MICROFLUIDIC COOLING FOR DENSELY INTEGRATED MICROELECTRONIC SYSTEMS

A Dissertation
Presented to
The Academic Faculty

By

Thomas E. Sarvey

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

Georgia Institute of Technology

December 2018

Copyright © Thomas E. Sarvey 2018

MICROFLUIDIC COOLING FOR DENSELY INTEGRATED MICROELECTRONIC SYSTEMS

Approved by:

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

Dr. Arijit Raychowdhury
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. Oliver Brand
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Yogendra K. Joshi
School of Mechanical Engineering
Georgia Institute of Technology

Date Approved: October 31, 2018

Dedicated to my parents,

who fostered the curiosity and patience

which I needed to complete this dissertation.

ACKNOWLEDGEMENTS

I would like to first thank my advisor, Dr. Muhannad Bakir, for the guidance, advice, and support he provided during my PhD. Dr. Bakir's infectious enthusiasm drew me to the area of microsystems research. He fostered an environment where I was free to explore ideas and test theories, propelled by his encouragement and his consistent confidence that I would succeed.

I would also like to thank Dr. Sudha Yalamanchili, Dr. Arijit Raychowdhury, Dr. Bruno Frazier, Dr. Oliver Brand, and Dr. Yogendra Joshi for sharing valuable insight with me and for serving on my PhD committee.

I would like to specifically acknowledge those who directly contributed to parts of the work shown in this thesis. This includes Mohamed Nasr and Reza Abbaspour for their considerable contribution to the work on microgap hotspot cooling, Yuanchen Hu for his help with the experiments with heterogeneous micropin-fins, and Ankit Kaul and Sreejith Rajan for their help with the chapter on cooling of a Stratix 10 FPGA. I would also like to thank those with whom I directly collaborated on work which did not make it into this thesis, including Yue Zhang, Casey Woodrum, Yang Zhang, William Wahby, Xuchen Zhang, Pouya Asrar, Xuefei Han, Craig Green, Daniel Lorenzini, and Peter Kottke. I would also like to thank Professor Yogendra Joshi, Professor Andrei Fedorov, Professor Suresh Sitaraman, Professor Sudha Yalamanchili, and Professor Saibal Mukhopadyay for their collaboration and guidance through multiple DARPA ICECool programs. I would like to thank my collaborators at Altera/Intel including Aravind Dasu, Ravi Gutala, Arif Rahman, and many others who helped with resources and guidance, both as part of our joint DARPA projects, and through my internship at Intel. I am also thankful to DARPA for funding my work throughout my time at Georgia Tech.

I am grateful that I had such good company throughout my PhD and I would like to thank all current and former I3DS group members for their support. I was fortunate to

be in such a collaborative environment, surrounded by humble and knowledgeable people willing to generously share their time and expertise. In addition to those mentioned above, I would like to thank Joe Gonzalez, Opu Hossen, James Yang, Paragkumar Thadesar, Chaoqi Zhang, Li Zheng, Hanju Oh, Muneeb Zia, Congshan Wan, Paul Jo, and Ting Zheng.

I am also thankful to the IEN staff, including, but not limited to, Vinny Nguyen, Gary Spinner, and John Pham, for providing support and fabrication resources which helped me fabricate many of the devices in this dissertation.

Lastly, I would like to thank my friends and family for their support and encouragement throughout the PhD experience. I especially want to thank Marissa Habeshy for her support, sacrifices, and patience which she generously gave so that I could finish the work in this thesis.

TABLE OF CONTENTS

Acknowledgments iv

List of Tables ix

List of Figures x

Acronyms xv

Chapter 1: Introduction 1

1.1 State of the Art Packaging Technologies 4

1.1.1 High Performance Memory Interface Standards 8

1.2 Microfluidic Cooling 9

1.2.1 Microchannel Cooling 11

1.2.2 Micropin-fin Cooling 12

1.2.3 Hotspot Cooling 13

1.2.4 Microfluidic Cooling in 3D Stacks 13

1.3 Research Objectives and Contributions 14

1.4 Organization of this Thesis 15

Chapter 2: Hotspot Cooling Using Dedicated Microgaps 16

2.1 Dedicated Hotspot Microgap 16

2.2 Combined Hotspot and Background Cooling Using a Dedicated Microgap and Micropin-fin Heat Sink 23

2.2.1 Testbed Design and Fabrication 24

2.2.2 Thermal and Hydraulic Results 26

2.3 Conclusions 29

Chapter 3: Integrated Circuit Cooling Using Non-Uniform Micropin-Fin Arrays for Non-Uniform Power Maps 31

3.1 Effect of Heat Spreading 32

3.1.1 Effect of Base Thickness on Hotspot Temperature 33

3.1.2 Measuring Thermal Resistance at the Hotspot 34

3.2 Fabrication of Multiple Micropin-fin Densities on a Single Wafer 36

3.3 Thermal Testbed and Heterogeneous Micropin-fin Samples 39

3.4 Experimental Results 46

3.4.1 Non-Uniform Heat Flux 50

3.5 Conclusion 52

Chapter 4: Monolithic Integration of a Micropin-fin Heat Sink in a 28 nm FPGA 54

4.1 Fabrication 57

4.2 Testing 60

4.2.1 Variable Flow Rate Testing 64

4.2.2 Elevated Inlet Temperature Testing 67

4.2.3 Clock Speed 70

4.2.4 Die Power 70

4.3 Conclusion 72

Chapter 5: Microfluidic Cooling of a 2.5D FPGA 73

5.1 Experimental Setup 75

5.1.1 FPGA Benchmark Application 76

5.1.2 Air Cooled Heat Sink 78

5.2 Heat Sink Design 80

5.3 Heat Sink Simulation 81

5.4 Heat Sink Fabrication and Assembly 84

5.5 Experimental Results 90

5.6 Conclusion 98

Chapter 6: Summary and Future Work 99

6.1 Summary of Work 99

6.2 Future Work 101

6.2.1 Low Profile Heat Sink for 2.5D Packages 101

6.2.2 Micropin-fin Heat Sink Optimization Framework 102

Appendix A: Microfluidic Cooling Test Device Design and Fabrication 112

A.1 Micropin-fin Test Device Fabrication 112

A.2 Micropin-fin test Device Design 116

A.2.1 First generation devices 116

A.2.2 Second generation devices 117

A.2.3 Third generation devices 118

References 126

LIST OF TABLES

1.1 Signaling energy per 32-bit word with various interconnects 3

1.2 Commercially available memory interfaces 9

3.1 Cylindrical micropin-fin dimensions built on a single wafer 38

3.2 Measurement uncertainties for heterogeneous micropin-fin experiments . . 45

4.1 FPGA thermal and power measurements with microfluidic heat sink and air cooled heat sink 62

5.1 Stratix 10 die powers, areas, and power densities 80

5.2 Change in die powers as a function of FPGA die power 96

5.3 Stratix 10 temperature measurements with air cooled heat sink and microfluidic cooled heat sink 97

A.1 Bosch Process Etching Parameters 113

A.2 NR5-8000 Lithography Parameters (for etching) 113

A.3 NR9-1500PY Lithography Parameters (for liftoff) 115

A.4 Metalization Parameters 115

LIST OF FIGURES

1.1 Microprocessor trends over the last 35 years. 2

1.2 Multi-die interconnection methods 5

1.3 Cross section of Xilinx 28 nm FPGA on Interposer 6

1.4 Cross section of an Nvidia Pascal GPU with interposer and HBM 7

1.5 Two processors utilizing an interposer with HBM 8

2.1 Cross-sectional diagram of microgap test device 17

2.2 Microgap test device fabrication process 18

2.3 Package for microgap test device 18

2.4 Experimental setup for microgap test devices 19

2.5 Microgap thermal resistance with mass flux of 3000 kg/(m^2 s) 20

2.6 Microgap thermal resistance with mass flux of 5000 kg/(m^2 s) 20

2.7 Microgap thermal resistance with mass flux of 7000 kg/(m^2 s) 21

2.8 Microgap with stratified flow 21

2.9 Microgap with vapor plume flow 22

2.10 Microgap with ultra thin film flow 22

2.11 Microgap pressure drop 23

2.12 Combined hotspot and background cooling test device diagram 25

2.13 Combined hotspot/background test device 26

2.14 Images of combined microgap/background test chip 26

2.15 Hotspot and background pressure drops 27

2.16 Hotspot temperatue vs. heat flux 29

3.1 COMSOL heat spreading model 33

3.2 Simulated surface averaged hotspot and background temperature rise above ambient vs. silicon base thickness 34

3.3 Simulated surface averaged hotspot and background temperature rise above ambient vs. silicon base thickness with three hotspot heat transfer coefficients 35

3.4 Direction of heat flow in the substrate for different power maps 36

3.5 Effective hotspot thermal resistance (R_{hs}) vs. hotspot heat flux (q_{hs}) 37

3.6 SEM images of (a) sparse and (b) dense, high aspect ratio micropin-fins before and after etching process optimization. 38

3.7 Six micropin-fin dice etched using a single batch process on the same wafer 39

3.8 Heterogeneous micropin-fin test chip cross section 40

3.9 Test chip (a) top view of etched silicon through pyrex cap (b) bottom view of heater and ports 41

3.10 Background micropin-fin dimensions 42

3.11 SEM images of of the four micropin-fin devices 42

3.12 Diagram of open loop system used to test chips 44

3.13 Photo of a packaged heterogeneous micropin-fin test device 45

3.14 Junction temperature rise above inlet temperature vs. axial position in direction of fluid flow with uniform power 47

3.15 Thermal resistance from the silicon to fluid (R_{jf}) vs. flow rate for all four dice 48

3.16 Pressure drop vs. flow rate for all four dice 49

3.17 Junction temperature rise above inlet temperature vs. axial position in direction of fluid flow with 500 W/cm^2 hotspot heat flux and 250 W/cm^2 background heat flux 50

3.18 Hotspot and background temperatures vs. flow rate of all four chips 51

4.1 (a) Traditional microelectronic system (b) Microelectronic system with monolithically integrated microfluidic heat sink 55

4.2 Microfluidic cooling integrated in (a) the interposer and (b) the back side of the die 56

4.3 Fabrication process for etching micropin-fins into the back side of a packaged FPGA die 58

4.4 Image of the etched back side of the silicon FPGA die along with the micropin-fin dimensions 59

4.5 SEM image of micropin-fins etched using the same process in a silicon wafer 59

4.6 Processed FPGA soldered to development board with silicon cap and Nanoports for fluid delivery 60

4.7 Diagram of the open loop system used to test the FPGA 62

4.8 Die temperature vs. die power 64

4.9 Die temperature vs. flow rate 65

4.10 Measured outlet temperature and theoretical outlet temperature (no heat loss) vs. flow rate. 66

4.11 Heat Loss vs. average die temperature. 67

4.12 Pressure drop vs. flow rate. 68

4.13 FPGA temperature range vs. inlet temperature. 68

4.14 Pressure drop vs. inlet temperature. 69

4.15 Maximum FPGA clock speed vs. maximum die temperature 69

4.16 FPGA power vs. average die temperature. 71

5.1 Chiplet microfluidic cooling 74

5.2 Cross-sectional diagram of micropin-fin heat sink for 2.5D package 75

5.3 Stratix 10 ES development board 76

5.4 Stratix 10 GX package 76

5.5 Stratix 10 ES development board with air cooled heat sink 79

5.6 Custom mounting for air cooled heat sink on Stratix 10 development board 79

5.7 Simulated micropin-fin heat sink models 81

5.8 Simulated temperatures of FPGA and transceiver regions with heterogeneous and uniform heat sinks 83

5.9 Silicon micropin-fin heat sink for Stratix 10 FPGA 84

5.10 SEM image of high and low density micropin-fins 85

5.11 Histogram of micropin-fin heat sink heights 86

5.12 Profile of dense micropin-fin base 87

5.13 3D printed enclosure and fluid routing device 88

5.14 Profile of micropin-fin heat sink enclosure 88

5.15 Assembled microfluidic heat sink 89

5.16 Delidded Stratix 10 FPGA on development board 89

5.17 Custom mounting for microfluidic heat sink on Stratix 10 development board 90

5.18 Heterogeneous micropin-fin heat sink die temperatures vs. flow rate 91

5.19 Heterogeneous micropin-fin heat sink pressure drop vs. flow rate 92

5.20 Air cooled heat sink die temperatures vs. FPGA core power 93

5.21 Air cooled heat sink die temperatures vs. transceiver power 94

5.22 Heterogeneous micropin-fin heat sink die temperatures vs. FPGA core power 95

5.23 Heterogeneous micropin-fin heat sink die temperatures vs. transceiver power 96

6.1 Diagram of optimization loop 104

6.2 Example 3×4 cell geometry in FreeCAD 105

6.3 Temperature and pressure vs. optimization iteration with fixed aspect ratio . 107

6.4 Temperature and pressure vs. optimization iteration with fixed aspect ratio and pressure constraint 108

6.5 Optimized micropin-fin heat sink geometry 109

A.1 Cross-sectional diagram of micropin-fin test device 113

A.2 Fabrication flow for micropin-fin test device 114

A.3 First generation micropin-fin test device 116

A.4 Second generation micropin-fin test device 118

A.5 Third generation micropin-fin test device 119

ACRONYMS

AMD Advanced Micro Devices. 7, 8, 78

API application programming interface. 107

BTS board test system. 77, 78

CAD computer-aided design. 107

CMOS complementary metal oxide semiconductor. 1, 14, 15, 56, 102

CPU central processing unit. 5, 8, 78

CSV comma-separated value. 108

DDR double data rate. 8, 9

DRAM dynamic random-access memory. 3, 6, 8

DRIE deep reactive ion etching. 11, 25

DSP double side polished. 25, 114, 116

DSP digital signal processing. 57, 60, 61, 63, 77

EMIB embedded multi-die interconnect bridge. 75

ES engineering silicon. 75, 76, 79

FFT fast fourier transform. 77, 91, 94, 98

FIFO first-in, first-out. 77

FPGA field-programmable gate array. 6, 15, 57–72, 74–82, 84, 86, 90, 91, 94–99, 102

FVM finite volume method. 104

GDDR graphics double data rate. 8, 9

GPU graphics processing unit. 4, 6–8

GUI graphical user interface. 107, 108

HBM High Bandwidth Memory. 7–9

HMC Hybrid Memory Cube. 7

I/O input/output. 3, 9

IC integrated circuit. 3, 12, 16, 29

IP intellectual property. 77, 91

PC personal computer. 10

PCB printed circuit board. 3, 5, 7, 8, 17, 18, 43, 73, 119

PCIe peripheral component interconnect express. 73, 98

PCS physical coding sublayer. 77

PECVD plasma enhanced chemical vapor deposition. 17, 25, 116

PoP package on package. 7

RAM random-access memory. 8

RF radio frequency. 14

RIE reactive ion etching. 118, 120

RTD resistance temperature detector. 16–19, 24–28, 43, 116–118

SEM scanning electron microscope. 25, 26, 84, 85

SIMD single instruction, multiple data. 3, 4

SRAM static random-access memory. 3

TIM thermal interface material. 10, 54, 57, 78, 89, 103

TSMC Taiwan Semiconductor Manufacturing Company. 2

TSV through-silicon via. 5–7, 13, 14, 56

VRM voltage regulator module. 69, 77

WLP wafer-level packaging. 7

SUMMARY

For decades, transistor scaling has been the driver of exponential advances in computing speed enjoyed by society. However, processor voltage and power levels have nearly saturated, making the relative value of power, as a resource fixed by the limits of current cooling solutions, grow exponentially with each scaling node. Additionally, the bottleneck for computational performance and energy has shifted from the switching of transistors for computation to the movement of data between various levels of storage and compute resources. 2.5D and 3D architectures have emerged as solutions to this interconnection problem, but these high-density architectures increase package power densities and only exacerbate the thermal challenge. This research aims to help enable the next generation of high performance computing architectures through the design, microfabrication, and characterization of microfluidic cooling technologies.

In this work, microfabrication processes were first developed for high aspect ratio micropin-fins and optimized for the fabrication of multiple densities of micropin-fins on a single die or wafer, enabling heterogeneous heat sink designs. Next, two avenues were explored for cooling of non-uniform power maps: hybrid heat sinks with micropin-fins and dedicated microgap hotspot coolers, and heterogeneous micropin-fin designs. Both methods effectively reduced hotspot temperatures, but the heterogeneous micropin-fins could be fabricate with a single etching step, while still effectively normalizing the temperatures between a background region dissipating 250 W/cm^2 and a hotspot region dissipating 500 W/cm^2.

To demonstrate system-level benefits of these technologies, micropin-fins were etched into the back side of a Stratix V FPGA. Running a pulse compression algorithm on the modified FPGA as a benchmark, a junction-to-inlet thermal resistance of 0.07 °C/W was demonstrated, along with improvements in device temperature, throughput, and efficiency. Lastly, an ultra-thin heterogeneous micropin-fin heat sink was developed for a 2.5D Stratix 10

FPGA, providing improvements in maximum temperature, temperature uniformity, computational efficiency, and computational density.

CHAPTER 1
INTRODUCTION

For over 50 years, transistor density in microprocessors has been doubling approximately every two years, following the trend predicted by Gordon Moore. This continuous shrinking of transistors has been the driver of improvements in computing power and cost, but computer architectures have had to adapt along the way to make full use of these transistors. While Moore's Law has been the economic blueprint for reducing transistor size and cost, "Dennard scaling" has been a technological blueprint for improving transistor performance with each scaling node.

In 1974, Dennard *et al.* described the way in which complementary metal oxide semiconductor (CMOS) transistor scaling could be used to achieve improvements in speed and efficiency [1]. If transistor length and width are scaled by $1/\kappa$ (where κ can be any scaling factor, but has often been approximately 1.4 between technology nodes), the gate oxide thickness is also scaled by $1/\kappa$ while doping concentration is increased by a factor of κ. Voltage and current can then both be scaled by $1/\kappa$, while clock frequency can be scaled up by κ. Since the power per transistor is proportional to $fC{V_{DD}}^2$, where f is the clock frequency and C is the gate capacitance, the power per transistor can be decreased by $1/\kappa^2$. With this proportional reduction in transistor size and power, overall power density can be kept constant.

Historically, however, voltage levels were not scaled down at the same rate as feature dimensions. Simultaneously, clock frequencies were scaled up at a rate greater than the inverse of transistor feature sizes. The effect was an exponential increase in power density until approximately 2005 when the situation became untenable. At this point, energy efficiency was low and power densities in some microprocessors had hit 100 W/cm^2, which is roughly the limit of conventional air cooling. Further scaling of the gate dielec-

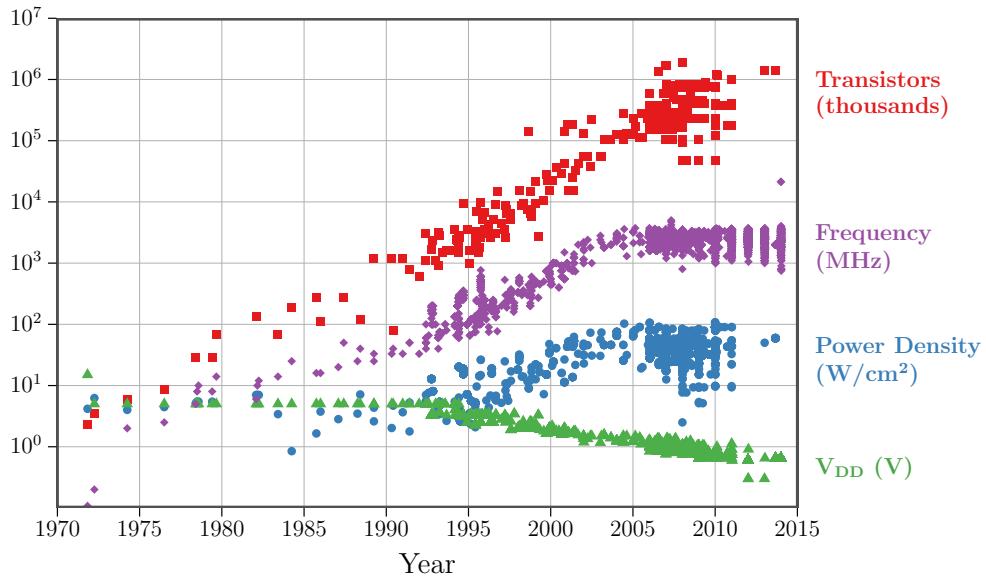


Figure 1.1: Microprocessor trends over the last 35 years. Data from CPU DB [3]

tric and threshold voltage was hampered by subthreshold leakage current, which accounted for a significant portion of total chip power [2]. Therefore, rather than increasing performance through increased clock frequencies, the industry made a shift to multi-core architectures. Clock frequencies and power dissipation flattened and further improvements in performance came primarily from exploitation of parallelism. These trends can be seen in Figure 1.1.

Nonetheless, with modern threshold voltages and clock frequencies remaining almost constant, the power per transistor only scales down at a rate of approximately $1/\kappa$. With the number of transistors increasing at a rate of $1/\kappa^2$ within a saturated power budget, the fraction of chip area which must be kept idle at any given time, sometimes called dark silicon, is increasing at an exponential rate. In other words, power and energy are becoming exponentially more expensive resources relative to chip area [4]. For example, Venkatesh *et al.* found that only 7% of a 300 mm^2 die can be switched at full frequency with a power budget of 80 W at the Taiwan Semiconductor Manufacturing Company (TSMC) 45 nm node and this fraction is decreasing by almost a factor of two with each technology generation [5].

Table 1.1: Signaling energy per 32-bit word, recreated from [7]

Interconnect Type	**Energy per 32-bit word**
On-chip	0.23 pJ
PCB	54 pJ
Interposer	17 pJ
TSV	1.1 pJ

With power and energy being primary constraints on performance, it is important to quantify where integrated circuit (IC) power is used. As on-chip cache sizes and off-chip dynamic random-access memory (DRAM) sizes have increased, the fraction of energy used for moving data has greatly eclipsed the amount of energy used to perform arithmetic operations. For example, in a 45 nm technology, fetching the instruction and operands from the register file for a 16-bit add operation uses twice as much energy as the operation itself [6]. Fetching data from the on-chip static random-access memory (SRAM) cache uses over 12× the energy of the add operation and a DRAM access requires over 1000× the energy.

Franzon *et al.* scaled the energy per operation to a future 7 nm logic technology generation and quantified the input/output (I/O) energy necessary for different methods of interconnection, which can be seen in Table 1.1 [7]. In this 7 nm technology, a 32-bit multiply-add operation would use 6.02 pJ. On the other hand, fetching an instruction and two operands from a 16 nm DRAM core would require 420 pJ, not including the energy for I/O, which would be 162 pJ using a printed circuit board (PCB) for interconnection. Scaling these numbers up to modern memory data rates, which are approaching 1 TB/s for high end single instruction, multiple data (SIMD) processors, reveals that using 16 nm DRAM with a PCB interconnect would use about 48.5 W. Increasing data rates far beyond this point without a change in architecture or cooling would be impossible.

Given these constraints, there are three important opportunities for increasing system performance, as follows.

1. *Increase hardware specialization and create very energy-efficient coprocessors which sacrifice generality for efficiency.* With a large transistor budget and only a small fraction of the chip being active at any one time, it is practical to make many highly optimized processors dedicated to specific tasks. This trend has been underway for some time with system-on-chip processors which contain large sets of specialized hardware integrated onto single chips. Another example of this is the rising use of SIMD coprocessors (such as graphics processing units (GPUs)) for workloads with a large amount of data parallelism which amortize the cost of instruction fetches over a large number of operands.

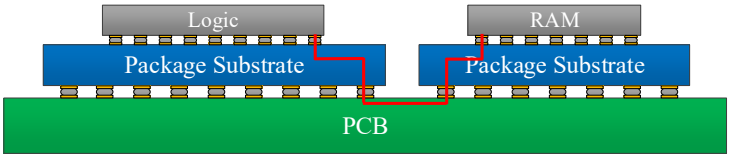
2. *Increase the efficiency of moving data by increasing system density.* Table 1.1 shows the energy per bit used for several types of interconnects. Current high end systems, such as high performance GPUs, use multi-die packages with short, high-density interconnects between chips, often using a silicon interposer or bridge as a substrate for these fine-featured interconnects. 3D architectures further improve efficiency, but also decrease the total power budget due to the reduced efficacy of conventional cooling techniques with these architectures.

3. *Increase the power budget and reduce leakage current with improved cooling.* When cooling is improved, the fraction of the chip which must be kept dark can be decreased. In addition, efficiency can be increased through reduced leakage current and architectural codesign enabled through reduced thermal design requirements.

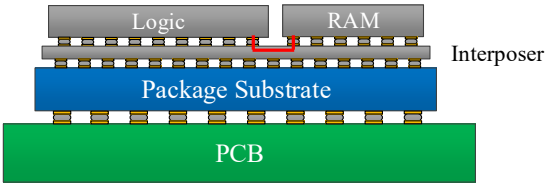
State of the Art Packaging Technologies

At the time of this writing, all three of the aforementioned techniques for improving performance are gaining industrial traction. The majority of processors incorporate dedicated blocks for efficiently executing specific workloads. Specialized coprocessors, and GPUs in particular, have become the norm in the field of deep learning, which has experienced

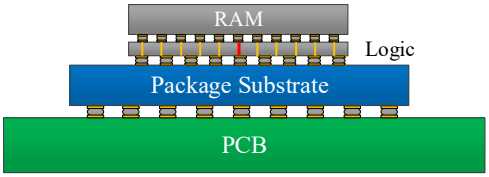
enormous progress in the last five years. This progress is largely due to the massive increase in parallel computing power offered by these more specialized pieces of hardware. A large number of new packaging techniques are also being introduced to increase system density for both high performance and mobile platforms. Liquid cooling, while still relatively niche, is also being adopted to cool a number of high performance processors [8, 9, 10].



(a) Traditional flip-chip bonded dice in separate packages



(b) 2.5D integration using a silicon interposer



(c) 3D integration using a thinned logic die with through-silicon vias (TSVs)

Figure 1.2: Multi-die interconnection methods

Figure 1.2 shows three methods of interconnecting high performance electronics. The majority of central processing units (CPUs) and memory in servers and personal computers are packaged in a fashion similar to that shown in Figure 1.2a. Signals which are transmitted between the processor and memory must pass through the package substrate, a PCB, and another package substrate before reaching the other die.

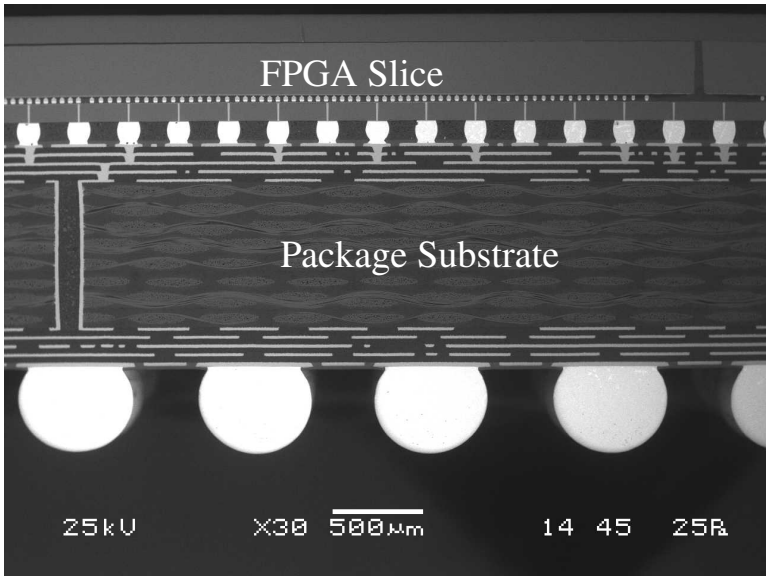


Figure 1.3: Cross section of Xilinx 28 nm FPGA on Interposer © 2012 IEEE [11]

Many high performance coprocessors, such as GPUs and field-programmable gate arrays (FPGAs), now use 2.5D architectures to package multiple dice in the same package with short, high density interconnects linking the dice. These high density interconnects can be on a silicon interposer, as shown in Figure 1.2b, or on embedded bridge chips which only span the small gaps between adjacent dice. Silicon dice can also be vertically stacked and interconnected with TSVs, as shown in Figure 1.2c. Due to cooling challenges, 3D stacking is currently limited to low power devices, such as DRAM stacks.

One of the first interposer-based products was released by Xilinx in 2011. By splitting their 28 nm FPGA into several die slices and integrating them in a single package with high density interconnects on a silicon interposer, Xilinx was able to double the amount of logic they could fit in a single package, while also improving yield relative to a single-die solution [12]. A cross section of the packaged FPGA slices can be seen in Figure 1.3. The microbumps used to connect the logic dice to the interposer are at a pitch of 45 µm while the interposer-to-package solder bumps have a pitch of 180 µm [11].

3D packaging has primarily been adopted in mobile applications and high performance DRAM. In mobile applications, a large number of methods have been used to shrink de-

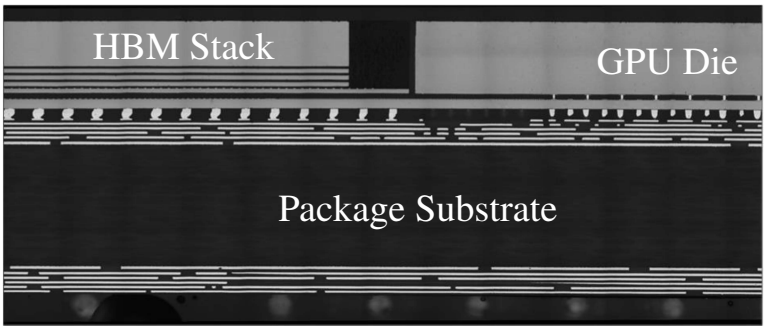
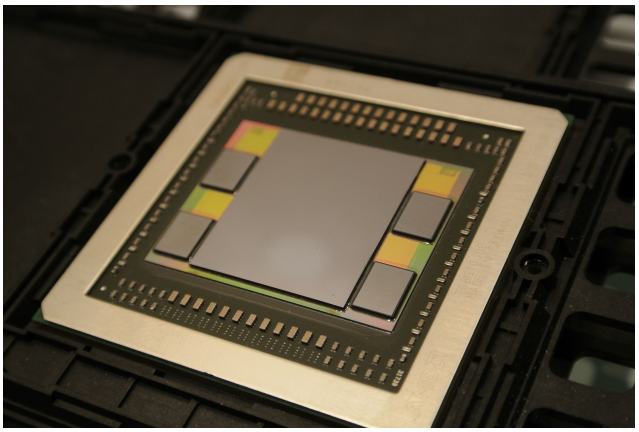


Figure 1.4: Cross section of an Nvidia Pascal GPU with interposer and High Bandwidth Memory (HBM) © 2016 IEEE [13]

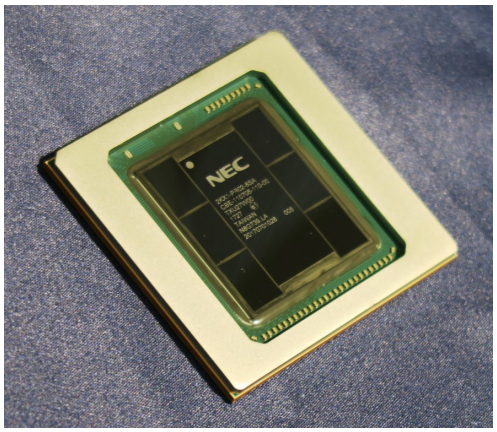
vices, including package on package (PoP) stacking, face-to-face stacking, and, more recently, a wide range of wafer-level packaging (WLP) techniques. Memory density has been increased by using die stacking with wire bonding and, more recently, 3D stacking with TSVs. The Hybrid Memory Cube (HMC) and HBM are two such 3D memory stacks utilizing TSVs.

The HMC utilizes a high-speed serialized bus which communicates through PCBs, where interconnect numbers are limited. HBM saves power and increases bandwidth by utilizing an extremely wide bus (1024-bit) at lower frequency, enabled through high density in-package interconnects, generally on an interposer. Figure 1.4 shows a cross section of an Nvidia Pascal GPU on an interposer next to an HBM stack.

The Advanced Micro Devices (AMD) Fiji GPU, shown in Figure 1.5a, was the first GPU to use HBM. Since the AMD Fiji GPU was released, several more devices have integrated HBM and HBM2 memory into the same package, generally with up to four HBM(2) stacks surrounding a larger logic die, interconnected with a silicon interposer. The NEC SX-Aurora TSUBASA vector processor, shown in Figure 1.5b, was the first processor to integrate 6 HBM stacks in a single package, for a total memory capacity of 48 GB and memory bandwidth of 1.2 TB/s per processor [14].



(a) AMD Fiji GPU with interposer and HBM



(b) NEC SX-Aurora TSUBASA vector processor with HBM2

Figure 1.5: Two processors utilizing an interposer with HBM

High Performance Memory Interface Standards

Double data rate (DDR)4 is currently the highest commercially available DRAM memory standard for desktops, notebooks, and servers. The highest bandwidth DDR4 memory currently supported by mainstream processors is DDR4-2666 with a data transfer rate of 21.3 GB/s per channel. At the high end of CPUs, the AMD EPYC processor has a total of eight memory channels, for a total memory bandwidth of approximately 170 GB/s [15]. Like DDR random-access memory (RAM), graphics double data rate (GDDR) RAM communicates with logic through a PCB. GDDR5X is currently the highest GDDR memory standard in use. The Nvidia TITAN Xp utilizes a 384-bit wide bus to achieve a data rate of 547.6 GB/s to GDDR5X memory [16].

Many accelerators requiring the highest memory bandwidth now utilize HBM and HBM2 memory integrated into the same package as the logic die. HBM2 RAM utilizes a very wide (1024-bit) bus to efficiently achieve data rates of up to 256 GB/s per 3D RAM stack. Several high end GPUs incorporate four HBM stacks in a single package. For example, the Nvidia V100 GPU includes four HBM2 stacks with an aggregate capacity of 32 GB and a total data rate of 900 GB/s [17].

Table 1.2: Commercially available high performance memory interfaces

Architecture	Standard	Bus Width	Max Commercial Channels	Aggregate Bandwidth
2D	DDR4	64 bits	8	170 GB/s
2D	GDDR5X	32 bits	12	547.6 GB/s
2.5D	HBM2	1024 bits	6	1200 GB/s

Table 1.2 summarizes these interconnect technologies and commercially available memory interfaces for high performance applications. 3D memory interfaces, such as Wide I/O, are currently limited to low power, mobile applications.

While moving the DRAM into the same package as the logic has greatly improved performance and efficiency, it has also introduced additional cooling challenges. Many of these high performance architectures operate with package heat fluxes at the limit of what traditional air cooled heat sinks can dissipate. In addition to increasing total package powers, integrating DRAM into the same package as the logic can also introduce the issue of thermal coupling between the logic and DRAM. This may cause the DRAM to refresh more often, reducing performance and increasing power. Continued progress along this trajectory of increasing system density will be severely limited by the ability to cool these architectures, particularly in the area of 3D logic stacking. By enabling high density architectures as well as lowering device temperatures, improving heat sink thermal conductance can improve computing system performance at a ratio greater than 1:1 by simultaneously increasing the power available for computational operations and decreasing the average power per operation.

Microfluidic Cooling

Microscale liquid cooling offers a solution to these thermal challenges. By using a liquid coolant, such as water, which has a volumetric heat capacity approximately three orders of magnitude higher than air, the necessary volume for heat exchange and fluid delivery

can be dramatically decreased compared to traditional air cooled heat sinks. The thermal conductivity of water is also approximately one order of magnitude higher than that of air, which further improves heat exchange to the fluid. Liquid cooled heat sinks have been commercially available for quite some time, particularly for enthusiast personal computers (PCs) and supercomputers. However, these existing liquid cooled heat sinks tend to be fabricated through macroscale fabrication techniques, such as skiving, and are mounted on top of packaged electronics with a thermal interface material (TIM), in a similar fashion to air cooled heat sinks.

An example of heat transfer in a channel is useful for demonstrating the utility of microfluidic cooling. The Nusselt number, Nu, is defined as the ratio of heat convection to heat conduction. For flow through a channel it is given by

$$Nu = \frac{hD_h}{k_f} \tag{1.1}$$

where h is the heat transfer coefficient, D_h is the hydraulic diameter of the pipe, and k_f is the thermal conductivity of the fluid. For fully developed laminar flow, Nu is constant, giving us the relation

$$h \propto \frac{k_f}{D_h} \tag{1.2}$$

which tells us that decreasing our hydraulic diameter will increase our heat transfer coefficient between the solid surface and the fluid. This tells us that decreasing the size of our heat sink not only improves the footprint, but also the convective thermal resistance. The heat sink can even be built at the microscale in the same silicon as the integrated circuit, so that it is physically located within a few hundred micrometers of the heat generating transistors.

Microfluidic cooling can have a number of electrical benefits. First, improved cooling can reduce dark silicon and enable computer architectures which are not feasible with traditional cooling solutions. Second, lowering temperatures of existing device architectures

can increase the clock frequency at which integrated circuits can be run and improve efficiency by reducing leakage current. The following sections will highlight some of the prior work in this area.

Microchannel Cooling

The aforementioned inverse relationship between length scale and heat transfer coefficient motivated Tuckerman and Pease to create the first silicon microfluidic heat sink in 1981 [18]. Microchannels with a width and wall thickness of 50 µm were etched into a silicon substrate to a depth of approximately 300 µm. Heaters and temperature sensors were deposited on the opposite side of the silicon. A thermal resistance of approximately 0.09 °C/W was achieved between the fluid inlet and heaters, over a 1 cm × 1 cm area. This thermal resistance is several times lower than the thermal resistance of the best air cooled heat sinks, and would therefore be sufficient to cool devices with heat fluxes several times higher than those cooled with air cooled heat sinks.

The thermal resistance from junction to fluid was conveniently broken down into three components as follows:

$$R_{tot} = R_{cond} + R_{conv} + R_{heat} \tag{1.3}$$

where R_{cond} is the thermal resistance due to conduction through the silicon, R_{conv} is the thermal resistance associated with transferring the heat from the silicon to the fluid, and R_{heat} is the effective thermal resistance due to heating of the fluid.

This microchannel heat sink from Tuckermann and Pease was fabricated using an anisotropic wet etching process. It had channel and fin widths of 50 µm which were nearly optimal dimensions for the applied fluid pressure gradient of 31 psi. After this seminal work, the advent of deep reactive ion etching (DRIE) paved the way for several other microfluidic heat sink geometries, such as micropin-fins with a wide range of diameters, depths, pitches, and cross sectional shapes.

Micropin-fin Cooling

While rectangular microchannels essentially provide two to three degrees of freedom for optimization, micropin-fin heat sinks have diameter, transverse pitch, lateral pitch, and an infinite number of cross sectional shapes that can be used. Several different empirical correlations have been created to predict their performance, each pertaining only to a small part of this large design space.

Prasher *et al.* performed single phase experiments on cylindrical silicon micropin-fin heat sinks with diameters between 50 and 150 µm with water as a coolant [19]. Correlations for friction factor and Nusselt number, which are the commonly used dimensionless terms relating to pressure drop and heat transfer coefficient, were developed to address the discrepancy between previously developed correlations and the observed experimental data. Koz *et al.* and Tullius *et al.* also performed parametric simulations of cylindrical micropin-fin arrays to study thermal and hydraulic performance as a function of micropin-fin design [20, 21].

Brunschwiler *et al.* investigated microchannels and micropin-fins with an emphasis on interlayer cooling in 3D IC stacks [22]. Many different types of microfluidic heat sinks were tested and compared in single phase experiments with deionized water as a coolant. Geometries included microchannels, inline cylindrical micropin-fins, staggered micropin-fins, and tear-drop micropin-fins. In addition to these variations in basic structure, two different heights (100 µm and 200 µm) and several pitches of micropin-fins and microchannels were fabricated, tested, and compared. Single phase thermal and hydraulic data was used to develop correlations for Nusselt number and friction factor as a function of micropin-fin geometry and flow conditions.

