DENSE INTERCONNECTION AND ADVANCED COOLING TECHNOLOGIES FOR 2.5D AND 3D HETEROGENEOUS INTEGRATED CIRCUITS

A Dissertation
Presented to
The Academic Faculty

By

Sreejith Kochupurackal Rajan

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology

May 2022

DENSE INTERCONNECTION AND ADVANCED COOLING TECHNOLOGIES FOR 2.5D AND 3D HETEROGENEOUS INTEGRATED CIRCUITS

Thesis committee:

Dr. Muhannad S. Bakir, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Albert B. Frazier
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Gary S May, Co-advisor
Office of the Chancellor
University of California, Davis

Dr. Tushar Krishna
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. John Cressler
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Yogendra Joshi
The George W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology

Date approved: April 18, 2022

The noblest pleasure is the joy of understanding

Leonardo da Vinci

To my parents

My work is just a testament to your love and support

ACKNOWLEDGMENTS

The work presented in this thesis is a direct result of support and contributions from a group of amazing people who supported me on various phases of this incredible roller-coaster ride. First and foremost, I wish to thank Dr. Muhannad Bakir for all the valuable guidance and his steadfast support during the course of my Ph.D. work. Dr. Bakir's vision, and the freedom he provided to explore different ideas and concepts have been instrumental in overcoming the many challenges I faced in this work. His infectious optimism, and his confidence in people working with him are traits that I would long cherish. I would forever be grateful for this incredible opportunity.

I also want to express my sincere gratitude to my co-advisor Dr. Gary May for his continued support throughout my Ph.D. He has always found time to provide timely guidance and support despite his very demanding schedule. His feedback on technical issues as well as on how to approach research problems has greatly influenced my work. It has been an absolute privilege to work with Dr. May.

I also want to thank Dr. John D. Cressler, Dr. Albert B. Frazier, Dr. Yogendra Joshi and Dr. Tushar Krishna for their willingness to serve on my Ph.D dissertation committee.

I am thankful for the opportunity to work with both previous and current members of I3DS group. In particular, I want to thank Dr. Hanju Oh for his mentoring during my first few months and his continued support even after his graduation from the group. I also want to acknowledge the incredible contributions from Dr. Joe Gonzalez on his insights on how to approach research, and properly teaching me the art and science of microfabrication. I want to thank Dr. Tom Sarvey for the help with the chapters on microfluidic cooling. I am especially in debt to Ankit Kaul for his assistance on a wide variety of the projects and being a supportive friend despite our at times heated disagreements on work. I could not have succeeded without the tremendous contributions from these people. I am also thankful for the opportunity to work with the previous group members including Dr. Xuchen Zhang,

Dr. Yang Zhang, Dr. William Wahby, Dr. Muneeb Zia, Dr. Reza Abbaspour, Dr. Congshan Wan, Dr. Paul Jo, Dr. Md Obaidul Hossen and Carl Li. I am grateful for the current members of the group including Ting Zheng, Shangtao Yu, Jiaao Lu, Jonathan Brescia, Madison Manley, Micheal Nieves Calderon, Ashita Victor and Philip Anschlutz for the many interesting discussions and creating a pleasant work environment. I also want to thank my collaborators from Microsoft Dr. Hussam Alissa, Dr. Bharath Ramakrishnan and Washington Kim for all the support and guidance.

I would also like to sincerely acknowledge the immense contributions from IEN staff and leadership, without which none of my experimental work would have been possible. Their continued support and availability even outside work hours have been instrumental in facilitating all the experiments. In particular, I want to thank Dr. Durga Rao Gajula, Dr. Chris Yang, Dr. Hang Chen, Dr. Mikkel Thomas, Charlie Turgeon, Tran-Vinh Nguyen, Charlie Suh, Thomas Averette, Devin Brown, Alex Gallmon, and Andrew Watkins for their tremendous assistance, and Dr. Oliver Brand and Gary Spinner for their leadership.

Finally, I want to thank my friends and family for their support that kept me going though this often challenging journey. I am especially indebted to my parents Rajan Nair and Indira R. Nair for their unconditional support and providing me with the freedom to pursue my passions without any questions. I am also thankful for the love and support from my brother Renjith, sister-in-law Malini, my grandmother Ponnamma and my amazing nephew Vaibhav.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

3DID Three-Dimensional Interconnect Density

AI Artificial Intelligence

ALD Atomic Layer Deposition

AMI Acetone, Methanol, Isopropanol

BEOL Back End of Line

BGA Ball Grid Array

BST Breakable Support Technology

BTS Board Test System

C4 Compressible Collapsible Chip Connect

CMOS Complementary Metal Oxide Semiconductor

CoWoS Chip-on-Wafer-on-Substrate

CPUs Central Processing Units

CTE Coefficient of Thermal Expansion

D2D Die-to-Die

D2W Die-to-Wafer

DBI Direct Bonded Interconnect

DI De-Ionized

DRAM Dynamic Random Access Memory

DRIE Deep Reactive Ion Etching

DSP Digital Signal Processing

EMIB Embedded Multi-die Interconnect Bridge

EPB Energy per Bit

EPS Electronics Packaging Society

ES Engineering Silicon

EUV Extreme Ultra-violet

FFT Fast Fourier Transform

FIFO First-in First-Out

FPGA Field Programmable Gate Array

GaAs Gallium Arsenide

GaN Gallium Nirtide

GDDRx Graphics Double Data Rate x

GIMPS Great Internet Mersenne Prime Search

GPUs Graphic Processing Units

GS Ground-Signal

GSG Ground Signal Ground

HAR High Aspect Ratio

HBM High Bandwidth Memory

HFSS HIgh-frequency Structure Simulator

HI Heterogeneous Integration

HIR Heterogeneous Integration Roadmap

HPC High Performance Compute

IC Integrated Circuit

IHS Integrated Heat Spreader

IMCs Inter-metallic Compounds

IoT Internet of Things

IP Intellectual Property

IR Infra-red

IRDS International Roadmap for DEvices and Systems

IT Information Technology

JTAG Joint Test Action Group

KGD Known Good Die

LPCVD Low Pressure Chemical Vapor Deposition

MCM Multi-Chip Module

MEMS Micro-Electro-Mechanical System

MOS Metal Oxide Semiconductor

NUMA Non-uniform Memory Access

OS Operating System

PCB Printed Circuit Board

PCS Physical Coding Sublayer

PUE power Utilization Effectiveness

RF Radio Frequency

RLGC Resistance, Incuctance, Conductance, Capacitance

RPM Rotations Per Minute

S Scattering Parameters

SAM Self-Assembled Monolayer

SEM Scanning Electron Microscope

Si Silicon

SiP System-in-Package

SLA Stereolithography

SMD Surface Mount Device

SoC System-on-Chip

SoIC System-on-Integrated Chip

SRAM Static Random Access Memory

SST Soluble Support Technology

TDP Thermal Design Power

TIM Thermal Interface Material

TSDs Temperature Sensing Diodes

TSV Through Silicon Via

V-Cache Vertical Cache

VRM Voltage Regulator Module

W2W Wafer-to-Wafer

WSE Wafer Scale Engine

WUE Water Utilization Effectiveness

xcvr Transceiver

Y Admittance Parameters

Z Impedance Parameters

SUMMARY

While the need for compute scaling continues unabated on an upward trajectory, the traditional drivers of this increase such as node scaling are showing reduced returns in terms of both performance and cost. This is placing increasing emphasis on system-level scaling approaches including heterogeneous integration. This approach relies on creating compute systems by tightly integrating multiple chiplets each containing a fractional element of the whole system. This recent surge towards heterogeneous integration where advanced 2.5D and 3D ICs complement System-on-Chip (SoC) innovations to provide high-performance, low cost, and more customizable polylithic integrated circuits. Ensuring the scalability of these systems requires two major enabling technologies: (1) providing high-density, efficient die-to-die interconnects, and (2) managing the thermal challenges arising from the increased device stacking density enabled by heterogeneous ICs. In this work, we investigate enabling technologies that can help address these issues in heterogeneous integration.

First, we propose and demonstrate the use of selective metal electroless plating in conjunction with mechanical self-alignment technologies for die-to-die interconnects. This method facilitates a low-temperature, low-pressure, and high interconnect density inter-die bonding in heterogeneous 2.5D and 3D ICs. This method is a highly scalable alternative to conventional solder-based interconnects but comes without the stringent requirements such as including high-temperature tolerance, high-pressure process, extreme surface planarity and cleanliness, and very accurate initial alignment requirements of Cu-Cu direct bonding. A proof-of-concept testbed was fabricated with 50 μm pitch area array of interconnects. Finally, the electrical properties of the technology was characterized by finite element simulations.

Secondly, the compute and cooling efficiency benefits of silicon-integrated monolithic microfluidic cooling were investigated on a high-power functional CPU running real-world benchmarks. Next, the technology was scaled to 2.5D architectures and was evaluated on

an Intel FPGA with five discrete dice. Finite volume simulations and measurement data were used to quantify the benefits in terms of managing higher aggregate package power and minimizing the thermal coupling between closely spaced dice in 2.5D ICs. Finally, the interconnection and thermal management co-design challenges were evaluated for a 3D stack with inter-layer microfluidic cooling. Fabrication optimizations for very high aspect ratio TSVs were developed and the thermal-electrical trade-offs for these vertical interconnects were analyzed using measurements and 3D-EM simulators.

CHAPTER 1
INTRODUCTION

The steady increase of compute capabilities since the invention of Metal Oxide Semiconductor (MOS) based Integrated Circuit (IC) in 1960s have been enabled by the ability to increase the density of transistors in a given footprint. While the basic tenets of this are often described using Moore's law, this increase in compute density and decrease in cost have been enabled by few different phases in technology scaling. The initial advancements in IC technology enabled by pure geometric scaling that lasted till late 1990s. This has then been augmented by equivalent or effective scaling approaches where the transistor density increase was fuelled by device and design innovations such as the strained gate channels, Finfet architecture etc, and material innovations such as high-k dielectrics, Cu interconnects, low-k dielectrics etc [1].

However, we are approaching the limits of compute scaling solely fueled by these techniques. There is a need to augment these with system level innovations such as Heterogeneous Integration (HI) to keep up with the ever increasing demand for compute. Heterogeneous Integration leverages the ability to tightly interconnect different smaller 'chiplets' placed laterally in close proximity or stacked on top each other. This can enhance aggregate system performance by both increasing the amount of logic that can be integrated within the package footprint, as well as by secondary benefits such as the ability to bring large amounts of memory closer to compute. These scaling trends are captured in Figure 1.1.

In the subsequent sections of this chapter, the driving factors behind the need for HI and the typical implementation architectures are discussed in greater depth. Two major technological challenges to ensure scalability of HI architectures: (1) the scalability of chip-to-chip interconnects, and (2) the thermal management challenges associated with concatenating chiplets in close proximity are introduced and the potential solutions would

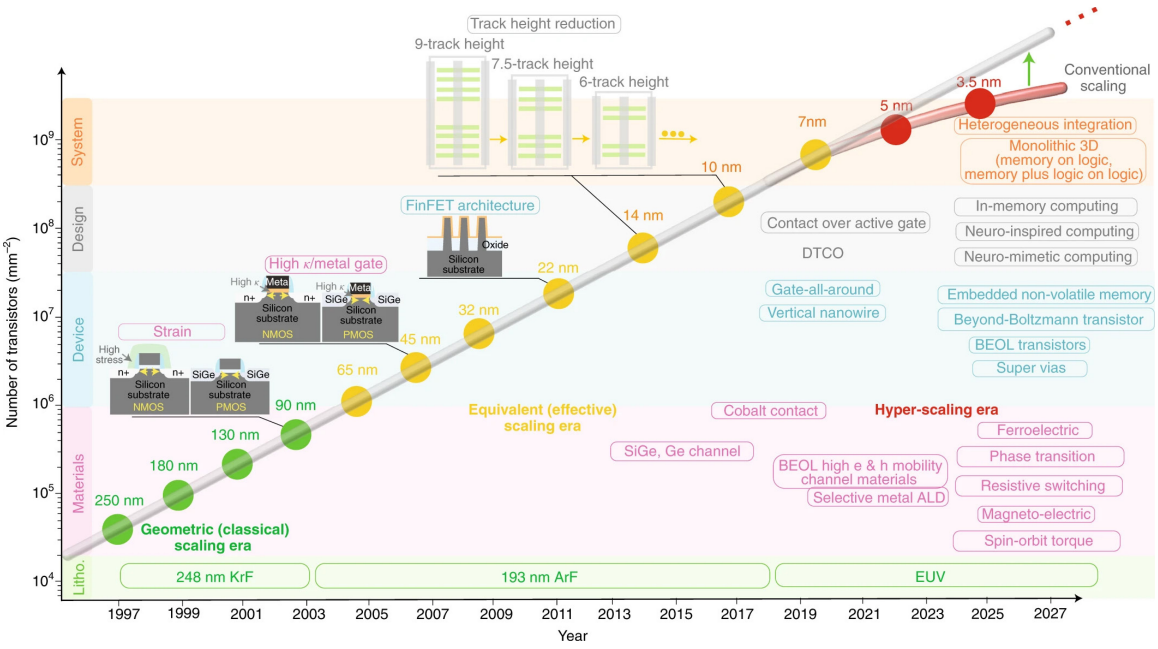


Figure 1.1: Trends in transistor density scaling [1].

be presented as the main focus of this thesis.

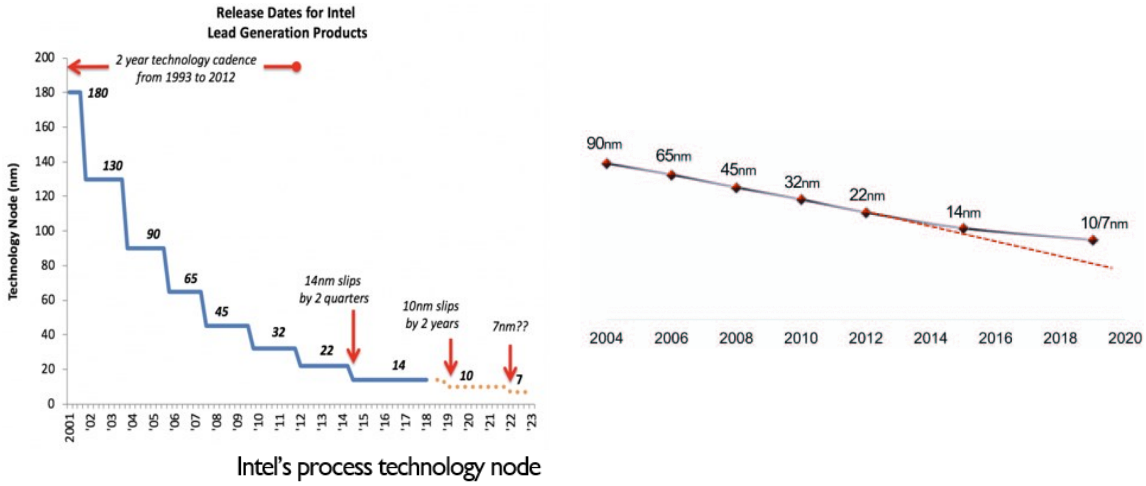
1.1 Need for Heterogeneous Integration

In one of the concluding sections of his paper *"Cramming More Components onto Integrated Circuits"* [2], Gordon Moore predicted that it would be more economical to create larger systems out of interconnected smaller functions. This economic advantage along with the ability to scale beyond what is possible with conventional single die systems are the main motivators for the use of HI as a new thrust vector for compute system scaling.

This need for multi-die systems for compute applications can be mainly attributed to three major trends: (1) the slowdown in Moore's law [3], (2) surging cost of development for advanced technology nodes [4] and (3) reticle size limitations for the dice.

The former is evident from Figure 1.2 which shows the delays in development of the leading technology node from both Intel and AMD. This increase in development time of newer technology nodes translates to the decay of both device density and cost scaling associated with Moore's law: i.e., the device density is not increasing as fast as it used to,

as well as per-transistor cost, even assuming fixed development costs, are not dropping at rates sustained over the last few decades.

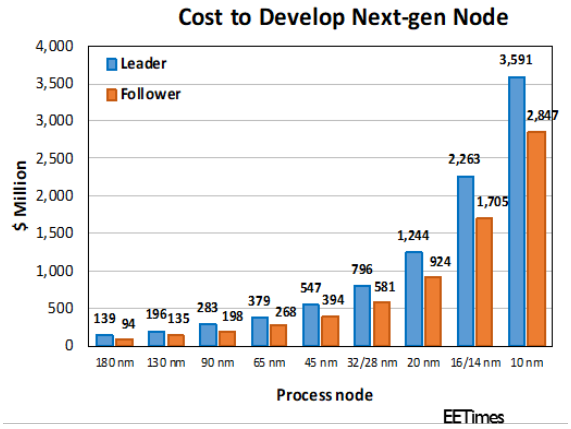


(a) (b)

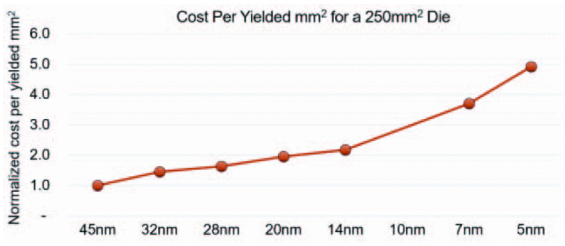
Figure 1.2: Slow down in Moore's law based scaling. Delay in introduction of leading node from (a) Intel [5] and (b) AMD [6].

This is compounded by the second issue of increase in cost of technology development as shown in Figure 1.3. Furthermore, the increased process complexity, the need for newer capital intensive equipment such as Extreme Ultra-violet (EUV) lithography, need for newer materials, and the need for more mask layers for multiple patterning compounds the cost increase associated with node scaling. This translates to an increase in cost per unit area of yielded die as captured in Figure 1.3b. This necessitates the need for new ways to enhance compute performance than solely relying on transistor scaling as it may be cost prohibitive.

Finally, another major advantage of using a chiplet based, multi-die design comes from the reticle size limitations that exists for single die systems. To circumvent the reduced rate of increase in device count due to slowdown of technology scaling, increasing the die size to add more transistors has been adopted in High Performance Compute (HPC) products such as server Central Processing Units (CPUs) and Graphic Processing Units (GPUs). The increasing prominence of this trend for high-end HPC products is captured in Fig-



(a)



(b)

Figure 1.3: (a) Surging development cost of technology nodes. (b) Normalized cost per chip versus technology node [6].

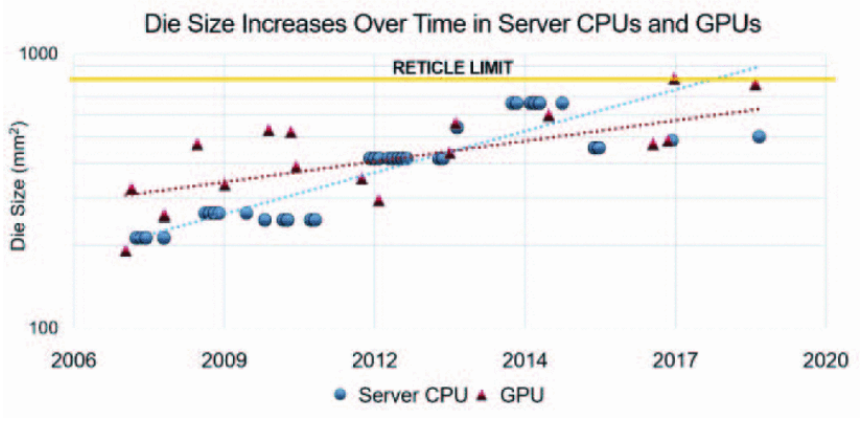


Figure 1.4: Die size trends for large SoCs over time. Adapted from [6].

ure 1.4. However, this trend is unsustainable due to two major drawbacks presented by the reticle size limitation in the manufacturing process. As shown in Figure 1.4, die sizes are already approaching the 858 mm^2 reticle size limit of current i193 and EUV lithography stepper systems. While it is possible to create larger dice with reticle stitching approaches demonstrated with implementations such as the Cerebras Wafer Scale Engine (WSE) [7] and TSMC Chip-on-Wafer-on-Substrate (CoWoS) interposers, these can prove to be extremely challenging with requirements such as customized design and production tools and approaches. Secondly, larger dice suffer from lower fabrication yields, thus increasing the overall cost [8]. An illustration of this effect of die size on overall yield is captured in

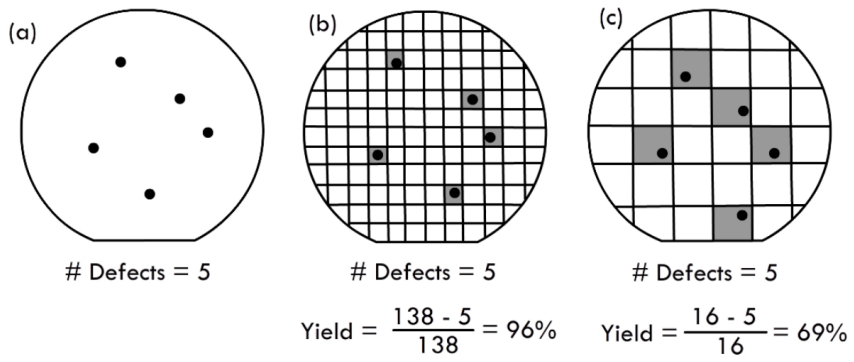


Figure 1.5: Five random defects on (a) wafer, (b) wafer with smaller die, and (c) wafer with larger die. Die with larger areas result in lower wafer yields [9].

Figure 1.5.

1.1.1 Advantages of Heterogeneous Integration

As described in the earlier part of this section, the conventional approach of relying on geometric scaling, combined with building all the requisite functionalities on a single large System-on-Chip (SoC) is becoming increasingly unsustainable. This has led to the increasing reliance on HI based system scaling approaches to ensure compute performance scaling. HI helps scale the overall system performance by (1) dis-aggregating a large SoC which is impractical due to aforementioned limitations into a System-in-Package (SiP) of tightly integrated, smaller chiplets, and (2) aggregating components which are distributed across a Printed Circuit Board (PCB) for a conventional system within a SiP, enabling tighter, more efficient integration of these blocks. These pathways for systems scaling is captured in Figure 1.6. Using multi-die heterogeneous integration also provides another vector for performance scaling by providing a more aggressive transistor count scaling with time when compared to its monolithic counterparts in the face of the slow-down of Moore's law, as shown in Figure 1.7.

Furthermore, beyond device count scaling, moving to a multi-die system by splitting the SoC into chiplets can help unlock multiple unique advantages as well. As explained

(a) (b)

Figure 1.6: Pathways for system scaling with heterogeneous integration. (a) SoC disaggregation to enable chiplet based systems beyond the limitations of single-die implementations [6]. (b) Component aggregation for a tighter integration of various components within a single package.

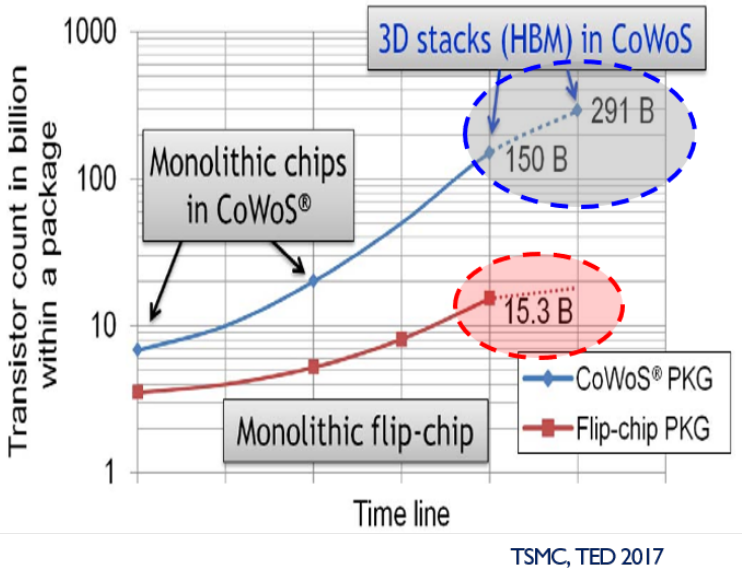


Figure 1.7: SiP Scaling benefits demonstrated for an interposer based package. The more aggressive transistor count realized by including HBM stacks within the package is shown. Adapted from [10].

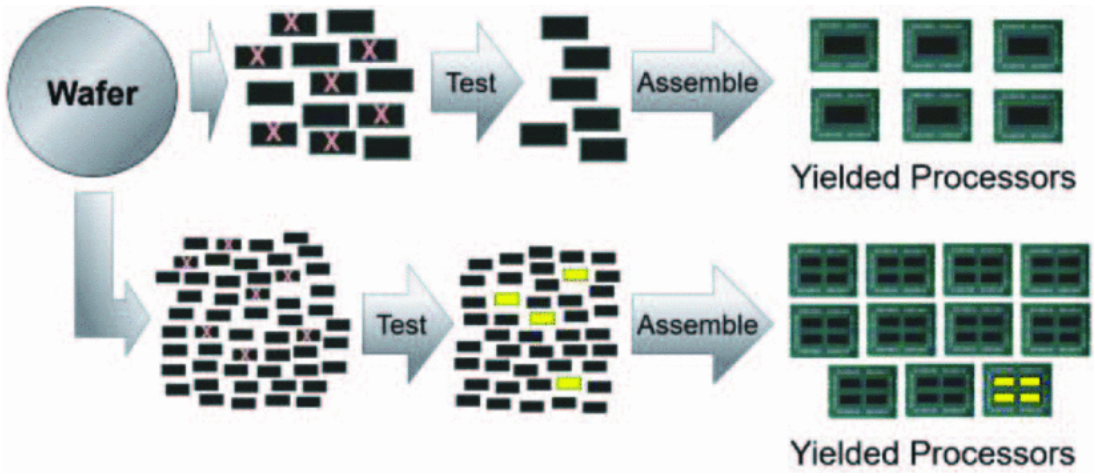


Figure 1.8: Illustrative construction of processors using (top) monolithic die and (bottom) reassembled chiplets. The light/yellow chiplets represent chiplets that can run at higher clock speeds. This shows the yield improvements as well as the potential pricing optimizations obtainable by KGD binning. Adapted from [6].

earlier, reducing the die size increases the overall yield, thereby bringing down the per-die cost. Even considering the silicon overhead needed to split the SoC, this can yield to considerable price savings. For example, AMD has reported an overall cost savings from yield improvement facilitated by splitting an SoC to a SiP to be as large as 41% [11]. Furthermore, the KGD testing before packaging can also be used for performance testing, which allows for better binning of individual chiplets that make up the final packaged product. This provides a higher level of granularity in controlling the performance, and correspondingly, pricing of the final packaged system, as shown in Figure 1.8.

Chiplet based architectures also facilitate Intellectual Property (IP) re-use, and mix-and-match of technology nodes which can lead to considerable cost savings along with added benefits such as lower time to market and an expansion of available product portfolios. As shown in Figure 1.9, various IP blocks have different times to yield technological maturity in advanced nodes. Using a chiplet based approach allows for optimal selection of technology nodes for different IP blocks thereby facilitating a faster and cost optimized pathway for creating the system.

This mix-and-match methodology has distinct benefits in applications such as high fre-

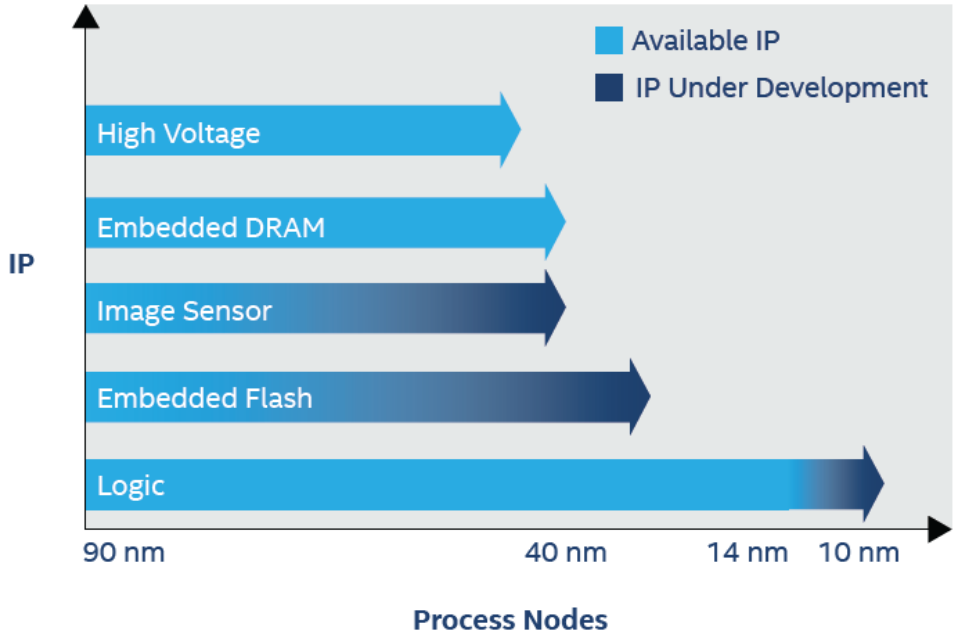
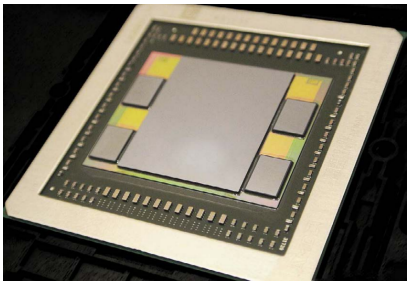


Figure 1.9: IP maturity for certain IP blocks at different process nodes [12].

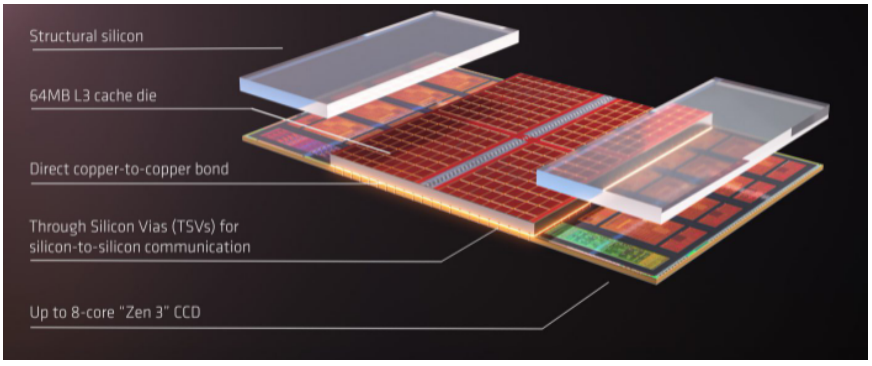
quency Radio Frequency (RF) front-ends wherein devices made from different processes such as Gallium Nirtide (GaN) or Gallium Arsenide (GaAs) power amplifiers might need to be integrated with Silicon (Si) logic circuitry[13, 14]. Integration of Micro-Electro-Mechanical System (MEMS) devices with CMOS for applications such as smart sensors and Internet of Things (IoT) is another area which could benefit from polylithic systems[15]. Another area of interest is integrating Complementary Metal Oxide Semiconductor (CMOS) devices with photonic chips to facilitate high speed off chip signalling links required by data intensive applications[13].

Finally, for HPC applications heterogeneous integration provides unique architectural benefits that can overcome the limitations of conventional designs. Highly parallel computing systems such as GPUs, Artificial Intelligence (AI) accelerators and multi-core CPUs typically have very high demand on their memory systems. Access to high bandwidth and low latency memory is critical to unlock the full processing potential of these processing units [16]. Furthermore, the energy requirements for data movements in these data intensive systems can drastically affect the overall energy requirements. To these ends, bringing

the memory systems into the package and using advanced packaging techniques can enable low latency, high throughput, efficient links to the processing units. Depending on the system requirements, this can include Dynamic Random Access Memory (DRAM) technologies such as High Bandwidth Memory (HBM) and Graphics Double Data Rate x (GDDRx) or Static Random Access Memory (SRAM) technologies such as stacked vertical cache as shown in Figure 1.10.



(a)



(b)

Figure 1.10: Example products showing various memory-in-package options for HPC. (a) HBM DRAM connected to GPU using a silicon interposer for AMD Fiji GPU. (b) Graphic showing the concept of AMD 3D V-Cache technology with an SRAM stacked over the CPU die.

1.1.2 Disadvantages of Heterogeneous Integration

Despite the many advantages of going the chiplet route, splitting the SoC into smaller chiplets may not be be best solution in all use cases. 'Chipletizing' adds additional complexity into the overall process flow with the need for more in-line testing steps, modification to design flows as well as adds silicon overheads [17, 18]. The overheads would

come from both the communication and testing overheads. Additional driver and receiver circuitry along with protocol conversions, circuits to cross voltage and timing domains, etc maybe required to drive the inter-chiplet links [6]. The testing silicon overhead comes from the inherently 'incomplete' nature of certain chiplets, necessitating the need for additional logic to facilitate KGD testing [17]. The heterogeneous equivalent may also suffer a performance penalty depending on the kind of die-stitching approach used.

Splitting an SoC can distribute IP that resides within a single chip and accessible through back-end metallization on traditional chips to non-uniform configurations with some blocks accessible over Back End of Line (BEOL), and others needing access through off-chip interconnect. Depending on the performance gulf between these interconnects, these may also introduce potential challenges to functionalities. For example, splitting a multi-core CPU into chiplets may result in Non-uniform Memory Access (NUMA) latency issues depending on the request being directed to a local or a remote resource [6].

Furthermore, KGD testing steps can also increase the overall testing costs [19]. Further, potential package failures caused due to the increased packaging complexity can increase the scrap cost as well as decrease the package assembly yields [18, 20].

1.1.3 Landscape Overview

While the concept of SoC partitioning has been around for a long time, the aforementioned need for system level scaling tools has led to the mainstream adoption of chiplet based systems in most major market segments. This in turn has led to a wide variety of integration architectures, each with its unique strengths and limitations. In this section, an overview of the major SiP integration architectures would be presented.

As per the classification of integration architectures as defined by the IEEE Electronics Packaging Society (EPS) Heterogeneous Integration Roadmap (HIR), multi-die systems can be classified into 2D, 2.5D or 3D architectures based on the die placement and interconnection type used. An overview of this classification is captured in Figure 1.11.

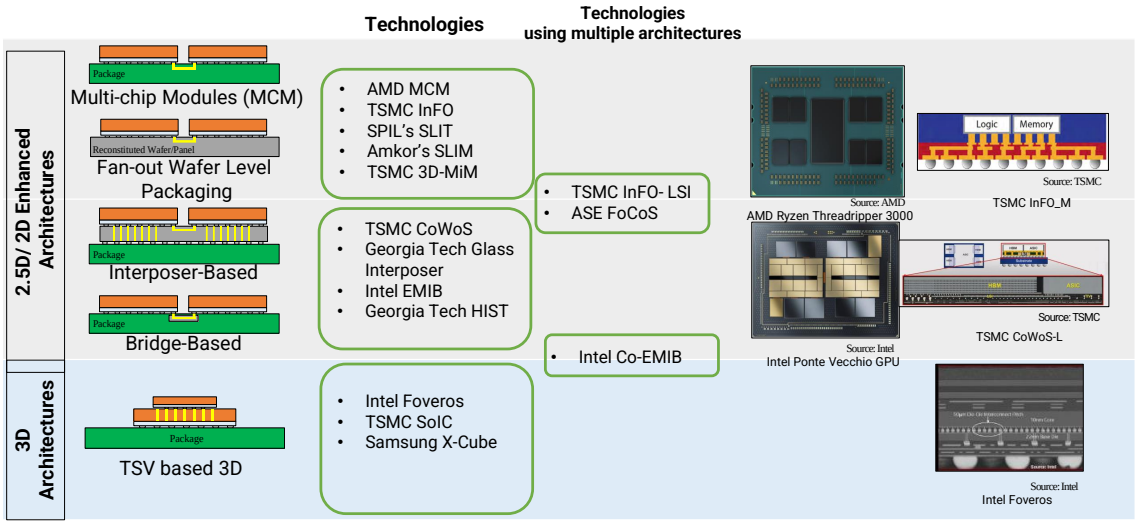


Figure 1.11: Converged nomenclature of heterogeneous ICs. Adapted from [21].

2D architecture consists of chiplets placed side-by-side on a conventional organic substrate to form a Multi-Chip Module (MCM). While this implementation offers the simplest process-flow for chiplet integration, the inter-chiplet communication density is limited due to the low wiring density of conventional organic substrates as well as the large pitch of solder based interconnection between silicon and the substrate. To this end, 2.5D architectures where the lateral interconnection density can be improved using substrates that can support finer wiring pitches. These may include fine pitch organic substrates, glass or silicon interposers, fine-pitch redistribution layers or even localized silicon or glass bridges. The fine wiring pitch available on these are supplemented with fine-pitch first level interconnections to the package using Compressible Collapsible Chip Connect (C4) bumps, microbumps or copper-to-copper direct bonds.

Finally, 3D architectures refer to the integration schemes where active chips are stacked on top of each other and connected through vertical interconnects. This is typically accomplished using Through Silicon Via (TSV) which are dielectric-clad copper filled vias running through the silicon. 3D architectures can offer the highest interconnect performance in comparison to other integration architectures owing to the very short distance

(die-thickness + interconnect height), area array of interconnects available for signalling and power delivery between chiplets [22]. For example, Beyne et.al. demonstrated that moving to 3D architecture with TSVs and hybrid bonds for a 256 core SoC can deliver up to 40% improvement in operating frequency, 48% improvement in power, and 12% improvement in total wire-length [23].

1.2 Enabling Technologies for Advanced Heterogeneous Integrated Devices

As illustrated, various applications and product categories demands very different integration architectures for HI SiPs with optimal performance. However, all of these require few basic technology enablers to ensure performance scaling to keep up with the ever increasing demands. Two major areas of particular interest are discussed in the subsequent parts of this section. Enabling technologies to potentially solve these are also discussed in more detail in the subsequent chapters of this thesis.

1.2.1 The Interconnection Challenges

Die-to-die interconnections play an important role in determining the performance of a polylithic system in comparison to its monolithic equivalent as shown in Figure 1.12. These include the vertical interconnects between the stacked dice in a 3D stack, as well as the interconnects between the dice and the substrate/interposer/bridge chip in case of 2.5D architectures. These interconnects are responsible for enabling all shared signal links as well as any shared power domains. To this end, optimizing the signal and power-delivery performances of these interconnect elements are required to enable effective stitching of chiplets.

The main performance metrics considered for signalling links include (1) the available interconnect density, (2) available communication bandwidth (3) link latency and (4) link energy consumption. While the exact values of these metrics for any given configuration would depend on the driving circuitry as well as the interconnection protocol used, a repre-

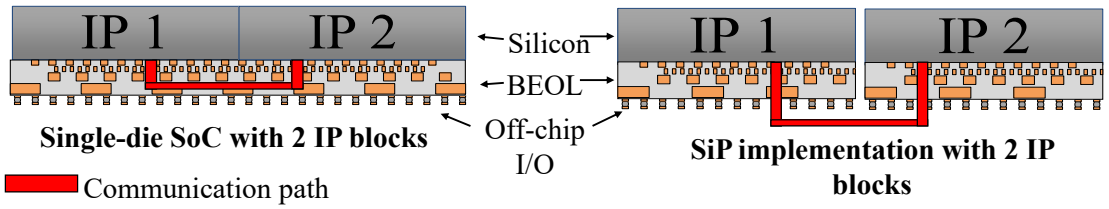


Figure 1.12: Interconnect bottleneck in a multi-die system. All inter-chiplet communication needs to pass through the off chip interconnects.

sentative trend can be estimated from just the physical construction used in implementation (i.e. the geometric parameters of the interconnect as well as the integration architecture type). Geometric scaling of interconnects is a powerful tool that can help improve these metrics, bringing the system performance comparable to monolithic implementations.

