °C to 400 °C. While such temperatures are typically compatible with CMOS fabrication, they may pose risks to temperature-sensitive components like high-bandwidth memory (HBM) dielets. Future research should explore alternative bonding techniques or process modifications that enable low-temperature annealing (<250 °C), particularly for HBM integration.
- **Advancement in tooling for high-yield assembly:** Assembly yield is significantly influenced by tooling-related parameters such as bond-head parallelism, alignment accuracy, and dielet dicing quality. Improvements in bonding and dicing tool precision, automation, and feedback control will be critical for achieving near-perfect assembly yields across full wafers.
- **Building a family of protocol translator chiplets:** The Anuvāda chiplet introduced in this work serves as a protocol translator for interfacing internal wafer-scale links with standard external interfaces such as PCIe and UCIe. However, Anuvāda represents just one member of a broader family of translator chiplets. Future development must extend this concept to support a variety of commercial protocols, forming a modular ecosystem of translator chiplets within the broader chiplet marketplace.
- **Power delivery at wafer scale:** Power delivery for wafer-scale systems remains an open and formidable challenge. Systems with thousands of dielets may require

power in excess of 50 kW, necessitating architectures that can provide low-noise supply rails, high transient response, and adequate decoupling. While this thesis does not address power delivery, Ren et al. [Hren23] have proposed a segmented architecture that includes 48 V-to-POL (point-of-load) conversion and high-frequency (>50 MHz) switching to support wafer-scale power demands. Interested readers are encouraged to refer to their work for further insight.

- **Thermal Management and Heat Extraction:** Another key challenge not addressed in this thesis is thermal management of wafer-scale system both during normal operation and transient computing spikes. Wafer-scale systems are expected to dissipate heat densities ranging from 1 W/mm² to 6 W/mm². Removing this heat effectively, particularly from within the shadow of the wafer (i.e., without bulky heat sinks or heat spreaders), is crucial for compatibility with rack-mounted deployments. Two-phase liquid cooling has emerged as a promising approach. Shah et al. [Usah25] have proposed a novel two-phase cooling scheme using a water-methanol mixture capable of extracting over 1 W/mm². Ongoing work aims to scale this capability up to 4 W/mm² [NMSu25].

These highlighted challenges represent few of the important directions for future research in the wafer-scale domain. While this work makes foundational contributions in assembly, yield optimization, and signaling interoperability, we anticipate that the insights and methodologies developed here will enable the next generation of engineers and researchers to further advance wafer-scale computing. Ultimately, these efforts will help meet the ever-growing demands for compute density, memory bandwidth, and energy-efficient interconnects in future AI and high-performance computing systems.

Chapter 6 References

[SJansFOM] S. Jangam and S. S. Iyer, “A Signaling Figure of Merit (s-FoM) for Advanced Packaging,” *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 10, no. 10, pp. 1758–1761, Oct. 2020, doi: 10.1109/TCPMT.2020.3022760.

[SJan17] SivaChandra Jangam, S. Pal, A. Bajwa, S. Parmarti, P. Gupta and S. S. Iyer, "Latency, Bandwidth and Power Benefits of the SuperCHIPS Integration Scheme", Proc. of 67th IEEE Electronic Components and Packaging Technology (ECTC) 2017, Orlando, FL, pp. 86-94. doi: 10.1109/ECTC.2017.246

[Hren23] H. Ren *et al.*, “Heterogeneous Power Delivery for Large Chiplet-based Systems using Integrated GaN/Si-Interconnect Fabric with sub-10 μm Bond Pitch,” in *2023 International Electron Devices Meeting (IEDM)*, Dec. 2023, pp. 1–4. doi: 10.1109/IEDM45741.2023.10413759.

[Usah25] U. D. Shah, “Transient Analysis of Flash Cooling for Dynamic Thermal Management for High-Performance Electronic Systems,” Ph.D., University of California, Los Angeles, United States -- California, 2023. Accessed: Apr. 24, 2025. [Online]. Available: https://www.proquest.com/docview/2835785749/abstract/1B23C4D66C8E4079PQ/1

[NMSu25] N. Meenakshi Sundaram, R. Pugazhendhi, S. Iyer, and T. Fisher, “Flash boiling of methanol/water mixtures in a microchannel cooler,” in 2025 24th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), May 2025. (accepted)