Kosar *et al.* also reported thermal and hydraulic measurements on micropin-fin heat sinks with multiple cross sectional shapes and reported the ideal shapes as a function of flow conditions [23]. A number of studies have also looked at flow boiling with microchannels and micropin-fins [24, 25].

Hotspot Cooling

In addition to large average heat fluxes, localized hotspots with heat fluxes many times higher than the average chip heat flux may exist, ultimately setting the thermal envelope for the entire package [26]. This makes heterogeneous cooling options, which separately target the hotspots and background regions, attractive.

Several methods of hotspot mitigation have been explored, including micro thermoelectric coolers [27], liquid jet impingement [28, 29, 30], thin film evaporation [31], and heat spreading through highly thermally conductive materials, such as graphene [32, 30].

Microfluidic Cooling in 3D Stacks

Brunschwiler *et al.* demonstrated a four tier die stack with a microfluidic heat sink in each of the four tiers, with all four tiers sharing an inlet and outlet region [33]. A total power of 390 W was dissipated in the stack. Zhang *et al.* demonstrated tier specific microfluidic cooling in a two tier stack, independently controlling the flow rate in each of the 1×1 cm micropin-fin arrays [34].

If a microfluidic heat sink is integrated into an interposer, or in a tier in a 3D stack, electrical signaling must still be achieved with TSVs, which can pass through the microfluidic heat sink. TSVs in silicon micropin-fins have been demonstrated as a means of facilitating interlayer electrical connectivity in a 3D stack with inter-tier microfluidic cooling [35]. Since heat sink thermal performance improves with increasing micropin-fin height, but TSV performance decreases with increasing height, a tradeoff exists between electrical and thermal performance when integrating TSVs in micropin-fins [36]. Oh *et al.* fabricated high aspect ratio (23:1) TSVs in a 300 µm tall micropin-fin heat sink [37]. By increasing the aspect ratio of the TSVs, both the TSV capacitance and footprint were reduced, partially sidestepping this thermal-electrical trade-off through improved TSV technology. Since this demonstration of high aspect ratio TSVs, Oh *et al.* have also demonstrated the radio frequency (RF) properties of these TSVs, both with and without surrounding coolant [38]. It

was found that water increases the TSV capacitance at high frequencies. In order to isolate the TSVs from the RF effects of the coolant, a ring of TSVs was fabricated around signal TSVs, mimicking a coaxial structure.

A simple and scalable method of delivering coolant is also necessary to build scalable 3D stacks with inter tier microfluidic cooling. Zheng *et al.* demonstrated simultaneous fabrication of fine pitch microbumps and fluid delivery ports sealed with solder rings of the same height as the microbumps [39]. These fluid ports can be used to deliver fluid to microfluidic heat sinks built into the backsides of dice, without adding additional assembly steps beyond those used for microbumps.

Research Objectives and Contributions

While significant work has focused on characterizing microfluidic heat sinks on passively heated silicon, there is still a substantial amount of work to be done characterizing system performance with microfluidic cooling. The following work attempts to fill this gap through a combination of fabrication advancements, microfluidic cooling designs for less idealized systems, and microfluidic cooling and benchmarking of functional CMOS systems.

The objective of this work is to create microfluidic cooling devices which specifically target the needs of modern, heterogeneous microelectronics. There are two parts to this. First, microfluidic cooling was integrated into functional CMOS circuitry to characterize the benefits and challenges associated with use with real electronics. Second, microfluidic cooling solutions were developed to target heterogeneous cooling needs within a single package. This encompasses both single-die heterogeneity, where on-die hotspot cooling must be balanced with full chip cooling, and the heterogeneity encountered in a 2.5D multi-die cooling scenario.

Organization of this Thesis

The chapters of this thesis are outlined below, each relating to the objectives mentioned above.

1. Microgap devices were developed to address localized hotspots on a silicon die. These microgaps were first fabricated and tested as stand-alone devices and then as part of a larger 1 cm $\times$ 1 cm chip. These devices were capable of dissipating several kW/cm^2.

2. Heterogeneous micropin-fin heat sinks were fabricated and tested with power maps consisting of a hotspot surrounded by a larger chip area. The hotspot temperature was effectively normalized to the average background temperature by using higher density micropin-fins directly over the hotspot region.

3. A monolithic micropin-fin heat sink was fabricated in the backside of a functional CMOS FPGA die. Improvements in temperature, throughput, and efficiency were measured. To our knowledge, this was the first time a microfluidic heat sink was successfully integrated into a functional CMOS die.

4. A heterogeneous micropin-fin heat sink was designed to cool each of the 5 dice in a 2.5D Stratix 10 FPGA package. The heat sink was fabricated and attached to the exposed backsides of the dice in a delidded Stratix 10 package. Improvements in device temperature, thermal decoupling between dice, and form factor were realized.

To support the above chapters, significant fabrication and test device development was performed. Some of this work, including the portions which were frequently reused or not specific to a certain experimental task, are outlined in the Appendix A.

CHAPTER 2

HOTSPOT COOLING USING DEDICATED MICROGAPS

As discussed in the introduction, the heat transfer coefficient in a channel is inversely proportional to the hydraulic diameter of the channel. Therefore, very high heat transfer coefficients can be achieved by pumping fluid through microgaps of only a few micrometers in depth. This approach, however, creates a very high pressure drop and is therefore impractical for cooling entire chips. Over the small area of a hotspot, however, this approach is feasible. In this chapter, dedicated microgaps are explored as a means of cooling local IC hotspots with extremely high heat fluxes. These experiments were conducted in two phases. First, dedicated microgap devices were fabricated and tested with R134a to extract heat transfer coefficients and determine feasibility for cooling large heat fluxes. Second, a dedicated microgap was integrated into a more complete test chip with a larger micropin-fin heat sink for background cooling. Extremely large heat fluxes up to 6.175 kW/cm^2 were dissipated from the hotspot heaters in these experiments, although it was found that much of this heat was spread to the surrounding bulk silicon and dissipated outside of the microgap.

Dedicated Hotspot Microgap

Thermal test devices were first developed to test cooling of hotspots using extremely thin microgaps. The devices were developed to cool a heated area of size 200 µm $\times$ 200 µm with a 200 µm $\times$ 300 µm gap. The devices were created with a gap of depth 5.6 µm over a serpentine platinum trace acting as both a heater and resistance temperature detector (RTD). Inlet and outlet channels with a nominal depth of 50 µm were designed to deliver coolant to the gap with a negligible pressure drop compared to the the pressure drop through the much smaller microgap. A cross-sectional diagram of the device can be seen in Figure 2.1.

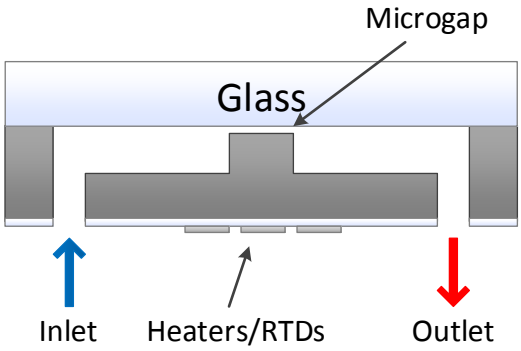


Figure 2.1: Cross-sectional diagram of microgap test device

The fabrication process used to create this device was similar to that developed for micropin-fin test devices (Appendix A), but there was an additional etching step. An outline of the process can be seen in Figure 2.2. First, a cavity with a footprint of 200 µm × 300 µm was etched into a 300 µm thick silicon wafer to a depth of 5.6 µm to create the microgap. Then the inlet and outlet channels were etched to a depth of 54.6 µm. The glass cap was then anodically bonded to the silicon to seal the microgap and fluid delivery channels. A 2 µm thick silicon dioxide layer was then deposited onto the silicon using plasma enhanced chemical vapor deposition (PECVD) to insulate the conducting traces from the silicon. 5 µm wide serpentine platinum traces were then deposited as heaters/RTDs, followed by copper and gold for the wirebonding pads and connections between pads and RTDs. Although not shown in the figure, a silicon dioxide passivation layer was deposited on top of the platinum heaters to protect them from the environment. This passivation was then etched away from the gold bonding pads to allow wirebonding. Ports were etched last to prevent photoresist used for other processing steps from clogging the narrow channels. This process flow was also constrained by the platinum heaters which are only 5 µm in width and require a relatively flat surface for processing.

The test device was packaged and tested with R134a as a two phase coolant in Dr. Andrei Fedorov's lab by Mohamed Nasr in collaboration with Craig Green and Dr. Yogendra Joshi, with results reported in [40]. The device was first mounted on a PCB with a

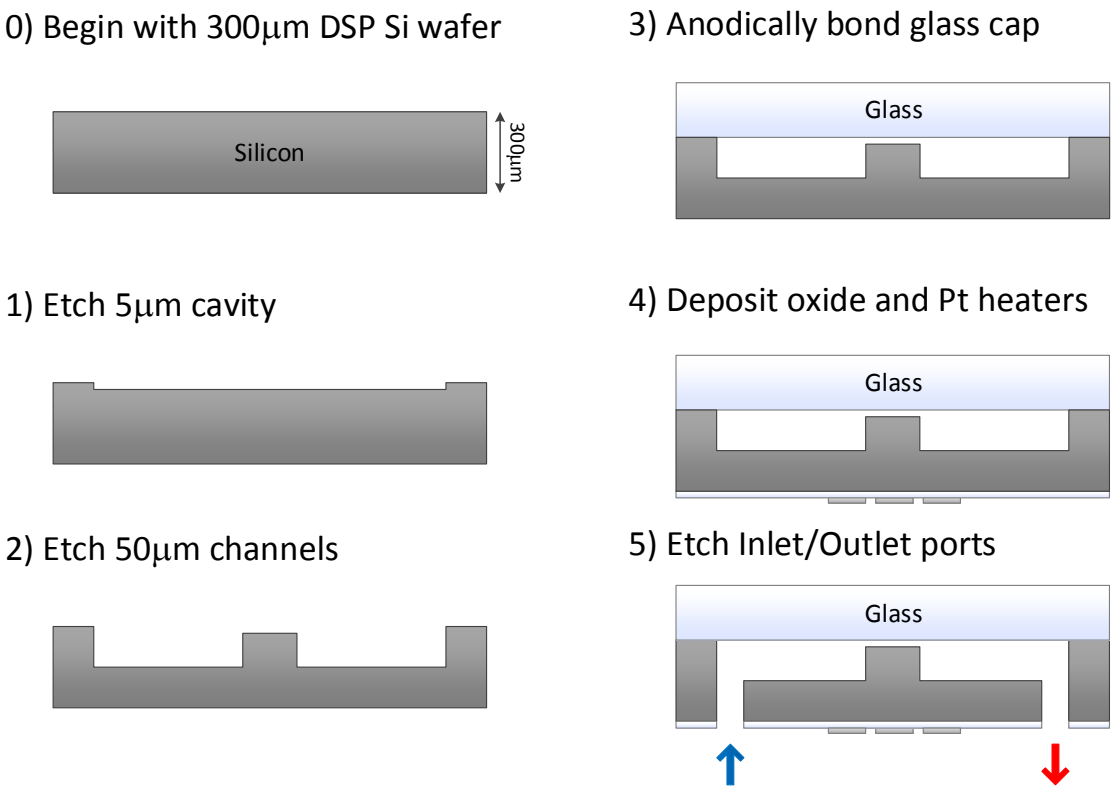


Figure 2.2: Microgap test device fabrication process

cutout for microgap viewing through the Pyrex device cap. Heaters/RTDs were connected to the PCB by wirebonding contact pads on the device to traces on the PCB. A machined PEEK package was then mounted on the other side of the device and used to apply pressure to O-ring seals around inlet and outlet ports. A diagram of the package can be seen in Figure 2.3.

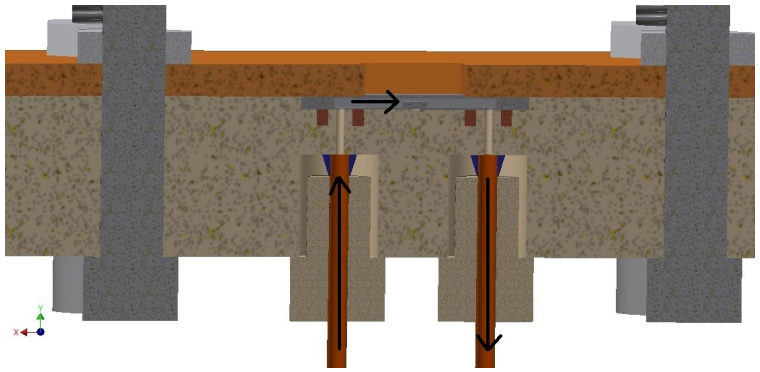


Figure 2.3: Package for microgap test device

The packaged test device was then placed in the closed loop system illustrated in Figure 2.4. A KDS Scientific Legato 270 syringe pump was used to deliver fluid to the microgap test devices. An Agilent 34970a was used to record flow rates, pressures, temperatures, and heater/RTD voltage current. Fluid temperatures were measured with Omega K-type thermocouples and pressures were measured with Omega PX 309 pressure transducers. Power was delivered to the heaters on the test device with an Agilent E3641A power supply. Lastly, a Keyence VH-Z100R microscope was used to obtain images of the boiling within the microgap under several flow conditions.

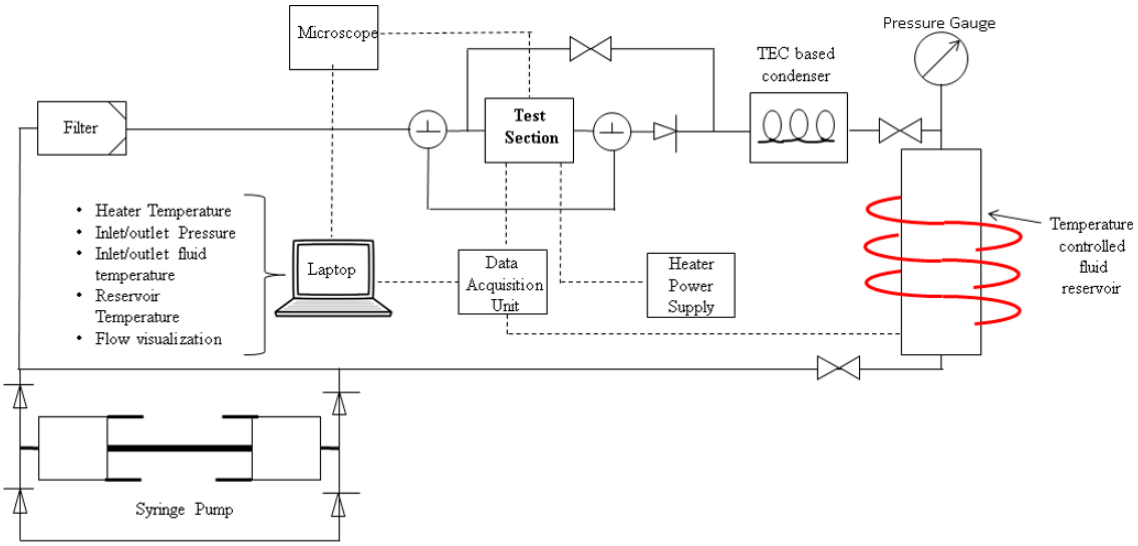


Figure 2.4: Experimental setup for microgap test devices

Thermal data was collected at three different mass flux rates: 3000 kg/(m^2 s), 5000 kg/(m^2 s), and 7000 kg/(m^2 s) with a a subcooled inlet temperature of 22.4 °C. Since characterization of the heat transfer coefficient within the microgap would require isolation of the heat flux entering the microgap, total thermal resistance to the ambient was computed instead. This thermal resistance was computed as

$$R'' = \frac{(T_h - T_a)}{Q''_h} \tag{2.1}$$

where T_h is the heater temperature, T_a is the ambient temperature, and Q''_h is the heat flux density from the 200 µm × 200 µm heater. This reported thermal resistance, therefore, includes the thermal resistance from the microgap, but also the parallel thermal resistance from the heat path through the surrounding bulk silicon to ambient. These thermal resistances are reported as a function of heat flux density in Figures 2.5, 2.6, and 2.7. Corresponding flow regimes, which were determined based on visualization of the gap, are labeled on these plots.

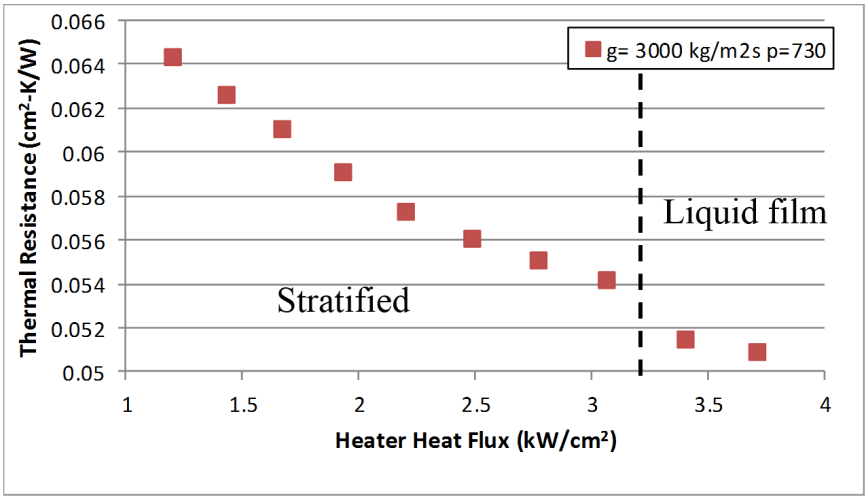


Figure 2.5: Microgap thermal resistance with mass flux of 3000 kg/(m^2 s)

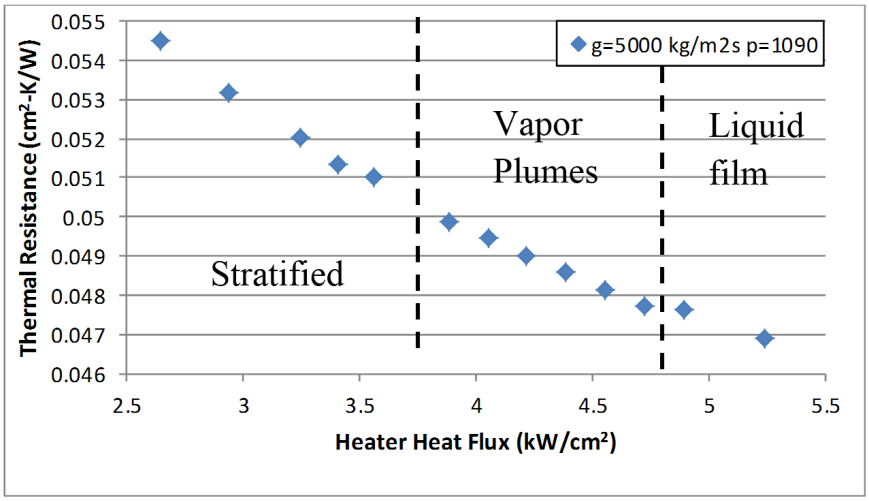


Figure 2.6: Microgap thermal resistance with mass flux of 5000 kg/(m^2 s)

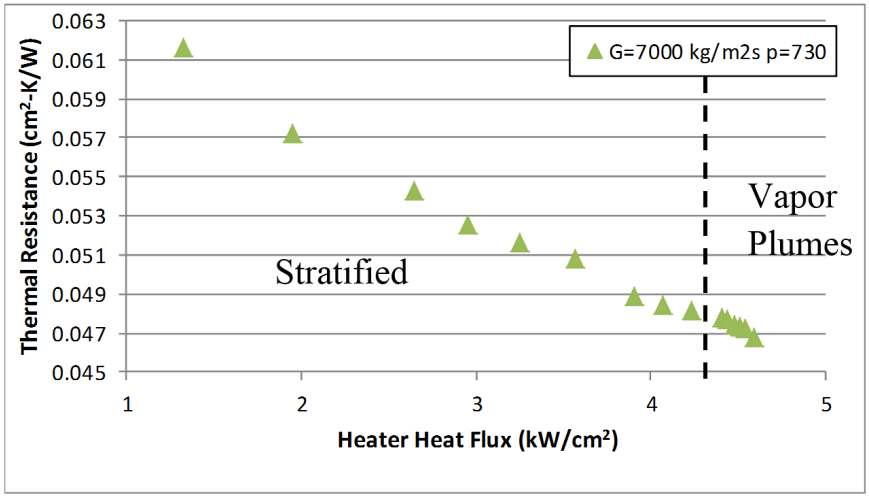


Figure 2.7: Microgap thermal resistance with mass flux of 7000 kg/(m^2 s)

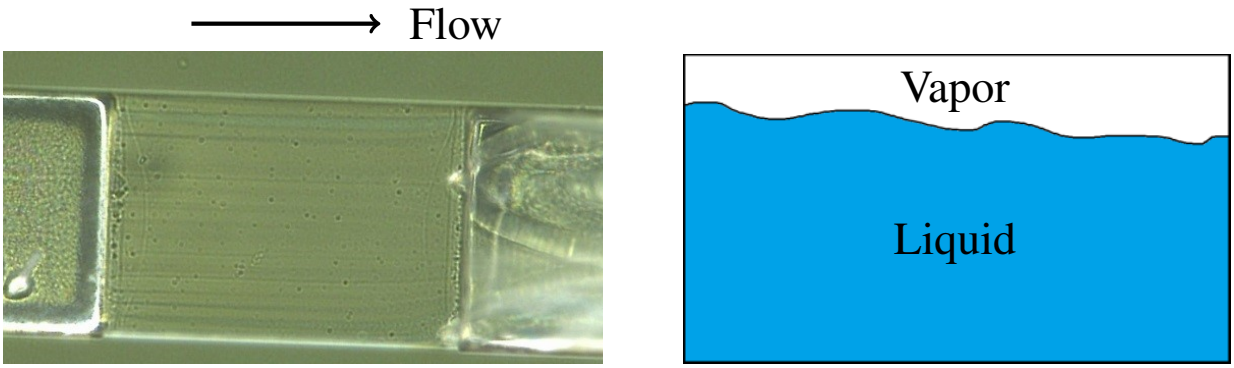


Figure 2.8: Microgap with stratified flow

Images of the microgap with different flow conditions can be seen in Figures 2.8, 2.9, and 2.10, along with diagrams interpreting these flow regimes. Based on these visualizations, the flow was classified into three separate regimes. At relatively low heat flux, the flow was stratified, with liquid coolant at the bottom of the microgap and a vapor layer at the top. At higher mass fluxes, the flow transitioned from stratified flow to vapor plume flow as the heat flux was increased. At the highest heat fluxes, an ultra-thin wavy liquid film was observed, which is thought to provide the highest heat transfer coefficients.

Pressure drop measurements can be seen for all three mass fluxes in Figure 2.11. The flow regime transitions, as observed visually, are marked on the plot and correspond to changes in the slope of the pressure versus heat flux. Pressure drop remained fairly constant

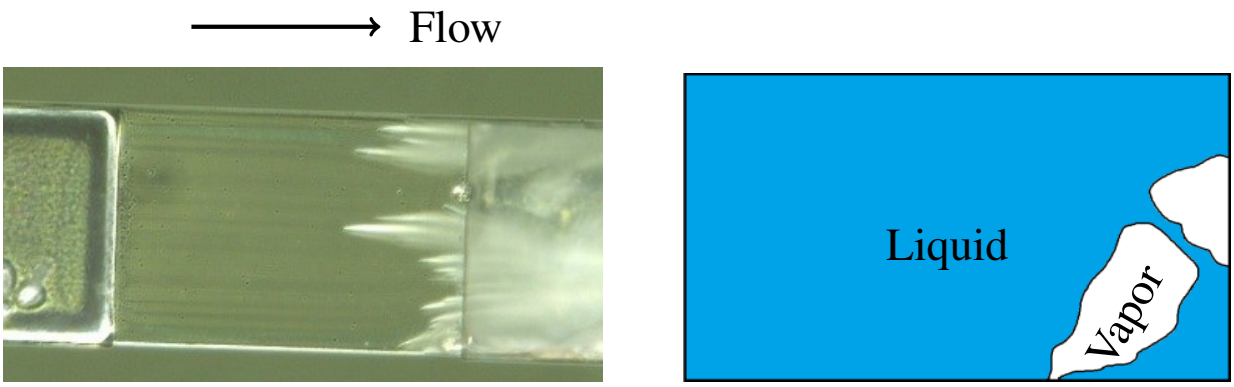


Figure 2.9: Microgap with vapor plume flow

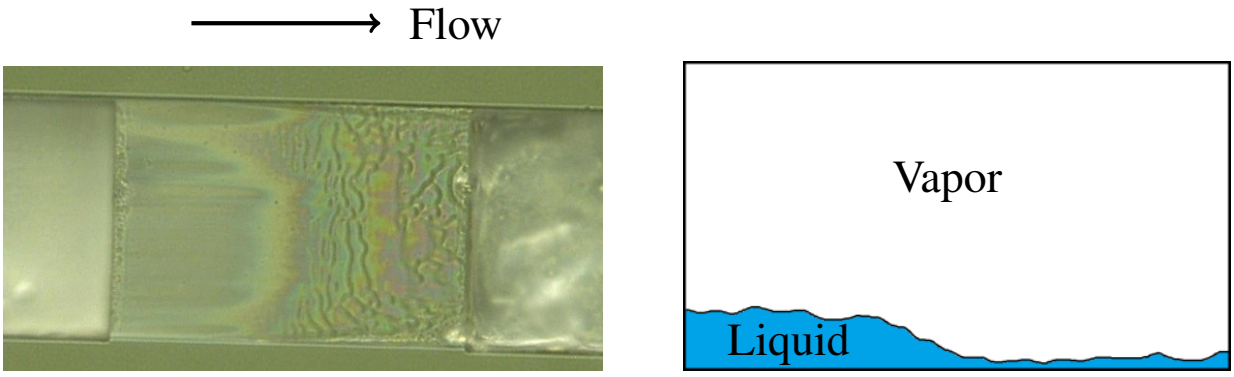


Figure 2.10: Microgap with ultra thin film flow

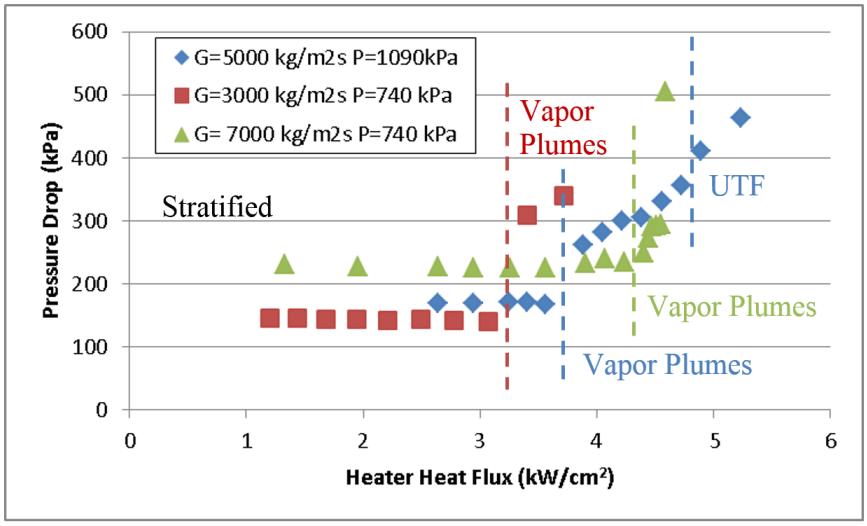


Figure 2.11: Microgap pressure drop

in the stratified flow regime and began to increase sharply in the the other flow regimes. This may indicate that a relatively small amount of boiling was occurring in the stratified flow regime. The large increase in pressure drop which occurred after these transitions was only accompanied by a small decrease in thermal resistance in Figures 2.5, 2.5, 2.5. However, the thermal resistance measurements include heat spreading through the bulk silicon, which can be significant. It is therefore difficult to draw conclusions about the heat transfer coefficient in the microgap.

This work covers extremely small microgaps at very high mass fluxes relative to prior work. It was extended in later work in which devices were developed with an air gap around the microgap to dramatically reduce the amount of heat spread through the bulk silicon [41, 42].

Combined Hotspot and Background Cooling Using a Dedicated Microgap and Micropin-fin Heat Sink

The microgap concept which was isolated in the previous section was also integrated into a test sample with both a hotspot region and a background region. A microgap depth of 10 µm was used and deionized water was used as a single phase coolant.

The single phase heat transfer coefficient in this microgap can be estimated analytically. For laminar flow in a channel, the Nusselt number is given by

$$Nu = \frac{hL}{k_f} \tag{2.2}$$

where h is the heat transfer coefficient on the wall of the channel, L is the width of the channel, and k_f is the thermal conductivity of the fluid. For fully developed flow, Nu is constant. For a channel with width much greater than depth, and heating from the bottom, such as the microgap, Nu is approximately 5.4[43, 44]. Although flow across the entire microgap will not be fully developed, this is used as a baseline to estimate the heat transfer coefficient in the channel. Taking the thermal conductivity of water to be 0.61 W/(m K) at 300 K[45], is estimated to be 3.3×10^5 W/(m^2 K).

Testbed Design and Fabrication

Both the background and hotspot regions of the test device had a dedicated inlet, but fluids merged into a shared outlet port for the two. The hotspot region was a 200 µm × 200 µm section in the center of the chip, surrounded by a 1 cm × 1 cm background region. This hotspot region was cooled with a dedicated microgap with a separate inlet, while the background region was cooled with a hydrofoil micropin-fin array. Five platinum heaters/RTDs were used to deliver power to the background region, while the hotspot was heated with a single RTD. The inlet and outlet plena contained larger diameter pin-fins for structural support, raising the pressure at which the device could be operated before the cap cracked or became separated from the bottom[46]. Four rows of dense micropin-fins were added at the inlet to stabilize fluid flow. This was primarily done for later testing with two phase coolants, which can experience unstable oscillations and dry out without this feature. A top-view diagram of the test device can be seen in Figure 2.12.

The fabrication process for the device can be seen in Figure 2.13. The process was

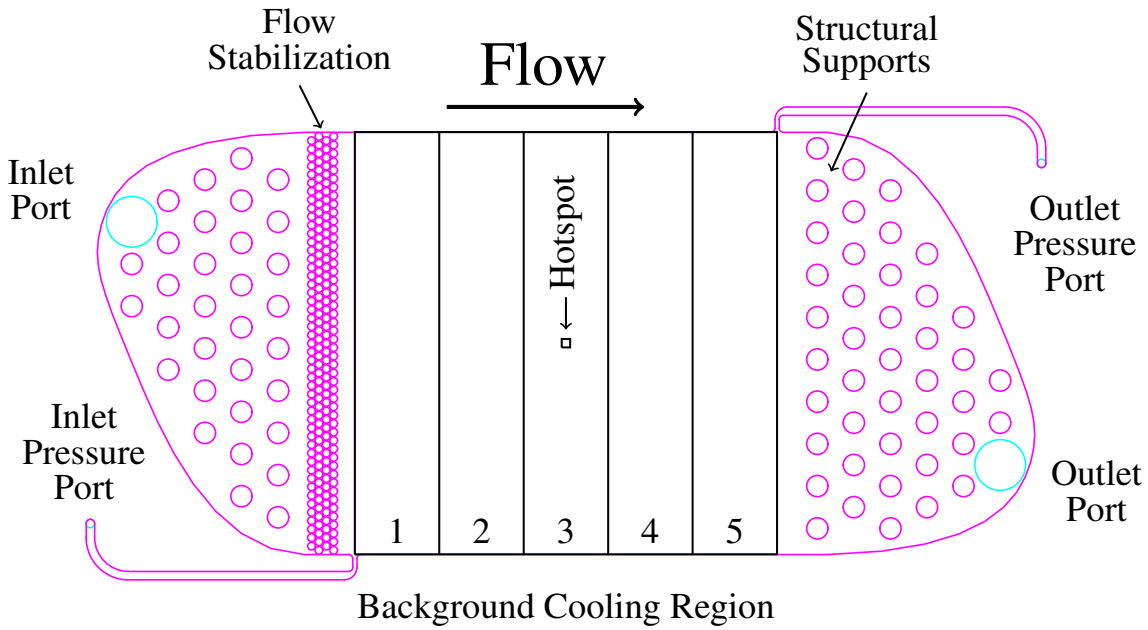


Figure 2.12: Combined hotspot and background cooling test device diagram

carried out in the cleanrooms at the Georgia Institute of Technology by Reza Abbaspour. Fabrication began with a 500 µm double side polished (DSP) wafer. First, the 10 µm deep microgap was etched using DRIE. Next, the purple regions in Figure 2.12 were etched to a nominal depth of 200 µm, including the background cooling micropin-fins, support structures, and flow stabilization micropin-fins. A scanning electron microscope (SEM) image of the etched silicon at this step can be seen in 2.14a. Next, a silicon cap was bonded to the silicon using direct silicon-silicon bonding. A 2 µm silicon dioxide layer was then deposited as an electrically insulating layer using PECVD before depositing platinum RTDs and gold bonding pads. An image of the hotspot RTD and surrounding background RTDs can be seen in 2.14b. Lastly, five ports were etched into the silicon cap: a background inlet, a hotspot inlet, a shared outlet, and two background pressure ports. The devices were built into a package which delivered fluid through the top of the device and power through wirebonds on the bottom. The platinum RTDs were calibrated in an oven with k-type thermocouples and a line was fit to each RTD to o convert resistance to temperature.

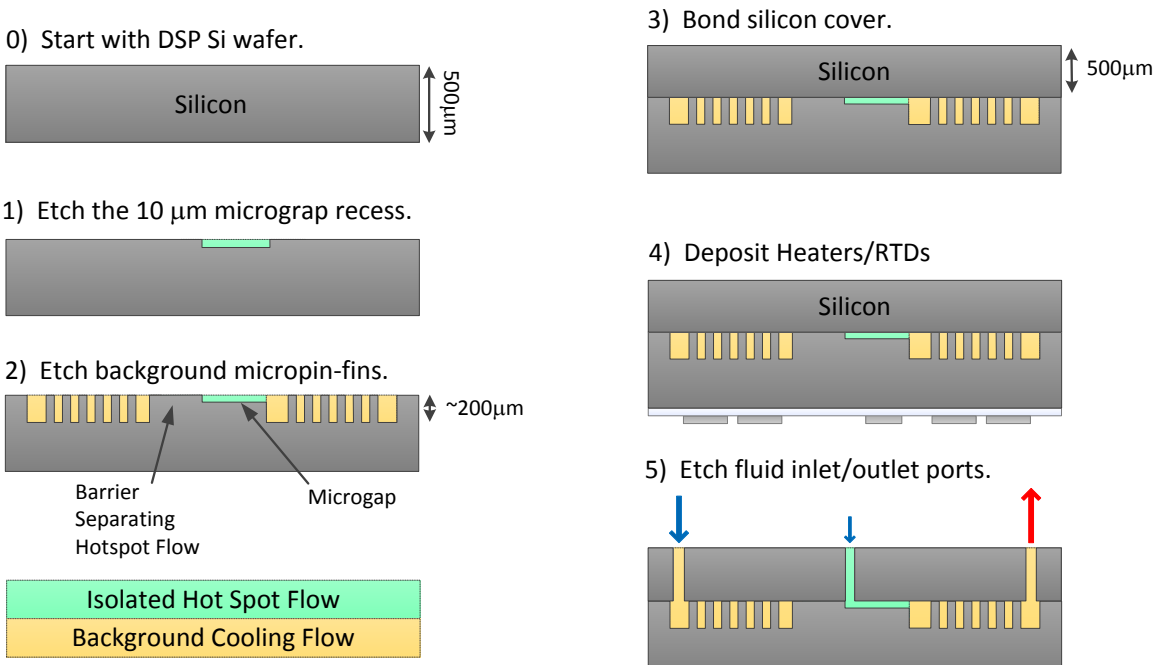
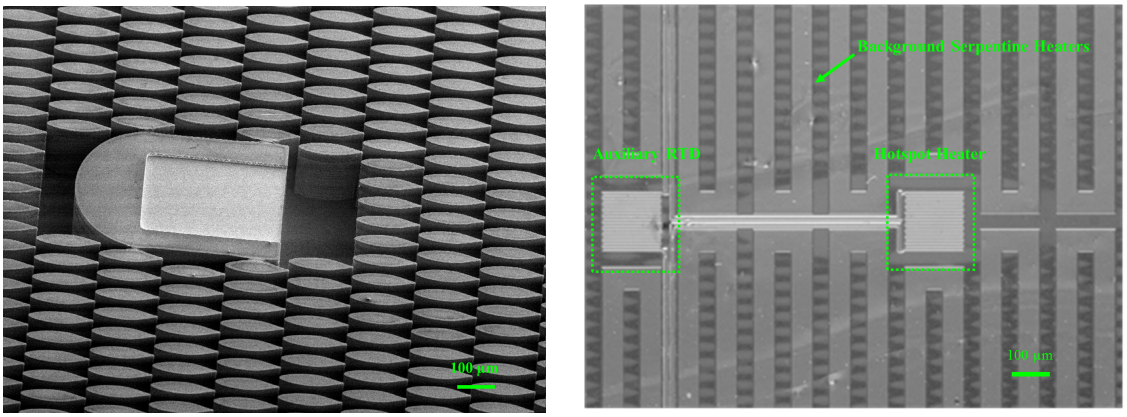


Figure 2.13: Combined hotspot/background test device



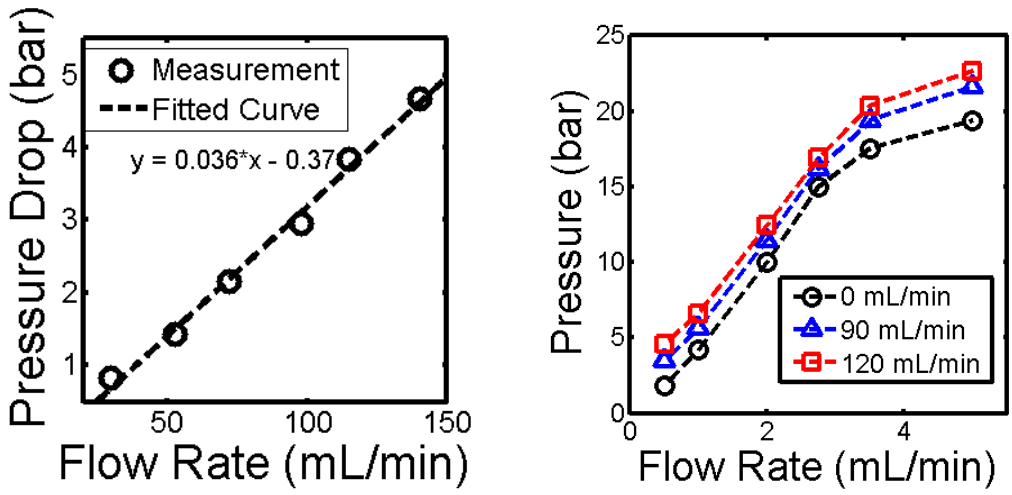
(a) SEM image of hotspot microgap surrounded by background hydrofoil micropin-fins

(b) Infrared image of hotspot RTD surrounded by background RTDs

Figure 2.14: Images of combined microgap/background test chip

Thermal and Hydraulic Results

The chip was tested using single phase deionized water as a coolant for both the hotspot and background regions. Three absolute pressure measurements were made: the hotspot inlet, the background inlet pressure port, and the combined outlet pressure port. A gear pump was



(a) Background pressure drop

(b) Hotspot pressure drop

Figure 2.15: Hotspot and background pressure drops

used to deliver fluid to the background inlet, while a syringe pump was used to deliver fluid to the hotspot inlet. An Agilent N6705B power analyzer was used to deliver 20 W to each of the five background heaters for an average heat flux density of 100 W/cm^2, while hotspot power was independently varied. Voltage and current across all heaters were recorded using an Agilent 74972A data logger. These values were used to compute resistance, which was then converted to temperature using the calibration for each RTD. The background flow rate was measured using a rotameter, while the hotspot flow rate was calculated from the syringe pump speed and syringe cross sectional area.

