HYBRID MICROFLUIDIC COOLING AND THERMAL ISOLATION TECHNOLOGIES FOR 3D ICS

A Thesis
Presented to
The Academic Faculty

by

Yue Zhang

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

Georgia Institute of Technology
May 2015

Copyright © 2015 by Yue Zhang

HYBRID MICROFLUIDIC COOLING AND THERMAL ISOLATION TECHNOLOGIES FOR 3D ICS

Approved by:

Professor Muhannad Bakir, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Oliver Brand
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Sudhakar Yalamanchili
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Hua Wang
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Yogendra Joshi
School of Mechanical Engineering
Georgia Institute of Technology

Date Approved: March 27, 2015

DEDICATION

To my husband,

Ming Yi, my son, Oliver Yi,

and my parents,

Shuqiang Zhang and Dengwen Zhang,

for their unlimited love and support.

ACKNOWLEDGEMENTS

The completion of this thesis would not be possible without the inspiration, support, guidance and help from many individuals. I would like to express my sincere gratitude toward them.

First of all, I would like to give special thanks to my advisor Dr. Muhannad Bakir for accepting me as a student in the Integrated 3D Systems Group and guiding me through my entire Ph.D. career. He has always been a great source of inspiration. His vision and advices have greatly impacted the direction of my research. All the accomplishments in my research could not be possible without his continuous support. I would like to thank him for his guidance that led to my successful moments. Even more, I would forever cherish his encouragement during tough times in my research. He has taught me how to become a good researcher. More importantly, he has taught me how to become a better person.

I also want to express my sincere gratitude to my Ph.D. dissertation committee members, Professor Oliver Brand, Professor Sudhakar Yalamanchili, Professor Hua Wang, and Professor Yogendra Joshi for giving me feedback and suggestions to improve my thesis.

I am very fortunate to have had the opportunity to work with the past and current Integrated 3D Systems Group members. I sincerely thank Dr. King and Mr. Zaveri for training me in the cleanroom and in the lab when I first joined the group. I would like to convey my special gratitude to Thomas Sarvey, Chaoqi Zhang, Yang Zhang, and Ashish Dembla for the close collaboration. My thanks are also extended to the other group members who are willing to share their knowledge with me: Hyung Suk Yang, Paragkumar Thadesar, Hanju Oh, Muneeb Zia, Xuchen Zhang, Prabha

Viswanathan, and Reza Abbaspour. I would like also to acknowledge the members from other groups for collaboration: Steven Isaacs, Yoon Jo Kim from Professor Yogendra Joshi's group , and Song Hu from Professor Hua Wang's group.

My experimental work in the cleanroom could not be possible without the Institute for Electronics and Nanotechnology (IEN) members. I would like to especially acknowledge Dr. Brand, Gary Spinner, John Pham, Vinny Nguyen, and Charlie Suh for answering my questions and solving my problems with the tools.

I would like to thank my parents for their endless love and support. It is their advice that made me pursue the Ph.D. degree in Georgia Tech. Also, none of this would happen without their financial support. I wouldn't have a chance to meet the most important person in my life, my husband, Ming Yi. Ming is not only my life partner, but he is also an excellent researcher who teaches me research skills such as how to create excellent schematics. Last but not least, I would like to thank my son Oliver for bringing so much happiness in my life.

TABLE OF CONTENTS

DEDICATION . . . . **iii**

ACKNOWLEDGEMENTS . . . . **iv**

LIST OF TABLES . . . . **x**

LIST OF FIGURES . . . . **xi**

SUMMARY . . . . **xviii**

I BACKGROUND AND INTRODUCTION OF THERMAL CHALLENGES IN 3D ICS **1**

1.1 Current 3D ICs Trend 1

1.2 Thermal Challenges in Stacks Containing Multiple High- Power Chips 4

1.3 Thermal Challenge in Heterogeneous 3D Stacks 8

1.4 Organization of the Thesis 14

II THERMAL-ELECTRICAL CO-ANALYSIS OF A TSV–COMPATIBLE MICROFLUIDIC HEAT SINK **16**

2.1 Introduction 16

2.2 Literature Review of Microfluidic Heat Sink 17

2.3 Heat Transfer Theory for Micropin-fin Heat Sink 20

2.4 Electrical Parasitics of TSVs Embedded in microfluidic heat sink . . 22

2.5 Thermal Resistance and Pressure Drop Trade-off Analysis of Microfluidic Heat Sink 24

2.6 Electrical-Thermal Trade-off Analysis of TSVs in Microfluidic Heat Sink 25

2.7 Alternative Methods to Reduce TSV Capacitance 29

2.7.1 Novel Liner Material 29

2.7.2 Heterogeneous TSV Integration 32

2.8 Conclusion 33

III TSV-COMPATIBLE MICROPIN-FIN HEAT SINK EXPERIMENTS **35**

3.1 Introduction 35
3.2 Fabrication of the TSV–Compatible Micropin-Fin Heat Sink 36
3.2.1 Bonding Process Selection 36
3.2.2 Fabrication Process of MPFHS 39
3.2.3 Integration of TSVs in MPFHS 44
3.3 Test Setup Description and Automated Data Collection in LabVIEW 48
3.4 Single Layer Thermal Measurements and Benchmarked with Air-Cooled Heat Sink 52
3.5 Data Extrapolations and Analysis 54
3.6 Conclusion 60
IV TIER-SPECIFIC MICROFLUIDIC COOLING EVALUATION IN 3D IC STACKS 61
4.1 Introduction 61
4.2 Thermal Testbed Preparation and Experimental Setup 63
4.3 Tier-Specific Microfluidic Cooling for Different Stacking Scenarios . 65
4.3.1 Processor-on-Processor and Memory-on-Processor Stack . . . 65
4.3.2 Tier-Specific Flow Rates in ICs with Different Power Dissipations 68
4.4 Microfluidic Cooling in Multi-core Processor Stacking 71
4.4.1 Preparation of the Thermal Testbed and Experimental Test Setup 71
4.4.2 Lateral Thermal Gradient 72
4.4.3 Electrical Implications Due to Lateral Thermal Gradient . . 77
4.4.4 Localized Coolant Delivery Method to Mitigate Lateral Thermal Gradient 82
4.4.5 Microfluidic Cooling Under Nonuniform Power Dissipation . 83
4.4.6 Vertical Thermal Coupling 85
4.5 Validation through ANSYS Simulations 86
4.5.1 Initial Simulation of Repeatable Cell Rows Assuming Even Flow Distribution 87
4.5.2 Adjusted Simulation of Repeatable Cell Rows 89

4.6 Conclusion 91

V THERMAL ISOLATION FOR HETEROGENEOUS 3D ICS . . 92

5.1 Introduction 92

5.2 Resistance Network Modeling 94

5.2.1 TSVs' Impact on Thermal Isolation 97

5.2.2 MFI Thermal Resistance 98

5.3 Finite Difference Modeling of the Proposed Stack with Thermal Isolation Technology 101

5.3.1 Thermal Bridge 103

5.3.2 Uniformly Distributed TSVs vs. Clustered TSVs 106

5.4 Design of the Testbed 109

5.5 Testbed Fabrication and Test Setup 110

5.5.1 Testbed fabrication 110

5.5.2 Assembly 113

5.5.3 Thermal and Electrical Test Setup 116

5.6 Thermal and Electrical Experimental Results 118

5.6.1 Thermal Testing I: Powering the high-power tier 120

5.6.2 Thermal Testing II: Minimize the Hotspot Coupling 124

5.6.3 Thermal Testing III: Bottom Tier Power Increases 126

5.6.4 Electrical Testing of MFIs 129

5.7 Validation by Finite Difference Modeling 130

5.8 Conclusion 134

VI SUMMARY AND FUTURE WORK 137

6.1 Summary of the Presented Work 137

6.1.1 Advanced Microfluidic Cooling Solution for 3D ICs Containing High Power Chips 137

6.1.2 Advanced Thermal Isolation Technology for Heterogeneous 3D ICs 139

6.2 Future Work 140

6.2.1 Explore a System with Interposer Cooling 140
6.2.2 Advancing the Thermal Isolation Technology 141
6.2.3 System Performance Implications 143
REFERENCES **145**

LIST OF TABLES

1 Selected optimal heat sink dimensions from the literature 27

2 Comparison of the area occupied by TSVs for homogeneous and heterogeneous TSV integration 33

3 Comparison of the measured and modeled normalized thermal resistance at a power density of 100 W/cm^2 55

4 Summary of the inlet water temperature (T_{in}), the outlet water temperature (T_{out}), and the chip junction temperature (T_j) at 40 W/cm^2 for different water flow rates 55

5 Electrical implications due to lateral thermal gradient 81

6 Summary of the simulation results vs. experimental results 88

7 ANSYS simulated thermal resistance of a single MFI with various designs 99

8 Summary of the temperature of the top and bottom tiers under different scenarios (Part I) 120

9 Summary of the temperature of the top and bottom tiers under different scenarios (Part II) 121

10 Parameters used in the finite difference model 130

11 Boundary conditions assumed in the finite difference model 131

12 Comparison between the measured and modeled junction temperatures 131

LIST OF FIGURES

1 Schematic of a typical 3D IC with two stacks. The homogeneous stack contains two processor tiers. The heterogeneous stack contains several memory stacks on top of a processor tier. 2

2 Examples of current commercialized 3D IC products. 3

3 Proposed 3D IC prototype solving thermal challenges of lack of effective cooling and lack of effective thermal isolation. 4

4 ITRS projections for the number of dice in a stack, number of TSVs, die thickness, and power of a single high-performance chip. 5

5 Illustration of conventional air cooling technology. 7

6 Illustration of prior integrated microfluidic cooling technology. 7

7 Illustration of the tier-specific microfluidic cooling technology in the present work. 7

8 Solder-based microfluidic chip I/Os and electric microbumps. 9

9 (a) Illustration of a 3D stacked memory module containing SDRAM, photonic transceivers, and associated driver circuitry. (b) Temperature impact on microring resonance frequency. 10

10 (a) DRAM on logic stack in [24]. (b) Temperature contour of DRAM and logic when logic has a uniform power dissipation [24]. 11

11 (a) TSV guard ring to reduce thermal coupling [28] and (b) SEM for a two-channel (de)multiplexer with an air cavity beneath to reduce the thermal coupling [29]. 12

12 Prototype shows the proposed thermal isolation technology that replaces microbumps and underfill with air gap and thermally degraded MFIs. 13

13 Selected single-phase microfluidic heat sink geometries in the literature. 18

14 TSV array integrated in a silicon micropin-fin 22

15 Impact of H_{fin} on TSV density and diameter. 23

16 Thermal resistance and pressure drop for different micropin-fin designs with flow rate 150 ml/min to 60 ml/min. 26

17 Thermal resistance and TSV capacitance as a function of microfluidic heat sink height at different pressure drop values. 28

18 The impact of microfluidic heat sink height on the number of TSVs and TSV capacitance. 28

19 TSV dielectric capacitance as a function of liner: oxide, Parylene-C, and air. 30

20 Homogeneous and heterogeneous approaches for TSV integration into microfluidic cooled chip. 31

21 Two-tier 3D IC stack with microfluidic heat sink and TSVs. 34

22 Schematic of a three-microprocessor chip stack each with interlayer microfluidic cooling. A 3D stack of memory chips resides above the microprocessors. High AR TSVs are integrated in the MPFHS. 36

23 Theory of anodic bonding. 38

24 Theory of fusion bonding. 39

25 Si–Si fusion bonding process. 40

26 Initial process flow of the MPFHS. 41

27 Modified process flow of the MPFHS. 43

28 SEMs of (a) the top and (b) cross-sectional view of the micropin-fin arrays. 44

29 (a) The tilted view of micropin-fins and (b) an infrared image of the bonded sample. 45

30 Overview of the MPFHS with a magnified angled view of the micropin-fins. Fluid flows from left to right. 46

31 Process flow of high aspect ratio TSV integration into the MPFHS . . 46

32 SEMs of (a) high-AR TSVs integrated in micropin-fins and (b) free standing high-AR TSVs. 47

33 Optical images of (a) the free standing TSVs after removing the Silicon and (b) a cross section of high-AR TSVs integrated in micropin-fins. 47

34 Characterization of a RTD's resistance as a function of temperature. . 48

35 The experimental test setup for single-layer microfluidic heat sink testing. 49

36 A photo of the test setup for microfluidic heat sink testing. The key instruments in the test setup include power analyzer, gear pumps, data logger, and a LabVIEW interface for data collection. 51

37 Photo of the microfluidic testbed. 52

38 Illustration of the resistance measurement of TSVs using four-point technique. The platinum pads are deposited selectively using focused ion beam (FIB) deposition. 53

39 Average junction temperature under air cooling and microfluidic cooling compared with ITRS projections. 54

40 Convective thermal resistance and heat transfer coefficient as a function of the flow rate. 58

41 Nusselt number and pressure drop as a function of Reynolds number. 59

42 Micropin-fin layout and dimensions (top and tilted view). 59

43 Prototype of (a) a general embedded microfluidic heat sink and (b) our tier-specific microfluidic cooling within a 3D stack. 62

44 SEM of solder microfluidic chip I/Os and electric microbumps. 63

45 Experimental setup for microfluidic heat sink evaluation in 3D stacks. 64

46 Junction temperature rise in a memory–processor stack under microfluidic heat sink and ACHS. 66

47 Junction temperature rise in a processor–processor stack under microfluidic heat sink and ACHS. 67

48 Junction temperature of the top layer (P1) and the bottom layer (P2) as a function of the flow rates. 69

49 Performance of multi-core processor compared with single-core processor [61]. 71

50 Intel Core i7 Processor. 72

51 (a) Image of the bonded two-tier thermal testbed and (b) layout of the four heaters. 73

52 Schematic of the tier-specific fluidic delivery mechanism. 73

53 Junction temperature rise at different heater locations on the chip for different power dissipations. ANSYS simulation for 100 W case is also plotted for reference. 74

54 Base temperature map in ANSYS simulation while the chip dissipates 100 W/cm^2. 75

55 Water temperature map in ANSYS simulation while the chip dissipates 100 W/cm^2. 76

56 Increase in leakage power as a function of chip temperature for a Intel 15 mm die with 100 nm technology [64]. 78

57 Normalized leakage current as a function of temperature. 80

58 A diagraph of Intel's Knights Landing CPU, consisting of up to 72 x86 cores for exascale supercomputing. 80

59 A diagraph illustration of the assumed 100-core CPU. 81

60 (a) Prototype of 3D stack with microfluidic chip I/Os for localized coolant delivery and (b) solder based microfluidic chip I/Os and electric microbumps. 82

61 (a) A polymer pipe and (b) a polymer socket for fluidic delivery. . . . 83

62 Evaluation of microfluidic cooling in chips with nonuniform power dissipation: (a) case 1 where heater 1 and 2 dissipate 100 W/cm^2 and heater 3 and 4 are off and (b) case 2 where heater 3 and 4 dissipate 100 W/cm^2 and heater 1 and 2 are off. 83

63 Junction temperature rise of heater 1 to 4 under the two different test cases shown in Figure 62 84

64 Vertical thermal coupling test cases. (Case A) Heaters 1 and 4 in upper tier are powered. (Case B) Heaters 1 and 4 in upper tier and heaters 2 and 3 in lower tier are powered. (Case C) Heaters 1 and 4 in upper tier and heaters 2 and 3 in lower tier are powered with DI water pumped into both tiers. 85

65 The junction temperature increase of the upper tier at different heater locations on the chip for the three cases. 86

66 The junction temperature increase of the lower tier at different heater locations on the chip for the three cases. 87

67 Created cell row in ANSYS to simulate MPFHS in a single-layer chip. 88

68 The temperature and pressure contour profiles after initial simulation 89

69 The created cell row in ANSYS to simulate MPFHS sink in a single-layer chip. 90

70 The temperature and pressure contour profiles from the adjusted simulation. 91

71 Prototype shows the proposed thermal isolation technology that replaces microbumps and underfill with air gap and thermally degraded MFIs. 94

72 (a) A 3D stack of processor and silicon nanophotonic chips with hybrid thermal management: within-tier microfluidic cooling in processor and air/vacuum cavity to thermally isolate the silicon nanophotonic chip. (b) The corresponding thermal resistance network. 95

73 The junction temperature increase of the upper tier at different heater locations on the chip for the three cases. T1 and T2 denote the temperature of the high-power and low-power tiers, respectively. 96

74 Illustration of the cross-sectional (left) and top (right) view of the structures simulated in ANSYS to represent TSVs through air cavity. 97

75 Temperature of both tiers in the simulated structure shown in Figure 74. T1 and T2 denote the temperature of the high-power and low-power tiers, respectively. 98

76 (a) The MFI structure created in ANSYS and (b) the corresponding thermal profile in a static thermal simulation. 99

77 Temperature of the high-power and low-power die with different interconnects: (a) Uniform MFIs within air gap and (b) microbumps and underfill. 101

78 Finite difference scheme: (a) general points inside the stack and (b) boundary points in the face of the stack [70]. 102

79 Thermal bridge on top of a memory tier simulated in ANSYS. 105

80 Memory tier temperature map for the calculation of the thermal resistance of the thermal bridge. 105

81 Power maps of the memory and processor tiers used in the finite difference modeling. 106

82 Power maps of the memory and processor tiers in (a) the clustered TSV case and (b) the uniform TSV case. 107

83 Schematic of the designed testbed for evaluation of the proposed thermal isolation technologies. 108

84 Layout (left) and schematic (right) of the power map designs of the top tier (low-power tier). 108

85 Layout (left) and schematic (right) of the power map designs of the bottom tier (high-power tier). 109

86 Layout of (a) the MFI array and (b) the connections for daisy-chain resistance and four-point resistance measurements. 110

87 Process flow for the low-power tier. 111

88 Images of (a) parts of the RTD array and (b) the pad of the RTD. . 113

89 Images of (a) the MFIs electroplated on top of the polymer dome, (b) the free standing MFIs after dome removal, (c) the MFI array with gold passivation, and (d) a single MFI with gold passiavtion. 114

90 Process flow for the high-power tier. 115

91 Optical image of one hotspot on the high-power tier. 115

92 Images of (a) the bottom tier after dicing and (b) the assembled two-tier testbed. 116

93 (a) Flip chip bonding assembly and (b) the alignment between the two tiers. 117

94 X-ray of (a) overall view of the boned chip and (b) a magnified view. 117

95 Microfluidic test setup to evaluate the thermal isolation technologies. 118

96 (a) Top and (b) bottom view of the stack assembled to a PCB board using wire bonding. 119

97 Four-point resistance measurement of MFI 119

98 (a) Initial case when the high-power tier dissipates 0 W and (b) Case E in Table 9 where the background power density is 30 W/cm^2 and the hotspot power density is 100 W/cm^2. 122

99 Junction temperature fluctuation before and after the high-power tier is powered. 123

100 (a) Uniform power density of 10 W/cm^2 in the bottom tier (Case A) and (b) background power of 10 W/cm^2 plus two hotspots each dissipates 150 W/cm^2 (Case C). 124

101 Junction temperature fluctuation of top and bottom tiers in Case H and Case I in Table 9. 125

102 (a) Zero background power with two hotspots each dissipates 150 W/cm^2 (Case H) and (b) zero background power with two hotspots each dissipates 200 W/cm^2 (Case I). 126

103 Junction temperature fluctuation of the top and bottom tiers in Case H and Case I in Table 9. 127

104 (a) Background power of 30 W/cm^2 plus two hotspots each dissipates 100 W/cm^2 (Case E) and (b) background power of 30 W/cm^2 plus two hotspots each dissipates 150 W/cm^2 (Case F). 128

105 Junction temperature fluctuation of the top and bottom tiers in Case E and Case F in Table 9 129

106 Junction temperature in Case E and Case F (as listed in Table 9 using the finite-difference model. This figure can be compared with the measured results shown in Figure 105. 132

107 The modeled heterogeneous stack with (a) MFI and air cavity and (b) microbumps and underfill. 133

108 Junction temperature in both tiers with and without the thermal isolation. In the case without thermal isolation, microbumps and underfill are integrated between the tiers. 134

109 A heterogeneous 3D stack with MFIs and independent microfluidic heat sink for the low-power die. 135

110 Benchmark the ideal thermal isolation technology with conventional 3D stacking approach. 135

111 Illustration of a 3D stack with a microfluidic cooled interposer. 141

112 Illustration of multi-optimized microfluidic heat sinks. 142

113 (a) MFI with a thickness of 2 μm and (b) MFI with a thickness of 4.5 μm. 143

SUMMARY

A key challenge for three dimensional (3D) integrated circuits (ICs) is thermal management. There are two main thermal challenges in typical 3D ICs. First, in the homogeneous integration with multiple high-power tiers, a cooling solution that scales with the number of dice in the stack is needed. Second, in the heterogeneous integration, a thermal isolation solution is needed to 'protect' the low-power tier from the high-power tier. This research focuses to address these two thermal challenges through hybrid microfluidic cooling and thermal isolation technologies.

Within-tier microfluidic cooling is proposed and demonstrated to cool a stack with multiple high-power tiers. Electrical thermal co-analysis is performed to understand the trade-offs between through silicon via (TSV) parasitics and heat sink performance. A TSV–compatible micropin–fin heat sink is designed, fabricated and thermally characterized in a single tier and benchmarked with a conventional air-cooled heat sink. The designed heat sink has a thermal resistance of 0.269 K·cm^2/W at a flow rate of 70 mL/min. High aspect ratios TSVs (18:1) are integrated in the micropin–fins. Within-tier microfluidic cooling is then implemented in 3D stacks to emulate different heating scenarios, such as memory-on-processor and processor-on-processor. Air gap and mechanically flexible interconnects (MFIs) are proposed for the first time to decrease the vertical thermal coupling between high-power (e.g. processor) and low-power tiers (e.g. memory or nanophotonics). A two-tier testbed with the proposed thermal isolation technology is designed, fabricated and tested. Compared with conventional 3D integration approach, thermal isolation technology helps reduce the temperature at a fixed location in the low-tier by 12.9 °C. The resistance of a single MFI is measured to be 46.49 mΩ.