The higher interconnect density offered by interconnect scaling allows for increased raw communication bandwidth without relying on higher signalling speeds. The lower speed removes the need for power hungry peripheral circuitry required to drive high-speed signal links. It also eliminates the need for complicated error-correction circuits and reduces the interconnect latency [21]. These improvements helps reduce the area required for the interconnect circuitry along with the power savings.

Quantitatively, this interconnect scaling can be represented in terms of Three-Dimensional Interconnect Density (3DID) which is the product of the highest horizontal interconnect density (number of wires per mm) and the highest vertical interconnect density (number of bonds per mm^2) between chips in a heterogeneous package. System scaling for continued performance improvements with HI architectures require an aggressive scaling of 3DID analogous to the geometric scaling of gate pitches that fueled the classical scaling era. However, as can be seen from Figure 1.13, most conventional packaging technologies show a much more modest improvements in interconnect density over time, requiring the development of novel interconnects that can bridge this gap.

Expected trends of interconnect scaling for 2.5D and 3D devices are captured in Table 1.1 and Table 1.2. Table 1.3 and Table 1.4 show the corresponding signal integrity

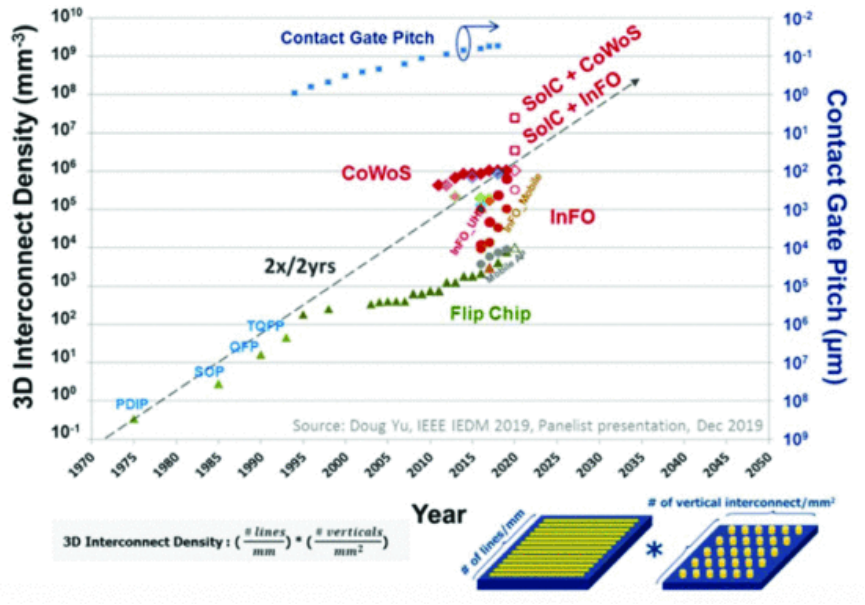


Figure 1.13: 3D interconnect density scaling of various packaging technologies from TSMC versus gate contact pitch scaling [24].

Table 1.1: Physical IO Scaling Roadmap for 2D and Enhanced-2D Architectures that use both solder and hybrid interconnects [21].

Generation Number		1	2	3	4	5
Raw Linear Bandwidth Density (GBps/mm)		125	250	500	1000	2000
Package Technology	Minimum Bump Pitch (μm)	55	40	30	20	10
	Linear Escape Density (IO/mm)	500	667	1000	2000	4000
	Areal Escape Density (IO/mm^2)	331	625	1111	2500	10000
Signaling Speed (Gbps)		2	3	4	4	4

attributes for these scaling trends. These show the aggressive improvement in die-to-die signalling that can be obtained by physical scaling of die-to-die interconnects. However, the interconnect scaling required to support the system-level scaling of heterogeneous devices presents multiple technological challenges.

While solder-based interconnects were primarily used for off-chip interconnects in traditional SiPs, their scalability to finer pitches is limited. Another commonly used alternative is solid-state diffusion bonding based copper-to-copper interconnects. These exhibit better scalability at the cost of a more stringent process-flow, typically requiring very high

Table 1.2: Physical IO Scaling Roadmap for 3D architectures that use both solder and hybrid interconnects.[21]

Generation Number		1	2	3	4	5
Raw Areal Bandwidth Density (GBps/mm^2)		125	250	500	1000	2000
Package Technology	Minimum Bump Pitch (µm)	40	30	20	15	10
	Areal Escape Density (IO/mm^2)	625	1111	2500	4444	10000
Signaling Speed (Gbps)		1.6	1.8	1.6	1.8	1.6

Table 1.3: Channel Signaling Characteristics for 2D and Enhanced-2D Architectures. [21]

Generation Number		1	2	3	4	5
Linear Bandwidth Density (GBps/mm)		125	250	500	1000	2000
Channel Performance	Channel Length (mm)	<2	<1.7	<1.4	<1.1	<0.8
	Bump-to-Bump Channel RC (ps)	<10	<10	<10	<10	<10

levels of planarity, surface cleanliness and high temperature/pressure bonding. Hence, there is a need for a more scalable approach to create fine-pitch die-to-die interconnects.

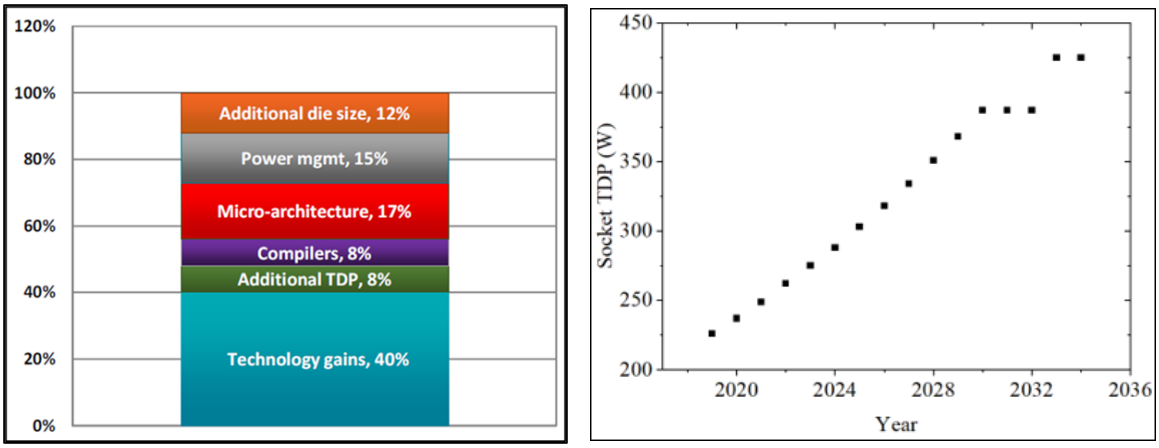
Table 1.4: Channel Signaling Characteristics for 3D Architectures. [21]

Generation Number	1	2	3	4	5
Areal Bandwidth Density (GBps/mm2)	125	250	500	1000	2000
Bump Capacitance (fF)	<30	<22	<15	<10	<7

1.2.2 The Thermal Challenges

Compute scaling enabled by heterogeneous integration relies on the ability to pack more functional silicon into a smaller footprint and interconnect the individual chiplets effectively to deliver improved performance. Furthermore, at a chiplet level, the stagnation in Dennard's scaling forces an increase in chiplet Thermal Design Power (TDP). This, alongside the more modest returns from node scaling is required to deliver the generational improvement in performance. In fact, increased device TDP is being used as a major driver for sustaining the generational improvement in compute throughput. For example,

AMD has stated that approximately 8% of performance enhancement for server CPUs per generation comes from raising the TDP envelope as shown in Figure 1.14a.



(a) (b)

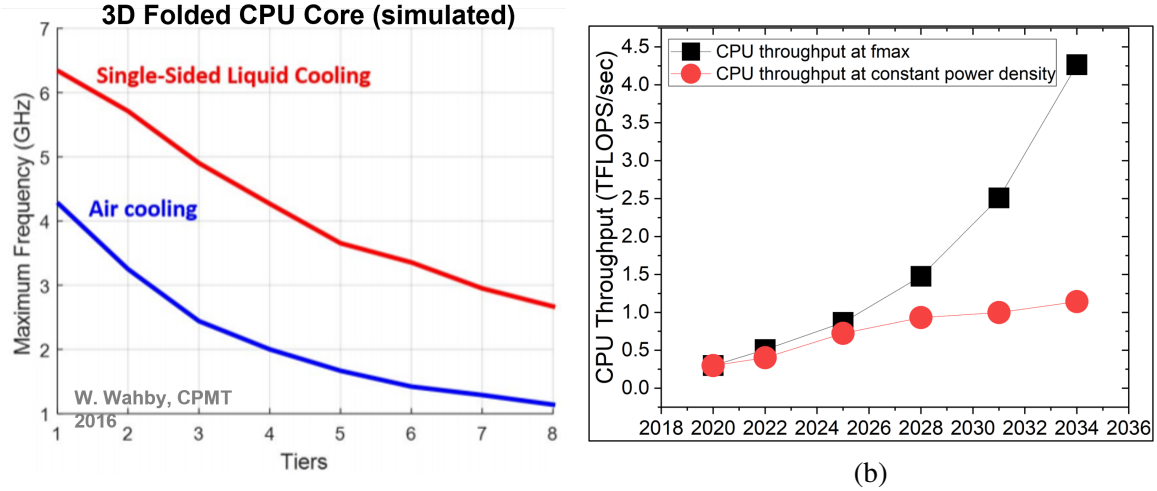
Figure 1.14: The increasing TDP of HPC products. (a) Elements of 2× in 2.4-year performance gain in server CPUs, showing the additional TDP added year-over-year to sustain performance gains. [11] (b) Scaling projection of CPU socket powers [25].

These trends have led to multiple thermal challenges. Firstly, the aforementioned reduced efficiency gains from technology scaling and the increased amount of silicon is leading to a considerable increase in package powers of leading edge compute chips. This is illustrated by the increasing trend in CPU socket powers for datacenter CPUs as projected by the IEEE International Roadmap for DEvices and Systems (IRDS) shown in Figure 1.14b.

This increase puts substantial strain on traditional thermal management approaches such as air-cooled heat sinks and coldplates. Specifically, the two main challenges are (1) ability to maintain the operating temperatures of the increasingly power dense devices to unlock the full device performance, [26, 27], and (2) to do so with minimal overheads (energy and water consumption) for thermal management [28]. The latter is particularly important for datacenters to operate in an environmentally sustainable manner [29]. Paradaigm shifts in cooling techniques are required to facilitate this transition, considering the power utilization efficiency of datacenters with conventional approaches are improving only at a small rate [30]. Further, device operation at lower junction temperatures can help reduce

the leakage power [31], which is an important component of power consumption in modern general-purpose processors [32]. Lower temperature operation can also increase the lifetime of the devices as major lifetime degradation pathways such as gate oxide breakdown and electromigration are proportional to the operational temperature [33–36].

Secondly, the close proximity of individual chiplets in a heterogeneous package can lead to thermal coupling between them. This increase in operational temperature of a chiplet due to the coupled power from a neighbor can lead to performance degradation in any thermally sensitive components. Finally, for 3D heterogeneous ICs, the increased device density from logic folding leads to increased power density, and correspondingly, higher device temperatures, which can throttle device performance, as shown in Figure 1.15a. Conversely, keeping the power density constant while increasing the device count can lead to reduced returns in computational throughput, as shown in Figure 1.15b. Therefore, the ability to extract heat from higher power density devices is essential to deliver the full benefits of system scaling.



(a) (b)

Figure 1.15: The effect of increasing power density on compute performance. (a) Effect of increasing power density due to logic folding under constant cooling conditions. [37] (b) Potential projection performance throttling from keeping power density constant [25].

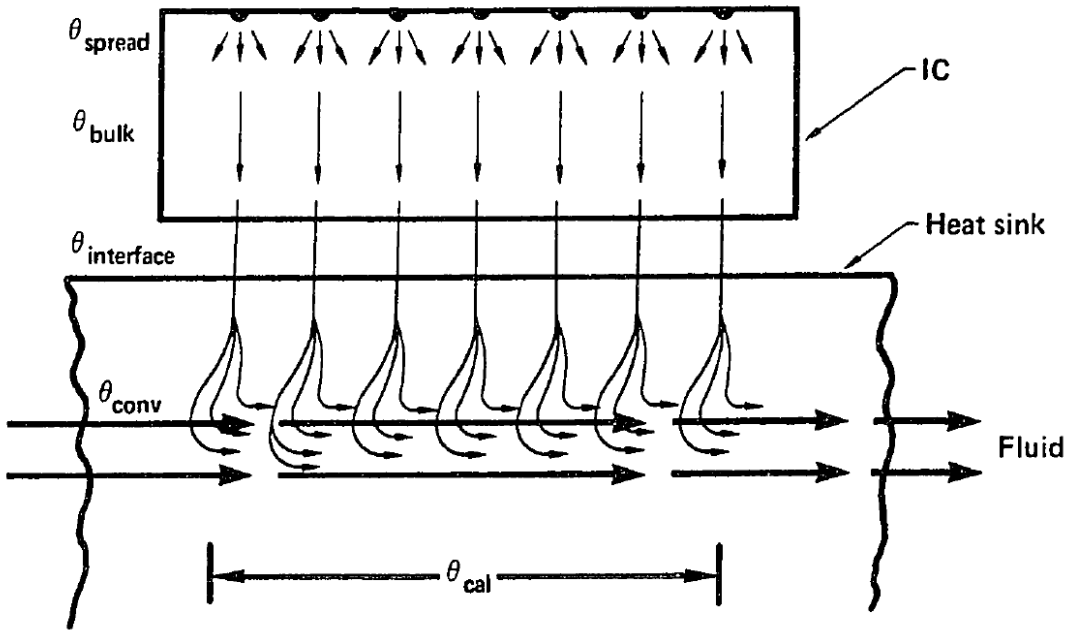


Figure 1.16: Components of thermal resistance in convectively cooled ICs [39].

Monolithic Microfluidic Cooling

Monolithic microfluidic cooling, first demonstrated by Tuckerman et.al [38], is a potential candidate to mitigate these challenges. This approach provides a very low overall thermal resistance by significantly reducing the major components of thermal resistances shown in Figure 1.16 when compared to conventional approaches. The conductive thermal resistance is reduced by bringing the working fluid closer to the source of heat generation, thereby reducing the bulk of material through which the heat has to travel through. Further, this removes the multiple interfaces and corresponding interface resistances along the heat-removal path. The caloric resistance is reduced by switching to a fluid with higher volumetric heat capacity such as water or other dielectric coolants. Finally, the large area-to-volume ratios that can be achieved by fabricating high aspect ratio microchannels or micropin-fins can be used to reduce the convective thermal resistance.

Sarvey et.al. expanded the concept demonstrations on non-functional silicon to a functional CMOS chip [40], showcasing the potential benefits. However, there is a need to

evaluate the approach at aggregate powers and heat fluxes more consistent with modern CPUs. Further, scaling to 2.5D architectures require taking into account the inherent differences from single die implementations such as the heterogeneity in the size, thickness, and cooling requirements of individual chiplets. Finally, while inter-layer microfluidic cooling has been demonstrated as a potential solution for the increased power density in 3D ICs [41], this approach requires co-optimizing the electrical performance of the vertical interconnects within the heatsink (TSVs within the microchannels or micropin-fins) to ensure acceptable electrical and thermal performance.

1.3 Research Objectives and Contributions

The main focus of this research is to develop enabling technologies that address the interconnect and thermal management issues in heterogeneous ICs. The main areas of focus detailed in this thesis are as follows:

1. **Selective metal electroless plating for die-to-die bonding.** A selective metal electroless plating based process-flow, in conjunction with mechanical self-alignment is developed for creating die-to-die interconnects. A more detailed discussion about this approach in contrast with the traditional solder and copper-to-copper based interconnect approaches would be presented. Results from an initial demonstration testbed would be presented. The electrical performance of these interconnects would be compared with respect to other I/Os using finite element modeling.

2. **Evaluation of operational benefits of microfluidic cooling for functional CMOS.** The implementation of monolithic microfluidic cooling on an off-the-shelf, functional CPU is presented. Heatsink design, including fluid delivery approaches with 3D printed manifolds, are evaluated. The improvements in computational performance unlocked by the increased power-handling capabilities are quantified by overclocking the chip to higher operational frequencies. The reduction in cooling power

and water requirements are also quantified and compared with other traditional cooling solutions.

3. **Demonstration of monolithic microfluidic cooling in a functional 2.5D IC.** The heatsink design discussed above is extended to a functional 2.5D Field Programmable Gate Array (FPGA). Additional design optimizations that cater for the heterogeneity of the chiplets are presented. The concept is experimentally demonstrated, and the absolute thermal resistance and thermal coupling values are quantified and compared with other cooling approaches.

4. **Development of a scalable process-flows for high aspect ratio TSVs in a micropin-fin heatsink, co-optimized for thermal and electrical performance.** Implementation of inter-layer microfluidic cooling in 3D stacks benefit from taller micropin-fins, and correspondingly, taller TSVs for good thermal performance. To achieve this without considerable reduction in TSV density or higher TSV parasitic capacitance associated with large diameter vias require an increase in TSV aspect ratio. A high aspect ratio TSV process flow is developed for this, and the fabrication and high frequency measurement results are presented.

1.3.1 Thesis Statement

This thesis introduces electroless plating in conjunction with mechanical self alignment technologies to create a scalable die-to-die interconnection platform. The thesis also explores the use of microfluidic cooling to deal with the high package powers and thermal coupling issues associated with 2.5D and 3D ICs.

1.4 Organization of this Thesis

The remainder of this thesis is organized as follows:

1. Chapter 2 details the fabrication of the testbed to evaluate the use of nickel electroless

plating for creating die-to-die interconnects. The testbed complements the selective electroless plating with mechanical self-alignment to create a more scalable platform. The fabrication results including the alignment accuracy data are presented.

2. Chapter 3 describes the fabrication and measurement results of the implementation of monolithic microfluidic cooling on an Intel i7-8700K CPU. The performance of the device and the cooling overheads is evaluated for operation under overclocked conditions.

3. Chapter 4 extends the microfluidic heatsink to a 2.5D IC. The design and evaluations are extended to include thermal coupling alongside aggregate power scaling.

4. Chapter 5 demonstrates the need for high aspect ratio TSVs to enable inter-layer cooling with results from HIgh-frequency Structure Simulator (HFSS) simulations. A scalable process flow developed for fabricating these structures is described. The electrical parasitics of the fabricated TSVs are extracted and compared with data from simulations.

5. Chapter 6 summarizes the key conclusions of this thesis, and potential avenues for future work are discussed.

CHAPTER 2
HIGH DENSITY OFF-CHIP INTERCONNECTIONS ENABLED BY SELECTIVE ELECTROLESS PLATING AND MECHANICAL SELF-ALIGNMENT TECHNIQUES

2.1 Limitations of Conventional Interconnects

Solder based interconnects has been the mainstay of chip level interconnects for the past few decades. This include the standard flip-chip bumps with 100 μm diameter and 150-200 μm pitch, fine pitch C4 bumps with 50 μm diameter and 100 μm pitch to micro C4 bumps with 20-30 μm diameter and 30-60 μm pitch [42]. An overview of these scaling trends are shown in Figure 2.1. However, as we start to scale to lower pitches, these have some major limitations. These include the formation of higher fraction of Inter-metallic Compounds (IMCs), making the joint brittle and introducing potential reliability concerns, solder extrusion, bridging etc [43].

For finer pitches, solder capped copper pillars are employed. These include copper pillar bumps with 15- 20 μm diameter and 20-40 μm pitch and solder capped flat copper pillars with 10 μm diameter and 20 μm pitch [45]. However, the above-mentioned limitations prevent scaling solder based interconnects to lower pitches, and requires using other types of interconnect techniques for higher density connectivity.

2.2 Copper-to-Copper Bonding Techniques

Using solder-less, copper-to-copper direct bonding techniques are a promising alternative to overcome these limitations of solder based interconnects. Furthermore, copper has the distinct advantage of the interconnect material already present in BEOL. Also, Cu based interconnects have superior electrical and thermal performance compared to solder

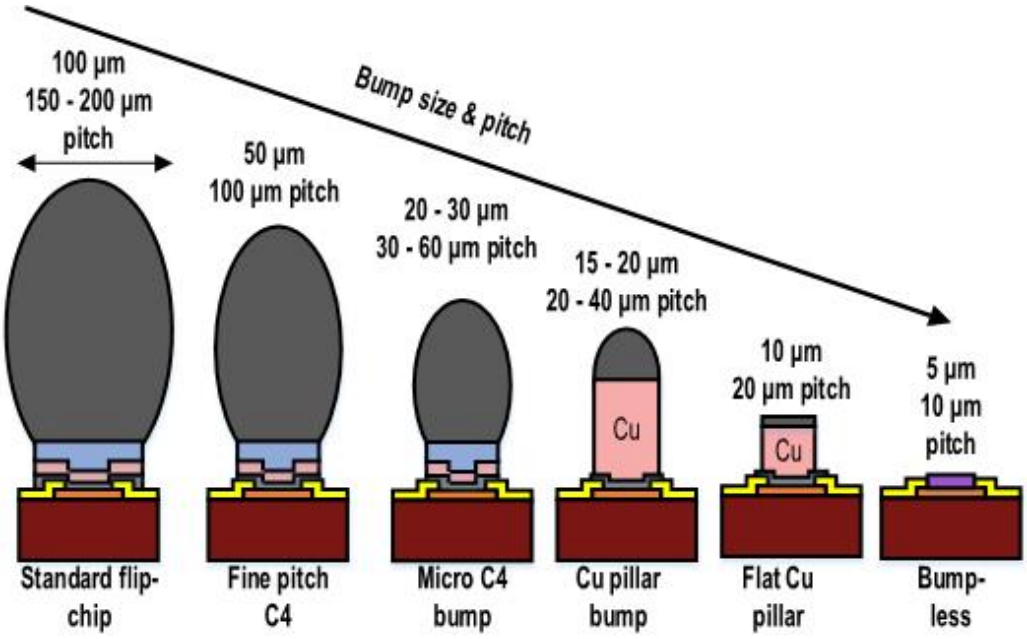


Figure 2.1: Interconnect scaling trends [44].

based interconnects. The lower resistivity of Cu results in lower parasitics, translating to superior performance, even compared to solder interconnects of similar dimensions.

The major techniques used to create direct Cu-to-Cu bonding involves direct thermo-compression bonding [46], Self-Assembled Monolayer (SAM) [47], surface activated bonding [48], insertion bonding etc [49]. The basic technique employed in all these techniques is to facilitate copper inter-diffusion between two pads/ pillars in contact to create a homogeneous bond. For thermocompression bonding, this is achieved by applying high temperature and pressure to force the diffusion.

Since all these techniques rely on inter-diffusion of copper atoms between two pads in contact [50], they place stringent requirements on the pads to be bonded. The ability of the atoms to diffuse, and hence the subsequent bond quality is determined by the planarity and cleanliness of the pads. Very accurate initial alignment is another requirement as these techniques doesn't have any self aligning capabilities like solder bonds. The dice / wafers used for these interconnection processes, thus, should be able to withstand these stringent planarization and cleaning processes. Furthermore, these techniques are typically high

temperature, high pressure processes [46–48], hence limiting the use to dice/wafers which can withstand these conditions.

2.2.1 Hybrid Bonding Techniques

Another solid state diffusion based approach, hybrid bonding is recently gaining traction as another promising alternative to achieve very fine interconnection pitches. This method involves simultaneous bonding of the interconnect pads (typically copper) and the surrounding dielectric material of two dice/wafer to create a very low profile, low parasitic, high performance interconnect between them. The two major formats of hybrid bonding used typically involve polymer dielectric materials [51] or silicon dioxide dielectric. Polymer based hybrid bonding typically helps in achieving lower temperature process flow but silicon dioxide or other interlayer dielectric based hybrid bonding is more commonly used in industry like Xperi Direct Bonded Interconnect (DBI) [52], TSMC System-on-Integrated Chip (SoIC) [24], and Intel Foveros direct [53].

The main advantage of hybrid bonding is the ability to create a monolithic like stack, which has superior electrical and thermal properties similar to a single die [54]. However, the process is dependent on the BEOL materials, especially the top layer dielectric being amenable to direct bonding. Furthermore, it comes with most of the stringent requirements associated with facilitating solid-state diffusion as in the case of traditional Cu-to-Cu direct bonding techniques.

2.3 Metal Deposition-Based Bonding

As discussed, the major challenges associated with scaling the copper-to-copper direct interconnect technologies comes from the requirements to facilitate solid-state diffusion of copper. While approaches like SAM or controlling the crystalline orientation of copper can help make these less stringent, it still is the major limiting factor that limits the scalability of these fabrication techniques. To this end, an alternative approach is to replace

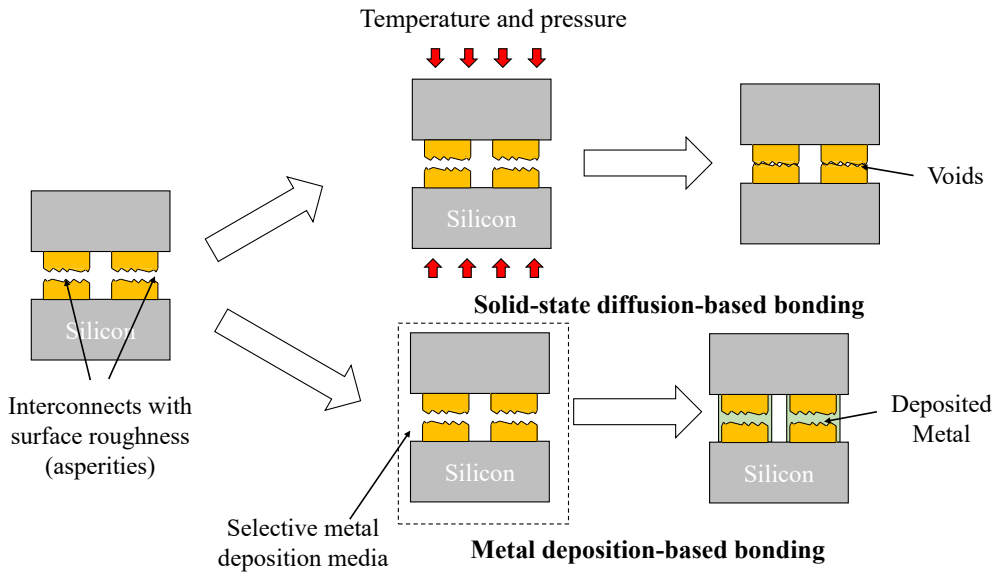


Figure 2.2: Basic conceptual difference between planarity requirements of diffusion-based bonding approaches and deposition-based bonding.

this metal diffusion based approach with a metal deposition based approach. By selectively depositing metal just between the structures to be connected, a good electrical contact can be made without needing to planarize them to levels that can facilitate solid state diffusion. Figure 2.2 shows this concept and the advantages of switching to a deposition based approach.

Selective electroless deposition and selective Atomic Layer Deposition (ALD) [55] are two promising techniques to achieve deposition based die-to-die interconnects. For both these approaches, a selective metal deposition process is used so as to bridge the gap between metal pads/pillars of pre-aligned dice. Process chemistry's which can deposit material only on the pads facing each other, with no deposition happening on the surrounding dielectric is critical for this application.

2.4 Selective Electroless Plating for Die-to-Die Interconnects

In this work, we propose the use of high density, low temperature electroless metal deposition based interconnects as an alternate technique to overcome the various limita-

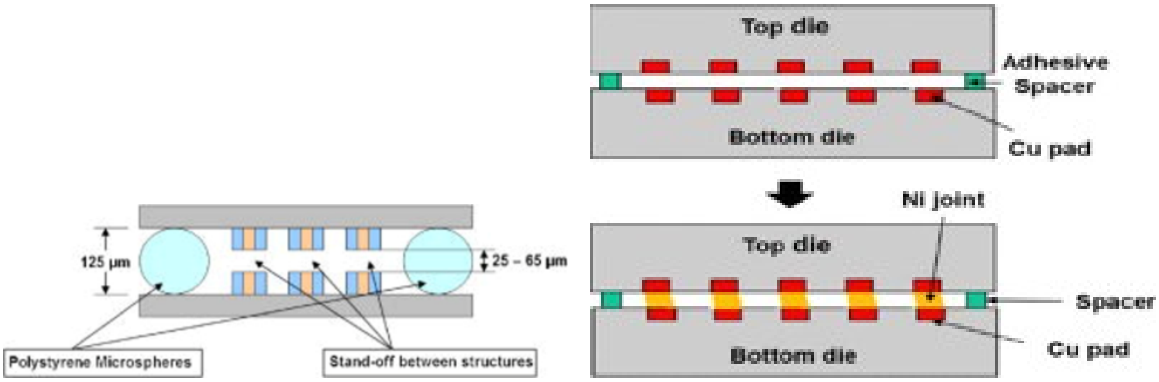
tions of other high temperature interconnection techniques discussed above. The idea is to do face-to-face alignment of Die-to-Die (D2D), Die-to-Wafer (D2W) or Wafer-to-Wafer (W2W) stacks such that the Cu pillars or pads to be bonded are aligned in the lateral directions but has a small, predefined vertical separation. This gap is then bridged by electroless deposited metal, deposited only on the pads using a metal-selective electroless plating chemistry. The experimental demonstration of this technology presented in the subsequent sections of this thesis is also published in [56].

Since the process does not require perfect contact, and subsequent inter-diffusion of Cu atoms, the planarity and cleanliness requirements are relaxed. Further, the process can be done at much lower temperatures, and atmospheric pressure. Another major benefit is the ability to create multi-pitch and/or multi-diameter interconnects simultaneously.

2.4.1 Prior Work

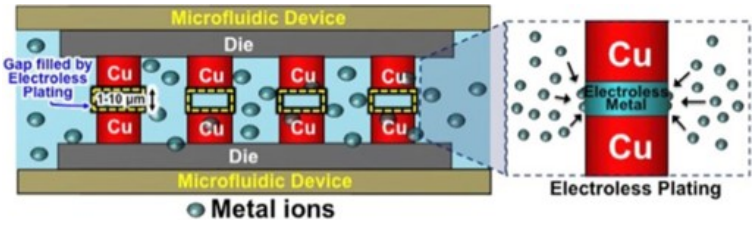
Due to these significant advantages discussed above, there has been multiple studies into the use of electroless plating for creating chip-to-chip interconnects. Osborn et.al used Cu electroless plating to develop a room temperature interconnection process at pitches greater than 100 μm [57]. The approach used polystyrene spheres to create the stand-off gaps between the pillars to facilitate the bonding process. Wu et. al. explored the use of Ni electroless plating for the same, but used an adhesive spacer for initial alignment and had a pillar to pillar pitch of 100 μm [58]. Yang et. al. used a specialized set-up wherein a microfluidic device with active fluid pumping to create electroless interconnects at 100 μm pitch [59].

All of these approaches provide a definitive proof of concept for the ability of electroless plating to be used as a potential method for low-temperature interconnection. However, the main limitation in all these approaches is the lack of scalability of the employed techniques. The non-standard approaches used to create and maintain the initial alignment severely limits the ability to scale to lower pitches while maintaining the required alignment



(a) Cu electroless plating based interconnects [57]

(b) Ni electroless plating facilitated by adhesive spacers [58]



(c) Ni electroless plating using microfluidic pumping [59]

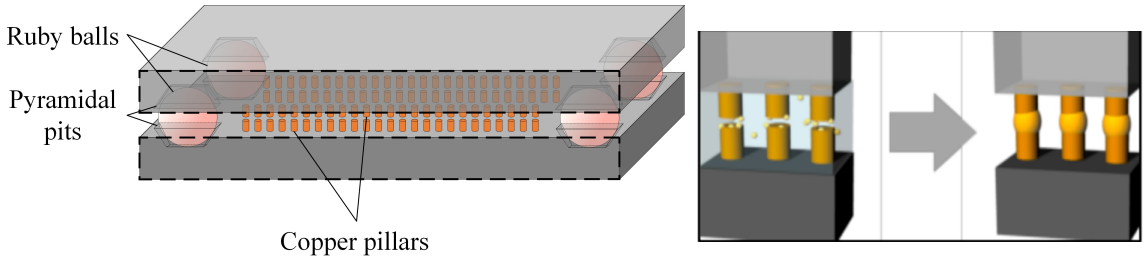
Figure 2.3: Prior work utilizing electroless plating.

accuracy, as well as the ability to create low-profile interconnects. Also, approaches which require active pumping through microfluidic cooling is not scalable to batch bonding and/or D2W and W2W bonding approaches. It is to this end that we propose a novel solution that combines mechanical self-alignment with an electroless deposition based bonding process to create a truly scalable low-temperature interconnection platform. Table 2.1 shows a comparison of our approach w.r.t the prior work in the area.

Mechanical self-alignment, where physical structures on the die or wafer can be used to

Table 2.1: Comparison of various electroless plating based bonding techniques

	This Work	**[57]**	**[58]**	**[59]**
Material	Ni	Cu	Ni	Ni
Temperature	95°C	20°C	70°C	95°C
Pitch	50 μm	$\geq$ 100 μm	Not specified	100 μm
Batch Bonding	Yes	Yes	Yes	No



(a) A conceptual image of the die-stack with self-alignment structures to be bonded using electroless plating

(b) Interconnect formation by electroless metal deposition

Figure 2.4: Electroless plating with mechanical self-alignment.

precisely self-align it to another die and/or wafer provides a scalable approach to accurately align dice for bonding. In the technology demonstration described in this thesis, we investigate the feasibility of using a conjunction of mechanical self-alignment with electroless plating to create low-temperature, high density interconnects between chips. Combining self-alignment techniques with selective electroless deposition helps address the scalability issues of the previous electroless deposition approaches. The conceptual diagram for the process is shown in Figure 2.4. Mechanical self-alignment structures are used to align the dice with pads facing each other with a small, pre-defined stand-off gap and the assembled structure is then put in an electroless plating solution to deposit metal and bond the pads.

2.4.2 Mechanical Self-Alignment and Electroless Plating

While various self-alignment techniques including electrostatic, magnetic, surface tension based approaches exist, mechanical self-alignment was selected for this demonstration owing to its simplicity [60]. In particular, a ball-in-pit technology, where a precise crystalline ball is fit into complementary pyramidal pits formed with anisotropic etching of silicon on both dice to be aligned [61]. The detailed design and fabrication techniques used for this is described in detail in the subsequent sections of this thesis.

2.5 Experimental Demonstration

2.5.1 Selection of Plating Chemistry

Electroless plating is a method for thin metal film deposition by simple immersion into an electrolyte bath containing an aqueous solution of a salt of the said metal without the need for external electric potentials. Electroless deposition occurs as a result of simultaneous cathodic deposition of metal and anodic oxidation of the reductant at the immersion potential, E_{im}. These can be expressed as the following partial reactions:

$$\textbf{Cathodic Reaction} : ML_m^{n+1} + ne^- \rightarrow M + mL \tag{2.1}$$

$$\textbf{Anodic Reaction} : Red \rightarrow Ox + ne^- \tag{2.2}$$

where *L* is a Ligand, *Red* a reductant, *Ox* an oxidant and *n* is the number of electrons [62]. For the electroless plating to take place, the oxidation potential of the reductant is less noble to the reversible potential of the metal to be deposited and there has to be enough catalytic activity on the substrate for the anodic oxidation to take place. Therefore, the selectivity of the deposition can be controlled using these factors. Most commonly used reductants for electrolytic deposition include hypophosphite, formaldehyde, borohydride, dialkylamine borane and hydrazine. Selection of a electroless chemistry with a reductant which favors the catalytic activity for the metal to be deposited on copper compared to the surrounding dielectric can provide selective deposition just on the metal pads.

Another important criterion in choosing the electroless plating chemistry is the temperature for the deposition process. Lower temperature deposition, within the CMOS compatible limits (sub 350 °C) is preferable as it ensures that this approach can used for functional CMOS. Furthermore, lower temperature bonding also ensures less thermo-mechanical issues during the bonding process when using dissimilar substrates with different Coefficient

Table 2.2: Comparison of common electroless chemistries

	Nickel (phosphite based)	**Nickel (boron based)**	**Copper**	**Gold**
Selectivity	Yes	No	Yes (certain chemistries)	Yes
Electrical conductivity	Fair	Fair	Good	Fair
Maximum Thickness	High	High	Low	Very Low
Temperature	$\approx$ 100 °C	$\approx$ 100 °C	$\approx$ 40 °C	$\approx$ 70 °C

of Thermal Expansion (CTE).

Electrical properties of the bonded interconnect also needs to be considered for selecting the appropriate electrolyte. Electroless deposited copper is favorable in this regard owing to its lower resistivity when compared to other commonly used metals for electroless deposition like nickel or gold.

Finally, another practical consideration for the selection of plating chemistry is the maximum thickness of metal that can be deposited. Many electroless chemistries are self-limiting to thin films due to the reduction of the catalytic activity for anodic oxidation after an initial film is deposited. A smaller maximum possible film thickness translates to the need to align the pillars to a correspondingly small vertical separation with the self-alignment mechanisms, thereby increasing the vertical alignment accuracy requirement. Therefore, a higher maximum film thickness was preferred for this demonstration to ease the constraints on the self-alignment process.

A comparison of various commonly used electroless chemistries across the metrics discussed above is captured in Table 2.2. While electroless copper has the lowest electrical resistance, the lower deposition thickness increases the required initial alignment accuracy. To this end, a nickel hypo-phosphate based chemistry was chosen for this initial technology demonstrator at the expense of increased electrical resistance.

2.5.2 Design of Self-Alignment Structures

To achieve electroless plated interconnects with pitches finer than conventional solder based approaches ($\approx$ sub-60 μm regime), the maximum height of the pillar, and correspondingly the die-to-die gap falls in the sub-120 μm regime. The ball-in-pit technology is ideal to achieve these gaps without sacrificing on the initial mis-alignment tolerance, and hence was chosen for this demonstration.

Alignment precision of mechanically self-aligned structures depends on the complementary structures precisely locking into each other, forcing alignment between them. Therefore, ensuring perfect fit while maintaining the requisite gap is required. A perfect fit can be characterized as the diametric chord of the ball being tangential to the inner pit wall as shown in Figure 2.5. Correspondingly, the geometric relations between the die-to-die gap *g*, the pit opening width *w*, and ball radius of *r* can be defined as:

$$g = 2 \times (\frac{r}{sin(54.7°)} - \frac{w}{tan(54.7°)}) \tag{2.3}$$

The relative angle of 54.7° between the $<100>$ and $<111>$ planes was chosen as KOH etching of a $<100>$ silicon wafer was used to create the pyramidal pits (represented by α in Figure 2.5).