The background flow rate was varied from approximately 20 mL/min to 140 mL/min. Pressure drops at these flow rates can be seen in 2.15a. The relationship seems to be nearly linear, which, along with the Reynolds number range, suggests that the flow regime remains purely laminar, likely without the vortex shedding sometimes seen around cylindrical micropin-fins at higher Reynolds numbers [47]. The hotspot pressure drop, measured from the hotspot inlet to the combined outlet pressure port, can be seen in Figure 2.15b. Since this pressure drop also includes pressure drop from the combined background/hotspot flow across the latter half of the device, the pressure increases with increasing background flow

rate as well as hotspot flow rate. At the highest flow rates of approximately 5 mL/min and 120 mL/min on the hotspot and background, respecticely, the pressure drop from hotspot inlet to the outlet pressure port was approximately 22 bar. At these very high pressures, both at the hotspot and across the entire test chip, good bonding between the cap and etched silicon is critical to prevent separation or cracking of the silicon [46]. For this test chip, the packaging involved encapsulation in epoxy, which further improved the reliability of the device at high pressures.

Using a background flow rate of 100 mL/s and a hotspot flow rate of 1.5 mL/min, with both inlet temperatures at 20.1 °C, the hotspot heat flux was varied from 100 W/cm^2 to 6.175 kW/cm^2. The hotspot temperature rise above inlet temperature can be seen as a function of heat flux in Figure 2.16. With a uniform heat flux of 100 W/cm^2 across both the background and hotspot regions, the temperature of the hotspot RTD was found to be 8.35 °C below the temperature of the surrounding background RTD. The hotspot heat flux was increased and reached the average background temperature at a heat flux density of approximately 500 W/cm^2. At a maximum hotspot heat flux density of 6.175 kW/cm^2, the temperature rise of the hotspot above the inlet temperature was 62.0 °C. As expected, the temperature rise of the hotspot increases linearly with heat flux, at a rate of 0.011 °C cm^2/W. It should be noted that, due to significant spreading to the surrounding silicon, this hotspot thermal resistance is only applicable to the specific conditions on both the hotspot and background regions. Changing the background flow rate or heat flux will affect this apparent hotspot thermal resistance [48].

Lastly, the temperature of the fluid pumped into the hotspot microgap was reduced below ambient. With a hotspot heat flux density of 6.1 kW/cm^2, the hotspot intlet temperature was reduced to 7.4 °C (measured close to the inlet port of the chip) at a flow rate of 5 mL/min. This caused the hotspot temperature measurement to drop by 4.8 °C relative to the temperature with the fluid inlet at ambient temperature. In this case, the total volume of fluid delivered to the hotspot is 5% of the background flow rate. Therefore,

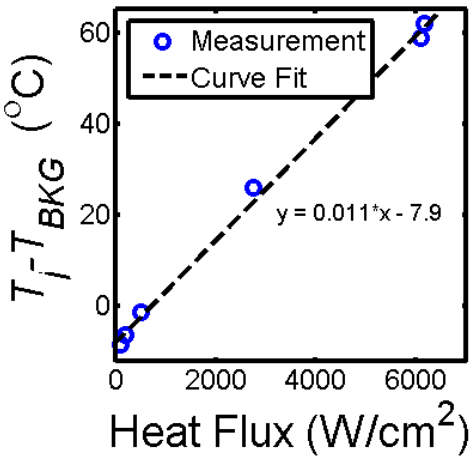


Figure 2.16: Hotspot temperatue vs. heat flux

cooling the hotspot flow requires significantly less energy and infrastructure than cooling the background flow.

Conclusions

In this chapter, microgaps of depth 10 µm, or less, were explored as a means of dissipating extremely high heat fluxes over a small region of a IC. A maximum heat flux density of 6.175 kW/cm^2 was dissipated across an area of 200 µm × 200 µm which represents a heat flux density far beyond those normally seen in modern ICs. Using a dedicated inlet for the hotspot cooler also allows independent and temporally dynamic control of the two temperatures at run time, which could be used to maintain constant temperatures with varying workloads. By chilling the fluid used for the hotspot, the temperature of the hotspot can be decreased while using a fraction of the energy which would be required to chill the fluid for the entire chip.

These dedicated microgap devices are also relatively complex devices, with the process flow in Figure 2.13 omitting many details for brevity. In addition to the complex fabrication process, the use of multiple inlets and outlets would complicate fluid delivery if this concept were to be implemented into a full system. Reducing the temperature of the inlet fluid

below ambient temperature also introduces the risk of condensation in other parts of the the system. Even with the microgap footprint being severely limited, the pressure drop across the gap was high enough to create challenges related to reliability and pump sourcing. The first generation of microgap, which had a depth of only 5.6 μm experienced clogging issues, where contaminants in the flow loop were deposited in the microgap as the refrigerant boiled, limiting the usable life of each test device. Lastly, it was found that the relatively thick silicon base of these devices allowed a substantial portion of the heat flux across the microgaps to be spread through the surrounding silicon. This effect is positive if the goal is to dissipate very large heat fluxes, but can be negative if we wish to prevent the hotspot from heating adjacent areas of the chip.

In the next chapter, a simpler, but still effective, approach for cooling combined background/hotspot heat fluxes will be explored.

CHAPTER 3

INTEGRATED CIRCUIT COOLING USING NON-UNIFORM MICROPIN-FIN ARRAYS FOR NON-UNIFORM POWER MAPS

As mentioned in the introduction to this thesis, many techniques have been explored for cooling hotspots, including [27], liquid jet impingement [28, 29, 30], thin film evaporation [31], dedicated microgap coolers [24, 49], and heat spreading through highly thermally conductive materials, such as graphene [32, 30]. While all of these approaches have their advantages and disadvantages, they all require relatively complicated fabrication processes. Solutions such as thermoelectric coolers (and to some extend heat spreaders) can reduce temperatures of small hotspots, but still require a means of dissipating large background heat fluxes.

Green et al. used a 3D strip model in ANSYS FLUENT to investigate the effect of micropin-fin clustering for cooling of hotspots [50]. A local doubling of pin-fin density was found to reduce local thermal resistance by a factor of roughly two. The strip model included symmetry constraints on either side, effectively simulating an infinite array of identical "cooling strips", with a high density pin-fin cluster spanning that entire channel.

Heterogeneous micropin-fin arrays are used in this work to cool a uniform background heat flux with a higher heat flux, hotspot region. Since flow could be diverted around a high density pin-fin cluster not spanning the entire width of the channel, two types of clusters were chosen for experimental investigation: local clustering over the hotspot, and clustering spanning the entire width of the channel. Pin-fins with both circular and hydrofoil shaped cross sections were also tested.

This chapter is organized as follows. First, an analysis of the effect of lateral heat spreading through the base of the heat sink is presented. Then fabrication results are shown for micropin-fin arrays fabricated with a wide range of dimensions on a single wafer. Next,

a thermal testbed and the four test chips with heterogeneous micropin-fin arrays are described. Lastly, thermal measurements of the four different micropin-fin test chips are presented and analyzed.

Effect of Heat Spreading

Even with a uniform micropin-fin array, heat spreading through the silicon bulk will act to partially mitigate hotspots. In order to quantify the effect of spreading as a function of base thickness and the ratio of hotspot heat flux to background heat flux, heat transfer simulations were performed with COMSOL 4.3b. A 3D model of a $1\,\text{cm} \times 1\,\text{cm}$ chip was created with a $500\,\mu\text{m} \times 500\,\mu\text{m}$ hotspot region in the center. Symmetry boundary conditions were utilized to only simulate a quarter of this geometry. A maximum mesh size of $10\,\mu\text{m}$ was used in the region over the hotspot and a maximum mesh size of $50\,\mu\text{m}$ was used over the background region, with a maximum growth rate of 1.1 between the two regions.

On the bottom of the chip slice, two heat fluxes were applied, one to the hotspot (q_{hs}) and one to the background region (q_{bg}). On the top side of the chip, two heat transfer coefficients were applied, one over the hotspot region (h_{hs}), and another over the remaining background region (h_{bg}). An illustration of the model can be seen in Figure 3.1. The reference temperature on the convective boundary conditions, h_{hs} and h_{bg}, was set to a constant $20\,^{\circ}\text{C}$. Although a microfluidic heat sink would have a gradient in this boundary temperature due to heating of the fluid, using a linear temperature gradient across the chip does not affect the average temperatures in the simulations. The hotspot heat flux can locally heat the fluid, but this effect is expected to be small since the hotspot area is 0.25% of the total chip area. Therefore, the hotspot power is a small fraction of the total power, even when the hotspot heat flux density is several times the average heat flux density.

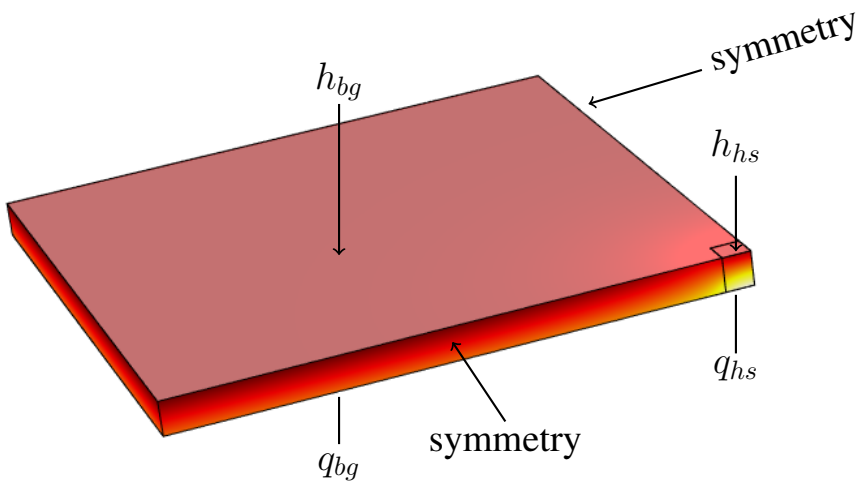


Figure 3.1: COMSOL heat spreading model

Effect of Base Thickness on Hotspot Temperature

In the first parametric study, the effect of base thickness on hotspot temperature was simulated for a silicon thickness range of 0 µm to 500 µm. The background heat flux was set to 250 W/cm^2, while the hotspot heat flux was set to 500 W/cm^2. The background and hotspot heat transfer coefficients were both set to 100,000 W/(°C m^2), representing a uniform micropin-fin array. Figure 3.2 shows the average background and hotspot temperature rise above ambient as a function of base thickness.

The background temperature rise above ambient is approximately a linear function of base thickness due to the conductive thermal resistance of the silicon. The hotspot temperature, on the other hand, initially decreases with increasing silicon base thickness due to lateral heat spreading through the silicon. Spreading then rapidly tapers off with increasing silicon base thickness and the hotspot temperature reaches a minimum at a thickness of approximately 130 µm. The hotspot temperature then increases with increasing thickness due to the increasing conductive thermal resistance of the silicon.

The hotspot temperature can also be influenced by changing the local heat transfer coefficient over the hotspot. Figure 3.3 shows the average hotspot and background temperatures when the hotspot heat transfer coefficient is $1 \times h_{bg}$, $2 \times h_{bg}$, and $5 \times h_{bg}$. The average

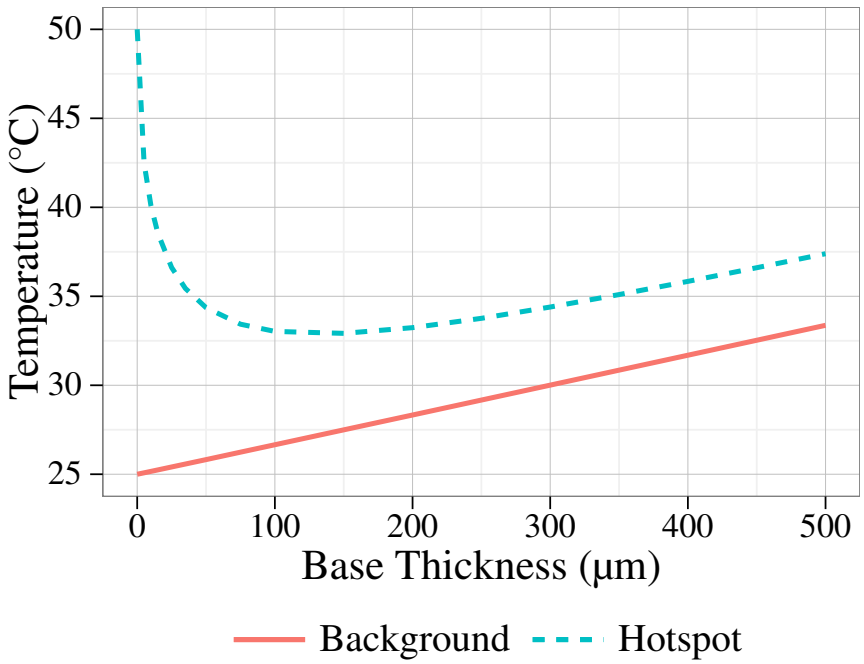


Figure 3.2: Simulated surface averaged hotspot and background temperature rise above ambient vs. silicon base thickness with uniform heat transfer coefficient (100,000 W/(m^2 °C)), hotspot heat flux of 500 W/cm^2 and background heat flux of 250 W/cm^2

background temperatures are relatively unaffected by the much smaller hotspot conditions, only varying by less than 0.3% between these three cases, so they are drawn as a single line, averaging the three cases. While heat spreading through the base can decrease hotspot temperature with uniform cooling, spreading increases the temperature in the cases with enhanced cooling over the hotspot shown in Figure 3.3.

Measuring Thermal Resistance at the Hotspot

Without heat spreading through the base silicon, one could characterize the heterogeneous heat sink by calculating separate background and hotspot thermal resistances (at specific flow conditions). However, when we define the hotspot thermal resistance as $R_{hs} = (T_{hs} - T_{amb}) / q_{hs}$, where T_{hs} is the average hotspot temperature and T_{amb} is the reference temperature on the convective boundary condition, the apparent thermal resistance at the hotspot (and to a much lesser extent, the background region) depends on the

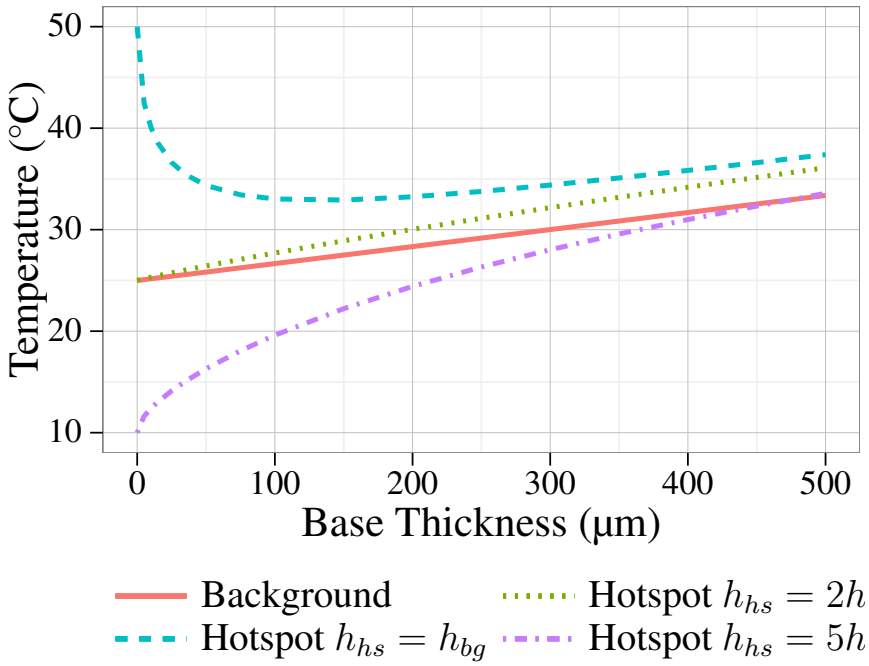
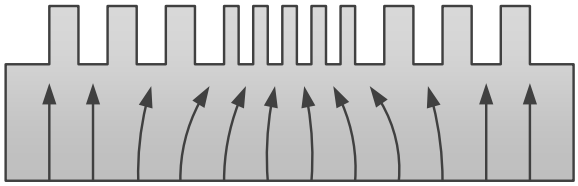
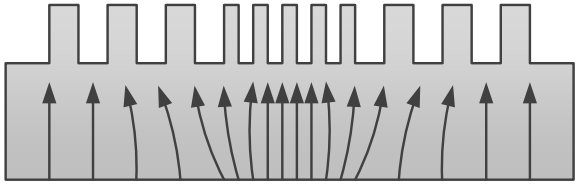


Figure 3.3: Simulated surface averaged hotspot and background temperature rise above ambient vs. silicon base thickness with a hotspot heat flux of 500 W/cm^2, a background heat flux of 250 W/cm^2, a background heat transfer coefficient (h_{bg}) of 100,000 W/(m^2 °C) and hotspot heat transfer coefficients of $1 \times h_{bg}$, $2 \times h_{bg}$, and $5 \times h_{bg}$



(a) Uniform heat flux



(b) High hotspot heat flux

Figure 3.4: Direction of heat flow in the substrate for different power maps

power map applied. When a uniform power map is applied, some heat from the background region is transferred to the high density cluster, making the apparent thermal resistance greater than it would be with no spreading (with a very thin base). However, as local heat flux increases, heat begins to leave the dense cluster and finds another path by spreading through the base, causing the apparent thermal resistance to decrease. This effect is illustrated in Figure 3.4.

The effective thermal resistance at the hotspot, R_{hs}, is shown in Figure 3.5 where the silicon base thickness is 300 μm and the background heat flux is 250 W/cm^2. Three lines are shown, one where $h_{hs} = 100{,}000\,\mathrm{W/(m^2\,^\circ C)}$, one where $h_{hs} = 200{,}000\,\mathrm{W/(m^2\,^\circ C)}$, and one where $h_{hs} = 500{,}000\,\mathrm{W/(m^2\,^\circ C)}$. Due to heat spreading to the surrounding background region, the apparent hotspot thermal resistance changes with hotspot heat flux. Therefore, in section IV, hotspot performance will be reported as temperatures relative to background temperatures.

Fabrication of Multiple Micropin-fin Densities on a Single Wafer

In order to locally tailor heat transfer to the fluid, micropin-fins of varying dimensions must be built on the same wafer and, ideally, in the same processing steps. The micropin-

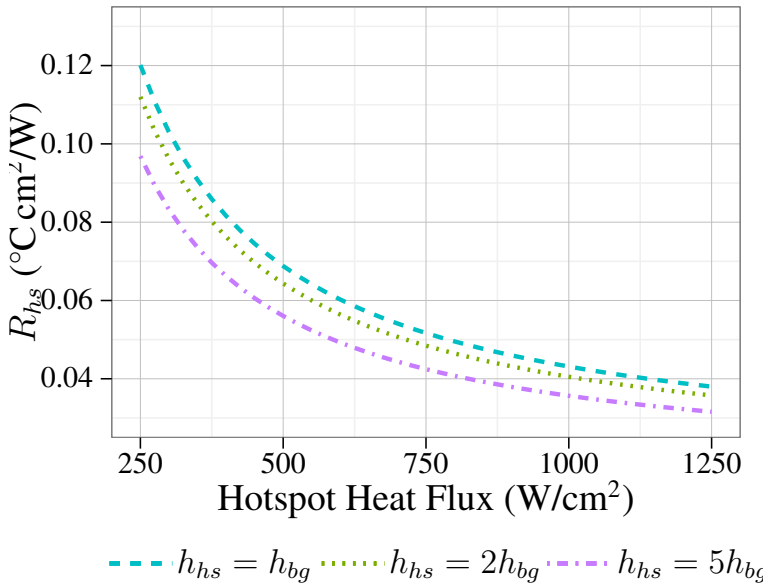


Figure 3.5: Effective hotspot thermal resistance (R_{hs}) vs. hotspot heat flux (q_{hs})

fins used in this study were etched into silicon using the Bosch process and an STS ICP (inductively coupled plasma) machine. The process consists of a reactive ion etching step with an SF_6 plasma, followed by a passivation step with C_4F_8. RF power is applied to the wafer platen in order to create a bias which accelerates ions towards the wafer.

To produce heterogeneous micropin-fin arrays, the Bosch process used for etching had to be optimized to produce reasonable results for all dimensions with a single wafer-level batch process. This is particularly challenging for high aspect ratio micropin-fins, which can be desirable from a thermal perspective for their large surface area. While ideal etching conditions vary between very dense and sparse micropin-fin arrays, a recipe was developed which produced acceptable results for both.

As seen in Figure 3.6, the largest issue was micropin-fin sidewall tapering on very sparse micropin-fin arrays. Significant tapering could decrease fin efficiency and produce results which deviate from expected results with cylindrical micropin-fins. This was improved by increasing the ratio of passivation time to etching time. Platen power was increased during the etch step, and the total cycle time was increased relative to the default trench etching recipe, yielding the final etching recipe consisting of a 14 second etch step, a

Table 3.1: Cylindrical micropin-fin dimensions built on a single wafer

Die Number	Diameter	Transverse Pitch	Lateral Pitch
Die 1	30 µm	90 µm	90 µm
Die 2	60 µm	240 µm	240 µm
Die 3	120 µm	420 µm	180 µm
Die 4	30 µm	90 µm	30 µm
Die 5	30 µm	75 µm	36 µm
Die 6	30 µm	120 µm	60 µm

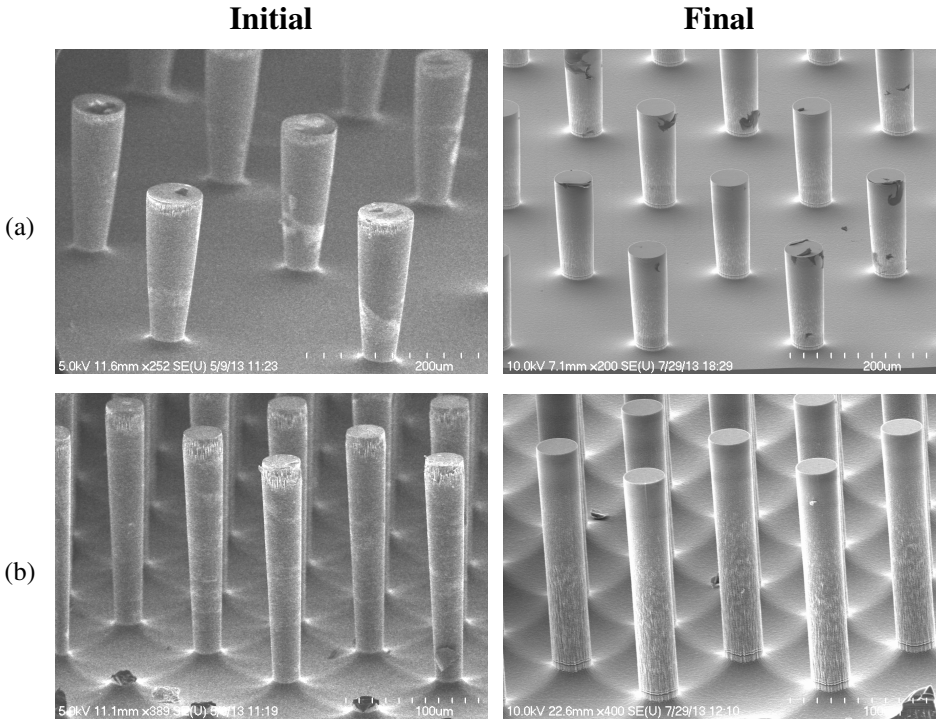
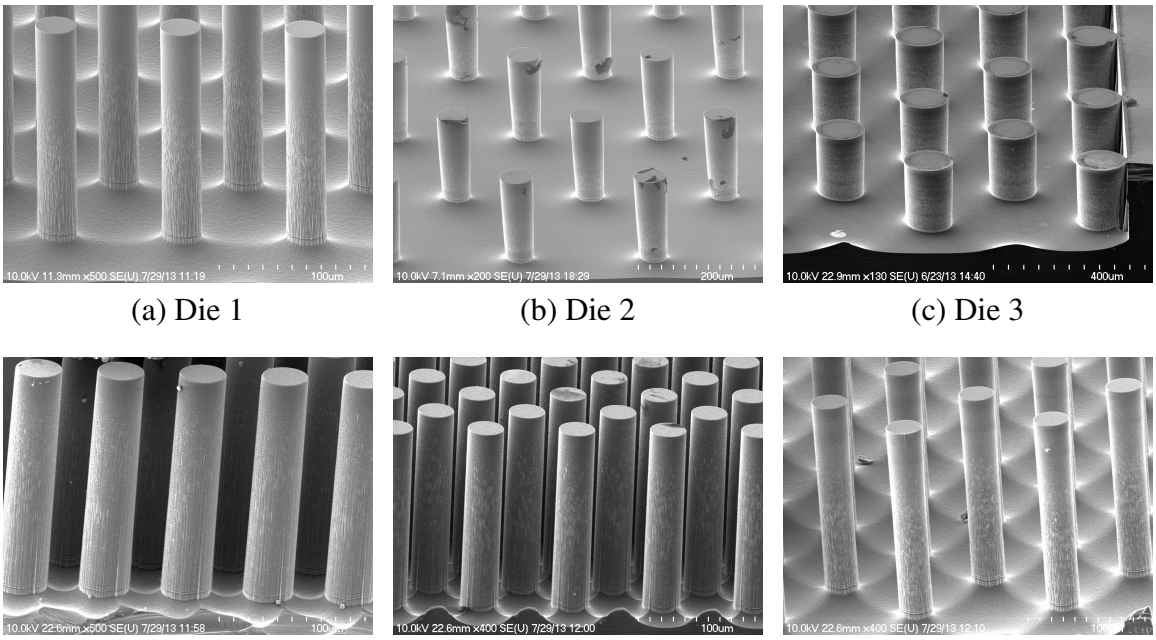


Figure 3.6: SEM images of (a) sparse and (b) dense, high aspect ratio micropin-fins before and after etching process optimization.



(a) Die 1 (b) Die 2 (c) Die 3

(d) Die 4 (e) Die 5 (f) Die 6

Figure 3.7: Six micropin-fin dice etched using a single batch process on the same wafer

17.5 second passivation step, and a platen power of 20 W. This recipe was used to produce the results shown in Figure 3.7, which shows sidewall profiles of six different micropin-fin arrays built on a single wafer. Dense, sparse, and high aspect ratio dimensions are included and can be seen in Table 3.1. Heat transfer and pressure drop studies were carried out with these samples in [51]. Four of the micropin-fin arrays have a micropin-fin diameter of 30 µm, which is substantially smaller than the majority of micropin-fins studied in literature. Since all micropin-fins were etched to a nominal height of 200 µm, these small diameter micropin-fins have an aspect ratio of 6.7:1, which is also necessary for high surface area enhancement, but makes sidewall profile crucial.

Thermal Testbed and Heterogeneous Micropin-fin Samples

In this work, chip heat flux is represented by a "background" heat flux and a "hotspot" heat flux. The hotspot is a smaller region of the chip, with a heat flux considerably higher than that of the background region. By locally clustering micropin-fins over a hotspot, it

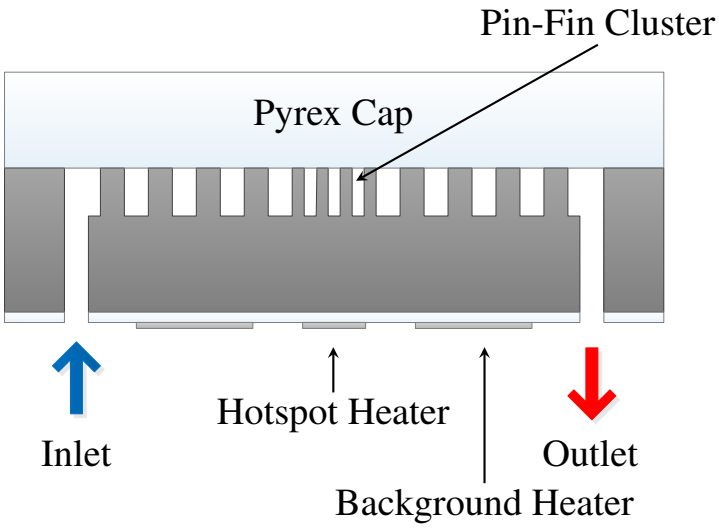


Figure 3.8: Heterogeneous micropin-fin test chip cross section

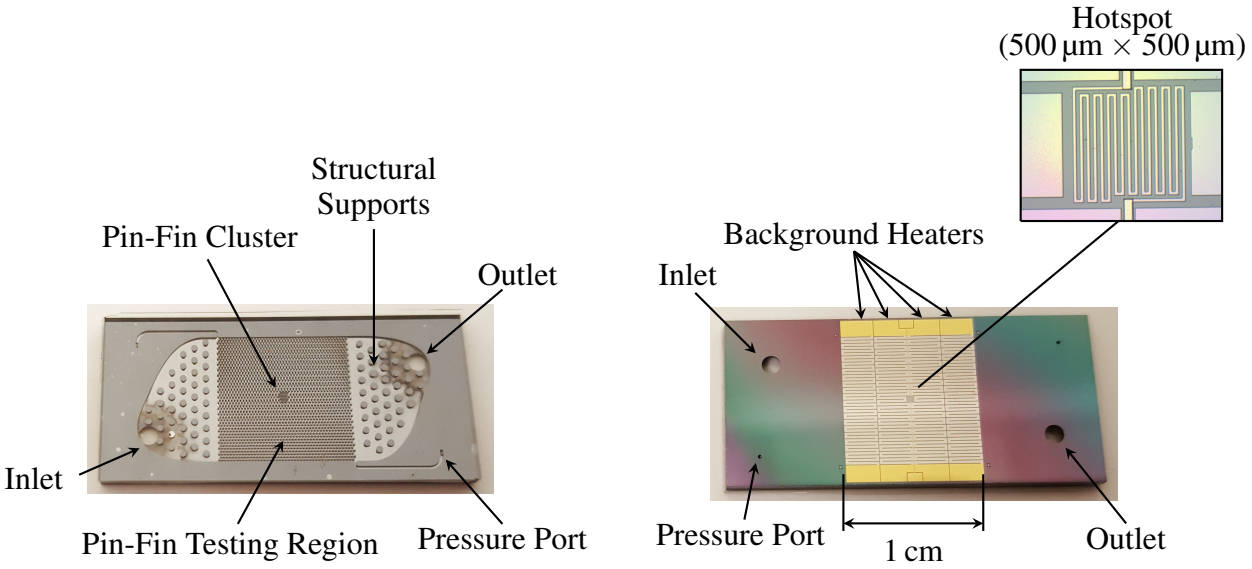
is expected that the local heat transfer coefficient can be increased to deal with the hotspot heat fluxes, while minimizing the pressure drop penalty compared to using this high density clustering over the entire background region.

The etching recipe discussed above was applied to heterogeneous micropin-fin arrays with local clustering of micropin-fins over hotspots. First, the micropin-fins were etched into a 500 µm double side polished silicon wafer. A pyrex cap was anodically bonded to the micropin-fins to seal the channels. Next, a 1.9 µm thick silicon dioxide layer was deposited on the back side of the wafer, on top of which 0.2 µm thick platinum heaters were deposited. The heaters also acted as resistance temperature detectors (RTDs). Finally, fluid inlet/outlet and pressure measurement ports were etched through the silicon. A cross-sectional diagram of the test chips can be seen in Figure 3.8.

While enhanced heat transfer is only necessary over the hotspot, local clustering over this region is expected to result in some flow diversion around the higher density clustering. Therefore, for both the hydrofoil and circular micropin-fin designs, clustering is done in a small region over the hotspot as well as in a region spanning the entire width of the channel to prevent this flow diversion. Four different background/hotspot test devices were fabricated, with two types of micropin-fin cross-sectional shapes, and two types of clustering

for hotspot cooling:

1. Cylindrical micropin-fins with local clustering
2. Cylindrical micropin-fins with span-wise clustering
3. Hydrofoil micropin-fins with local clustering
4. Hydrofoil micropin-fins with span-wise clustering.

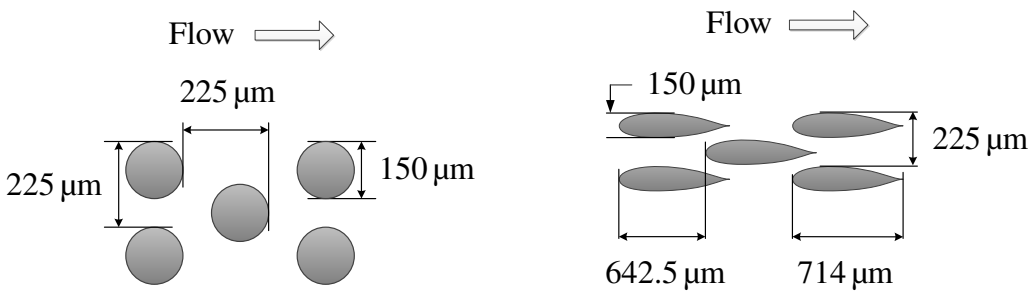


(a) Top view of micropin-fins

(b) Bottom view of heaters/RTDs

Figure 3.9: Test chip (a) top view of etched silicon through pyrex cap (b) bottom view of heater and ports

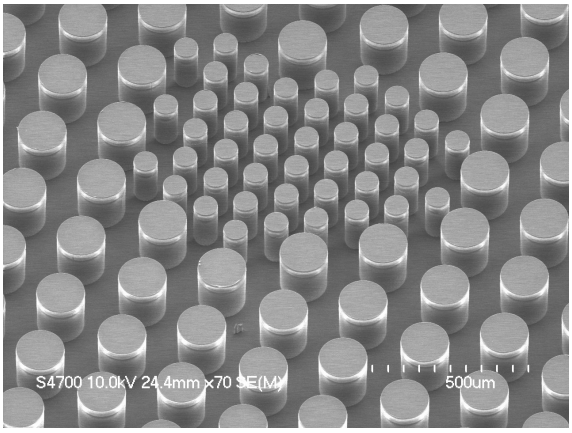
Images of the top and bottom of a micropin-fin test device can be seen in Figure 3.9. The background micropin-fin array covers an area of 1 cm $\times$ 1 cm. The hotspot consists of an area of 500 µm $\times$ 500 µm in the center of the chip. This region is heated by a dedicated serpentine platinum heater on the bottom of the chip while background heat flux is applied through four heaters spanning the chip. Pressure ports at the inlet and outlet sides of the background micropin-fin array provide accurate pressure drop measurements.



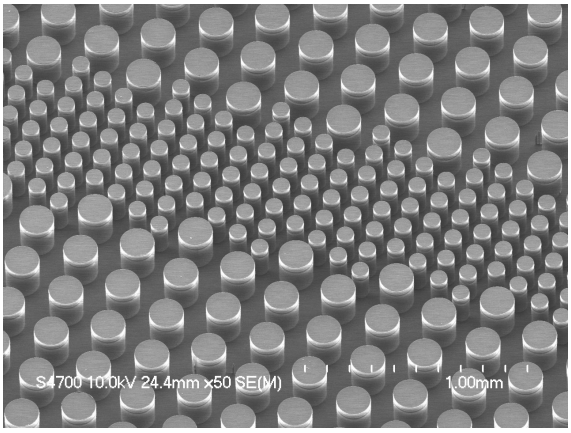
(a) Background circular micropin-fin dimensions

(b) Background hydrofoil micropin-fin dimensions

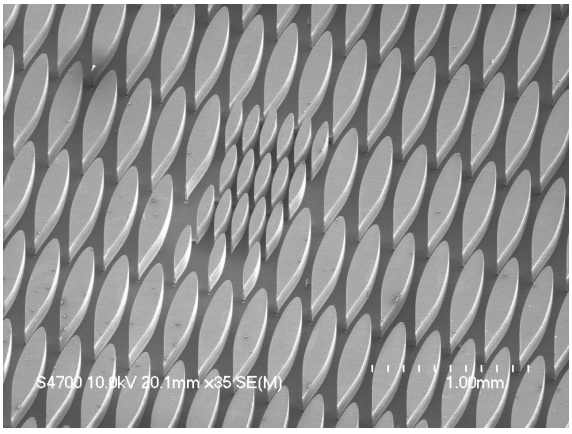
Figure 3.10: Background micropin-fin dimensions



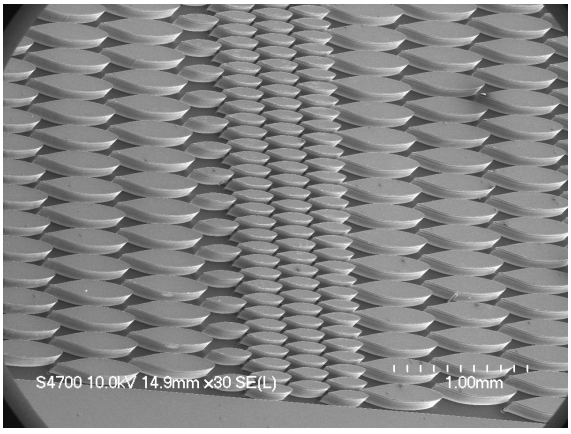
(a) Cylindrical micropin-fins locally clustered over hotspot



(b) Cylindrical micropin-fins spanning entire channel



(c) Hydrofoil micropin-fins locally clustered over hotspot



(d) Hydrofoil micropin-fins spanning entire channel

Figure 3.11: SEM images of of the four micropin-fin devices

Dimensions of the background micropin-fins can be seen in Figure 4.4. The diameter of the cylindrical micropin-fins is 150 µm. The transverse and longitudinal pitches are both 225 µm. The width of the background hydrofoil micropin-fins is 150 µm, and the transverse and longitudinal pitches are 225 µm and 642.5 µm, respectively. All micropin-fins were etched to a nominal height of 200 µm. Pitches and diameters of micropin-fins in the high density clusters were half of the background pitches and diameters. SEM images of the micropin-fin clustering over the hotspots for the four different heat sinks can be seen in 3.11a, 3.11b, 3.11c, and 3.11d.

Each chip was tested in an open loop system with deionized water as a coolant, with an inlet temperature of $21.0 \pm 0.5\,^{\circ}\mathrm{C}$. A diagram of the open flow loop can be seen in Fig 3.12. RTD resistance vs. temperature calibration lines were first generated by measuring RTD resistances at temperatures ranging from approximately 20 °C to approximately 110 °C. The chips were placed in a package with O-ring seals on the ports for fluid delivery and pressure measurement. Fluid temperature was measured at the inlet and outlet of the package with k-type thermocouples calibrated to an accuracy of 0.1 °C over the range of temperatures used in experimentation. Flow rate was measured with a rotameter with a maximum uncertainty of less than 2 mL/min. Pressure drop across the micropin-fin arrays was measured using a digital differential pressure gauge calibrated with an Omega DPI610 calibrator to within 0.1 kPa. RTD power and resistance were recorded with an Agilent data logger.