CHAPTER I

BACKGROUND AND INTRODUCTION OF THERMAL CHALLENGES IN 3D ICS

1.1 Current 3D ICs Trend

With the continued aggressive scaling of transistors, interconnect performance and power dissipation have become limiting factors for high-performance integrated circuits. This is true for both on-chip and off-chip interconnects. In the latter, the inability to provide high density off-chip wires with low latency, low energy-per-bit, and large bandwidth density has greatly exacerbated the memory wall problem for multi-core processors. This is critical because off-chip bandwidth between multiprocessors and DRAM impacts system performance. To overcome this interconnect limit, three-dimensional integrated circuits (3D ICs) technology has been pursued in recent years, as it represents a promising solution to the interconnect problem by significantly shortening the interconnect length as well as enabling heterogeneous integration of logic, memory, microelectromechanical systems (MEMS), and optoelectronics.

A typical 3D IC with homogeneous and heterogeneous stacks is shown in Figure 1. In the homogeneous 3D integration, two processor tiers are stacked as an example. In the heterogeneous integration, several memory tiers are stacked on top of a processor tier. Both stacks are bonded to a silicon interposer through microbumps and can communicate with each other through fine-pitch interposer-level interconnects.

A key challenge for such high-power 3D applications is thermal management. There are two main thermal challenges in a typical 3D IC system. First, in the homogeneous integration where more than one high-power tier is integrated, an effective cooling method that can scale with the number of tiers is needed. This is needed

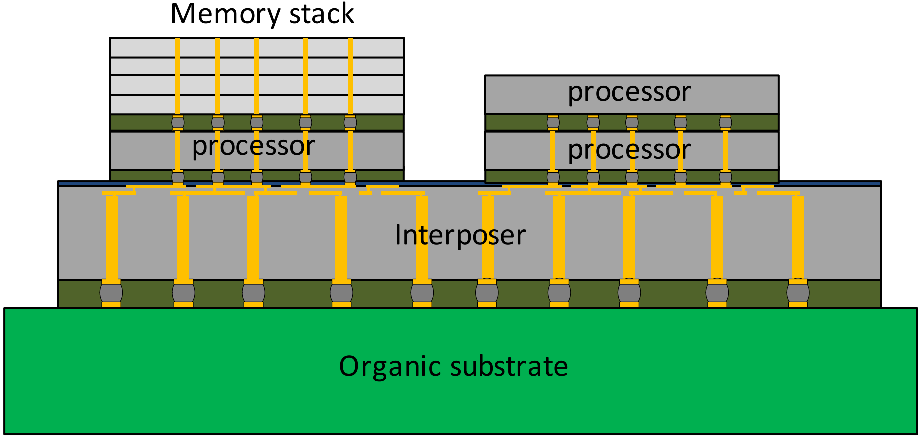
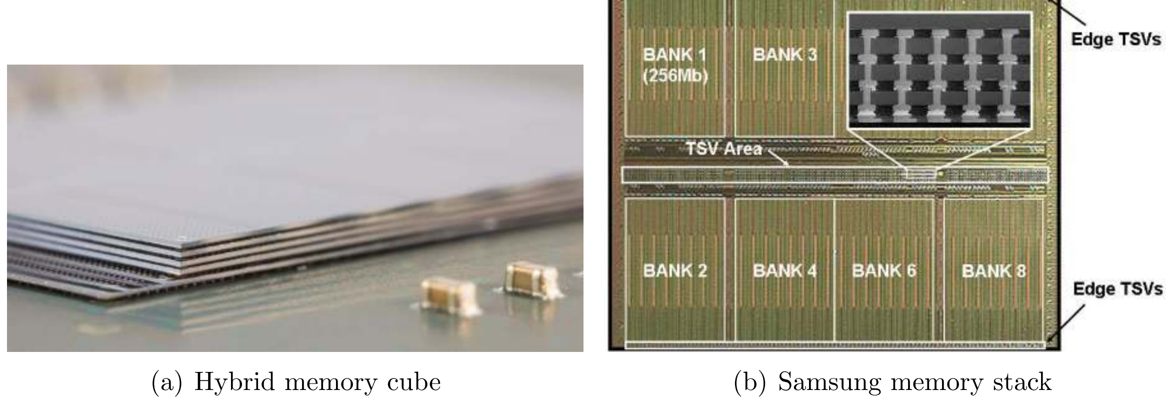


Figure 1: Schematic of a typical 3D IC with two stacks. The homogeneous stack contains two processor tiers. The heterogeneous stack contains several memory stacks on top of a processor tier.

because the power density in 3D ICs and the thermal resistance of the dice within the stack increases as the number of tiers increases. The latter is due to the fact that the 'inner' dice do not have direct access to the top-attached heat sink. The second thermal challenge in the system shown in Figure 1 is the thermal coupling effect. This challenge is especially significant in heterogeneous integration where high-power tiers are stacked with low-power tiers, such as a memory stack on a processor stack, a processor on a silicon photonics stack, and MEMS on a processor stack. In 3D ICs, the thermal coupling between the vertical tiers is critical because 3D ICs bring dice closer than in conventional ICs. Without an effective thermal isolation between the tiers, the thermal coupling will cause the low-power tier to follow the same temperature trend as that of the high-power tier and degrade its performance.

The performance and power consumption of high performance unit systems depends on the junction temperature. Currently, commercialized 3D IC products do not contain high power dice due to lack of effective cooling and thermal isolation technology. Figures 2(a) and 2(b) show examples of Micron Technology's hybrid memory cube (HMC) with four DRAM tiers stacked on top of a logic tier [1] and a stack of



(a) Hybrid memory cube (b) Samsung memory stack

Figure 2: Examples of current commercialized 3D IC products.

Samsung's dynamic random-access memories (DRAMs) [2], respectively. Note the central processing unit (CPU) is not included in the stack in either case because of thermal challenges. The objective of this research is to

- Demonstrate an effective cooling solution that scales with the number of dice.
- Demonstrate an effective thermal isolation solution that 'protects' the low-power tiers from the high-power tiers to enable heterogeneous 3D stacking.

A schematic illustrating our proposed prototype that solves the two thermal challenges is shown in Figure 3. For the homogeneous stack, embedded microfluidic cooling is adopted. Each high-power tier has its own microfluidic heat sink. Since these heat sinks are microscale, they can be integrated into any high-power tier as needed. The detailed integration method for this idea is discussed in Section 1.2. In the heterogeneous stack, the proposed solution includes using an air cavity and mechanically flexible interconnects between the heterogeneous elements. The details of the proposed thermal isolation concept is discussed in Section 1.3.

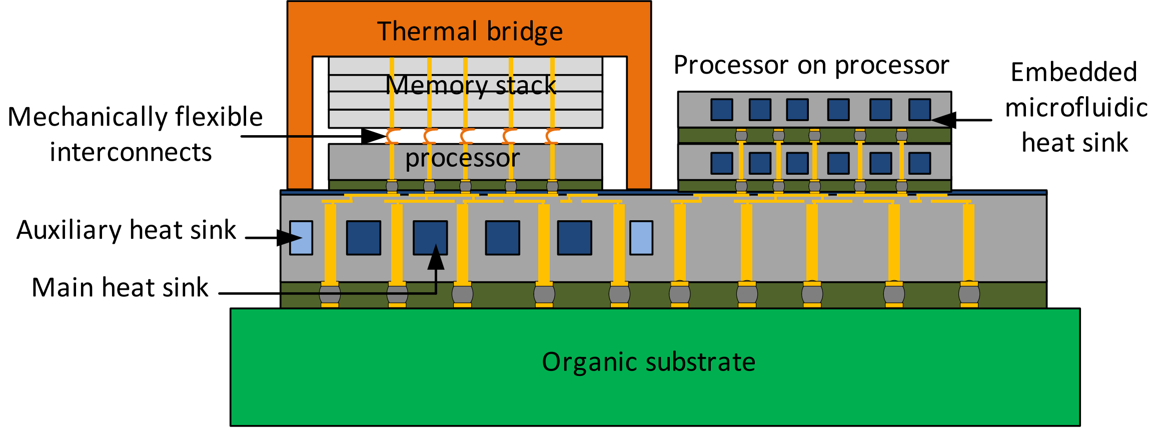


Figure 3: Proposed 3D IC prototype solving thermal challenges of lack of effective cooling and lack of effective thermal isolation.

1.2 Thermal Challenges in Stacks Containing Multiple High-Power Chips

Three dimensional integration is considered to be a promising solution to continue Moore's Law in the vertical direction, and offer possibilities of increased device density, shorter interconnects, smaller foot print, and heterogeneous integration. Although the concept has been under research for several years since it was first introduced in the 1980s, the products in the market do not yet include high-power units in the stack. The thermal challenge is one of the biggest hurdles.

In applications where multiple high-power chips (e.g. processors) are stacked together, the thermal challenge is mainly due to the increased power density and thermal resistance of the 'inner' dice as the number of tiers increases. In these cases, conventional air cooling is not sufficient for the high power stack. In [3], dividing a traditional microprocessor design into two dice to form a logic+logic stack, the peak temperature increases by 14 °C while the total power remains the same. The problem is exacerbated when the power density and number of tiers increases. According to ITRS projections, each high-power unit will dissipate 130 W by 2020 and each stack may have six tiers (possibly a combination of high-power tiers and low-power tiers) [4].

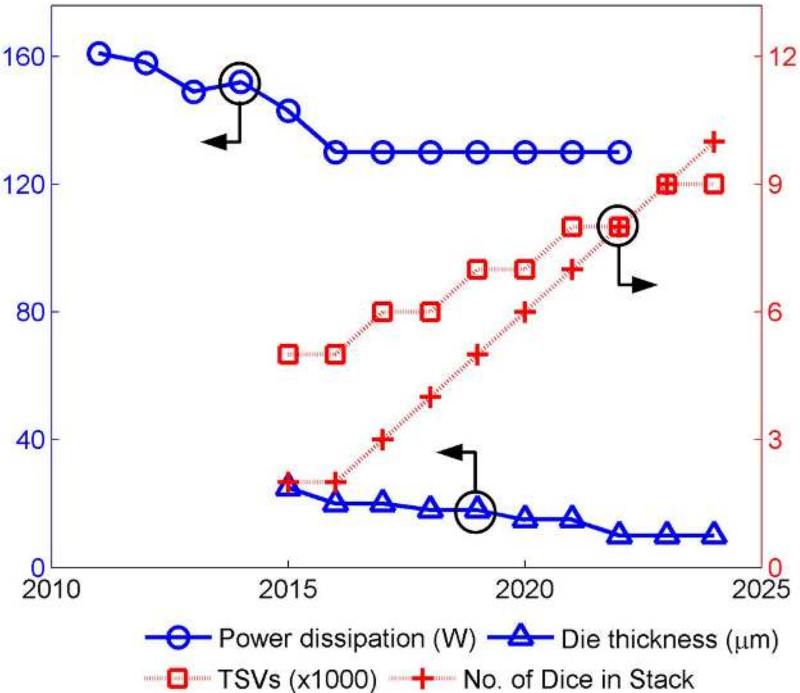


Figure 4: ITRS projections for the number of dice in a stack, number of TSVs, die thickness, and power of a single high-performance chip.

Since air cooling has limited cooling capabilities [5], researchers have shown the possibility of using embedded within-tier microfluidic cooling for 3D ICs [6]–[10]. The advantages of interlayer microfluidic cooling compared to air cooling are as follows:

1. Microfluidic cooling has higher cooling capability since water has much higher heat capacity than air.

2. The fabrication of the interlayer microfluidic heat sink is compatible with current CMOS microfabrication technology.

3. Since the microfluidic heat sink is chip-scale, it has vertical stackability. The air-cooled heat sink (ACHS), on the other hand, can not scale with the number of tiers.

Heat removal (790 W) by a silicon microfluidic heat sink was first demonstrated by Tuckerman and Pease in 1981 [6]. Because of its relative ease of fabrication, single-phase cooling and two-phase cooling using the microchannel heat sink have been widely investigated [9], [11], [12]. Our research mainly focuses on single-phase microfluidic cooling. As microfabrication technology has advanced, more complex microfluidic heat sink designs have become possible, bringing the possibility of outperforming the microchannel heat sink [13]. One method to enhance single-phase cooling utilizes the fabrication of obstructions in the flow direction [8], [13]. In [8], multiple heat sink designs were analyzed and compared including microchannels, inline micropin-fins, and staggered micropin-fins. A staggered micropin-fin heat sink (MPFHS) is demonstrated to have the lowest thermal resistance at a constant flow rate [8].

Figure 6 depicts a typical chip configuration with embedded microfluidic heat sink in the literature, in which the fluid is supplied through a single inlet from the top of the stack [14], [15]. The authors of [14] and [15] have demonstrated the cooling of a four-tier and a two-tier stack with total power dissipation of 390 W and 200 W, respectively. With this approach, it is not possible to control or tailor the flow rate in each tier. However, in a realistic 3D stack, especially in a heterogeneous stack, the power dissipation in each tier may be different (workload dependent). Thus, one needs the capability to control the coolant flow rate in each tier independently. To address this need, wafer-level batch fabricated solder microfluidic chip I/Os and fine pitch electrical microbump I/Os have been recently demonstrated, as shown in Figure 8 [16]. Based on this innovative chip I/O technology, we propose and experimentally demonstrate tier-specific microfluidic cooling where the flow rate in each tier is chosen based on the power dissipation of each tier (Figure 7).

The height of most reported interlayer microfluidic heat sinks ranges from 200 μm to 400 μm. Because of the insertion of these microfluidic heat sinks, a wafer

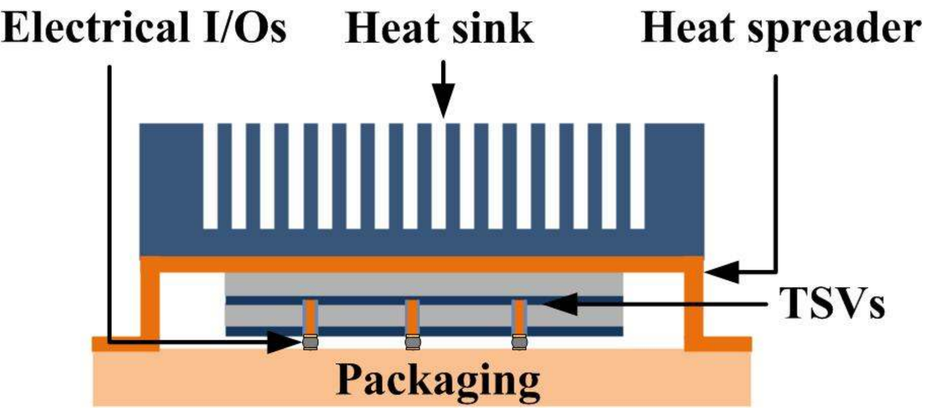


Figure 5: Illustration of conventional air cooling technology.

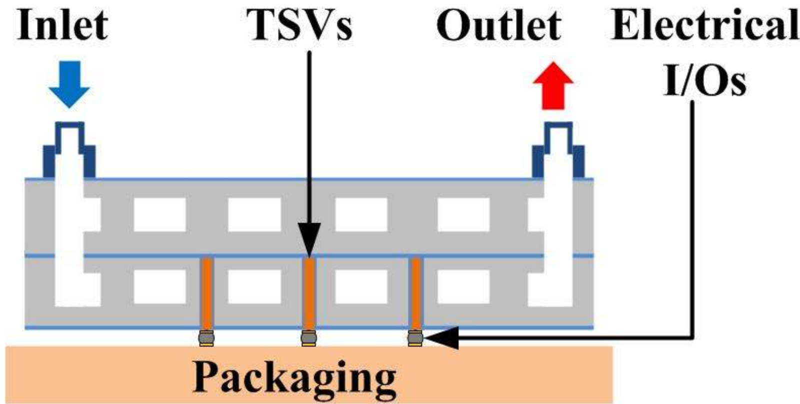


Figure 6: Illustration of prior integrated microfluidic cooling technology.

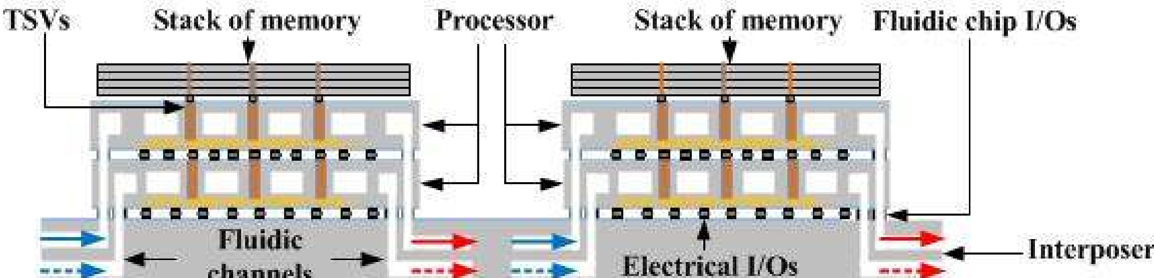


Figure 7: Illustration of the tier-specific microfluidic cooling technology in the present work.

will have a typical thickness of a few hundred microns, which presents challenges to through silicon via (TSV) fabrication and electrical performance. Therefore, designing a microfluidic heat sink without considering TSV fabrication compatibility and TSV parasitics greatly diminishes the advantages of 3D ICs. There are very few studies focusing on designing a microfluidic heat sink while accounting for the impact on TSVs. In [7], microchannel dimensions were designed to maximize the TSV density while meeting the thermal constrains. In [17], a microchannel infrastructure with microchannel bends was designed to bypass the region that contains TSVs. However, to date, there has been neither an attempt to optimize the heat sink design while accounting TSV fabrication compatibility nor an attempt to analyze the corresponding impact of the microfluidic heat sink on the electrical performance of TSVs. In this work, electrical and thermal co-analysis of trade-offs between TSV parasitics and heat sink performance was done and the results are shown in Chapter II.

For TSVs in CMOS, where wafer thickness is typically less than 100 μm, aspect ratios as high as 15:1 have been demonstrated [18], [19]. For thick silicon wafers (greater than 100 μm), TSVs with aspect ratios greater than 10:1 have been shown for the application of silicon interposer [20]. TSVs with higher aspect ratio need to be developed for thicker silicon (with embedded microfluidic heat sinks). Moreover, experimental integration of fine-pitch and high aspect ratio TSVs within microfluidic heat sinks is missing from the literature. In Chapter III, integration of high aspect-ratio TSVs with microfluidic cooling will be shown.

1.3 Thermal Challenge in Heterogeneous 3D Stacks

Thermal coupling has always been an issue for ICs. Intra-socket thermal coupling has been investigated between CPU and DIMMs that are in the same socket [21]. Temperature increment was observed in memory when the workload of the CPU increases. Increased power dissipation in the CPU causes the heat sink temperature

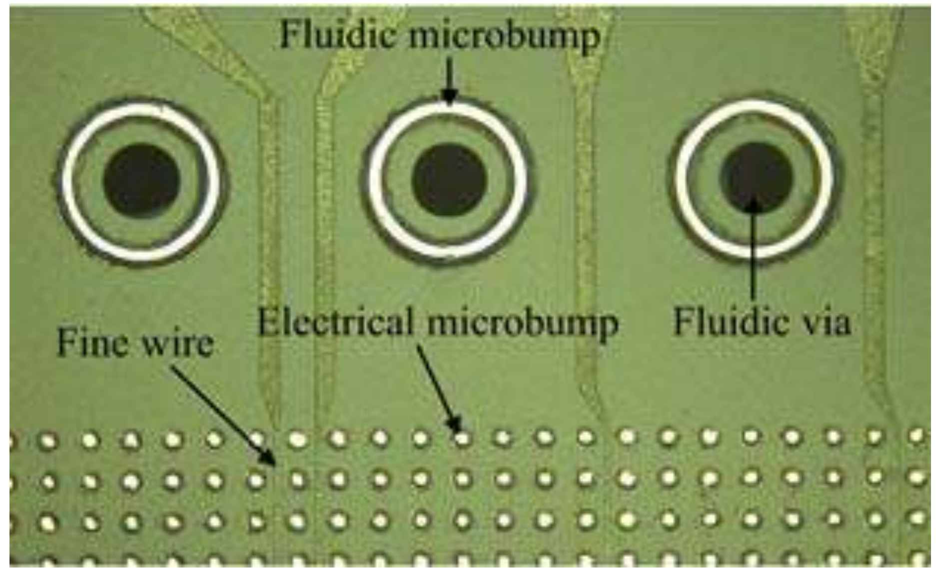
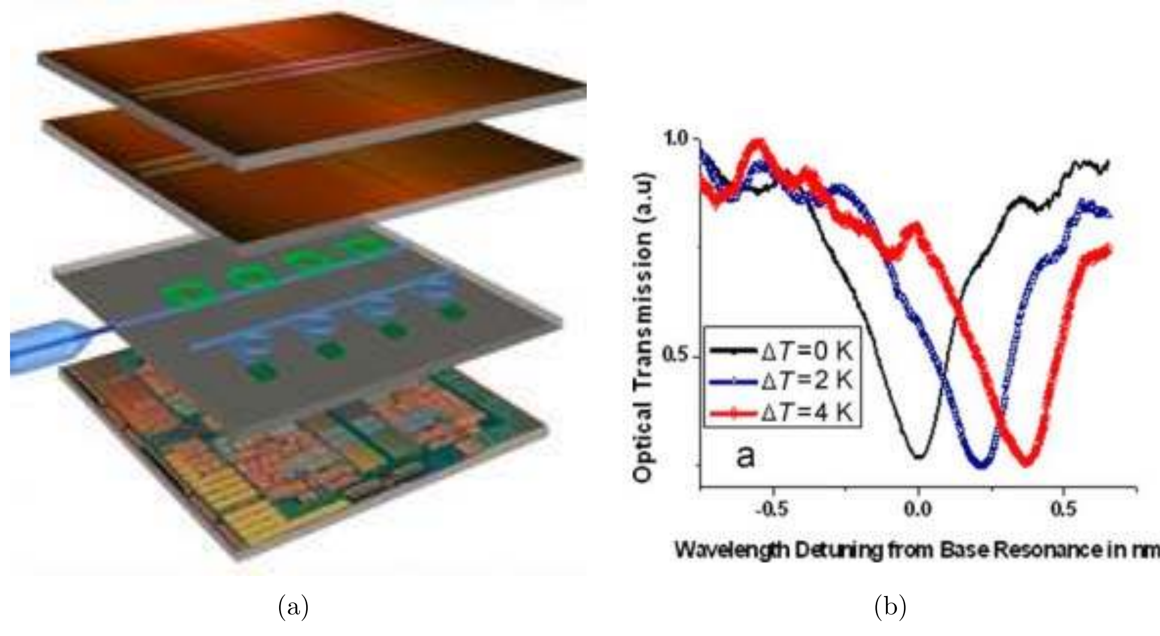


Figure 8: Solder-based microfluidic chip I/Os and electric microbumps.

to increase and, thus, causes the downstream memory to become warmer. Within-chip thermal coupling between CPU and GPU has been studied in [22]. An AMD Trinity APU consisting of two CPU cores and one GPU core was used to study the thermal coupling. The moment the two CPU cores were allocated with more power, the GPU core power also increases because of the thermal coupling and the impact of temperature on leakage power. In 3D ICs applications, not only the above mentioned lateral thermal coupling need to be investigated, but the vertical die-to-die level thermal coupling is also critical because 3D ICs bring dice closer than conventional ICs [23], [24]. In applications where high-power chips (e.g. processors) are stacked along with low-power and temperature-sensitive components (memory or silicon nanophotonic chips, for example), thermal management will not only require effective cooling, but may also require effective thermal isolation to 'protect' the temperature-sensitive components from the high and time-varying power dissipation of other chips in the stack. By placing such tiers next to each other, the thermal coupling between them will be significant, leading to possibly undesirable junction-temperature variation in the temperature-sensitive tier as a result of the high-power



(a) (b)

Figure 9: (a) Illustration of a 3D stacked memory module containing SDRAM, photonic transceivers, and associated driver circuitry. (b) Temperature impact on microring resonance frequency.

chips.