2.5.3 Testbed Description

For evaluation, we designed a test-bed emulating face-to-face die-die bonding. The dice were designed to have an area array of 50 $\times$ 50 copper pillars each with a diameter of 20 μm and at a pitch of 50 μm. A pillar height of 2.5μm and a pillar-to-pillar gap of 1 μm was chosen. The pillar height and pillar-to-pillar gap was fixed at these values to provide a scalable solution that can achieve sub-10 μm pitch without having to modify the mechanical self-alignment structures.

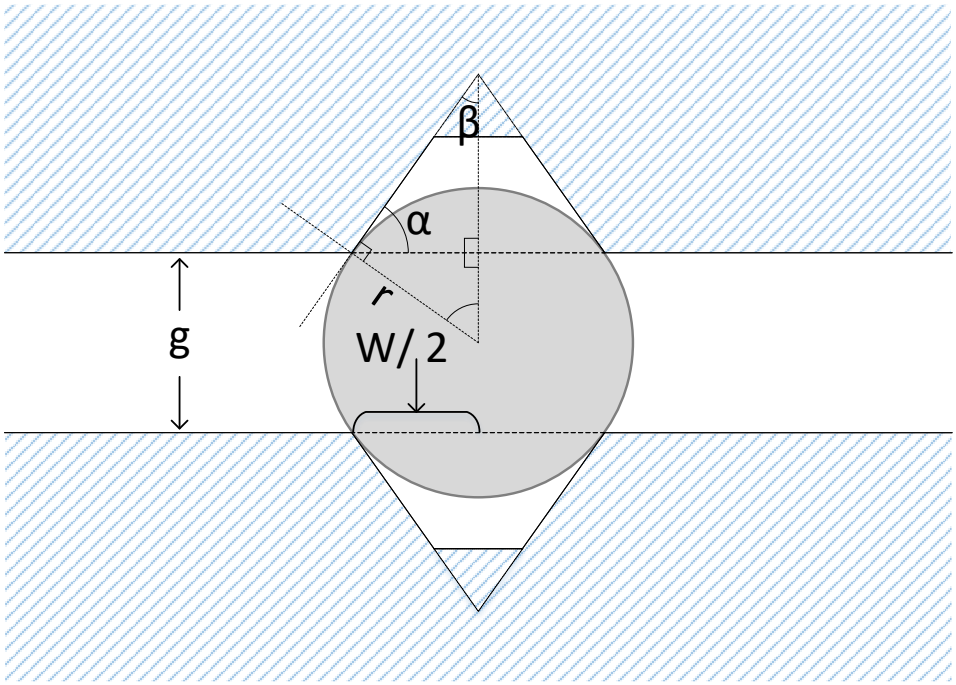
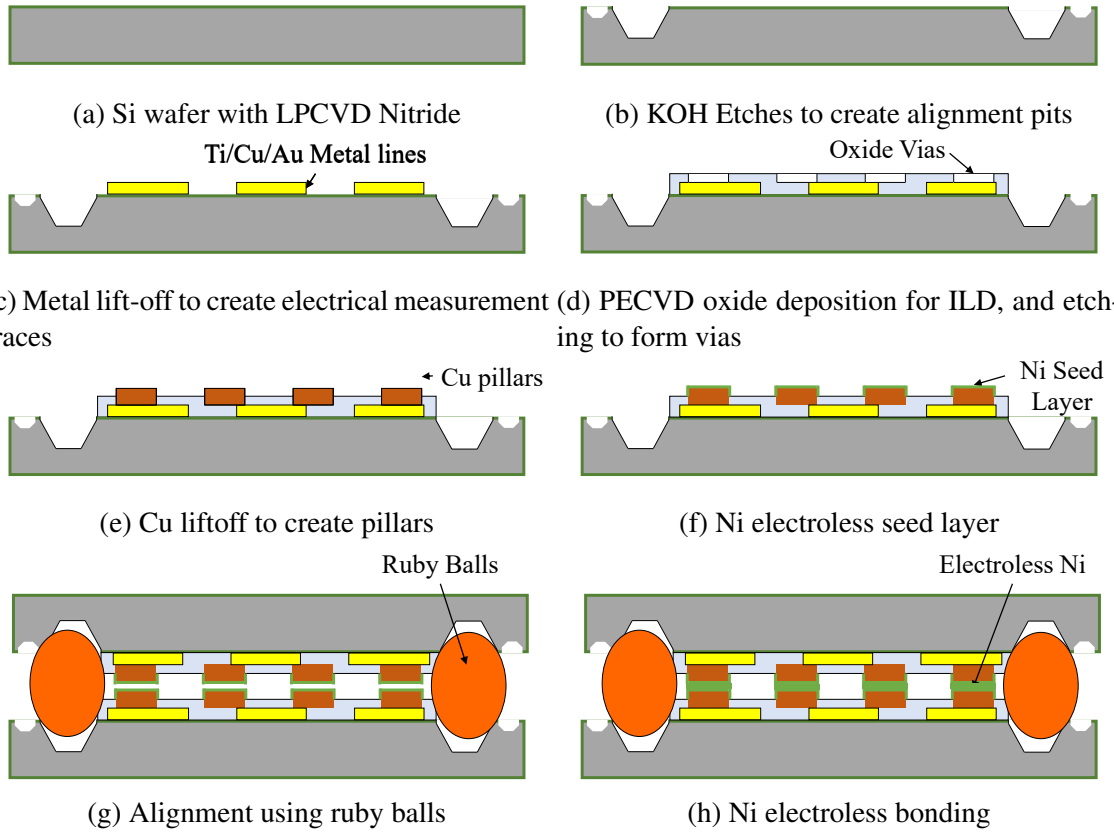


Figure 2.5: Ball-in-pit alignment mechanism.

2.5.4 Fabrication

The detailed fabrication and assembly flow is shown in Figure 2.6. The dice were fabricated on a 500 μm thick <100> silicon wafer. Wafers were initially cleaned in a piranha bath (3 parts H_2SO_4 : 1 part H_2O_2) followed by a HF acid dip to remove the native oxide layer and De-Ionized (DI) water rinse. The wafers were dried in a a spin rinse drier to remove the water. Next, a 130 nm thick silicon nitride (Si_3N_4) layer was deposited using a Low Pressure Chemical Vapor Deposition (LPCVD) process to act as the mask for the subsequent KOH etching step.

Anisotropic Si etching with KOH creates pyramidal pits with a uniform, repeatable undercut on all four sides for a specific etching process if the etch mask is perfectly aligned to the crystalline planes. Any mis-alignment in this step leads to distortion of the etched pits, with non-uniform undercut on the sides, and subsequently leading to poor overall alignment accuracy. Figure 2.7 shows the effect of crystalline plane mis-alignment on the



(a) Si wafer with LPCVD Nitride

(b) KOH Etches to create alignment pits

(c) Metal lift-off to create electrical measurement traces

(d) PECVD oxide deposition for ILD, and etching to form vias

(e) Cu liftoff to create pillars

(f) Ni electroless seed layer

(g) Alignment using ruby balls

(h) Ni electroless bonding

Figure 2.6: Process flow for fabrication and assembly.

etched features. Since Si wafer flats may have an error in the crystallographic orientation from the fabrication process, these cannot be used to determine precise crystalline planes.

Various techniques have been discussed in the literature to address this issue. Most of these rely on an initial feature etch used to reveal the crystalline planes, and then doing the subsequent alignment to these features. These techniques present a trade-off between the area-requirement for the initial-alignment structures, process complexity, and final alignment accuracy. A similar approach using a two-step etch process: an initial pre-etch of circular patterns neat the bottom of the wafer to reveal the crystalline plane and a subsequent etch of the actual pits by aligning to the revealed silicon plane is used. An NR-5 8000 photoresist mask was used to pattern both etch-masks. A Deep Reactive Ion Etching (DRIE) process was used to transfer the photoresist mask on to the Si_3N_4.

For pre-etch, a 45% KOH bath at 70°C was used for approximately 5 hours. This reveals

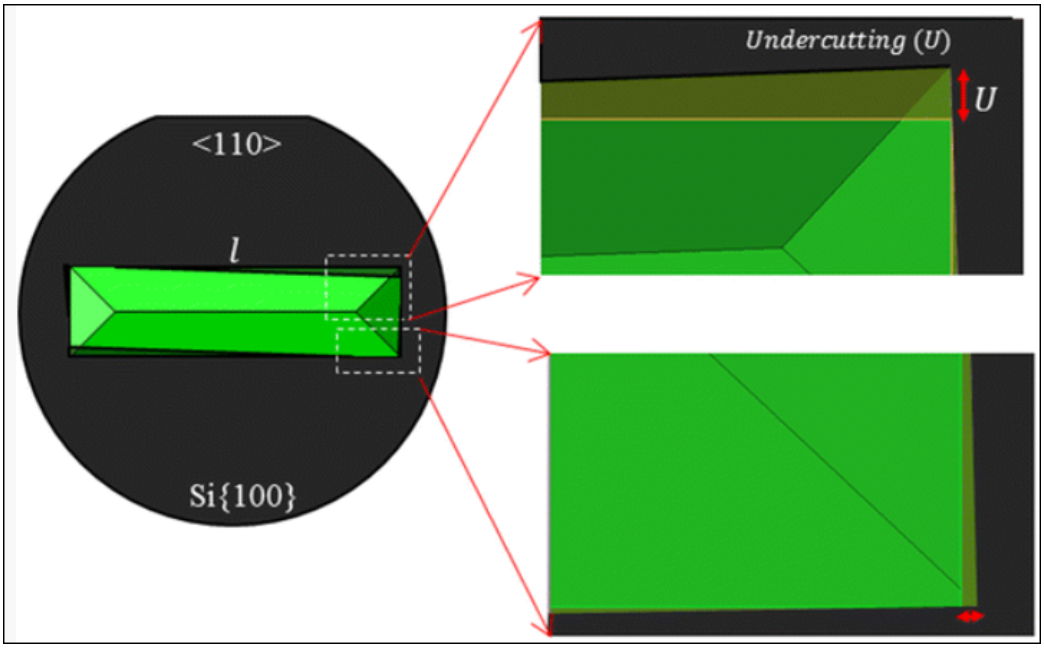


Figure 2.7: Effect of the misalignment of mask edges on the etched profile on <100> Si wafer [63].

the plane for subsequent alignment. The actual etch step was done in a a 45% KOH bath at 90°C. The different bath temperatures were chosen to control the undercut (etching under the Si_3N_4 mask) during the etch process. For the pre-etch step, enhancing the undercut is beneficial to properly reveal the crystalline planes for subsequent alignment. However, during the main etch step, increased undercut may lead to widening of the pit structures, and correspondingly, change the alignment gap. An average etch depth of 428 μm was targeted for the 300 μm $\times$ 300 μm square pits to get a die-to-die separation of 6 μm. The actual etch depth was characterized by contact profilometer measurements of the pits after the etch step.

A four step cleaning process using Acetone, Methanol, Isopropanol (AMI) and DI water was used before every photolithography step. This was followed by a dehydration bake at 150 °C for two minutes. After photoresist development a 30 second de-sum using oxygen plasma in a DRIE was used as well before subsequent etch or metal deposition steps. Finally, 2.5 μm thick Cu pillars are created using a lift-off process.

Next, the traces for electrical measurements are defined using a 15 nm Ti/ 200nm Cu/

50nm Au metal lift-off process. The Ti layer acts as an adhesion layer, and the Au layer serves as the passivation for the Cu traces. A PECVD silicon dioxide layer is deposited on top of the metal lines to serve as the inter-layer dielectric (ILD). Vias are patterned and etched using silicon dioxide RIE to access the metal lines.

The pillars are then cleaned with 50% hydrochloric acid for 30 seconds to remove the native oxide formation. Subsequently, the pillars are coated with a thin layer of nickel by immersing the wafer in Transene electroless Ni strike [64] solution at 95°C for 30 seconds. This deposits a thin seed layer of nickel around the pillars to enhance further nickel growth and also covers the pillars to prevent further copper oxidation. The wafer is then diced to separate the dice. The individual dice are cleaned initially in an ultrasonicated methanol bath followed by an isopropanol bath to remove the silicon debris from the dicing process sticking on to the silicon dioxide layer. This is critical as these silicon debris can act as a potential site for electroless nickel deposition on the silicon dioxide.

The dice are then assembled manually using Grade 3 [65] precision ruby balls with 500 μm diameter. The assembly is secured with room temperature cure epoxy and is then bonded by electroless plating in Transene Nickelex [66] electroless plating solution at 95°C for 10 minutes. The electroless plating is done in a temperature stabilized water bath within a closed container, and is agitated using a magnetic stirrer to ensure repeatability and minimize the variation of the electrolyte properties during the plating step.

2.5.5 Results and Evaluations

Electroless Ni Deposition

Cross-sectional SEM imaging of bonded samples was performed to do a qualitative analysis of the plating. Figure 2.8 shows a close-up cross-section of a plated pillar-pillar joint. Cross-sectional inspection shows that electroless deposition of nickel occurs selectively on copper pillars without plating on the surrounding dielectric region. An average pillar-to-pillar gap of $\sim$ 680 nm was measured from cross-sectional Scanning Electron Mi-

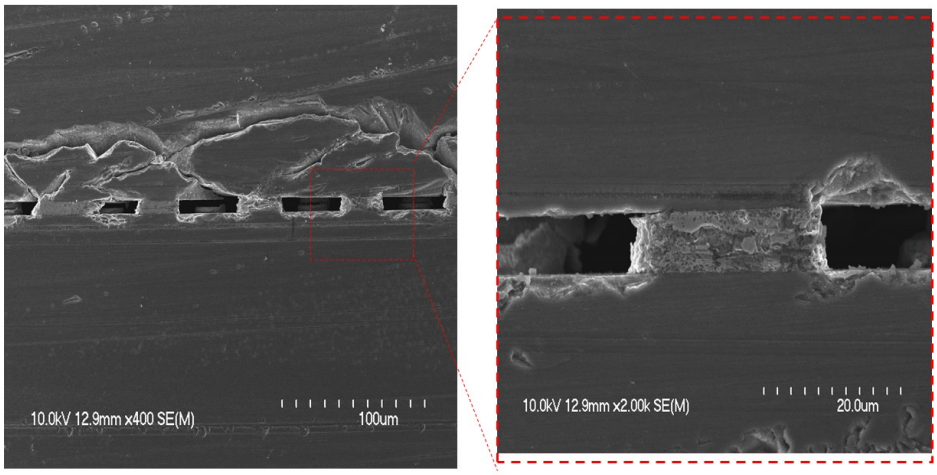
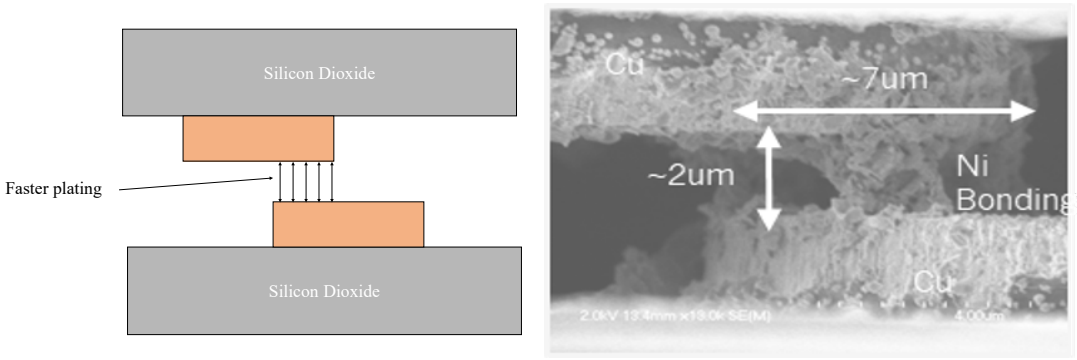


Figure 2.8: Cross-sectional SEM images showing pillars bonded by electroless plating.

croscope (SEM) images. This provides initial verification of nickel electroless plating as a means to achieve low temperature, pillar-to-pillar bonding at fine pitches. Further, cross-sectional analysis across a diagonal cut reveals that the plating is mostly uniform across the entire area array. This verifies that the assembly process allows for the fluid to pass through narrow gaps to facilitate plating.



(a) (b)

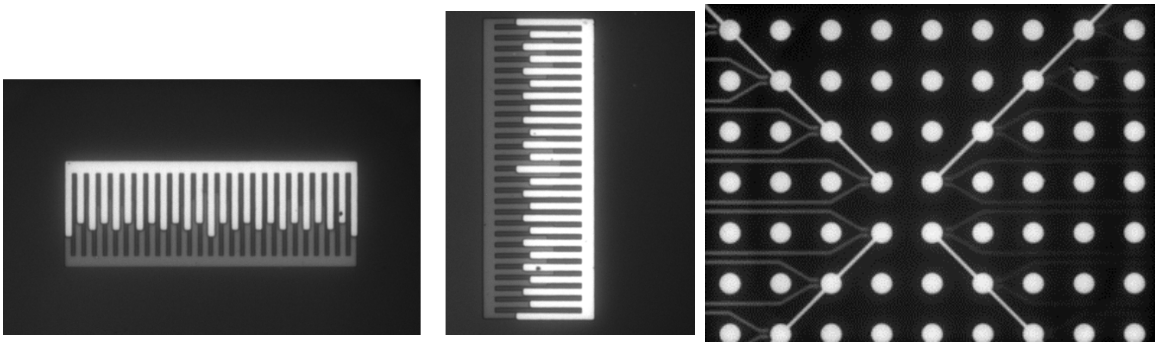
Figure 2.9: Enhanced plating in the smaller gaps. (a) Conceptual diagram. (b) Cross-sectional SEM showing enhanced plating in smaller gaps.

Moreover, we observed enhanced plating on the overlap areas between the pillars when compared to the sides and mis-aligned cases (Figure 2.9). This effect is reported previously in the literature [67] as potentially due to the mass-transport limitations of reaction inter-

mediate products in narrow gaps, leading to changes in localized plating conditions. This, however, is beneficial in this application as this anisotropy in plating can help us achieve bonding between dice without lateral shorting of the pillars.

Alignment Measurements

Die-to-Die lateral alignment was measured using metal vernier marks patterned on the assembled dice alongside the electrical traces. Post assembly, Infra-red (IR) microscopy was used to inspect the alignment between the vernier marks on the top and the bottom dice. The designed marks had a minimum resolution and accuracy of 1μm. Sets of two vernier marks, to measure X and Y misalignment, were patterned on all the four corners to get an average lateral alignment measurement. Average X and Y mis-alignment of $\leq$ 1μm was observed across multiple tests. IR images showing the vernier marks, as well as the electrical measurement traces are shown in Figure 2.10 (a). Figure 2.10 (b) shows the pillars aligned to each other, as well as the electrical measurement traces patterned on the bottom die.



(a) IR microscope image of Vernier scales used for alignment measurements

(b) IR microscope image of assembled stack

Figure 2.10: Infra-red microscope images used for alignment measurements.

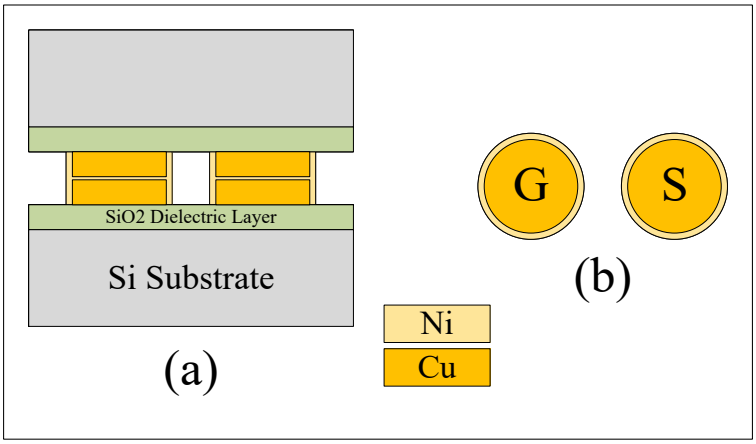


Figure 2.11: Ground-Signal (GS) interconnect configuration used for simulations.

2.6 Electrical Modeling

The various fabrication and scalability benefits of selective electroless plated interconnects have been elucidated in the previous sections. However it is important to benchmark the electrical performance of these interconnects, and quantitatively compare the electrical performance of these with respect to other interconnect technologies. To this end, in this section, the Resistance, Incuctance, Conductance, Capacitance (RLGC) parasitics of the electroless plated interconnects up to 30 GHz are extracted from HFSS simulations. This is compared with simulations of other commonly used interconnects including solder-capped copper microbumps, copper-to-copper direct bonds and hybrid bonds.

Figure 2.11 shows the Ground-Signal (GS) configuration and the geometric dimensions used for the simulation. The dimensions of the structures used for the simulations are captured in Table 2.3. These dimensions were chosen to replicate the fabrication results presented in the previous section. A two-port simulation was performed and the structure was excited with 50 Ω characteristic impedance lumped ports.

2.6.1 Frequency Dependent Parasitics Extraction

Extracting the frequency-dependent parasitics of these interconnect structures is important to quantify their signalling performance as discussed in the introduction. To this end,

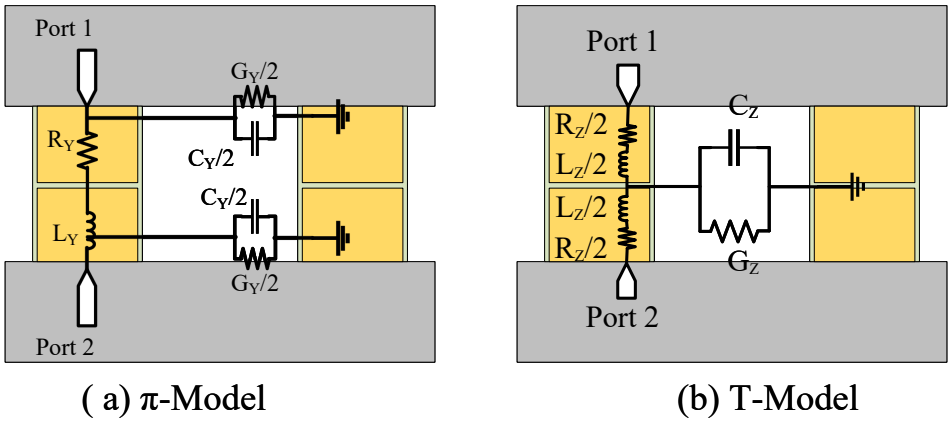


Figure 2.12: Equivalent lumped circuit models used for extracting parasitics from Y and Z parameters.

Table 2.3: Geometric dimensions used for the simulation

Dimension	**Value**
Pitch	50 μm
Cu pillar diameter	20 μm
Cu pillar height	2.5 μm
Deposited Ni thickness	1 μm
SiO_2 thickness	5 μm
Si thickness	500 μm

the lumped equivalent R, L, G, C components were extracted from the simulated Scattering Parameters (S) and the two-port lumped equivalent circuit models shown in Figure 2.12. As shown , using the π and T models, the lumped parasitics can be expressed in terms of the Admittance Parameters (Y) and Impedance Parameters (Z) respectively.

The following sets of equations were used to convert S parameters into Z and Y equivalents, and to extract the corresponding parasitics $R_Z L_Z G_Z C_Z$ and $R_Y L_Y G_Y C_Y$ respectively [68]:

$$Z_{11} = Z_0 \frac{(1+S_{11})(1-S_{22}) + S_{12}S_{21}}{(1-S_{11})(1-S_{22}) - S_{12}S_{21}} \tag{2.4}$$

$$Z_{12} = Z_0 \frac{2S_{12}}{(1-S_{11})(1-S_{22}) - S_{12}S_{21}} \tag{2.5}$$

$$Z_{21} = Z_0 \frac{2S_{21}}{(1-S_{11})(1-S_{22}) - S_{12}S_{21}} \tag{2.6}$$

$$Z_{22} = Z_0 \frac{(1-S_{11})(1+S_{22}) + S_{12}S_{21}}{(1-S_{11})(1-S_{22}) - S_{12}S_{21}} \tag{2.7}$$

$$Y_{11} = Y_0 \frac{(1-S_{11})(1+S_{22}) + S_{12}S_{21}}{(1+S_{11})(1+S_{22}) - S_{12}S_{21}} \tag{2.8}$$

$$Y_{12} = Y_0 \frac{-2S_{12}}{(1+S_{11})(1+S_{22}) - S_{12}S_{21}} \tag{2.9}$$

$$Y_{21} = Y_0 \frac{-2S_{21}}{(1+S_{11})(1+S_{22}) - S_{12}S_{21}} \tag{2.10}$$

$$Y_{22} = Y_0 \frac{(1+S_{11})(1-S_{22}) + S_{12}S_{21}}{(1+S_{11})(1+S_{22}) - S_{12}S_{21}} \tag{2.11}$$

$$R_Z = Re(Z_{11} + Z_{22} - (Z_{12} + Z_{21})) \tag{2.12}$$

$$L_Z = \frac{Im(Z_{11} + Z_{22} - (Z_{12} + Z_{21}))}{\omega} \tag{2.13}$$

$$G_Z = Re\left(\frac{2}{Z_{12} + Z_{21}}\right) \tag{2.14}$$

$$C_Z = \frac{Im\left(\frac{2}{Z_{12}+Z_{21}}\right)}{\omega} \tag{2.15}$$

$$R_Y = Re\left(\frac{-2}{Y_{12}+Y_{21}}\right) \tag{2.16}$$

$$L_Y = \frac{Im\left(\frac{-2}{Y_{12}+Y_{21}}\right)}{\omega} \tag{2.17}$$

$$G_Y = Re(Y_{11}+Y_{22}+Y_{12}+Y_{21}) \tag{2.18}$$

$$C_Y = \frac{Im(Y_{11}+Y_{22}+Y_{12}+Y_{21})}{\omega} \tag{2.19}$$

The extracted parsitics data from both approaches is shown in Figure 2.13. As can be seen, data obtained from both of the lumped models shows good agreement with each other. It is worth noting that these components are obtained considering the aforementioned substrate dimensions and materials, and changing the underlying substrate can affect the fringe fields passing through the substrate. Correspondingly, any change in the underlying substrate can have an effect on the overall parasitics even at the same interconnect dimensions. Further, no underfill material was considered between the interconnects for these simulations.

2.6.2 Effect of Pitch Scaling

Next, the effect of scaling down the interconnect pitch was evaluated. For this evaluation, the Cu pillar height was kept constant, but the pillar diameter and pillar-to-pillar pitch was scaled down. This approach was chosen so that all the cases considered would still fall within the vertical alignment accuracy constraints of the fabricated testbed, and therefore, could be fabricated without any major modifications. The interconnect pitch was scaled down to 10 μm, and the same pitch : diameter ratio of 2.5 : 1 was used for all cases.

As the results from the previous sections showed good agreement between both Z and

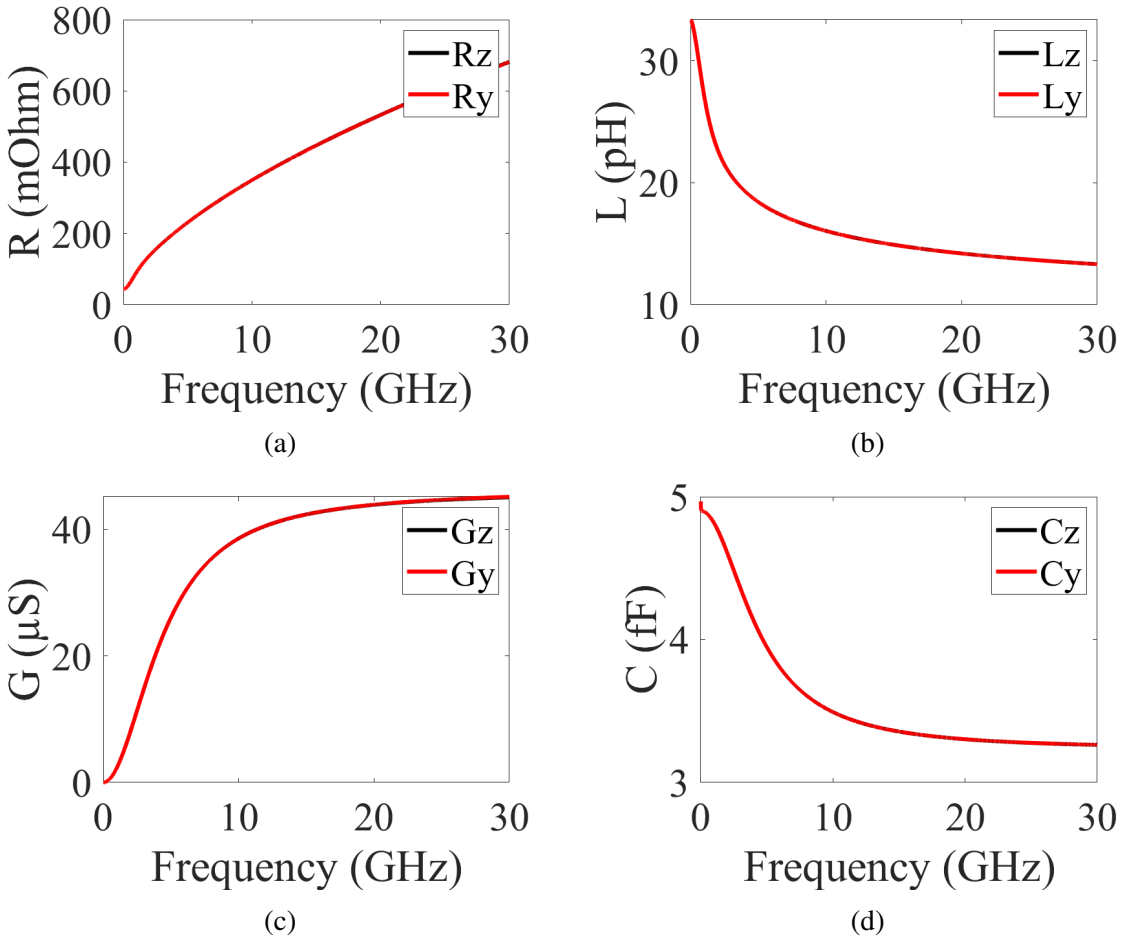
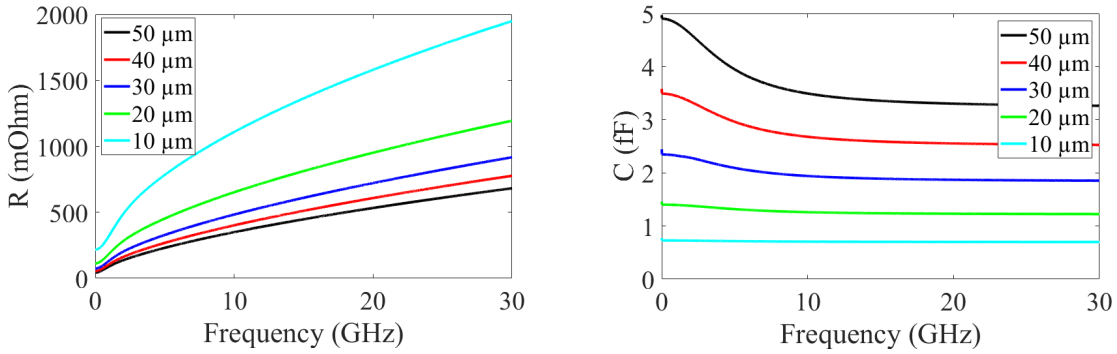


Figure 2.13: Extracted parasitics from the S parameters and equivalent circuit models. (a) Parasitic resistance, (b) Parasitic inductance, (c) Parasitic conductance, (d) Parasitic capacitance.

Y parameter based extraction methodologies, only Z parameter-based extraction was used for this analysis. The results are captured in Figure 2.14. The interconnect diameter shrinks as the pitch scales down, and correspondingly, we can see an increase in the parasitic resistance. However, the parasitic capacitance reduces with pitch scaling as shown.

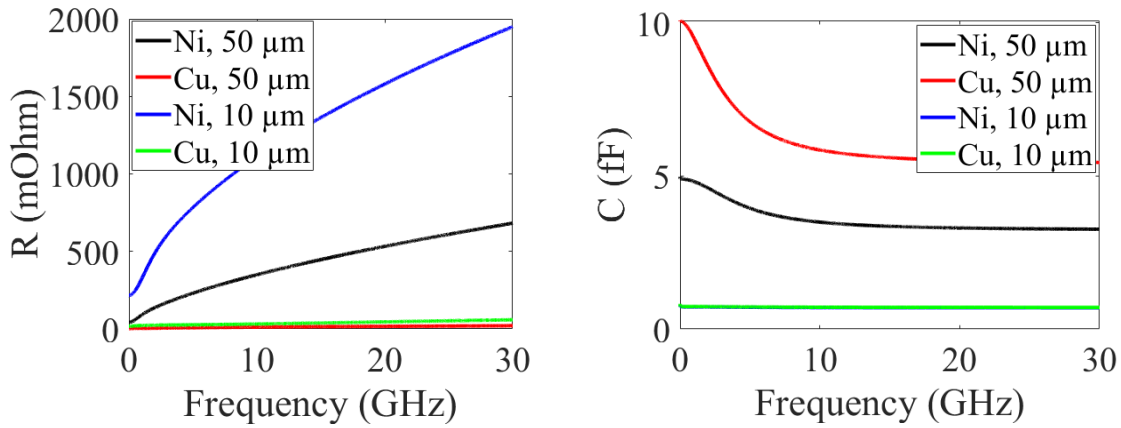


(a) (b)

Figure 2.14: Effect of interconnect pitch scaling on parasitics (a) Parasitic resistance, (b) Parasitic capacitance.

2.6.3 Comparison of Electroless Chemistries

While electroless nickel based chemistry was selected for this technology demonstrator due to the reasons discussed in the previous sections, Cu electroless chemistries could also be used by optimizing the process-flow. To quantify the potential benefits of such a process, the simulations were repeated by replacing Ni with Cu at both 50 μm and 10 μm pitches, and the corresponding results are shown in Figure 2.15.



(a) (b)

Figure 2.15: Effect of electroless plating chemistry on parasitics (a) Parasitic resistance, (b) Parasitic capacitance.

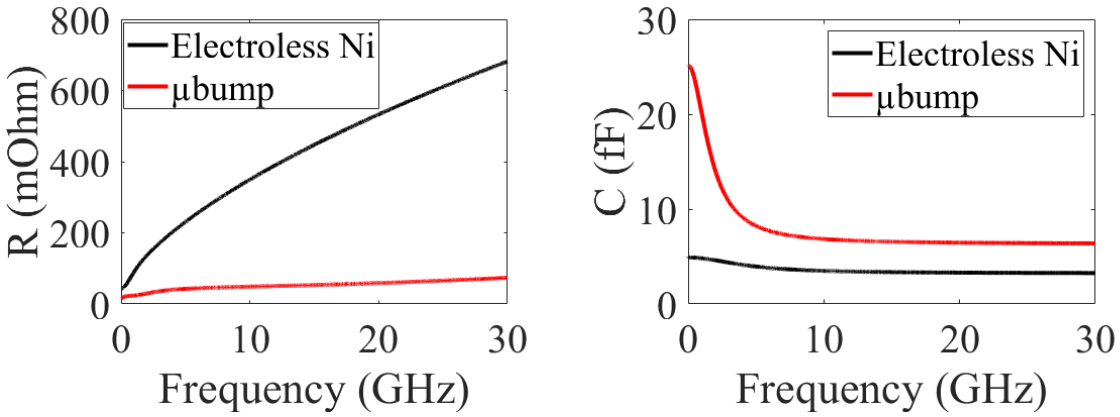


Figure 2.16: Comparison with solder-capped copper microbumps. (a) Parasitic resistance, (b) Parasitic capacitance.

2.6.4 Comparison with Conventional Interconnects

Finally, to compare the performance of electroless based interconnect to conventional approaches, a 50 μm pitch GS configuration of solder-capped copper microbumps were simulated. A copper pillar height of 10 μm, pillar diameter of 25 µm, and a re-flowed solder height of 5 μm were used for the simulation. The results from this analysis are presented in Figure 2.16.

The lower height of electroless interconnects helps considerably reduce the overall parasitic capacitance. However, as majority of the signal propagates through the highly resistive Ni layer due to skin effects at high frequencies, the overall parasitic resistance increases considerably despite the shorter electrical length. Switching to a more conductive material like Cu can help mitigate this, and extract the full benefits of the shorter interconnects provided by this approach as shown in Table 2.4.

Table 2.4: Extracted parasitic components for electroless I/Os with Ni and Cu

	R @10 GHz (mΩ)	**R @20 GHz** (mΩ)	**C @10 GHz** (fF)	**C @20 GHz** (fF)
Electroless Ni, 50 µm pitch	352.5	535.5	3.49	3.29
Electroless Cu, 50 µm pitch	11.87	16.71	5.84	5.5
Microbump, 50 µm pitch	48.68	59.05	6.85	6.46

2.7 Conclusion

This chapter demonstrates the use of a low temperature electroless nickel deposition process along with mechanical self-alignment for high density interconnection between chiplets. Void free and selective nickel deposition was observed across a 50 $\times$ 50 area array of 50 μm pitch copper pillars. Mechanical self-alignment structures were used to achieve alignment accuracies of less than 2 μm. The scalability of the approach was evaluated using RLGC parasitics extracted from high frequency simulations. These results demonstrate that electroless plating, in conjunction with self-alignment, can be used as a potential alternative to meet the rising interconnection demands of future 2.5D and 3D polylithic integrated devices.

CHAPTER 3

MONOLITHIC MICROFLUIDIC COOLING WITH INTEGRATED MICROPIN-FINS AND 3D PRINTED MANIFOLDS FOR EFFICIENT COOLING

3.1 Evaluation of Microfluidic Cooling on a Functional Testbed

As discussed in the introduction, the increasing socket power of compute systems puts substantial strain on traditional thermal management approaches such as air-cooled heat sinks and coldplates. Direct liquid cooling approaches, where the coolant is directly in contact with the backside of the active silicon have been extensively explored in literature as an attractive alternative [39, 69–72]. These techniques offers a considerable reduction in bulk and interface thermal resistances in the heat removal path. Combining these with the large surface area-to-volume ratio provided by microfabricated channels or micropin-fin structures on silicon enhances the ability to extract very large heat fluxes, while using only a fraction of coolant flow rates compared to conventional approaches. Concept demonstrations on passive silicon testbeds with heating elements [38, 73] as well as considerable numerical and finite element modeling simulation based evaluations [74–77] have been documented. This includes microchannel based devices [73, 78, 79], micropin-fin heat sinks [80–84], as well as other approaches such as jet impingement cooling [85–87]. The potential energy savings with microfluidic cooling for datacenter applications has also been extensively studied. However, while advanced silicon CMOS CPUs has been discussed as the prime use case for most of these studies, no demonstration exists on a comparable active device, with most functional demonstrations focusing on non-CMOS devices or devices with lower power and/or power densities [40, 88]. Further, compared to a few static power maps evaluated in many of these studies, state-of-the art general purpose CPUs can have a wide range of power dissipation profiles. These also vary based on the widely dif-

ferent classes of applications handled by them. Therefore, an evaluation with real-world devices and benchmarks is required to fully quantify the benefits of this technology.