An image of the packaged device can be seen in Figure 3.13. The test chip was mounted to a PCB with copper traces and a cutout which enabled viewing of the $1\,\mathrm{cm} \times 1\,\mathrm{cm}$ micropin-fin array through the transparent Pyrex cap. Gold bonding pads were wirebonded to the PCB traces, which, in turn, were soldered to wires which delivered power to the device. A PEEK package was clamped to the opposite side of the test chip with 8 screws. The package applied pressure to the O-ring seals on the chip and connected the chip's four ports to the fluid supply and pressure gauge. The package also housed the inlet and outlet

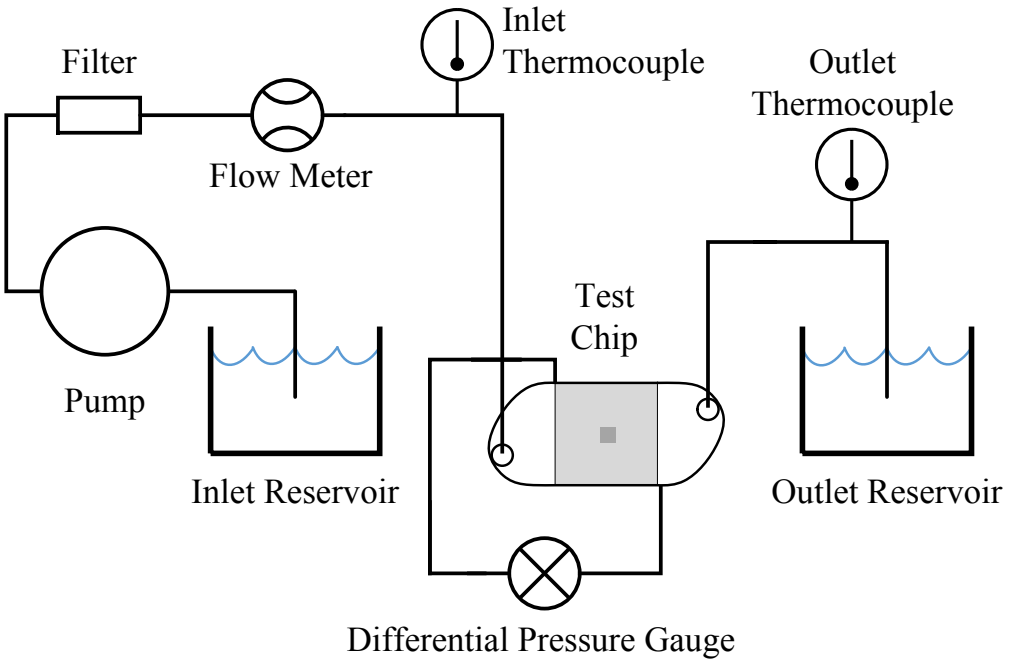


Figure 3.12: Diagram of open loop system used to test chips

thermocouples, which were mounted in the inlet and outlet streams, immediately outside of the test chip.

Each experiment was run twice and the averages of the results from the two runs are reported in the next section. In order to quantify the uncertainty in the RTD temperature measurements as well as the pressure measurements, combined standard deviations across these repeated runs were calculated and are reported in Table 3.2 along with the measurement device uncertainties.

The amount of heat lost to the ambient surroundings can be calculated from measured quantities according to

$$Q_{loss} = Q_{in} - \dot{m} C_p \left(T_{out} - T_{in} \right) \tag{3.1}$$

where $\dot{m}$ is the water mass flow rate, C_p is the specific heat of water (approximately

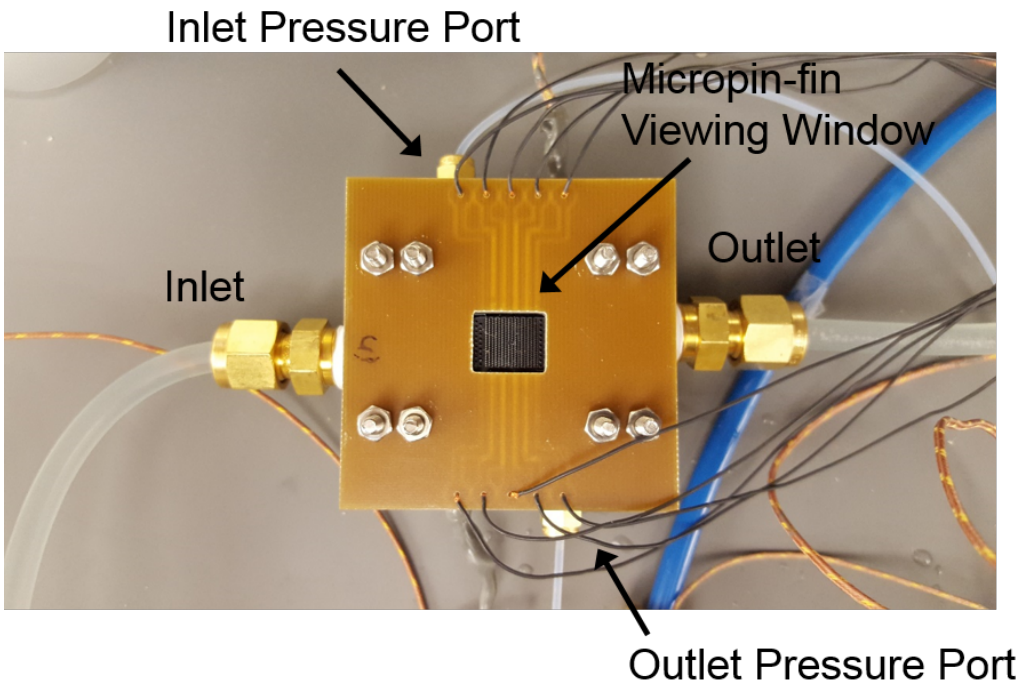


Figure 3.13: Photo of a packaged heterogeneous micropin-fin test device

Table 3.2: Measurement uncertainties

Pressure Gauge Accuracy	0.1 kPa
Thermocouple Measurement Accuracy	0.1 °C
Flow Rate Measurement Accuracy	2 mL/min
Combined Temperature Standard Deviation	0.33 °C
Combined Pressure Standard Deviation	1.3 kPa

4.18 J/(g °C)), and T_{in} and T_{out} are the measured water inlet and outlet temperatures, respectively. Heat loss to ambient was found to be below 3.35% for all data points.

Due to the high heat fluxes and low convective thermal resistances in these experiments, temperature drop across the 1.9 µm thick silicon dioxide insulation under the platinum heaters was a substantial portion of the temperature difference between the RTDs and the fluid. This temperature drop across the silicon dioxide layer was estimated through heat conduction simulations in COMSOL, taking the thermal conductivity of PECVD silicon dioxide to be 1.1 W/(m °C) [52, 53]. Since the serpentine heaters do not provide

completely uniform heat flux, the temperature drop across the silicon dioxide is higher than 1D conduction calculations would predict. The temperature drop from the RTD to the silicon was found to be 7.9 °C and 15.8 °C for nominal hotspot heat fluxes of 250 W/cm^2 and 500 W/cm^2, respectively. From the background RTDs to the silicon, this number was 5.2 °C at 250 W/cm^2. These numbers were used to compute the temperatures at the silicon surface, called the junction temperature, or T_j.

Experimental Results

Figure 3.14 shows junction temperature rise above inlet temperature ($T_j - T_{in}$) vs. axial position in the direction of fluid flow (x) for all four types of chips at a flow rate of 208 mL/min and a uniform power of 246 W across the 1 cm $\times$ 1 cm chips. Maximum variation in heat flux between the four heaters was found to be between 1% and 4%. With a uniform micropin-fin array, uniform power, constant fluid heat capacity, and no edge effects, one would expect the junction temperature to increase linearly with axial position. A line was fit to the four background junction temperatures using least squares regression. As can be seen, the temperature does rise linearly, although the first and last temperatures tend to be below the fit line, while the center heaters tend to be above. This could be a result of heat spreading to the inlet and outlet plena as well as flow development in these regions.

The slopes of the lines in Figure 3.14 have units of °C/mm and depend primarily on flow rate and heat flux rather than the micropin-fin dimensions. As expected, all four lines have the same slope. By multiplying the slope by the length of the chip, the change in fluid temperature across the 1 cm length of the micropin-fin region can be estimated. Since this gradient across the micropin-fin region of the chip only accounts for approximately 85% of the measured $T_{out} - T_{in}$, it is estimated that approximately 15% of the heat flux was spread to the inlet and outlet plena where it was transferred to the water.

As can be seen, the background T_j values for all four chips are very similar. A larger difference can be seen in the hotspot temperatures. Since the hotspots have the same power

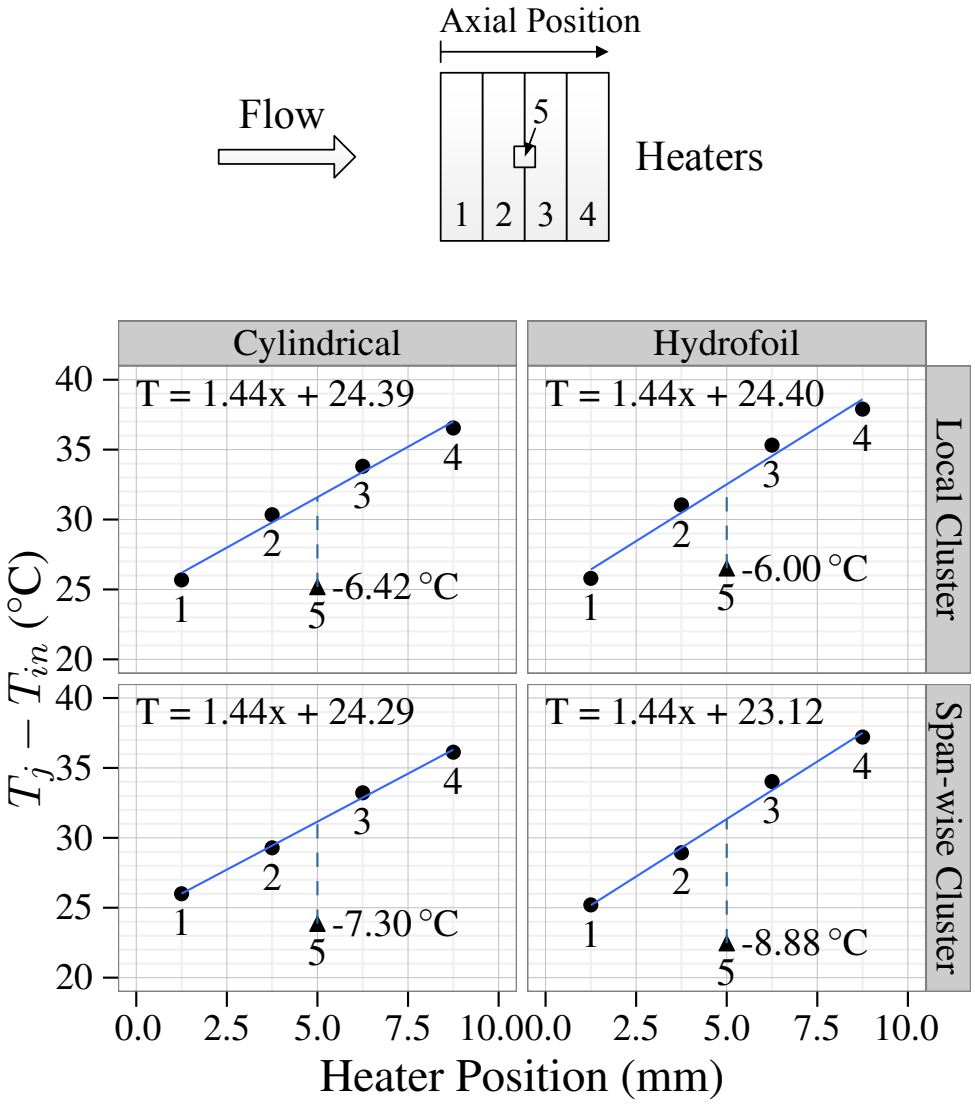


Figure 3.14: Junction temperature rise above inlet temperature vs. axial position in direction of fluid flow with a uniform 246 W power. Hotspot labels represent temperature deviation from expected temperature with a uniform micropin-fin density.

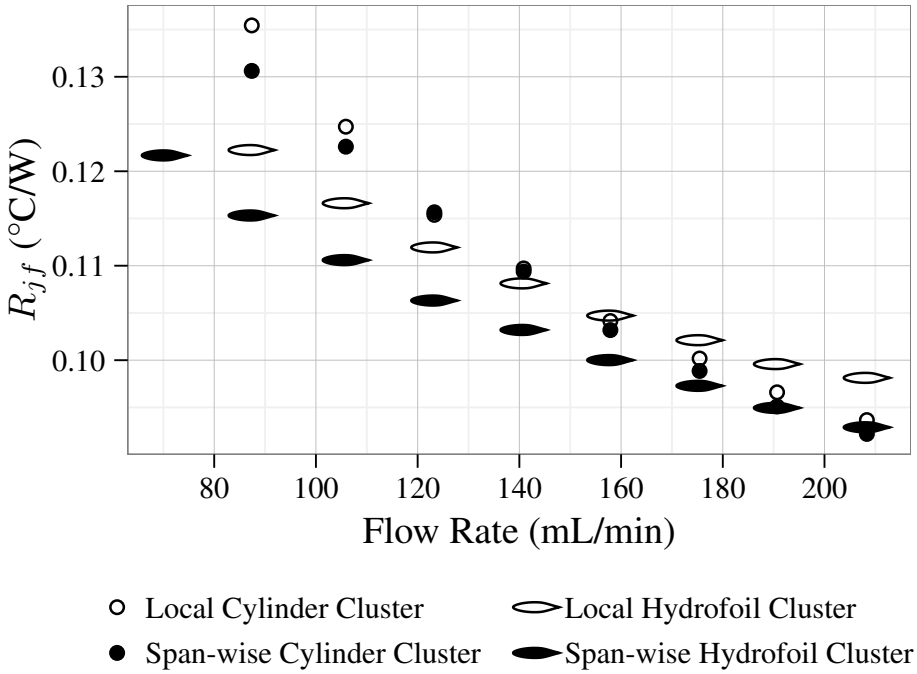


Figure 3.15: Thermal resistance from the silicon to fluid (R_{jf}) vs. flow rate for all four dice

density as the rest of the chip for this dataset, the temperature would be expected to fall on the best fit line with uniform micropin-fins. The hotspot temperature deviation from the line is shown on each plot and illustrates the performance improvement from the high density clustering. All four designs offered a performance improvement through clustering, but the hydrofoil cluster spanning the entire width of the chip performed best, reducing temperature by 8.88 °C vs. the background.

The average background junction-to-fluid thermal resistance, $R_{jf} = R_{cond} + R_{conv}$, can be found by subtracting the average fluid temperature from the average junction temperature and dividing by the chip power. Figure 3.15 shows background R_{jf} vs. flow rate for all four test chips. It should be noted that this includes any effects of the higher density micropin-fin cluster on the background measurements, which can be significant for the micropin-fins spanning the width of the channel, as they cover a larger surface area and are located directly above sections of the background RTDs. Therefore, it is unsurprising that the average background thermal resistance of the hydrofoil sample with span-wise clustering is lower than that of the hydrofoil sample with local clustering, despite the same

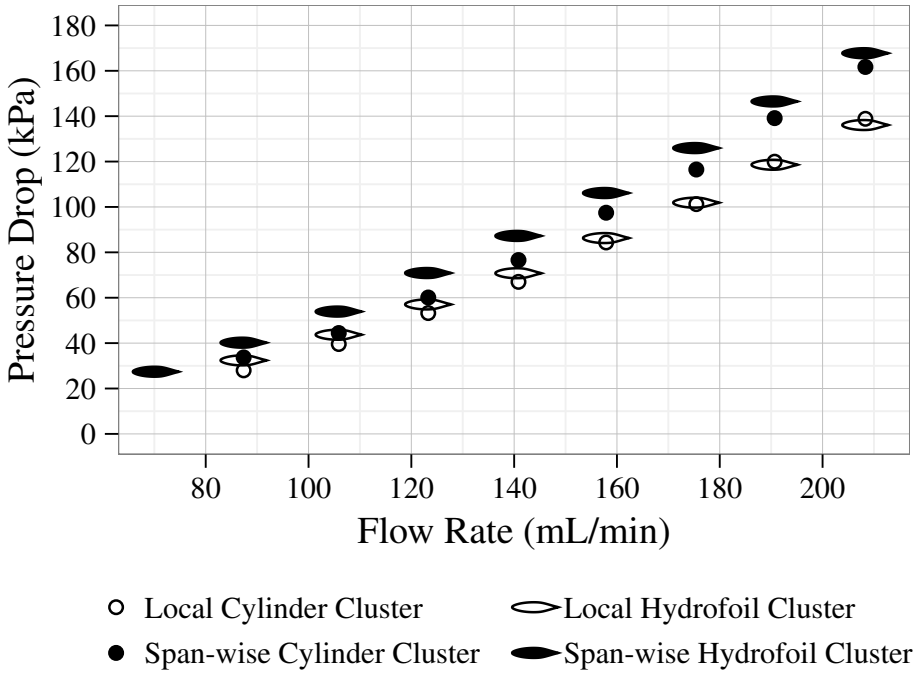


Figure 3.16: Pressure drop vs. flow rate for all four dice

dimensions of micropin-fins over the rest of the sample.

Pressure drop vs. flow rate for the four dice can be seen in Figure 3.16. The four samples have very similar pressure drops, but, unsurprisingly, the two samples with the high density clustering spanning the width of the channel have the highest pressure drops. The hydrofoil sample with span-wise clustering had the lowest background and hotspot temperatures, but also had the highest pressure drop.

The Reynolds number was calculated for all four dice. With the characteristic length defined as the hydraulic diameter of the gap between micropin-fins, the Reynolds number ranges from a minimum of 190.5, at the lowest flow rate, to a maximum of 873, at the highest flow rate. Beyond a Reynolds number around 100-300, a flow transition leading to higher pressure drops has previously been observed in micropin-fin arrays[19, 22]. This transition was shown by Renfer et al. to occur when the flow transitions from steady laminar flow to a regime with vortex shedding [47]. Since most of the Reynolds numbers in the current work were close to or above this range for all of the flow rates and chips tested, the flow was likely in this flow regime with vortex shedding.

Non-Uniform Heat Flux

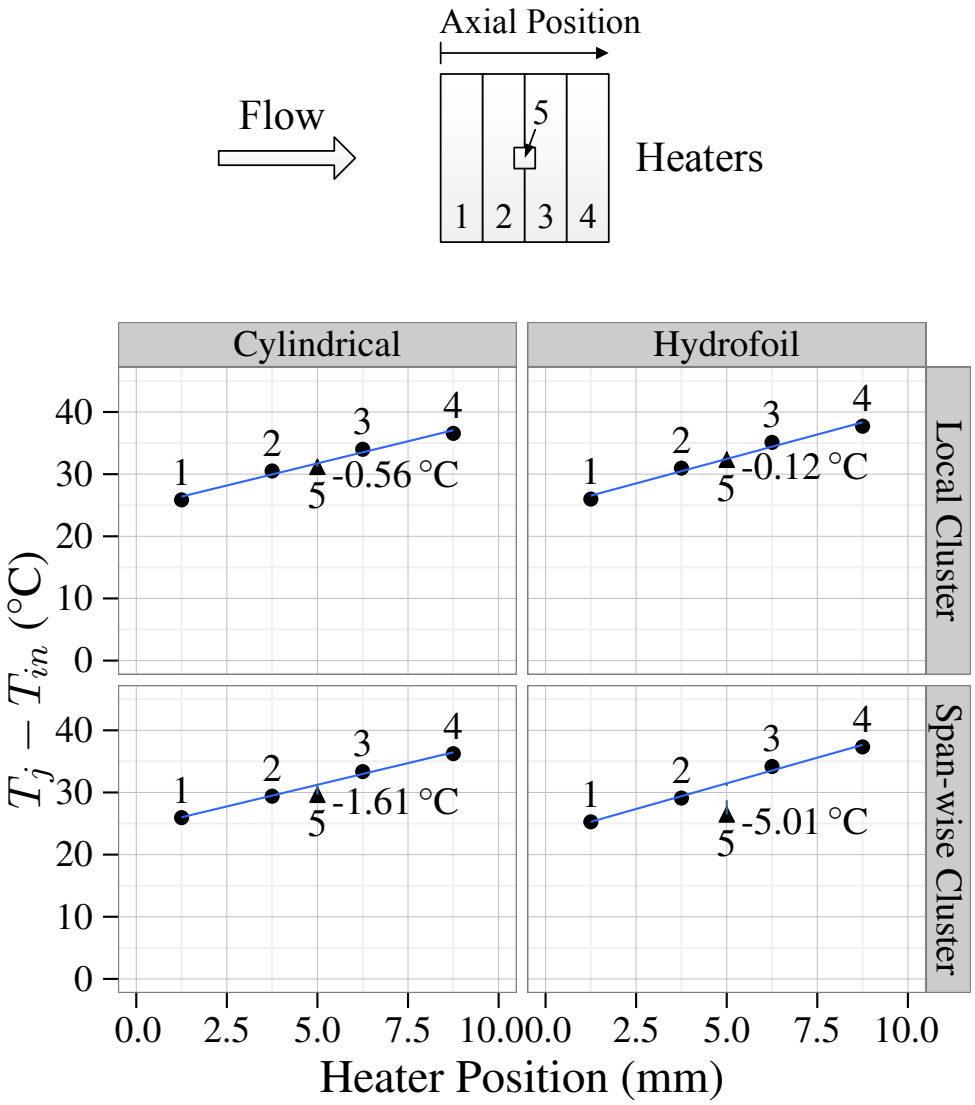
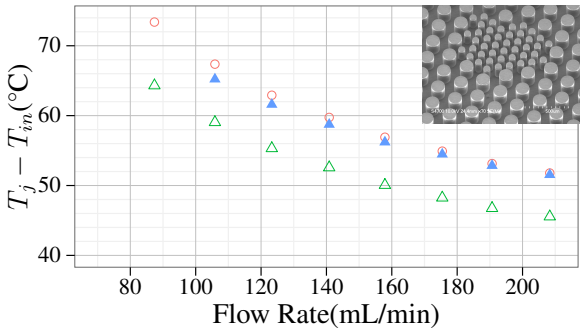


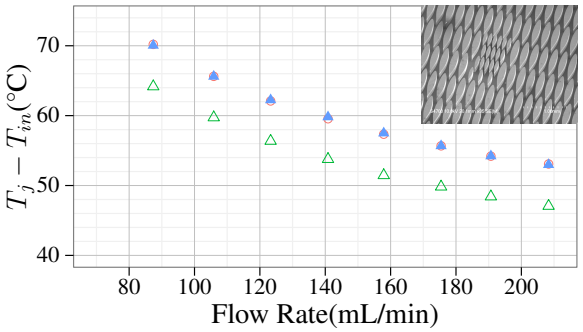
Figure 3.17: Junction temperature rise above inlet temperature vs. axial position in direction of fluid flow with 500 W/cm^2 hotspot heat flux and 250 W/cm^2 background heat flux. Hotspot labels represent temperature deviation from expected temperature with a uniform micropin-fin density.

Lastly, all four devices were tested with a nominal background heat flux of 250 W/cm^2 and a hotspot heat flux of 500 W/cm^2. Results at the highest flow rate of 208 mL/min can be seen in Figure 3.17. Despite doubling the heat flux relative to the background region, hotspot temperatures remain below the background centerline temperatures.

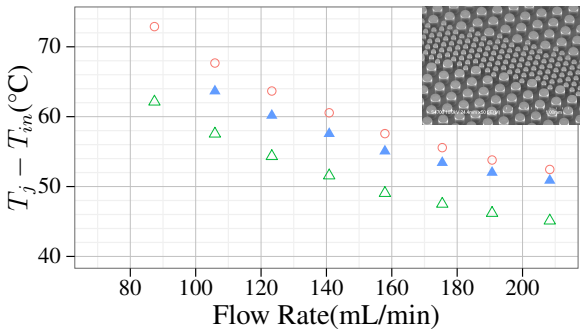
Background and hotspot temperatures are shown for all four devices as a function of flow rate in Figure 3.18. Hotspot junction temperature rise above inlet temperature is shown for nominal hotspot heat fluxes of 250 W/cm^2 and 500 W/cm^2 with a background heat flux of 250 W/cm^2. The average background temperature across all four background heaters is also shown. Since the total hotspot power was much lower than the background power (due to its smaller size), the average background temperature was minimally effected by the hotspot heat flux. Therefore, a single set of background temperature points is shown, representing the average background temperatures between the experiments with the two hotspot heat fluxes. Each of these average background temperature values varies by less than 0.5% from the measured temperatures in either of the experiments.



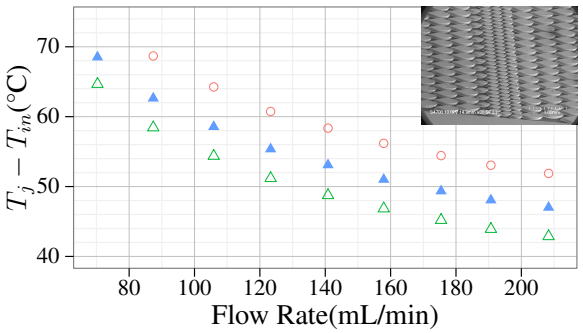
(a) Cylindrical Micropin-fins with local clustering



(b) Hydrofoil Micropin-fins with local clustering



(c) Cylindrical Micropin-fins with span-wise clustering



(d) Hydrofoil Micropin-fins with span-wise clustering

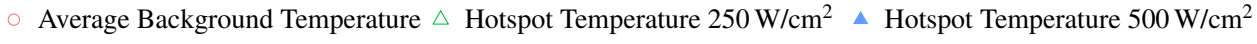


Figure 3.18: Hotspot and background temperatures vs. flow rate of all four chips

As can be seen in Figure 3.18, the hotspot temperature is lower than the average background temperature for all uniform power maps (250 W/cm^2 background and hotspot). When the hotspot heat flux is increased to 500 W/cm^2, the hotspot temperature increases, but remains below average background temperature for most of the four device types. The hydrofoil micropin-fins with local clustering have a hotspot temperature which matches the background temperature very closely when the hotspot heat flux is twice the background heat flux (maximum difference of 0.25 °C, or 0.4%). As seen before, the hydrofoil device with high density micropin-fins spanning the entire width of the channel has the lowest hotspot temperatures and would likely be ideal for even higher hotspot heat fluxes relative to background heat fluxes.

The two devices with micropin-fins spanning the entire width of the channel had lower hotspot temperatures than those with clustering directly over the hotspot. This could be partially due to flow diversion around the cluster when the cluster does not span the entire width of the channel, but also because of heat spreading to the larger high density cluster when the cluster extends beyond the hotspot region.

Conclusion

In this work, heterogeneous micropin-fin arrays were fabricated and tested for the cooling of integrated circuits with hotspots. Two types of micropin-fins (cylindrical and hydrofoil) were tested, each with clustering directly over the hotspot, and clustering spanning the entire width of the channel to prevent flow bypass (a total of four devices).

Of the four devices, the device with the hydrofoil shaped micropin-fins and the dense cluster spanning the entire width of the channel had the lowest hotspot temperatures and the lowest background temperatures at many of the flow rates tested. With a uniform power map and the highest flow rate of 208 mL/min, this sample had a background junction-to-fluid thermal resistance of 0.093 °C cm^2/W.

All four devices effectively reduced hotspot temperatures. With a nominal hotspot heat

flux of 500 W/cm^2 and a nominal background heat flux of 250 W/cm^2, the average junction temperatures between the hotspot and background were matched very closely for all flow rates on both the samples with only local clustering directly over the hotspot. Background and hotspot average temperatures, relative to inlet tempetature, differed by a maximum of 0.4% for all flow rates with the hydrofoil device.

The two devices with micropin-fins spanning the entire width of the channel had lower hotspot temperatures than those with local clustering, at the cost of higher pressure drop. This difference in hotspot thermal performance may partially result from flow diversion around the local clusters, but also likely arises from heat spreading to areas of the high density clusters which are not directly above the hotspot.

Heat spreading to larger clustered areas could be utilized in future work with higher density clusters extending beyond the limits of the hotspot in both dimensions. In this work, average background and hotspot temperatures were used, but more detailed temperature maps could also be investigated to resolve the boundaries between these regions and mitigate maximum temperatures. In fact, continuously varying micropin-fin pitches could also be utilized, potentially eliminating almost all temperature variation across the chip for a given power map.

CHAPTER 4

MONOLITHIC INTEGRATION OF A MICROPIN-FIN HEAT SINK IN A 28 nm FPGA

As more functionality and higher density logic continue to be packed into increasingly dense systems, traditional cooling systems are being pushed to their limit, leading to the problem of dark silicon and throttled performance. Microfluidic cooling, first demonstrated by Tuckerman and Pease [18], has the potential to solve this cooling challenge for high power and high performance integrated circuits. Microfluidic cooling has the potential for very low junction-to-fluid thermal resistance in a very small form factor. Low thermal resistance opens the possibility of cooling very high heat flux integrated circuits with a moderate inlet temperature, or moderate heat fluxes with an elevated inlet temperature. Cooling with elevated inlet temperatures can reduce or eliminate the need for chilling of the coolant below maximum outside ambient temperatures and open the possibility of waste heat reuse, increasing data center energy efficiency [54].

The most common method of cooling microelectronics has long been an air cooled heat sink mounted on top of the packaged integrated circuit, as shown in Fig. 4.1 (a). Efficacy is limited with this solution and can be improved in several ways through the use of microfluidic cooling. First, due to the small heat sink dimensions and properties of the liquid coolants, much lower convective thermal resistances can be achieved with microfluidic cooling when compared to direct air cooling. Additionally, in the traditional configuration shown in Fig. 4.1 (a), there is a large conductive thermal resistance due to the large distance and, more importantly, several material interfaces through which heat must conduct in order to reach the heat sink. In order to improve thermal resistance at these interfaces, two levels of TIM are used, but these interfaces still remain a major bottleneck in total junction-to-ambient thermal resistance. By etching the heat sink directly into the silicon

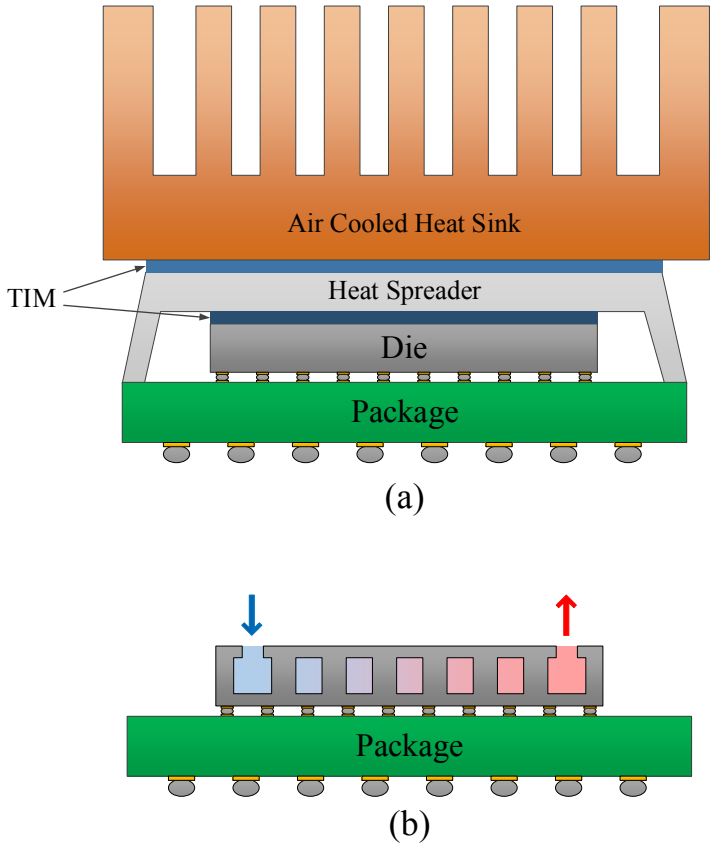


Figure 4.1: (a) Traditional microelectronic system (b) Microelectronic system with monolithically integrated microfluidic heat sink

die, conductive thermal resistance between the heat source and heat sink is minimized.

The very low profile achievable with microfluidic heat sinks also makes them compatible with many dense 2.5D and 3D systems, as shown in Fig. 4.2. The examples shown in Fig. 4.2 use a silicon interposer for signal and fluid routing, but a traditional organic package could also be used.

Integrating the microfluidic heat sink in an interposer, as shown in Fig. 4.2 (a), can offer thermal resistances superior to typical package level air cooled heat sinks, without modifying the logic dice [55]. In order to bring the heat sink as (thermally) close to the area of heat generation as possible, the microfluidic heat sink can be etched into the back side of the active silicon dice, as shown in Fig. 4.2 (b). These microfluidic cooled dice can

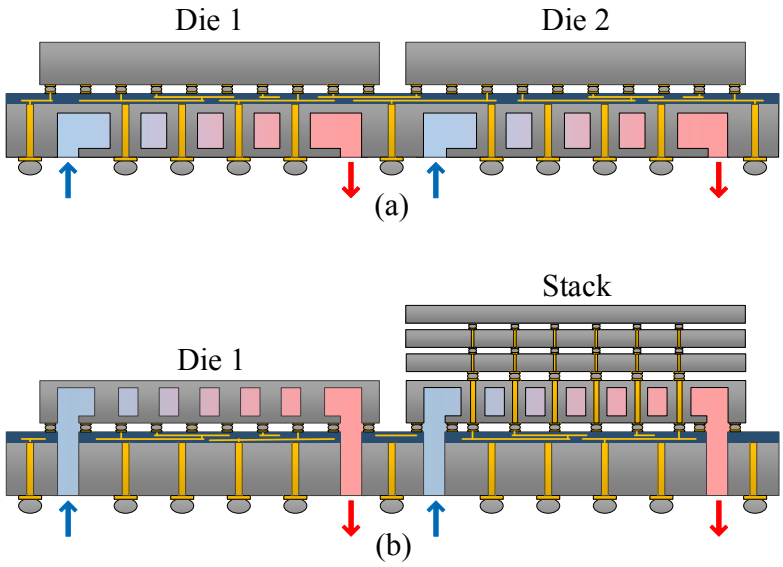


Figure 4.2: Microfluidic cooling integrated in (a) the interposer and (b) the back side of the die

then be stacked to form a 3D stack. Signaling and power delivery are achieved with TSVs passing through the microfluidic heat sinks. Although microfluidic cooling limits how far silicon dice can be thinned, high aspect ratio TSVs can be fabricated in micropin-fins to limit TSV capacitance [37]. Unlike traditional cooling methods implemented on the top of the active silicon, microfluidic heat sinks can be integrated into multiple tiers in a 3D stack, allowing cooling to scale with the number of high power tiers [34] [33]. This could potentially enable the stacking of multiple high power tiers which could not be cooled with a single heat sink.

Since Tuckerman and Pease achieved a thermal resistance of 0.09 °C cm^2/W using microchannels etched into silicon, a great deal of effort has focused on achieving improved thermal resistance and characterizing microchannel or micropin-fin heat sinks with correlations to predict performance [19, 22, 23]. However, research to date has focused on passive silicon dice with resistive heaters representing the heat producing circuitry. In this chapter, we present a functional microfluidic-cooled CMOS circuit.

An Altera Stratix V FPGA, built in a 28 nm process, was post processed to integrate

a micropin-fin heat sink directly into the back of the flip-chip bonded silicon die, a few hundred micrometers from the active circuitry. This microfluidic-cooled FPGA was then tested with deionized water as a coolant at several flow rates and inlet temperatures. All testing was performed with Altera Stratix V digital signal processing (DSP) development boards. A comparison was made with a stock board with the default air cooled heat sink. Testing was performed at flow rates ranging from 0.15 mL/s to 3.0 mL/s and with inlet temperatures ranging from 21 °C to 50 °C.

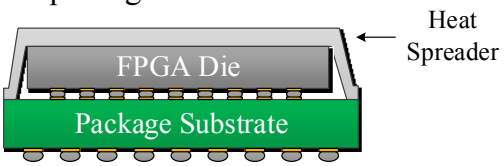
Fabrication

The Stratix V FPGA consists of a silicon die that is flip-chip bonded onto an organic substrate. The back side of the die was used to etch a micropin-fin heat sink in the same bulk silicon as the active circuitry. The Bosch process was used in order to etch silicon with vertical micropin-fin sidewalls. This batch process could be completed at the wafer level, but in this case was applied to a single chip at the die level. This fabrication flow was used for this proof of concept due to the relative ease of acquiring packaged parts, but may be different when optimized for scalability and manufacturing throughput.

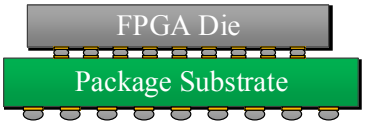
The process used to add microfluidic cooling to a packaged Stratix V FPGA is shown in Fig. 4.3. First, the metal lid was removed along with TIM on the back side of the die (TIM 1). The flip-chip bonded die, along with the package substrate, was then attached to a carrier wafer with cool grease and Kapton tape to protect the package substrate and sides of the die. Photoresist was then spin coated on the exposed back side of the silicon die and the Bosch process was used to etch micropin-fins to a depth of approximately 240 µm. Inlet and outlet plena were formed in the same etching step by etching regions on either side of the micropin-fin array without any micropin-fins. A photograph of the etched die along with the micropin-fin dimensions can be seen in Fig. 4.4.

An SEM image of identical micropin-fins fabricated with the same process in a silicon wafer can be seen in Fig. 4.5. In general, aspect ratio and surrounding features are known

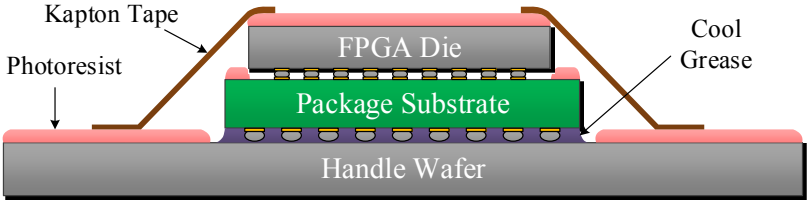
0) Start with a packaged FPGA.



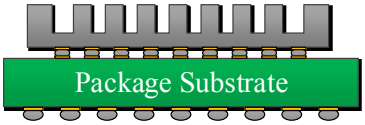
1) Remove the heat spreader and TIM.



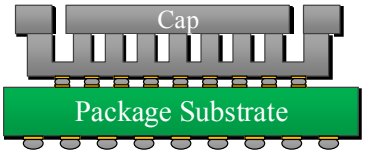
2) Mount to handle wafer and spin coat photoresist.



3) Etch Micropin-fins and remove handle wafer.



4) Attach cap with epoxy.



5) Solder to PCB and attach Nanoports.