In the case of silicon nanophotonics, stacking such chips in a stack adjacent to logic and memory has been explored [25], [26]. Figure 9(a) shows an example of a 3D stacked memory module with optical interconnections [25]. However, the temperature sensitivity of the optical elements presents significant challenges for integration; for example, a microring modulator with 5 μm diameter is reported to have a wavelength drift of 0.11 nm/°C in [27]. Figure 9(a) shows the transmission spectra with varying ambient temperature over 4 °C. A temperature change of 13.5 °C will result in a complete passband mismatch between transmitter–receiver pairs in 64-channel wavelength-division multiplexing (WDM) [26].

In applications involving a memory die, it has been shown that stacking logic on SRAM causes a 30 to 40 °C temperature increase in the SRAM die [23]. The increased temperature not only causes the leakage power to increase by approximately two times, but also causes the average cache access time to increase by 50 ps (28%

performance degradation) [23]. A stack of DRAM-on-logic is investigated (Figure 10(a)) in [24]. When logic tier dissipates a uniform power, the temperature difference between the DRAM and logic is very small, as shown in Figure 10(b), which indicates strong thermal coupling between the two tiers. As such, there is a need for wafer-level batch-fabricated thermal isolation technologies in order to minimize thermal coupling between the high-power logic chip and the low-power and temperature-sensitive chips in the stack.

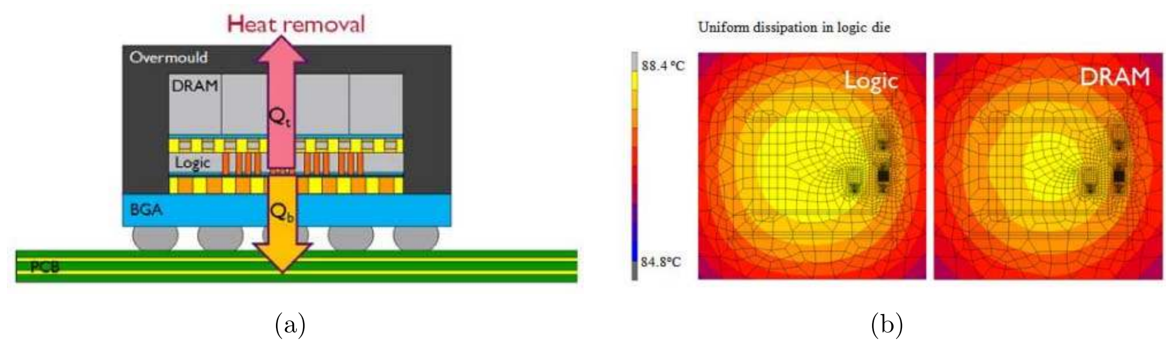
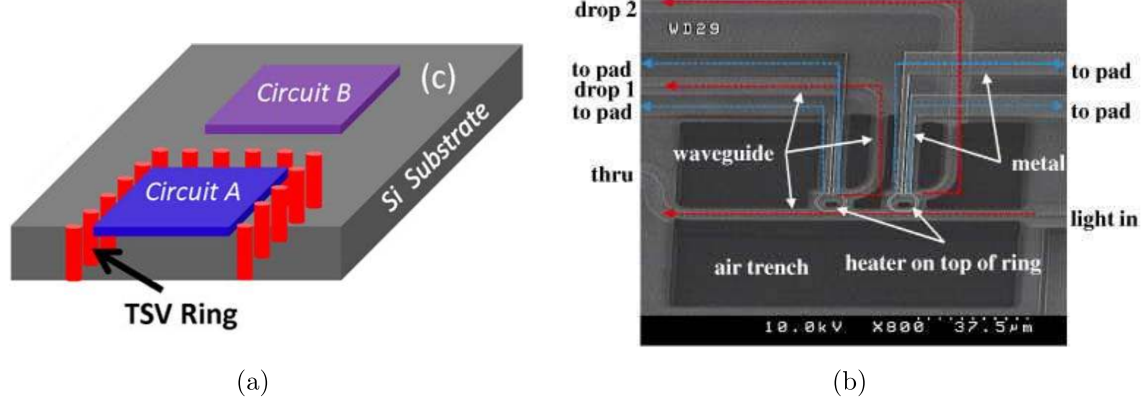


Figure 10: (a) DRAM on logic stack in [24]. (b) Temperature contour of DRAM and logic when logic has a uniform power dissipation [24].

There is very little effort to investigate thermal isolation in 3D ICs. Researchers have proposed to use a set of TSV guard rings to thermally isolate to some level two circuits side by side (Figure 11(a)) [28]. In their simulation, the bottom of the Si substrate is set to a constant temperature of 25 °C. Because of the high thermal conductivity of the TSVs, a large portion of the heat generated from circuit B tends to flow downwards to the bottom through TSVs rather than flows to the circuit A and, thus, creates thermal isolation between the two circuits. Circuit A has a ring oscillator, whose resonant frequency will shift because of the influence of circuit B. It is claimed that the TSV-based guard ring can alleviate the thermal coupling so that the resonant frequency shift is reduced by 65%. Because of the temperature sensitivity of silicon phonics devices, a local heater is often used to create a constant



(a) (b)

Figure 11: (a) TSV guard ring to reduce thermal coupling [28] and (b) SEM for a two-channel (de)multiplexer with an air cavity beneath to reduce the thermal coupling [29].

temperature environment in a local region. Extra tuning power is needed from the heater if the generated heat spreads to the adjacent area because of thermal coupling. Researchers from Oracle [29] have demonstrated decreased thermal decoupling with an air trench right below the microring resonators and heaters (Figure 11(b)). The tuning power is reduced from 27 mW to 21 mW. However, the thermal isolation between tiers in 3D ICs is missing from the literature.

In the current technology, tiers in 3D ICs are bonded through microbumps. Two tiers in a stack may expand differently due to different junction temperature, and may cause stress on the microbumps which leads to cracking in thermal cycles. To address this issue, underfill is applied between the two tiers to alleviate the stress on the solder microbumps, as shown in Figure 1. However, the thermal conductivity of underfill is usually around 0.4 W/mK– 1.3 W/mK. This will introduce a small thermal resistance between the two tiers and cause thermal coupling between the tiers. Thus, we propose to integrate an air gap and thermally degraded mechanically flexible interconnects (MFIs) to replace both microbumps and underfill. The proposed prototype is shown in Figure 12. When the two tiers expand differently because of different junction temperatures, stress will occur on MFIs. Unlike rigid solder

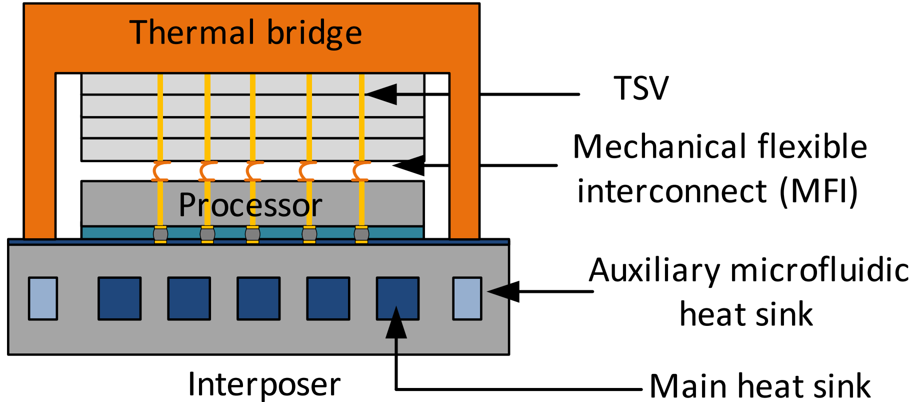


Figure 12: Prototype shows the proposed thermal isolation technology that replaces microbumps and underfill with air gap and thermally degraded MFIs.

microbumps, MFIs will deform elastically under stress, which helps maintain the electrical connectivity between tiers. Thanks to this phenomenon, MFIs can help get rid of the underfill and thus reduce the thermal coupling between tiers. The thermally degraded MFIs are a type of MFI that are specially designed to have large thermal resistance and small electrical parasitics.

Once the low-power dice are thermally isolated from the high-power die, it is also isolated from the interposer-level cooling path. Therefore, a new cooling path is provided to the memory dice from the top. A thermal bridge is attached to the top of the memory and thermally interconnects the memory dice to the auxiliary microfluidic heat sink in the interposer. The thermal bridge can be made of metals of high thermal conductivity such as copper. Thermal interface material is applied between the thermal bridge and memories. Also to be noted, the interposer has multi-optimized microfluidic heat sinks. The main heat sink is for cooling the processor, and the auxiliary heat sink is dedicated to the memory stack. The dark blue color represents the heat sink dedicated for the processor, and the light blue color represents the auxiliary heat sink. The two heat sinks have separate flow paths and may have different flow rates and designs depending on their power loads. For example, the

auxiliary heat sink for the memory stack may be less dense and may have a smaller flow rate. The proposed concept is demonstrated with a two-tier thermal testbed. The thermal isolation technology with MFIs is implemented. The results will be shown in Chapter V.

1.4 Organization of the Thesis

In this thesis, a hybrid thermal management solution is proposed to overcome the above mentioned challenges. The key contributions of this work include:

1. A silicon micropin-fin heat sink (MPFHS) is designed to be TSV–compatible. Electrical-thermal trade-off analysis is performed to study the impact of heat sink design on cooling capability and electrical parasitics of TSVs. The selected design provides a thermal resistance of 0.2 K$\cdot$cm^2/W, a TSV dielectric capacitance of 0.4 pF, and a TSV density of 4×10^4 TSVs/cm^2.

2. The designed TSV-compatible MPFHS is thermally characterized in a single tier and benchmarked with a conventional air-cooled heat sink. High aspect ratio TSVs with a diameter of 10 μm and a height of 178 μm (18:1) are integrated in the micropin-fins.

3. Within-tier microfluidic cooling is then implemented in 3D stacks to emulate different heating scenarios, such as memory-on-processor, processor-on-processor with the same power load, and processor-on-processor with different power loads. The microfluidic heat sink maintains the stack temperature below 50 °C for a total power density of 200 W/cm^2 in a two-tier processor-on-processor stack.

4. A tier-specific cooling mechanism is proposed and implemented in a two-tier stack where the flow rate in each tier is chosen based on the power dissipation and temperature target for each tier. With tier-specific cooling, two tiers with different power levels can have the same junction temperature, mitigating the

thermal mechanical stress between the tiers. Pumping power reduction of 37.5% can be achieved by preventing over-cooling.

5. Within-tier microfluidic cooling is also implemented in a multi-core stack. The lateral and vertical thermal coupling are analyzed. The vertical thermal coupling is minimal when each tier contains its own microfluidic heat sink. The leakage power is analyzed with the presence of lateral thermal gradient.

6. A new heterogeneous architecture is proposed for the first time featuring thermal isolation technology using an air gap, thermally degraded MFIs, and novel cooling structures. The architecture is experimentally demonstrated with a two-tier testbed, and benchmarked with conventional 3D stacking approach.

The thesis is organized as follows. Chapter II discusses the electrical–thermal co-analysis of the microfluidic heat sink and TSV parasitics. Chapter III presents the co-integration of the TSV–compatible microfluidic heat sink with 18:1 aspect ratio TSVs. Thermal and electrical testing are also included in Chapter III. In Chapter IV, the microfluidic cooling is evaluated in 3D stacks. Scenarios including processor-on-processor, memory-on-processor, and processors with different power loads are emulated. A tier-specific cooling mechanism is proposed and implemented to minimize the vertical thermal coupling within the stack. Chapter V discusses the modeling and experimental implementation of the thermal isolation technology based on air cavity and MFIs.

CHAPTER II

THERMAL-ELECTRICAL CO-ANALYSIS OF A TSV–COMPATIBLE MICROFLUIDIC HEAT SINK

2.1 Introduction

Transistor scaling along with continued material innovation in ICs has propelled the semiconductor industry during the past 50 years in terms of improvements in IC performance, power dissipation, and cost [30]. However, with the continued aggressive scaling, interconnect performance and power dissipation have become limiting factors for higher-performance integrated circuits [31]. This is true for both on-chip and off-chip interconnects [32]. In the latter, the inability to provide high density off-chip wires with low latency, low energy-per-bit, and large bandwidth density has greatly exacerbated the memory wall problem for multicore processors. This is critical because off-chip bandwidth between multiprocessors and DRAM impacts system performance [33]. To overcome this interconnect limit, 3D IC technology has been pursued in recent years, as it represents a promising solution to the interconnect problem by significantly shortening the interconnect length as well as enabling heterogeneous integration of logic, memory, MEMS, and optoelectronics [34], [35].

TSVs are the key enabling technology for 3D ICs as they provide inter-layer communication and power delivery between stacked chips. Shorter TSVs (in thinner dice) have lower capacitance and, thus, lower latency and energy-per-bit. Therefore, thinner dice are preferred in 3D ICs. According to the International Technology Roadmap for Semiconductors (ITRS), die thickness is expected to decrease from 25 μm to 10 μm by 2024 [6] (Figure 4). A key challenge for high-power 3D applications is cooling. The reason is that both the power density in 3D ICs and the thermal resistance of

the dice within the stack increase as the number of tiers increases. The latter is due to the fact that the inner dice do not have direct access to a heat sink. Prior studies have shown the possibility of using interlayer microfluidic cooling instead of conventional air-cooling for 3D ICs to overcome this challenge [6]–[10]. The height of most reported interlayer microfluidic heat sinks ranges from 200 μm to 400 μm in [6]–[10]. Because of the insertion of these microfluidic heat sinks, a wafer will have a typical thickness of a few hundred micrometers, which presents significant challenges to TSV fabrication and electrical performance. Therefore, designing a microfluidic heat sink without considering TSV fabrication compatibility and TSV parasitics greatly diminishes the advantages of 3D ICs. Microchannel dimensions were designed to maximize the TSV density while meeting the upper limit of thermal constraints in [7]. A microchannel infrastructure with bent microchannels was designed to bypass the region that contains TSVs in [17]. However, to date, there has been neither an attempt to optimize the heat sink design while accounting TSV fabrication compatibility nor an attempt to analyze the corresponding impact of the microfluidic heat sink (MFHS) on the electrical performance of TSVs. This section of the thesis investigates the thermal and electrical co-design of an interlayer MFHS. Design trade-offs between heat removal capability and the associated TSV parasitics are analyzed.

2.2 Literature Review of Microfluidic Heat Sink

Figure 13 summarizes the evolution of microfluidic heat sink design. A few key efforts are included for comparison. Microfluidic heat sinks for heat removal from IC were first demonstrated by Tuckerman and Pease in 1981 [6]. In their work, a heat removal of 790 W is demonstrated in a 1 cm^2 area. The lowest thermal resistance that is reported in their work is 0.09 K/W under a pressure drop of 213.9 kPa (corresponding to 512 mL/min). The height of the heat sink is 302 μm. Owing to advancement in microfabrication, more complicated structures can be designed to enhance the

Structure		Reference	Conclusions	Dimensions (μm)
Plain Microchannel		D. B. Tuckerman and R. F. W. Pease (Stanford), 1981	0.09 K/W at 213.9 kPa (516 mL/min)	W_{ch}=50; W_w=50; H=302
Enhanced Microchannel		E. G. Colgan et. al. (IBM), 2005	0.21 K·cm^2/W at 27 kPa (1250 ML/min)	W_{ch}=75; W_w=25; L_w=250; H=195
Staggered Micropin-fin		Y. Peles et. al. (RPI), 2005	0.03 K/W at 203 kPa	D=100; Porosity=0.65; H=400
Pearl Chain		T. Brunschwiler et. al. (IBM), 2009	0.3 K/W at 18 kPa (100 mL/min)	D=100; Pitch=200; H=200
Staggered Micropin-fin		T. Brunschwiler et. al. (IBM), 2009	0.2 K/W at 40 kPa (100 mL/min)	D=100; Pitch=200; H=200

Figure 13: Selected single-phase microfluidic heat sink geometries in the literature.

cooling performance. One important method is to fabricate an obstruction along the flow path so that the fluid is continuously disturbed, and provides a higher heat transfer coefficient. An enhanced microchannel structure using offset microchannels is proposed by Colgan et al. [36]. The heat transfer coefficient in the staggered fashion is significantly higher than in continuous microchannels [37]. Changing from offset square channels to offset circular pins can further enhance the heat transfer. Peles et al. have demonstrated a single-phase microfluidic heat sink using staggered micropin-fins, as shown in the figure. Compared to the plain microchannel, the thermal resistance decreases by 33% at a similar pressure drop. More geometrical variations including microchannels, in-line and staggered micropin-fins, and drop-shaped micropin-fins structures are investigated in a recent IBM work (shown in the last two rows in the figure). Two representative structures are selected for comparison. According to their evaluation, the staggered micropin-fin structure provides the lowest thermal resistance at a constant flow rate.

In addition to the method of introducing obstructions in the flow direction, other novel methods have been investigated. For example, [38] shows that microchannels with a sinusoidal roughness profile can significantly increase the heat transfer coefficient with little pressure drop penalty. Two-phase microfluidic cooling [39]–[41] and active thermoelectric coolers to address hotspots [42] are also investigated.

Owing to the ease of implementation of single-phase microfluidic cooling and relatively lower pressure drop compared to two-phase cooling, the single-phase MPFHS is chosen in this work for the applications in 3D ICs. Although different micropin-fin layout has been studied in previous work, very little work has looked at the impact of heat sink design on the interconnect performance. The following sections will focus on the trade-off between thermal performance (including thermal resistance and pressure drop) and interconnect electrical performance.