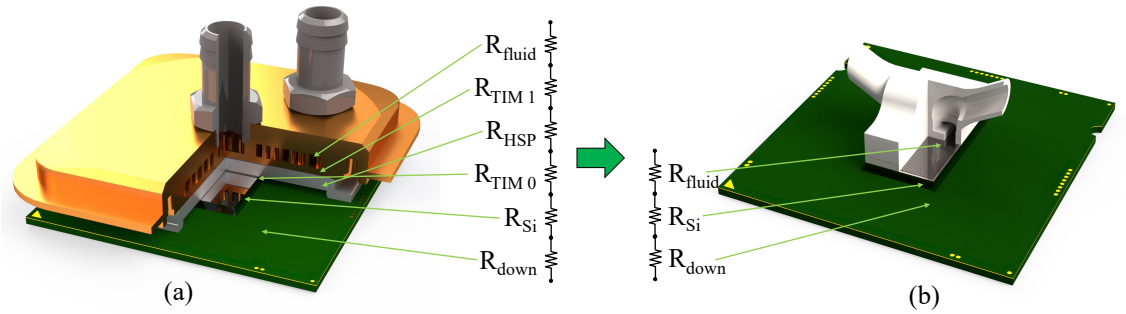


Figure 3.1: Traditional cold-plate based cooling (a) compared with microfluidic cooling (b). Switch to microfluidic heatsink represents considerable reduction in thermal resistances as well as form factor reductions.

To this end, in this chapter, we evaluate the impact of monolithic microfluidic cooling on the thermal, electrical, and compute performance of a high-power Intel CPU. A monolithic micropin-fin heatsink is etched directly on the backside of an off-the-shelf CPU, and is capped with 3D printed fluid delivery manifolds, as shown in Figure 3.1(b). The device is then tested while being overclocked to the highest stable frequency for two different workloads, while being cooled with de-ionized (DI) water. Evaluations of the impact of fluid flow rate as well as the feasibility of using elevated inlet temperatures to cool the device are studied. Comparisons with conventional cooling approaches are also presented. This work is: (1) the first study of its kind, implementing a microfluidic heatsink directly on the back of an off the shelf consumer CPU, (2) representing the highest power densities evaluated with microfluidic cooling on active CMOS devices, and (3) is the first to evaluate the impact of such cooling with multiple consumer benchmark programs. This data helps more accurately quantify the potential of microfluidic cooling in real-world use cases, when compared to existing literature.

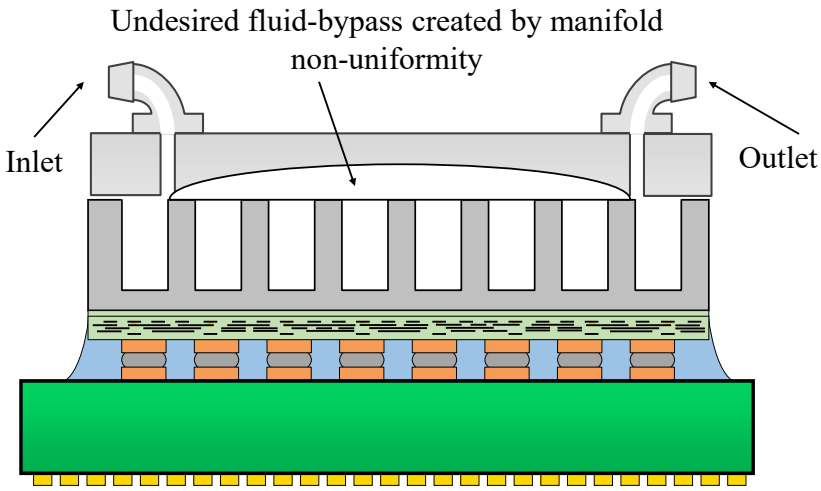


Figure 3.2: Conceptual diagram showing the 3D print non-planarity leading to flowbypass.

3.2 3D printed Fluidic Manifolds for Monolithic Microfluidic Cooling

As discussed, advances in microfabrication enables fabricatrion of high aspect ratio, and correspondingly, large area : volume ratio structures directly on the backside of the silicon. This helps create heatsinks with very low convective thermal resistance within a small volumetric footprint. A considerable reduction in overall size of the heatsink can be achieved by combining it with 3D printed fluid delivery manifolds. A conceptual diagram showing this is captured in Figure 3.1. The ability to accurately control the fluid exchange to the etched areas can also serve as an extra design parameter to improve the heatsink performance. In this section, the basic design considerations for the 3D printed manifolding structure is presented.

3.2.1 Material Selection

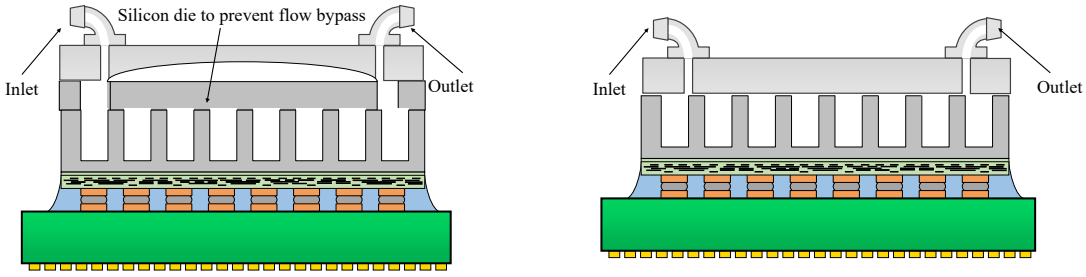
3D printing techniques and material selection choices for such a manifold are presented in Table 3.1. Among the different characteristics, the main vectors that we focused on in this work are (a) the ability to create high-detail features with acceptable amount of precision and (b) printing technique with a support material technology that allows fabrication of enclosed channels without any obstructions. The high precision in printing is required to

ensure an acceptable level of planarity in the manifold area, which seals the etched silicon micropin-fin region to prevent flow bypass. This potential flow-bypass effect is shown in Figure 3.2. In terms of the support material, the use of a Breakable Support Technology (BST) as opposed to a Soluble Support Technology (SST) may lead to the presence of constricting support pillars inside the enclosed fluidic channels, thereby increasing pressure drop within the fluid flow channels.

Table 3.1: Comparison of various relevant 3D printing techniques

3D Printing Process	Cost	Material	Support Material		Finest Dimension (μm)		Advantages
			Support required?	Type of support (solid, soluble)	Layer Height	Wall Thick-ness	
FDM (Fused Deposition Modeling)	Low [89], [90]	Thermoplastic filaments, elastomers, ABS (acrylonitrile butadiene styrene) [89]	Dependent on model geometry [91]	Soluble [91]	100 - 300 [90], [92]	800 [92]	low-cost prototyping [89], low manufacturing time [92]
Stereolithography (SLA)	Low [90]	Laser cured liquid photopolymer resin [93]	Required [91]	Solid [91]	25 - 100 [90], [92]	500 [92]	Smooth surface finish, high feature detail [93], [92]
SLS (Selective Laser Sintering)	Higher than FDM [92]	Plastic (often nylon), metals [93]	Not required [91]	N/A	100 [92]	700 [92]	Printing complex objects with fine details [93]
Material Jetting	Higher than FDM [90]	UV curable plastic, metals [93]	Required [91]	Soluble [91]	32 [90], [92]	1000 [92]	Highest dimensional accuracy, smooth surface finish, high feature detail [93], [92]

As evidenced by the trade-offs listed in Table 3.1, 3D printing technologies which provides highest dimensional accuracy, and hence the least fluid-bypass, typically uses low temperature, expensive polymers making them less ideal candidates for a heatsink enclosure. To this end, a hybrid enclosure, where a 3D printed manifold provides the fluid routing and connections, along with an etched silicon base added to provide good sealing to the top of micropin-fins is used in this demonstration. More details of this capping structure is provided in subsequent sections of this chapter. Conceptual images showing both these approaches are shown in Figure 3.3.



(a) (b)

Figure 3.3: Methods to reduce flow bypass (a) two-part manifold, (b) single part manifold with high accuracy printing.

3.3 Testbed Details

3.3.1 Intel Core i7 8700K Processor

For this demonstration, we use an off-the-shelf Intel Core-i7 8700K processor (Table Table 3.2). This particular processor was chosen for this demonstration due to: (1) The high device TDP, coupled with the small die size offers a very high power density to help test the device under extreme flux conditions, (2) the device is unlocked, allowing for easy overclocking, and (3) the Integrated Heat Spreader (IHS) is connected to the die using a thermal grease, as opposed to a solder Thermal Interface Material (TIM), which allows for easy de-lid and further processing on the backside of the silicon CPU. While the base TDP may not require advanced cooling approaches, the high power and power density values

Table 3.2: Specifications of the Intel Core i7-8700K processor used for this demonstration

Technology	14nm
Microarchitechture	Coffee Lake
No. of cores	6
Socket	FCLGA1151
Die Area	149 mm^2
Core Area	54 mm^2 *
Base TDP	95 W

* Estimated value based on open-source die-shots

that the device can support under overclocked operation makes it an ideal testbed for this evaluation.

Overclocking

The objective of this work is to evaluate the performance of the CPU when operating at the most compute intensive settings that it can handle as required by each kind of workload being evaluated. For this purpose, the CPU is run at a clock frequency higher than it's rated base. In real world applications, overclocking helps speed-up compute intensive workloads and can help improve user experience and even reduce the cloud provider's overall cost in a datacenter setting [94].

However, this comes with a power overhead. The overall power dissipation in an active CMOS device can be characterized by (Equation 3.1).

$$P_{total} = \underbrace{\alpha C V_{dd} \Delta V f_{clk}}_{Dynamic} + \underbrace{\alpha E_s f_{clk}}_{ShortCircuit} + \underbrace{V_{dd} I_{leakage}}_{Leakage} \tag{3.1}$$

where α is the activity factor, C is the total load capacitance, V_{dd} is the bias voltage, ΔV is the voltage swing during switching, f_{clk} is the clock frequency, E_s is the short circuit energy per switch operation and $I_{leakage}$ is the total leakage current. Sustained stable

overclocking typically requires increasing the bias voltage with increasing frequency for stable operation. Consequently, it follows that all three components of the power dissipation, including dynamic, short circuit and leakage powers increase for an overclocked device. Therefore, the typical ceiling for stable overclocked frequency is determined by the ability of the cooling system to handle the increased power dissipation, while maintaining the junction temperatures within safe operating limits. For this reason, characterizing the highest stable overclocked frequency is a direct indicator of the performance boost that can be achieved by using a more effective cooling system.

Table 3.3: Testbed Details

Motherboard	GIGABYTE Z390 AORUS ULTRA Motherboard
Memory	32GB (2 x 16GB) 288-Pin DDR4 4000
Power Supply Unit	850W Modular Power Supply
Operating System	Microsoft Windows 10

3.3.2 Baseline Data and Benchmarking Procedure

Details of the set-up used to evaluate the CPU performance is captured in Table Table 3.3. Two different open source benchmark programs, Great Internet Mersenne Prime Search (GIMPS) Prime95 and Cinebench R20, were used to evaluate the performance of the device under different operating conditions. Prime95 represents a higher per-core power dissipation at any given frequency compared to Cinebench R20. The latter however typically can run at a much higher clock-rate. So analyzing data from these distinct benchmarks helps identify the cooling solution's performance under some variants of the vastly different applications handled by modern day general purpose CPUs.

For each test case of the cooling set-up, the highest stable frequency point was identified. This was characterized by the operating conditions at which all the six cores, each with two threads, were able to operate without any stability issues (thread getting killed/ any errors being thrown by the application), as well as the core being able to maintain the

set frequency for the duration of the test without throttling down the clock. For Prime95, the total run-time was set at five minutes. This duration is considerably lower than what's considered necessary to ensure absolute system stability. But the experimental limitations of running an open-loop cooling set-up for longer duration, coupled with the fact that thermal measurements remain mostly stable for even longer runs of the application help achieve acceptable data points for a thermal evaluation with this benchmarks. For Cinebench R20, the application was run for the duration of one complete render test. It is worth noting that this methodology of testing would put all the highest stable frequency point of each test-case near the temperature limits of stable operation. This was chosen as the focus of the study was to obtain the limits of achievable performance with each cooling set-up rather than comparing operating temperature for each at the same workload.

Table 3.4: Intel Core i7-8700K and Core i9-9900K Processors

	i7-8700K	**i9-9900K**
Socket	FCLGA1151	FCLGA1151
No. of cores	6	8
Die Area	149 mm^2	180 mm^2
Core Area*	54 mm^2	92 mm^2
Base Clock	3.7GHz	3.6 GHz
Base TDP	95 W	95 W

* Estimated value based on open-source die-shots

For finding the highest stable frequency point at each test condition, starting from the base operating frequency, the CPU clock was manually changed from BIOS in increments of 100MHz, and it's stability evaluated for the corresponding application as described above. In case of a failure, the CPU core voltage was increased from BIOS in increments of 0.01 volts and the stability tests were re-run until the highest stable operating frequency across all cores was obtained.

Finally, to compare the results from the overclocking experiments, two distinct ap-

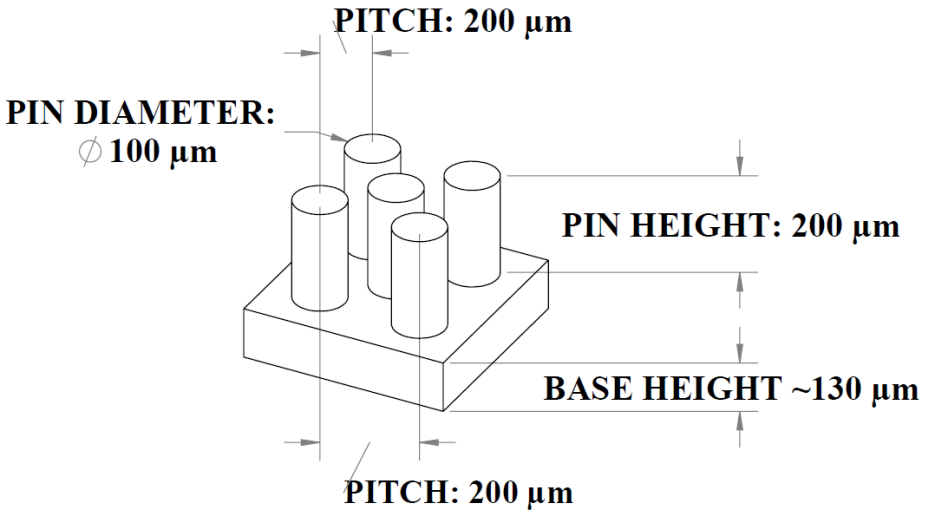


Figure 3.4: Geometric design of the micropin-fin heatsink used for this demonstration.

proaches were used. First, before de-lid, the device was tested with a conventional air-cooled heat-sink, operating at room temperature. An Intel XTS100H heatsink was used for this purpose, and it was mounted on the package using a thermal grease with a thermal conductivity value of 8.5 $W(mK)^{-1}$. Secondly, data from an Intel Core i9-9900K as presented in [95] was used for comparison. As shown in Table 3.4, the i9-9900K represents an analogous test-case with die size and power values similar to the i7-8700K used in this study.

3.3.3 Micropin-fin Heatsink

Various heatsink structures have been explored in the literature for monolithic heatsinks including microchannels, jet impingement cooling, hydrofoils, and in-line and staggered micropin-fins. For this demonstration, we chose a staggered micropin-fin heatsink due to it's high thermal performance, scalability of fabrication process, as well as the ability to fine tune thermal performance by easily changing the geometric design [96, 97].

The dimensions of the micropin-fins are shown in Figure 3.4. The micropin-fin height was fixed subject to the die-height of this off-the-shelf CPU. In an ideal setting, this could be tuned for better thermal performance by varying the die height. The lateral dimensions

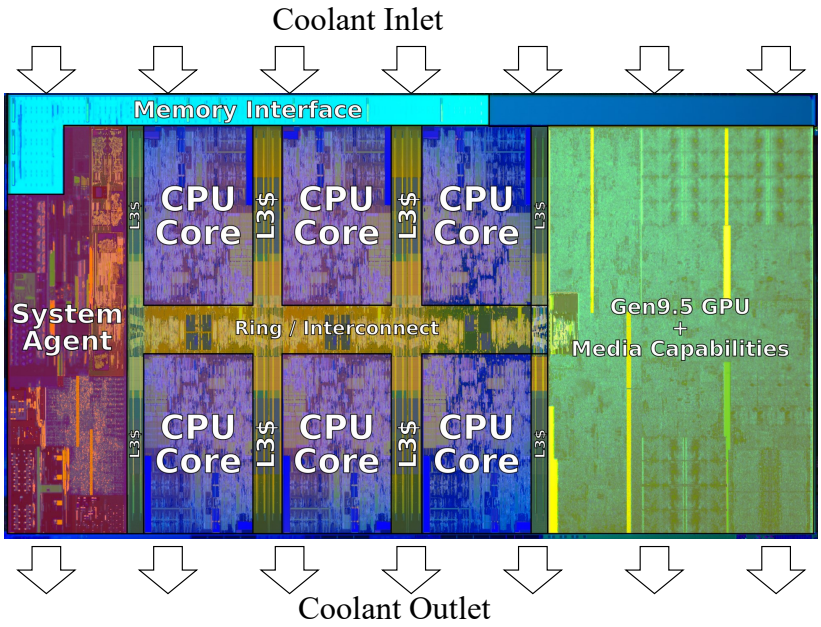


Figure 3.5: Direction of coolant flow shown superimposed on the i7-8700K die-shot from [98].

were selected to give acceptable thermal performance (in terms of junction-to-inlet thermal resistance as well as pressure drop) while staying within the aspect ratio limits of Bosch DRIE process to etch micropin-fins without large amounts of taper, which may reduce the thermal performance.

Further, it is worth noting that the selected micropin-fin design does not represent the performance limits of this technology, but serves as a technology demonstrator to study the benefits of monolithic microfluidic cooling on a high power, commercial off-the shelf device. Further improvements including optimizing the micropin-fin dimensions, improved design for fluidic manifolds etc [97, 99].

3.4 Heatsink Fabrication and Assembly

The monolithic heatsink used for this demonstration is made of two distinct components: (1) the micopin-fin heatsink, etched directly on the back of the dice, and (2) a capping and fluid delivery manifold, which encapsulates and delivers coolant to the etched area. The direction of fluid-flow was chosen to be parallel to the shorter edge, as shown

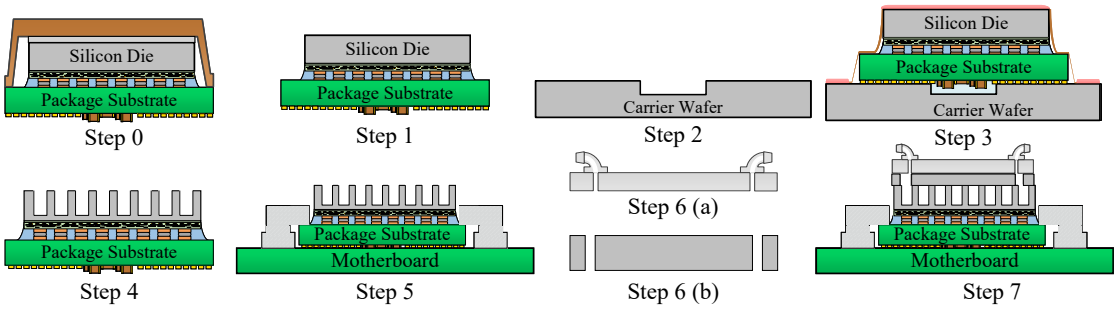


Figure 3.6: Fabrication and assembly process flow used for the monolithic microfluidic cooling. **Step 0:** Off-the-shelf processor package. **Step 1:**Remove the heatspreader and TIM. **Step 2:** Carrier wafer with a cavity that corresponds to SMD capacitor's profile prepared by Bosch etching of a silicon wafer. **Step 3:** Mount to the carrier wafer and spin coat photoresist. **Step 4:** Etch micropin-fins and remove from carrier wafer. **Step 5:** Mount the etched device into the motherboard socket. **Step 6(a):** 3D print the fluidic manifold. **Step 6(b):** Etch ports into a silicon wafer to create a capping layer. **Step 7:** Attach the silicon cap and the 3D printed manifold using epoxy.

in Figure 3.5. This was chosen to minimize the apparent increase in thermal resistance for any downstream components due to caloric heating [40].

Detailed fabrication flow used to create the micropin-fin heat sink on the pre-packaged processor is shown in Figure 3.6. After benchmarking the device with an air-cooled heatsink as described in earlier sections, the device was de-lidded to remove the IHS. The thermal grease between the device and IHS was removed using acetone wipes, followed by an acetone/isopropanol spray cleaning. A custom carrier wafer was prepared with cavities to accommodate for the backside Surface Mount Device (SMD) components. The package was then mounted on the carrier with a thermally conductive paste (cool grease 7016) layer in-between. This helps reduce the device heating during the subsequent etch steps. The exposed organic substrate layer is then covered with kapton tape, which protects it from plasma damage. Micropin-fins patterns are created on the backside of the die using an NR5-8000 photoresist mask, which is then used in a Bosch DRIE process to etch the micropin-fins to the desired depth. The mean etch depth was measured to be 223.53 μm using data collected from across the die using an Olympus LEXT optical profilometer. SEM images of the etched CPU, as shown in Figure 3.7, points to a good etch profile with

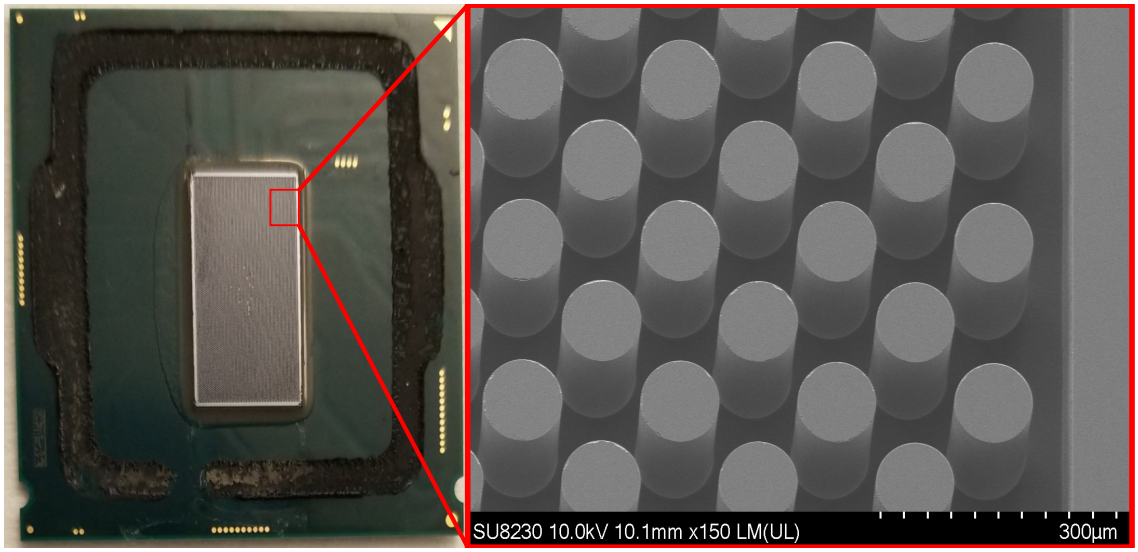


Figure 3.7: Photograph of the etched package with the insert showing an SEM of the micropin-fins.

minimal taper.

The encapsulation structure in itself was designed to have two parts: (1) a fluid delivery and capping manifold, designed using Solidworks 2020 and 3D printed with PA2200 polymer using a selective laser sintering process, and (2) a silicon capping structure, diced to the exact dimensions of the CPU die, and having DRIE etched fluid delivery ports. The two part structure for the capping layer leverages the excellent dimensional accuracy of silicon microfabrication techniques to help create a good seal above the micropin-fins without requiring expensive 3D printing techniques and materials with high printing accuracy. Figure 3.8 shows both these components. These are assembled on the etched die using a waterproof epoxy after mounting the CPU on the motherboard. Finally, the fluid delivery tubing is connected using the barbed ports of the 3D printed manifold. An image of the device with the heatsink assembled is shown in Figure 3.9.

3.5 Testing Set-up

The CPU was tested in an open loop configuration, as shown in Figure 3.10. DI water was used as the coolant for this evaluation. Alternates like dielectric coolants can also

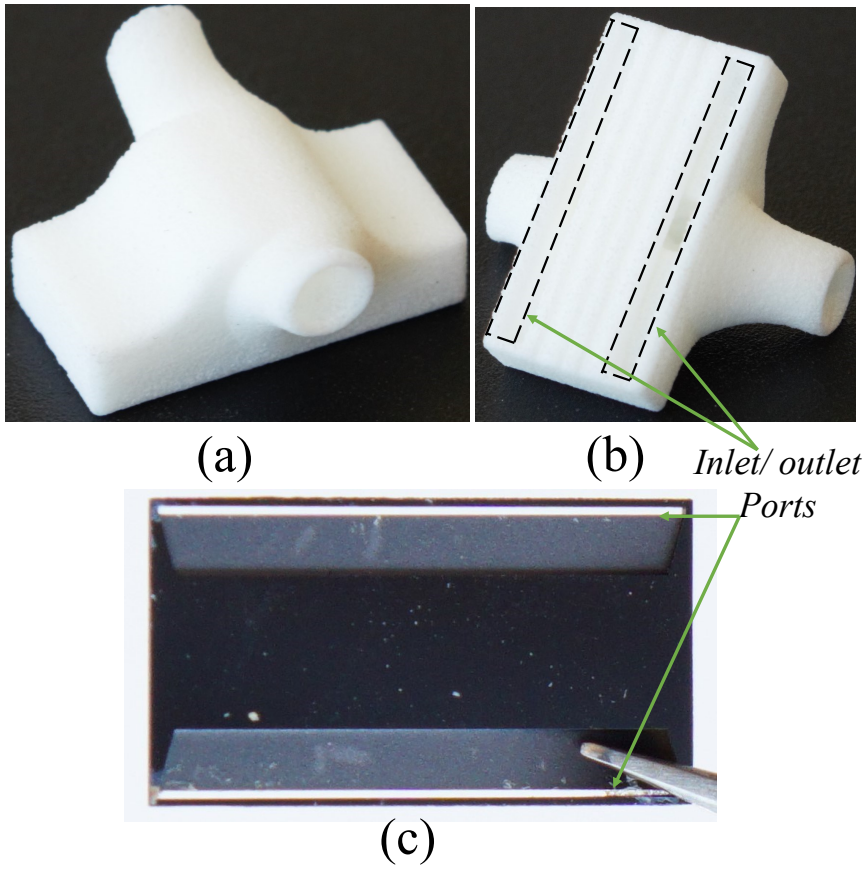


Figure 3.8: (a) 3D printed manifold top side (b) 3D printed manifold bottom side (c) etched Si cap.

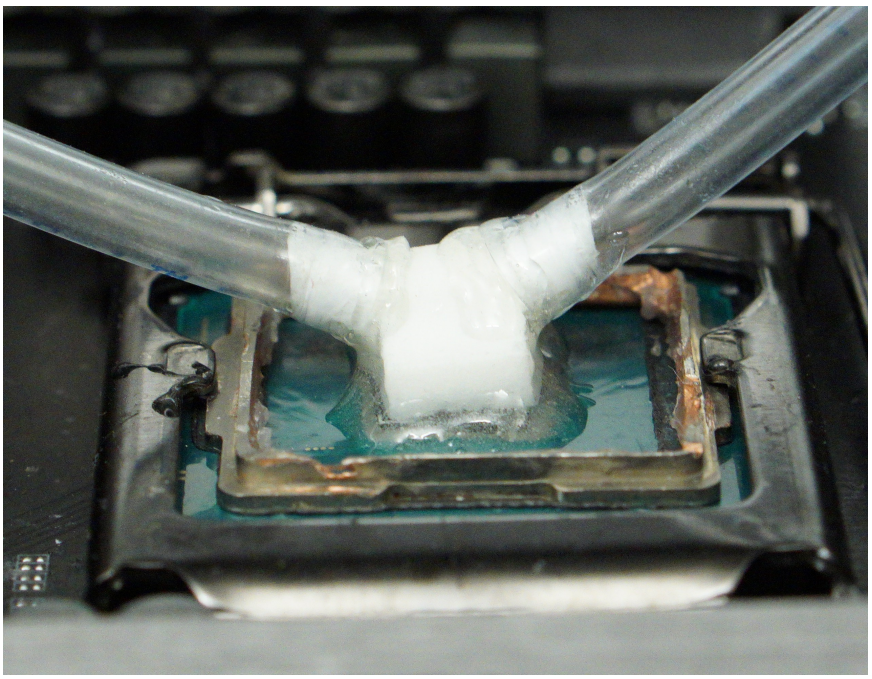


Figure 3.9: Device after attaching of manifold and fluid delivery tubes connected.

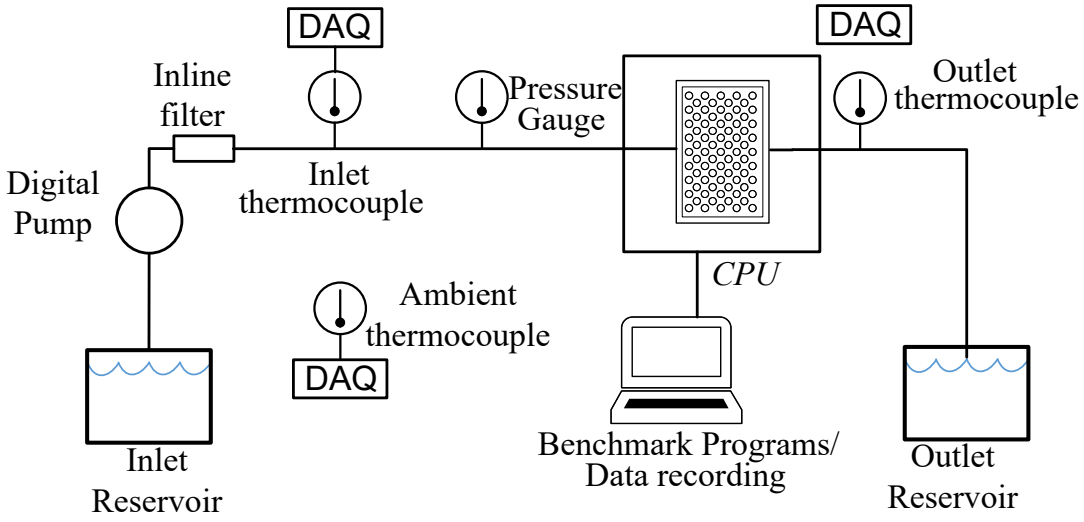


Figure 3.10: Open-loop measurement set-up. DI water is used as the coolant.

be used instead of DI water in an operational setting. However, as long term operational reliability was not a primary concern for this technology demonstration, these were not considered. Inlet, outlet and ambient temperatures were recorded using data from four J type thermocouples (one each for inlet and outlet, two for ambient, placed along different points beside the motherboard). The thermocouple outputs were read out using an Agilent 34972A LXI Data Acquisition system. The temperature values were recorded for the entire duration of the experiment at 250 ms intervals. The mean value of these recordings for the duration of the run time of corresponding benchmarks will be presented in the subsequent results section of this article. A differential pressure gauge was used to measure the coolant's pressure drop at the heatsink inlet with respect to atmosphere. The coolant flow rate was controlled using the rotations of the digital gear pump and corresponding flow rate was calibrated by mapping the gear's Rotations Per Minute (RPM) with the volumetric output for a known time. A second order polynomial fit was used to map the gear rotations to corresponding flow rates. Four different pump RPMs, spanning the range of flow rates used for the study, with two distinct data points collected at each RPM were used to create this mapping. The calibration procedure was repeated for all the fluid temperatures used in

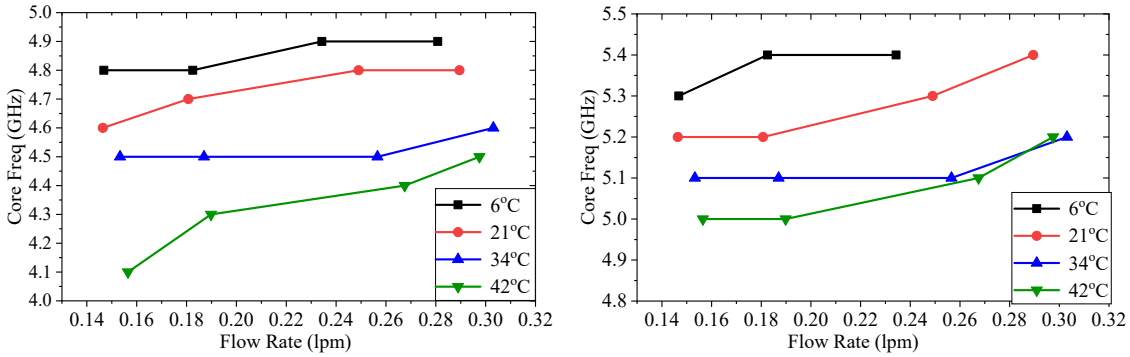
this study.

Data from the motherboard and the CPU's on chip sensors including the individual core clock rates, voltage and temperature read-outs, as well as the overall package and core power dissipation values were recorded using HWInfo64 program, with data sampled at a resolution of 250 ms. A similar approach of obtaining the arithmetic mean of relevant parameters over the runtime of the benchmarks was used to create the data plots. An open build was used for this benchmarking with natural convection being used to cool the majority of the motherboard. A fan was used to blow air over the Voltage Regulator Module (VRM) on the motherboard and VRM temperatures were monitored from the Operating System (OS) to ensure stable operation.

3.6 Results

As discussed, the maximum power that can be dissipated, and consequently, the maximum compute performance of the CPU is limited by the heat sink's ability to transfer this power to the coolant without increasing the device's junction temperature above the safe operating limits. This can be quantified as the thermal resistance of the heatsink. The overall thermal resistance associated with the monolithic micropin-fin heatsink can be described as the summation of four distinct components: spreading resistance, bulk resistance, convective resistance, and caloric resistance. While the first three components are determined by the geometry and material properties of the heatsink, the caloric resistance component has a strong relation to the coolant flow rate. To characterize this effect, the device was tested with varying flow rates. Further, to evaluate the heatsink performance under different inlet temperatures, these experiments were repeated for different inlet temperatures varying from 6 °C to 42 °C. The test was repeated with both benchmarks, and the data points corresponding to the highest stable frequency point is reported in the subsequent sections.

While the exact power dissipation of each core is near impossible to estimate without



(a) Prime95 (b) Cinebench R20

Figure 3.11: Highest stable frequency points for both benchmarks under various cooling conditions. Legend shows the coolant inlet temperatures. An increase in computational throughput, as signified by the highest stable frequency point, can be obtained by either reducing the inlet temperature, increasing the flow rate, or both depending on the requirements.

exact knowledge of the manufacturer's proprietary device micro-architecture and layout details, inference based on multiple experiments and reported data predicts mostly uniform temperatures across the different cores. However, the measurements with the monolithic micropin-fin heat sink showed one of six cores of the CPU to be consistently at a higher temperature compared to others. This effect is particularly pronounced for Prime95. This is potentially caused by a visual defect in the micropin-fin region directly above this core introduced during the assemble process, increasing the localized thermal resistance. As a result, this core gets thermally throttled much earlier than the other five and considerably skews the overall results. To minimize the impact of this, and to present a more accurate picture of the device performance, two different notations will be used for the temperature values: (1) average core temperature computed as the mean of temperatures across the cores as measured using HWInfo64 while the benchmark is active, and (2) maximum core temperature, which represents the instantaneous maximum value of temperatures across the cores as measured using HWInfo64, while the benchmark is active.

The maximum overclocked frequency point for each benchmark is also used as an indicator for improvement in compute throughput. This data, for both benchmarks, is captured

in Figure 3.11. It shows the high performance of monolithic microfluidic cooling even at an elevated inlet temperature of 42°C. Stable operation was observed for up to 5.2 GHz for Cinebench R20 and 4.5 GHz for Prime95, compared to the rated base clock of 3.7 GHz for the processor. Furthermore, by lowering the inlet temperature, higher compute potential can be unlocked even under lower coolant flow rates. However, this comes with the extra overhead of requiring power intensive heat exchange and/or refrigeration loops for lowering the inlet temperature.

Flow rates between 2.4 mL/s and 5.1 mL/s were tested for all the inlet temperatures. For all inlet temperatures, the highest stable clock frequency, and correspondingly, the highest sustained power dissipation increases with increasing flow rate for both benchmarks. These results are shown in Figure 3.12. Similarly, the increase in the peak throughput by reduced temperature coolant inlet can be seen in Figure 3.14. These two figures also show the reduction in device power dissipation with better cooling when operating at the same frequency.

The pressure drop of the monolithic heatsink is also characterized for all the fluid temperatures and flow rates and is presented in Figure 3.13. This data is recorded with the device turned off to prevent any temperature fluctuations and corresponding change in the recordings. The reduced viscosity of water at higher temperatures leads to a decrease in pressure drop when using elevated inlet temperatures. This can lead to considerable reduction in the required pumping power. Combining this with the excellent thermal performance even at elevated inlets, more efficient cooling configurations with minimal overheads can be unlocked. Further, it's worth noting that micropin-fin and manifold design optimizations can be used to further reduce the pressure drop and correspondingly, the overall pumping power without compromising on thermal performance.

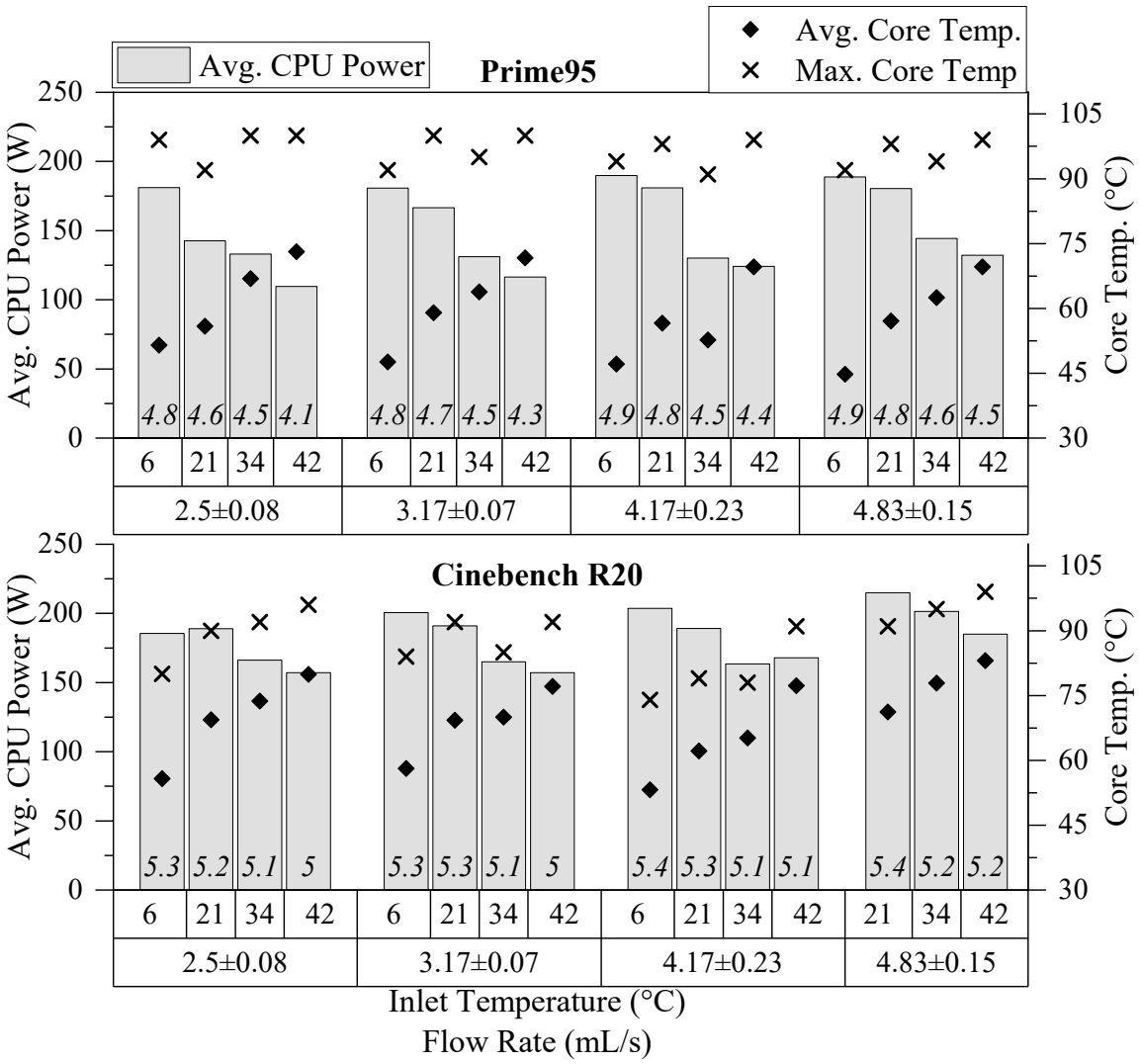


Figure 3.12: Variation of highest sustained power dissipation for all test-cases with the coolant flow rate. The corresponding frequency point is shown inside each bar. An increase in the highest sustained power can be observed at higher flow rates while maintaining similar core temperatures.