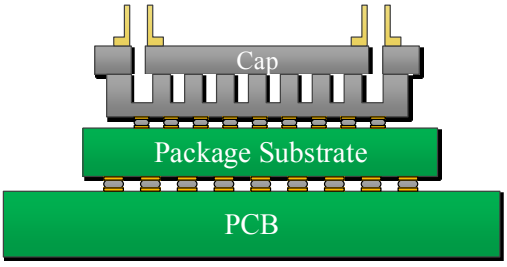


Figure 4.3: Fabrication process for etching micropin-fins into the back side of a packaged FPGA die

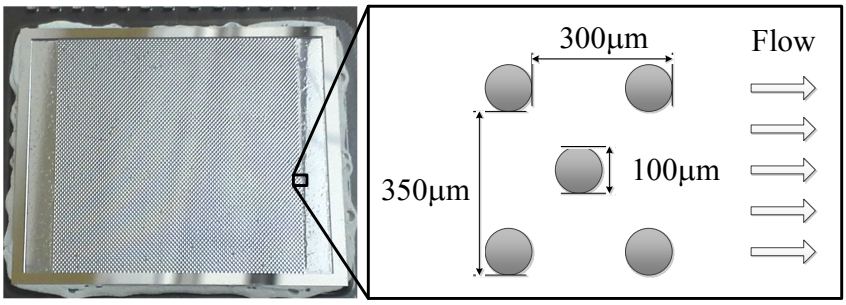


Figure 4.4: Image of the etched back side of the silicon FPGA die along with the micropin-fin dimensions

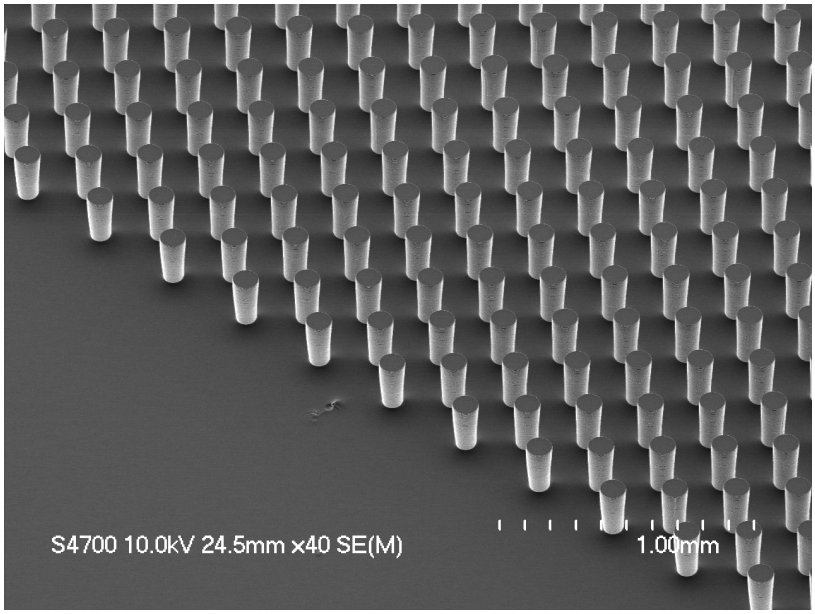


Figure 4.5: SEM image of micropin-fins etched using the same process in a silicon wafer

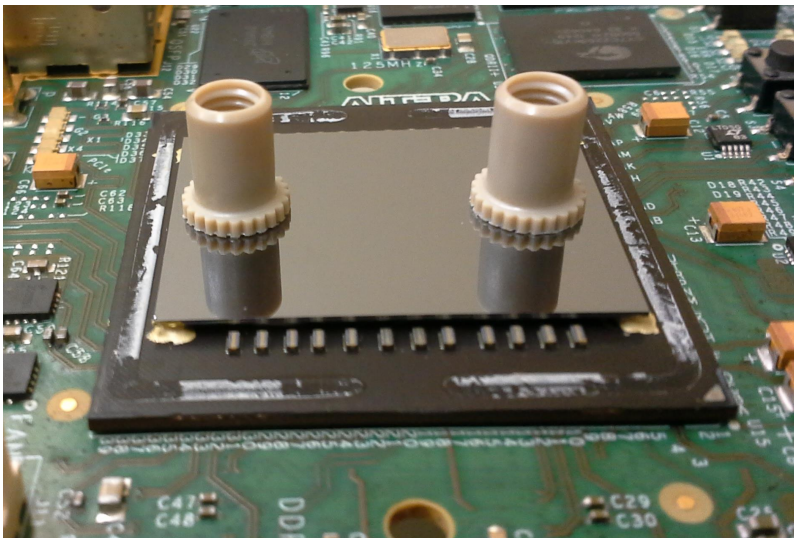


Figure 4.6: Processed FPGA soldered to development board with silicon cap and Nanoports for fluid delivery

to affect the profile and depth of etched cavities[56, 57]. Tapering of the micropin-fin sidewalls is visible on the micropin-fins closest to the inlet and outlet regions, but is minor in the rest of the array. Significant tapering, where the micropin-fin base is narrower than the top, could reduce fin efficiency and thermal performance by limiting the cross sectional area through which heat can conduct up the micropin-fins.

A separate silicon lid was fabricated with an inlet and an outlet port. The lid was first tacked on to the top of the etched FPGA with high temperature epoxy in order to provide a smooth surface for resoldering to the development board. After soldering, the lid was permanently secured with epoxy and Nanoports were attached to deliver coolant. A photograph of the resoldered FPGA with Nanoports can be seen in Fig. 4.6.

Testing

The FPGA was loaded with a custom pulse compression algorithm designed to mimic common DSP-style use cases of FPGAs and also to utilize a large amount of the FPGA resources. The algorithm consisted of 9 soft computing cores which could be toggled on and off during run time. The FPGA was tested in an open loop system, shown in Fig. 4.7,

with deionized water as a coolant. Testing was conducted with the Altera Stratix V DSP development board. The voltage regulator module on the board was run at a current higher than its datasheet rating, so air was blown over it in order to prevent overheating.

Flow rate was measured with a rotameter, which was calibrated by repeatedly filling a known volume of deionized water at the experiment temperature. Variation between these repeated measurements was found to be less than 0.03 mL/s for the flow rates used in this paper. The pressure gauge used to measure pressure drop across the heat sink was calibrated using an Omega DPI610 calibrator to within 0.1 kPa. K-type thermocouples were used to make fluid temperature measurements at the inlet and outlet of the micropin-fin heat sink. The relative uncertainty between temperature measurements was found to be 0.1 °C in the temperature ranges used.

Die temperature measurements were taken using the Altera Power Monitor tool, which retrieves measurements from an on-die temperature diode with a resolution of 1 °C. At 20 °C the temperature sensor on the FPGA die was found to have an offset from the thermocouples that was smaller than this resolution.

The FPGA was first tested with a flow rate of 2.4 mL/s, running zero to nine cores in order to vary the FPGA power. The inlet water temperature was 20.5 °C and the ambient air temperature was 19.3 °C. Temperature measurements can be seen in Table 4.1. A stock Stratix V DSP development board was also tested for comparison, using the stock air cooled heat sink with which it was bundled.

The pulse compression algorithm uses 80% of the logic, 93% of memory blocks, and 98% of the DSP blocks on the FPGA. In addition to many subtraction, addition, multiplexing, look-up-table, and memory operations, 346 18-bit multiplications are done every clock cycle. Counting only these multiplications, 934 GOPS are performed when operating all 9 cores at 300 MHz.

In order to capture the thermal gradient produced by heating of the fluid, measurements were taken with fluid flowing in both directions, as the temperature diode is located at

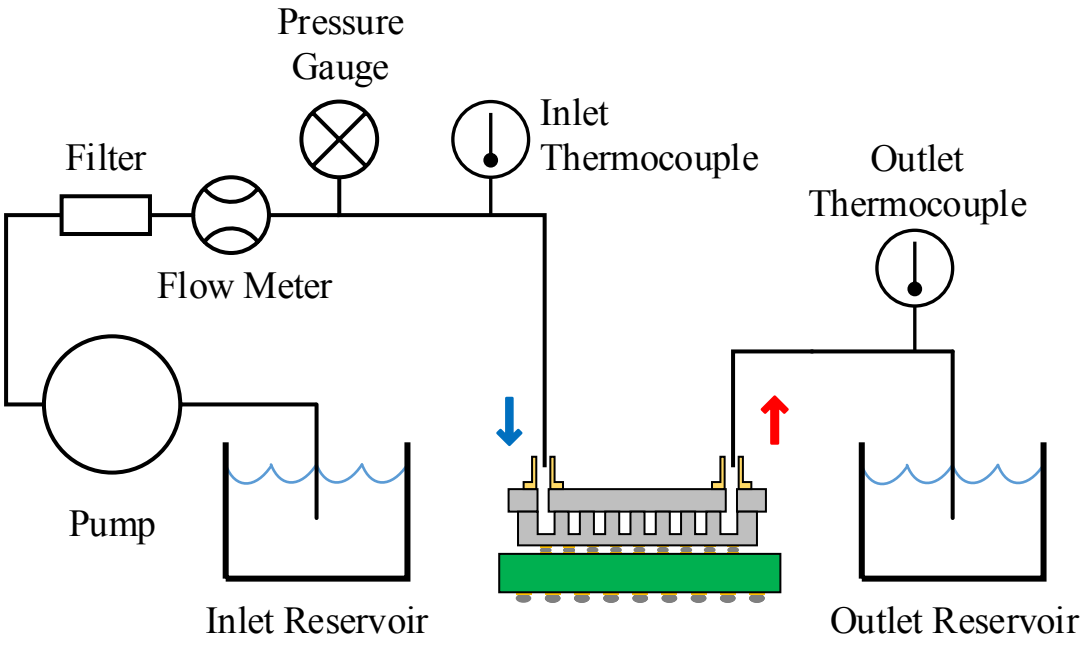


Figure 4.7: Diagram of the open loop system used to test the FPGA

Table 4.1: FPGA thermal and power measurements with microfluidic heat sink (MFHS) and air cooled heat sink (ACHS)

Cores	MFHS FPGA Power (W)	ACHS FPGA Power (W)	MFHS FPGA Temp (°C)	ACHS FPGA Temp (°C)
0	13.2	13.7	21–22	43
1	15.4	16.0	21–23	46
2	17.6	18.3	22–23	49
3	19.8	20.5	22–23	51
4	21.9	22.8	22–23	53
5	24.0	25.1	22–23	56
6	26.2	27.5	22–23	59
7	28.3	29.8	22–24	61[I]
8	30.4	—	22–24	—
9	32.4	—	22–24	—

[I]Temperature warning on board illuminated.

the edge of the chip. Therefore, the temperatures of the microfluidic cooled FPGA are all reported as a range of two values representing flow in both directions.

A maximum die temperature of 60 °C was set to match the default on-board temperature warning indicator of the Stratix V DSP development kit. Although better thermal results could undoubtedly be achieved with a larger, more powerful air cooled heat sink, it should be noted that the stock air cooled heat sink ran six computing cores before reaching this maximum temperature, while the microfluidic cooled FPGA ran all nine cores (a 1.5x improvement in throughput) while maintaining a die temperature below 24 °C (with additional power as per Table 4.1). Although the air cooled solution results in a higher junction temperature, it meets market requirements for Stratix V target applications.

The FPGA heat flux is lower than many high power processors and the low profile air cooled heat sink with which it came is significantly less effective than the best available air cooled heat sinks. Since the temperature has a linear relationship with thermal resistance and power ($T_j = T_{in} + R_{th}P$), the temperature can be predicted for higher power and higher performance air cooling, assuming a constant thermal resistance. This is demonstrated in Fig. 4.8, where the average temperatures and powers from Table 4.1 are plotted. Lines are fit to the temperature versus power data points of the liquid cooled FPGA at 2.4 mL/s and the stock air cooled FPGA. An additional line shows the projected temperature with a powerful hypothetical air cooled heat sink with a junction-to-ambient thermal resistance of 0.25 °C/W.

As can be seen, at an FPGA power of 160 W, the microfluidic cooled FPGA in this work would have a die temperature of 31.5 °C, while the die cooled with the hypothetical high performance air cooled heat sink would have a temperature of 61.3 °C. At 300 W, these temperatures would be 40.6 °C and 96.3 °C, respectively.

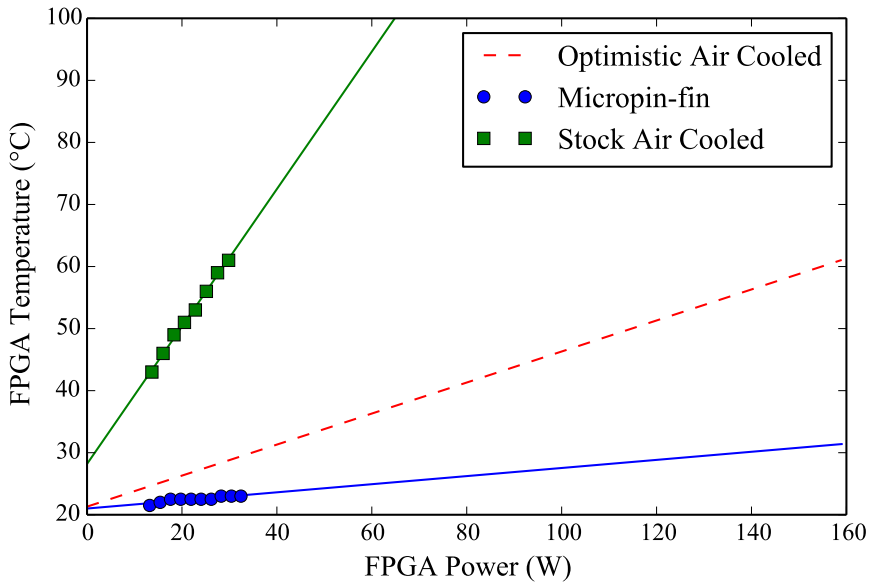


Figure 4.8: Die temperature vs. die power

Variable Flow Rate Testing

As flow rate through a micropin-fin heat sink increases, the convective thermal resistance decreases. This relationship between Nusselt number, which is proportional to heat transfer coefficient, and Reynolds number, which is proportional to flow rate, has been measured for a variety of micropin-fin geometries [19, 22, 23, 51].

The microfluidic cooled FPGA was tested with several different flow rates, running the same pulse compression algorithm with all nine cores and an inlet temperature of 20.3 °C to 20.9 °C. The results can be seen in Fig. 4.9. As flow rate increases, the FPGA temperature decreases due to decreasing heating of the fluid as well as decreasing convective thermal resistance. At a maximum flow rate of 3.0 mL/s, a minimum average thermal resistance of 0.07 °C/W was achieved.

As flow rate increases, the temperature gradient from inlet to outlet due to heating of the fluid also decreases. For a given power, the temperature gradient across the chip is approximately equal to the temperature rise of the fluid, which is related to flow rate as $\Delta T \propto 1/\dot{m}$, where ΔT is the temperature rise of the fluid and $\dot{m}$ is the mass flow rate.

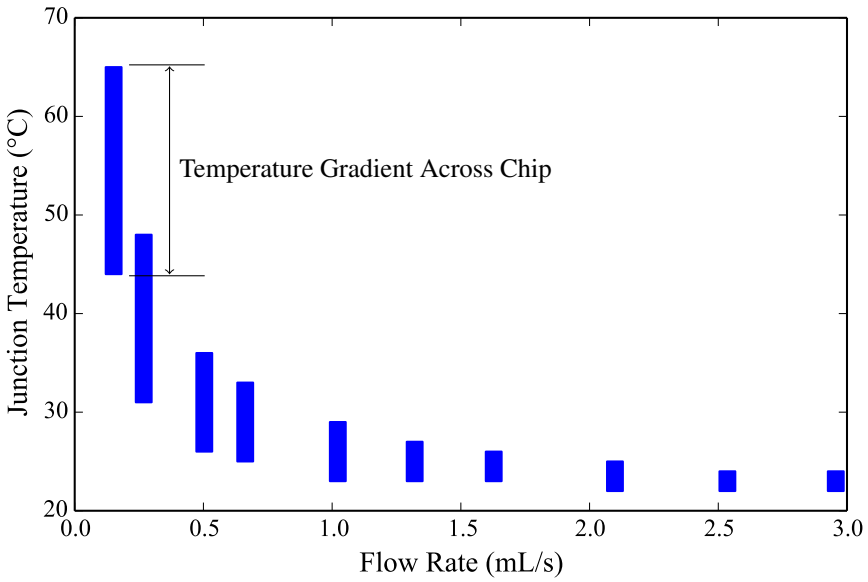


Figure 4.9: Die temperature vs. flow rate. Die temperature is a range representing the measurements with the temperature diode near the inlet and near the outlet.

The measured outlet water temperature as well as the predicted outlet water temperature are plotted in Fig. 4.10. The difference in the measured and calculated outlet water temperatures is due to heat loss through alternate heat paths to ambient air, such as the board and tubes.

Heat loss to the surrounding ambient air was quantified as

$$Q_{loss} = Q_{in} - \dot{m} C_p \left(T_{out} - T_{in} \right) \tag{4.1}$$

where Q_{in} is the measured power of the FPGA chip, $\dot{m}$ is the water mass flow rate, C_p is the specific heat of water, and T_{out} and T_{in} are the measured outlet and inlet water temperatures, respectively. C_p is a relatively weak function of temperature and was taken to be 4.18 J/(°C g). The density of the water was taken to be 1 g/mL for the purposes of converting measured volumetric flow rate to mass flow rate. The heat loss is plotted versus average die temperature in Fig. 4.11. Data points were used from this variable flow rate experiment as well as the elevated inlet temperature experiment presented in the next subsection.

Heat produced on the FPGA die has many thermal paths: through the microfluidic heat

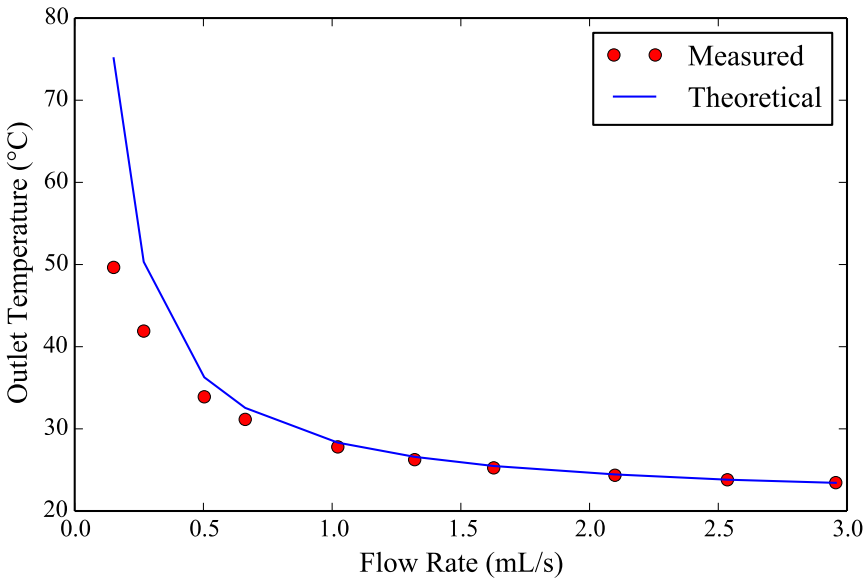


Figure 4.10: Measured outlet temperature and theoretical outlet temperature (no heat loss) vs. flow rate.

sink and into the liquid, or through the package, board, etc. to the surrounding ambient air. When the microfluidic heat sink is operated with a high flow rate and the coolant is near the temperature of ambient air, as is the case in the left side of Fig. 4.11, the majority of the die heat is captured in the fluid. If the efficacy of the heat sink is limited through a restricted flow rate, or an elevated inlet temperature, the die temperature rises relative to the surrounding ambient air and more heat is lost through these alternate heat paths to the ambient air. A higher ambient temperature, reduced airflow around the board, and insulation would all increase the fraction of heat captured by the coolant (and increase die temperature).

After fitting a line to the points in Fig. 4.11, the slope can be used to calculate the thermal resistance from the FPGA die to the ambient air through the board, tubes, etc. It was calculated to be 1.8 °C/W for this test setup.

Pressure drop was also measured as a function of flow rate while running all nine computing cores, which can be seen in Fig. 4.12. These pressure measurements were made outside of the chip and therefore include pressure drop across the inlet and outlet ports. A

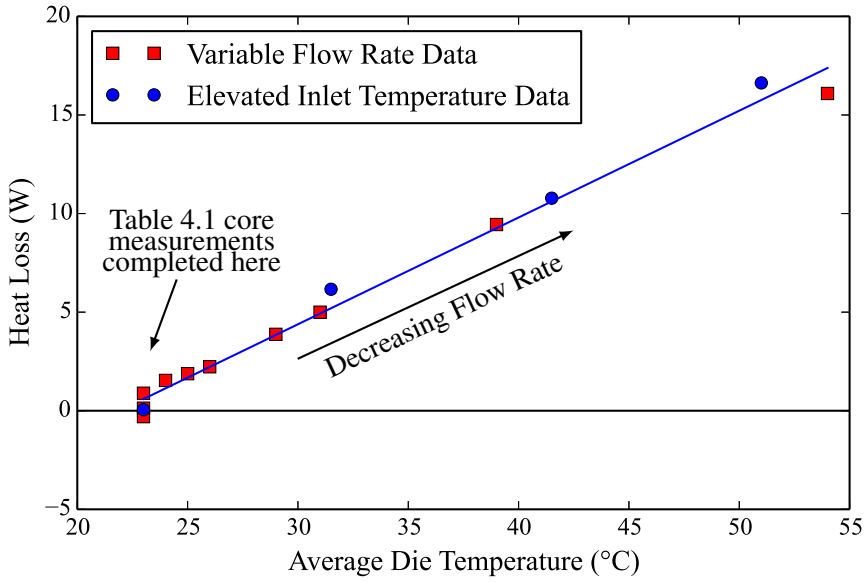


Figure 4.11: Heat Loss vs. average die temperature.

maximum pressure drop of 100 kPa was set as a conservative limit to prevent fluid leakage. A much higher pressure drop could be sustained with improved cap bonding [46].

Elevated Inlet Temperature Testing

The FPGA was also tested with an elevated inlet temperature, varying from 21 °C to 50 °C at a flow rate of 3.1 mL/s. These temperatures can be seen in Fig. 4.13. As expected, the FPGA die temperature tracks the water inlet temperature very closely, with an average junction temperature rise above inlet of 2.1 °C and 0.8 °C at average inlet temperatures of 20.9 °C and 50.1 °C, respectively. Temperature rise above inlet (and hence apparent junction-to-inlet thermal resistance) decreases with increasing temperature due to increased heat loss to the surrounding ambient air, which was 19.7 °C.

Pressure drop is also plotted as a function of inlet temperature in Fig. 4.14. As water temperature increases, its viscosity decreases, leading to the downward trend in pressure drop versus water temperature shown in Fig. 4.14.

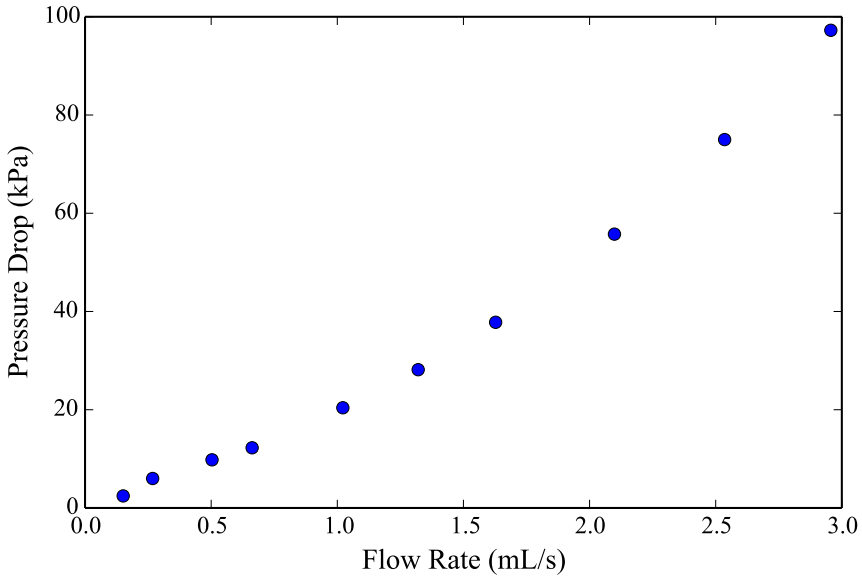


Figure 4.12: Pressure drop vs. flow rate.

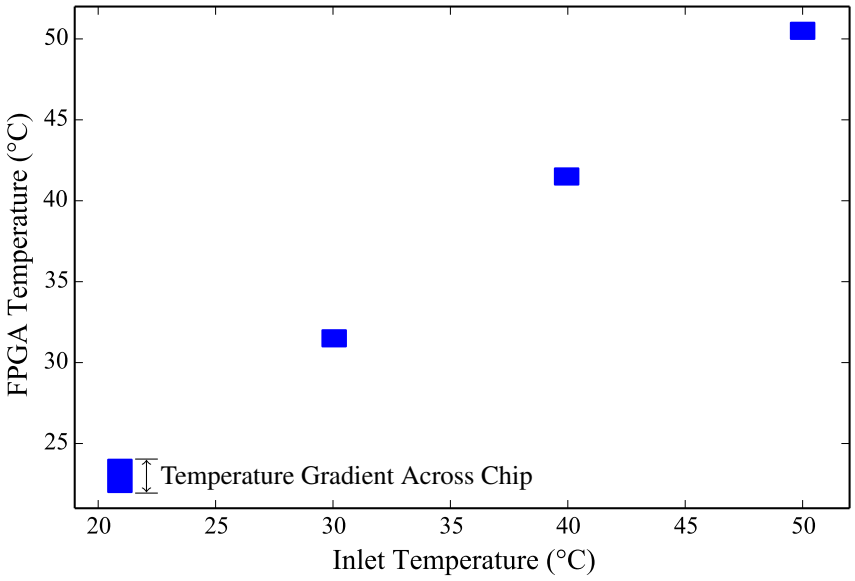


Figure 4.13: FPGA temperature range vs. inlet temperature.

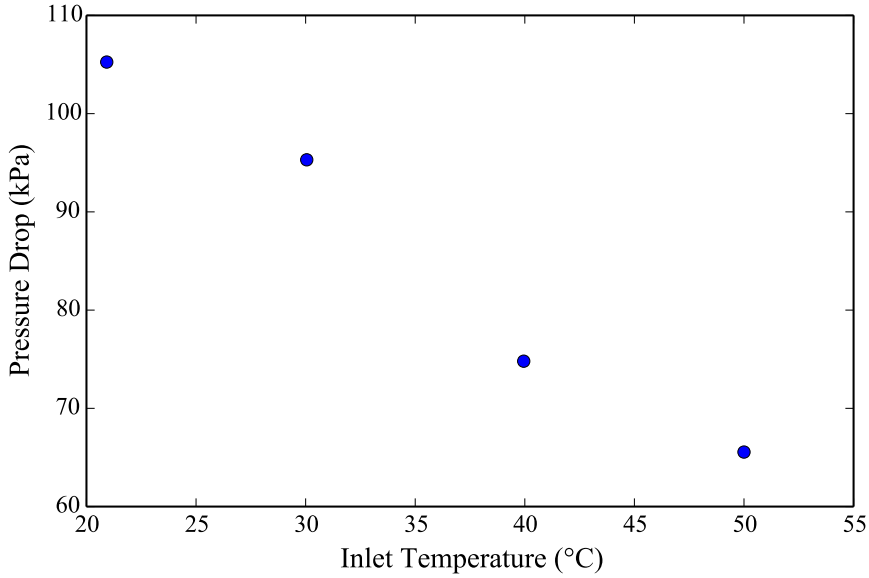


Figure 4.14: Pressure drop vs. inlet temperature.

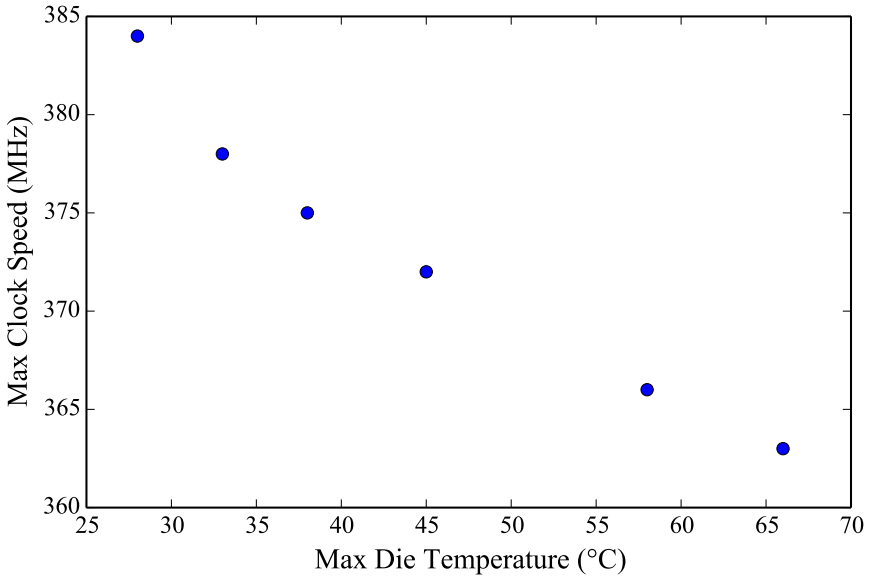


Figure 4.15: Maximum FPGA clock speed without glitches vs. maximum die temperature.

Clock Speed

In addition to increasing performance through increased silicon utilization, improved cooling can also benefit performance in terms of clock speed. Both the transistors and interconnects experience enhanced performance with decreased temperature in planar bulk CMOS. Transistor threshold voltage and mobility tend to decrease as temperature increases [58]. Although decreased threshold voltage partially counteracts decreased mobility, a net decrease of drive current is observed by Lin et al. in simulations of a 45 nm technology. Lin et al. observed a 9% increase in delay time between simulations of a nine-stage inverter chain at 25 °C and 125 °C. The dependence of total critical path delay can, however, vary widely depending on chip design and process technology.

In order to test the dependence of maximum clock frequency on die temperature with the microfluidic cooled FPGA, temperature was varied by varying flow rate with an inlet temperature of 24.3 °C to 24.7 °C. Due to current limitations from the on-board voltage regulator module (VRM), seven of the nine cores were run. The output from all cores were monitored through the Altera Signaltap tool to detect glitches which occurred in the output waveforms. The maximum clock speeds at which all seven cores operated with no glitches can be seen in Fig. 4.15 as a function of the die temperature measured on the side of the chip closest to the outlet. Decreasing the maximum die temperature from to 66 °C to 28 °C yielded an improvement of 21 MHz, a 6% improvement in clock speed (with an accompanying increase in power).

Die Power

Chip power consists of dynamic power, which has little dependence on temperature, and static/leakage power, which comes from several components, such as subthreshold leakage, gate leakage, and reverse bias junction current. Subthreshold leakage current tends to be the most significant temperature dependent component of the power [59, 60] and is given

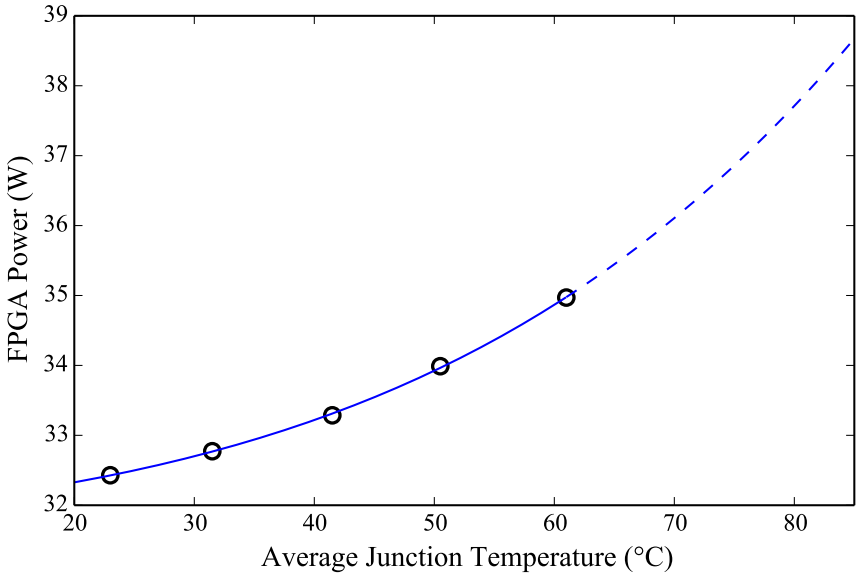


Figure 4.16: FPGA power vs. average die temperature.

by [61]

$$I_{ds} = \mu_0 C_{ox} \frac{W}{L} (m-1)\, v_T^2 \mathrm{e}^{(V_{gs}-V_{th})/mv_T} \times \left(1 - \mathrm{e}^{-V_{ds}/v_T}\right) \tag{4.2}$$

where μ_0 is the zero bias mobility, C_{ox} is the gate oxide capacitance, W/L is the channel width to length ratio, m is the subthreshold swing coefficient, V_{th} is the threshold voltage, and v_T is the thermal voltage, given by $v_T = k_b T/q$.

The total FPGA power versus average die temperature is plotted in Fig. 4.16. Die temperature was varied by varying inlet temperature at a constant flow rate of 3.1 mL/s since this provided nearly uniform die temperatures (Fig. 4.13). Due to the temperature-dependent leakage power, the measured total FPGA power increases by 2.6%, 4.8%, and 7.8% at 41.5 °C, 50.5 °C, and 61 °C relative to power dissipation at 23 °C. A trend curve using a first order approximation of equation 4.2 is also shown in Fig. 4.16. From an efficiency standpoint the measured increase in FPGA power at elevated temperatures provides another strong motivation for effective cooling.

Conclusion

In this work, a micropin-fin heat sink was etched into the back side of an Altera Stratix V FPGA. The FPGA was tested with a pulse compression algorithm to demonstrate functionality and perform thermal benchmarks. Die temperature and power were measured as a function of flow rate and inlet temperature. An average junction-to-inlet thermal resistance of 0.07 °C/W was achieved at a flow rate of 3.0 mL/s and pressure drop of 97 kPa. This thermal resistance is sufficiently low to cool future generations of FPGAs and other high heat flux processors. The FPGA was also cooled with inlet water temperatures up to 50 °C, enabling high efficiency through heat exchange directly to ambient air, or waste heat reuse.

Future work may focus on both enhancement of the microfluidic heat sink as well as the benchmarking algorithm loaded on to the FPGA. Pressure drop and thermal resistance could be improved through optimization of the micropin-fin heat sink and ports. The FPGA offers an opportunity to benchmark the performance gained through increased clock speed and chip utilization with many algorithms and architectures.

CHAPTER 5
MICROFLUIDIC COOLING OF A 2.5D FPGA

As mentioned in the introduction, interconnection has become a bottleneck in computing performance and efficiency at every level of integration. While the obvious solution to this problem is decreasing interconnect length and increasing system density, the ability to remove heat already limits this approach. Computational density is primarily limited by the large volume of air that must be used to capture heat with a reasonable increase in temperature.

By switching to a liquid coolant, such as water, which has orders of magnitude higher volumetric heat capacity than air, the necessary volume for heat exchange and fluid delivery can be dramatically decreased. In addition to enabling higher density at the die-level in this fashion, microfluidic cooling can be applied at the package level to enable high density stacking of PCBs.

For example, high power computing accelerators are often integrated onto boards and connected to a main PCB through peripheral component interconnect express (PCIe) slots. The air cooled heat sink on these boards is the largest component on the board and limits the pitch at which the boards can be mounted, and therefore the number of accelerator boards that can be integrated into a single computer/server.

In addition to addressing the challenge of decreasing heat sink volume, these heat sinks must be able to address the needs of modern high power packages. Most importantly, these high power accelerator packages no longer include a single monolithic die, but several dice mounted in close proximity to one another, usually with an interposer or embedded bridge chips for high bandwidth interconnection. These dice implement heterogeneous functionalities and have heterogeneous thermal requirements. Therefore, there is a need for low-profile, low thermal resistance heat sinks which can simultaneously address the

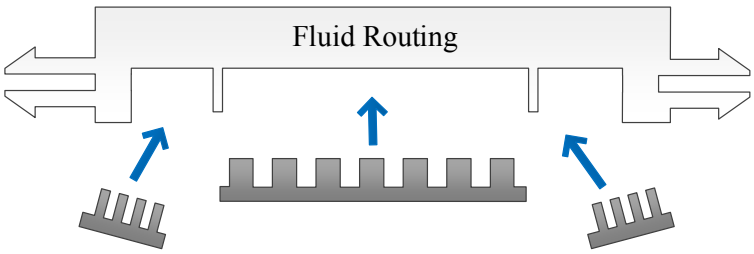
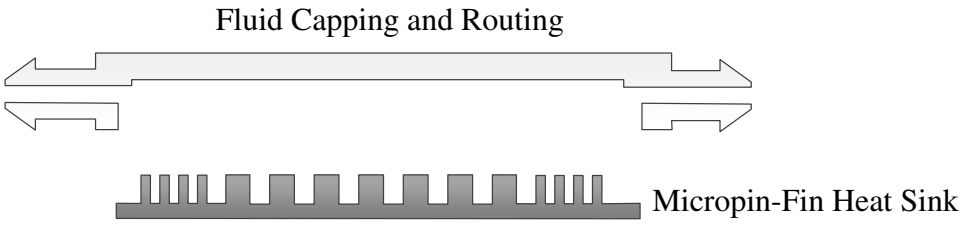


Figure 5.1: Chiplet microfluidic cooling

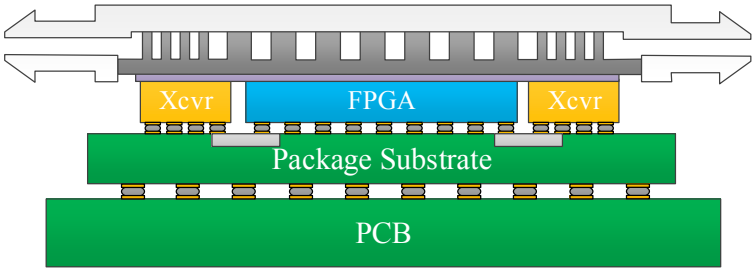
thermal needs of heterogeneous dice mounted in close proximity to one another within a single package.

As microelectronics transition from monolithic dice to large packages of heterogeneous chiplets, heat sinks can also be modified to match these heterogeneous designs. A low cost polymer manifold can be used to assemble microfluidic cooling chiplets to construct a large heterogeneous heat sink that can be attached to a 2.5D heterogeneous collection of chiplets, as shown in Figure 5.1. Alternatively, functional chiplets could be assembled in a plastic (or similar) manifold before flip chip bonding the package to the assembled dice. Bridge chips or interposers could then be flip chip bonded to the chiplets, followed by the package substrate. Chiplets could either be etched, as shown in Figure 5.1, or jet impingement cooling could be used to obtain a substantial heat transfer coefficient across the backside of unetched dice.

In this work, a single silicon micropin-fin heat sink is designed and fabricated to cool a Stratix 10 GX FPGA consisting of five heterogeneous dice. Following previous work which successfully cooled local hotspots by locally increasing micropin-fin geometry, the micropin-fin geometry is locally varied to match the heat flux of the underlying dice. The silicon micropin-fin heat sink is embedded in a 3D printed plastic piece which seals the top of the silicon micropin-fins and connects the heat sink to inlet and outlet tubing. A conceptual cross-sectional diagram of the heat sink concept can be seen in Figure 5.2.



(a) Two parts of the microfluidic heat sink



(b) Microfluidic heat sink assembled on dice

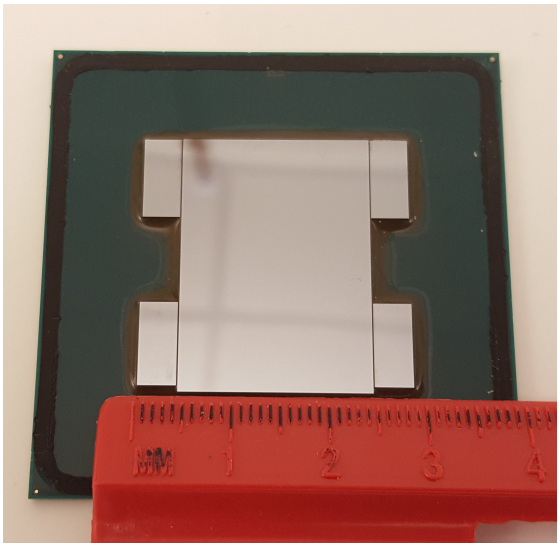
Figure 5.2: Cross-sectional diagram of micropin-fin heat sink for 2.5D package

Experimental Setup

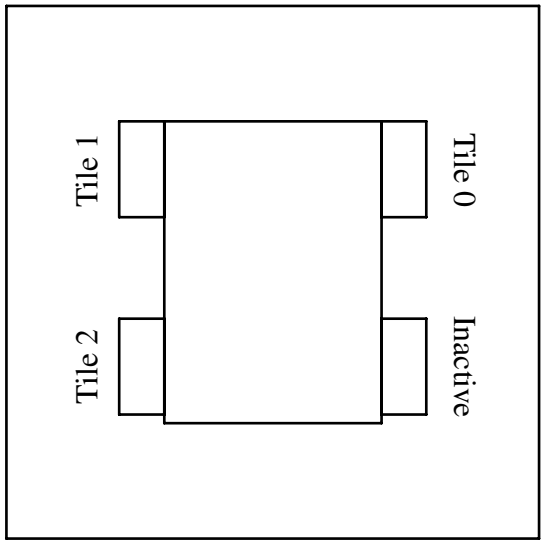
The experiments in this work were carried out using a Stratix 10 engineering silicon (ES) development kit from Intel. A photograph of the board can be seen in Figure 5.3. The board carries a Stratix 10 GX FPGA which is a 2.5D device consisting of a 14 nm FPGA core die surrounded by four transceiver dice, connected through Intel embedded multi-die interconnect bridge (EMIB). A photo of a delidded package can be seen in Figure 5.4a.

Each transceiver tile (die) contains 24 transmitters and receivers, for a total of 96 in the package. Each receiver requires a dedicated reference clock connected directly through the package to the transceiver tile on which it resides. The Stratix 10 ES development board only has reference clocks connected to three of the four transceiver tiles, so only three of the four transceiver tiles are used in this work.