2.3 Heat Transfer Theory for Micropin-fin Heat Sink

Thermal resistance and pressure drop across the heat sink are key metrics for evaluating a heat sink. The total thermal resistance R_{total} consists of three parts: R_{cond} is due to the conductance from the circuit through the substrate and the heat sink interface; R_{conv} accounts for the convection from the substrate to the liquid; R_{heat} is due to the temperature increase of the cooling fluid as it flows across the heat sink [6]. For most cases, R_{cond} has a small contribution since the heat sink is close to the heat source and can be neglected. R_{total} is derived as follows:

$$R_{total} = R_{conv} + R_{cond} + R_{heat} \approx R_{conv} + R_{heat} \tag{1}$$

$$R_{total} \approx \frac{1}{h_{ave}A_t} + \frac{1}{W_t c_p} \tag{2}$$

where W_t and c_p are mass flow rate and specific heat capacity, respectively. The average heat transfer coefficient, h_{ave}, is calculated as

$$h_{ave} = Nu \cdot k_f / D_h \tag{3}$$

$$Nu = CRe^m Pr^{0.36} (Pr/Pr_s)^{0.25} \tag{4}$$

where k_f and D_h are the thermal conductivity of the fluid and the micropin-fin hydraulic diameter, respectively. The Nusselt number, Nu, can be evaluated using (4), where Re and Pr are the Reynolds number and Prandtl number evaluated using the bulk fluid properties, respectively. Pr_s is the Prandtl number using the fluid property at the surface temperature. For the Reynolds number studied in this work, C and m take the value of 0.9 and 0.4 in (4), respectively [43]. A_t is the effective heat transfer area described as follows:

$$A_t = A_b + \eta A_{fin} \tag{5}$$

$$\eta = \frac{\tanh(2H_{fin}\sqrt{h_{ave}/k_{si}D})}{2H_{fin}\sqrt{h_{ave}/k_{si}D}} \tag{6}$$

where k_{si}, H_{fin}, D, and η are the thermal conductivity of silicon, micropin-fin height, micropin-fin diameter, and fin efficiency, respectively. A_b is the base area exposed to the fluid, and A_{fin} is the aggregate surface area of the pin-fins exposed to the fluid, and are calculated as follows:

$$A_b = A_{tot} - 1/4n\pi D^2 \tag{7}$$

$$A_{fin} = n\pi DH_{fin} \tag{8}$$

$$N_{fin} = \frac{(W - W_s)(L - L_s)}{P_w P_l} \tag{9}$$

where W and L are the width and length of the entire chip, respectively; W_s, L_s, P_w, and P_l are the horizontal and vertical spacing and the pitch between the pins. N_{fin} is the total number of pins. This correlation-based model is valid under the following conditions: $10 \leq Re \leq 1000$, $2 \leq H_{fin}/D_h \leq 20$, $20 \leq L/H_{fin} \leq 200$, and with a pitch to diameter ratio of 1.25 to 3 [43].

In this work, we use (10) to calculate the Darcy friction factor (f), which is an empirical correlation model derived by Short et al. [44]. The pressure drop (ΔP) is then calculated using (11) [43].

$$f = 104.4(\frac{P_w}{D_h})^{-1.3}(\frac{P_l}{D_h})^{-0.78}(\frac{H_{fin}}{D_h})^{-0.55}Re^{-0.65} \tag{10}$$

$$\Delta P = N\rho\frac{V_{\max}^2}{2}f \tag{11}$$

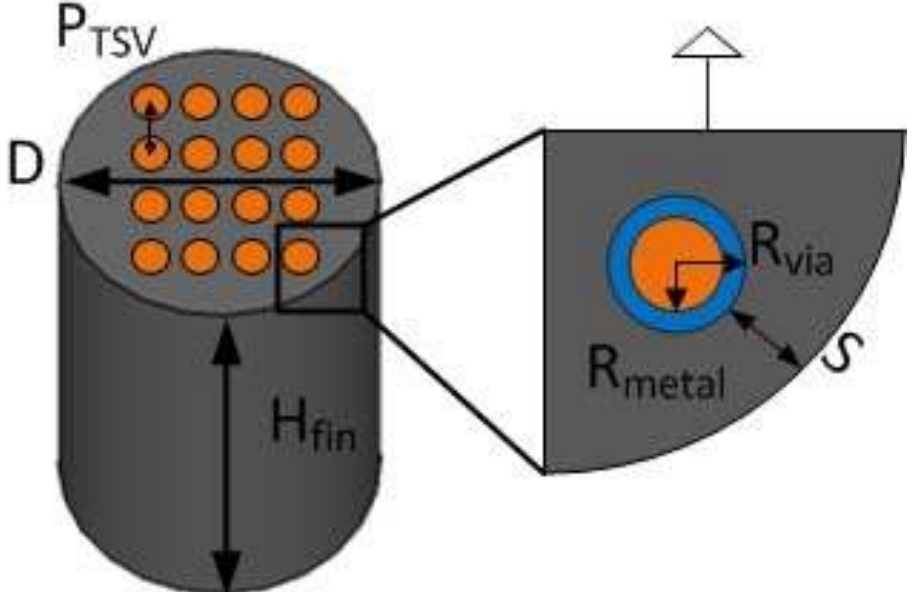


Figure 14: TSV array integrated in a silicon micropin-fin .

where V_{max} is the mean fluid velocity at the minimal cross-section, and N is the number of rows in the horizontal direction. Please note the above models are developed for laminar flow, which is dominant in microscale structures that are studied in the present work.

2.4 Electrical Parasitics of TSVs Embedded in microfluidic heat sink

As shown in the 3D IC prototype (Figure 7), TSVs need to be routed through the silicon micropin-fins. Figure 14 is a schematic of an array of TSVs embedded within a silicon micropin-fin. The total number of TSVs (N_{TSV}) that can be integrated within the heat sink can be calculated based on geometrical considerations. The number of TSVs per micropin-fin (n_{TSV}) is calculated by

$$n_{TSV} = \left\lfloor \frac{D - 2 \times (S + R_{via})}{\sqrt{2} P_{TSV}} \right\rfloor^2 \tag{12}$$

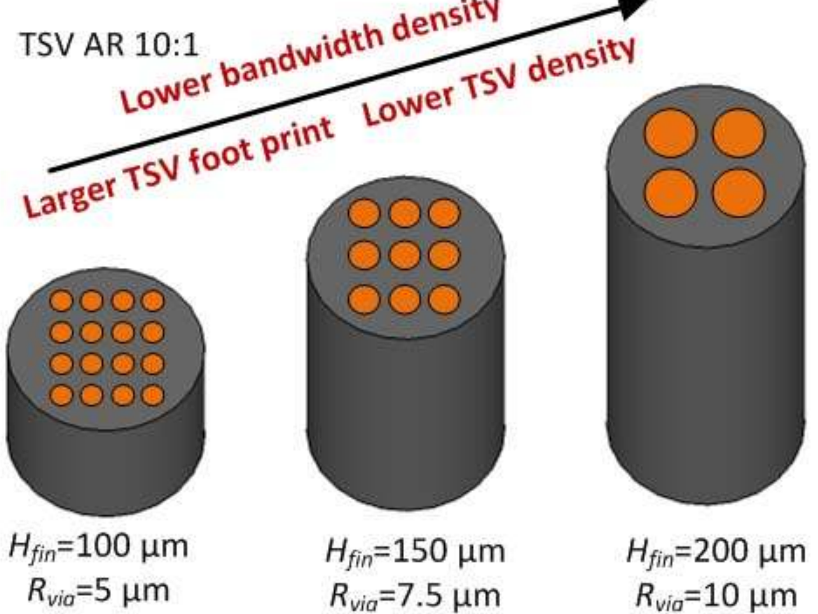


Figure 15: Impact of H_{fin} on TSV density and diameter.

where R_{via} and P_{TSV} are the radius and pitch of the TSVs, respectively. S denotes the spacing between the TSVs and the micropin-fin edge. This spacing is used to compensate for the misalignment during the fabrication. S is assumed to be 10 μm in this work. P_{TSV} is assumed to be $3 \times R_{via}$, while R_{via} is determined from H_{fin} and the TSV aspect ratio (AR) by (13).

$$R_{via} = \frac{1}{2}(H_{fin}/AR) \tag{13}$$

N_{TSV} is given by (14).

$$N_{TSV} = N_{fin} \times n_{TSV} \tag{14}$$

The empirical expression for TSV dielectric capacitance is as follows:

$$C_{ox} = \frac{2\pi\varepsilon_{ox}L_{TSV}}{\ln\left(\frac{R_{via}}{R_{metal}}\right)} \tag{15}$$

where ϵ_{ox} is the oxide permittivity, L_{TSV} is the TSV length, and R_{metal} (Figure 14) is the radius of the copper. The dielectric capacitance model is verified against measured results from [45]. The modeled dielectric capacitance for a TSV with $R_{via} = 5.2$ μm, $R_{metal} = 5.07$ μm, and $L_{TSV} = 26$ μm is 115.5 fF while the measured value is 126 fF, thus showing reasonable agreement.

2.5 Thermal Resistance and Pressure Drop Trade-off Analysis of Microfluidic Heat Sink

Before taking the electrical parasitics of TSVs into account, thermal analysis is first done. There is a trade-off between thermal resistance and pressure drop. As flow rate increases, thermal resistance will decrease. However, the increased flow rate will result in a larger Darcy friction factor, as shown in Equation (11), and the pressure drop across the microfluidic heat sink will increase. Figure 16 shows the trade-offs between ΔP (y axis) and the R_{total} (x axis) for different configurations as the volumetric flow rate decreases from left to right for a 1 cm $\times$ 1 cm chip area. The parameters that are varied in the heat sink design are channel height (H_{fin}), pin diameter (D) and pitch to diameter ratio. Since the thermal resistance should be low to be sufficient to cool the high power-density chip, the target that we set is R_{total} <0.2 K/W based on ITRS projections. Meanwhile, we set the upper limit for pressure drop (ΔP) to be 40 kPa due to the pump power and size limitations. High flow rate provides better cooling capability and is accompanied with higher friction factor which brings high pressure drop. Thus, R_{total} decreases and ΔP increases as the flow rate goes higher for all configurations. In general, high H_{fin} provides larger effective heat transfer area and lower friction factor. It is consistent with the result in the plot: R_{total} and ΔP

decrease as the heat sink height increases. However increasing H_{fin} is not ideal for a 3D IC system because it will increase TSV length and possibly TSV diameter if we assume a fixed TSV aspect ratio. Consequently, TSV density decreases and electrical performance degrades. In conclusion, our design rules are the following:

- Provide a low total thermal resistance.
- Provide a low pressure drop.
- Maintain the heat sink height as small as possible.
- Maximize available silicon area for TSV routing for potential applications that require it.

2.6 Electrical-Thermal Trade-off Analysis of TSVs in Microfluidic Heat Sink

Larger H_{fin} means higher effective heat transfer area until the decrease of η causes a degradation. Since larger H_{fin} provides a larger flow path, the friction factor is reduced [46], [47]. Therefore, a larger H_{fin} is generally preferred to obtain smaller thermal resistance and pressure drop. The optimized heat sink design for either microchannels or inline/staggered MPFHS has been derived previously, and a few key results are summarized in Table 1 [12], [46].

In 3D IC applications, the heat sink design not only needs to achieve the target heat removal capability and pressure drop, but it should also be compatible with TSV fabrication and their target electrical parasitics. The most important variable in their co-design is the height of the micropin-fin (H_{fin}). H_{fin} greatly impacts TSV diameter, TSV density, and TSV capacitance, which in turn influences interconnect latency, bandwidth density, and power consumption. As seen in Table 1, prior heat sink designs have a large height. Although a high H_{fin} value may decrease the thermal resistance of the heat sink, assuming a fixed TSV AR, it results in a large TSV

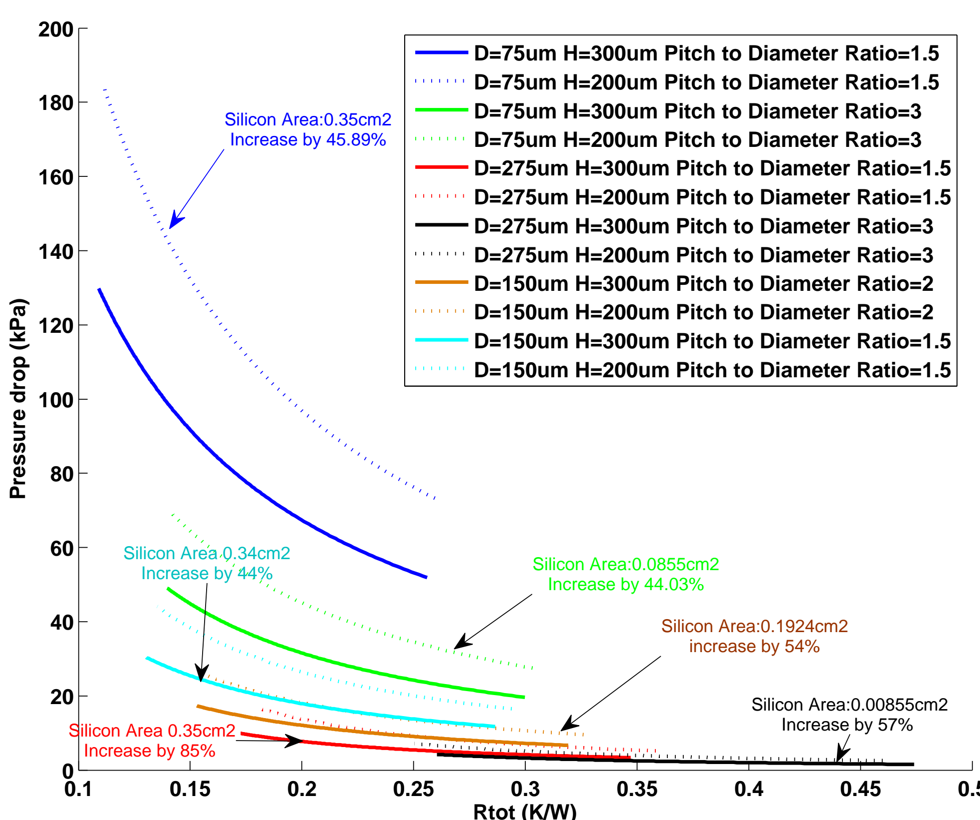


Figure 16: Thermal resistance and pressure drop for different micropin-fin designs with flow rate 150 ml/min to 60 ml/min.

Table 1: Selected optimal heat sink dimensions from the literature

Heat Sink Type	Dimensions (μm)
Microchannel [12]	Channel width (w_c) = 65 Wall width (w_c) = 65 Channel height(H_{ch}) = 399.75
Microchannel [46]	$w_c = 65$ $w_c = 63.7$ $H_{ch} = 929.5$
Staggered micropin-fin [46]	$D = 196$ $Pitch = 305.8$ $H_{fin} = 3155$

diameter and a large TSV capacitance (assuming the TSV is still manufacturable). TSV technology is normally limited by AR. Etching, sidewall passivation, and metal filling of high AR TSVs are very challenging. Figure 15 illustrates how H_{fin} impacts TSV diameter and density qualitatively. For example, in a 100 μm tall micropin-fin, the TSV diameter is 10 μm with a limited AR of 10:1. This allows the integration of 16 TSVs per micropin-fin. When the micropin-fin height increases to 200 μm, the TSV diameter increases to 20 μm, allowing only four TSVs per micropin-fin.

In Figure 17, for a constant pressure drop of 80 kPa, the total thermal resistance (R_{total}) decreases from 0.34 K·cm^2/W to 0.23 K·cm^2/W as H_{fin} increases from 130 μm to 220 μm. Meanwhile, the TSV dielectric capacitance (C_{ox} with 500 nm liner) with a 10:1 AR is observed to increase from 0.352 pF to 1.025 pF, leading to larger latency and energy consumption. This is, of course, not a desirable consequence. In addition, as H_{fin} increases, the flow rate needs to increase to maintain a constant pressure drop, resulting in higher pumping power. Another trade-off shown in Figure 17 is between the pressure drop and the thermal resistance. For the same H_{fin}, higher pressure (higher flow rate) corresponds to lower thermal resistance. At H_{fin} = 200 μm, R_{total} decreases from 0.38 K·cm^2/W to 0.2 K·cm^2/W as the pressure drop increases from 40 kPa to 120 kPa. The assumptions made in these models are: 1)

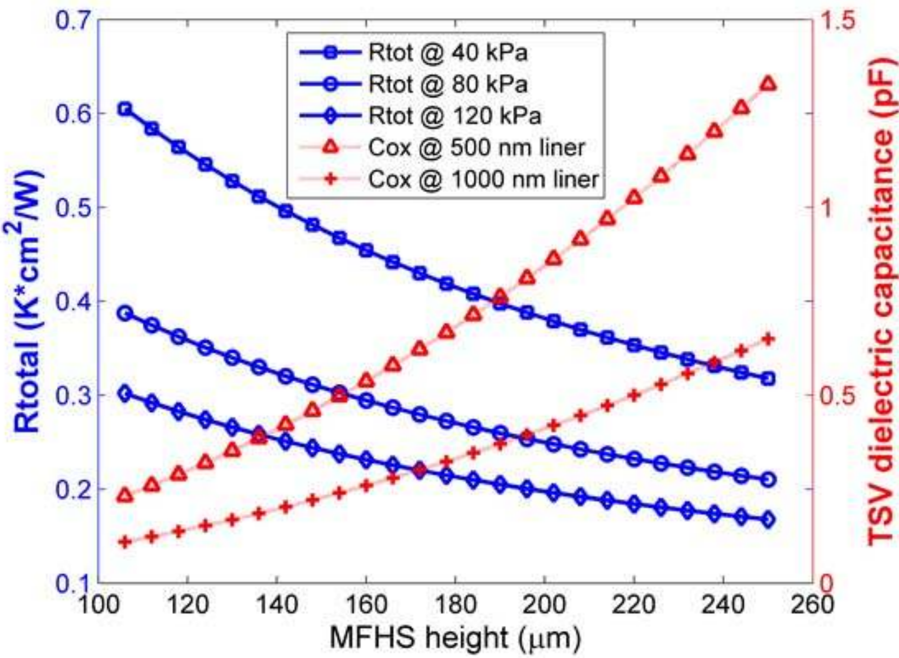


Figure 17: Thermal resistance and TSV capacitance as a function of microfluidic heat sink height at different pressure drop values.

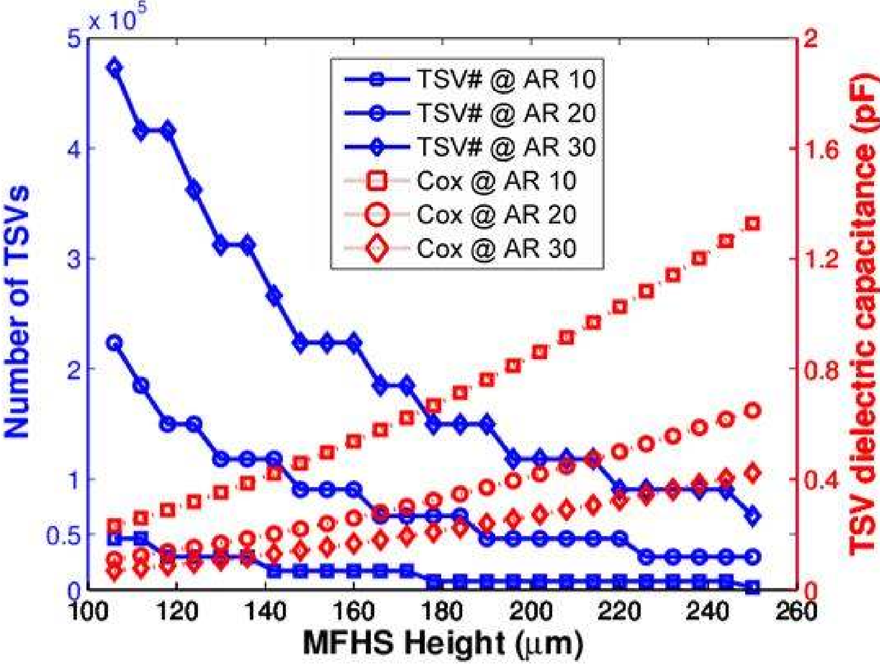


Figure 18: The impact of microfluidic heat sink height on the number of TSVs and TSV capacitance.

the heat sink spans 1 cm $\times$ 1cm, 2) the oxide liner of TSVs is 500 nm thick, 3) TSV pitch is 1.5$\times$TSV, 4) base silicon thickness is 50 μm, and 5) pressure drop across the micropin-fin array is 80 kPa unless otherwise noted in the analysis.

Figure 18 captures the impact of TSV AR and H_{fin} on N_{TSV} and C_{ox}. One trend observed is that C_{ox} increases and the number of TSVs decreases for higher H_{fin}. Another trend shown in Figure 18 is that as TSV AR increases, N_{TSV} increases and C_{ox} decreases. For example, as the TSV AR increases from 10:1 to 20:1 in a 200 μm tall MPFHS, N_{TSV} increases from 7,396 to 4.62×10^4, and C_{ox} decrease from 862 fF to 419 fF. Further increasing the TSV AR results in a larger TSV density and a smaller C_{ox}. Hence, H_{fin} should be designed as small as possible to achieve the best TSV performance. Yet, H_{fin} needs to be greater than a certain value (100 μm) in order to keep the pressure drop tolerable. To obtain a thermal resistance smaller than 0.2 K$\cdot$cm^2/W (Figure 4) and maintain the smallest die thickness (in order to get the low TSV parasitics and high TSV density), H_{fin} is chosen to be 200 μm in this work. The other selected MPFHS geometries are $D = 150$ μm, $P_w = P_s = P = 225$ μm.

2.7 Alternative Methods to Reduce TSV Capacitance

2.7.1 Novel Liner Material

Although H_{fin} can be designed small, the total thickness still increases drastically because of the insertion of interlayer MFHS. As a result, for a fixed aspect ratio, TSV diameter will increase, leading to larger TSV capacitance. Although increasing the oxide liner thickness may reduce TSV capacitance, it is not cost effective nor easily manufacturable. We propose replacing the silicon dioxide liner with polymer or air. Polymer and air liner concepts have been shown in [45], [48]. TSV dielectric capacitance is plotted as a function of H_{fin} for different liner materials in Figure 19. The assumed dielectric constant of silicon dioxide, Parylene-C, and air is 3.90, 2.95 (at 1 MHz) [49], and 1.00, respectively . By replacing the 500 nm oxide liner with

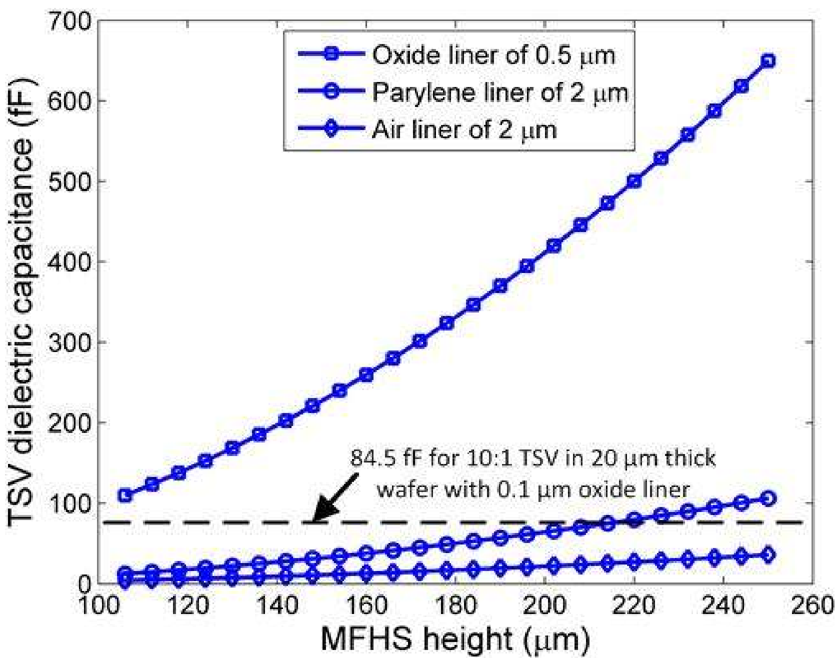


Figure 19: TSV dielectric capacitance as a function of liner: oxide, Parylene-C, and air.