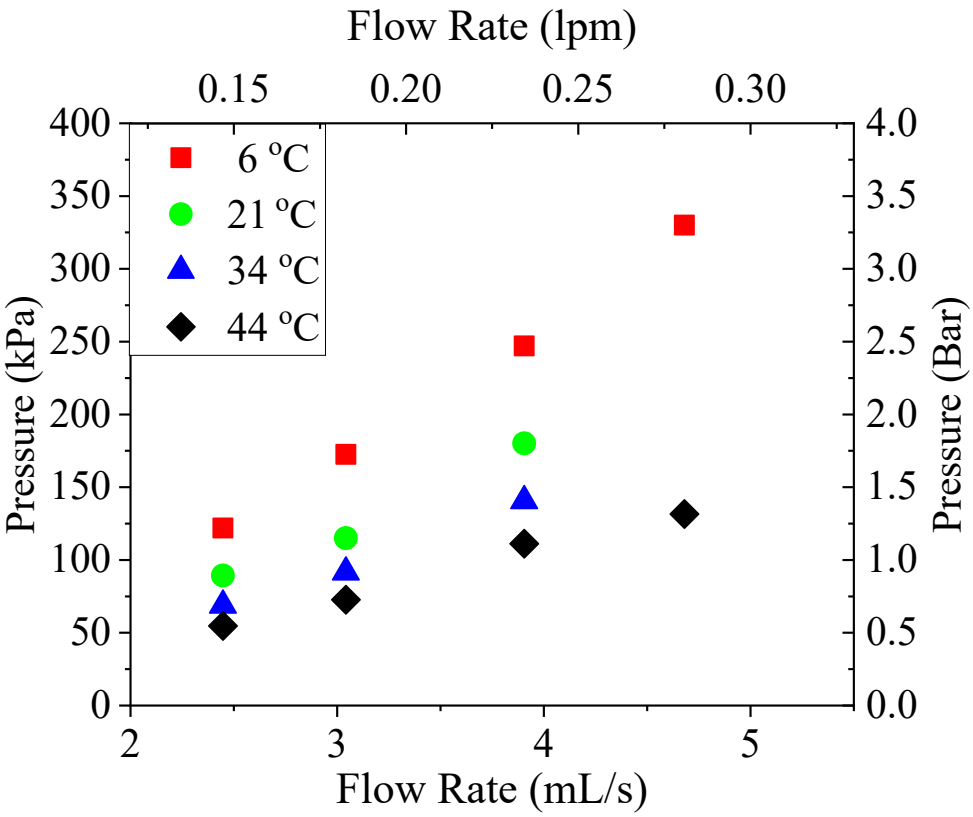


Figure 3.13: Pressure drop versus flow rate. The reduction in pressure drop at higher temperatures can be clearly observed, which translates to a reduction in the pumping power required to sustain the same flow rate at elevated temperatures.

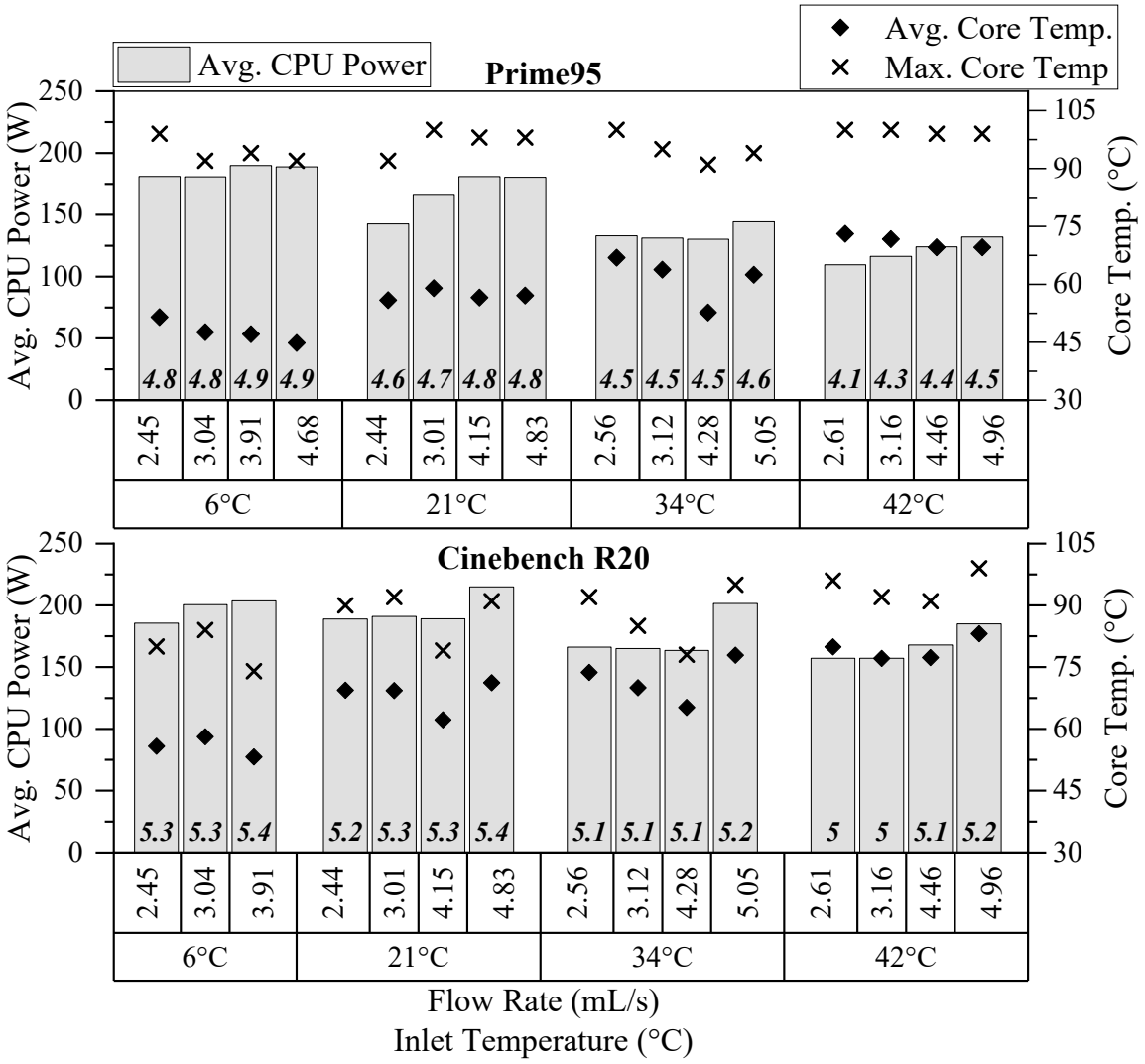
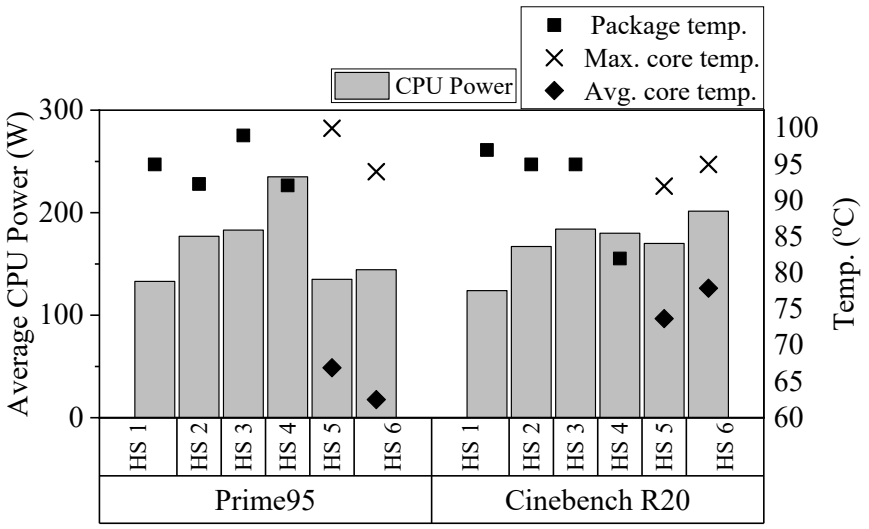
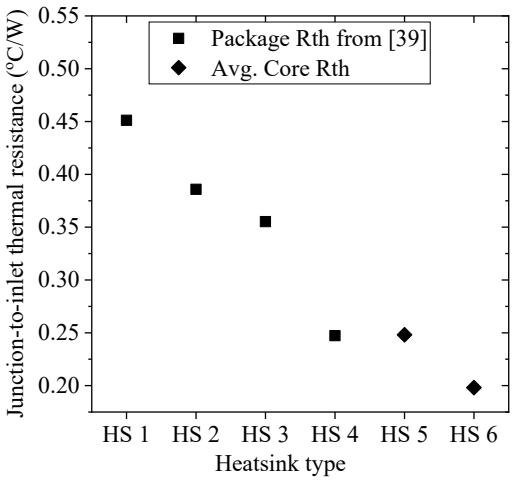


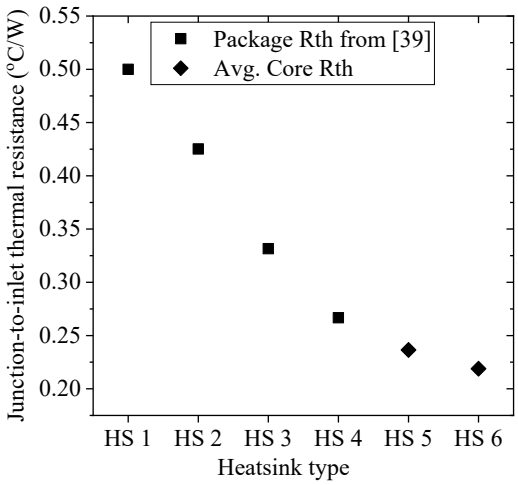
Figure 3.14: Variation of highest sustained power dissipation for all test-cases with the inlet temperature. The corresponding frequency point is shown inside each bar. Highest stable point drops slightly for elevated inlet temperatures. This effect is more prominent for a higher flux application such as Prime95. This performance penalty however helps reduce the cooling system power consumption.



(a) Peak sustained power dissipation



(b) Junction-to-inlet thermal resistance for Prime95



(c) Junction-to-inlet thermal resistance for Cinebench R20

Figure 3.15: Comparison of monolithic heatsink performance with other approaches. **HS 1:** Air cooled heatsink at 34°C on an i9-9900K, **HS 2:** Air cooled heatsink at 21°C on an i7-8700K, **HS 3:** Cold-plate at 34°C on an i9-9900K, **HS 4:**Two-phase immersion cooling at 34°C on an i9-9900K, **HS 5:** Monolithic microfluidic cooling at 34°C on an i7-8700K, at 0.1533 lpm flow rate, **HS 6:** Monolithic microfluidic cooling at 34°C on an i7-8700K, at 0.3031 lpm flow rate. The similar die-sizes of these chips makes a first-order thermal comparison feasible, but compute throughput cannot be directly compared due to differences in functionalities between the chips.

3.6.1 Performance Comparison with Other Heatsinks

The performance of the monolithic microfluidic cooling set-up was compared to other conventional and upcoming CPU cooling techniques. Measurements using room temperature air-cooled heatsink on the i7-8700K and data reported by Ramakrishnan *et.al.* in [95] for various state-of-the-art cooling techniques on an i9-9900K was used for this comparison. As discussed in Section subsection 3.3.2, thermal performance comparison across these chips can be done owing to the similar power values and die area. However, the differences in underlying functionalities do result in differences in local heat-flux values, and prevents an exact one-to-one comparison between the i7-8700K and i9-9900K. This approach however helps evaluate different cooling approaches more accurately than comparison with thermal resistance numbers reported in literature for non-functional testbeds. This is also evident when comparing the thermal resistance values of both applications under the same cooling conditions. When running different workloads, due to the differences in powermaps and the activity factor across different areas of the chip, thermal resistance values even under same cooling conditions would be different for different applications.

As can be seen from Figure 3.15a, monolithic heatsink with an elevated inlet temperature can sustain comparable, or even higher power dissipation than other cooling solutions due to it's low thermal resistance. This is more apparent from the Cinebench R20 data as the slight fabrication issues in this demonstration test-bed, as discussed earlier, prevented pushing the device to it's performance ceiling when running Prime95. This superior thermal performance is more evident for both benchmarks from the average junction-to-inlet thermal resistance data shown in Figure 3.15b and Figure 3.15c. We demonstrate up-to 44.4% reduction in junction-to-inlet thermal resistance when compared to a conventional cold-plate operating with the same inlet temperature of 34°C. Further, this reduction is achieved while pumping only 30.3% of DI water when compared to the cold-plate, representing a 69.7% reduction in coolant flow rate per kW of chip power. Furthermore, while the pressure drops associated with the microfluidic heatsink is higher than traditional

cold-plate based approaches, the reduction in the volumetric flow rate reduces the required pumping power, and improves operational efficiency.

3.7 Datacenter Efficiency Metrics and Microfluidic Cooling

With the exponential increase in cloud based computing services, datacenters consumes an ever increasing portion of energy and water consumption[100] in many countries. Therefore, an improvement in the power and water utilization efficiency of datacenters is required to minimize it's increasingly significant environmental impact.

To this end, microfluidic cooling can help improve the operational efficiency of datacenters across multiple commonly adopted metrics. As discussed earlier, the reduced operational temperature due to the low thermal resistance of this approach helps reduce the overall power consumption of the CPU while delivering the same performance. This corresponds an improvement in the computational efficiency of the CPUs (represented as number of compute operations per watt of power). Further, the reduction in required coolant flow-rate coupled with the ability to cool efficiently with elevated temperature inlets helps reduce the energy and water consumption overheads of the cooling system. These represent improvements in the power Utilization Effectiveness (PUE) (defined as the ratio of total power to the Information Technology (IT) power), as well as the Water Utilization Effectiveness (WUE) (defined as the ratio of water usage to the power consumed by IT equipment, typically in lpm/kW) of the datacenter. Further, the reduction in required components like bulky heat-sinks, TIMs etc, coupled with the smaller form-factor, also translates to a considerable reduction in the overall embodied carbon. Along with the power and water savings, this can help reduce the carbon footprint and environmental impacts associated with datacenter cooling.

3.8 Conclusion

In this chapter, we presented the first demonstration of monolithic microfluidic cooling on a functional CPU running real-world benchmarks. Thermal benchmarks showed junction-to-inlet thermal resistance values as low as 0.197°C/W for a 34°C inlet. Further, stable operation while dissipating up-to 215W of power within the 1.5 cm^2 area of the chip was shown. This represents a power density in excess of 200 W/cm^2 for the hotspot areas in the core region, showing the efficiency of the system to handle extremely high heat fluxes. This, to the best of our knowledge, is the highest heat flux demonstrated with microfluidic cooling on a functional CMOS device running real-world benchmarks. These results also offer more insights into the application based variability in the thermal performance at these very high power levels. The technique also uses considerably lower coolant flow compared to traditional coldplates, and is efficient even at elevated coolant inlet temperatures, both of which could lead to considerable energy and water savings for datacenter operations.

This initial demonstration shows considerable potential both from a performance and energy standpoint. Future work can further expand on this by implementing micropin-fin and fluid delivery manifold design optimizations. This work also serves as a starting point for evaluations of monolithic microfluidic cooling for 2.5D and 3D CPUs and other power dense SoCs.

CHAPTER 4

MONOLITHIC MICROFLUIDIC COOLING FOR A 2.5D FUNCTIONAL FPGA

Many modern high performance compute devices have multiple smaller dice stitched together with high bandwidth on-package interconnects to form heterogeneous 2.5D ICs with monolithic-like performance [6, 101]. This approach provides a broader design freedom to mix and match chiplets from different foundries and technology nodes while reducing the overall cost due to the higher fabrication yield of smaller chiplets [6].

Typically, in 2.5D ICs, placing the chiplets in close proximity is beneficial to provide monolithic-like link bandwidth, Energy per Bit (EPB) etc. This, however, results in higher aggregate power dissipation within a smaller footprint and puts higher strain on the cooling solutions [102]. Further, the heterogeneity of chiplets makes keeping each individual chiplet in it's ideal thermal operating regime challenging, leading to potential performance penalties. For example, the higher operational temperature due to coupling from an adjacent die can lead to increased power consumption in CPU chiplets [32], or increased adaptive refresh rates and correspondingly, lower bandwidths in HBM stacks [103].

Traditional solutions such as air-cooled heatsinks and cold-plates are limited in terms of scalability and often require bulky solutions for higher power use-cases. Furthermore, the large shared heat-spreader can further increase thermal coupling between chiplets with different power dissipation profiles. An ideal cooling solution for 2.5D ICs should rectify these and can also scale with it's power and size increases.

Monolithic microfluidic cooling, with it's very low thermal resistance and ease of customization, has been proposed as an alternative [69, 70]. A conceptual diagram showing the reduction in resistance path and coupling paths is shown is Figure 4.1. This includes etching heat-exchange surface such as microchannels or micropin-fins directly on the backside of the dice [72, 83] or other approaches such as multi-jet impingement direct cooling.

Bringing the coolant directly to the source of heat generation helps remove the thermal resistances associated with intermediate layers. Further, the ability to tune cooling efficacy at the chiplet level, as opposed to package level, helps reduce thermal coupling [104].

In this chapter, we further extend the the evaluation of the performance benefits of monolithic microfluidic cooling presented in the previous chapter to 2.5D devices. We explore the performance benefits of monolithic microfluidic cooling using finite volume modeling of a multi-chip CPU package. Design trade-offs between electrical and thermal performance using different cooling solutions are compared. Finally, we discuss the experimental implementation on a 2.5D FPGA, and the performance data is compared with a stock cold-plate solution.

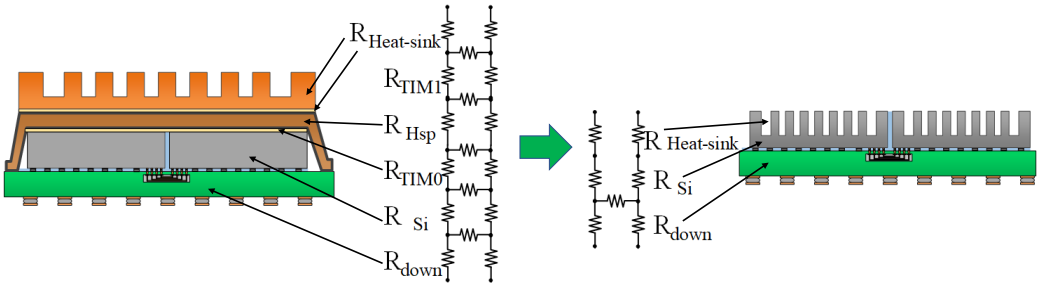


Figure 4.1: Conceptual diagram showing the reduction in absolute thermal resistances and thermal coupling paths with monolithic microfluidic cooling.

4.1 Thermal Challenges in 2.5D ICs

As discussed in the introduction, monolithic microfluidic cooling holds promise for performance improvement in 2.5D ICs. Therefore, it is beneficial to conduct early analyses to estimate the impact of dense integration of active chiplets in a single package. In this section, we describe the thermal analysis of a hypothetical multi-chiplet CPU [105].

4.1.1 Design Setup

For our analyses, we considered a hypothetical four chiplet system-in-package (SiP) based on the AMD EPYC 'Naples' CPU [107]. The considered model geometry and the

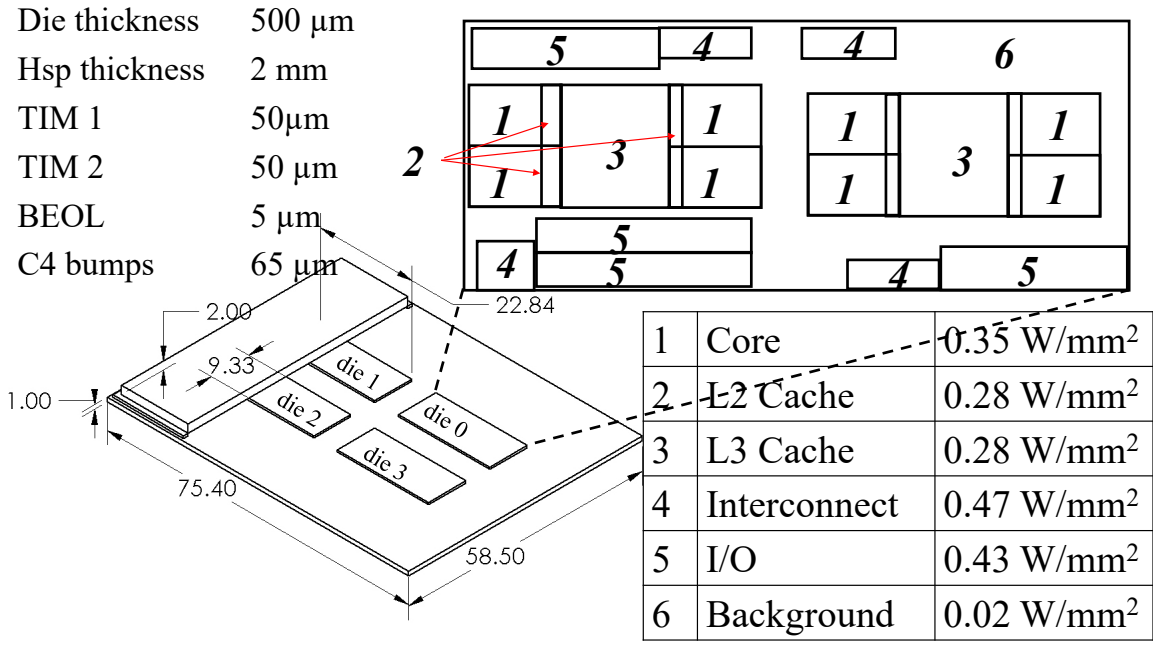


Figure 4.2: Modeled 4 chiplet SiP along with SoC power density assumptions. A heat transfer coefficient of 2×10^3 W/m^2 ^{o}C was used for the air-cooled heatsink and 3.3×10^5 W/m^2 ^{o}C for microfluidic cooling.

block-level power assumptions are shown in Figure 4.2 and summarized in Table 4.1. The block-level dimensions were estimated using the 'Zeppelin' SoC floor plan [106]. The block-level powers were estimated using the TDP [108] and the server power breakdown available in [11]. The power for 'clocks' was assigned as a background power to the SoC, as shown in Figure 4.2. With four dice in the system, the total system power with all four SoCs active is assumed as 200 W.

The simulation model conditions are summarized in Figure 4.2. The package and heat spreader sizes were obtained from [109]. We assume an organic package substrate. C4 bumps with 130 μm pitch and a 65 μm diameter were assumed as the bonding interface between the die and package. We assumed both air and monolithic microfluidic cooling for our analyses, and the assumed ambient temperature was 38 ^{o}C.

Table 4.1: 'Zeppelin' SoC power and area assumptions [11, 106]

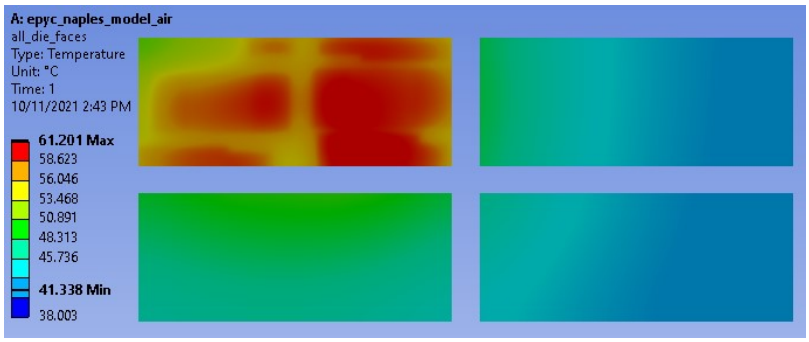
Functionality	**Total Power (W)**	**Total Area (mm^2)**
Computation	16.5	47
I/O	15	34.6
CPU Caches (L2+L3)	11	39.6
Interconnect	6	12.8
Clocks	1.5	79.1
Total	50	213.1

4.1.2 Steady State Thermal Analysis

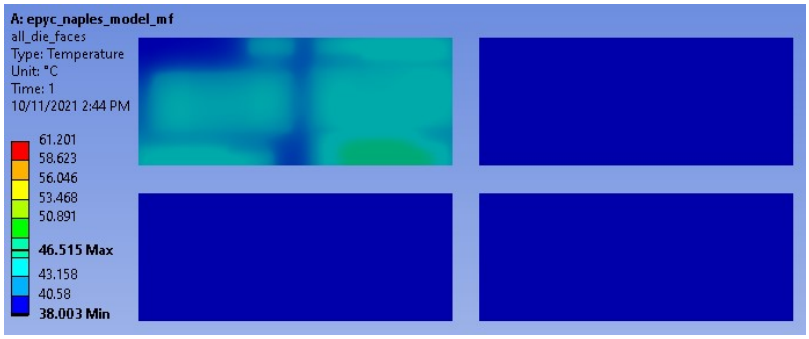
We modeled the 2.5D CPU using the finite-element ANSYS Mechanical APDL solver (ver. 2021 R1). For a die-to-die lateral spacing (both x and y dimensions) of 2 mm, the maximum junction temperatures were observed as 86.94 ^{o}C (air) and 46.91 ^{o}C (microfluidic). Next, the results from thermal coupling and impact of die spacing analyses are discussed as follows:

Die-to-Die Thermal Coupling

Quantifying thermal coupling can either improve product performance or save energy either through thermal-aware design-time optimization or run-time workload scheduling. Here we discuss one such method to quantify steady state die-to-die thermal coupling. *die0* was activated using the power map from Figure 4.2 with no power dissipation in the other chiplets. We define thermal coupling as the ratio of victim chiplet (die1, die2, die3) junction-to-ambient temperature to aggressor's (die0) power dissipation. The steady-state temperature contours for 2 mm inter-die spacing are shown in Figure 4.3. The maximum coupling for this case, observed for die3, was found to be an order of magnitude higher using air-cooling (0.246 ^{o}C/W) compared to microfluidic cooling (0.011^{o}C/W).



(a)



(b)

Figure 4.3: Steady state coupling for (a) air and (b) monolithic microfluidic cooling configurations with die0 (top left) active, all other dice not powered, and 2 mm inter-die lateral (x and y) spacing. Die face view looking from below the package.

Effect of Die Spacing

Reducing die-to-die link length is beneficial in 2.5D integration to reduce channel EPB and latency [110]. However, the thermal performance implications of this depends on the type of cooling used. Therefore, to analyze this thermal-electrical trade-off, we varied the inter-die lateral (x and y) spacing from 2 mm to 8 mm and analyzed the following two cases:

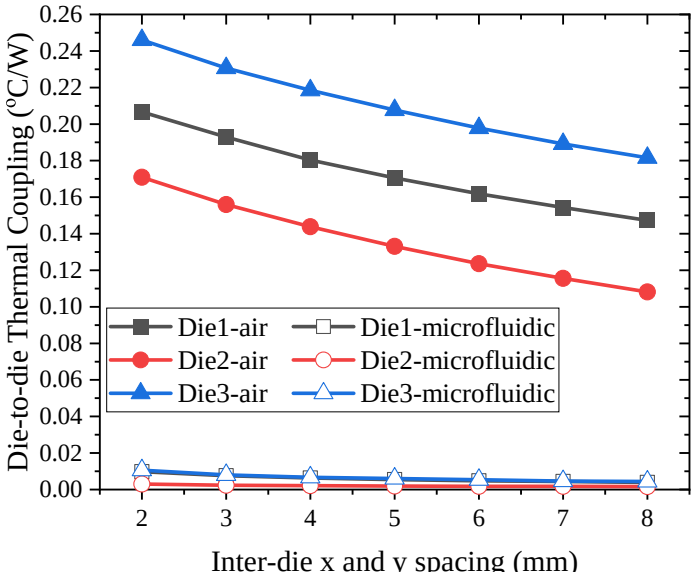
1. Chiplet (die0) active with other chiplets (die1, die2, die3) dissipating no power, to quantify the effect of spacing on thermal coupling.

2. All chiplets active to estimate the worst-case impact on absolute junction temperature.

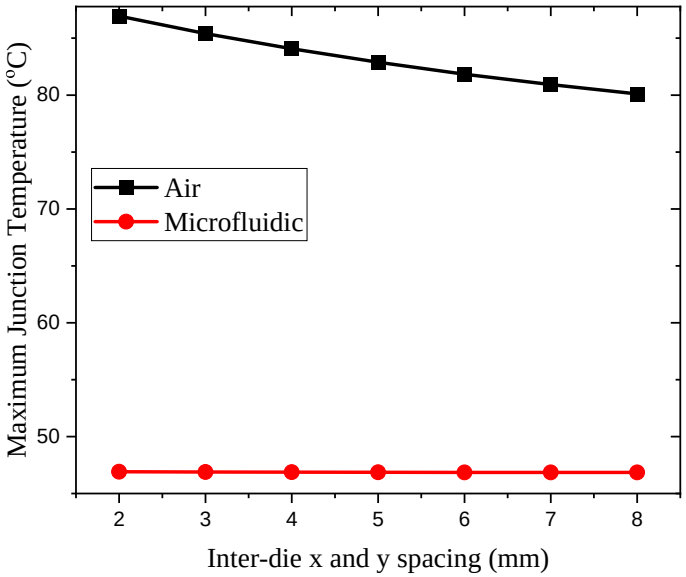
For case 1), the results are summarized in Figure 4.4a. Although increasing separation from 2 mm to 8 mm causes the coupling in air-cooling to reduce, it still remains significantly higher compared to monolithic microfluidic cooling. This is because the presence of a single heat spreader in an air-cooled package provides a low resistance lateral path for energy from one chiplet to couple to a neighbouring chiplet.

For case 2), an increased inter-die spacing leads to a higher reduction in maximum junction temperatures for air-cooling (6.83^{o}C) vs microfluidic cooling (0.06^{o}C), as illustrated in Figure 4.4b. However, a higher thermal resistance path in the upward direction in air-cooling leads to higher junction temperatures for all values of die-spacing compared to microfluidic cooling.

The range of die-to-die spacing was selected based on the typical values seen in the MCM package based integration used for this analysis. While other 2.5D integration technologies have tighter die-to-die spacing, the range was chosen to be compliant within the limits of MCM technology.



(a)



(b)

Figure 4.4: (a) Die-to die thermal coupling and (b) maximum junction temperature as a function of die spacing.

Figure 4.5: Intel Stratix 10 ES development kit with the Stratix 10 GX FPGA package attached.

4.2 Experimental Set-up

The experimental demonstration described in this section is based on the results published in [104]. An Intel Stratix 10 ES development kit, shown in Figure 4.5 was used for this thermal demonstration. The on-board Stratix 10 GX FPGA consists of a 14nm FPGA core die connected to four surrounding transceiver tiles using the Intel Embedded Multi-die Interconnect Bridge (EMIB) as can be seen from the delidded image shown in Figure 4.6. The package contains a total of 96 transmitters and receivers, distributed equally across the four tiles. Each transceiver channel requires a reference clock, provided to its corresponding tile for operation. These can also be adjusted to control the data rate of the transceivers. However, the Engineering Silicon (ES) development board version used for this demonstration does not have clock connected to tile t3 (Figure 4.6). Consequently, the tile t3, and the corresponding tile t2 on its opposite direction (referenced as on either sides of the direction of coolant flow along the main core) would not be used for benchmarking in this article.

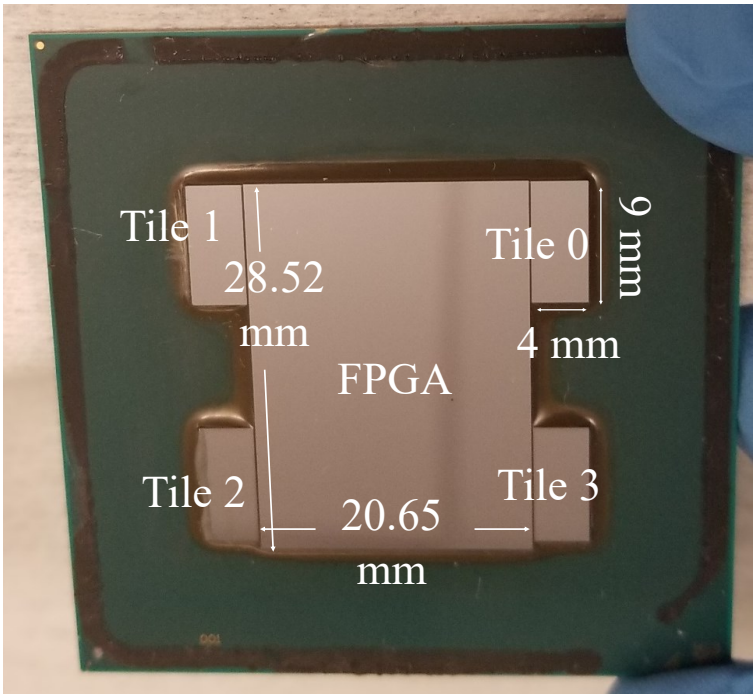


Figure 4.6: De-lidded Startix 10 GX FPGA showing the five chiplets.

4.2.1 Benchmark Application

Thermal issues are most pronounced in the high power use cases of the FPGA core and the transceiver chiplets. Hence, we designed a benchmark program to emulate these. For the FPGA core, we designed units that consist of a streaming Fast Fourier Transform (FFT) block followed by six First-in First-Out (FIFO) buffers operating on random hard coded inputs. The implementation utilizes the programmable logic blocks as well as the Digital Signal Processing (DSP) blocks on the core die. The arithmetic operations on the floating point inputs performed by the DSP blocks account for majority of the power dissipation. This unit design was repeated 160 times across the FPGA core to utilize near 100 % of available compute resources. The computational throughput, and correspondingly, the power dissipation was controlled by varying the reference clock, with a 475 MHz base frequency used for initial benchmarks. The second unit of the benchmark program was designed to mimic a data intensive use case of the transceivers. All 24 transceiver channels on both transceiver tiles used for benchmarking were programmed to run in enhanced

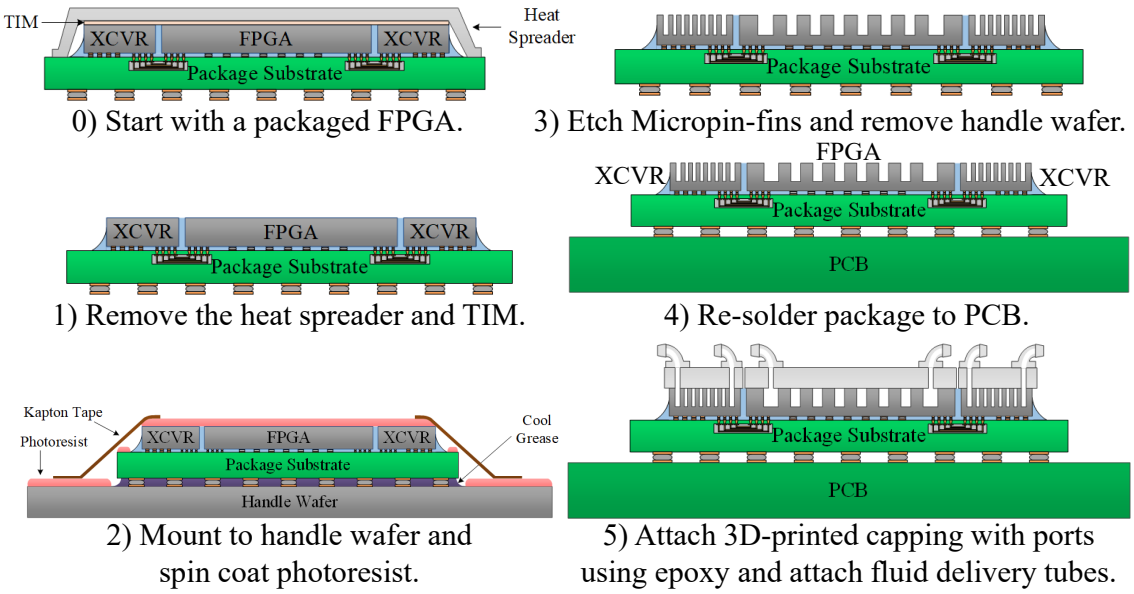


Figure 4.7: Fabrication process for etching the 2.5D package and attaching manifold structures.

Physical Coding Sublayer (PCS) mode with serial loopback enabled. The design had a base data rate of 16Gb/s per channel at a transceiver reference clock rate of 400 MHz. For increasing the power density, we overclocked this to up-to 550 MHz for a data rate of 22 Gb/s per channel.

We also programmed a Nios II soft processor in the FPGA core design, which could read the output data from the Temperature Sensing Diodes (TSDs) included in each chiplets; more details are provided in later sections. This Nios processor communicates over a Joint Test Action Group (JTAG) interface with a Nios terminal running on a host computer for temperature read-outs. The package was powered by a total of 5 power rails which were measured by onboard sensors and interfaced to an Intel Board Test System (BTS) software for data logging. Two of these rails (VCC, VCCRAM) fed the core die and the power dissipation on the remaining three (VCCR, VCCT, 1.8V) were assumed to be split equally between the transceiver tiles (equal split for static power on all four tiles, remaining dynamic power split equally between the clocked and operational two tiles). This assumption gives a near estimate of the power dissipation on each tile, without accounting for the die-

to-die variation which may cause slight differences between the transceiver chiplets.

4.3 Heat Sink Design

Reduced absolute temperatures and thermal coupling are the main design requirements for ideal heat sinks in 2.5D ICs. Lower thermal resistance, which can help manage higher power dissipation can be achieved by micropin-fin area enhancement structures etched directly on top of the dice. Further, a staggered micropin-fin array geometry is used, owing to its excellent thermal performance [96, 111].