Figure 5.3: Stratix 10 ES development board



(a) Delidded Stratix 10 FPGA



(b) Stratix 10 FPGA die layout

Figure 5.4: Stratix 10 GX package

FPGA Benchmark Application

The goal of the benchmark program used in this work was to mimic a high power use case of the FPGA. The benchmark application consists of a portion which runs on the FPGA,

and a portion which exercises the transceivers. The core of the design used consists of a streaming fast fourier transform (FFT) block followed by six first-in, first-out (FIFO) buffers operating on random inputs hard coded into the FPGA. This design is implemented on the FPGA through a combination of programmable logic blocks and DSP blocks. Much of the computational throughput as well as power dissipation comes from the FPGA's DSP blocks, which perform arithmetic operations on floating point operands. The FFT core was replicated 160 times across the FPGA and clocked at 475 MHz. The clock (and power) could be raised higher, but the VRMs on the board are only designed to supply a maximum of 100 A of current on the `VCC` rail and the board begins to shut off at clock frequencies much higher than this. Even at 475 MHz, the FPGA used well over 100 A on the `VCC` rail and air was therefore blown over the VRMs to prevent them from overheating during experiments.

In addition to this FFT design, 72 transceiver channels were utilized to dissipate power on three of the four transceiver tiles. Each of the 72 transceiver channels were programmed to run in Enhanced physical coding sublayer (PCS) mode with serial loopback enabled. This design was modified from a publicly available design from the Intel FPGA Wiki. The maximum data rate within the Intel transceiver intellectual property (IP) for the GX series of Stratix 10 FPGAs is 16 Gbit/s per channel, but the transceiver clocks were increased during runtime to overclock the transceivers to 22 Gbit/s. The future Stratix 10 TX FPGA will be available with up to 56 Gbit/s bandwidth per channel and will likely dissipate even more power.

One temperature sensor on each die is read using a Nios II soft processor on the FPGA which feeds these numbers back to an attached computer for logging. Voltage and current on the power rails of the board were measured through on-board sensors which interface to a board test system (BTS) interface that comes with the development board.

The board has six power rails which can be monitored in the BTS interface. For power estimates, it is assumed that two of these rails represent power dissipated on the FPGA

tile, while the sum of the remaining rails is assumed to be the power dissipated equally across the transceiver tiles. More granular, die specific power measurement is not possible because the individual dice within the package are not connected to independent power rails. Therefore, it is further assumed for the remainder of this work that the transceiver tiles dissipate identical amounts of power when running identical transceiver configurations. In actuality, there will be manufacturing variations across dice, but this assumption will allow us to estimate the power dissipation on each die for the purposes of heat sink design.

Air Cooled Heat Sink

A Cooler Master MA621P TR4 air cooled heat sink was used as a baseline for comparison. The heat sink is designed for a large AMD Threadripper CPU package and covers the entire 5 cm $\times$ 5 cm FPGA package. The heat sink was mounted using custom designed, 3D printed mounting brackets. MasterGel Pro TIM, which has a thermal conductivity of 8 W/(m K), was applied between the base of the heat sink and the FPGA heat spreader. The assembled heat sink and FPGA development board can be seen in Figure 5.5. The 3D printed mounting brackets and springs can be seen in Figure 5.6.

Baseline power and temperature figures were first measured on the development board with the air cooled heat sink. First, power was measured before loading the benchmark design, with the clocks disabled to estimate leakage power and, in particular, the power dissipated on the unused transceiver die. Temperature measurements were made near the inlet of the air cooled heat sink using two k-type thermocouples, as well as on the FPGA die using the external temperature measurement on the board, read through BTS.

Power on the FPGA die was 6.1 W while power across the transceiver tiles was 2.8 W. The board was enclosed and warmed with a hotplate to increase the FPGA temperature an additional 5 °C, as measured by BTS, to better mimic the temperatures of the transceivers expected with a full design and microfluidic cooling. This increased the powers to 6.8 W and 2.9 W, respectively.

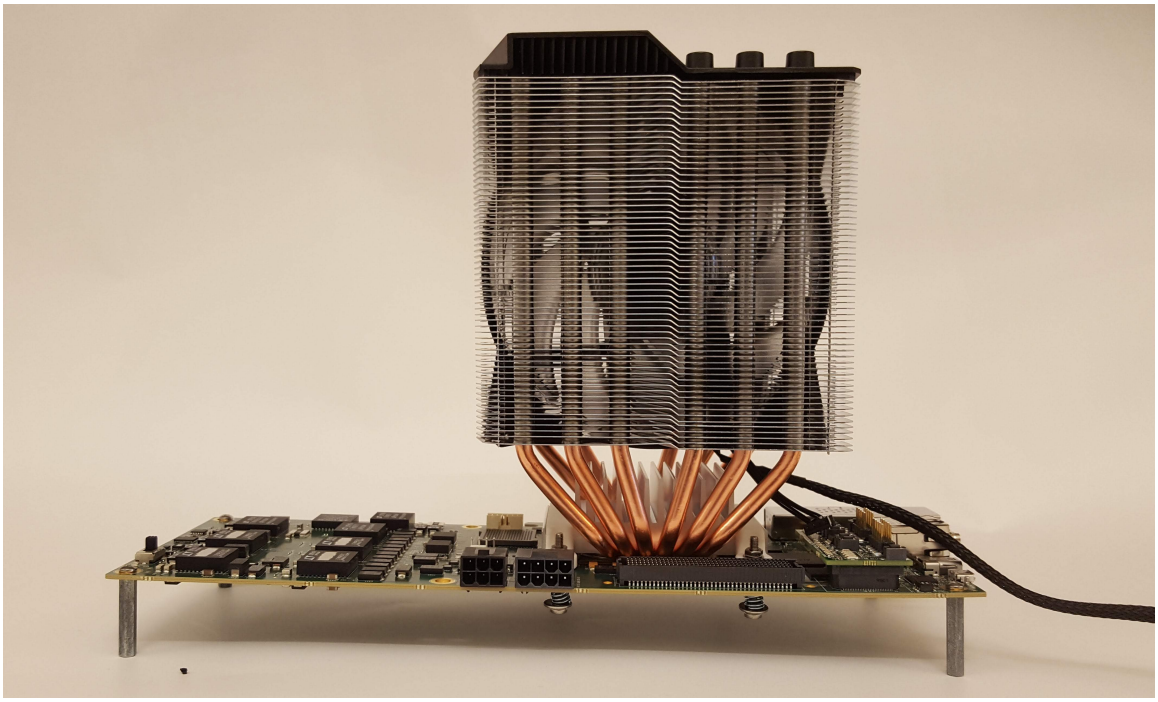
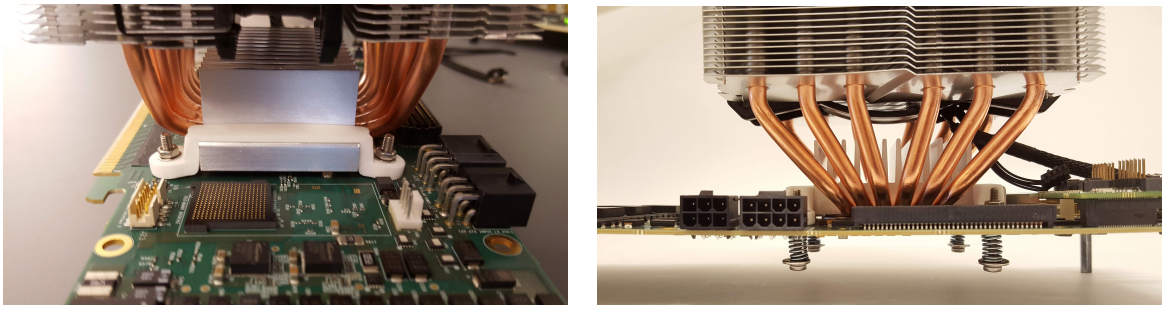


Figure 5.5: Stratix 10 ES development board with air cooled heat sink



(a) 3D printed mount (b) Screws and springs applying pressure

Figure 5.6: Custom mounting for air cooled heat sink on Stratix 10 development board

Next, the benchmark application was loaded onto the FPGA and power and temperature were measured across the FPGA and three active transceiver dice. The total power on the FPGA die was 110.4 W while the total power across transceivers was 64.2 W. These initial measurements were made without waiting the full stabilization period used for the air cooled measurements reported later. As will be shown, this initial power estimate, which was used to design the microfluidic heat sink, was within 0.2 W of the actual power measured with the microfluidic cooled heat sink at the highest flow rate.

Table 5.1: Stratix 10 die powers, areas, and power densities

Die	Power	Area	Power Density
FPGA	110 W	5.8 cm^2	19 W/cm^2
Transceiver Tile 0	21 W	0.38 cm^2	56 W/cm^2
Transceiver Tile 1	21 W	0.38 cm^2	56 W/cm^2
Transceiver Tile 2	21 W	0.38 cm^2	56 W/cm^2
Inactive Tile 3	0.72 W	0.38 cm^2	1.9 W/cm^2

Heat Sink Design

The aforementioned assumption that power is equally distributed across the transceiver dice which run identical applications was used to estimate the power on each transceiver die for the purpose of designing the microfluidic heat sink, which targets individual dice. It was assumed that the leakage power determined in the previous section was equally split across all four dice, while the dynamic power was equally split across the three active transceiver dice. It was also assumed that the leakage power, measured without any design running, was equal to the leakage power when running the benchmark application. The power breakdown for each of the five physical dice used for simulations can be seen in Table 5.1.

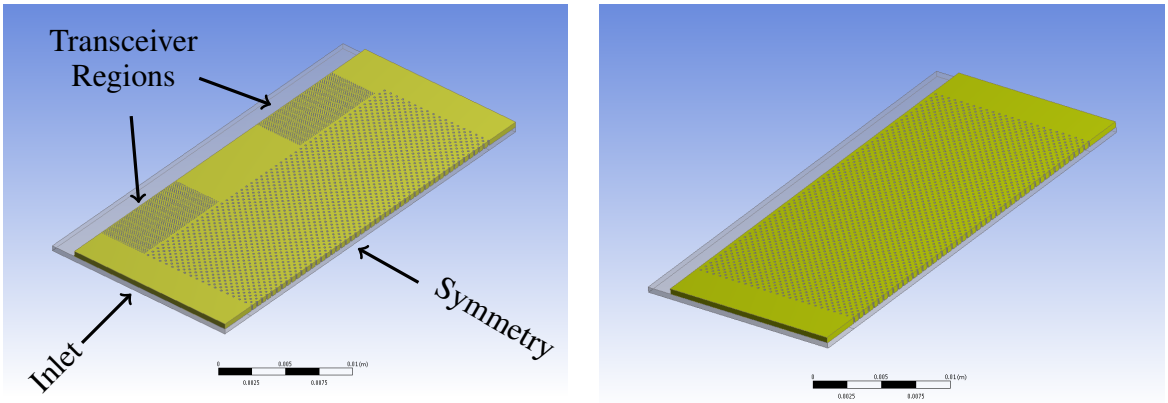
An initial heterogeneous micropin-fin design was created by beginning with a high aspect ratio micropin-fin geometry which could be etched to the desired depth while maintaining near-vertical sidewalls. This high aspect ratio micropin-fin design was used to populate the regions of the heat sink directly over the transceiver dice, while a second micropin-fin design was used to cool the region above the FPGA die. The pitch and diameter were scaled by the relative power density of the transceiver and FPGA dice to create the dimensions used to cool the center FPGA die.

Heat Sink Simulation

Ansys FLUENT was used to simulate a silicon microfluidic heat sink with an applied power map matching the values shown in Table 5.1, applied to the underside of the silicon micropin-fin heat sink. The micropin-fins are 460 µm tall, while the base silicon underneath the micropin-fins is 440 µm thick.

A top view of the simulated 3D model can be seen in Figure 5.7a, with the fluid region highlighted in yellow. A symmetry plane down the center of the chip was used to reduce the model size. This means that the fourth transceiver tile is modeled as active and identical to the other transceivers, although the experimental results do not include this transceiver tile. The model assumes a uniform inlet velocity and pressure outlet boundary condition, which do not include fluid delivery effects of the plastic enclosure which must be used in experimentation.

Simulation results were compared with simulations of a uniform micropin-fin geometry, where the micropin-fin geometry used over the FPGA was extended across the entire heatsink. This model can be seen in Figure 5.7b.



(a) Heterogeneous micropin-fin design (b) Uniform micropin-fin heat sink

Figure 5.7: Simulated micropin-fin heat sink models

Temperature dependent material properties were used within FLUENT for water viscosity and thermal conductivity as well as the thermal conductivity of silicon. The specific

heat of water was taken to be a constant 4182 J/(kg K).

Specific heat and viscosity values for water were taken from the NIST WebBook[62] and fit to polynomials over the temperature range 273.16 K to 370 K. Second and fourth order polynomial functions were fit to the thermal conductivity and viscosity values, respectively. The following polynomial function was generated for the thermal conductivity of water

$$k_w = -0.7491 + (7.430 \times 10^{-3})T - (9.658 \times 10^{-6})T^2 \tag{5.1}$$

where k_w is the thermal conductivity of water in W/(m K) and T is the temperature in K. The dynamic viscosity of the water is given by

$$\mu_w = 0.4635 - (5.384 \times 10^{-3})T + (2.357 \times 10^{-5})T^2 - (4.602 \times 10^{-8})T^3 + (3.376 \times 10^{-11})T^4 \tag{5.2}$$

where μ_w is the dynamic viscosity of water in units of Pa s. The following correlation from [63] was used for the thermal conductivity of silicon

$$k_{Si} = 2122.1 - 16.765T + (4.818 \times 10^{-2})T^2 - (4.744 \times 10^{-5})T^3 \tag{5.3}$$

where k_{Si} is the thermal conductivity of silicon.

Simulations were performed in ANSYS FLUENT for both geometries with inlet flow velocities of 0.1 m/s to 0.9 m/s, corresponding to flow rates of 1.3 mL/s to 11.9 mL/s. Average temperatures across the FPGA and transceiver heat flux regions were extracted and can be seen in Figure 5.8.

At low velocities, the temperatures between the uniform and non-uniform heat sinks are similar. However, as the flow rate increases, the temperature of all three dice cooled with the heterogeneous micropin-fin heat sink, and the transceiver dice in particular, drop below the temperatures of the dice cooled with the uniform micropin-fin heat sink. Therefore, the absolute temperatures, as well as the temperature differences between dice are lower with

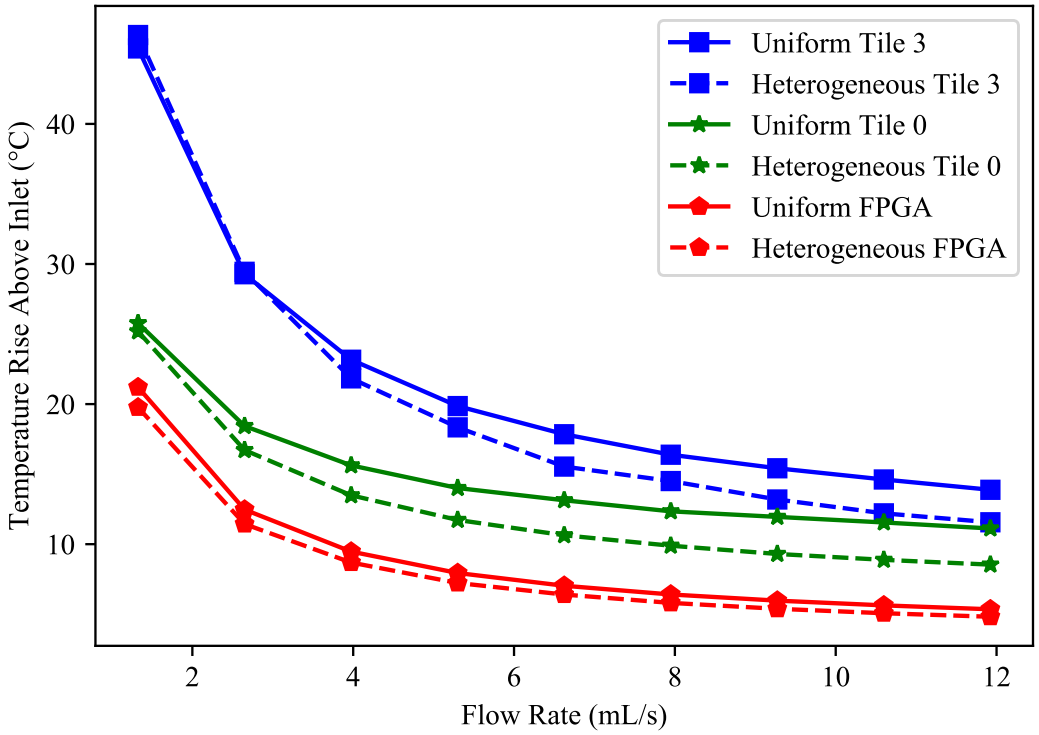


Figure 5.8: Simulated temperatures of FPGA and transceiver regions with heterogeneous and uniform heat sinks

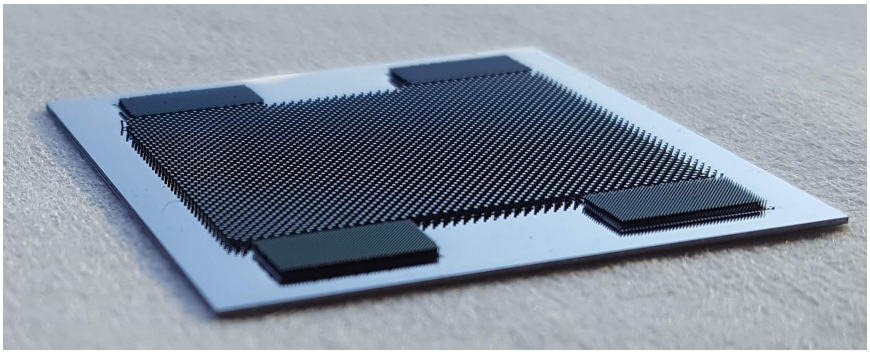


Figure 5.9: Silicon micropin-fin heat sink for Stratix 10 FPGA

the non-uniform heat sink than with the uniform heat sink.

Heat Sink Fabrication and Assembly

The complete microfluidic heat sink assembly consists of two parts: the etched silicon heat sink, through which heat is transferred to the fluid, and a 3D printed plastic piece which encases the silicon insert and routes fluid between the inlet/outlet tubes and the silicon micropin-fins.

A photograph of the silicon micropin-fin heat sink can be seen in Figure 5.9. The heat sink was fabricated through Bosch process etching of a 900 µm thick silicon wafer to a depth of approximately 460 µm.

Since both sets of micropin-fin dimensions are etched to approximately the same depth, the micropin-fins over the transceivers are much higher aspect ratio than those over the FPGA. Therefore, the micropin-fin dimensions used for the transceivers were chosen first with a pitch-to-diameter ratio which yielded good results when etched to approximately 460 µm. The micropin-fin dimensions over the FPGA region were then chosen by scaling both the pitch and diameter by the respective heat flux densities of the two regions. An SEM image of both micropin-fin regions can be seen in 5.10. The profile of the etched sidewalls depends on surrounding features. Therefore, the micropin-fins nearest to the edges of the arrays have significant taper and many of the high aspect ratio micropin-fins in the outermost two rows broke off during fabrication. As can be seen in Figure 5.10, the

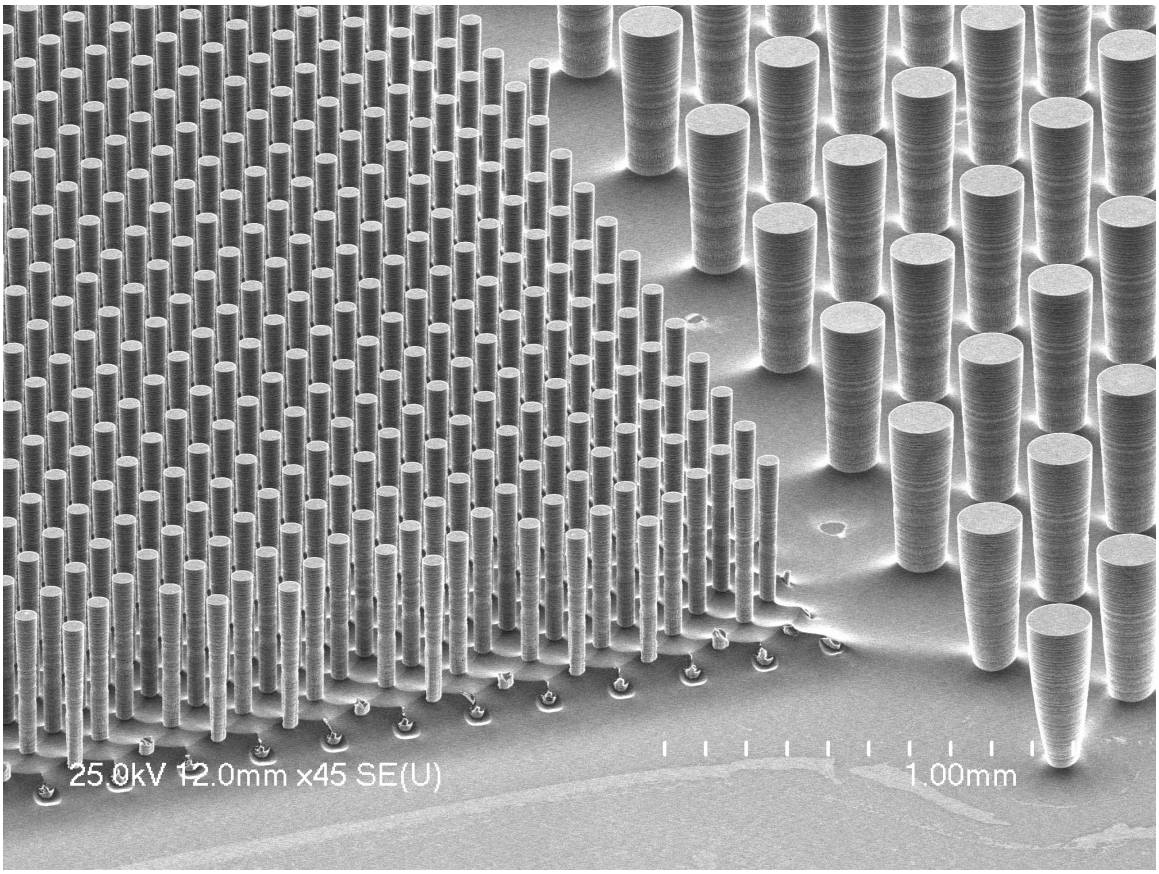


Figure 5.10: SEM image of high and low density micropin-fins

majority of the micropin-fins remain and have minimal taper.

The height of the micropin-fins were measured using an Olympus LEXT optical profilometer. Since etch depth can locally vary based on local features, measurement windows were stitched together to collect height data across approximately one quarter of the chip, capturing one full transceiver micropin-fin region and approximately one quarter of the FPGA micropin-fin region. Since the entire micropin-fin heat sink is nearly symmetrical across two planes, this quarter of the chip is assumed to be representative of the entire chip. 13.3 million height measurements across this region are aggregated into a histogram in Figure 5.11 with a bin width of 0.1 µm. From this histogram it can be estimated that the base silicon has a mode height of 400 µm and the micropin-fins have a mode height of 457 µm.

It can be seen in Figure 5.10 that the regions between the micropin-fins are also not completely flat. Therefore, a profile of the region between the high density micropin-fins

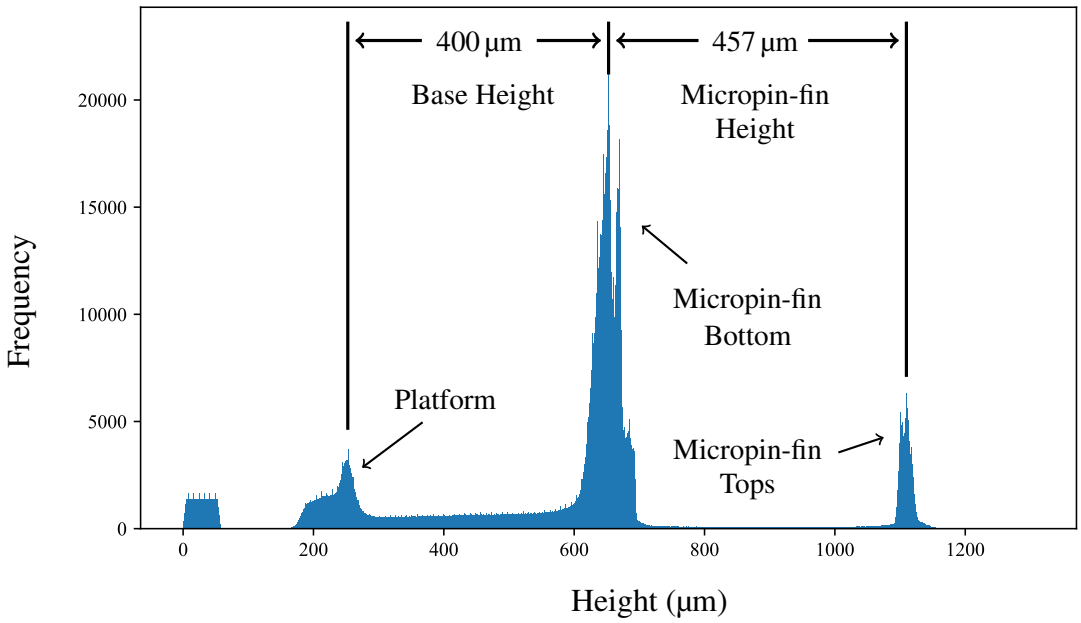
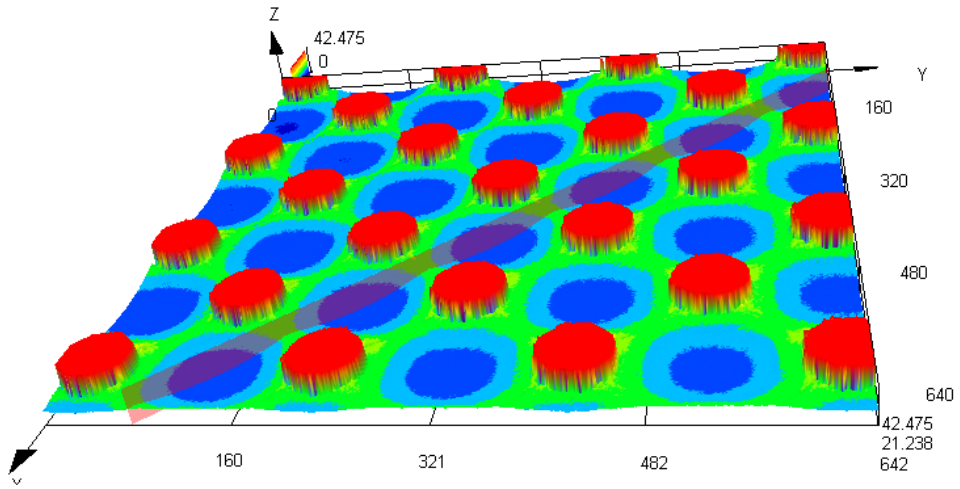


Figure 5.11: Histogram of micropin-fin heat sink heights

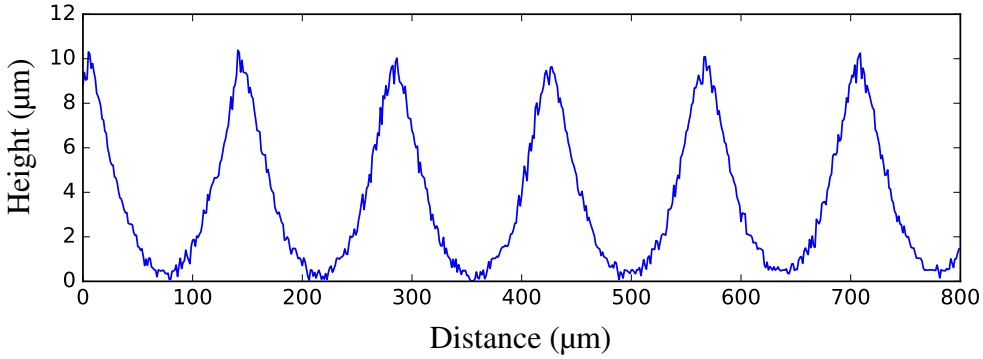
was taken and can be seen in Figure 5.12. Figure 5.12a shows a 3D map of the micropin-fins with the tops truncated so that the base can be better seen. Figure 5.12b shows the profile along the diagonal axis shown in Figure 5.12a. The maximum height variation along this path was found to be 10.4 µm.

The 3D printed enclosure can be seen in Figure 5.13. A recess with a nominal depth of 450 µm exists for the silicon heat sink. A further recess with an additional nominal depth of 450 µm exists to enclose the micropin-fins. The height of the micropin-fins was made to be approximately 10 µm taller than this recess to reduce the likelihood of a gap existing between the tops of the micropin-fins and the plastic. With the micropin-fins taller than the cavity, the edges of the die will not touch the edge of the plastic where epoxy is used to form a seal, but this gap can be filled with epoxy.

Epoxy was dispensed with a syringe along the edges of the cavity before inserting the silicon die. A groove was added between the edges where epoxy was applied and the micropin-fin region, so that epoxy which was pushed out during assembly would fill this



(a) 3D height map of dense micropin-fin base

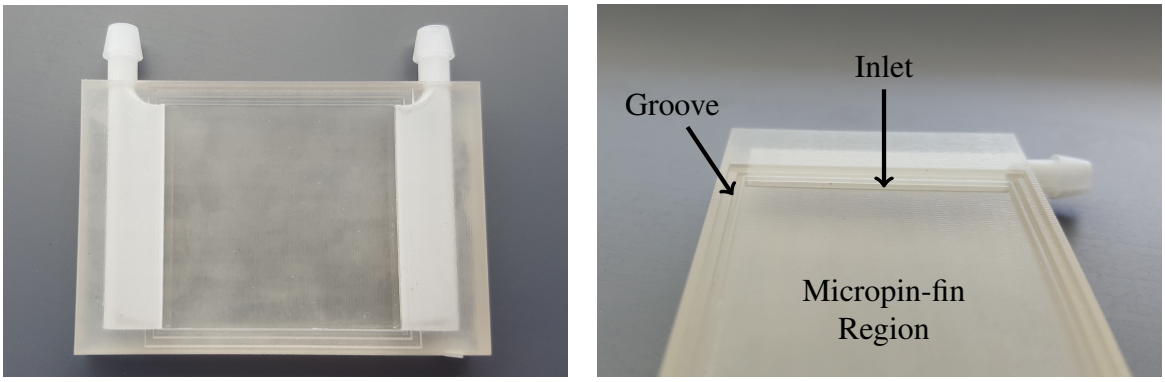


(b) Base height vs. position

Figure 5.12: Profile of dense micropin-fin base

groove before clogging the micropin-fins. A close up photograph of the two levels of recess, the groove, and an inlet slit can be seen in 5.13b.

The height profile across the enclosure was taken using an Olympus LEXT optical profilometer (from left to right in Figure 5.13b) and can be seen in Figure 5.14. As the plots shows, the region of the enclosure which sits atop the micropin-fins is not completely



(a) Top view of 3D printed enclosure (b) Close up of enclosure features

Figure 5.13: 3D printed enclosure and fluid routing device

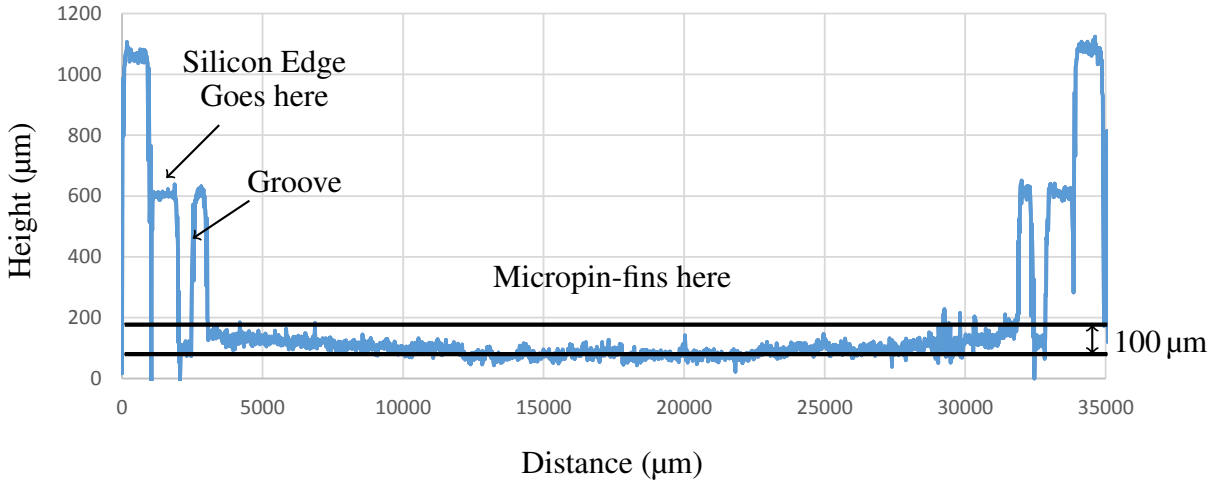


Figure 5.14: Profile of micropin-fin heat sink enclosure

uniform. Therefore, the micropin-fins are likely to make contact at the edges of the cover, leaving up to approximately 100 µm above some regions of the heat sink where fluid can flow outside of the micropin-fin arrays.

After applying epoxy to the edges of the plastic cover, the silicon heat sink was inserted into the cavity and excess epoxy was wiped from the outside of the assembly. A small number of high density micropin-fins were broken over the inactive transceiver region before assembly, but the effect on the active regions is expected to be small. A photograph of the assembled heat sink can be seen in Figure 5.15. The entire heat sink was designed to be nearly planar so that it could make contact with the tops of the silicon dice without

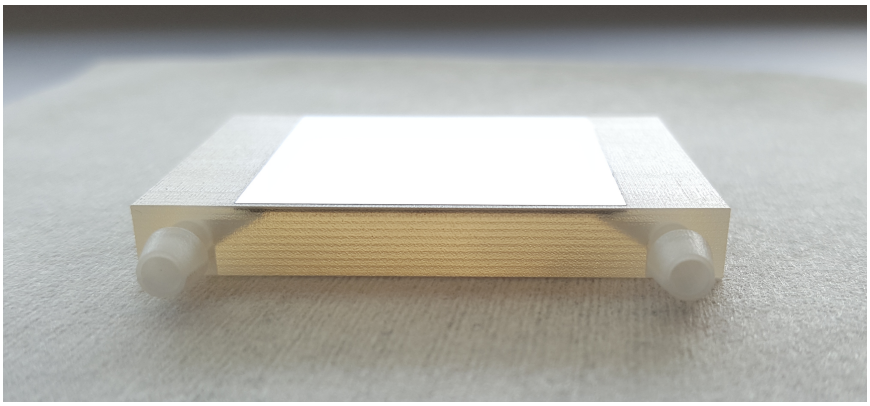


Figure 5.15: Assembled microfluidic heat sink

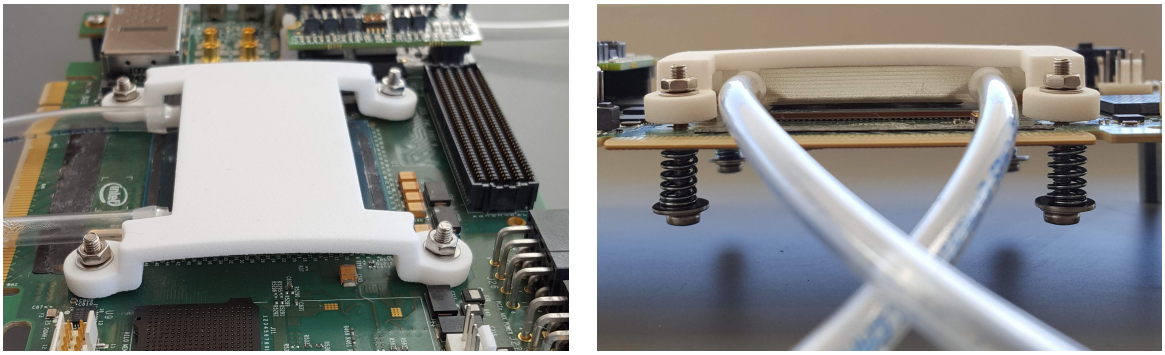
Figure 5.16: Delidded Stratix 10 FPGA on development board

touching any surrounding components or the board.

The lid of the Stratix 10 package on the development board was removed using razor blades and the backsides of the five dice were cleaned with isopropyl alcohol. The board with the delidded die can be seen in Figure 5.16.

After assembling the complete microfluidic heat sink, MasterGel Pro TIM was applied to the back sides of the five Stratix 10 dice and the heat sink was mounted on top. Pressure was applied using a custom 3D printed mounting bracket and the same screws and springs used to previously mount the air cooled heat sink. The edges of the heat sink nearest to the inlet and outlet were visually aligned to the edges of the package, while the other two edges

were aligned to the edges of the mounting bracket, which was designed to be situated in the center of the package and have the same width as the heat sink enclosure. An image of the mounted heat sink can be seen in Figure 5.17.



(a) 3D printed mount (b) Screws and springs applying pressure

Figure 5.17: Custom mounting for microfluidic heat sink on Stratix 10 development board

Experimental Results

The assembled heat sink and board were tested in an open loop system with deionized water as a coolant. Temperature measurements were made at the inlet, outlet, and in the surrounding ambient air using k-type thermocouples. Flow rates were measured with a Kobold rotameter and Ohmega electronic flow meter, both calibrated through repeated filling of a known volume of fluid. Calibration took place with deionized water at $21.2 \pm 0.3\,^{\circ}\mathrm{C}$, which is within $2.4\,^{\circ}\mathrm{C}$ of all temperatures used for testing. Die temperatures were measured using on-die temperature diodes with temperature measurement IP integrated into the FPGA design.

Die temperatures were recorded every second. An initial experiment was conducted at the lowest flow rate to find the time necessary for the temperatures to stabilize. Little systematic variation was observed after the first measurement. Nonetheless, temperatures were allowed to stabilize for one minute prior to taking all measurements with the microfluidic heat sink. A similar experiment was conducted with the air cooled heat sink, which

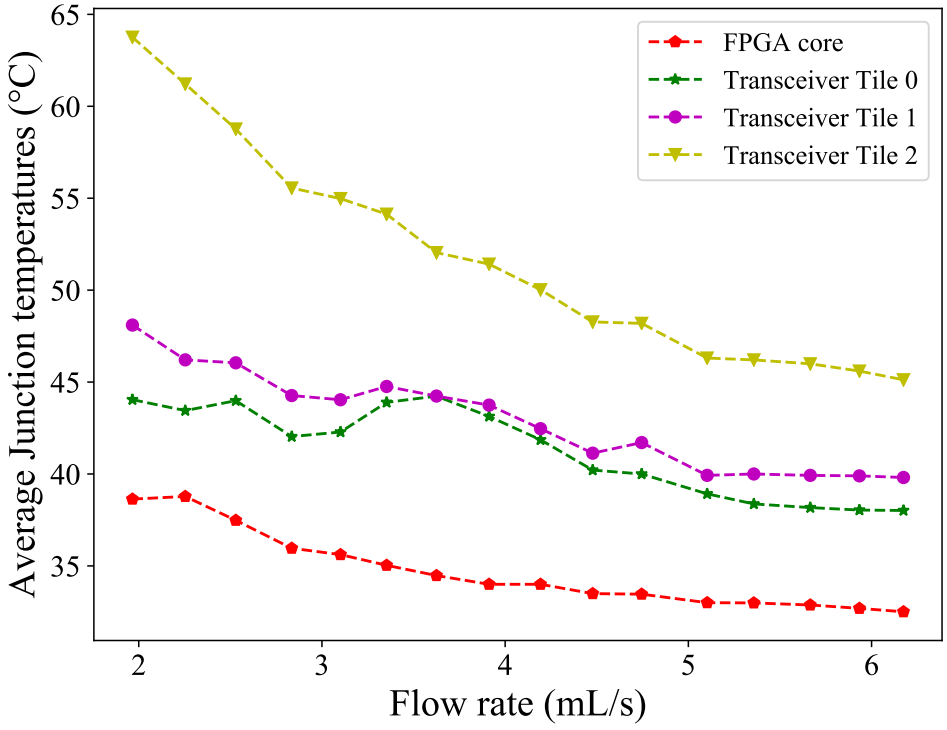


Figure 5.18: Heterogeneous micropin-fin heat sink die temperatures vs. flow rate

took significantly longer to stabilize. Therefore, all of the air cooled measurements in this section were recorded with 15 minutes of wait time.