500 nm Parylene-C liner, the dielectric capacitance decreases from 419 fF to 317 fF for $H_{fin} = 200$ μm. This is due to the reduction of the dielectric constant. By further increasing the thickness of Parylene-C liner to 2 μm, the TSV dielectric capacitance decreases to 65.7 fF for $H_{fin} = 200$ μm. For the same dimensions, C_{ox} is reduced to 22 fF by replacing the oxide liner with a 2 μm air liner. For reference, C_{ox} of a 10:1 AR TSV with a 0.1 μm oxide liner in a 20 μm thick die (i.e., without embedded MPFHS) is plotted in Figure 19. As can be seen, the reduced capacitance values are comparable with the reference value. Another benefit of using a polymer liner is that the polymer has a lower Young's modulus and may serve as a stress buffer between silicon and the metal-filled TSVs [48]. However, the resistance of the TSVs will increase from 52 mΩ to 115 mΩ when the liner thickness increases from 0.5 μm to 2 μm.

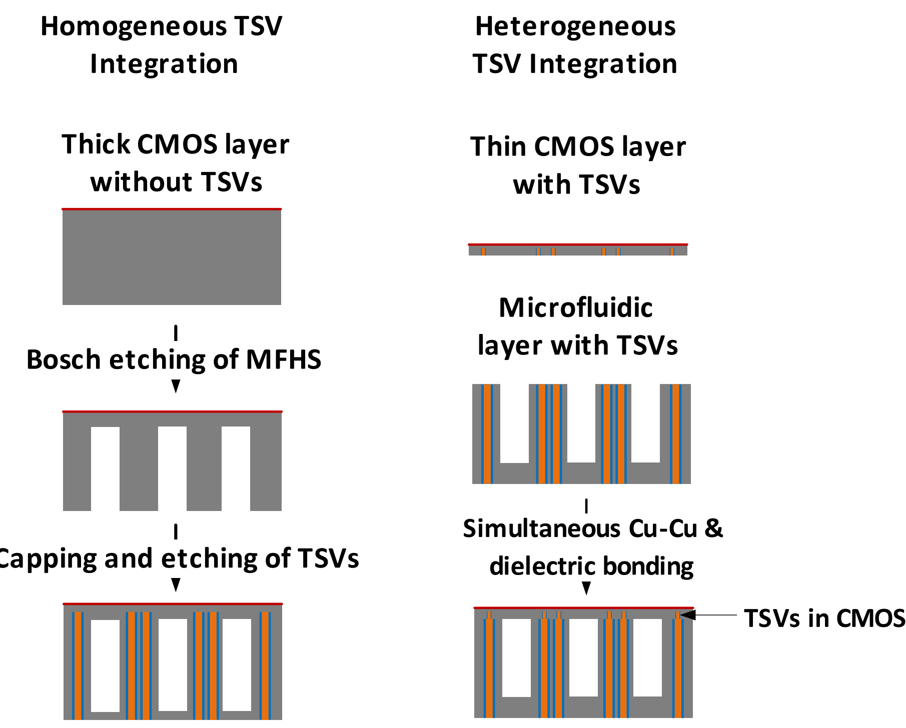


Figure 20: Homogeneous and heterogeneous approaches for TSV integration into microfluidic cooled chip.

2.7.2 Heterogeneous TSV Integration

Two possible methods of integrating MFHS and TSVs into 3D ICs are described in this section. As shown in Figure 20, one is denoted as the homogeneous TSV approach. This approach starts with a thick CMOS chip without TSVs or a MFHS. MFHS is then etched on the back side of the CMOS chip. After capping the MFHS, TSVs are etched into the silicon containing CMOS and the microfluidic layer. The second proposed approach is based on heterogeneous TSV integration [50]. In this case, the CMOS layer and the microfluidic layer, each with their own TSVs, are fabricated independently. The two layers are electrically and mechanically bonded simultaneously at the end, for example, using a hybrid bonding technique discussed in [51]. The advantages of heterogeneous TSV integration mainly involve three aspects. Firstly, the TSVs in the CMOS layer become much smaller, leading to conservation of precious silicon area in the CMOS chip. Table 2 illustrates how much area is saved by the heterogeneous integration method. The chip area occupied by TSVs is reduced by two orders of magnitude with heterogeneous TSV integration. Secondly, since the microfluidic layer is fabricated independently, restrictions on temperature and materials are eliminated. This creates more flexibility in the processing of the microfluidic layer. For example, it enables one to pursue bottom-up electroplating for the TSVs (different from the superfill process adopted in CMOS layer) in the MFHS silicon wafer, which can provide much higher aspect ratio TSVs. Lastly, heterogeneous integration brings about greater opportunity to explore novel dielectric liner materials and processes, e.g. thick oxide liner or polymer liner; the benefit being that the total TSV capacitance can be reduced despite the larger TSV dimensions in the microfluidic layer.

Figure 21 illustrates the bonding of a two-tier 3D IC stack in which each tier has TSVs integrated with microfluidic cooling. The top tier is electrically and mechanically bonded to the bottom tier by flip-chip bonding. The electrical microbumps are

Table 2: Comparison of the area occupied by TSVs for homogeneous and heterogeneous TSV integration

	Assumptions	Area occupied by TSVs in CMOS layer
Homogeneous Integration	Chip area=1 cm × 1 cm $H_{cooling}$=200 μm H_{CMOS}=10 μm	4.3%
Heterogeneous Integration	TSV aspect ratio=20 : 1 N_{TSV}=0.5x10^5	0.0098%

for signaling and power delivery. Fluid is supplied through the interposer and delivered into each tier through solder- or polymer-based microfluidic chip I/Os [52]. A reliable method to deliver fluid into each tier without leakage is a critical challenge for implementing microfluidic cooling in 3D ICs. Recent advances in wafer-level batch-fabricated solder microfluidic chip I/Os and fine-pitch electrical microbump I/Os have been made [16]. The solder-based microfluidic chip I/Os have been experimentally shown to withstand a pressure drop of 100 kPa without leakage.

2.8 Conclusion

In this chapter, the thermal–electrical modeling of a microfluidic heat sink and TSVs is performed. Trade-offs between cooling capability and TSV parasitics are analyzed for the first time. It should be noted that when the heat sink height increases, the thermal resistance of the microfluidic heat sink will decrease. However, TSVs will have larger diameter and larger parasitics. In this sense, the heat sink should be designed as low as possible to ease TSV integration. To compensate for the thickness increment, high aspect-ratio TSVs should be developed. The results in Section 2.6 shows that C_{ox} decreases by 51.4 % when TSV AR increases from 10:1 to 20:1. In this case, TSV density also increases from 7,396 to 4.62×10^4 /cm^2. Even more, a novel liner such as SU-8 can further reduce the oxide capacitance. A heterogeneous TSV integration method is also proposed to save valuable silicon area in the CMOS

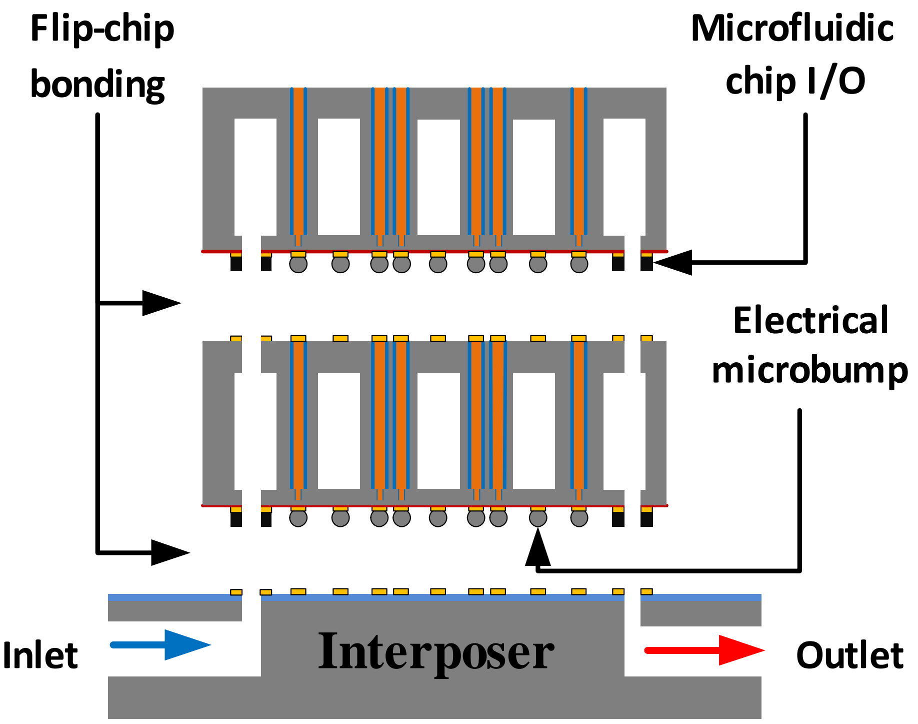


Figure 21: Two-tier 3D IC stack with microfluidic heat sink and TSVs.

layer.

CHAPTER III

TSV-COMPATIBLE MICROPIN-FIN HEAT SINK EXPERIMENTS

3.1 Introduction

As stated in Chapter II, very little effort is made to analyze the thermal–electrical trade-offs between TSVs and microfluidic heat sink designs. In addition, there are very few research efforts that demonstrate TSVs and microfluidic heat sink co-integration. The reason is that the die thickness increases drastically because of the insertion of the microfluidic heat sink. While the aspect ratio (AR) of most TSVs today is limited to 10:1, the TSVs will have large diameters and thus large parasitics. The key to compensating for this increased TSV height is to increase TSV AR. As shown in Figure 18, for example, as the TSV AR increases from 10:1 to 20:1 in a 200 μm tall MPFHS, N_{TSV} increases from 7,396 to 4.62×10^4, and C_{ox} decreases from 862 fF to 419 fF. Further increasing the TSV AR results in a larger TSV density and a smaller C_{ox}. However, very little work focuses on high AR TSV development.

For TSVs in CMOS, where wafer thickness is typically less than 100 μm, aspect ratios as high as 15:1 have been demonstrated [18], [19]. For thick silicon dice (greater than 100 μm), TSVs with aspect ratio greater than 10:1 have been shown for the application of silicon interposer [20]. TSVs with higher aspect ratio need to be developed for thicker silicon die (with embedded MFHS). Moreover, experimental integration of fine-pitch and high AR TSVs within MFHS is missing. A 3D IC system featuring TSV-compatible interlayer microfluidic cooling and high AR TSVs is shown in Figure 22. The proposed system features processor and memory stacks enabled through interlayer microfluidic cooling and low-parasitic TSVs. Since the microfluidic

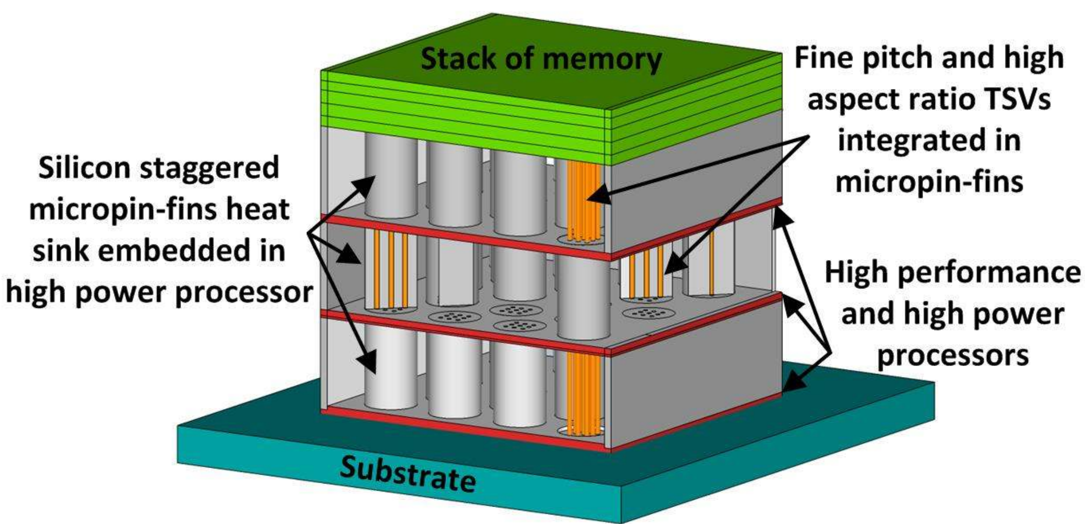


Figure 22: Schematic of a three-microprocessor chip stack each with interlayer microfluidic cooling. A 3D stack of memory chips resides above the microprocessors. High AR TSVs are integrated in the MPFHS.

cooling solution is chip scale, the 3D chip stacks can be placed virtually side-by-side and, thus, shorten off-chip interconnects.

In this chapter, a TSV-compatible staggered MPFHS is developed and fabricated. In order to demonstrate the TSV compatibility, 18:1 AR copper TSVs are integrated in the MPFHS. Thermal experiments for the fabricated MPFHS and benchmarks against a high-performance air-cooled heat sink (ACHS) are included. Thermal testing results and four-point resistance measurements of TSVs are also reported.

3.2 Fabrication of the TSV–Compatible Micropin-Fin Heat Sink

3.2.1 Bonding Process Selection

The most important and complicated step in the fabrication process is to select an appropriate bonding method to encapsulate the fluid. A suitable bonding process in the microfluidic heat sink application will have the following characteristics:

- Hermetic

- Stable under high pressure
- Stable when interacting with fluidic
- Mechanically reliable during thermal cycles

Several wafer bonding methods have been performed and tested in the literature [53], [54]. One well-studied method uses an intermediate layer (such as SU-8, polyimide, parylene-C) to bond two layers. Such intermediate polymers typically require a low bonding temperature. For example, ≤200 °C is required for wafer bonding using SU-8. The bonding quality is fairly independent of the surface roughness and planarity [53]. However, in order to prevent the degradation of SU-8, the bonded sample cannot be exposed to a temperature higher than ∼380 °C [55]. Moreover, adding an intermediate bonding layer adds a conductive thermal resistance to the ICs in the stack, which will degrade the cooling capability of the heat sink. Furthermore, TSVs need to be routed through the MPFHS layer for interlayer communication. Therefore, a bonding method without additional material is preferred in 3D IC applications.

Another well-established bonding method is anodic bonding. Anodic bonding can be used to bond a silicon wafer to a pyrex wafer without any intermediate polymer layer. Figure 23 shows the bonding theory [56]. When a high electrical voltage is applied between the silicon and the glass, charge separation occurs in sodium dioxide. The sodium ions (Na^+) drift toward the top of the glass wafer while the oxygen ions (O^{2-}) drift toward the bottom of the glass wafer. When the oxygen ions reach the boundary, they react with silicon and form SiO_2. This thin SiO_2 layer provides good bonding strength between the two substrates. The advantage of using anodic bonding is to allow a transparent view of the fluid. This is especially important to evaluate the flow in the channel. For example, in two-phase cooling, anodic bonding allows for flow regime visualization [11]. However, this bonding is not suitable for 3D stacking of chips because glass introduces large thermal resistance to the 3D stack. In addition,

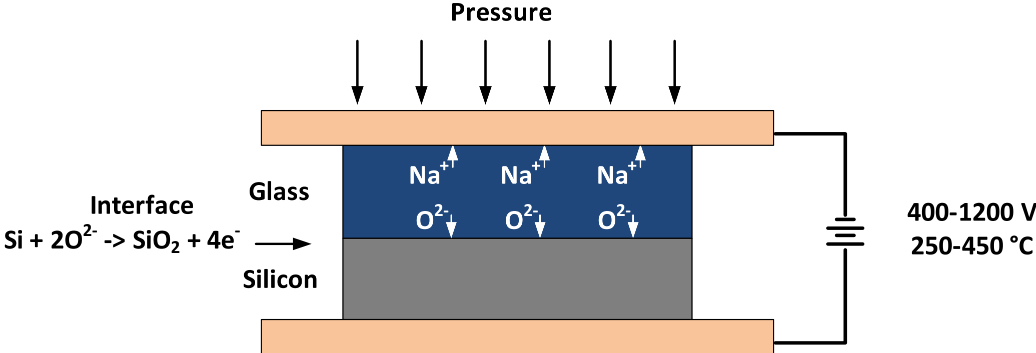


Figure 23: Theory of anodic bonding.

forming TSVs through glass is still challenging. Lastly, because of the coefficient of thermal expansion (CTE) mismatch between glass and silicon, thermal cycles can induce large stress at the interface and increase reliability concerns.

The last method is direct fusion bonding. Direct fusion bonding is used to join two clean silicon wafers together. Figure 24 shows the bonding diagram. Two clean silicon wafers bond at room temperature while pressure is applied. The Si–O–Si bond ensures the bonding strength. Figure 25 illustrates the detailed bonding process flow. Two silicon wafers first go through standard RCA clean to remove residual and make the surface hydrophilic. Oxygen plasma is then used to activate the surface and create hydroxyl radicals (HO). At room temperature, when the two cleaned wafers are brought together, the hydroxyl radicals will form the initial weak Si–O–Si bond. To strengthen the bond, an annealing at 400 °C is performed afterward. As can be seen, there is no intermediate material involved in this process. Since the other layer is also silicon, it is suitable for 3D stacking. In addition, fabricating TSVs is not a problem in this case. Lastly, since there is no CTE mismatch issue, the reliability can be improved.

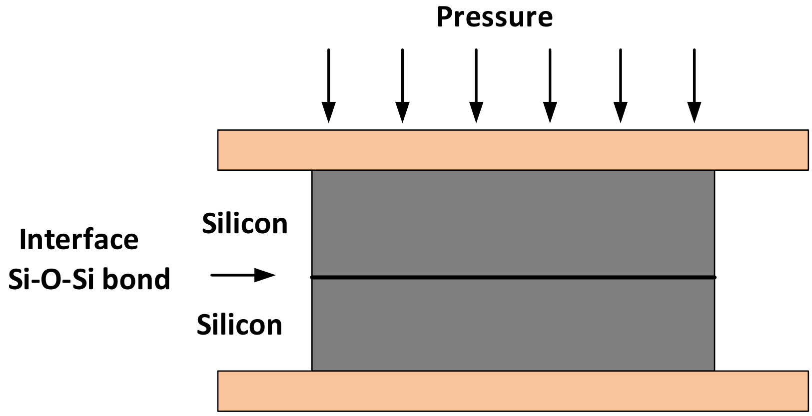
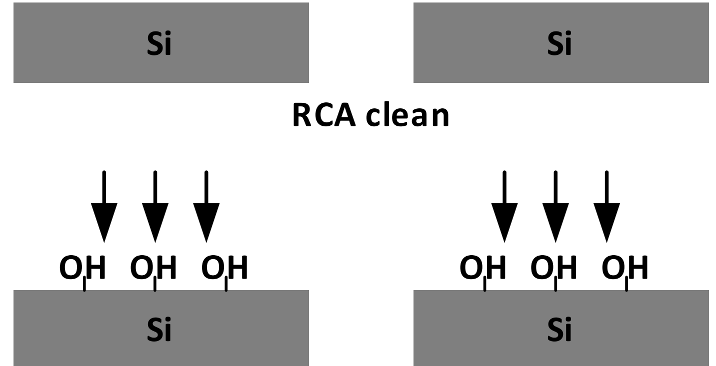


Figure 24: Theory of fusion bonding.

3.2.2 Fabrication Process of MPFHS

After selecting the bonding process, the fabrication process is initially developed, as shown in Figure 26. The fabrication starts with a cleaned double-side polished wafer. A standard Bosch process, which alternates between SF_6 (plasma etching step) and inert C_4F_8 (passivation step) is used first to etch silicon. A 200 μm ($\pm$2 μm) deep micropin-fin array is etched. Next, the wafer is flipped over and a platinum (Pt) spiral heater is sputter-coated and patterned onto the back side of the chip to emulate the heating of a microprocessor. Owing to the linear resistance–temperature relationship, the Pt heater also serves as the resistance thermal detector (RTD) during the thermal measurements. The next step is to encapsulate the micropin-fins using Si–Si fusion bonding. Fluidic vias are then etched to enable liquid circulation into and out of the microfluidic heat sink. The final step is to attach the Nanoports (from Idex Health and Science) to provide consistent fluid connections to the testbed.