A two-fold design approach was used to reduce the thermal coupling. The first approach aims to minimize the thermal coupling through conductive heat transfer. This mode of coupling is driven by the relative temperature difference between different chiplets, owing to their difference in power dissipation. In this work, we minimize this by targeting a more uniform temperature profile across different chips by tailoring the micropin-fin cooling efficacy proportional to the power dissipation in individual chiplets. To achieve this, micropin-fin heat sink's geometric dimensions over each die were designed according to the die's power density. Similar approaches have been explored for creating a heterogeneous cooling efficacy in chiplets with different power densities [112] as well as for managing hot spots within a die[99]. First, leakage power was estimated by running the benchmark program with all clocks disabled. The measurement from the transceiver rails was assumed to be equally divided between all four tiles. Next, measurements were made while the FPGA and transceiver clocks are set to 475 MHz and 550 MHz respectively while being cooled by an air-cooled heat-sink. These measurements, for the transceivers, represent the sum of static power on all four tiles, and dynamic power only the two clocked tiles. These measurements, estimated the FPGA core power density at 19 W/cm^2 and the active transceiver tiles at 56 W/cm^2(110 W in FPGA core, 21 W in active transceivers and 0.72 W static power in non-clocked transceiver tiles).

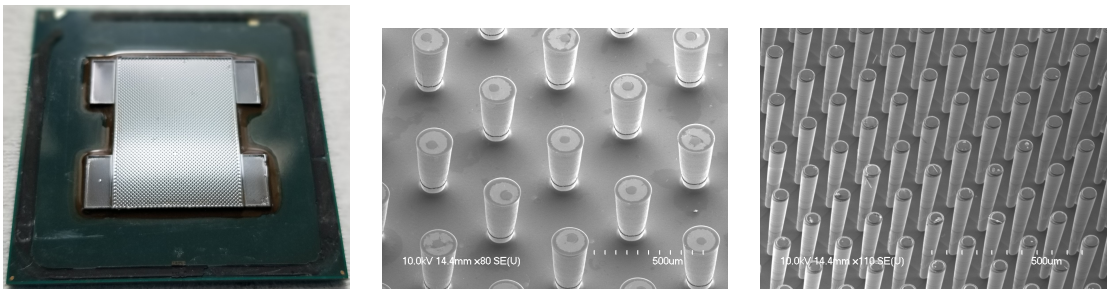
These were used as the baseline power densities for the heat sink design. Based on

fabrication constraints for etching the pre-packaged die on the etching tools available to the authors, the micropin-fins over the transceiver dice were designed to have 60 µm diameter and 200 µm pin-to-pin pitch. These corresponded to the ability to etch micropin-fins with minimal taper and near-vertical side-walls. Correspondingly, the scaled dimensions for the FPGA core region were set to 175.6 µm diameter and 585.1 µm pin-to-pin pitch. While this linear scaling of geometries does not translate to an equivalent linear translation in die temperatures, the increased density over the transceivers help ensure a more uniform temperature profile between all the dice.

Secondly, to prevent thermal coupling to dice in the coolant flow-path (i.e., downstream from inlet), a 3D printed manifold was designed to deliver fluid to individual dice. Allowing a shared flow where outlet of one feeds into the inlet of another die as in [112] would result in thermal coupling through caloric heating of the coolant. The rise in fluid temperature due to self heating can be described using

$$(T_{out} - T_{in}) = \frac{Q_{die}}{C_p f} \tag{4.1}$$

where T_{out} and T_{in} are the fluid outlet and inlet temperatures, Q_{die} is the power being removed through water, C_p is the specific heat of water, and f is the mass flow rate. For the transceiver tiles, using the power as 21 W (assuming no loss due to natural convection) and the specific heat of water as 4.18 J/(^{o}C g), we estimate a fluid heating of 10.04 ^{o}C at the lowest used flow rate of 0.5 mL/s. The temperature rise drops to 1.57 ^{o}C at the highest flow rate of 3.2 mL/s. This highlights the thermal coupling issues that may arise when the dice share a common fluid flow, and the potential mitigation by separating the flow to individual dice.



(a) Photograph of the etched dice.

(b) Close-up SEM of micropin-fins over FPGA core.

(c) Close-up SEM of micropin-fins over transceiver tiles.

Figure 4.8: Etched monolithic micropin-fins.

4.4 Fabrication and Heat Sink Assembly

The Stratix 10 GX package consists of 5 discrete chips as discussed earlier. The monolithic cooling solution was implemented by etching micropin-fins directly on the backside of these silicon dice, and then capping them with 3D printed fluidic manifolds. The process flow, as shown in Figure 4.7 used for this study is designed for a proof-of concept demonstration on a pre-packaged multi-die solution. In production, this could be modified to do wafer level etching before individual chiplets are singulated and processed.

The Ball Grid Array (BGA) package was first de-soldered from the development board, and its heat-spreader and TIM were removed. An optimized BOSCH silicon etching process was used to etch micropin-fin heat sinks of different dimensions simultaneously. The etch-passivation cycles were controlled to prevent excessive tapering on either micropin-fin structures due to the aspect ratio dependent variation of the DRIE etch process. The process optimization was also done to account for the increased complexity in etching of a packaged dice. This includes accounting for increased thermal stress during the process due to additional layers beneath the silicon chiplets. The device was mounted on a carrier silicon wafer with cool grease to prevent overheating during the process. A photoresist mask was used to define the micropin-fin pattern on all the dice simultaneously, and the organic substrate of the package was covered with photoresist and Kapton tape to prevent plasma damage. The etch was done to get a micropin-fin height of approximately 300

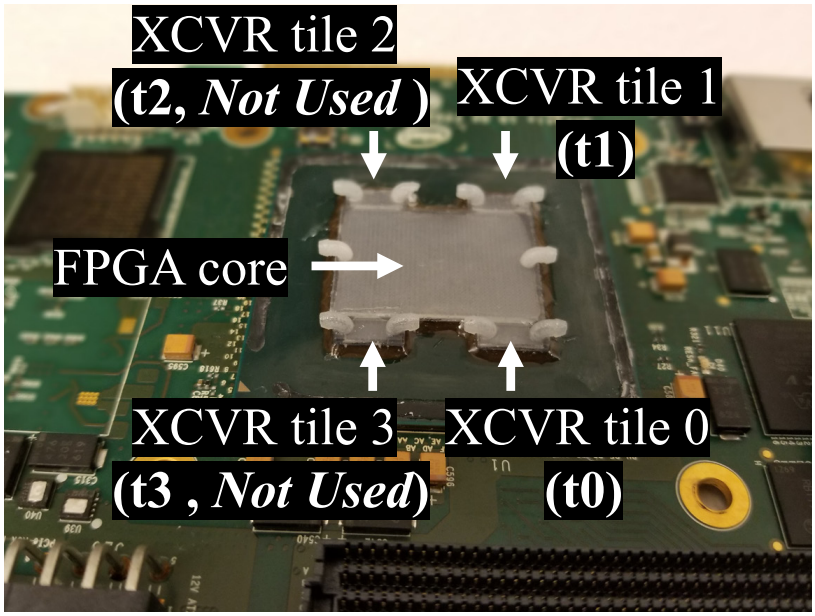


Figure 4.9: Stratix 10 GX development board with monolithic micropinfin heatsink etched and 3D printed manifolds for fluid delivery.

μm as characterized by multiple contact profilometry measurements using a Tencor profilometer. Scanning Electron Microscope images of the etched micropin-fins are shown in Figure 4.8. The etched BGA was then re-soldered on to the development board.

3D printed capping manifolds that conform to the dimensions of individual dice were designed using Solidworks 2018 and printed with 3D Systems Visijet M3 Crystal material jetting polymer from 3D systems[113], using a Projet 3510HD industrial 3D printer. The printing type (material jetting) and material was chosen so as to have acceptable levels of dimensional accuracy, as well as printed surface planarity. These help prevent leaks post attachment, as well as help ensure a mostly flushed capping layer above the micropin-fin region (not accounting for the slight gaps which may be introduced between the micropin-fins and the manifold wall due to surface roughness). This is particularly important as any gap between the micropin-fin region and the cap is a potential lower-resistance fluid bypass zone as compared to the dense, higher pressure drop micropin-finned region. Allowing larger bypass can potentially lead to lower cooling performance as most of the coolant may bypass the micropin-finned region, and hence may not contribute to cooling.

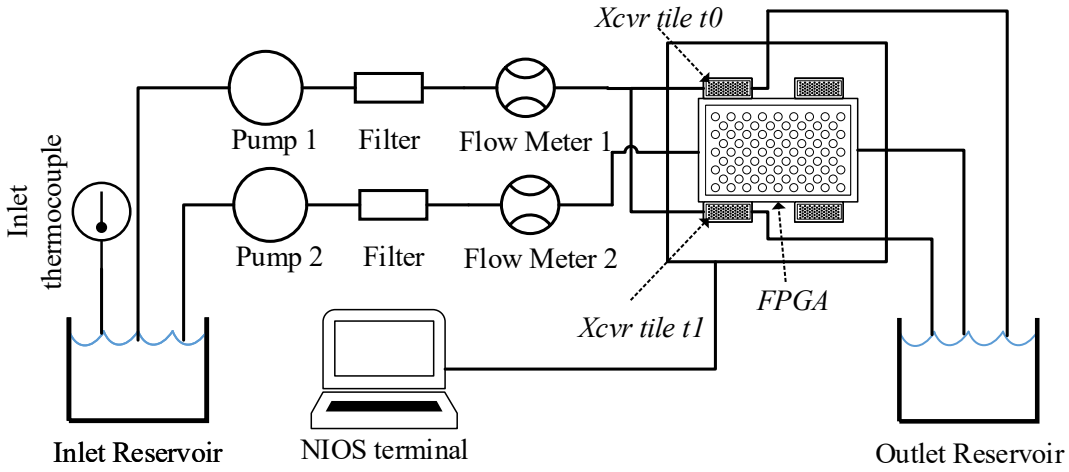


Figure 4.10: Open loop measurement set-up.

The caps were designed with lateral fluidic ports to have a very small form factor cooling solution. The caps were attached on individual dice using waterproof epoxy, and tubes were connected for coolant delivery. Figure 4.9 shows an image of the board post manifold attachment. Ports for tile t2 were damaged during the initial testing, so the tile t2, as well as the non-clocked tile t3 are not included in the benchmarking.

4.5 Testing Set-up

For benchmarking, we tested the system in an open loop configuration with room temperature DI water as the coolant. A block diagram of the testing set-up is shown in Figure 4.10. The FPGA core and transceiver tiles were cooled using two independent flow paths in accordance with the design considerations discussed earlier. All the measurements were done in a controlled environment with ambient temperature kept constant at 20^{o}C ± 1^{o}C. Ambient and inlet fluid temperatures were measured using k-type thermocouples, and the temperature measurement IP integrated into the FPGA design was used to read out the die temperatures from all the on-die temperature measurement diodes. In-line filters from McMaster-Carr were used directly after the pumps. A Kobold rotameter and elec-

Figure 4.11: FPGA development board with stock heatsink.

tronic flow meter were used to measure the flow rates in FPGA and the transceiver flow paths respectively, and both were calibrated by repeatedly filling a known volume of fluid. All calibrations were done at 21.2 ^{o}C ± 0.3 ^{o}C. It is worth noting that this would result in some systematic errors in flow rate measurements at elevated temperatures as the reduced viscosity would cause the measured flow rate to be lower than its actual value.

The benchmarking was limited to only two transceiver tiles along with the FPGA core die. For illustration purposes, the results are compared to a separate development board with the stock liquid-cooled solution (Figure 4.11) that was provided with it. The comparison with the closed loop stock solution is done as a technology demonstrator and for initial cooling benefit analysis.

Furthermore, the on-die TSDs are located on the edges of each die as shown in Figure 4.12. Also, for a microfluidic heat sink, the thermal performance deteriorates along the direction of the flow due to the fluid heating up. To limit the issues from this bias, as well as for a better estimation of overall chip temperature, we take measurements by reversing the direction of flow as well to get a better approximation of average temperature gradient across the dice. The average die temperature and the temperature gradient were estimated from these experiments as described in the following equations:

$$T_{jn,avg} = \frac{T_{jn,fwdflow} + T_{jn,revflow}}{2} \tag{4.2}$$

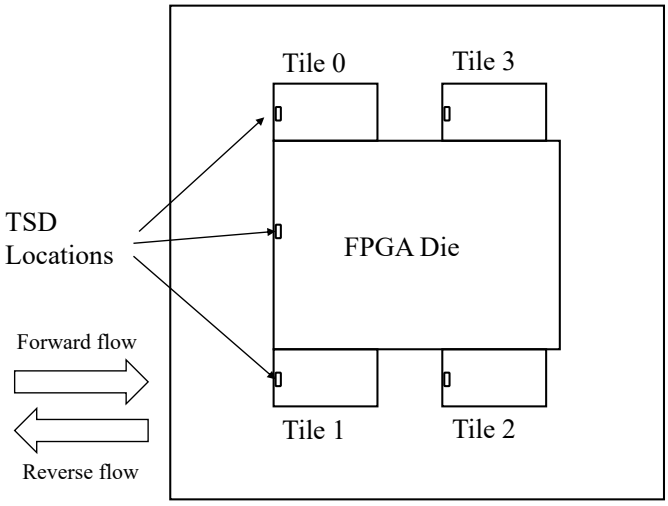


Figure 4.12: Temperature Sensor diode (TSD) position on the package floor-plan.

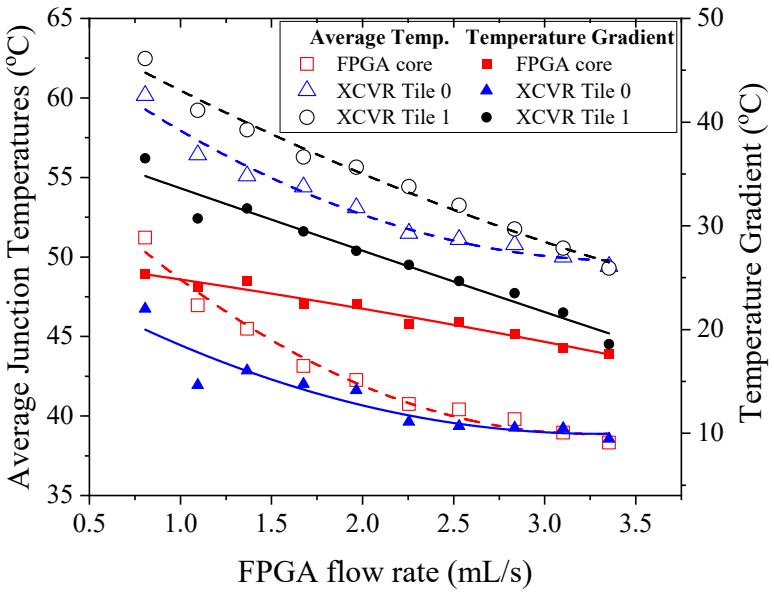
$$T_{jn,gradient} = T_{jn,revflow} - T_{jn,fwdflow} \tag{4.3}$$

Estimation of statistical variations across measurements was done by observing variations across all measured parameters. Each on-die temperature measurement was performed by continuously recording the Nios output for approximately 200 points after the temperature stabilizes. This stabilization period was set as 1 minute for the monolithic microfluidic heat-sink and 5 minutes for the stock liquid-cooled heat sink. We observed a less than 1^oC standard deviation in recordings taken after the stabilization period for all the test cases. The flow measurements were kept constant without any noticeable changes, subject to the resolution of the flow-meters for each experiment. Inlet and ambient temperature measurements were also kept within 1^oC variation. Power readouts from the BTS output were recorded for approximately 10 seconds on each of the 5 power rails. These corresponded to less than 1% variation within the recorded values.

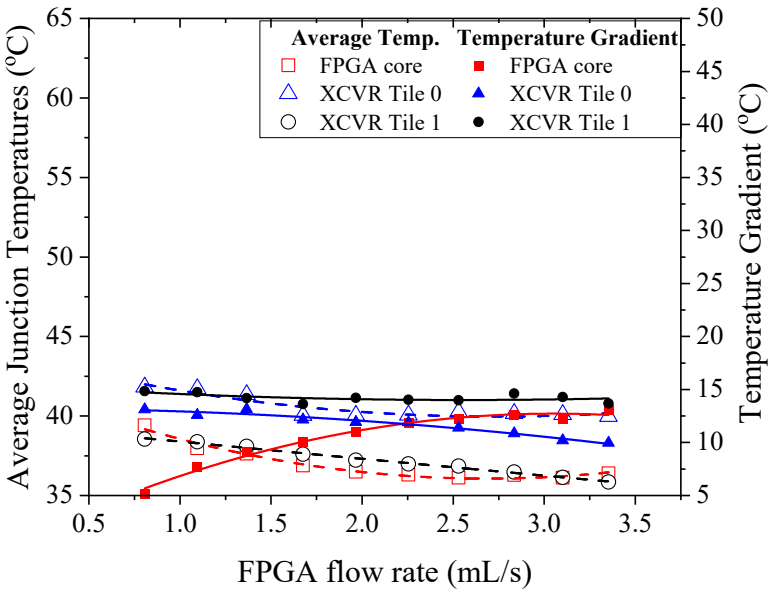
4.6 Results and Discussion

4.6.1 Variable Flow Rate Testing

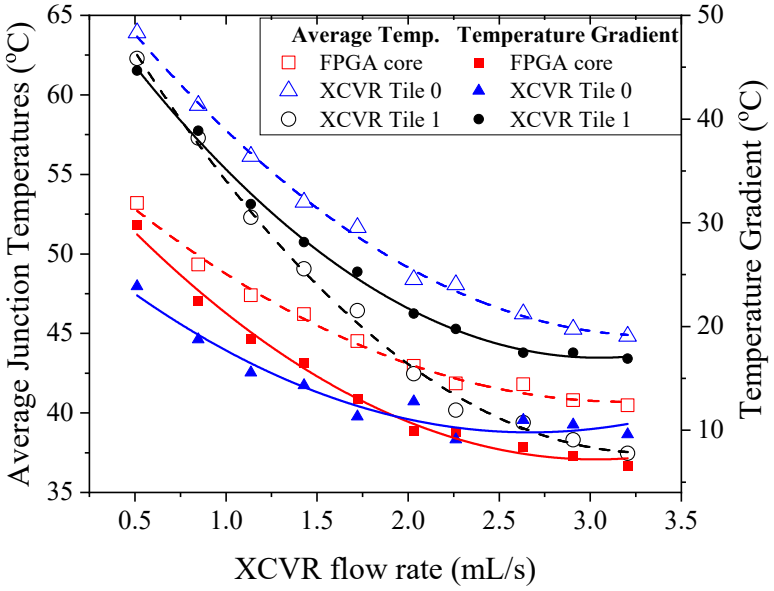
The convective thermal resistance, and correspondingly the overall performance of a fluidic heat sink with micropin-fin structures depends on the flow rate of the coolant in the channels. Hence, it is important to characterize the thermal performance under various flow conditions for both the FPGA core and the transceiver channels. Measurements were done for both flow directions, and sweeping both flow rates within the corresponding bounds have been done. The data from these 8 sets of sweeps are presented in Figure 4.13.



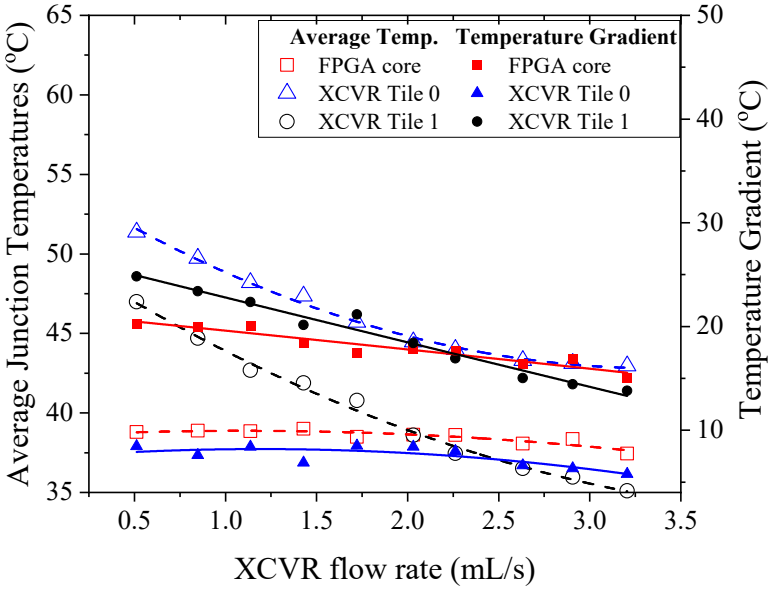
(a) Transceiver flow fixed at 0.539 mL/s.



(b) Transceiver flow fixed at 3.194 mL/s.



(c) FPGA flow fixed at 0.806 mL/s.



(d) FPGA flow fixed at 3.352 mL/s.

Figure 4.13: Die temperatures as a function of FPGA and XCVR flow rates for monolithic microfluidic heatsink. Temperature gradient is the range of measurements with the temperature diode near the inlet and near the outlet.

These measurements were done by setting the FPGA core frequency to 475 MHz and the transceiver clocks to 550 MHz for a corresponding overclocked data rate of 22 Gb/s. This corresponds to a total package power of approximately 152W.

The average temperatures for all chiplets drop as the flow rate increases, due to the reduction in convective thermal resistance. Further, the temperature gradient across the chip also drops as the fluid heats up less with increasing flow rates. It is worth noting that the fluid heating also captures the effect of underlying non-uniform power densities along the direction of flow, which may increase the measured thermal gradient.

These results also illustrates the ability to manage thermal performances of dice in a single flow path without affecting the nearby dice. For example, varying the FPGA flow rate while keeping transceiver flow constant changes the minimum core temperature by 6.96 ^{o}C, while the corresponding change in the transceiver tiles is 2.12 ^{o}C. This is particularly of interest if the relative workloads between individual dice can change. Since the pumping power for a given flow path is proportional to the coolant flow rate through it, this ability to control individual coolant paths with minimal effect to the performance of the other path can help tailor the cooling solution for target temperatures while optimizing the cooling power.

The slight differences in the temperatures of the transceiver tiles may come from the differences stemming from manual capping as well as from the small etch non-uniformities. Also, as explained earlier, the power dissipation in both tiles may have slight differences which could not be captured using the BTS measurements.

4.6.2 Thermal Coupling Measurements

To quantify the thermal coupling performance, we modulated the power in the FPGA core by changing its reference frequency, while the transceivers were operated at a constant data rate of 22 Gbps. The increase in the transceiver tile temperature with the the FPGA power represents the thermal coupling from the FPGA to the transceiver tiles. Following

equations were used to calculate the junction-to-inlet thermal resistance and the FPGA-to-Transceiver (xcvr) thermal coupling:

$$R_{jn-to-inlet} = \frac{T_j - T_{in}}{P} \tag{4.4}$$

$$Coupling_{fpga-to-xcvr} = \frac{\Delta T_{j,xcvr}}{\Delta P_{FPGA}}\bigg| P_{xcvr} = constant \tag{4.5}$$

Results (Figure 4.14, Table 4.2) show a 7× reduction in the FPGA core's junction-to inlet thermal resistance compared to the stock package level closed-loop liquid cooling. Since the transceiver operation was kept constant in this experiment, any change in transceiver temperature was caused by thermal crosstalk from the FPGA core die. The slope of the lines in Figure 4.14 quantifies this cross-talk, and as summarized in Table 4.2, demonstrates up to 43.3× reduction in core-to-transceiver thermal crosstalk for the monolithic 2.5D microfluidic cooling solution (while being cooled at the highest flow rates), when the core was dissipating approximately 107 W of power.

The comparatively lower thermal performance of the stock solution can be mainly attributed to the presence of multiple layers between the heat source and the coolant, all of which contributes to increasing the thermal resistance. These include the layer of thermal interface material (TIM) between the dice and the heat-spreader, the heat-spreader itself, the TIM between the heat-spreader and the cold-plate, and finally, the thick base plate of

Table 4.2: Junction to inlet thermal resistance and FPGA-to-xcvr thermal coupling comparison, calculated for FPGA flow rate of 3.35mL/s and xcvr flow of 3.21 mL/s

	Thermal Resistance (°C/W)		**Thermal Coupling(°C/W)**	
	Monolithic	Stock	Monolithic	Stock
FPGA	0.074	0.518	-	-
XCVR 0	0.910	2.504	0.042	0.403
XCVR 1	0.398	2.906	0.01	0.433

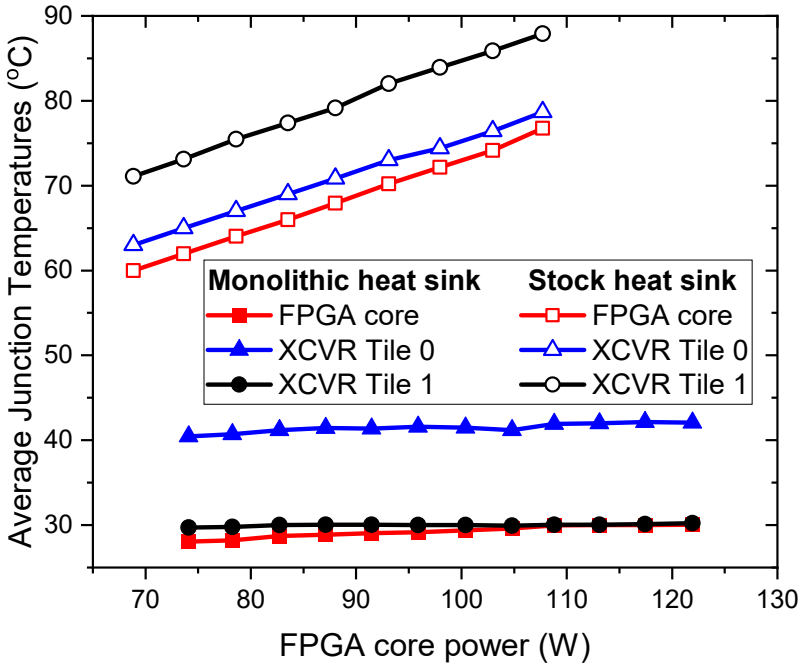
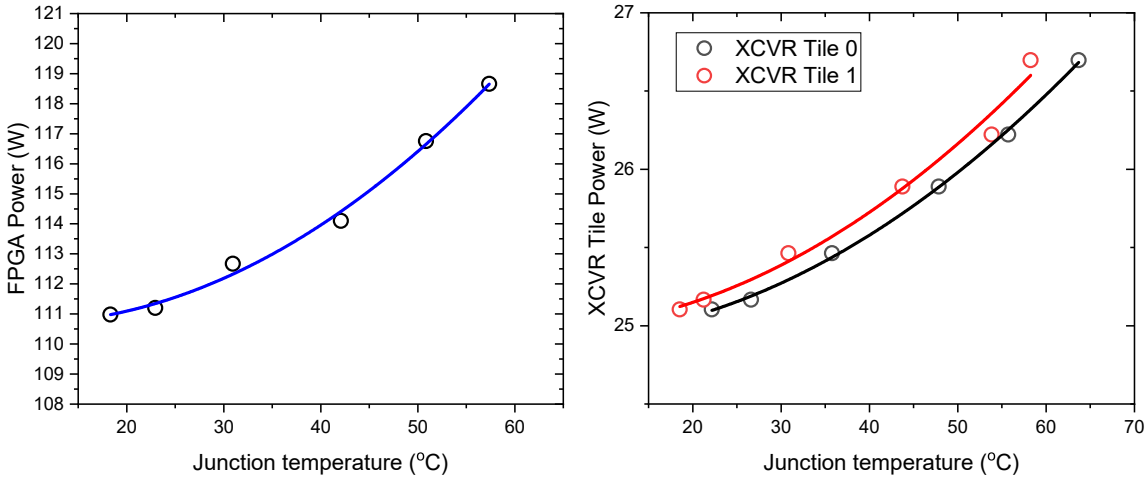


Figure 4.14: Heat sink performance comparison at different core powers. Both transceivers operating at a constant power of approximately 23W. Comparison with closed-loop stock solution provided as reference to conventional cooling techniques.

the cold-plate. An apparent increase in thermal resistance also comes from the closed-loop nature of the system, where the heat-exchanger efficiency also plays a role. The increased thermal coupling values can be attributed to large lateral heat-spreading capabilities of the package heat spreader as well as the base of the cold-plate.

Further, while the computed thermal resistance assumes the power to be transferred entirely to the liquid, there is some heat loss to the ambient through natural convection. This has been quantified for a monolithic microfluidic cooled device by Sarvey et. al [40] and the observations show less than 5 % heat loss to the surroundings at higher flow rates and lower inlet temperatures. While the heat loss increases at elevated inlet temperatures, at higher flow rates, this loss is still minimal compared to the heat removed through the liquid.



(a) FPGA core power variation at different die temperatures. Temperatures were controlled by varying the inlet fluid temperature at a constant flow rate of 3.35 mL/s.

(b) Transceiver chiplet power variation at different die temperatures. Temperatures were controlled by varying the inlet fluid temperature at a constant flow rate of 3.21 mL/s.

Figure 4.15: Power versus junction temperature.

4.6.3 Varying Coolant Temperatures

We also tested the device under various inlet temperature conditions from 5^{o}C to 50 ^{o}C. The FPGA core was operated at a frequency of 500 MHz and the transceiver clock was set to 600 MHz for total package power dissipation in excess of 161 W. The FPGA flow rate was kept constant at 3.35 mL/s and the transceiver flow was fixed at 3.21 mL/s.

These results point to two major advantages of using a high performance thermal management solution such as monolithic microfluidic cooling. First, as shown in Figure 4.15, a reduction in chip temperatures help reduce the overall power consumption of the device when performing the same operation. The static power component of the overall power includes components like sub-threshold leakage, which is strongly correlated to device temperature [31] and is a major contributor of this temperature dependent change in power. This points to the power saving benefits of operating compute and data intensive silicon at lower temperatures. We see an increase of 6.47% in power consumption for a 39^{o}C change in temperature for the FPGA and an increase of 5.96% on average for the transceiver tiles for an average temperature rise of 40.61^{o}C.

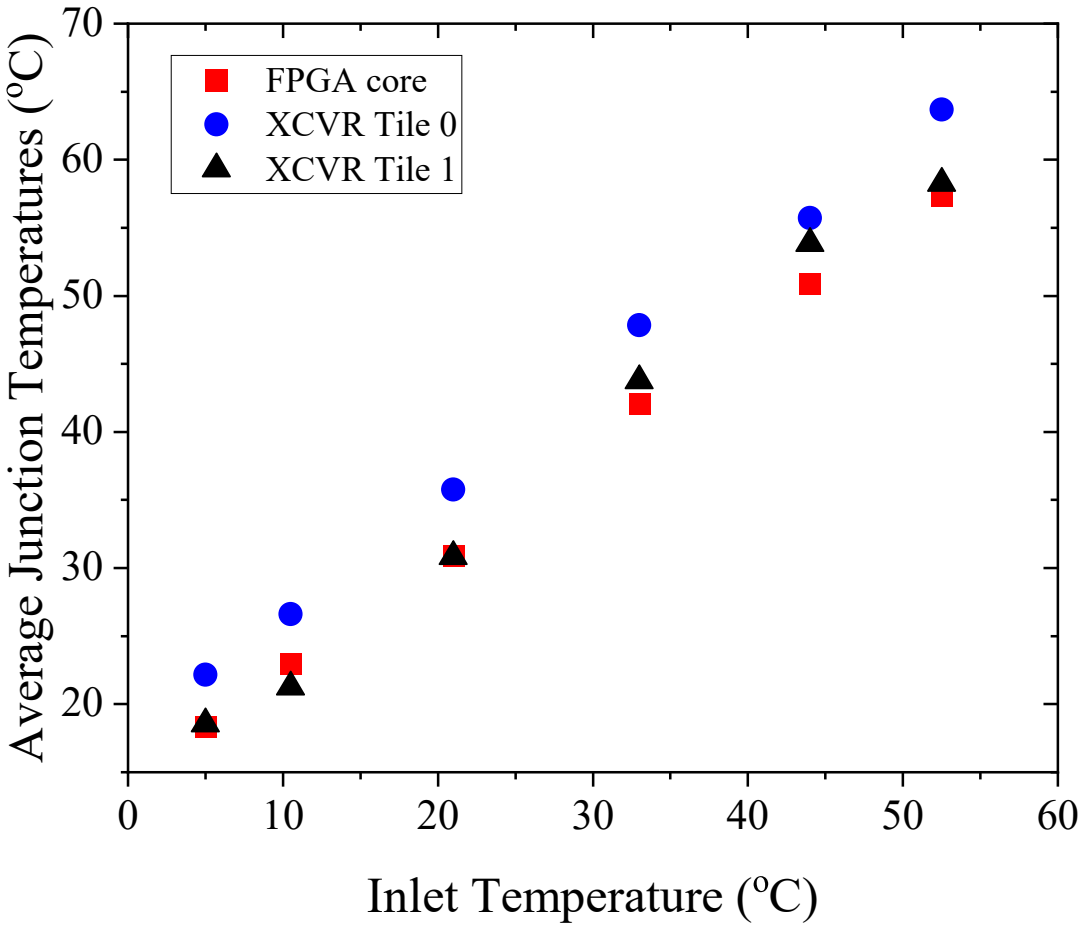


Figure 4.16: Junction temperature versus inlet fluid temperature.

Table 4.3: Comparison between Stock heat-sink and monolithic microfluidic heat-sink operating with elevated temperature inlet

	Stock	**Monolithic with inlet at 52.5°C**
Total power (W) (5 % increase in number of compute cores and a further 5% increase in core clock frequency)	114.32	172.06
FPGA core Temp. (°C)	59.98	57.37
XCVR 0 Temp.	63.02	63.68
XCVR 1 Temp.	71.09	58.26

Second, another major inference from these experiments can be observed from Figure 4.16 and Table 4.3. As can been seen from Figure 4.16, the heat-sink was able to keep all chiplets within 70^{o}C even with an inlet fluid temperature of 52.5^{o}C. This ability for high performance thermal management without the need for chilled inlet can be particularly useful for data-center like applications to increase the overall efficiency[114]. Further, as can be seen from Table 4.3, even with a 52.5^{o}C inlet, the monolithic heat sink can maintain roughly similar temperatures while dissipating 174 W that the stock liquid cooled heat-sink does for a 33$\%$ lower aggregate package power.

Table 4.4: Comparison of different single phase microfluidic cooling techniques

Source	**Test bench**	**Heat Sink Type**	**Total Package Power (W)**	**Die Area (cm^2)**	**Average Power Density (W/cm^2)**	**$R_{th,junc.-to-inlet}$ for highest power die in system ($^oC/W$)**
[115]	2.5D Multi-die, Simulation	Attached Micropin-fin	50.0 ($\times$1 die) + 30.0 ($\times$4 dice)	6.25 ($\times$1 die), 0.36 ($\times$4 dice)	8.0 ($\times$1 die), 83.3 ($\times$4 dice)	0.606*
[40]	Single-die, Experimental, Active	Monolithic Micropin-fin	32.4	8.4	3.9	0.070
[88]	Single-die, Experimental, Active	Monolithic Microchannel	$\approx$ 40.0	0.081	$\approx$ 493.8	0.170
[112]	2.5D Multi-die, Experimental and Simulation, Active	Attached Micropin-fin	114.0 ($\times$1 die) + 21.2 ($\times$3 dice)	5.8 ($\times$1 die), 0.38 ($\times$3 dice)	19.7 ($\times$1 die), 55.7 ($\times$3 dice)	0.094*
[116]	2.5D Multi-die, Experimental and Simulation, Acive	Jet Impingement	50.0 ($\times$2 dice)	0.64 ($\times$2 dice)	Upto 150.0 (Hotspots of size 240μm $\times$ 240μm over an 8mm$\times$8mm die)	0.203*
This work	2.5D Multi-die, Experimental, Active	Monolithic Micropin-fin	121.9 ($\times$1 die) + 23.1 ($\times$2 dice)	5.8 ($\times$1 die), 0.38 ($\times$3 dice)	20.7 ($\times$1 die), 60.9 ($\times$2 dice)	0.074

*Computed from the junction to inlet temperature rise, as well as the power and area values reported in the paper

4.6.4 Comparison with Other Microfluidic Coolers

Table 4.4 shows a comparison of this work with other microfluidic coolers from literature. To the best of our knowledge, this is the first demonstration of monolithic microfluidic cooling on active 2.5D CMOS. Also, the aggregate package power is among the highest of previous works, while demonstrating very low thermal resistance values. This demonstrates the superior performance of monolithic microfluidic cooling in high power 2.5D packages.

4.7 Conclusion

In this chapter, for the first time, we demonstrate the design, fabrication and testing of a monolithic microfluidic heat sink etched directly on the backside of chiplets in a functional 2.5D package. We use finite element modeling and experimental demonstration to show the electrical and performance benefits of monolithic microfluidic cooling for high power 2.5D ICs. Using Ansys modeling, we showed that monolithic microfluidic cooling provides up-to 40.03^{o}C reduction in maximum junction temperature, and up to 23.3 $\times$ lower die-to-die thermal coupling for a four-chiplet server class CPU package. We demonstrate the use of heterogeneous micropin-fin design for managing thermal performance of chiplets with disparate power densities kept in proximity. Furthermore, an ultra-small form factor overall thermal solution is achieved by creating 3D printed plastic manifold with lateral inputs for fluid delivery and encapsulation of individual chiplets. The demonstrated cooling technology achieves a thermal resistance of 0.074^{o}C/W for the core and a core-to-transceiver thermal coupling reduction of up to 43.3$\times$ compared to the stock liquid cooled solution, as well as considerable power savings. We also demonstrated the feasibility for high efficiency thermal management without the need for large and energy intensive heat exchangers or radiators by cooling with inlet fluid temperatures up to 52.5^{o}C, while dissipating 172W of power.

CHAPTER 5

SCALING MONOLITHIC MICROFLUIDIC COOLING TO 3D SYSTEMS: MICROPIN-FIN HEATSINKS WITH EMBEDDED TSVS

5.1 Thermal Challenges in 3D ICs and Microfluidic Cooling

The system-level scaling benefits of switching to a high-bandwidth heterogeneous integration architecture such as TSV-based 3D is increasing in relevance as discussed in the introduction. The key benefit of this approach is the ability to stack chiplets within the same lateral footprint without compromising on the interconnection density between them. This increasing circuit density however corresponds to a proportional increase in power density as well the the aggregate package power. As discussed in the previous chapters, the thermal challenges presented by both of these can lead to potential performance penalties. To this end, there has been considerable effort in the literature to extend microfluidic cooling to 3D stacks to provide a more aggressive cooling solution.