The temperature of the FPGA die as well as the three active transceiver dice can be seen as a function of flow rate in Figure 5.18. The FPGA die temperature measurement is the lowest of the four dice, while Tile 2 has the highest temperature due to it's close proximity to the outlet. The temperature difference between Tile 2 and the other tiles decreases as flow rate is increased and the fluid temperature at the outlet drops.

Pressure drop, measured across the heat sink, including inlet/outlet ports and short lengths of tubing, can be seen in Figure 5.19, with no power applied to the FPGA.

An additional experiment was performed in which the power on the FPGA die was modulated by varying the clock frequency to the FFT computational blocks. The results of the experiment with air cooling can be seen in Figure 5.20. As expected, the temperature of

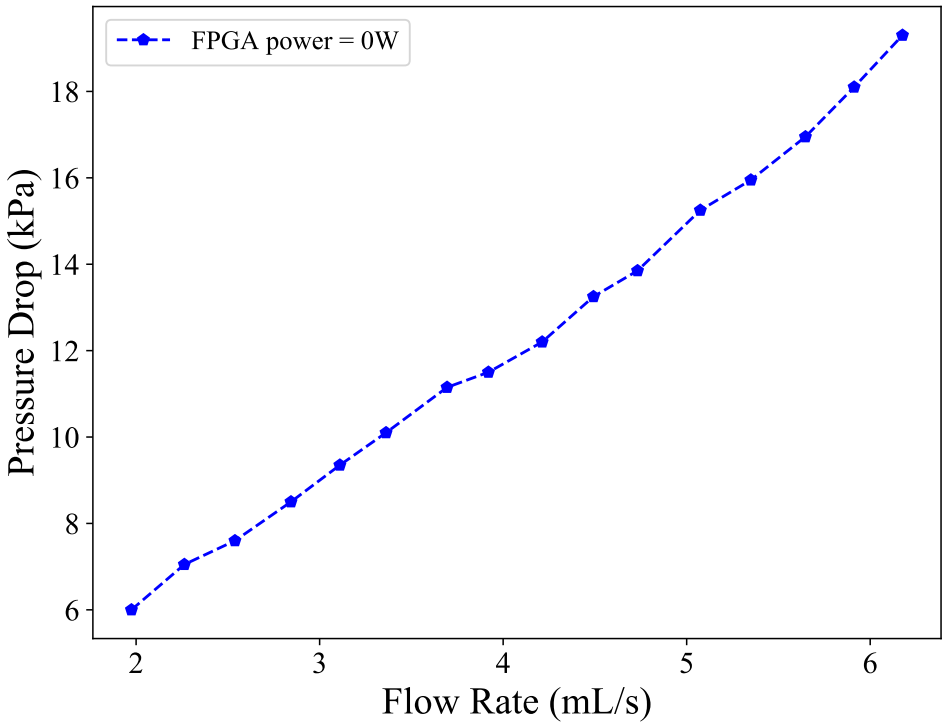


Figure 5.19: Heterogeneous micropin-fin heat sink pressure drop vs. flow rate

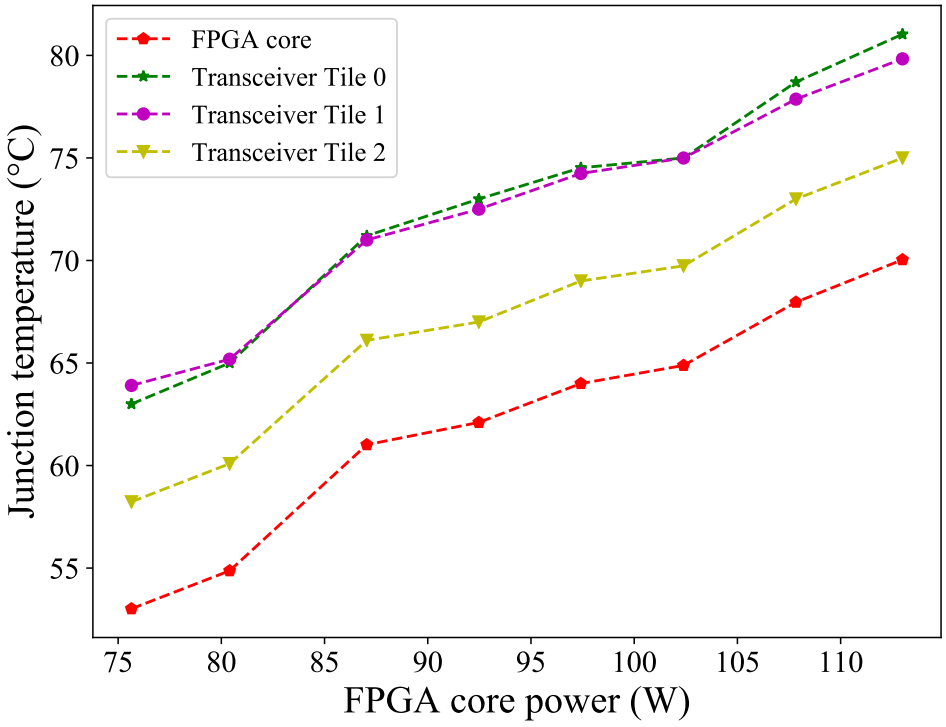


Figure 5.20: Air cooled heat sink die temperatures vs. FPGA core power

the FPGA die increases with increasing power, with a slope of approximately 0.46 °C/W, computed from the first and last points in the plot. However, it can also be seen that the slopes of the transceiver die temperatures are all approximately equal to the slope of the FPGA die temperature. From the lowest FPGA die power of 75.6 W, corresponding to an FFT clock frequency of 300 MHz, to the highest FPGA die power of 113 W, corresponding to 475 MHz, the temperature rose 17.0 °C degrees on the FPGA die while the temperatures of transceiver tiles 0, 1, and 2 rose by 18.0 °C, 15.9 °C, and 16.8 °C, respectively. This indicated strong thermal coupling between adjacent dice, where the conditions on the FPGA die have a large effect on the temperatures of the surrounding dice.

Similarly, the power of the transceivers was changed to three different values by modulating the data rates of the transceivers while the FPGA core power was kept constant. Due to the details of the FPGA design, varying the transceiver data rates also effected the

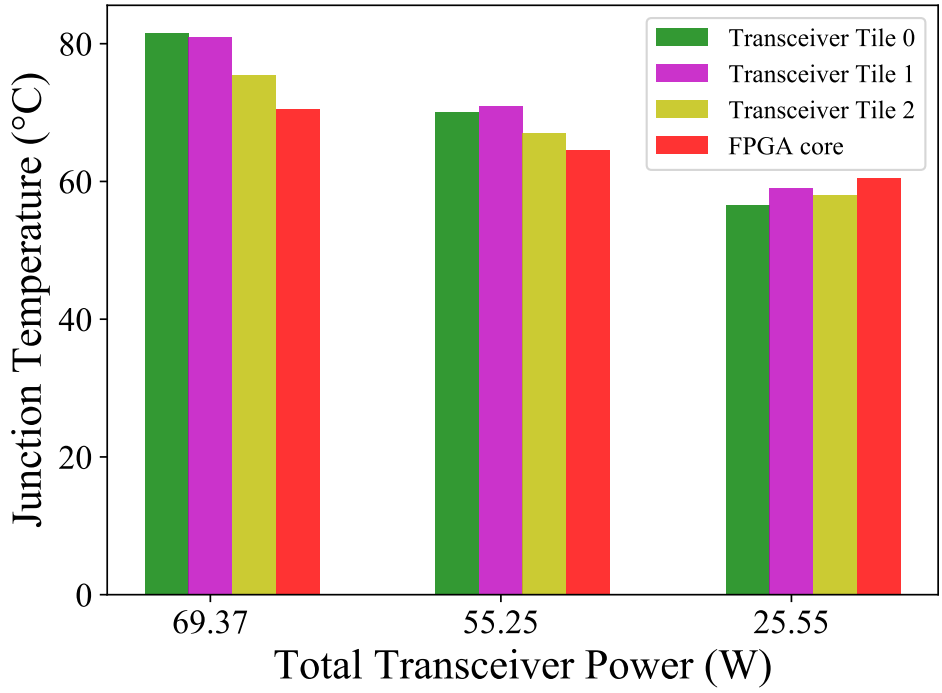


Figure 5.21: Air cooled heat sink die temperatures vs. transceiver power

FPGA die power. Therefore, the FFT clock rate was modulated to keep FPGA die power within 1.7% of 111 W for all three transceiver powers shown in Figure 5.21, thus isolating the effect of transceiver power on die temperatures. The transceiver temperatures change by 17.5 °C to 25 °C when the power is varied from 69.37 W to 25.55 W. This change in transceiver power, while holding FPGA die power nearly constant, led to a change in FPGA die temperature of 10 °C. Hence there is strong thermal coupling from the FPGA dice to surrounding dice (Figure 5.20) and from surrounding dice to the FPGA (Figure 5.21). It is believed that this strong thermal coupling is primarily due to the very thick heat spreaders, both integrated into the lid of the die, and the base of the air cooled heat sink.

The power of the FPGA die was similarly varied with the microfluidic cooled heat sink with a flow rate of 6.18 mL/s and an inlet temperature of 19.5 ± 0.3 °C. The results can be seen in Figure 5.22. The FPGA die temperature increased with increasing power, with a slope of approximately 0.054 °C/W, which is significantly lower than the slope seen with the air cooled heat sink because of the comparatively lower thermal resistance of the mi-

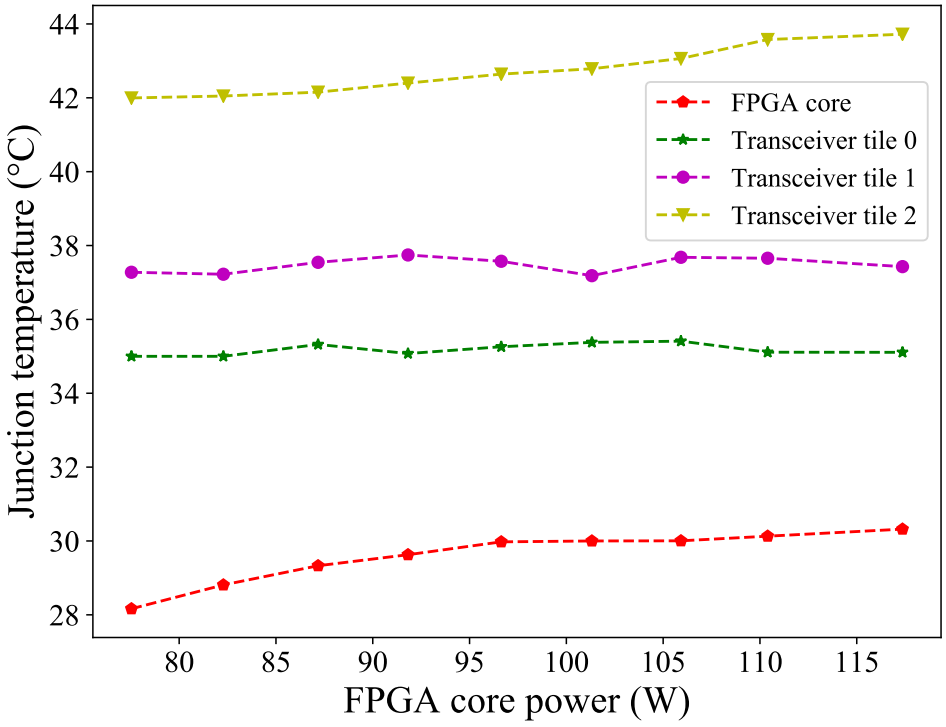


Figure 5.22: Heterogeneous micropin-fin heat sink die temperatures vs. FPGA core power

crofluidic heat sink. While the temperature of the transceiver tiles with air cooling closely followed the temperature of the FPGA die, the temperatures of transceiver tiles 0 and 1 are nearly constant with microfluidic cooling as the FPGA die power changes from 77.5 W to 117.4 W. The average Tile 2 temperature measurement increases by 1.7 °C because it is located close to the outlet and likely receives more fluid which has been warmed by the FPGA die.

Transceiver powers were similarly varied with the microfluidic heat sink with a flow rate of 6.19 mL/s and inlet temperature of 19.2 ± 0.2 °C. The results can be seen in Figure 5.23. While transceiver temperatures dropped by 10.1 °C to 14.6 °C when total transceiver power was changed from 63.34 W to 22.61 W, the average temperature measurement on the FPGA die only changed by 0.13 °C. The effects of FPGA die power on the temperatures of all four dice with both the air cooled heat sink and microlfuidic cooled heat sink are summarized

Table 5.2: Change in die powers as a function of FPGA die power

Die	Air	Microfluidic
FPGA	0.46 °C/W	0.054 °C/W
Transciever Tile 0	0.48 °C/W	0.003 °C/W
Transciever Tile 1	0.43 °C/W	0.004 °C/W
Transciever Tile 2	0.45 °C/W	0.043 °C/W

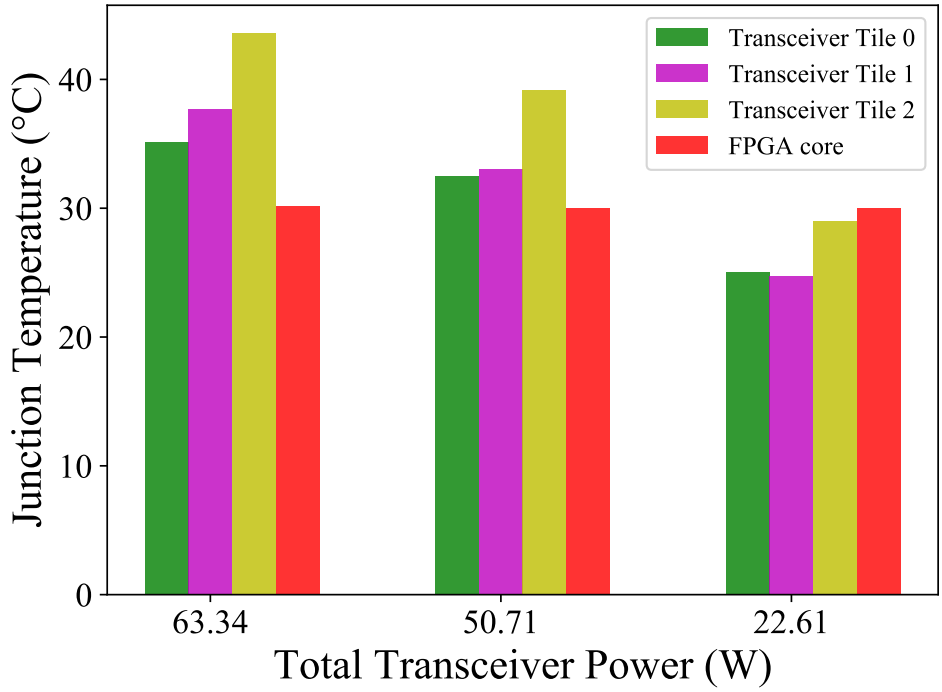


Figure 5.23: Heterogeneous micropin-fin heat sink die temperatures vs. transceiver power

in Table 5.2.

This significant reduction in thermal coupling between adjacent dice is likely a result of the significantly reduced thickness of the microfluidic heat sink. Heat is rapidly transferred to the fluid and extracted, whereas heat must conduct through a large shared thermal mass with the air cooled heat sink. This also has an effect on the thermal time constant of the two systems. Before taking measurements with either heat sink, the temperature of both were measured over a period of approximately thirty minutes to determine the time that was required for the temperature to stabilize. It was observed that the thermal time constant of the air cooled heat sink was approximately 10 min to 15 min while the thermal time

Table 5.3: Stratix 10 temperature measurements with air cooled heat sink and microfluidic cooled heat sink

Die	Air	Liquid	Air with no Compute
FPGA	70.5 °C	30.1 °C	38.5 °C
Transciever Tile 0	81.5 °C	35.1 °C	48.5 °C
Transciever Tile 1	81 °C	37.7 °C	52 °C
Transciever Tile 2	75.5 °C	43.6 °C	46 °C

constant of the microfluidic cooled heat sink was approximately 1 s to 3 s.

The temperature measurements corresponding to the baseline design with 475 MHz FFT clock and transceivers at 22 Gbit/s are summarized in Table 5.3 for both the air cooled heat sink and the microfluidic heat sink at the highest flow rate. The temperatures of all four dice with the microfluidic cooled heat sink were lower than those with the air cooled heat sink. For reference, the results of an additional test with the air cooled heat sink can also be seen in Table 5.3 where all 160 FFT compute cores were deactivated, with the transceivers remaining on. It can be seen that to reach temperatures similar to those of the microfluidic-cooled dice, all FFT computation needed to be shut down.

In addition to the benefits of low temperatures and low thermal coupling, the microfluidic heat sink offers the potential for very high density computing. The height of the air cooled heat sink used for baseline measurements was 172.2 mm, while the height of the microfluidic cooled heat sink is 6.5 mm. If the boards were to be stacked at high density using, for example, PCIe connections, the pitch of these boards would not be limited by the height of the heat sink, but rather by other connectors on the board, which give the board a total height of approximately 17.4 mm. This would still allow the boards to be mounted at approximately $10\times$ the density of boards with the air cooled heat sink. Lower profile air cooled heat sinks are available, but using them would likely require further sacrifices in other areas, such as die temperature.

Modest improvements in total power draw were also realized through reduced temperatures provided by the microfluidic cooled heat sink. When running the baseline design

with the air cooled heat sink, a total power of 182.5 W was measured. With the microfluidic cooled heat sink, this power was 174.4 W.

Conclusion

In this chapter, a microfluidic heat sink was designed, fabricated, and tested for the cooling of a 2.5D Stratix 10 FPGA. Results were compared with a high performance air cooled heat sink and a temperature reduction ranging from 31.9 °C to 46.4 °C was observed across the four active dice under test. It was also found that thermal coupling between the FPGA die and surrounding transceiver dice was very strong with the air cooled configuration. For example, the temperatures of the transceiver dice varied by approximately the same amount as the FPGA die temperature as the FPGA power was varied. With the microfluidic cooled heat sink this effect was reduced on the transceiver tile near the outlet and nearly eliminated on the other transceiver tiles. In addition to the reduction in temperature and thermal coupling between dice, a large increase in compute density was also enabled through the use of the low profile microfluidic heat sink. The final heat sink was limited in its height by the diameter of the tubes delivering coolant, rather than the height of the micropin-fins.

Although the temperatures of all of the dice were reduced relative to experiments with the air cooled heat sink, the transceiver tiles remained warmer than the FPGA die. This runs counter to one of the goals of the microfluidic heat sink design, which was to normalize temperatures across all of the dice by varying the micropin-fin density. This could be due to a number of factors. First, fluid may have bypassed these higher density regions, either through the gap between the tops of the micropin-fins and the cover, or through the lower density micropin-fins in the center region of the heat sink. These challenges will be addressed in the following chapter on future work.

CHAPTER 6
SUMMARY AND FUTURE WORK

This thesis aims to demonstrate several microfluidic cooling technologies to enable future high density, high performance microelectronics. The following sections highlight the contributions presented in this thesis.

Summary of Work

The following research projects were completed and presented in the previous chapters.

- Chapter 2 explored the concept of dedicated microgaps for hotspot cooling. First, a dedicated microgap device was designed, fabricated, and tested with R134a as a two phase coolant with a heat flux up to 5.2 kW/cm^2. Next, a test vehicle was designed and fabricated which consisted of a combined microgap and micropin-fin heat sink for simultaneous cooling of a background heat flux of 100 W/cm^2 and hotspot heat flux of up to 6.175 kW/cm^2. It was found that the dedicated microgap reduced the temperature of the hotspot relative to the average background temperature, allowing higher heat flux dissipation on the hotspot, particularly if chilled fluid was used at the inlet of the hotspot cooler. Since the hotspot uses only a small fraction of the fluid used to cool the background, cooling the fluid entering the microgap would use a fraction of the energy necessary to cool fluid for the entire device.

- Chapter 3 presents heterogeneous micropin-fin heat sinks used to cool non-uniform power maps. The test chips were used to cool a background heat flux over the 1 cm $\times$ 1 cm background region as well as a 500 µm $\times$ 500 µm hotspot region in the center of the chip. Micropin-fins were locally clustered over the hotspot with a density which was twice that of the background micropin-fins. Four micropin-fin heat

sinks were tested, two with cylindrical micropin-fins and two with hydrofoil shaped micropin-fins. Two of the heat sinks only had the higher density micropin-fins over the hotspot region of the chip, while the other two had a high density region spanning the entire width of the chip to prevent flow bypass. All four chips effectively cooled both the background and hotspot, each dissipating 250 W/cm^2 and 500 W/cm^2, respectively. The test chip with hydrofoils directly over the hotspot maintained an average hotspot and average background temperature difference of less than 0.25 °C.

- Chapter 4 presents a Stratix V FPGA with a monolithically integrated micropin-fin heat sink. Though microfluidic heat sinks have been studied for many years, this is the first demonstration of a CMOS processor with monolithically integrated microfluidic cooling. Thermal results are recorded and compared with those with the stock air cooled heat sink as well as a hypothetical best-case air cooled heat sink. A thermal resistance of 0.07 °C/W as well as substantial improvements in temperature, throughput, and leakage power are demonstrated.

- Chapter 5 presents a low-profile heterogeneous micropin-fin heat sink used to cool a Stratix 10 FPGA which consists of five dice mounted adjacent to one another in a single package. The heat sink consists of an etched silicon insert and a 3D printed plastic cover/manifold. The low-profile heat sink is able to provide a lower thermal resistance than an air cooled heat sink which is orders of magnitude larger while also separately targeting the cooling needs of the FPGA die and transceiver dice and reducing thermal coupling between the dice. The thin profile facilitates high-density stacking of accelerator boards within a system for a small footprint and low signaling overhead.

- Appendix A presents much of the microfabrication process development that went into the work shown in this thesis as well as several other microfluidic cooling experiments which were not included. The design of a 1 cm $\times$ 1 cm micropin-fin test

device is presented, which uses a fabrication process which is representative of many of the processes used to fabricate the devices shown in earlier chapters. Three generations of devices were created, which included several innovations such as strategically placed support structures, pressure ports, flow stabilization pin-fins, and a small footprint facilitated through wirebonding and O-ring seals

Future Work

Low Profile Heat Sink for 2.5D Packages

In designing the microfluidic heat sink for the Stratix 10, we assumed a uniform inlet flow rate. In reality, the manifold does not deliver a uniform velocity across the width of the micropin-fin heat sink and a uniform velocity may not be optimal. For example, it may be desirable to direct more coolant to the sides of the heat sink where the higher heat flux density transceivers are located. In future work, the manifold could be better optimized to deliver coolant to where it needs to be, potentially using area above the micropin-fin heat sink to deliver fluid, which could reduce pressure drop and temperature.

The base of the heat sink could also be thinned to reduce heat spreading and more accurately target the cooling needs of each die. As shown in the beginning of the chapter 5, separate cooling chiplets could be integrated into the plastic manifold, which could further reduce thermal coupling between separate parts of the die through conduction and by separating the coolant flow.

Additionally, the cooling module used to cool the 2.5D package could be migrated from a design which uses TIM to conduct heat to the heatsink, to a design in which the coolant makes contact with the back side of the silicon. This could be achieved by etching all of the silicon dice, similar to the Stratix V demonstration, or by using jet impingement to achieve a high heat transfer coefficient without etching the silicon. Either design would use the attached manifold to deliver coolant to each of the dice.

Micropin-fin Heat Sink Optimization Framework

Some of the work presented in this thesis has shown that a background region and hotspot region can be cooled separately by simply increasing the density of micropin-fins over the hotspot region. However, this still leaves a temperature gradient across the chip due to heating of the fluid and could become complicated if one is to manually design a heat sink for a power map which consists of more than a few uniform regions. This method of locally varying micropin-fin density could be automated to cool arbitrary non-uniform power maps.

Unfortunately, performing a full global optimization of micropin-fin placement and geometry for heat sinks containing thousands of micropin-fins is infeasible. With n micropin-fins and m dimensional parameters for each micropin-fin, a global optimizer must search a nm dimensional design space. If each dimension is discretized into d points for sampling, the total run time would be proportional to d^{mn}. Hence, the running time grows exponentially with the number of micropin-fins if the entire design space is to be searched. While a general purpose finite volume method (FVM) simulation using a tool such as ANSYS FLUENT may take several hours for a large geometry, even specialized simulators which decrease this simulation time by orders of magnitude will not be able to explore the entire design space, even for reasonably small d and m.

It is, however, possible to find local optima in the micropin-fin design space. Prior work shown in this thesis has demonstrated that local clustering of micropin-fins can be an effective means of simultaneously meeting the cooling needs of integrated circuit hotspots and background heat fluxes[50, 63, 48]. This prior work indicates that a greedy algorithm, which locally increases micropin-fin density to reduce temperature, may yield a desirable local design optimum in the micropin-fin heat sink design space. Inspired by these results, the following preliminary work investigates optimization of micropin-fin arrays using a local optimization algorithm. An optimization loop was built around ANSYS FLUENT which was used to simulate chip temperatures for silicon micropin-fin heat sink designs.

The following subsections will present this optimization framework and some preliminary results.

The structure being optimized in this work consists of a silicon micropin-fin heat sink with water as a coolant. The heat sink has a total height of 500 µm and is consistent with dimensions that could be etched into active or inactive silicon dice.

Optimization Code Design

ANSYS includes built-in automation and optimization routines, but they require that the design be broken down into a manageable number of parameters, which is not practical for a large micropin-fin geometry. Additionally, converting these high level parameters into a heat sink geometry is not trivial within ANSYS. Performing the geometry generation outside of ANSYS within a custom program allows a large degree of flexibility to explore algorithms for optimization and pin-fin placement.

To perform the optimization, the micropin-fin heat sink area was split into a grid of "cells," each containing a number of micropin-fins with a fixed pitch and diameter. Given a pitch and diameter for each cell, a 3D model of the heat sink geometry is generated in FreeCAD. This design is imported into ANSYS workbench, where a simulation is performed to generate temperatures and an inlet pressure. The average pressure at the inlet as well as average temperatures on all of the cell faces are then extracted. These values are then used to update the geometry and the cycle is repeated. A diagram of this optimization loop can be seen in Fig. 6.1.

Geometry Creation

Geometry creation begins with an area to be populated with micropin-fins. This area is split into a number of "cells," which each contain micropin-fins. Each cell has a high level micropin-fin "density" which is later converted into actual micropin-fin dimensions. A custom Python script creates this cell array and updates the density numbers based on

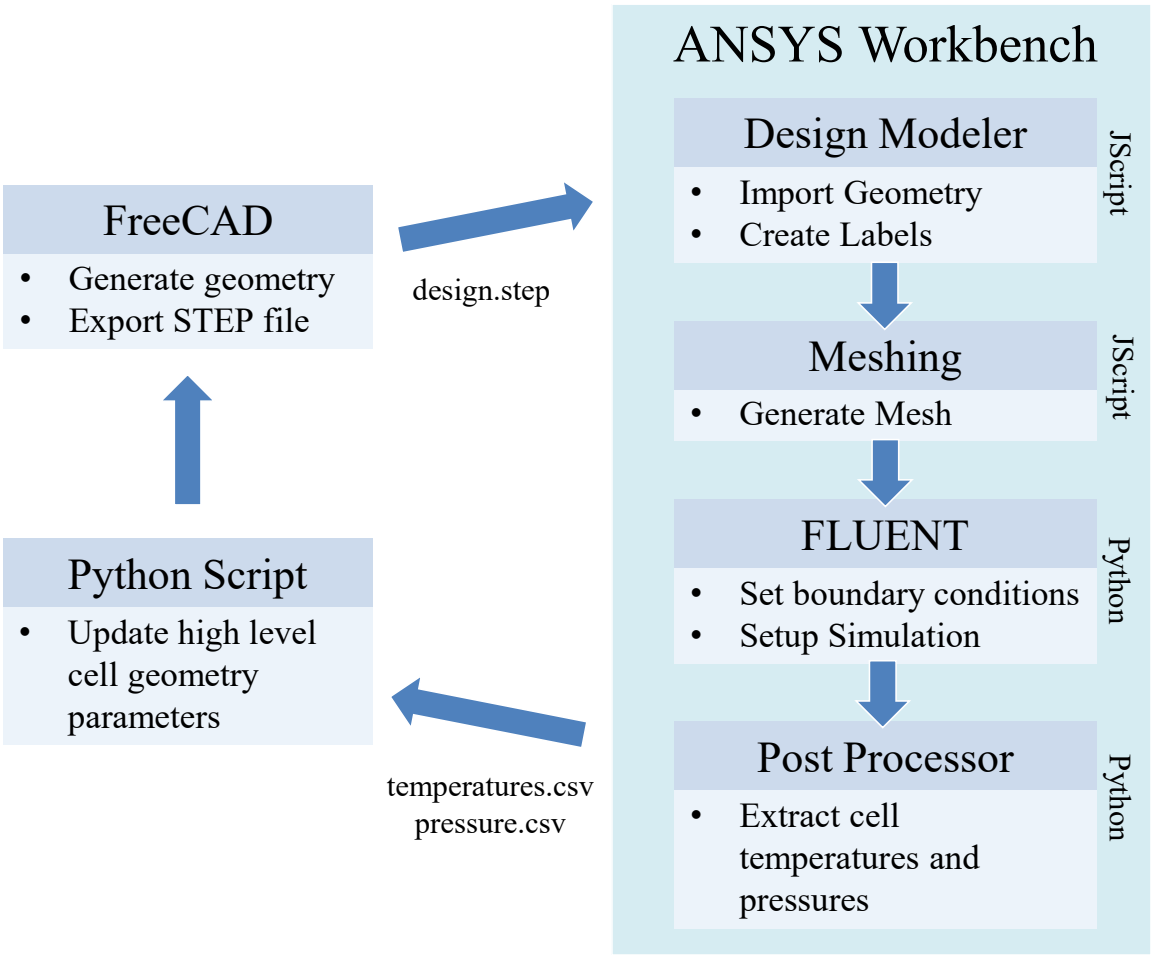


Figure 6.1: Diagram of optimization loop

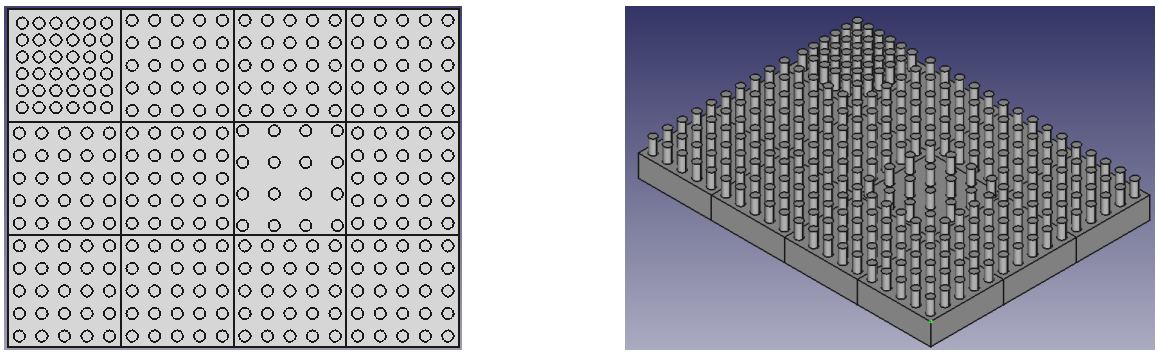


Figure 6.2: Example 3×4 cell geometry in FreeCAD

temperatures from the previous iteration. Once density numbers are updated, micropin-fin dimensions and locations are calculated. These numbers are used to create a 3D model using FreeCAD v0.17, an open-source computer-aided design (CAD) tool with a Python back end and application programming interface (API). The API is used to generate a a geometry and export it as a STEP file which is then imported into ANSYS Design Modeler. An example of a micropin-fin geometry with 3×4 cell can be seen in Fig. 6.2.

ANSYS Automation

The workflow within ANSYS Workbench begins with Design Modeler, then Meshing, and then FLUENT simulation and post processing. ANSYS Design Modeler and Meshing internally use Microsoft JScript to perform operations within the graphical user interface (GUI). Although these tools support running custom JScript code which can hook into some of this native functionality, there is no documented API. Therefore the majority of the functionality used for this automation was discovered by examining ANSYS JScript and XML source files and by stepping through the Visual Studio debugger while using the Design Modeler GUI.

The ANSYS simulation consists of a number of steps which are all automated within ANSYS Workbench. ANSYS workbench supports Python-based journaling to control functionality within workbench. This was used to start each of the workbench modules and

feed them commands in their respective languages. First, the Design Modeler is opened and fed a script written in JScript to control Design Modeler functionality. The STEP file is imported and faces on the geometry are matched to ID numbers in a comma-separated value (CSV) file based on face centroids and surface area. Named Selections are added to these faces to identify them in FLUENT for the application of boundary conditions. Next, ANSYS meshing is used to create a mesh for the geometry, which is also controlled using custom JScript.

After the geometry is imported, labeled, and meshed, the FLUENT simulation is set up. Unlike Design Modeler and Meshing, FLUENT commands executed in the GUI can be captured into a Python script using the Workbench journaling feature. Within FLUENT, appropriate model parameters and material properties are first selected. A velocity inlet boundary condition is applied to the inlet surface and heat fluxes are applied to the bottom faces of the geometry, identified using the aforementioned CSV file which matches face geometry, ID number, and heat flux for the model.

After the FLUENT simulation completes ANSYS Post Processing is opened. A table is created with cell surfaces and their average surface temperatures. These cell temperatures, along with an average pressure measurement across the inlet surface, are then exported to CSV files which are fed back to the geometry updating algorithm.

Preliminary Results

A 5 mm $\times$ 5 mm silicon micropin-fin heat sink was used as a test case for optimization. A constant background heat flux of 100 W/cm^2 was set, except for a 1 mm $\times$ 1 mm cell in the center of the chip, which had a heat flux of 200 W/cm^2. The same temperature dependent material properties mentioned in chapter 5 were used.

As mentioned previously, there are many options of parameters to optimize. In general, pressure drop, pumping power, average temperature, maximum temperature, and temperature variation may be parameters which would benefit from minimization. Temperature

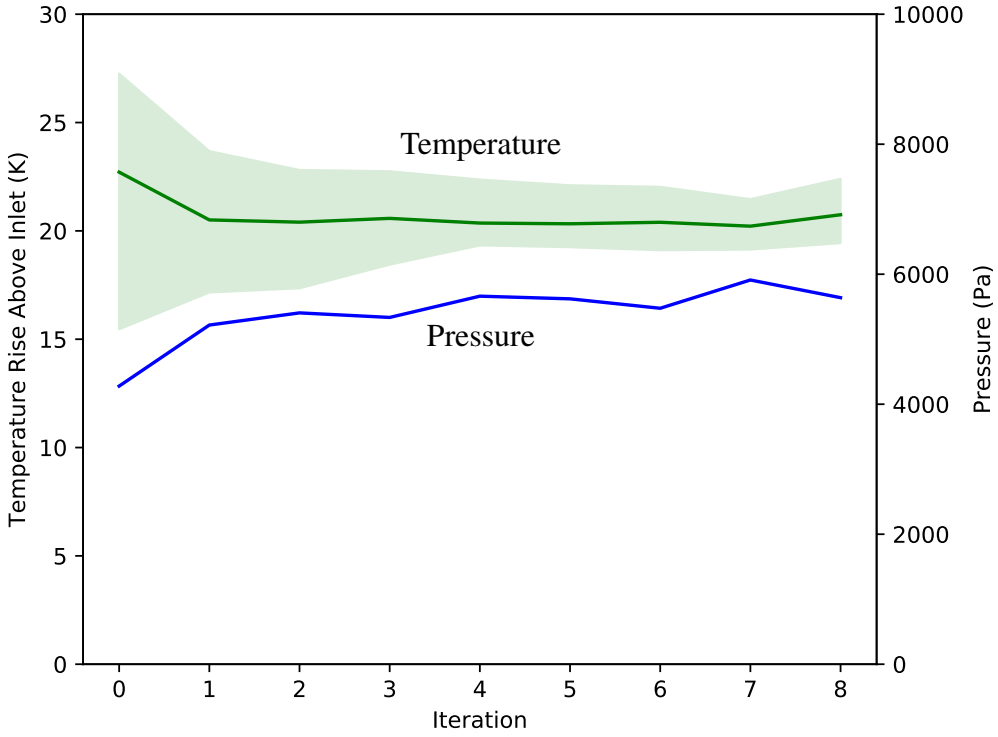


Figure 6.3: Temperature and pressure vs. optimization iteration with fixed aspect ratio

variation was directly targeted for optimization in this demonstration. This will also indirectly entail minimization of the maximum temperature, which will often be a limiter of performance.

The geometry can also be modified in multiple ways. First, a pitch to diameter ratio was fixed, so that both pitch and diameter of micropin-fins was simultaneously varied to modify density. This sometimes lead to extremely large micropin-fins, so a maximum aspect ratio of 1 was set. For each iteration of the optimization, the average temperature of each cell was recorded as well as the inlet pressure. The results can be seen in Figure 6.3. The average cell temperature is shown as a line, while the range from the minimum cell temperature to the maximum cell temperature is shown as a shaded region around this line.

Both maximum temperature and temperature variation across the chip were successfully reduced. The minimum temperature was increased because it is considered less important than the maximum temperature and decreasing the micropin-fin density can have

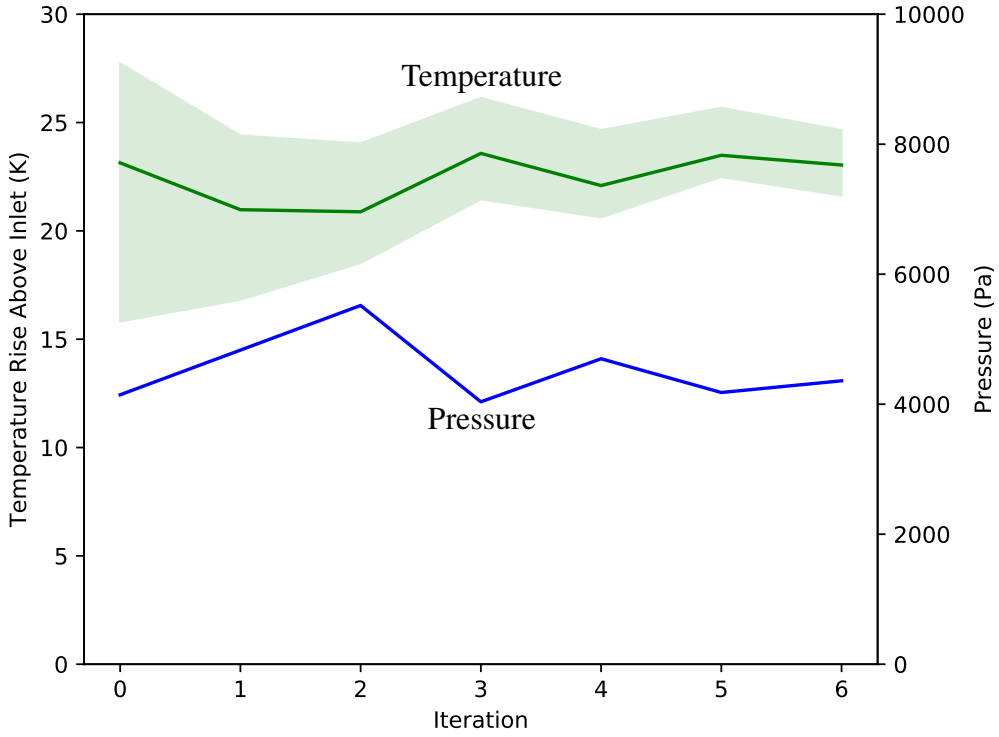


Figure 6.4: Temperature and pressure vs. optimization iteration with fixed aspect ratio and pressure constraint

pressure drop benefits. Since no explicit pressure constraint was set, however, the pressure also increased from 4.3 kPa to 5.6 kPa. Therefore, another optimization was performed with an explicit limit on pressure drop. When pressure extracted from the ANSYS simulations exceeded this prescribed limit, micropin-fin density was reduced across the entire chip to reduce pressure drop. The results of this simulation can be seen in Figure 6.4.