During the Si–Si fusion bonding process, no cracking is observed after the initial contact at room temperature. However, after the 400 °C annealing, cracking occurs. The reason may be that photoresist residual is trapped between the bonded surfaces and causes initial cracking. After the hermetic seal is formed, air is trapped in the microchannels and expands when temperature increases, which induces large stress



Oxygen plasma activation

Initial contact at room temperature

Annealing at 400 °C to strengthen Si-O-Si bond

Figure 25: Si–Si fusion bonding process.

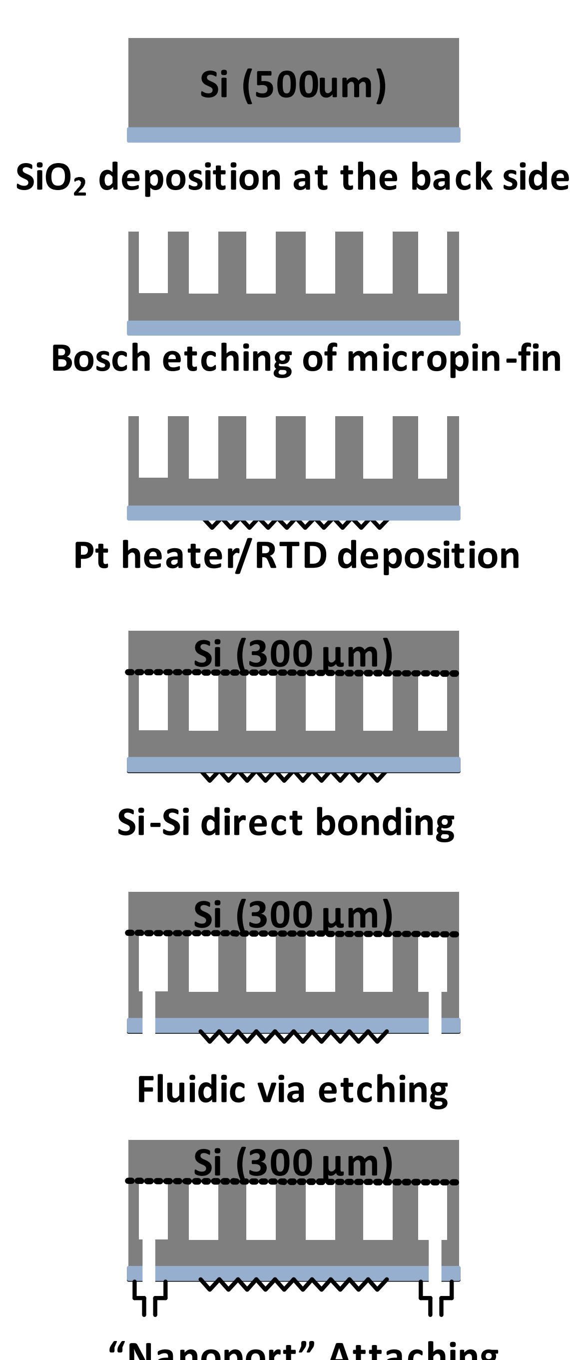


Figure 26: Initial process flow of the MPFHS.

on the two wafers. This cracking may propagate because of the increased stress. In order to alleviate the stress during annealing, the key is to create a path for the air to escape. To this end, the process is modified, as shown in Figure 27. In this new process, the Si–Si bonding step is moved after the 'fluidic via etching' step. Therefore, the trapped air can escape from the fluidic vias and alleviate the stress. No cracking is observed after the annealing.

The etched micropin-fins are shown in Figure 28. In Figure 28(a), a top-view of an array of staggered micropin-fins is shown. The diameter of a single pin is 150 μm. The vertical pitch is 225 μm. The next column is displaced upwards by 112.5 μm. These parameters are selected based on the thermal–electrical co-analysis in Chapter II. A tilted view is also included in Figure 29(a). After the MPFHS is encapsulated with another silicon wafer, the sample is diced and a cross-sectional image is taken (Figure 28(b)). After inspecting with SEM, it is confirmed that the two surfaces remain in intimate contact after dicing. The height of the etched micropin-fin array is 200 μm. Figure 29(b) is an infrared image of the top-view of the bonded wafer from which the high bonding yield can be observed. Air and metal appear brighter than the silicon surface in the IR image. From the IR image, we conclude that air is not trapped between the top of the micropin-fins and the capping wafer (the brighter color is the underlying Pt heater). Figure 30 shows an overall view of the MPFHS sink with the micropin-fin arrays in the middle. The micropin-fin array spans 1 cm $\times$ 1 cm and is used to cool a chip that is 1 cm $\times$ 1 cm. On the two sides, there are two large rectangles that serve as mechanical support. The fluid flows from left to right across the micropin-fin array. Fluid absorbs heat from the heat sink and the temperature increases gradually, as shown in Figure 30.

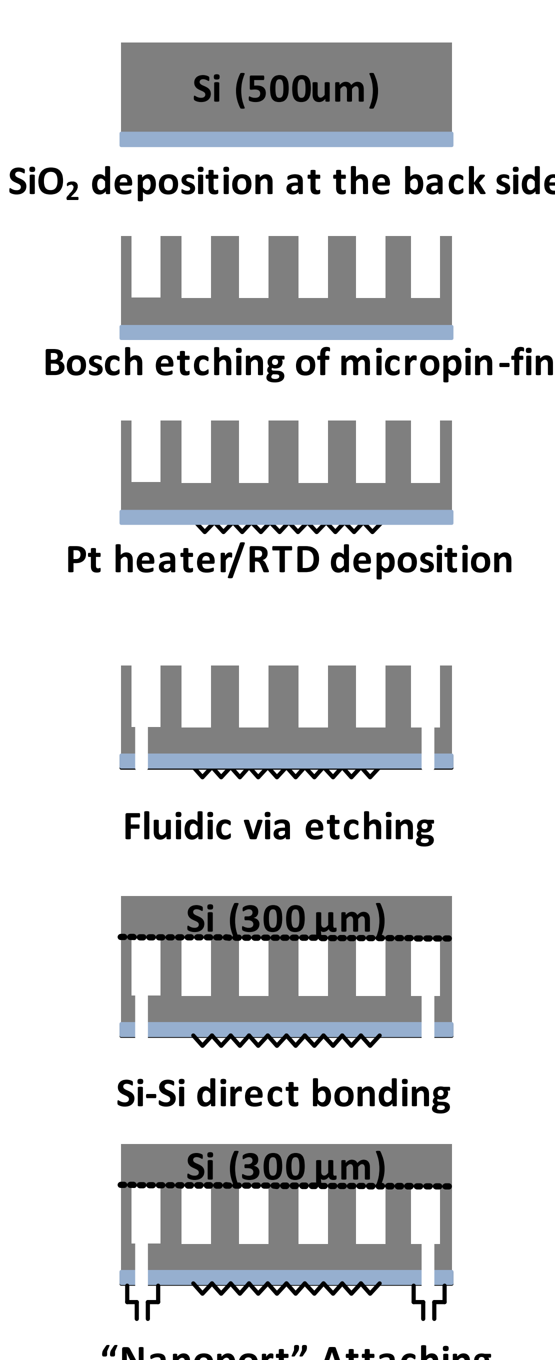


Figure 27: Modified process flow of the MPFHS.

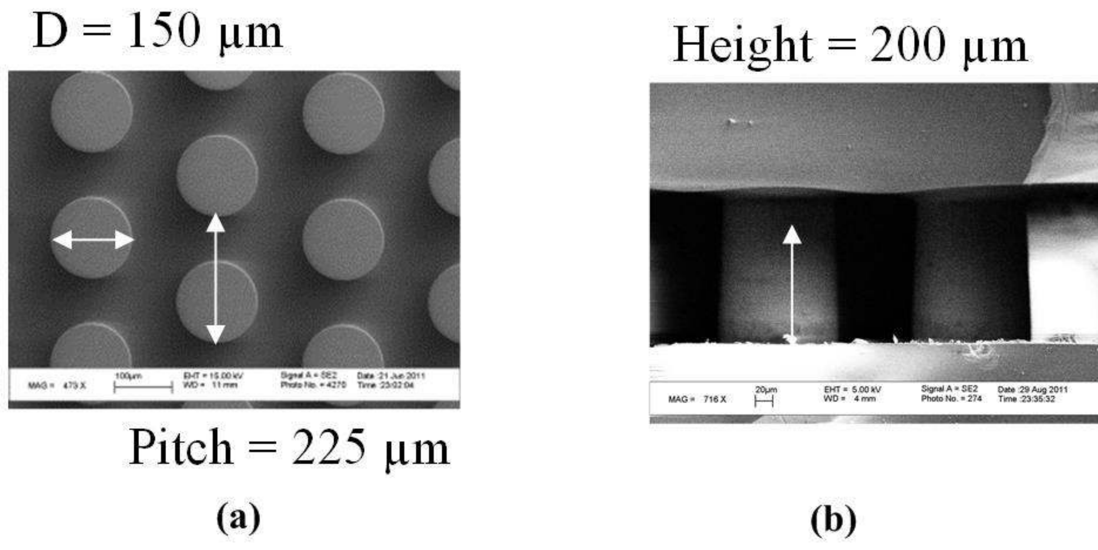
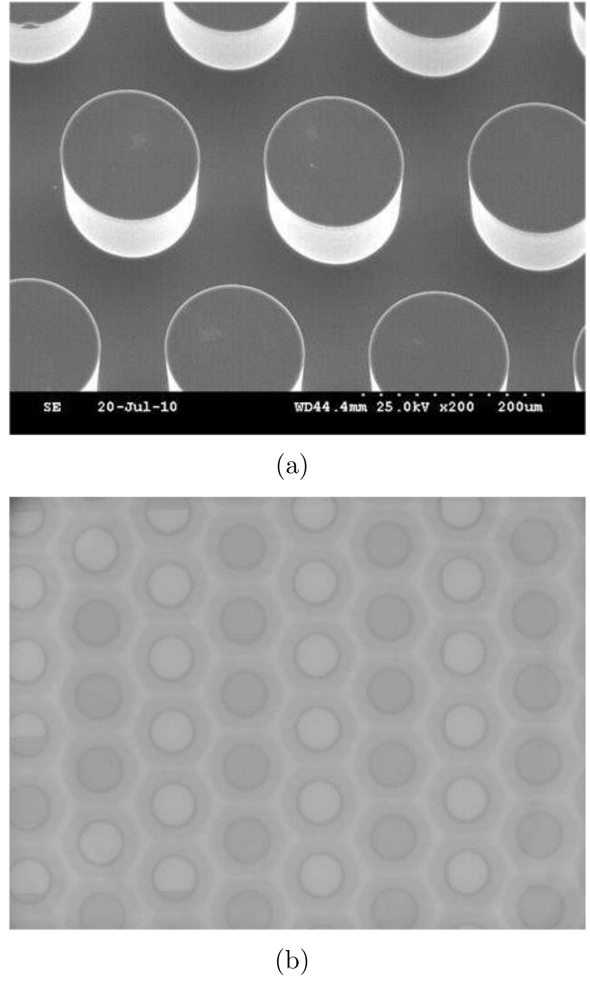


Figure 28: SEMs of (a) the top and (b) cross-sectional view of the micropin-fin arrays.

3.2.3 Integration of TSVs in MPFHS

The fabrication process flow of TSVs in MPFHS is shown in Figure 31. High-AR (18:1) TSVs are etched using the Bosch process. A thermal oxide liner is then grown to isolate the TSVs from the substrate. In Step 3, bottom-up pulsed electroplating with Enthone DVF plating solution is used to fill the vias with copper. Following electroplating, the sample is polished using iCue 5001 provided by Cabot Microelectronics Corp. The MPFHS is patterned and etched from the top side. The fabricated TSVs have a diameter of 10 μm and a length of 178 μm (18:1) [57]. A top-view of the fabricated high AR TSVs within the micropin-fin is shown in Figure 32(a). A 3 $\times$ 3 TSV array per micropin-fin is shown. The fabricated die spans 1 cm $\times$ 1 cm and has 1,936 micropin-fins. Each micropin-fin has nine electrical TSVs, providing a total of 17,424 electrical I/Os. The TSVs consume only 1.36% of the die area. The fabricated structure is then dipped in KOH to remove the silicon and to leave behind free standing high AR copper electroplated pillars (TSVs) (SEM shown in Figure



(a)

(b)

Figure 29: (a) The tilted view of micropin-fins and (b) an infrared image of the bonded sample.

32(b) and optical image shown in Figure 33(a)). To further inspect the copper electroplating, optical images are taken as shown in Figure 33. Figure 33(b) shows the cross-sectional image of the electroplated copper. This verifies the absence of voids in the electroplated TSVs. The copper contact at the bottom of the TSVs in Figure 32 and Figure 33(b) short-circuits the TSVs so that a four-point resistance measurement can be performed to measure single TSV resistance.

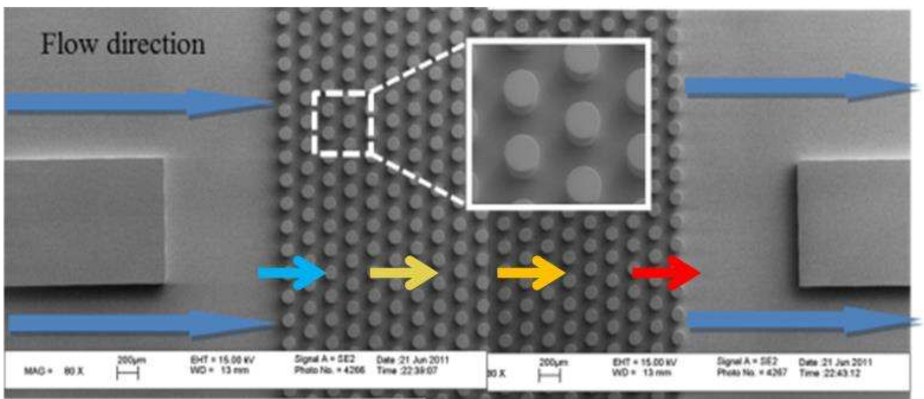


Figure 30: Overview of the MPFHS with a magnified angled view of the micropin-fins. Fluid flows from left to right.

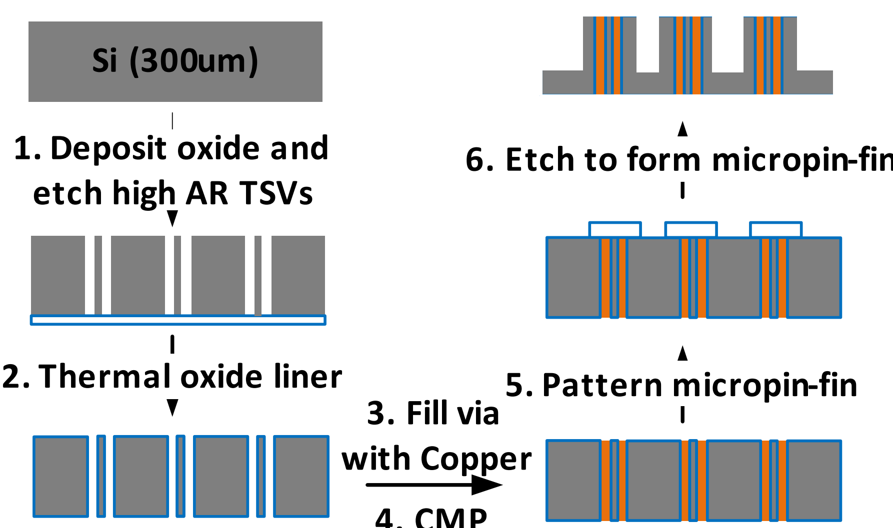
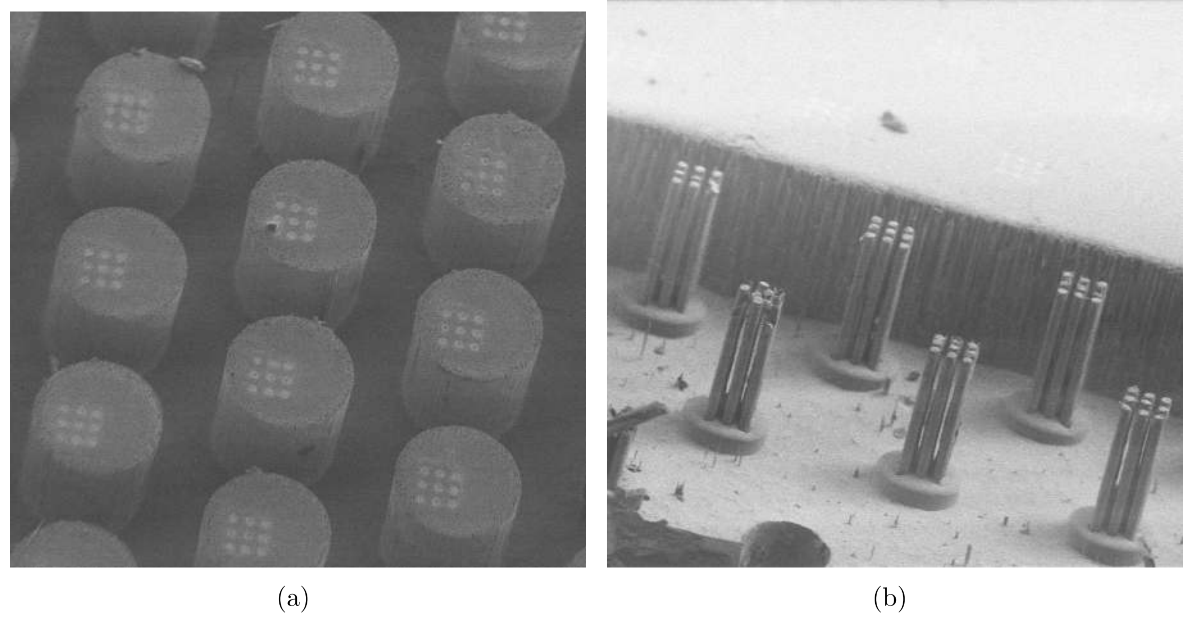
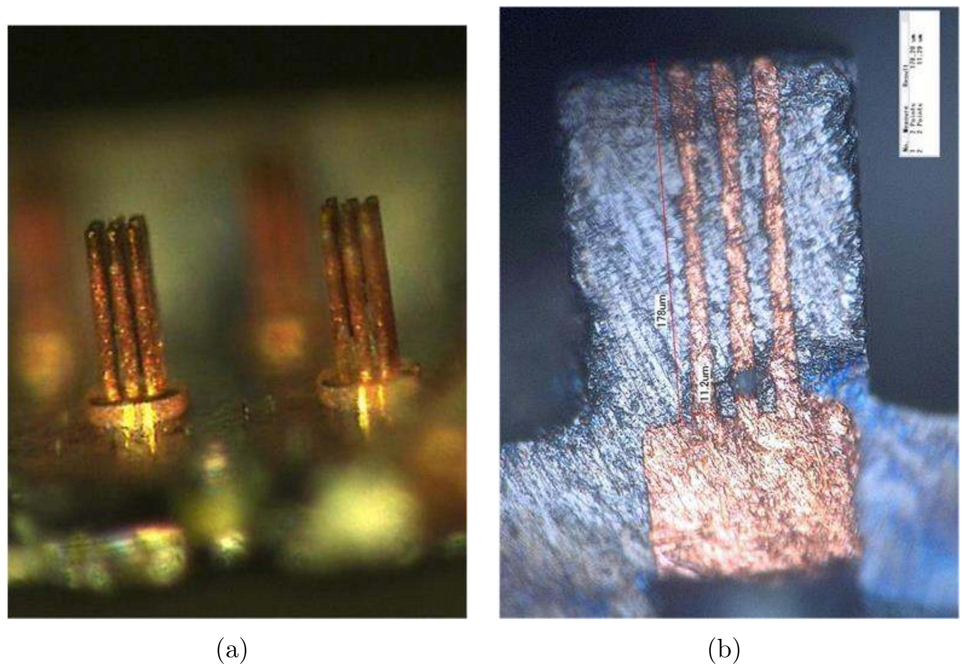


Figure 31: Process flow of high aspect ratio TSV integration into the MPFHS



(a) (b)

Figure 32: SEMs of (a) high-AR TSVs integrated in micropin-fins and (b) free standing high-AR TSVs.



(a) (b)

Figure 33: Optical images of (a) the free standing TSVs after removing the Silicon and (b) a cross section of high-AR TSVs integrated in micropin-fins.

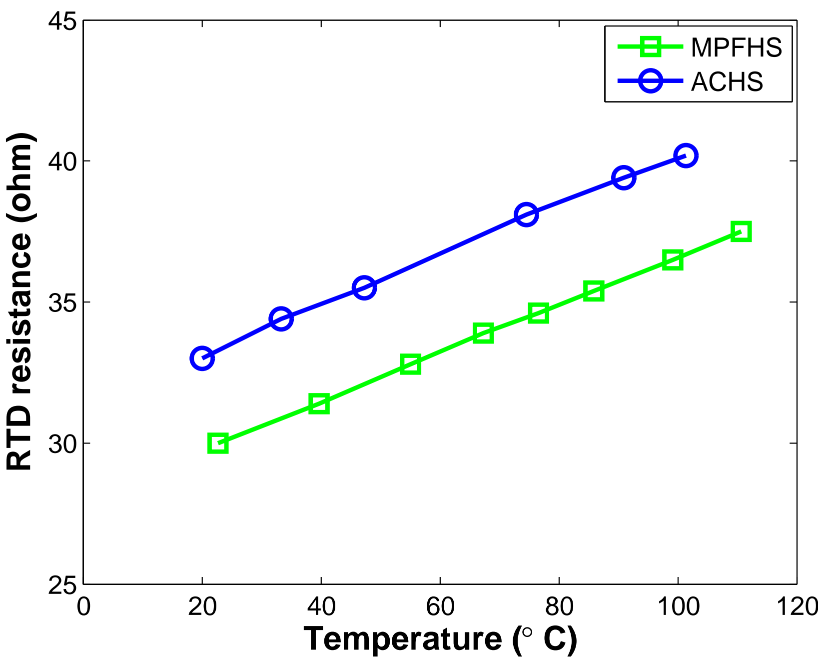


Figure 34: Characterization of a RTD's resistance as a function of temperature.