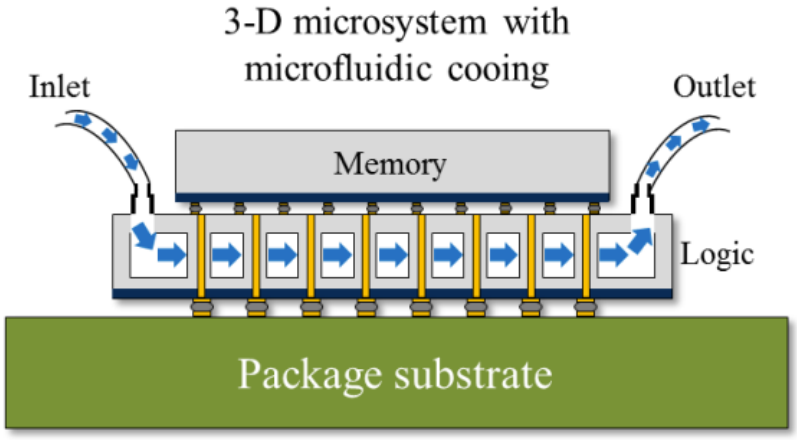


Figure 5.1: 3-D integration with embedded inter-layer microfluidic cooling

Multiple simulations and experimental evaluations of monolithic microfluidic cooling have been demonstrated in the existing literature [117–120]. In particular, inter-layer mi-

crofluidic cooling approaches where fluid flows between multiple active tiers as shown in Figure 5.1 can be beneficial for very high power devices. Bringing the working fluid in-between multiple heat sources helps to more effectively cool them, as well as reduce the thermal coupling between the tiers [121].

5.2 Inter-Layer Cooling: TSVs and Design Trade-offs

As shown in Figure 5.1, integrating an inter-layer cooling approach requires TSVs integrated within the micropin-fins or microchannels to facilitate electrical connectivity between different tiers. This introduces the need to co-optimize the thermal performance of the heatsink with the electrical performance of the integrated TSVs. The dependence of the thermal and electrical properties on the geometry is captured in the subsequent parts of this section.

5.2.1 Micropin-fin Thermal Performance

The performance of a micropin-fin heatsink can be described using its thermal resistance, as shown in the previous chapters. Considering a single phase cooling approach, the following sets of relations from [122] can be used to describe the heatsink performance.

Considering the negligible contribution of conductive thermal resistance in a micropin-fin heatsink, the overall thermal resistance can be quantified as:

$$R_{total} = \underbrace{\frac{1}{h_{ave}A_t}}_{R_{convective}} + \underbrace{\frac{1}{W_t c_p}}_{R_{caloric}} \tag{5.1}$$

where h_ave is the average convective heat transfer coefficient, A_t is the effective heat transfer area, W_t is the mass flow rate, and c_p is the specific heat capacity of the cooling fluid, respectively. The components of convective resistance can be described using the following sets of equations:

$$h_{ave} = Nu.\frac{k_f}{D_h} \tag{5.2}$$

where *Nu* is the Nusselt number and D_h micropin-fin hydraulic diameter. Nu can be described in terms of Reynolds number *Re*, Prandtl number *Pr*, and Prandtl using fluid surface temperature Pr_s as:

$$Nu = CRe^m Pr^{0.36} \frac{Pr}{Pr_s}^{0.25} \tag{5.3}$$

The effective heat transfer area for convective heat transfer, A_t:

$$A_t = A_b + \eta A_{fin} \tag{5.4}$$

$$\eta = \frac{tanh(2H_{fin}\sqrt{h_{ave}/k_{Si}D})}{2H_{fin}\sqrt{h_{ave}/k_{Si}D}} \tag{5.5}$$

where k_{Si}, H_fin, *D*, and η are the thermal conductivity of silicon, micropin-fin height, micropin-fin diameter, and micropin-fin efficiency respectively. It is worth noting that these relations are simplified assuming a uniform Si micropin-fin. Embedding the non-homogeneous TSV structure with materials having different thermal properties than Si into the fin can change the effective thermal conductivity, and correspondingly, the fin efficiency. A_b is the base area exposed to the fluid, and A_{fin} is the aggregate surface area of the micropin-fins exposed to the fluid and are calculated as follows:

$$A_b = A_{tot} - \frac{1}{4\pi D^2} \tag{5.6}$$

$$A_{fin} = n\pi D H_{fin} \tag{5.7}$$

Finally, considering the overall chip length and breadth as *L* and *B* and W_s, L_s, P_l, and

P_t as the horizontal and vertical spacing and the lateral and transverse pitches associated with the staggered micropin-fin array, the total number of micropin-fins, *n* can be computed as:

$$n = \frac{(W - W_s)(L - L_s)}{P_l P_t} \tag{5.8}$$

5.2.2 TSV Electrical Performance

The interconnect density, along with the electrical properties of the via, can be used to characterize the vertical bandwidth for a TSV based 3DIC. The electrical equivalent circuit of a TSV is shown in Figure 5.2.

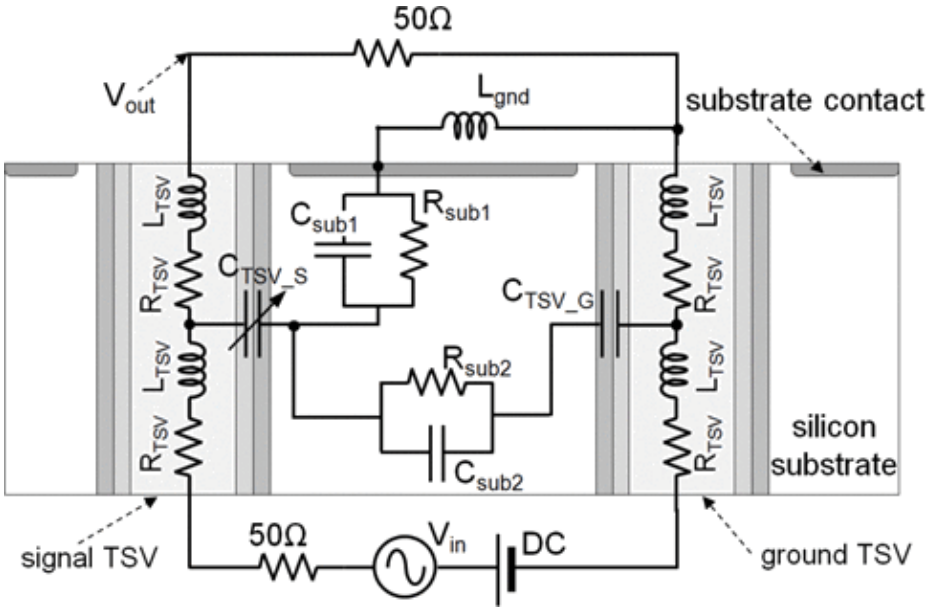


Figure 5.2: The electrical equivalent circuit of a ground-signal TSV pair showing the various parasitic components [123].

The electrical parasitic components include as the resistance and inductance associated with the filled via, the conductance component due to the finite resistivity of the dielectric liner and silicon substrate, and the TSV capacitance components. The capacitance component captures two major effects: (1) the electric field emanating from the TSV, being terminated in the ground-biased silicon and, (2) the electric field emanating from the TSV,

being terminated in nearby ground-biased TSV(s). For both cases, the total capacitance is a sum of the oxide capacitance (due to the dielectric barrier), the MOS capacitance due to the formation of a depletion region around the TSV, and the substrate capacitance. Of these, the MOS capacitance is voltage dependent and can therefore introduce non-linearities in the TSVs electrical performance. The following sets of equations can be used to quantify the relation of these parasitic components to the geometric dimensions.

The TSV electrical resistance can be described using [124, 125] as:

$$R_{AC} = \frac{\rho l}{\pi(2r\delta - \delta^2)} \tag{5.9}$$

where *l* is the TSV length, *r* the radius and δ the skin depth. Katti *et.al.* describes the TSV inductance *L* in [126] as

$$L = \frac{\mu_0}{4\pi}\left[2l\ln\left(\frac{2l+\sqrt{r^2+(2l)^2}}{r}\right) + r - \sqrt{r^2+(2l)^2}\right] \tag{5.10}$$

where μ_0 is the metal permeability. The via parasitic capacitance can be described using the relations in [127] as

$$\frac{1}{C} = \underbrace{\frac{1}{2\pi l\epsilon_{ox}}\ln\left(\frac{r+t_{ox}}{r}\right)}_{oxide-capacitance} + \underbrace{\frac{1}{2\pi l\epsilon_{Si}}\ln\left(\frac{r+t_{ox}+w_{dep}}{r+t_{ox}}\right)}_{depletion-capacitance} \tag{5.11}$$

where t_{ox} is the oxide thickness, and w_{dep} is the width of the depletion region. Finally, the depletion region conductance is characterized in [128] as

$$G_{dep} = \frac{\sigma_{Si}}{\ln\left(\frac{r+t_{ox}+w_{dep}}{r}\right)\sqrt{1-\frac{V}{V_{th}}}}2\pi l \tag{5.12}$$

where σ_{Si} is the silicon conductivity, *V* is the applied voltage, and V_{th} is the threshold voltage.

5.2.3 Thermal-Electrical Co-Optimization

As can be seen from the equations, a taller micropin-fin can lead to lower convective thermal resistance and improves the thermal performance. The traditional approach of keeping the aspect ratio constant (typically between 10:1 and 15:1) would therefore lead to larger via diameter to accommodate the height increase. This however can degrade the electrical performance by increasing the parasitic capacitance as shown in Equation 5.11. The larger area requirement would also lead to a reduction in via density. Therefore, increasing the TSV aspect ratio corresponding to the thickness increase is needed to ensure optimal electrical and thermal performance simultaneously. Zhang *et.al.* quantified these trade-offs for different via aspect ratios, depicting these trade-offs as shown in Figure 5.3 [122].

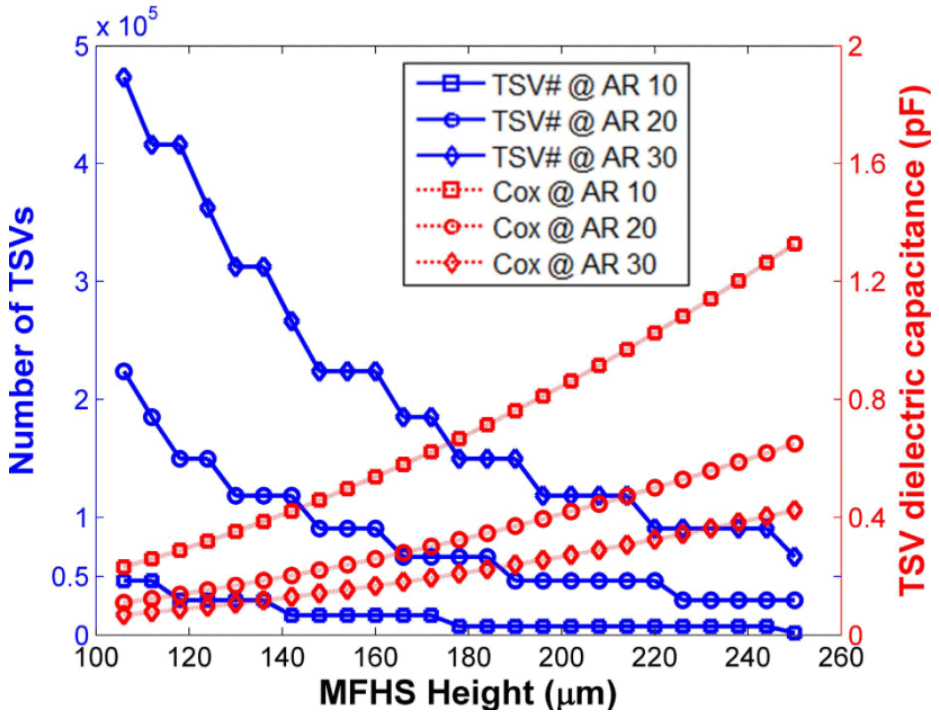


Figure 5.3: Impact of MFHS height on the number of TSVs and TSV capacitance [122].

5.3 High Aspect Ratio TSVs in a Micropin-fin Heatsink

A summary of existing implementations of TSVs in microfluidic heatsinks is shown in Table 5.1. As can be seen from the geometric parameters presented, most of these implementations use very large vias with stadard aspect ratios, and therefore suffer from the electrical trade-offs discussed earlier. While Oh *et. al.* scaled the TSV aspect ratio to ensure better electrical performance, the via diameter of 13 μm used in this implementation is relatively large to provide the interconnect density required for current generation products. For example, most commercial products using TSV based 3D stacking utilize vias with sub-10 μm diameters.

To address these issues, this work focuses on developing a high-aspect ratio TSV fabrication flow, while scaling down the via diameters. Furthermore, individual process steps including the via etching and filling are optimized to ensure the scalability of the process to lower via diameters / higher aspect ratios. Finally, the electrical performance of the scaled down vias would be investigated using finite element simulations.

Table 5.1: Comparison of TSVs integrated with microfluidic cooling technology. Adapted from [129]

	King et.al [130]	Madhour et al. [131]	Khan et al. [132]	Zhang et al. [122]	Oh et al. [117]	This Work
TSV Diameter	50 μm	60 μm	150 μm	13 μm	13 μm	6.4 μm
TSV Height	300 μm	380 μm	400 μm	400 μm	320 μm	155 μm
TSV Pitch	200 μm	200 μm	500 μm	24 μm	200 μm	50 μm
Heatsink Type	Channel	Channel	Channel	Micropin-fin	Micropin-fin	Micropin-fin
Aspect Ratio	6:1	6.33:1	2.66:1	23:1	23:1	24:1

5.4 Fabrication Process

The overall process-flow developed for this demonstration is shown in Figure 5.4. A 300 μm thickness <100> Si wafers were used. Individual process-steps and optimizations employed for each are discussed in the subsequent parts of this section.

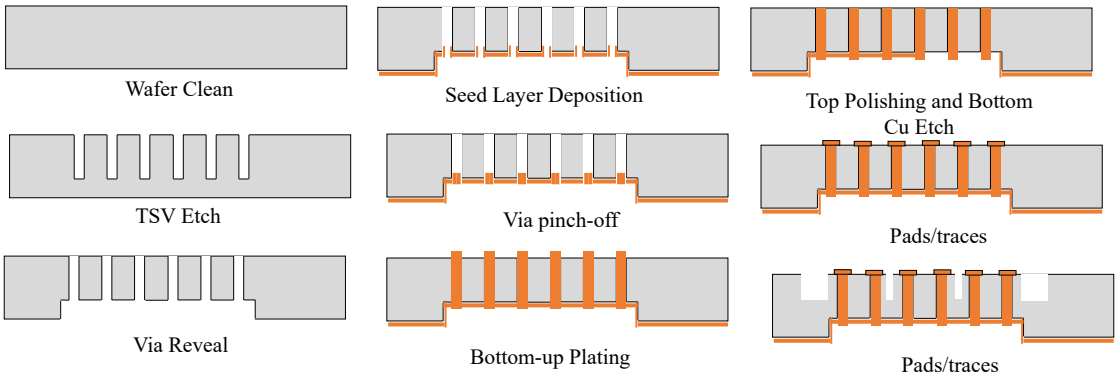


Figure 5.4: The process-flow used for TSV fabrication.

5.4.1 High-Aspect Ratio Via Etch

Optimal etching of the high aspect ratio vias is a major challenge in creating the High Aspect Ratio (HAR) TSV flow. Typically, a Bosch process based etch, where alternating etch and passivation steps, primarily using sulfur hexaflouride (SF_6) and octaflourocyclobutane (C_4F_8) gases respectively, is used.

The cyclical nature of the Bosch process can lead to scalloping of the via sidewalls. These rough sidewalls consequently leads to potential issues in liner and barrier deposition. Furthermore, the sidewall roughness can degrade the electrical signal propagation speed through the via, especially at higher frequencies [133]. This effect is predominant when the skin-depth of signal propagation becomes comparable to the root-mean-square value of the conductor roughness. As the signal propagation frequency increases, the skin depth reduces, requiring a corresponding reduction in the sidewall roughness to prevent signal degradation. Therefore, optimizing the sidewall roughness by controlling the key process parameters in Bosch etch is critical [134, 135].

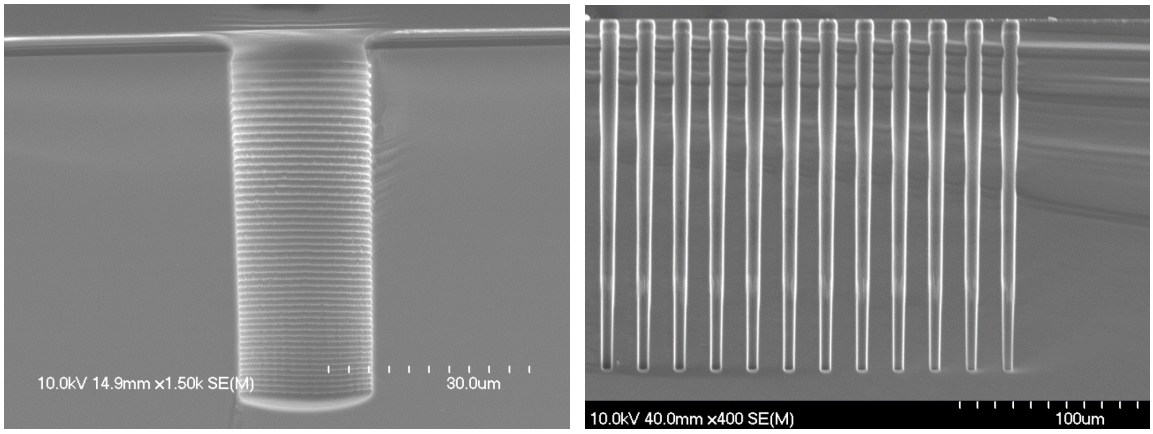
Table 5.2: Process parameters used for the optimized Bosch etching

Parameter	**Dep**	**Etch**
Coil Power	2000 W	2800 W
Platen Power	0 W	60 W
Pressure	25 mTorr	30 mTorr
C4F8 Flow	250 sccm	30 to 0 sccm
SF6 Flow	0 sccm	390 sccm
O2 Flow	0 sccm	39 sccm
Time	2 s	3.1 to 3.5 s
Chuck Temperature	10°C	

While the scalloping can be reduced by using shorter etch and passivation cycle times, this approach would to a reduction in the Si-to-mask selectivity of Bosch process, necessitating thicker and/or hard masks. To avoid these, the Bosch process parameters including the coil and platen power, as well as the gas flows were optimized to compensate for the reduction in selectivity from lower cycle times. The optimized parameters used for the Bosch etch are shown in Table 5.2.

The wafer was cleaned using a 3:1 H_2SO_4:H_2O_2 piranha solution at 120 °C for 20 minutes. The etch mask was then patterned using SPR-220 7.0 photoresist. The lithography process was optimized to minimize the undercut in the developed resist so as to prevent transferring this to the Si during the very long etch step. Further, the mask opening was adjusted to account for the expected lateral etching from the Bosch process.

The patterned wafer was mounted on a carrier wafer using spin-coated Crystalbond 509. For this, a 20:80 Crystalbond : Acetone mixture was prepared, and spin-coated on the carrier wafer at 1500 RPM. Acetone spray was used to remove the crystalbond edge bead from the wafer edges. The coated wafer was soft baked at 100 °C for 10 minutes on a hotplate to remove the acetone. The patterned wafer was then mounted on the carrier wafer at 120 °C on a hotplate.



(a) (b)

Figure 5.5: Cross-sectional SEM showing the etched vias. (a) Non-optimized recipe showing the side-wall scallops. (b) Optimized recipe with negligible scalloping and sidewall roughness.

The vias are then etched in the ICP using the aforementioned optimized Bosch process. The photoresist mask is then removed using resist remover 1165. The wafer is then cleaned with acetone, methanol, isopropyl alcohol, and water and dried under nitrogen. Figure 5.5 shows SEM cross-sections of the etched vias.

5.4.2 Backside Via Reveal

While the traditional TSV process-flows use a blind-via filling approach, the very high aspect ratios needed for this application make this approach challenging. More details regarding this would be discussed in the future work section. To circumvent these issues, a modified through-via based approach, derived from [117], is used in this demonstration.

To this end, the HAR vias etched in the last step are revealed using a backside ICP etch process. Similar wafer mounting, etch and cleaning processes were used for this step.

5.4.3 Oxide Liner

To remove the etch residue from the insides of the etched via, the wafer was cleaned in ultrasonicated acetone, metanol and isopropyl alcohol. The ultrasonic agitation was used to ensure that the solvents displace air from within the high aspect ratio vias and clean them.

After the AMI clean, a piranha clean and a hydroflouric acid (HF) dip were performed to remove all residues and any native oxide from the wafer. The wafer was then thermally oxidized in a furnace at 1100 °C using a wet oxidation process.

5.4.4 Seed Layer Deposition

After the liner oxide growth, a metal sputtering process was used to deposit the electroplating seed layer. The oxidized wafer was cleaned using a 30 seconds oxygen-plasma based descum process, and a 15 nm Ti / 350 nm Cu layer was sputtered on the backside. The aspect ratio limitations of the sputtering process for the narrow vias causes only partial coating of the sidewalls as illustrated in the process flow.

5.4.5 Via Electroplating

A two-step electroplating process was used to fill the TSVs. First, a pinch-off step was performed, which closes the bottom of the vias, and then a bottom-up plating to fill the remaining portion was used. An Elevate Cu Electrolye 2.0 solution with 0.6% v/v brightener additive and 0.1% v/v carrier additive, was used for both steps.

Via Pinch-off

A reverse-pulsed electroplating process, with process parameters shown in Table 5.3, is used for this step. This was used to ensure the uniform pinch-off of the bottom side as shown in the process- flow illustrations, without the corners closing off. The dissolution pulse helps reduce the corner pinch-off. The increased current density in the via corners would typically enhance plating in those regions, which may cause it to pinch-off before the insides are filled. By adding the dissolution pulse, which removes at a higher rate from the corners due to the same increased current density, pinch-off effects are be mitigated.

Table 5.3: Process parameters for pinch-off process

Process Parameter	**Value**	**Unit**
Forward Current Density	30	mA/cm^2
Forward Cycle Duration	25	ms
Forward Pulse Duty Cycle	50	%
Forward Pulse Frequency	200	Hz
Reverse Current Density	60	mA/cm^2
Reverse Cycle Duration	2	ms
Reverse Pulse Duty Cycle	50	%
Reverse Pulse Frequency	1000	Hz

Bottom-Up Plating

Bottom-up plating was then performed using a forward pulsed electroplating approach. The process parameters used are shown in Table 5.4. The pinched-off side was covered with an adhesive tape, and the edges were sealed with a photoresist to prevent any plating on the backside. A vacuum wetting process was done in water before bottom-up plating to ensure any trapped air was replaced with water inside the vias.

Table 5.4: Process parameters for bottom-up process

Process Parameter	**Value**	**Unit**
Forward Current Density	30	mA/cm^2
Forward Pulse Duty Cycle	50	%
Forward Pulse Frequency	200	Hz

5.4.6 Excess Cu Removal and Electrical Measurement Pads

The overplated Cu stubs on the top side were removed using a mechanical polishing step. For the TSVs where an open structure on the backside was required, a wet etch using APS-100 Cu etchant was performed to removed the electroplated and sputtered Cu seed.

Cross-sectional SEM images of the filled TSVs are shown in Figure 5.6. A 15 nm Ti / 350 nm Cu metal lift-off process was used to create 25 μm diameter probing pads around the TSVs.

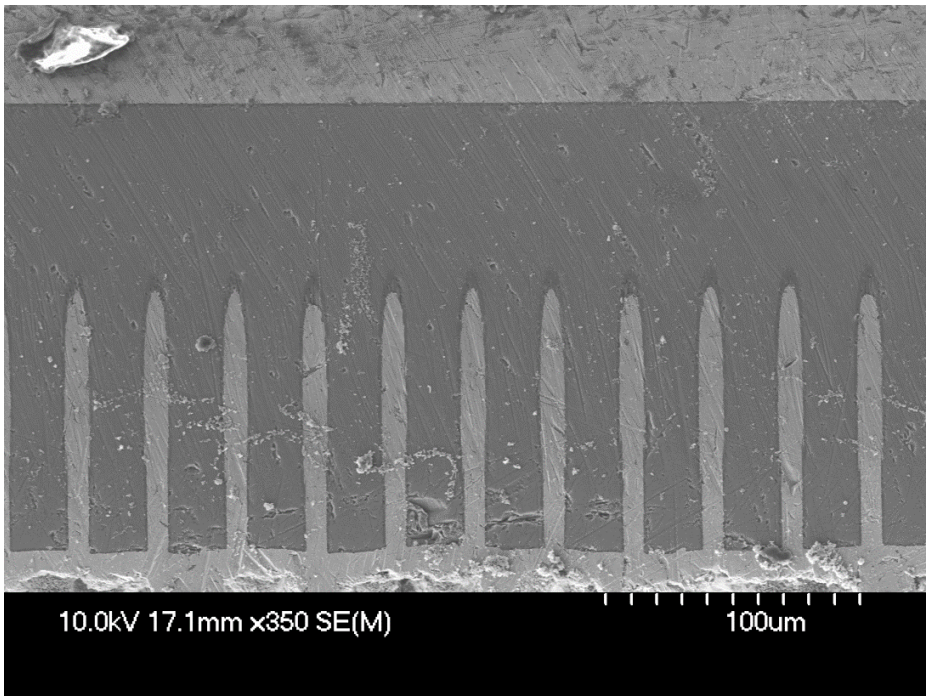


Figure 5.6: Cross-sectional SEM showing the plated TSVs.

5.4.7 Micropin-fin Etch

Finally, the micropin-fins are patterned using a photoresist mask. A dry oxide etch step followed by Bosch Si etching was used to create the micropin-fins. A mean pin-fin height of 55.3 μm was measured using a contact profilometer.

5.4.8 Fabrication Results

Cross-sectional SEM imaging as well as X-ray inspections were used to characterize the fabricated TSVs. A Dage X-Ray Inspection System (XD7600NT model) was used to capture X-ray images of the filled vias. Images from multiple orientations were used to ensure the void-free filling. Figure 5.7 shows the X-ray image of a pair of three TSV groups,

with the dark uniform coloration showing the complete filling of the vias. Figure 5.8 shows the SEM images of the TSVs within a micropin-fin. The slight non-uniformities from the mechanical polishing process are also visible in this image.

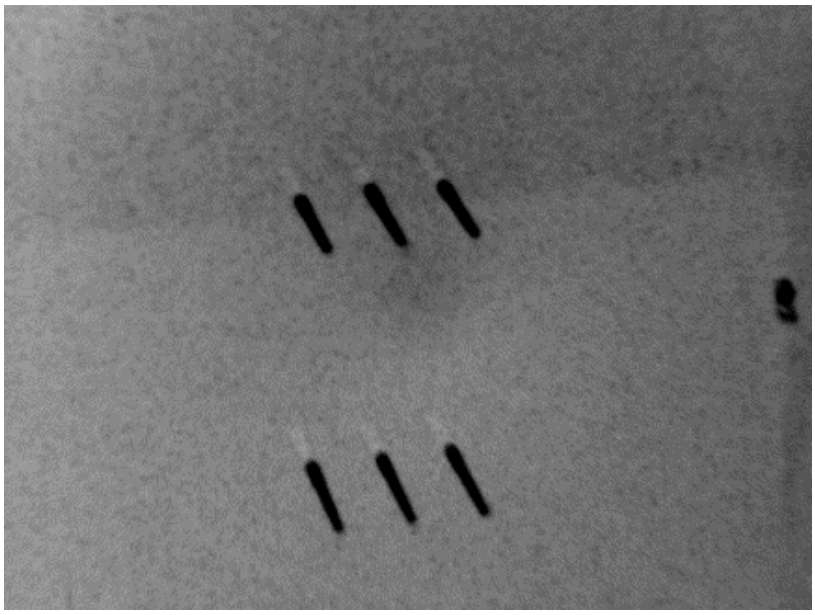
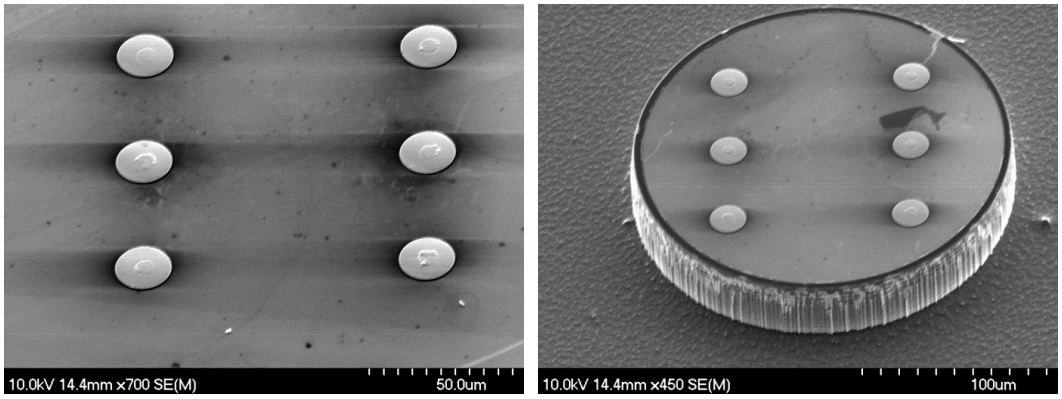


Figure 5.7: X-ray image showing void-free filling of the TSVs.



(a) (b)

Figure 5.8: SEM images of the fabricated test-bed. (a) TSVs in bulk Si. Slight non-planarities from the polishing process visible. (b) TSVs in a 55 μm tall micropin-fin. TSV height of 155 μm

The via dimensions were measured by a combination of cross-sectional SEM, profilometry of the backside via -reveal trench, and drop-gauge characterization of the wafer

thickness. A mean via diameter of 6.4 μm and height of 155 μm, corresponding to an approximately 24:1 aspect ratio via were measured, as shown in Figure 5.9. The taper from the etch process is caused due to the reducing etch characteristic of the Bosch process as the via gets deeper and could be optimized by ramping up the etch for deeper vias.

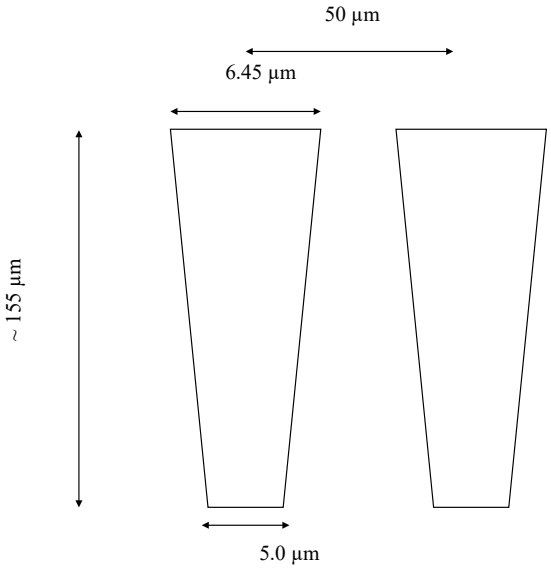


Figure 5.9: Measured geometric dimensions of the fabricated via, showing the taper from the etch process.

5.5 High Frequency Electrical Performance

The electrical performance of the TSVs within the micropin-fin heatsinks were characterized using 3D full-wave HFSS simulations. A Ground Signal Ground (GSG) structure with 50 μm pitch and via dimensions corresponding to the fabrication results from the previous section (6.45 μm diameter, 155 μm height) were used for the simulations. A lumped-port based model was used for these simulations.

Figure 5.10 shows the simulation set-up used for this evaluation. Single port S-parameter simulations were used to characterize the electrical performance. The simulations were performed for both open and short structures and the R, L, G, C parameters were extracted

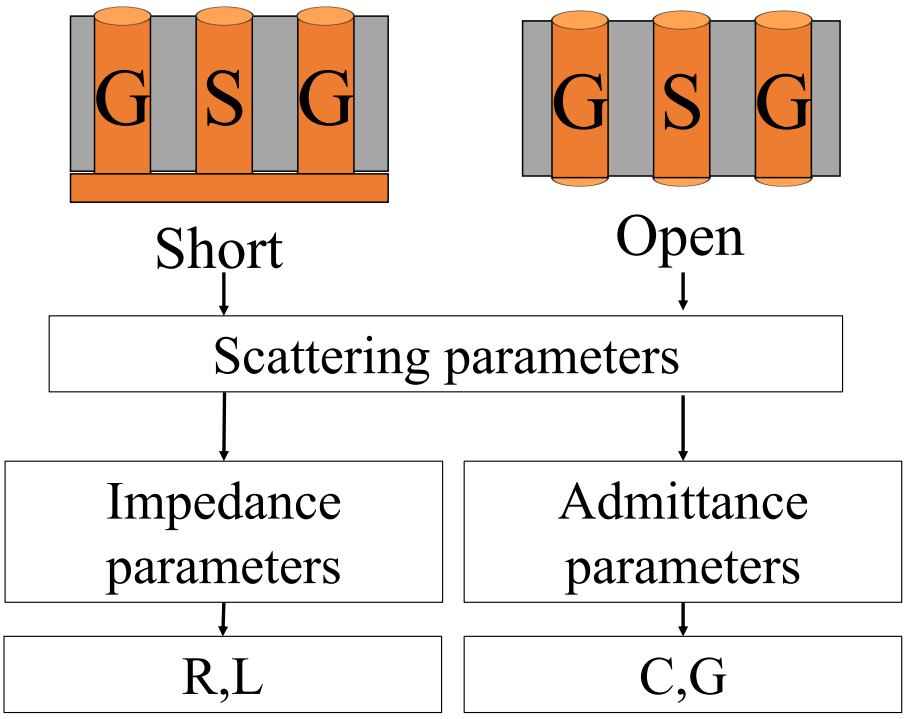


Figure 5.10: Simulation structures and extraction methodologies used to quantify TSV parasitics.

from these using the following sets of equations.

For the capacitance and conductance values, the S parameter (S_{11}) of the open structure was converted to corresponding admittance parameter (Y_{11}) using:

$$Y_{11} = \frac{1}{Z_0}.\frac{1 - S_{11}}{1 + S_{11}} \tag{5.13}$$

where Z_0 is the characteristic impedance. The lumped parasitic capacitance and conductance components of the GSG structure were then extracted using the equations:

$$C = \frac{Im(Y_{11})}{\omega} \tag{5.14}$$

$$G = Re(Y_{11}) \tag{5.15}$$

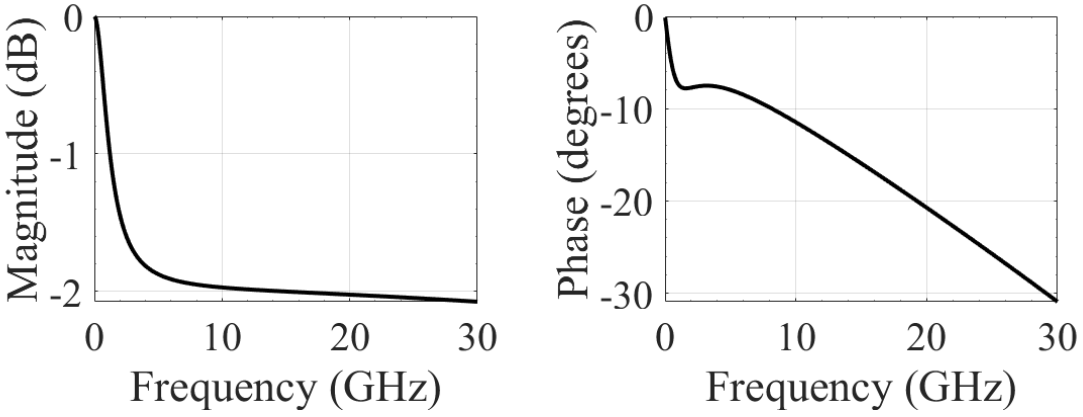
Similarly, the S-parameter of the short structure was converted into impedance parameter (Z_{11}), and the resistance and inductance components were extracted using the following sets of equations:

$$Z_{11} = Z_0 . \frac{1 + S_{11}}{1 - S_{11}} \tag{5.16}$$

$$L = \frac{Im(Z_{11})}{\omega} \tag{5.17}$$

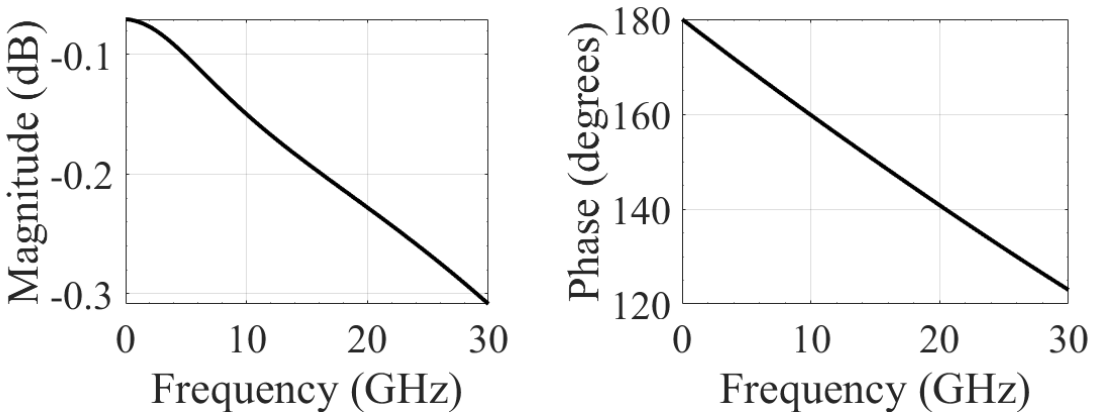
$$R = Re(Z_{11}) \tag{5.18}$$

The magnitude and phase of the S-parameter measurements from the open and short simulations are captured in Figure 5.11 and Figure 5.12, respectively. The extracted parasitic values from these simulations are captured in Figure 5.13. The single port extraction methodology introduces some error in the extracted data, especially at higher frequencies as can be seen in the inductance plot. More details regarding the same is discussed in the Future Work section.



(a) (b)

Figure 5.11: Magnitude and phase of S_{11} for open GSG structure.



(a) (b)

Figure 5.12: Magnitude and phase of S_{11} for short GSG structure.

5.5.1 Effect of Coolant on TSV Properties.

The effect of embedding the TSVs inside a micropin-fin heatsink was evaluated using simulations. A micropin-fin structure of 200 μm diameter and 55 μm height with the signal pin's centre coinciding with the micropin-fin center was used for the simulation. The results from the simulation are shown in Figure 5.14 and compared with the values from bulk Si. As can be seen from the data, there is no significant effect of the micropin-fin etch on the TSV properties. Embedding the signal and return paths within the same micropinfin, thereby shielding the field lines within the Si bulk itself, helps isolate the TSVs electrical properties from the surrounding fluid.

This is in contrast to other approaches where the signal and return TSVs reside in different micropin-fins, thereby getting affected by the surrounding fluid. The smaller via diameters demonstrated in this chapter helps pack more vias within the same micropin-fin region, thereby making such an approach easier.