Once again the maximum temperature is reduced while the inlet pressure changes from an initial pressure of 4.1 kPa to a pressure of 4.4 kPa on the last iteration. The current technique of limiting pressure results in a pressure which can oscillate around the set point before converging. The fifth iteration demonstrates both a decrease in maximum temperature as well as a decrease in pressure drop relative to the initial geometry. These saving are primarily obtained by sacrificing heat transfer in the locations where it is less needed. The optimized design can be seen in Figure 6.5. Micropin-fin geometry is reduced towards

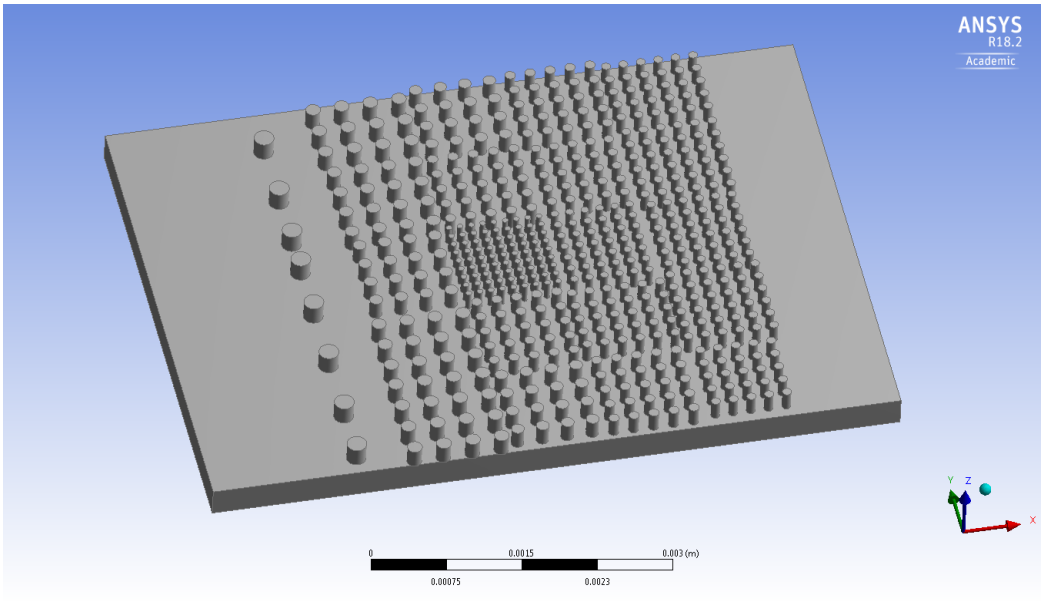


Figure 6.5: Optimized micropin-fin heat sink geometry

the inlet of the heat sink, where fluid temperature is lower, and increased towards the outlet, where fluid temperature is higher. The center of the chip also contains higher density micropin-fins to cool the increased heat flux located there.

Future Work

The current results indicate that local optimization of micropin-fins can be an effective means of reducing maximum temperature and pressure drop. In future work, this optimization can be enhanced in a number of ways.

Currently, the final iteration does not necessarilly yield the best results. The updates to the geometry could be modified to smooth these variations. Additionally, different stopping criteria could be explored. This may also involve improving the efficiency of the simulations in terms of mesh size and computation, since each simulation can take several hours due to their size and complexity.

One of the main advantages of the optimized design is that it counters the temperature gradient caused by heating of the fluid, even with a uniform power map. This effect, however, is relatively small in the small geometries presented above and would increase

in a longer chip. Therefore, a range of micropin-fin heat sink sizes could be explored to demonstrate the benefit as a function of heat sink size.

Additionally, the manner in which the high level parameter of cell density is converted into actual micropin-fin geometries could be experimented upon. Currently, the pitch-to-diameter ratio is kept constant, except when the diameters become "too large." This limit is currently defined as an aspect ratio limit of 1. There may be a benefit to keeping the diameter constant, at the maximum practical aspect ratio, and simply varying the number and pitch of these micropin-fins.

Since these are local optimizations, they may be sensitive to the starting point of the parameters to be optimized. The sensitivity of the final design could therefore be studied as a function of the initial design.

Lastly, although it was mentioned that a global optimization is not practical, a comparison could be made between a stochastic global optimization scheme within a fixed time constraint. With a reasonable time limit in place, it is expected that the presented local optimization scheme would outperform a global optimization scheme.

Appendices

APPENDIX A

MICROFLUIDIC COOLING TEST DEVICE DESIGN AND FABRICATION

Throughout this work, a number of microfluidic heat sink test devices have been designed and fabricated to test various types of heat transfer structures. Several generations of micropin-fin test devices were developed for various single and two phase experiments with micropin-fins, as well as some other heat transfer structures. Many of these designs consisted of a 1 cm $\times$ 1 cm micropin-fin array. This appendix will highlight the fabrication recipe and three generations of such micropin-fin test device designs. The fabrication recipe is also representative of the recipes used to fabricate the other devices shown in this Thesis.

A cross section of a micropin-fin test device can be seen in Figure A.1. Fluid is pumped through an inlet port on one side of the device, passes over the micropin-fins on the device, and then leaves through the exit port. Heat is generated through serpentine platinum traces on the bottom side of the test device. The cavity around the micropin-fins is sealed using an anodically bonded Pyrex cap.

Micropin-fin Test Device Fabrication

A diagram depicting the fabrication steps for the micropin-fin test device can be seen in Figure A.2. The process begins with a DSP prime grade 4 inch silicon wafer. First, a cavity is etched into into one side of the wafer using the Bosch process (Table A.1) and NR5-8000 photoresist as a masking layer (Table A.2). This cavity defines the region which will contain fluid, leaving behind the micropin-fins and support structures. An etch depth of 200 µm is targeted by etching in two steps. First, an etch depth of approximately 190 µm is targeted assuming an etch rate of 0.88 µm per etch cycle, the average for many test runs. Then the height of the etched cavities is measured on several points across the wafer using

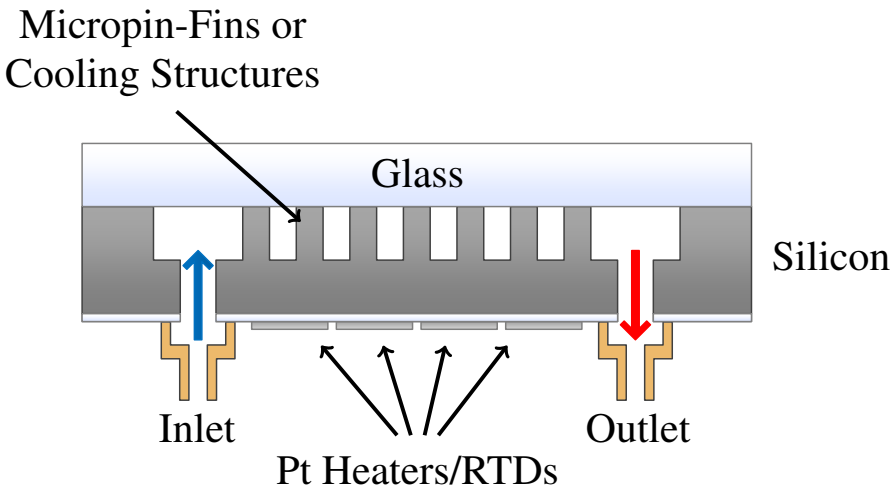


Figure A.1: Cross-sectional diagram of micropin-fin test device

Table A.1: Bosch Process Etching Parameters

Tool	STS ICP
Etch Time	14 s
Passivation Time	17.5 s
Platten Power	20 W
Plasma Power	800 W

a profilometer to find the average etch depth. Photoresist thickness is assumed to remain constant because the etch/photoresist recipes yield high selectivity in this process. The remaining etch cycles are carried out, assuming an etch rate consistent with the etch rate up to that point. The height can then be fine tuned one more time with another measurement/etch cycle because the etch rate usually decreases with etch depth.

Next a Pyrex cap is anodically bonded to the tops of the micropin-fins to seal the fluid cavity using a 350 °C process with a 1200 V under vacuum. Reliable bonding requires very

Table A.2: NR5-8000 Lithography Parameters (for etching)

Spin Coat	1100 rpm, 40 s
Thickness	16 µm
Pre exposure bake	5 min@80 °C oven + 60 s@150 °C hotplate
Exposure	340 mJ/cm^2, 365 nm
Post-exposure bake	60 s@100 °C hotplate
Develop	90 s RD6 with agitation, more as needed

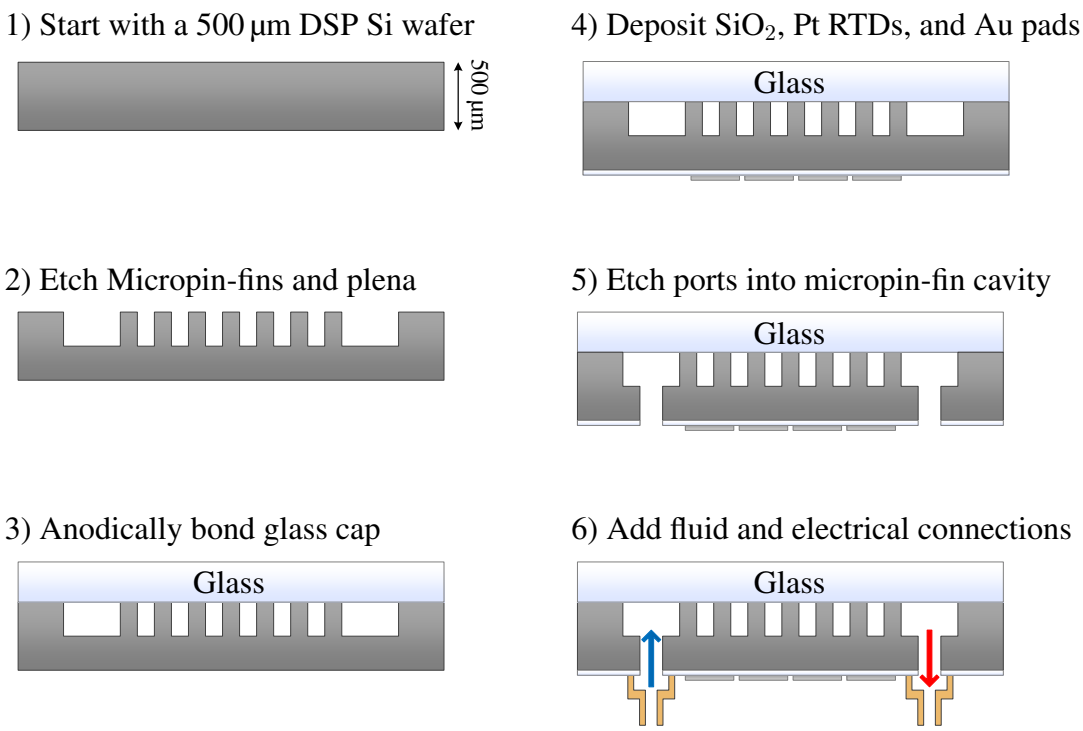


Figure A.2: Fabrication flow for micropin-fin test device

clean surfaces, so both the silicon and Pyrex are first cleaned with Piranha and an oxygen plasma to remove any residue. A 2 µm silicon dioxide layer is then deposited through PECVD across the entire silicon side of the bonded wafers. This dielectric layer acts as an electrical insulator on which conducting traces can be deposited.

The etching and bonding processes, to this point, have left a very smooth silicon surface on which to deposit RTDs and electrical traces. NR9-1500PY is used for a lift off process which defines the serpentine RTD traces as well as the bonding pads and thicker traces connecting these elements (Table A.3). An oxygen plasma descum is performed after photodefinition, before metal deposition to clean the surface of any remaining organic residue and promote adhesion. This is achieved through backside alignment with the etched features visible through the Pyrex cap. A titanium adhesion layer is sputtered on the wafer, followed by platinum. Liftoff is done by soaking the wafer in acetone for several hours, followed by up to 15 min of ultrasonic agitation in acetone. After performing lift off, a sub-

Table A.3: NR9-1500PY Lithography Parameters (for liftoff)

Spin Coat	1000 rpm, 40 s
Thickness	2.5 µm
Pre exposure bake	120 s@150 °C hotplate (to account for thermally resistive glass)
Exposure	475 mJ/cm^2, 365 nm
Post-exposure bake	120 s@100 °C hotplate
Develop	15 s RD6 with agitation, more as needed

Table A.4: Metalization Parameters

Adhesion Layer	30 nm titanium
Platinum Layer	0.2 µm
Copper Layer	0 µm to 2 µm (depending on resistance requirements)
Gold Layer	0.5 µm

sequent photodefinition and lift off process is carried out to deposit copper and gold onto the bonding pads and traces connecting these pads to the RTDs. A summary of the metal layers can be found in Table A.4. In later generations of devices, a second silicon dioxide passivation layer was then deposited on top of the metal layers to protect the platinum RTDs from the surrounding environment. This was found to reduce drifting of the RTD temperature-resistance calibration which occurred in earlier devices without passivation. After depositing the passivation oxide, an additional masking step is required to open the bonding pads for electrical connection.

The ports are then etched last. Leaving this step for last leaves a relatively smooth surface on which to do the previous metalization steps, and also prevents the fluid channels from being clogged with photoresist during subsequent processing. A Bosch process is used to etch the ports and the wafer is periodically inspected through the Pyrex cap to determine when the etching penetrates into the channel. Each cycle in the Bosch process consists of two steps: an etching step and a passivation step. The passivation step is substantially more anisotropic than the etching step which removes it, which can therefore cause passivation material to enter the fluid cavity if allowed to etch too long. A small of amount of passivation material near the inlet/outlet is unlikely to cause issues, but deposi-

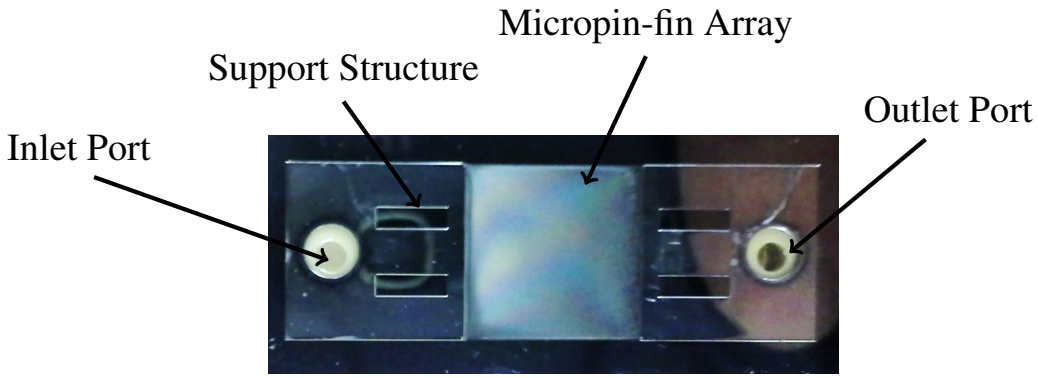


Figure A.3: First generation micropin-fin test device

tion on the micropin-fins deeper in the channel could affect heat transfer. To prevent this, later generations of devices used a standard Bosch process to etch the fluid ports to a depth of approximately 20 µm above the cavity. This was followed by reactive ion etching (RIE) without the passivation step to penetrate into the cavity without depositing passivation material inside the cavity. Lastly, the wafer is diced and the die is packaged for testing.

Micropin-fin test Device Design

First generation devices

A top-view image through the glass cap of a first generation test chip can be seen in Figure A.3. Thermal test results using these devices can be found in [51]. These test devices initially had a glass cap which was epoxied on to the silicon heat sink. The inlet and outlet plena each contained two mechanical supports. The fluid was delivered through NanoPorts which were attached onto the backside of the chip with epoxy and wires were soldered onto copper pads to deliver power to the heaters/RTDs. This first generation of devices could not withstand the high pressures necessary for two phase refrigerant cooling, or single phase cooling with very high flow rates. Additionally, wires soldered directly to the copper bonding pads often caused pads to delaminate when wires were moved.

Second generation devices

Images of a second generation micropin-fin heat sink test device design can be seen in Figure A.4. These devices were used for a number of two phase cooling experiments, including experiments with water below atmospheric pressure[64]. The goals of this design were to increase the number of test chips per 4 inch wafer, increase reliability, ensure uniform flow, and allow accurate measurement of pressure drop across the 1 cm $\times$ 1 cm micropin-fin array under test. These goals were accomplished through a number of changes.

First, the design footprint was reduced and a package was used for the chips to deliver coolant and power. The die shrink allowed 14 devices to be fabricated on a single wafer. After fabrication, the chips were mechanically attached to a PCB. Power was delivered from the PCB to the chip through wirebonds, which enabled substantial reduction in the size of the bonding pads as well as increased reliability. While soldered wires often pulled off copper traces on the silicon, the rigid PCB prevented this from happening. Additionally, a package was used to deliver fluid with O-rings sealing the inlet, outlet, and pressure ports of the device. This further increased reliability as the epoxy seals were one of the primary points for leaks in the first generation of devices. An anodically bonded cap further reduced leaks and increased maximum operating pressures.

Pressure measurements with the first generation of devices, which had only an inlet and outlet port, could only provide pressure measurements which included the pressure drop across the NanoPorts, and inlet/outlet plena. The second generation of test devices received two pressure ports to accurately measure pressure drop across the micropin-fin array. These pressure measurements function analogously to four point electrical resistance measurements. In steady state conditions, there is no fluid flow through the pressure ports, so the pressures measured at the ends of the ports are equal to the pressures at the inlets of the pressure measurement channels, which are placed precisely at the beginning and end of the micropin-fin array.

For mechanical reliability, oval supports were added around the inlet and outlet ports,

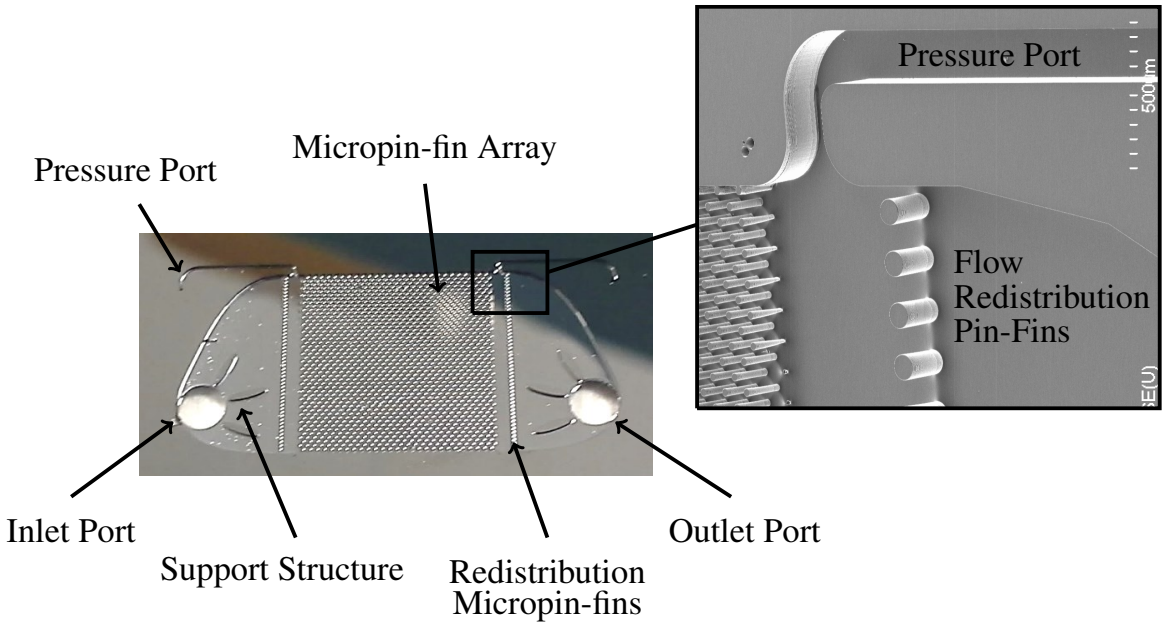


Figure A.4: Second generation micropin-fin test device

where pressure from the O-ring could otherwise cause cracking. Additional rows of micropin-fin at the inlet and outlet, but outside of the pressure ports, were added to help promote uniform flow through the micropin-fin array.

Third generation devices

An image of a third generation micropin-fin test device can be seen in Figure A.5. Passivation material from the Bosch process can be seen around the inlet and outlet ports of the device because the no-passivation RIE step was skipped for the particular device shown. This design was used for several experiments using R245fa as a coolant[65, 66]. A slight modification of this design was also used for experimentation with heterogeneous micropin-fin arrays, as shown in chapter 3.

The third generation of devices have a few changes and innovations beyond the second generation of devices, primarily for experimentation with high pressure two phase refrigerants. First, a number of support pillars were added to the inlet and outlet plena to minimize the length of any unsupported area. The maximum pressure the device can withstand is

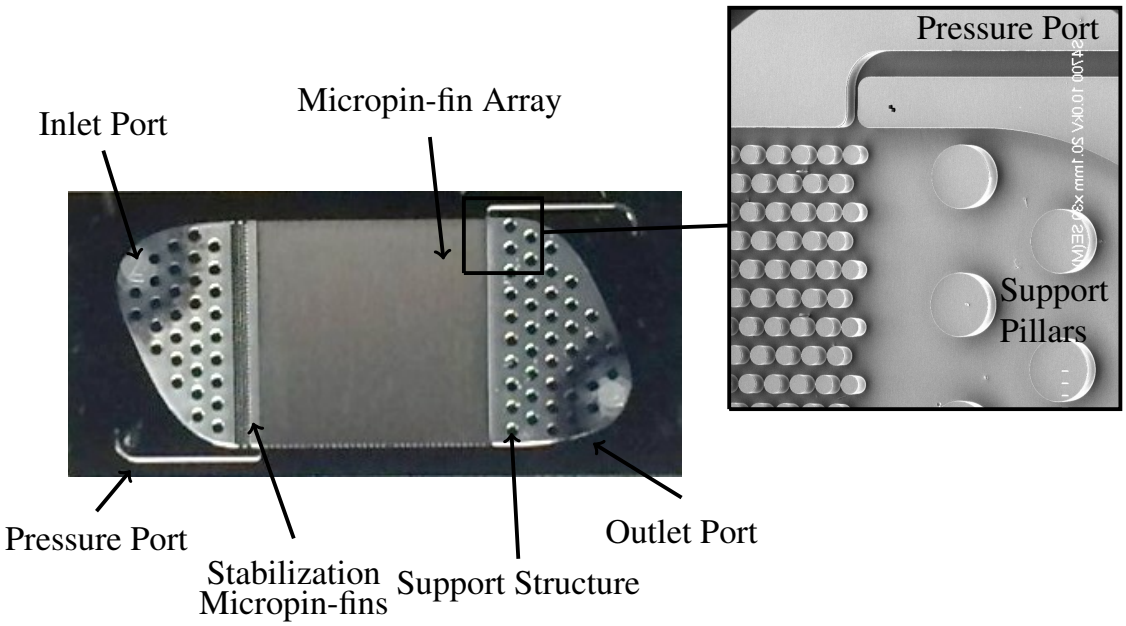


Figure A.5: Third generation micropin-fin test device

related to maximum distance between support structures[46]. Second, the flow stabilization micropin-fin density was increased substantially and they were moved only to the inlet of the device. The purpose of this column of micropin-fins was to stabilize oscillations which can occur in two phase cooling, where vapor may be pushed back into the inlet of the device[24]. Lastly, the inlet and outlet ports were moved to opposite sides of the device to further promote uniform flow.

REFERENCES

[1] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted mosfet's with very small physical dimensions," *IEEE Journal of Solid-State Circuits*, vol. 9, no. 5, pp. 256–268, 1974.

[2] M. Horowitz, "1.1 Computing's energy problem (and what we can do about it)," in *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, Feb. 2014, pp. 10–14.

[3] A. Danowitz, K. Kelley, J. Mao, J. P. Stevenson, and M. Horowitz, "CPU DB: Recording Microprocessor History," *Queue*, vol. 10, no. 4, 10:10–10:27, Apr. 2012.

[4] M. B. Taylor, "Is dark silicon useful? Harnessing the four horsemen of the coming dark silicon apocalypse," in *DAC Design Automation Conference 2012*, Jun. 2012, pp. 1131–1136.

[5] G. Venkatesh, J. Sampson, N. Goulding, S. Garcia, V. Bryksin, J. Lugo-Martinez, S. Swanson, and M. B. Taylor, "Conservation Cores: Reducing the Energy of Mature Computations," in *Proceedings of the Fifteenth Edition of ASPLOS on Architectural Support for Programming Languages and Operating Systems*, ser. ASPLOS XV, New York, NY, USA: ACM, 2010, pp. 205–218, ISBN: 978-1-60558-839-1.

[6] A. Pedram, S. Richardson, S. Galal, S. Kvatinsky, and M. A. Horowitz, "Dark Memory and Accelerator-Rich System Optimization in the Dark Silicon Era," *IEEE Design & Test*, vol. 34, no. 2, pp. 39–50, Apr. 2017, arXiv: 1602.04183.

[7] P. Franzon, E. Rotenberg, J. Tuck, W. R. Davis, H. Zhou, J. Schabel, Z. Zhang, J. B. Dwiel, E. Forbes, J. Huh, and S. Lipa, "Computing in 3d," in *2015 IEEE Custom Integrated Circuits Conference (CICC)*, Sep. 2015, pp. 1–6.

[8] Google Inc. (May 2018). Effective machine learning using cloud tpus (google I/O '18), (visited on 05/15/2018).

[9] Nvidia, *Nvidia dgx station, Personal ai supercomputer*, Jul. 2018.

[10] AMD. (Sep. 2018). Radeon™vega frontier edition (liquid-cooled), (visited on 09/01/2018).

[11] B. Banijamali, S. Ramalingam, H. Liu, and M. Kim, "Outstanding and innovative reliability study of 3d TSV interposer and fine pitch solder micro-bumps," in *2012 IEEE 62nd Electronic Components and Technology Conference*, May 2012, pp. 309–314.

[12] P. Quinn, "Fpga based silicon innovation exploiting "more than moore" technology," in *Ph.D. Research in Microelectronics and Electronics (PRIME), 2013 9th Conference on*, 2013, pp. 11–12.

[13] J. Danskin and D. Foley, "Pascal gpu with nvlink," in *2016 IEEE Hot Chips 28 Symposium (HCS)*, 2016, pp. 1–24.

[14] NEC, *Sx-aurora tsubasa*, SX-Aurora, Oct. 2017.

[15] AMD, *Amd epyc 7000 series processors*, EPYC, Sep. 2018.

[16] Nvidia. (2018). Nvidia titan xp, (visited on 05/14/2018).

[17] Nvidia, *Nvidia tesla v100 gpu accelerator*, Mar. 2018.

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

[19] R. Prasher, J Dirner, J. Chang, A Myers, D Chau, D. He, and S Prstic, "Nusselt number and friction factor of staggered arrays of low aspect ratio micropin-fins under cross flow for water as fluid," *ASME Trans. Journal of Heat Transfer*, vol. 129, no. 2, pp. 141 –153, 2007.

[20] M. Koz and A. Kosar, "Parameter optimization of a micro heat sink with circular pin-fins," in *ASME 2010 8th Int. Conf. Nanochannels, Microchannels, and Minichannels*, Montreal, Canada, 2010, pp. 531–539.

[21] J. Tullius, T. Tullius, and Y. Bayazitoglu, "Optimization of short micro pin fins in minichannels," *International Journal of Heat and Mass Transfer*, vol. 55, no. 15 - 16, pp. 3921 –3932, 2012.

[22] T Brunschwiler, B Michel, H Rothuizen, U Kloter, B Wunderle, H Oppermann, and H Reichl, "Interlayer cooling potential in vertically integrated packages," *Microsystem Technologies*, vol. 15, no. 1, pp. 57 –74, 2008.

[23] A. Kosar and Y. Peles, "Micro scale pin fin heat sinks: parametric performance evaluation study," *IEEE Trans. Components and Packaging Technologies*, vol. 30, no. 4, pp. 855 –865, 2007.

[24] C. E. Green, P. A. Kottke, X. Han, D. C. Woodrum, T. E. Sarvey, P. Asrar, X. Zhang, Y. K. Joshi, A. G. Fedorov, S. K. Sitaraman, and M. S. Bakir, "A review of two-phase forced cooling in three-dimensional stacked electronics: Technology integration," *J. Electron. Packag.*, vol. 137, no. 4, p. 040 802, 2015.

[25] X. Han, A. Fedorov, and Y. Joshi, "Flow boiling in microgaps for thermal management of high heat flux microsystems," *J. Electron. Packag.*, vol. 138, no. 4, pp. 040 801–12, 2016.

[26] A. Shakouri and Y. Zhang, "On-chip solid-state cooling for integrated circuits using thin-film microrefrigerators," *IEEE Trans. Compon, and Packag. Technol.*, vol. 28, no. 1, pp. 65–69, 2005.

[27] I. Chowdhury, R. Prasher, K. Lofgreen, G. Chrysler, S. Narasimhan, R. Mahajan, D. Koester, R. Alley, and R. Venkatasubramanian, "On-chip cooling by superlattice-based thin-film thermoelectrics.," *Nature Nanotechnology*, vol. 4, no. 4, pp. 235 – 238, 2009.

[28] C. E. Green, A. G. Fedorov, and Y. K. Joshi, "Fluid-to-fluid spot-to-spreader (f2/s2) hybrid heat sink for integrated chip-level and hot spot-level thermal management," *J. Electron. Packag.*, vol. 131, p. 025 002, 2009.

[29] D. Nikolic, M. Hutchison, P. T. Sapin, and A. J. Robinson, "Hot spot targeting with a liquid impinging jet array waterblock," in *Thermal Investigations of ICs and Systems, THERMINIC 2009*, 2009, pp. 168–173.

[30] Y. Han, B. L. Lau, and X. Zhang, "Package-level microjet-based hotspot cooling solution for microelectronic devices," *IEEE Electron Device Letters*, vol. 36, no. 5, pp. 502–504, 2015.

[31] S. Narayanan, A. G. Fedorov, and Y. K. Joshi, "On-chip thermal management of hotspots using a perspiration nanopatch," *J. Micromechanics and Microengineering*, vol. 20, no. 7, p. 075 010, 2010.

[32] Z. Yan, G. Liu, J. M. Khan, and A. A. Balandin, "Graphene quilts for thermal management of high-power gan transistors," *Nature Communications*, vol. 3, p. 827, May 2012.

[33] T. Brunschwiler, S. Paredes, U. Drechsler, B. Michel, W. Cesar, Y. Leblebici, B. Wunderle, and H. Reichl, "Heat-removal performance scaling of interlayer cooled chip stacks," in *Proc. 12th IEEE Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, 2010, pp. 1–12.

[34] Y. Zhang, L. Zheng, and M. Bakir, "3-d stacked tier-specific microfluidic cooling for heterogeneous 3-d ics," *IEEE Trans. Components, Packaging and Manufacturing Technology*, vol. 3, no. 11, pp. 1811–1819, 2013.

[35] M. S. Bakir, C. King, D. Sekar, H. Thacker, B. Dang, G. Huang, A. Naeemi, and J. D. Meindl, "3d heterogeneous integrated systems: Liquid cooling, power delivery,

and implementation," in *2008 IEEE Custom Integrated Circuits Conference*, 2008, pp. 663–670.

[36] Y. Zhang, A. Dembla, and M. Bakir, "Silicon micropin-fin heat sink with integrated tsvs for 3-d ics: tradeoff analysis and experimental testing," *IEEE Trans. Components, Packaging and Manufacturing Technology*, vol. 3, no. 11, pp. 1842–1850, 2013.

[37] 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.

[38] H. Oh, G. S. May, and M. S. Bakir, "Analysis of signal propagation through tsvs within distilled water for liquid-cooled microsystems," *IEEE Transactions on Electron Devices*, vol. 63, no. 3, pp. 1176–1181, 2016.

[39] L. Zheng, Y. Zhang, G. Huang, and M. S. Bakir, "Novel electrical and fluidic microbumps for silicon interposer and 3-d ics," *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 4, no. 5, pp. 777–785, 2014.

[40] M. H. Nasr, C. E. Green, P. A. Kottke, X. Zhang, T. E. Sarvey, Y. K. Joshi, M. S. Bakir, and A. G. Fedorov, "Extreme-microgap (x-gap) based hotspot thermal management with refrigerant flow boiling," in *2016 15th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, 2016, pp. 1466–1476.

[41] M. H. Nasr, C. E. Green, P. A. Kottke, X. Zhang, T. E. Sarvey, Y. K. Joshi, M. S. Bakir, and A. G. Fedorov, "Flow regimes and convective heat transfer of refrigerant flow boiling in ultra-small clearance microgaps," *International Journal of Heat and Mass Transfer*, vol. 108, pp. 1702 –1713, 2017.

[42] M. H. Nasr, C. E. Green, P. A. Kottke, X. Zhang, T. E. Sarvey, Y. K. Joshi, M. S. Bakir, and A. G. Fedorov, "Hotspot thermal management with flow boiling of refrigerant in ultrasmall microgaps," *Journal of Electronic Packaging*, vol. 139, pp. 011 006–8, 2017.

[43] P. S. Lee, J. C. Ho, and H. Xue, "Experimental study on laminar heat transfer in microchannel heat sink," in *ITherm 2002. Eighth Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (Cat. No.02CH37258)*, 2002, pp. 379–386.

[44] W. M. W. M. Kays, *Convective heat and mass transfer*, 3rd ed., ser. McGraw-Hill series in mechanical engineering. New York: McGraw-Hill, 1993, ISBN: 0070337217.

[45] E. Lemmon, M. McLinden, and D. Friend, *Thermophysical Properties of Fluid Systems*, ser. NIST Chemistry WebBook, NIST Standard Reference Database Number 69. Gaithersburg MD, 20899: National Institute of Standards and Technology, Dec. 2013.

[46] D. C. Woodrum, T. Sarvey, M. S. Bakir, and S. K. Sitaraman, "Reliability study of micro-pin fin array for on-chip cooling," in *Proc. 65th IEEE Electronic Components and Technology Conf. (ECTC)*, 2015, pp. 2283–2287.

[47] A. Renfer, M. K. Tiwari, T. Brunschwiler, B. Michel, and D. Poulikakos, "Experimental investigation into vortex structure and pressure drop across microcavities in 3d integrated electronics," *Experiments in Fluids*, vol. 51, no. 3, pp. 731–741, 2011.

[48] T. E. Sarvey, Y. Hu, C. E. Green, P. A. Kottke, D. C. Woodrum, Y. K. Joshi, A. G. Fedorov, S. K. Sitaraman, and M. S. Bakir, "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.

[49] M. H. Nasr, C. E. Green, P. A. Kottke, X. Zhang, T. E. Sarvey, Y. K. Joshi, M. S. Bakir, and A. G. Fedorov, "Hotspot thermal management with flow boiling of refrigerant in ultra-small microgaps," *J. Electron. Packag*, vol. 139, no. 1, p. 011 006, 2017.

[50] C. E. Green, P. A. Kottke, T. E. Sarvey, A. G. Fedorov, Y. Joshi, and M. S. Bakir, "Performance and integration implications of addressing localized hotspots through two approaches: clustering of micro pin-fins and dedicated microgap coolers," in *Proc. ASME Int. Tech. Conf. and Exhibition on Packaging and Integration of Electronic and Photonic Microsystems collocated with ASME 13th Int. Conf. on Nanochannels, Microchannels, and Minichannels*, 2015.

[51] T. E. Sarvey, Y. Zhang, Y. Zhang, H. Oh, and M. S. Bakir, "Thermal and electrical effects of staggered micropin-fin dimensions for cooling of 3d microsystems," in *Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2014 IEEE Intersociety Conference on*, 2014, pp. 205–212.

[52] T. Yamane, N. Nagai, S.-i. Katayama, and M. Todoki, "Measurement of thermal conductivity of silicon dioxide thin films using a 3ω method," *J. Applied Physics*, vol. 91, no. 12, pp. 9772–9776, 2002.

[53] M. B. Kleiner, S. A. Kuhn, and W. Weber, "Thermal conductivity measurements of thin silicon dioxide films in integrated circuits," *IEEE Trans. Electron Devices*, vol. 43, no. 9, pp. 1602–1609, 1996.

[54] 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," *Int. J. Heat and Mass Transfer*, vol. 55, 63846390, 2012.

[55] T. E. Sarvey, Y. Zhang, L. Zheng, P. Thadesar, R. Gutala, C. Cheung, A. Rahman, and M. S. Bakir, "Embedded cooling technologies for densely integrated electronic systems," in *Proc. IEEE Custom Integrated Circuits Conf. (CICC)*, 2015.

[56] J. Karttunen, J. Kiihamaki, and S. Franssila, "Loading effects in deep silicon etching," in *Proc. SPIE*, vol. 4174, Santa Clara, CA, 2000, pp. 90–97.

[57] I. W. Rangelow, "Critical tasks in high aspect ratio silicon dry etching for microelectromechanical systems," *Journal of Vacuum Science & Technology A: Vacuum, Surfaces, and Films*, vol. 21, no. 4, pp. 1550–1562, 2003.

[58] S. C. Lin and K. Banerjee, "Cool chips: Opportunities and implications for power and thermal management," *IEEE Transactions on Electron Devices*, vol. 55, no. 1, pp. 245–255, 2008.

[59] 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.

[60] Y. Liu, R. P. Dick, L. Shang, and H. Yang, "Accurate temperature-dependent integrated circuit leakage power estimation is easy," in *2007 Design, Automation Test in Europe Conference Exhibition*, 2007, pp. 1–6.

[61] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer cmos circuits," *Proceedings of the IEEE*, vol. 91, no. 2, pp. 305–327, 2003.

[62] E. Lemmon, M. McLinden, and D. Friend, *Thermophysical Properties of Fluid Systems*, ser. NIST Chemistry WebBook, NIST Standard Reference Database Number 69. Gaithersburg MD, 20899: National Institute of Standards and Technology.

[63] D. Lorenzini, C. Green, T. E. Sarvey, X. Zhang, Y. Hu, A. G. Fedorov, M. S. Bakir, and Y. Joshi, "Embedded single phase microfluidic thermal management for non-uniform heating and hotspots using microgaps with variable pin fin clustering," *Int. J. Heat and Mass Transfer*, vol. 103, pp. 1359 –1370, 2016.

[64] X. Zhang, X. Han, T. E. Sarvey, C. E. Green, P. A. Kottke, A. G. Fedorov, Y. Joshi, and M. S. Bakir, "Three-dimensional integrated circuit with embedded microfluidic cooling: Technology, thermal performance, and electrical implications," *Journal of Electronic Packaging*, vol. 138, no. 1, pp. 010 910–9, 2016.

[65] P. Asrar, X. Zhang, C. E. Green, P. A. Kottke, T. E. Sarvey, A. Fedorov, M. S. Bakir, and Y. K. Joshi, "Flow visualization of two phase flow of r245fa in a microgap with integrated staggered pin fins," in *2016 32nd Thermal Measurement, Modeling Management Symposium (SEMI-THERM)*, 2016, pp. 86–89.

[66] P. Asrar, X. Zhang, C. D. Woodrum, C. E. Green, P. A. Kottke, T. E. Sarvey, S. Sitaraman, A. Fedorov, M. Bakir, and Y. K. Joshi, "Flow boiling of r245fa in a microgap with integrated staggered pin fins," in *2016 15th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, 2016, pp. 1007–1012.