3.3 Test Setup Description and Automated Data Collection in LabVIEW

The thermal experiment begins with the characterization of the Pt RTD (shown in Figure 34). This verifies the linear resistance–temperature relationship of Pt. The relationship between the resistance and the temperature is expressed in (16),

$$R(T) = R(T_0) + \alpha R(T_0)(T - T_0) \tag{16}$$

where $R(T)$ and $R(T_0)$ are the resistance of the Pt RTD at T and T_0, respectively. α is the temperature coefficient. Based on the calibration, α of the heaters on the ACHS and the MPFHS samples is 0.00267 and 0.002864 K^{-1}, respectively. From various fabricated Pt heaters, α varies from 0.0026 to 0.0029 K^{-1}, showing good consistency.

An ACHS testbed is constructed similarly, as shown in Figure 27. The only

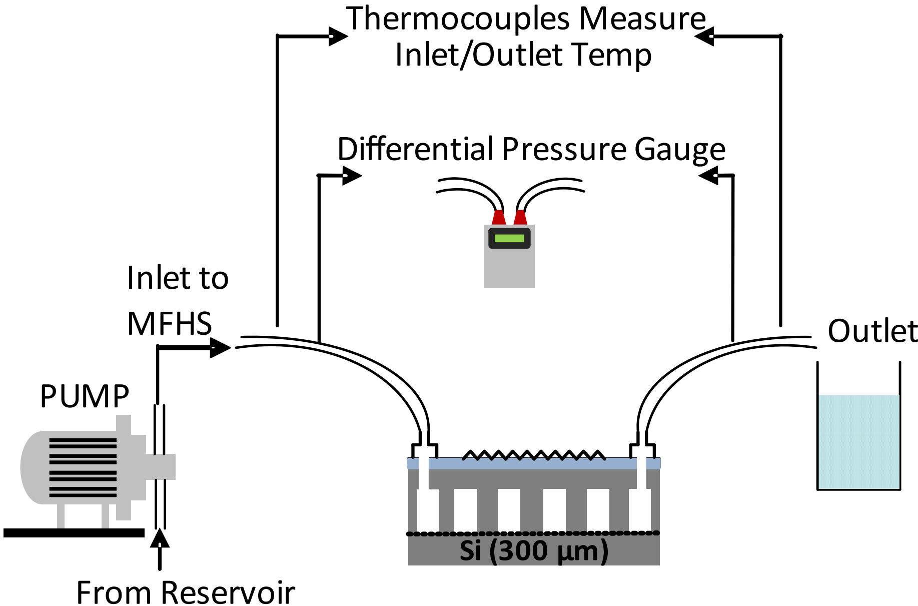


Figure 35: The experimental test setup for single-layer microfluidic heat sink testing.

difference is that there is no embedded microfluidic heat sink. This will be used to benchmark the thermal results to the MPFHS cooled chip. In the ACHS experiment, a high-performance ACHS containing three copper heat pipes and 45 aluminum fins designed for the Intel i5/i7 CPU is attached to the back side of the ACHS chip through thermal interface material (TIM). The ACHS chip is tested while the fan rotates at its maximum speed (2500 rpm $\pm$ 15%). The corresponding air flow is 54.8 CFM.

In the microfluidic heat sink experimental setup, a gear pump is connected to the inlet of the testbed in the stack. De-ionized (DI) water is pumped from a nearby reservoir. Polyester-based filters are connected to the outlet of the pump to eliminate any particles ($\geq$20 μm) that may potentially block the microfluidic heat sink. An acrylic block flow meter that measures up to 100 mL/min is connected to each inlet serially to measure the flow rate. An Agilent N6705B power analyzer with four outputs is used to source current to the thin-film Pt heaters/RTDs in order to emulate chip power dissipation. The heater resistance in each tier is measured and tracked using an Agilent 34970A data logger at 1 Hz. The measured resistance is used to

calculate the average junction temperature of the chip. Note that we use a single heater/RTD and, thus, the junction temperature represents the average junction temperature. A differential pressure gauge and J-type thermocouples are connected in parallel with the MPFHS in order to measure the pressure drop across the heat sink and the inlet/outlet temperature of DI water, respectively.

A photo of the experimental set up is included in Figure 36. The key instruments in the experiments include:

- Cole-Parmer gear pump system with 0.092 mL/rev. The flow rate ranges from 5.5 to 331.2 mL/min. The pump's physical dimensions are 7-3/4"L × 11-1/2"W × 7-1/4"H. The pump allows a remote control by voltage regulation.

- Agilent N6705B power analyzer that features four outputs. Each output allows a maximum power of 100 W. The maximum voltage and current are 60 V and 1.66 A, respectively. Thus, a resistance of ∼36.1 Ω should be designed to use the maximum power.

- Agilent 34970 data acquisition system with three slots. Each slot features a 20-channel multiplexer. There is a built-in 6.5 digit DMM for current, voltage, and temperature measurements. The DC voltage measurement accuracy is 0.004% up to 300 V. For measuring temperature using J-type thermocouples, the offset is 1 °C. For measuring resistance less than 1000 Ω, the accuracy is 0.01 % of reading +0.001% of reading range. The multiplexing rate is up to 60 channels/sec. The USB interface allows remote data logging.

- NI myDAQ data acquisition system with ±15 V output that can be used to regulate the flow rate of the pump.

- Comark C9557 differential pressure gauge with measuring range of 690 kPa. The accuracy is 0.2% of the full measurement scale (±1.38 kPa).

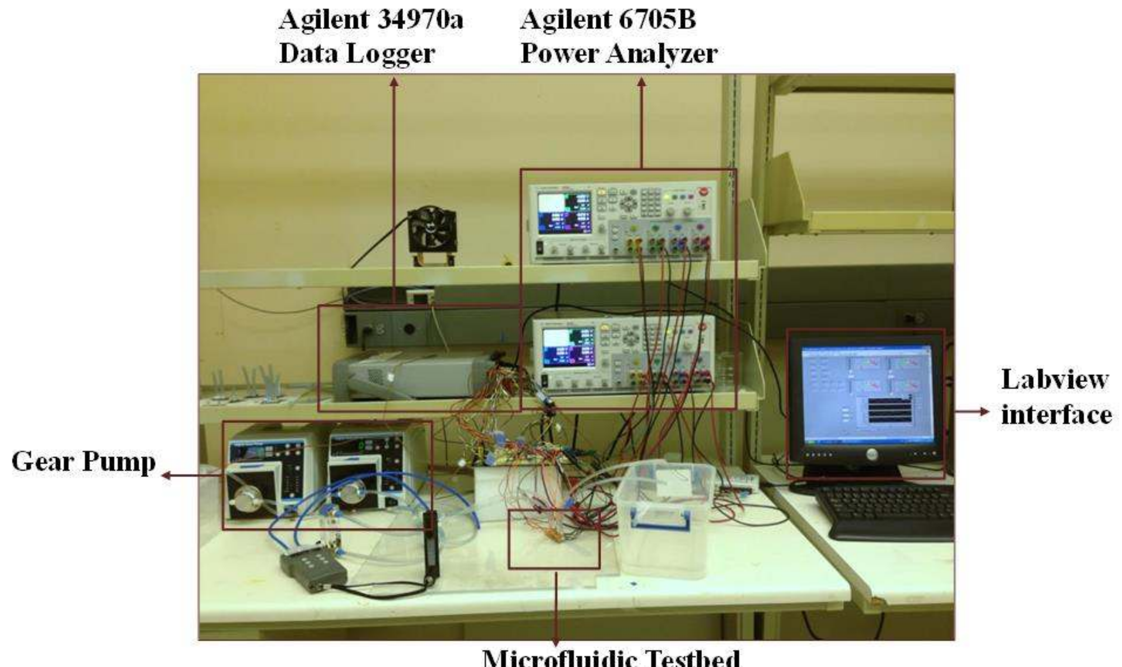


Figure 36: A photo of the test setup for microfluidic heat sink testing. The key instruments in the test setup include power analyzer, gear pumps, data logger, and a LabVIEW interface for data collection.

- KOBOLD Model KFR-2110NS acrylic flow meter with flow measuring range of 10 to 100 mL/min. The accuracy is ±5% of the full measurement range (±5 mL/min).

A labVIEW program is developed to automate the data collection. During the experiment, a fixed power density value is entered to the program. When the power source is turned on, an initial current is pumped to the on-chip heater. The data logger then measures the heater resistance and feeds it back to the program and calculates the needed current for the target power density. When power is on, the chip junction temperature will increase and causes the heater resistance to increase. The heater resistance will be measured in real time and is used to adjust the supply current.

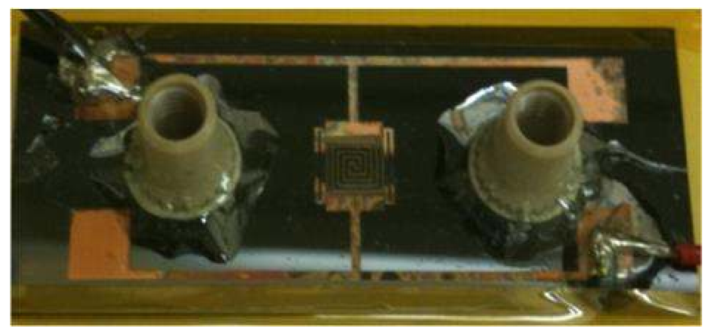


Figure 37: Photo of the microfluidic testbed.

The thermal measurements are made at two flow rates: 45 mL/min and 70 m-L/min. The heating area for both testbeds is 0.6 cm × 0.6 cm. The DI water temperature is 20 ± 1 °C and is used to compute the increase in junction temperature. To evaluate the TSVs' electrical resistance, four-point resistance measurements are performed on the high-AR TSVs. Platinum pads are deposited selectively using focused ion beam (FIB) deposition, as shown in Figure 38.

3.4 Single Layer Thermal Measurements and Benchmarked with Air-Cooled Heat Sink

The corresponding average junction temperatures as a function of power dissipation for the ACHS and the MPFHS thermal testbeds are plotted in Figure 39. The average junction temperature under the ACHS is 77.6 °C at 109 W/cm^2 for an air flow rate of 54.8 CFM. In contrast, the average junction temperature with the embedded MPFHS is 53.5 °C at 105 W/cm^2 for a flow rate of 45 mL/min and 47.9 °C at 103.4 W/cm^2 for a flow rate of 70 mL/min. It is expected that the junction temperature decreases as the flow rate increases. However, the decreasing rate will be smaller when R_{conv} and R_{cond} start to dominate. At the same power density, the chip junction temperature with MPFHS is lower than that with ACHS. The junction temperature reduction by MPFHS is more pronounced at high power densities. At a lower operating temperature, the leakage current in CMOS circuits is smaller, which results in lower power

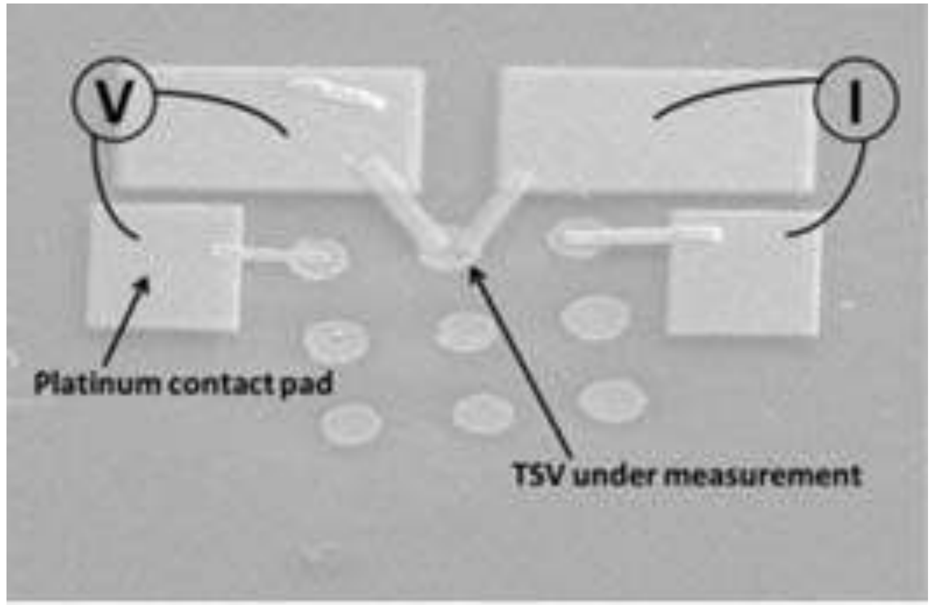


Figure 38: Illustration of the resistance measurement of TSVs using four-point technique. The platinum pads are deposited selectively using focused ion beam (FIB) deposition.

consumption. Sekar et al. [58] have shown that by reducing the chip temperature from 88 °C to 47 °C, the total power of a high-performance chip decreases from 102 W to 83 W for the same operating frequency.

The thermal resistance is normalized to 1 cm^2 area. The normalized thermal resistance of the ACHS (including the thermal resistance of the TIM layer) and the MPFHS under different flow rates is tabulated in Table 3. The thermal resistance obtained from the compact physical modeling discussed in Section 3.1.1 is also included in the table. Since the experimental design is at the edge of the validation range of the correlation-based model, there is an error of ∼20% between modeling and experimental data. Error is also induced by the non-uniform flow distribution in the testbed, as discussed in Section 4.5. The measured pressure drop is 38.5 kPa and 83 kPa for 45 mL/min and 70 mL/min, respectively. The measured pressure drop includes the pressure drop across the micropin-fin array as well as the pressure drop over the relatively long embedded leading microchannels.

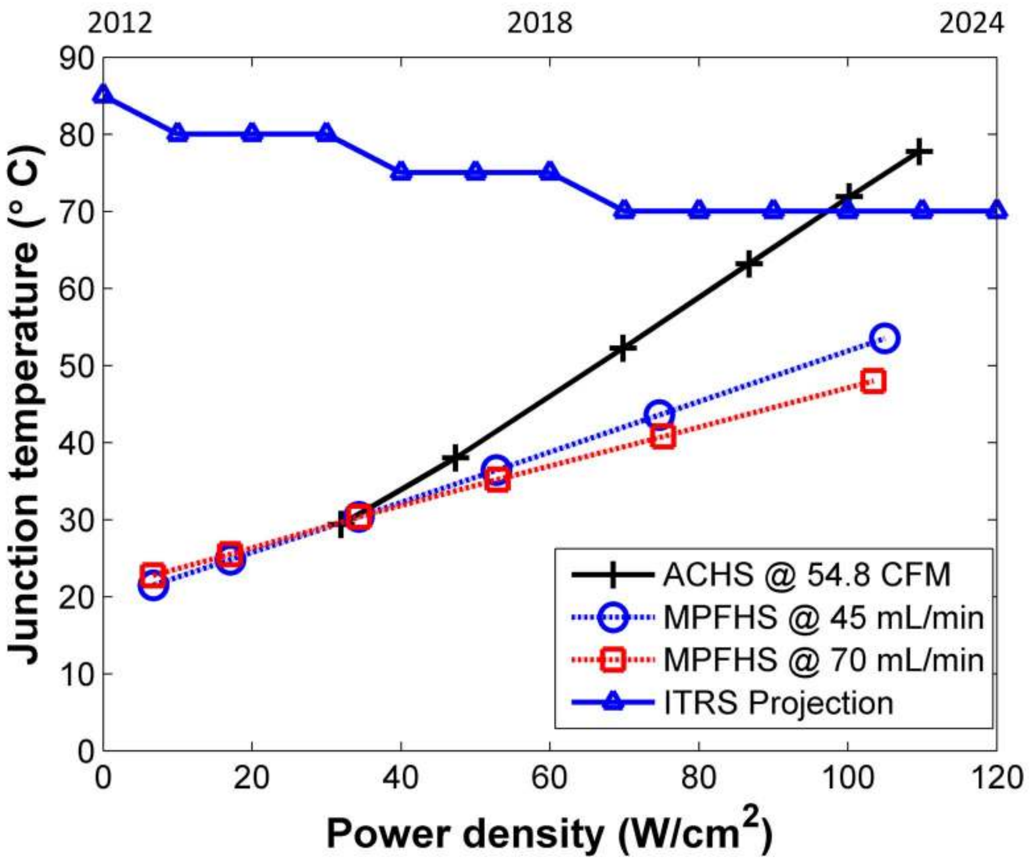


Figure 39: Average junction temperature under air cooling and microfluidic cooling compared with ITRS projections.

Four-point resistance measurements are performed. The theoretical value of resistance for the fabricated TSVs is 38 mΩ. The average measured TSV resistance is 36.5 ± 1.5 mΩ, which is close to the modeled value.

3.5 Data Extrapolations and Analysis

In this section, single-tier measurements are presented. These measurements are used to extract the main metrics of the heat sinks including the heat transfer coefficient, Nusselt number, and pressure drop.

DI water with different flow rates (37 to 110 mL/min) is pumped into the tier. The power density (P) is kept at 40 W/cm^2 for all flow rates. DI water inlet temperature (T_{in}), outlet temperature (T_{out}), and chip junction temperatures (T_j) are

Table 3: Comparison of the measured and modeled normalized thermal resistance at a power density of 100 W/cm^2

	ACHS	MPFHS@ 45 mL/min	MPFHS@ 70 mL/min
Thermal resistance (K·cm^2/W)	0.518	0.326	0.269
Power density (W/cm^2)	100.2	104.9	103.4
Modeled Rtotal (K·cm^2/W)	-	0.253	0.2
Modeling Error	-	22.4%	25.6%

Table 4: Summary of the inlet water temperature (T_{in}), the outlet water temperature (T_{out}), and the chip junction temperature (T_j) at 40 W/cm^2 for different water flow rates

Flowrate (mL/min)	T_j (°C)	T_{in} (°C)	T_{out} (°C)	R_{tot} (K/W)
37	39.9	18.7	35.2	0.54
50	36.0	18.7	31.1	0.44
60	33.9	18.7	28.7	0.39
72	32.4	18.7	27.1	0.35
85	31.4	18.7	25.9	0.32
97	30.6	18.7	25.1	0.30
110	30.0	18.7	24.4	0.29

monitored. The temperatures at different flow rates are documented in Table 4 and used to calculate the convective thermal resistance (R_{conv}) of the MFHS using 17 [8].

$$R_{conv} = \frac{(T_j - T_f)}{P} - R_{cond} \tag{17}$$

where R_{cond} represents the conductance from the circuit through the base to the heat sink interface given by (18);

$$R_{cond} = \frac{t_{base}}{k_{si} A_{base}} + \frac{t_{ox}}{k_{ox} A_{base}} \tag{18}$$

T_f is the average fluid temperature calculated by:

$$T_f = \frac{1}{2}(T_{in} + T_{out}) \tag{19}$$

R_{cond} is dependent on the thickness of the base (t_{base}) and silicon dioxide (t_{ox}), the area of the base (A_{base}), and the thermal conductivity of silicon (k_{si}) and silicon dioxide (k_{ox}). It is a constant throughout the experiments and the value is calculated to be 0.05 K/W. Figure 40 plots R_{conv} as a function of the flow rates. The heat transfer coefficient (h) is derived from using R_{conv} and the effective total heat transfer area (A_t) by 20,

$$h = \frac{1}{R_{conv} A_t} \tag{20}$$

A_t is calculated by equation (21)

$$A_t = A_b + \eta A_{fin} \tag{21}$$

where A_b is the base area exposed to the fluid, η is the fin efficiency and is a function of the micropin-fin height (H_{fin}) and diameter (D) given by (22), and A_{fin} is the aggregate surface area of the micropin-fins exposed to the fluid [59].

$$\eta = \frac{\tanh(2H_{fin}\sqrt{h_{ave}/k_{si}D})}{2H_{fin}\sqrt{h_{ave}/k_{si}D}} \tag{22}$$

Furthermore, the Nusselt number (Nu) as a function of Reynolds number (Re) is plotted in Figure 41. The Nusselt number and Reynolds number are calculated using (23) and (24), respectively.

$$Nu = \frac{h}{D_h \cdot k} \tag{23}$$

$$Re = \frac{V_{max} D_h}{\nu} \tag{24}$$

where D_h represents the hydraulic diameter and is calculated as follows:

$$D_h = \frac{2H_{fin} w_c}{H_{fin} + w_c} \tag{25}$$

$$V_{max} = \frac{Q}{H_{fin}(W - n \cdot D)} \tag{26}$$

where k and ν are the thermal conductivity and the kinematic viscosity of the fluid, respectively. The hydraulic diameter (D_h) is calculated using (25), where w_c is the

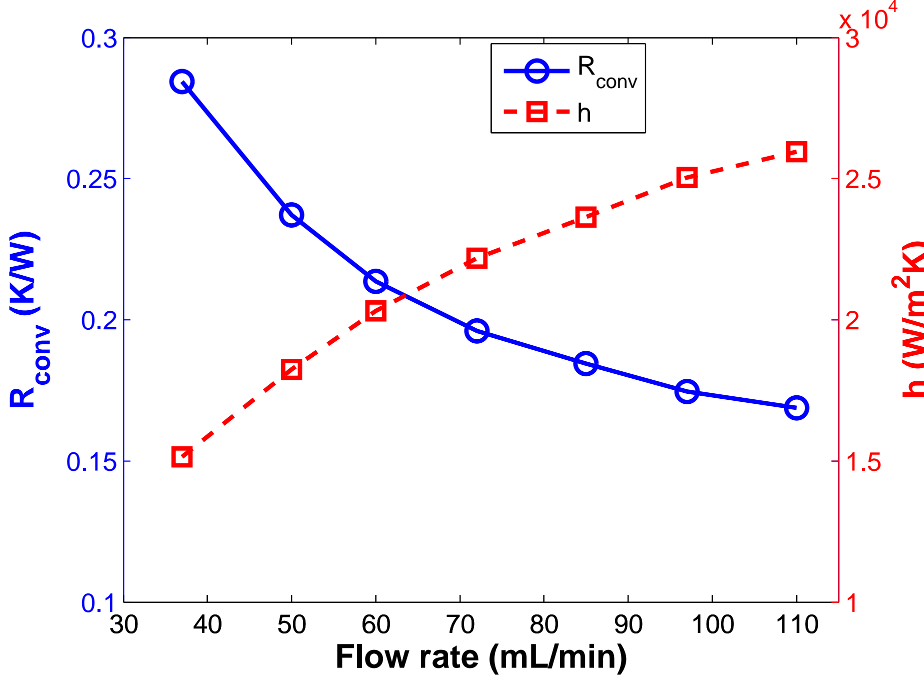


Figure 40: Convective thermal resistance and heat transfer coefficient as a function of the flow rate.

diagonal pitch (shown in Figure 42) [8]. The maximum velocity (V_{max}) crossing the minimum cross-section is calculated using (26), where Q is the volumetric velocity, W is the width of the micropin-fin heat sink, and n is the number of the micropin-fins in the vertical direction (shown in Figure 42). Although increasing flow rate (Reynolds number) decreases R_{conv} and increases h, the pressure drop (ΔP) will increase, which is not desirable in electronic systems since it increases the pumping power and may introduce reliability issues. The measured pressure-drop data for different Reynolds numbers is also plotted in Figure 41 in order to show the trade-off between the heat transfer characteristics and the pressure drop.