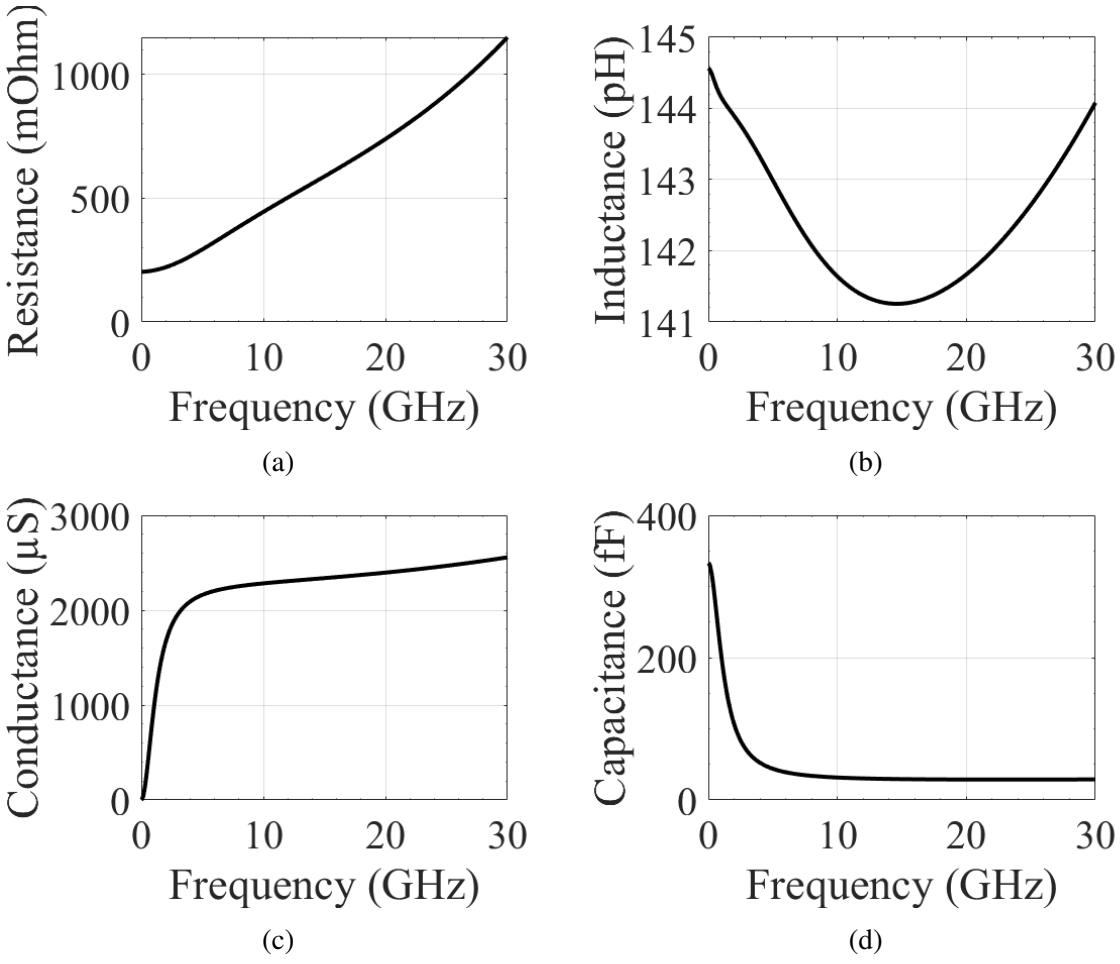


Figure 5.13: Extracted parasitics (a) Parasitic resistance, (b) Parasitic inductance, (c) Parasitic conductance, (d) Parasitic capacitance.

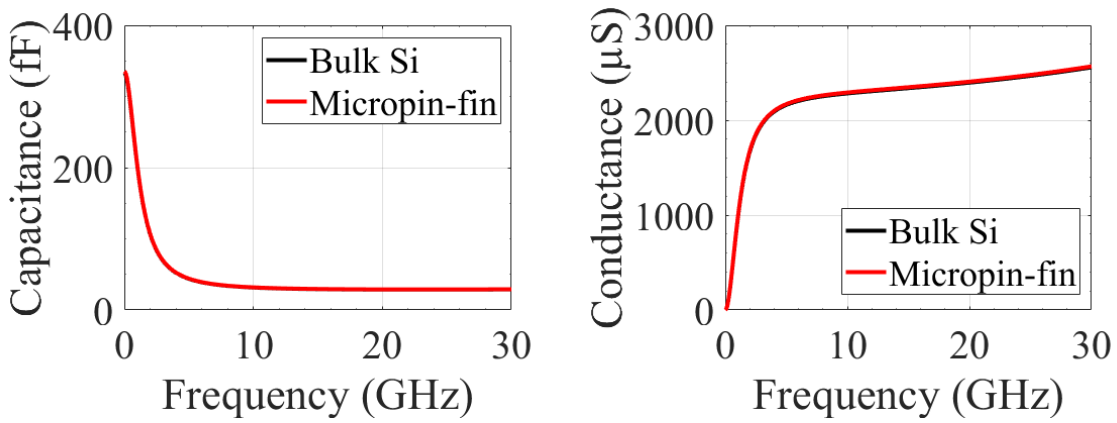


Figure 5.14: Effect of coolant on electrical parasitics. Comparison of (a) Capacitance, and (b) Conductance in bulk Si and a 200 μm diameter micropin-fin heatsink.

5.6 Conclusion

In this chapter, we demonstrated an optimized process flow for the integration of 6.5μm diameter high aspect ratio (24:1) TSVs in a micropin-fin heatsink. The Bosch etching process was optimized to selectively etch high aspect-ratio vias with smooth sidewalls using a photoresist mask. This approach greatly simplifies the fabrication flow compared to other similar demonstrations existing in literature. Void free TSV plating was done using a modified bottom-up plating approach, and was verified using cross-sectional SEM as well as X-ray imaging.

Furthermore, the electrical properties of the TSVs in bulk Si as well as in micropin-fin heatsinks were characterized using single port full-wave simulations. Electrical parasitics from DC to 30GHz were extracted and shows that embedding both signal and return paths within the same micropin-fin can help mitigate the effect of surrounding fluid on the TSV electrical performance.

CHAPTER 6
SUMMARY AND FUTURE WORK

The chapter describes the summary of key research contributions as well as the potential future activities extending the work presented in this thesis.

6.1 Summary of the Work

In this thesis, various thermal and interconnection technologies that can help mitigate the cooling and interconnection challenges in heterogeneous ICs were proposed, experimentally demonstrated, and characterized. The following are the major contributions presented in this thesis:

6.1.1 Selective Electroless Plating for Die-to-Die interconnects

We demonstrate a scalable platform to use selective metal electroless deposition as an alternate approach to create die-to-die interconnects. This approach helps bypass the scalability limits of solder based interconnects and the stringent process requirements for Cu-Cu solid-state diffusion based bonds. While electroless interconnects had been previously proposed in the literature, we proposed and demonstrated the use of mechanical self-alignment technologies in conjunction with this technology to create a much more scalable platform. An area-array with 50 μm pitch was demonstrated. The alignment tolerances as well as the pillar height used for the demonstration provides a path for scaling to sub-10 μm pitch regimes.

The proposed interconnect was also bench-marked against conventional techniques like solder-capped Cu microbumps using HFSS simulations. The superior electrical performance, owing to the small form-factor of the interconnect, was demonstrated.

6.1.2 Evaluation of Performance Benefits of Microfluidic Cooling

This work presents a monolithic microfluidic heatsink implemented on an off-the-shelf Intel 8700K CPU. Use of 3D printed manifolds to control fluid delivery, and to create a very low-profile heatsink, was evaluated. This demonstration scaled the existing literature on monolithic microfluidic cooling to a functional device running real-world benchmarks. Thermal, compute as well as cooling overhead measurements were used to compare this approach with other commonly used techniques like air-cooling, cold-plates as well as two-phase immersion-cooling. We demonstrated the suitability of monolithic microfluidic cooling for the steady state and thermal loads associated with modern CPUs and showed the reduction in cooling overheads that can be achieved by this technique.

6.1.3 Scaling Monolithic Microfluidic Cooling to 2.5D ICs

This involves the first demonstration of monolithic microfluidic cooling on a functional 2.5D device. The cooling efficacy of different dice within the package was tailored to match with their corresponding power dissipation values. The 3D printed manifold was also used to separate the coolant flow paths for different dice. These design optimizations helps reduce the thermal coupling between the dice in addition to reducing the overall thermal resistance.

6.1.4 Scaling Monolithic Microfluidic Cooling to 3D ICs: TSV Co-optimization

In this task, we present an optimized fabrication flow for fabricating a 24: 1 aspect ratio TSV with a via opening of 6.4 μm. This represents almost 50 $\%$ reduction in the via diameter, with similar aspect ratio when compared to previous demonstrations in literature. The via scaling helps improve the interconnect density and the electrical parasitics of the via without compromising on the thermal performance.

6.2 Future Work

The enabling technologies demonstrated in this thesis can be advanced into different avenues. Few of the potential options that directly relate to the discussion already presented in this thesis are summarized here.

6.2.1 Scaling Electroless Interconnects to Lower Pitches

As discussed in Chapter 2, the lateral and vertical alignment limits of the developed testbed allows for scaling the interconnect pitch to sub-10 μm values without the need for redesigning the process-flow. However, this has to be experimentally verified to rule-out any potential issues that may result from the higher interconnect density as we scale down.

This can also be coupled with more detailed electrical characterization of the fabricated structures. This includes (1) Kelvin resistance measurements of individual interconnects, (2) Kelvin resistance measurements for daisy chain of interconnects of varying lengths to ascertains yield and uniformity data, and (3) high frequency measurements to extract the frequency dependent RLGC parasitics.

6.2.2 Fluid Delivery Manifold Optimization to Improve Thermal Performance

The advancements in 3D printing provides the ability to create highly specialized fluid flow paths within the fluid delivery manifold. This provides an extra design vector that can be leveraged to improve the performance of the monolithic microfluidic heatsinks discussed in this thesis. A first order design optimization provided by this flexibility was used in this demonstration to optimize the port placement based on the die dimensions as well as the power-maps.

This however, can be expanded to include evaluations like port splitting, various flow paths based on the underlying powermaps etc. These approaches can help reduce the thermal resistance as well as the pumping power by optimizing the fluid flow, thereby improv-

ing the overall efficacy and efficiency of the heatsink.

6.2.3 Detailed Characterization of TSV Testbed

The single-port electrical characterization used in this thesis just serves to provide only the basic electrical data about these interconnects. Further, the lumped circuit models used to extract the via parasitics tend to loose accuracy at higher frequencies. Therefore, there is a need to have more detailed characterization of these TSVs using two-port measurement and de-embedding techniques. Further, the testbed should be modified to include simultaneous thermal and electrical measurements, while being cooled to more accurately represent the intended testcase. This approach would help significantly improve the understanding of interlayer cooling on both electrical and thermal performance.

REFERENCES

[1] S. Salahuddin, K. Ni, and S. Datta, "The era of hyper-scaling in electronics," *Nature Electronics*, vol. 1, no. 8, pp. 442–450, 2018.

[2] G. E. Moore *et al.*, "Cramming more components onto integrated circuits," 1965.

[3] C. Mack, "The multiple lives of moore's law," *IEEE Spectrum*, vol. 52, no. 4, pp. 31–31, Apr. 2015.

[4] K. Flamm, "Measuring moore's law: Evidence from price, cost, and quality indexes," 2018.

[5] "What's the future of technology scaling?." Available: https://www.sigarch.org/whats-the-future-of-technology-scaling/, Accessed 01-March-2022.

[6] S. Naffziger *et al.*, "Pioneering chiplet technology and design for the amd epyc™ and ryzen™ processor families: Industrial product," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, IEEE, 2021, pp. 57–70.

[7] "Cerebras wafer scale engine: An introduction." Available: https://www.cerebras.net/wp-content/uploads/2019/08/Cerebras-Wafer-Scale-Engine-An-Introduction.pdf, Accessed 07-November-2019.

[8] "The five technical challenges cerebras overcame in building the first trillion-transistor chip," Available: https://techcrunch.com/2019/08/19/the-five-technical-challenges-cerebras-overcame-in-building-the-first-trillion-transistor-chip/, Accessed 07-November-2019.

[9] P. V. Zant, *Microchip fabrication*. McGraw-Hill Education, 2014.

[10] S. Hou *et al.*, "Wafer-level integration of an advanced logic-memory system through the second-generation cowos technology," *IEEE Transactions on Electron Devices*, vol. 64, no. 10, pp. 4071–4077, 2017.

[11] L. T. Su, S. Naffziger, and M. Papermaster, "Multi-chip technologies to unleash computing performance gains over the next decade," in *2017 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2017, pp. 1–1.

[12] "Enabling next-generation platforms using intel's 3D system-in-package technology," Available: https://www.intel.com/content/dam/www/programmable/us/en/pdfs/literature/wp/wp-01251-enabling-nextgen-with-3D-system-in-package.pdf, Accessed 03-March-2022.

[13] "TeraPHY: A chiplet technology for low-power, high-bandwidth in-package optical i/o," Available: https://newsroom.intel.com/wp-content/uploads/sites/11/2019/08/TeraPHY-HotChips-presentation.pdf, Accessed 07-November-2019.

[14] S. Raman, C. Dohrman, T. Chang, and J. Rodgers, "The darpa diverse accessible heterogeneous integration (dahi) program: Towards a next-generation technology platform for high-performance microsystems," in *Proc. 23rd–26th, CS MANTECH Conf*, 2012, pp. 10–13.

[15] "Heterogeneous Integration Roadmap - Chapter 4: Medical, health and wearables," Available: https://eps.ieee.org/images/files/HIR_2019/HIR1_ch04_health.pdf, Accessed 07-November-2019.

[16] "Heterogeneous Integration Roadmap - Chapter 2: High Performance Computing and data centers," Available: https://eps.ieee.org/images/files/HIR_2021/ch02_hpc.pdf, Accessed 10-March-2022.

[17] A. Mastroianni *et al.*, "Proposed standardization of heterogenous integrated chiplet models," in *2021 IEEE International 3D Systems Integration Conference (3DIC)*, IEEE, 2021, pp. 1–8.

[18] "Odsa cost models," Available: https://drive.google.com/drive/folders/1kphneR4UElaZTmnOYgh
b/, Accessed 01-March-2022.

[19] S. Abdennadher, "Testing inter-chiplet communication interconnects in a disaggregated soc design," in *2021 IEEE International Conference on Design & Test of Integrated Micro & Nano-Systems (DTS)*, IEEE, 2021, pp. 1–7.

[20] M. Hutner, R. Sethuram, B. Vinnakota, D. Armstrong, and A. Copperhall, "Special session: Test challenges in a chiplet marketplace," in *2020 IEEE 38th VLSI Test Symposium (VTS)*, IEEE, 2020, pp. 1–12.

[21] "Heterogeneous Integration Roadmap - Chapter 22: Interconnects for 2D and 3D Architechtures," Available: https://eps.ieee.org/images/files/HIR_2021/ch22_2D-3D.pdf, Accessed 10-March-2022.

[22] "IRDS 2020: Packaging integration," Available: https://irds.ieee.org/editions/2020/packaging-integration, Accessed 18-August-2021.

[23] E. Beyne, D. Milojevic, G. Van der Plas, and G. Beyer, "3d soc integration, beyond 2.5d chiplets," in *2021 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2021, pp. 3–6.

[24] Y. Chen *et al.*, "Ultra high density soic with sub-micron bond pitch," in *2020 IEEE 70th Electronic Components and Technology Conference (ECTC)*, IEEE, 2020, pp. 576–581.

[25] "IRDS 2020: Systems and architectures," Available: https://irds.ieee.org/editions/2020/systems-and-architectures, Accessed 18-August-2021.

[26] A. Kanduri, A. M. Rahmani, P. Liljeberg, A. Hemani, A. Jantsch, and H. Tenhunen, "A perspective on dark silicon," in *The Dark Side of Silicon*, Springer, 2017, pp. 3–20.

[27] J. Liu, B. Jaiyen, Y. Kim, C. Wilkerson, and O. Mutlu, "An experimental study of data retention behavior in modern dram devices: Implications for retention time profiling mechanisms," *ACM SIGARCH Computer Architecture News*, vol. 41, no. 3, pp. 60–71, 2013.

[28] M. Dayarathna, Y. Wen, and R. Fan, "Data center energy consumption modeling: A survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 732–794, 2015.

[29] N. Jones, "How to stop data centres from gobbling up the world's electricity," *Nature*, vol. 561, no. 7722, pp. 163–167, 2018.

[30] T. Persoons and J. A. Weibel, "Foreword: Special section on data center cooling," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 7, no. 8, pp. 1189–1190, 2017.

[31] A. Agarwal, S. Mukhopadhyay, A. Raychowdhury, K. Roy, and C. H. Kim, "Leakage power analysis and reduction for nanoscale circuits," *IEeE Micro*, vol. 26, no. 2, pp. 68–80, 2006.

[32] A. Guler and N. K. Jha, "Mcpat-monolithic: An area/power/timing architecture modeling framework for 3-d hybrid monolithic multicore systems," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 28, no. 10, pp. 2146–2156, 2020.

[33] F. M. d'Heurle, "Electromigration and failure in electronics: An introduction," *Proceedings of the IEEE*, vol. 59, no. 10, pp. 1409–1418, 1971.

[34] D. Young and A. Christou, "Failure mechanism models for electromigration," *IEEE Transactions on Reliability*, vol. 43, no. 2, pp. 186–192, 1994.

[35] A. M. Yassine, H. Nariman, M. McBride, M. Uzer, and K. R. Olasupo, "Time dependent breakdown of ultrathin gate oxide," *IEEE Transactions on Electron Devices*, vol. 47, no. 7, pp. 1416–1420, 2000.

[36] E. Wu *et al.*, "Interplay of voltage and temperature acceleration of oxide breakdown for ultra-thin gate oxides," *Solid-State Electronics*, vol. 46, no. 11, pp. 1787–1798, 2002.

[37] W. Wahby, L. Zheng, Y. Zhang, and M. S. Bakir, "A simulation tool for rapid investigation of trends in 3-dic performance and power consumption," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 6, no. 2, pp. 192–199, 2016.

[38] D. B. Tuckerman and R. F. W. Pease, "High-performance heat sinking for vlsi," *IEEE Electron Device Letters*, vol. 2, no. 5, pp. 126–129, 1981.

[39] D. B. Tuckerman, "Heat-transfer microstructures for integrated circuits (cooling, heat-sink)," Ph.D. dissertation, Stanford University, 1984.

[40] T. E. Sarvey *et al.*, "Monolithic integration of a micropin-fin heat sink in a 28-nm FPGA," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 7, no. 10, pp. 1617–1624, 2017.

[41] T. Brunschwiler *et al.*, "Interlayer cooling potential in vertically integrated packages," *Microsystem Technologies*, vol. 15, no. 1, pp. 57–74, 2009.

[42] F. Li, A. W. Owens, and Q. Li, "Microbump processing for 3d ic integration," in *Additional Conferences (Device Packaging, HiTEC, HiTEN, & CICMT)*, International Microelectronics Assembly and Packaging Society, vol. 2019, 2019, pp. 001 028–001 049.

[43] Y. Wang, I. M. De Rosa, and K. Tu, "Size effect on ductile-to-brittle transition in cu-solder-cu micro-joints," in *2015 IEEE 65th Electronic Components and Technology Conference (ECTC)*, IEEE, 2015, pp. 632–639.

[44] "Advanced packaging- future challenges," Available: https://www.techcet.com/wp-content/uploads/2016/05/Session-I_Ramalingam_CMCconf2016.pdf, Accessed 20-November-2021.

[45] W. Koh, B. Lin, and J. Tai, "Copper pillar bump technology progress overview," in *2011 12th International Conference on Electronic Packaging Technology and High Density Packaging*, IEEE, 2011, pp. 1–5.

[46] S. Jangam *et al.*, "Fine-pitch ($\leq$ 10 μm) direct cu-cu interconnects using in-situ formic acid vapor treatment," in *2019 IEEE 69th Electronic Components and Technology Conference (ECTC)*, IEEE, 2019, pp. 620–627.

[47] L. Peng, H. Li, D. F. Lim, S. Gao, and C. S. Tan, "High-density 3-d interconnect of cu–cu contacts with enhanced contact resistance by self-assembled mono-

layer (sam) passivation," *IEEE Transactions on Electron Devices*, vol. 58, no. 8, pp. 2500–2506, 2011.

[48] T. Kim, M. Howlader, T. Itoh, and T. Suga, "Room temperature cu–cu direct bonding using surface activated bonding method," *Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films*, vol. 21, no. 2, pp. 449–453, 2003.

[49] C. Okoro, R. Agarwal, P. Limaye, B. Vandevelde, D. Vandepitte, and E. Beyne, "Insertion bonding: A novel cu-cu bonding approach for 3d integration," in *2010 Proceedings 60th Electronic Components and Technology Conference (ECTC)*, IEEE, 2010, pp. 1370–1375.

[50] C.-M. Liu *et al.*, "Low-temperature direct copper-to-copper bonding enabled by creep on (111) surfaces of nanotwinned cu," *Scientific reports*, vol. 5, p. 9734, 2015.

[51] A. Jourdain, P. Soussan, B. Swinnen, and E. Beyne, "Electrically yielding collective hybrid bonding for 3d stacking of ics," in *2009 59th Electronic Components and Technology Conference*, IEEE, 2009, pp. 11–13.

[52] G. Gao *et al.*, "Development of low temperature direct bond interconnect technology for die-to-wafer and die-to-die applications-stacking, yield improvement, reliability assessment," in *2018 International Wafer Level Packaging Conference (IWLPC)*, IEEE, 2018, pp. 1–7.

[53] J. H. Lau, "Recent advances and trends in advanced packaging," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, 2022.

[54] A. Agrawal, S. Huang, G. Gao, L. Wang, J. DeLaCruz, and L. Mirkarimi, "Thermal and electrical performance of direct bond interconnect technology for 2.5d and 3D integrated circuits," in *IEEE 67th Elec. Comp. and Tech. Conf. (ECTC)*, May 2017, pp. 989–998.

[55] M.-J. Li *et al.*, "Cu–cu bonding using selective cobalt atomic layer deposition for 2.5-d/3-d chip integration technologies," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 10, no. 12, pp. 2125–2128, 2020.

[56] S. K. Rajan, M. J. Li, M. S. Bakir, and G. S. May, "High density and low-temperature interconnection enabled by mechanical self-alignment and electroless plating," in *2019 International 3D Systems Integration Conference (3DIC)*, IEEE, 2019, pp. 1–4.

[57] T. Osborn, A. He, N. Galiba, and P. A. Kohl, "All-copper chip-to-substrate interconnects part i. fabrication and characterization," *Journal of The Electrochemical Society*, vol. 155, no. 4, pp. D308–D313, 2008.

[58] C. Wu *et al.*, "Sub-micron electrical interconnection enabled ultra-high i/o density wafer level sip integration," in *2017 IEEE 67th Electronic Components and Technology Conference (ECTC)*, IEEE, 2017, pp. 1231–1236.

[59] S. Yang, H. Hung, P. Wu, Y. Wang, H. Nishikawa, and C. Kao, "Materials merging mechanism of microfluidic electroless interconnection process," *Journal of The Electrochemical Society*, vol. 165, no. 7, pp. D273–D281, 2018.

[60] J. L. Gonzalez, "Heterogeneous integration of chiplets using socketed platforms, off-chip flexible interconnects, and self-alignment technologies," Ph.D. dissertation, Georgia Institute of Technology, 2021.

[61] I. Shubin *et al.*, "Optical proximity communication," in *Optoelectronic Integrated Circuits XI*, L. A. Eldada and E.-H. Lee, Eds., International Society for Optics and Photonics, vol. 7219, SPIE, 2009, pp. 11–18.

[62] I. Ohno, "Electrochemistry of electroless plating," *Materials Science and Engineering: A*, vol. 146, no. 1-2, pp. 33–49, 1991.

[63] S. S. Singh, P. Pal, A. K. Pandey, Y. Xing, and K. Sato, "Determination of precise crystallographic directions for mask alignment in wet bulk micromachining for mems," *Micro and Nano Systems Letters*, vol. 4, no. 1, pp. 1–29, 2016.

[64] "Electroless nickel plating strike," Available: https://transene.com/ni-strike/, Accessed 17-September-2019.

[65] "Ball grade tolerances and terminology," Available: https://www.swissjewel.com/wp-content/uploads/2017/10/Ball-Grade-Sheet.pdf, Accessed 17-September-2019.

[66] "Nickelex: Improved electroless nickel plating solution ammonia free," Available: https://transene.com/ni/, Accessed 17-September-2019.

[67] T. Yokoshima, Y. Yamaji, K. Kikuchi, H. Nakagawa, and M. Aoyagi, "Anisotropic deposition of localized electroless nickel for preferential bridge connection," *Journal of The Electrochemical Society*, vol. 157, no. 1, pp. D65–D73, 2010.

[68] Z. Xu, X. Gu, and J.-Q. Lu, "Parasitics extraction, wideband modeling and sensitivity analysis of through-strata-via (TSV) in 3d integration/packaging," in *2011 IEEE/SEMI Advanced Semiconductor Manufacturing Conference*, IEEE, 2011, pp. 1–6.

[69] A. Bar-Cohen, M. Arik, and M. Ohadi, "Direct liquid cooling of high flux micro and nano electronic components," *Proceedings of the IEEE*, vol. 94, no. 8, pp. 1549–1570, 2006.

[70] A. Bar-Cohen, "Gen 3 "embedded" cooling: Key enabler for energy efficient data centers," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 7, no. 8, pp. 1206–1211, 2017.

[71] F. P. Incropera, *Liquid cooling of electronic devices by single-phase convection*. Wiley-Interscience, 1999, vol. 3.

[72] A. Mohammed Adham, N. Mohd-Ghazali, and R. Ahmad, "Thermal and hydrodynamic analysis of microchannel heat sinks: A review," *Renewable and Sustainable Energy Reviews*, vol. 21, pp. 614–622, 2013.

[73] R. J. Phillips *et al.*, "Microchannel heat sinks," *The Lincoln Laboratory Journal*, vol. 1, no. 1, pp. 31–48, 1988.

[74] W. Qu and I. Mudawar, "Analysis of three-dimensional heat transfer in microchannel heat sinks," *International Journal of heat and mass transfer*, vol. 45, no. 19, pp. 3973–3985, 2002.

[75] C. Zhao and T. Lu, "Analysis of microchannel heat sinks for electronics cooling," *International Journal of Heat and Mass Transfer*, vol. 45, no. 24, pp. 4857–4869, 2002.

[76] R. S. Prasher *et al.*, "Nusselt number and friction factor of staggered arrays of low aspect ratio micropin-fins under cross flow for water as fluid," 2007.

[77] D. Copeland, "Manifold microchannel heat sinks: Numerical analysis," *American Society of Mechanical Engineers(Paper). 6 ppp. 1995.*, p. 6, 1995.

[78] D. Mundinger *et al.*, "Demonstration of high-performance silicon microchannel heat exchangers for laser diode array cooling," *Applied physics letters*, vol. 53, no. 12, pp. 1030–1032, 1988.

[79] D. Liu and S. V. Garimella, "Analysis and optimization of the thermal performance of microchannel heat sinks," in *International Electronic Packaging Technical Conference and Exhibition*, vol. 36908, 2003, pp. 557–565.

[80] M. I. Hasan, "Investigation of flow and heat transfer characteristics in micro pin fin heat sink with nanofluid," *Applied thermal engineering*, vol. 63, no. 2, pp. 598–607, 2014.

[81] A. Koşar and Y. Peles, "Boiling heat transfer in a hydrofoil-based micro pin fin heat sink," *International Journal of Heat and Mass Transfer*, vol. 50, no. 5-6, pp. 1018–1034, 2007.

[82] T. Yeom, T. Simon, T. Zhang, M. Zhang, M. North, and T. Cui, "Enhanced heat transfer of heat sink channels with micro pin fin roughened walls," *International Journal of Heat and Mass Transfer*, vol. 92, pp. 617–627, 2016.

[83] M. Liu, D. Liu, S. Xu, and Y. Chen, "Experimental study on liquid flow and heat transfer in micro square pin fin heat sink," *International Journal of Heat and Mass Transfer*, vol. 54, no. 25-26, pp. 5602–5611, 2011.

[84] J.-N. Hung *et al.*, "Advanced system integration for high performance computing with liquid cooling," in *2021 IEEE 71st Electronic Components and Technology Conference (ECTC)*, IEEE, 2021, pp. 105–111.

[85] T. Wei *et al.*, "High-efficiency polymer-based direct multi-jet impingement cooling solution for high-power devices," *IEEE Transactions on Power Electronics*, vol. 34, no. 7, pp. 6601–6612, 2018.

[86] C. Glynn, T. O'Donovan, D. B. Murray, and M. Feidt, "Jet impingement cooling," in *Proceedings of the 9th UK National Heat Transfer Conference*, 2005, pp. 5–6.

[87] M. K. Sung and I. Mudawar, "Single-phase hybrid micro-channel micro-jet impingement cooling," *International Journal of Heat and Mass Transfer*, vol. 51, no. 17-18, pp. 4342–4352, 2008.

[88] R. van Erp, R. Soleimanzadeh, L. Nela, G. Kampitsis, and E. Matioli, "Co-designing electronics with microfluidics for more sustainable cooling," *Nature*, vol. 585, no. 7824, pp. 211–216, 2020.

[89] P. Dudek, "FDM 3D printing technology in manufacturing composite elements," *Archives of Metallurgy and Materials*, vol. 58, no. 4, pp. 1415–1418, 2013.

[90] Y. Li, B. S. Linke, H. Voet, B. Falk, R. Schmitt, and M. Lam, "Cost, sustainability and surface roughness quality–a comprehensive analysis of products made with personal 3d printers," *CIRP Journal of Manufacturing Science and Technology*, vol. 16, pp. 1–11, 2017.

[91] "Selecting right 3D printing process — 3D Hubs," Available: https://www.3Dhubs.com/knowledge-base/selecting-right-3D-printing-process, Accessed 03-March-2019.

[92] "FDM 3D printing — 3D Hubs," Available: https://www.3Dhubs.com/3D-printing/processes/, Accessed 03-March-2019.

[93] A. Pirjan and D.-M. Petrosanu, "The impact of 3D printing technology on the society and economy," *Journal of Information Systems & Operations Management*, vol. 8, no. 2, 2013.

[94] M. Jalili *et al.*, "Cost-efficient overclocking in immersion-cooled datacenters," in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, IEEE, 2021, pp. 623–636.

[95] B. Ramakrishnan *et al.*, "CPU overclocking: A performance assessment of air, cold plates, and two-phase immersion cooling," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, pp. 1–1, 2021.

[96] Z. Wan and Y. Joshi, "Pressure drop and heat transfer characteristics of pin fin enhanced microgaps in single phase microfluidic cooling," *International Journal of Heat and Mass Transfer*, vol. 115, pp. 115–126, 2017.

[97] D. Lorenzini *et al.*, "Embedded single phase microfluidic thermal management for non-uniform heating and hotspots using microgaps with variable pin fin clustering," *International Journal of Heat and Mass Transfer*, vol. 103, pp. 1359–1370, 2016.

[98] "Intel core i7-8700k," Available: https://en.wikichip.org/wiki/intel/corei7/i7-8700K, Accessed 18-August-2021.

[99] T. E. Sarvey *et al.*, "Integrated circuit cooling using heterogeneous micropin-fin arrays for nonuniform power maps," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 7, no. 9, pp. 1465–1475, 2017.

[100] B. Whitehead, D. Andrews, A. Shah, and G. Maidment, "Assessing the environmental impact of data centres part 1: Background, energy use and metrics," *Building and Environment*, vol. 82, pp. 151–159, 2014.

[101] R. Mahajan, "Quiet revolutions: How advanced microelectronics packaging continues to drive heterogeneous integration," in *2020 19th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, IEEE, 2020, pp. 1408–1412.

[102] Y. Zhang, X. Zhang, W. Wahby, and M. S. Bakir, "Design considerations for 2.5-d and 3-d integration accounting for thermal constraints," in *2016 IEEE International 3D Systems Integration Conference (3DIC)*, IEEE, 2016, pp. 1–5.

[103] K. Sohn *et al.*, "A 1.2 v 20 nm 307 gb/s hbm dram with at-speed wafer-level io test scheme and adaptive refresh considering temperature distribution," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 250–260, 2016.

[104] S. K. Rajan, A. Kaul, T. E. Sarvey, G. S. May, and M. S. Bakir, "Monolithic microfluidic cooling of a heterogeneous 2.5-d fpga with low-profile 3-d printed manifolds," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 11, no. 6, pp. 974–982, 2021.

[105] S. K. Rajan, A. Kaul, G. S. May, and M. S. Bakir, "Electrical and performance benefits of advanced monolithic cooling for 2.5 d heterogeneous ics," in *2021 IEEE International 3D Systems Integration Conference (3DIC)*, IEEE, 2021, pp. 1–5.

[106] T. Burd *et al.*, ""zeppelin": An soc for multichip architectures," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 1, pp. 133–143, 2019.

[107] "Amd epyc™ processors," Available: https://www.amd.com/en/processors/epyc-server-cpu-family, Accessed 11-October-2021.

[108] "Naples - cores - amd," Available: https://en.wikichip.org/wiki/amd/cores/naples, Accessed 11-October-2021.

[109] "Thermal design guide for socket sp3 processors," Available: http://developer.amd.com/wordpress/media/2013/12/ThermalGuideforSP3.pdf, Accessed 11-October-2021.

[110] Y. Zhang, X. Zhang, and M. S. Bakir, "Benchmarking digital die-to-die channels in 2.5-D and 3-D heterogeneous integration platforms," *IEEE transactions on Electron Devices*, vol. 65, no. 12, pp. 5460–5467, Dec. 2018.

[111] A. Koşar, C. Mishra, and Y. Peles, "Laminar flow across a bank of low aspect ratio micro pin fins," 2005.

[112] T. E. Sarvey, A. Kaul, S. K. Rajan, A. Dasu, R. Gutala, and M. S. Bakir, "Microfluidic cooling of a 14-nm 2.5-d fpga with 3-d printed manifolds for high-density computing: Design considerations, fabrication, and electrical characterization," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 9, no. 12, pp. 2393–2403, 2019.

[113] "Visijet 3D crystal material data sheet," Available: https://www.shapeways.com/wp-content/uploads/sites/3/2020/12/Material-Data-Sheet-FUD.pdf, Accessed 18-March-2021.

[114] S. Zimmermann, M. K. Tiwari, I. Meijer, S. Paredes, B. Michel, and D. Poulikakos, "Hot water cooled electronics: Exergy analysis and waste heat reuse feasibility," *International Journal of Heat and Mass Transfer*, vol. 55, no. 23-24, pp. 6391–6399, 2012.

[115] Y. Hu and Y. Joshi, "Single-phase microfluidic cooling of 2.5d-sics for heterogeneous integration," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 10, no. 9, pp. 1499–1506, 2020.

[116] T. W. Wei, H. Oprins, V. Cherman, E. Beyne, and M. Baelmans, "Experimental and numerical investigation of direct liquid jet impinging cooling using 3d printed man-

ifolds on lidded and lidless packages for 2.5d integrated systems," *Applied Thermal Engineering*, vol. 164, p. 114 535, 2020.

[117] H. Oh, J. M. Gu, S. J. Hong, G. S. May, and M. S. Bakir, "High-aspect ratio through-silicon vias for the integration of microfluidic cooling with 3d microsystems," *Microelectronic Engineering*, vol. 142, pp. 30–35, 2015.

[118] T. Brunschwiler *et al.*, "Forced convective interlayer cooling in vertically integrated packages," in *2008 11th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, IEEE, 2008, pp. 1114–1125.

[119] A. Sridhar, A. Vincenzi, M. Ruggiero, T. Brunschwiler, and D. Atienza, "3d-ice: Fast compact transient thermal modeling for 3d ics with inter-tier liquid cooling," in *2010 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, IEEE, 2010, pp. 463–470.

[120] Y. J. Kim, Y. K. Joshi, A. G. Fedorov, Y.-J. Lee, and S.-K. Lim, "Thermal characterization of interlayer microfluidic cooling of three-dimensional integrated circuits with nonuniform heat flux," 2010.

[121] Y. Zhang, H. Oh, Y. Zhang, L. Zheng, G. S. May, and M. S. Bakir, "Thermal isolation and cooling technologies for heterogeneous 3d-and 2.5 d-ics," *Handbook of 3D Integration: Design, Test, and Thermal Management*, vol. 4, pp. 347–373, 2019.

[122] Y. Zhang, A. Dembla, and M. S. Bakir, "Silicon micropin-fin heat sink with integrated TSVs for 3-d ics: Tradeoff analysis and experimental testing," *IEEE Transactions on Components, packaging and manufacturing technology*, vol. 3, no. 11, pp. 1842–1850, 2013.

[123] J. Cho *et al.*, "Nonlinear effects of TSV and harmonic generation," in *2012 IEEE 62nd Electronic Components and Technology Conference*, IEEE, 2012, pp. 834–838.

[124] J. Kim *et al.*, "High-frequency scalable electrical model and analysis of a through silicon via (TSV)," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 1, no. 2, pp. 181–195, 2011.

[125] I. Savidis and E. G. Friedman, "Closed-form expressions of 3-d via resistance, inductance, and capacitance," *IEEE Transactions on Electron Devices*, vol. 56, no. 9, pp. 1873–1881, 2009.

[126] G. Katti, M. Stucchi, K. De Meyer, and W. Dehaene, "Electrical modeling and characterization of through silicon via for three-dimensional ics," *IEEE transactions on Electron Devices*, vol. 57, no. 1, pp. 256–262, 2009.

[127] C. Xu, R. Suaya, and K. Banerjee, "Compact modeling and analysis of through-silicon-via-induced electrical noise coupling in three-dimensional ics," *IEEE Transactions on Electron Devices*, vol. 58, no. 11, pp. 4024–4034, 2011.

[128] K. Salah, H. Ragai, Y. Ismail, and A. El Rouby, "Equivalent lumped element models for various n-port through silicon vias networks," in *16th Asia and South Pacific Design Automation Conference (ASP-DAC 2011)*, IEEE, 2011, pp. 176–183.

[129] H. Oh, "Silicon microsystem platform with integrated microfluidic cooling and low-loss through-silicon vias," Ph.D. dissertation, Georgia Institute of Technology, 2017.

[130] C. R. King, D. Sekar, M. S. Bakir, B. Dang, J. Pikarsky, and J. D. Meindl, "3d stacking of chips with electrical and microfluidic i/o interconnects," in *2008 58th Electronic Components and Technology Conference*, IEEE, 2008, pp. 1–7.

[131] Y. Madhour *et al.*, "Integration of intra chip stack fluidic cooling using thin-layer solder bonding," in *2013 IEEE International 3D Systems Integration Conference (3DIC)*, IEEE, 2013, pp. 1–8.

[132] N. Khan *et al.*, "3-d packaging with through-silicon via (TSV) for electrical and fluidic interconnections," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 3, no. 2, pp. 221–228, 2013.

[133] M. A. Ehsan, Z. Zhou, L. Liu, and Y. Yi, "An analytical through silicon via (TSV) surface roughness model applied to a millimeter wave 3-d ic," *IEEE Transactions on Electromagnetic Compatibility*, vol. 57, no. 4, pp. 815–826, 2015.

[134] Y. Morikawa, T. Murayama, Y. N. T. Sakuishi, A. Suzuki, and K. Suu, "Total cost effective scallop free si etching for 2.5 d & 3d TSV fabrication technologies in 300mm wafer," in *2013 IEEE 63rd Electronic Components and Technology Conference*, IEEE, 2013, pp. 605–607.

[135] W.-W. Shen and K.-N. Chen, "Three-dimensional integrated circuit (3d ic) key technology: Through-silicon via (TSV)," *Nanoscale research letters*, vol. 12, no. 1, pp. 1–9, 2017.