It is expected that the junction temperature decreases as the flow rate increases. However, the decreasing rate will be smaller when R_{conv} and R_{cond} start to dominate. Continuing to increase the flow rate will bring less benefit in reducing the total thermal resistance. However, the pressure drop keeps increasing in an exponential fashion as

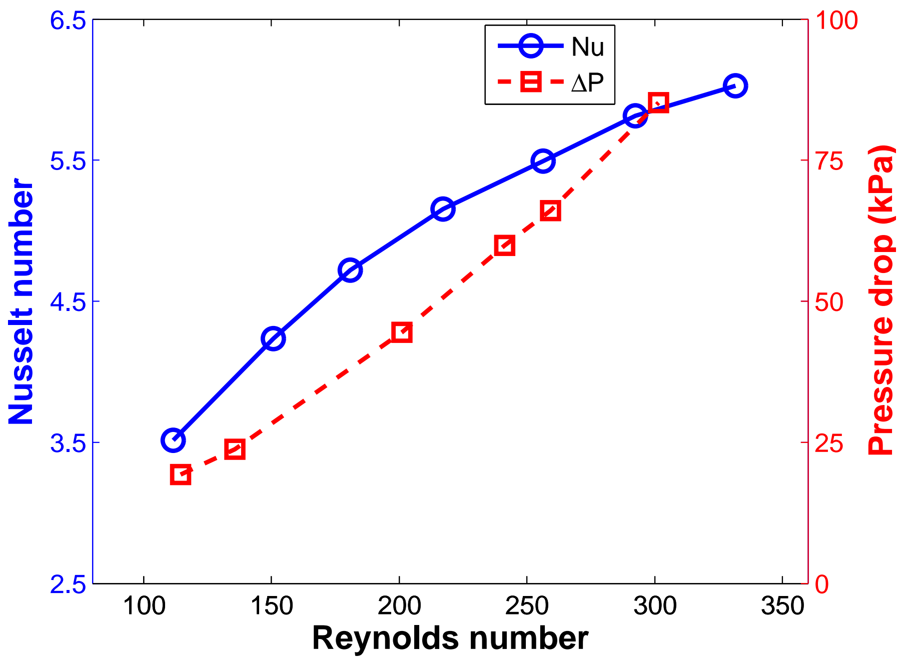


Figure 41: Nusselt number and pressure drop as a function of Reynolds number.

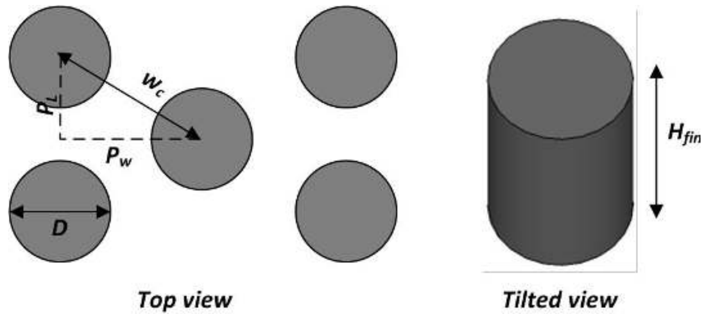


Figure 42: Micropin-fin layout and dimensions (top and tilted view).

the flow rate increases and will soon reach the system limit. Thus, it is critical to choose the flow rate in order to provide sufficient cooling capability with an acceptable pressure drop.

3.6 Conclusion

This chapter presents an experimental demonstration of a TSV–compatible microfluidic heat sink for high-power and high-performance chips. This is a solution that addresses the cooling needs of 3D ICs while accounting for TSV fabrication compatibility and electrical performance (minimizing TSV parasitics). In the test case, a staggered MPFHS is shown to provide a thermal resistance as low as 0.269 K·cm^2/W at a flow rate of 70 mL/min for a heat sink height of 200 μm. In addition, this result is benchmarked against a state-of-art air cooled heat sink. Based on the experimental data, microfluidic cooling provides lower chip junction temperature with a much smaller heat sink volume compared to air cooling. Finally, in order to demonstrate the compatibility with TSVs, high aspect-ratio (18:1) TSVs are integrated in MPFHS. The four-point resistance of a single TSV is found to be 36.5 $\pm$ 1.5 mΩ.

CHAPTER IV

TIER-SPECIFIC MICROFLUIDIC COOLING EVALUATION IN 3D IC STACKS

4.1 Introduction

In the previous chapter, single-layer thermal measurement has been performed for MPFHS. Microfluidic cooling is shown to maintain the chip at a lower junction temperature compared to an air-cooled heat sink. This benefit can be even more significant when implementing microfluidic cooling in 3D ICs. The biggest advantage of microfluidic cooling is its chip scale. In a multiple-tier chip stack, microfluidic cooling allows each layer to have its own heat sink. However, all the tiers need to share one air-cooled heat sink because of its large form factor. To quantify the benefits, microfluidic cooling is evaluated in 3D ICs in this chapter.

Microfluidic cooling has been implemented is 3D chip stacks in [14] and [15]. In [14], a four-tier stack is built where a microfluidic heat sink in integrated into each tier. Heat removal of 390 W was shown with a junction temperature rise of 54.7 °C and a pressure drop of 100 kPa, respectively. Figure 43 (a) illustrates the proposed stack with microfluidic heat sink in literature. One set of inlet and outlet is adopted in this work. With this approach, it is not possible to control or tailor the flow rate in each tier. However, in a realistic 3D stack, especially in a heterogeneous stack, the power dissipation in each tier may be different (workload dependent). Thus, one needs the capability to control the coolant flow rate in each tier independently. Even more, there is likely a need to control the flow rate locally within a single tier, as discussed later in the chapter. To address this need, wafer-level batch fabricated solder microfluidic chip I/Os and fine pitch electrical microbump I/Os have been demonstrated, as shown

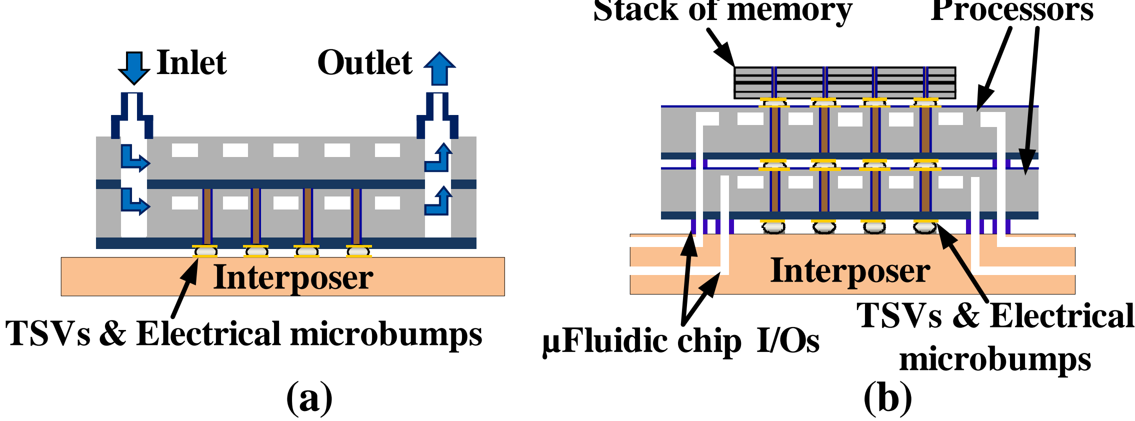


Figure 43: Prototype of (a) a general embedded microfluidic heat sink and (b) our tier-specific microfluidic cooling within a 3D stack.

in Figure 44 [16]. Based on this innovative chip I/O technology, this chapter proposes and experimentally implements tier-specific embedded microfluidic cooling in a two-tier stack (Figure 43 (b)). The proposed 3D IC stack features a silicon interposer with embedded fluidic delivery microchannels and an array of 3D stacked processor and memory tiers. Each processor tier contains an embedded MFHS. TSVs are routed through the integrated MFHS. Each tier has its dedicated microfluidic chip I/Os, that are formed using solder for fluid delivery from the interposer. The coolant flow rate in each tier can be tailored independently, according to the heat dissipation of each tier, i.e. tier-specific cooling. This approach helps minimize the vertical thermal gradient across the stack when power dissipation varies in the stack. Pumping power may be reduced by adjusting the flow rate to the needed value for a given power dissipation per tier. The proposed local coolant delivery mechanism, which is also based on the solder chip I/O technology (discussed later in this chapter), may minimize the lateral thermal gradient within a single tier, as well.

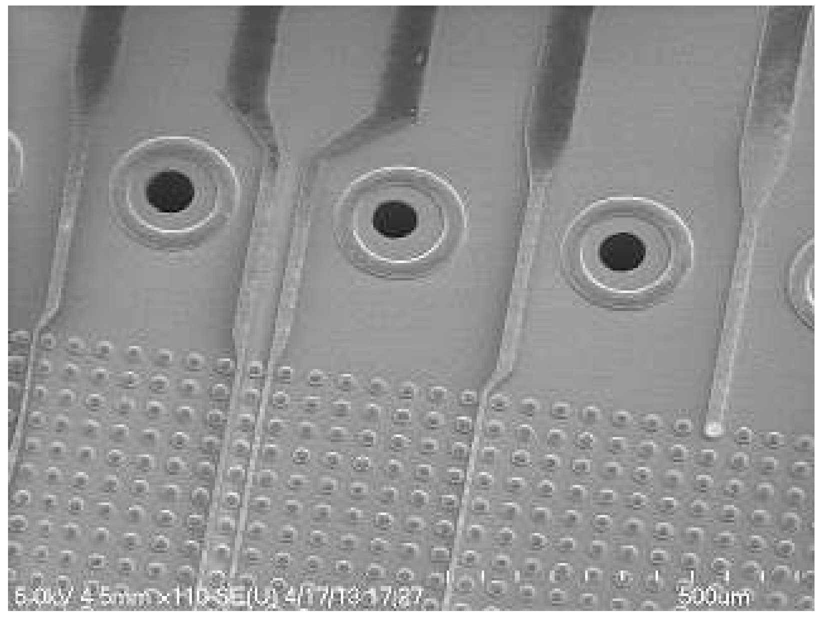


Figure 44: SEM of solder microfluidic chip I/Os and electric microbumps.

4.2 Thermal Testbed Preparation and Experimental Setup

Section 3.1 describes the process flow to prepare a single-tier microfluidic testbed. In this section, two of these testbeds are stacked orthogonally with a thermal interface material (TIM) to form a 3D thermal testbed. The thermal resistance of the TIM is 0.25–0.28 K/W (depending on the pressure applied during the experiments). The total heating area of each tier is 1 cm $\times$ 1 cm. The 3D testbed is illustrated in the experimental setup, as shown in Figure 45. Each tier has its own power supply, pump, flow meter, microfilter, and thermal couples. For the sake of simplified port access, the two tiers are stacked orthogonally such that the inlets and outlets are easily accessible (Figure 45). To attain an initial insight into the benefits of the embedded microfluidic cooling, a 3D ACHS testbed is constructed similarly without the embedded microfluidic heat sink. The same ACHS is used as in the single tier measurement reported previously.

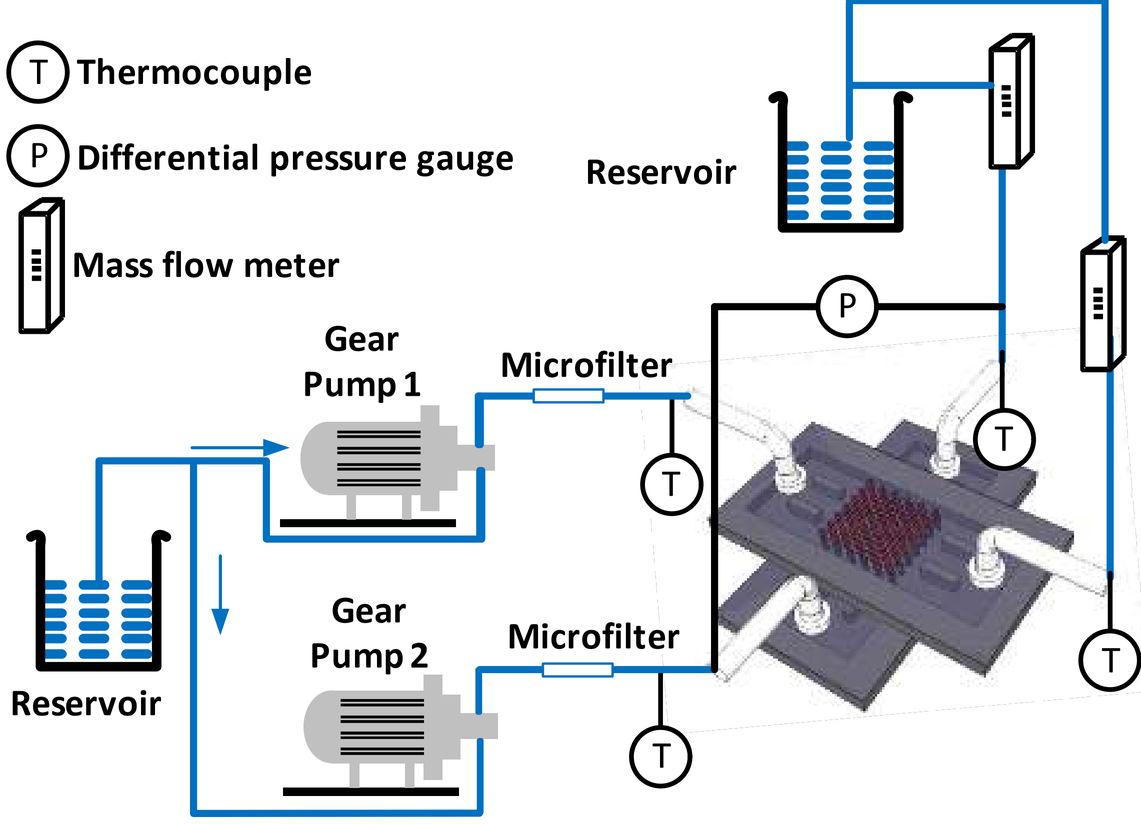


Figure 45: Experimental setup for microfluidic heat sink evaluation in 3D stacks.

4.3 Tier-Specific Microfluidic Cooling for Different Stacking Scenarios

4.3.1 Processor-on-Processor and Memory-on-Processor Stack

In Figure 46, the fluid is pumped only into the processor tier of the memory–processor stack at a flow rate of 100 ± 5 mL/min. In this experiment, the power density of the memory chip is held at 5 W/cm^2. Since the memory tier is stacked on the processor tier with integrated microfluidic cooled heat sink, the microfluidic heat sink serves as a path for cooling of the memory tier as well. The junction temperature rise of the memory and processor tiers are 15.3 °C and 28.7 °C, respectively, when the power density of the processor tier is 99.2 W/cm^2. As a comparison, a memory–processor stack is tested under ACHS (Figure 46(b) and 46(c)). For the case where memory is placed close to the ACHS (Figure 46(b)), the junction temperature rises of the memory and processor tier are 30.6 °C and 59.3 °C, respectively, when the power density of the processor is 49.3 W/cm^2. For the case where the processor is placed close to the ACHS (Figure 46(c)), the junction temperature rises of the memory and processor tier are 39.0 °C and 41.1 °C when the power density of the processor tier is 57.1 W/cm^2. For the same power density, the absolute junction temperatures of the chips under microfluidic heat sink are lower than those under ACHS by at least 12.0 °C and by 48.0 °C in the worst case. In the ACHS experiments, because of the over-heating of the chips, the power densities of the two tiers are limited to below 60 W/cm^2.

In Figure 47, the two-tier chip stack dissipates up to 100 W/cm^2 per tier to emulate the stacking of processors. A microfluidic heat sink is integrated into each tier. The flow rate in each tier is 100 mL/min. The junction temperature increase above the inlet coolant temperature in each tier is plotted in Figure 47. As seen from the plots, when the power density in each tier is 100 W/cm^2, the junction temperature rise in either tier is 30 °C. In contrast, the testbed under ACHS has a temperature

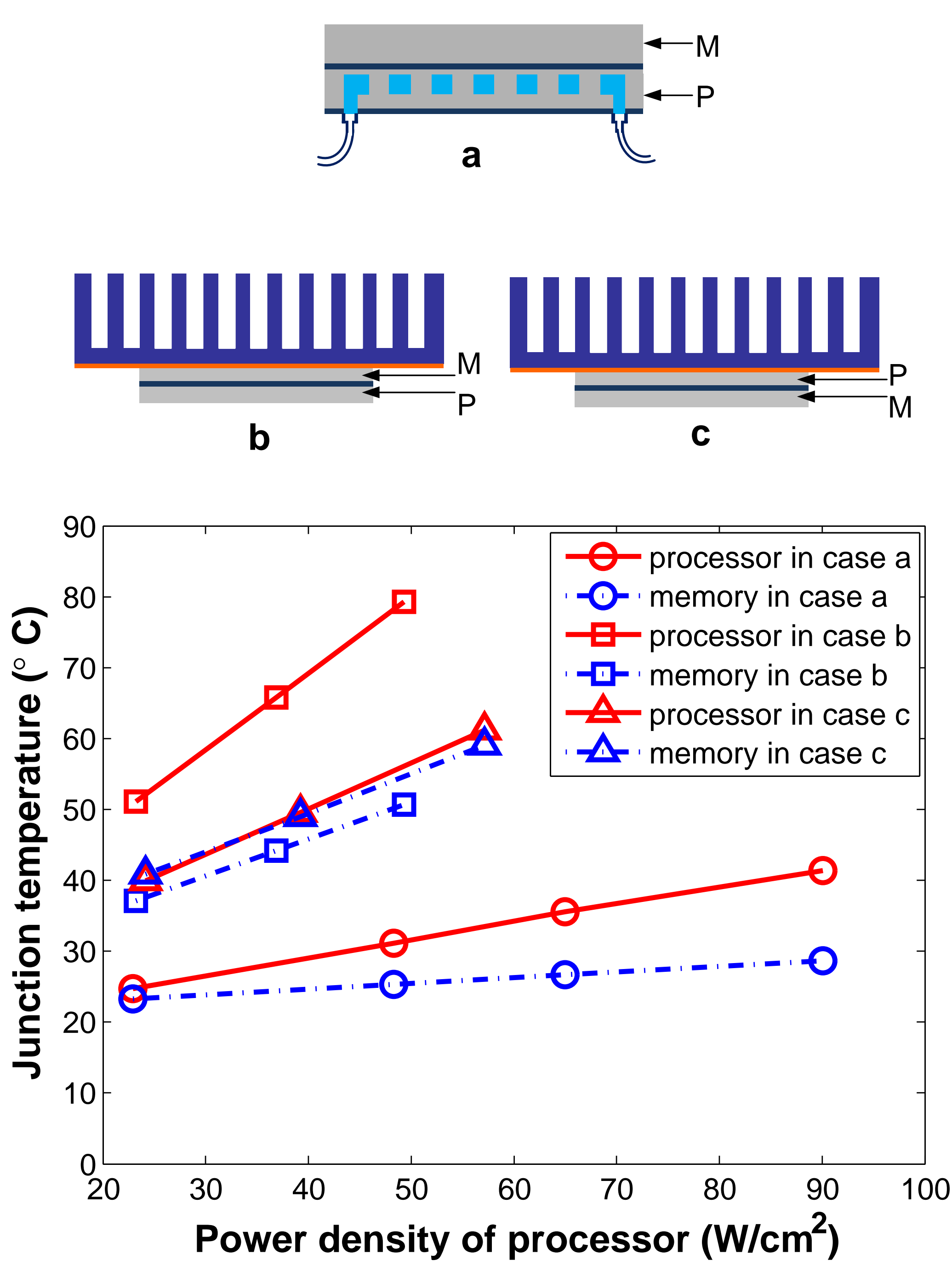


Figure 46: Junction temperature rise in a memory–processor stack under microfluidic heat sink and ACHS.

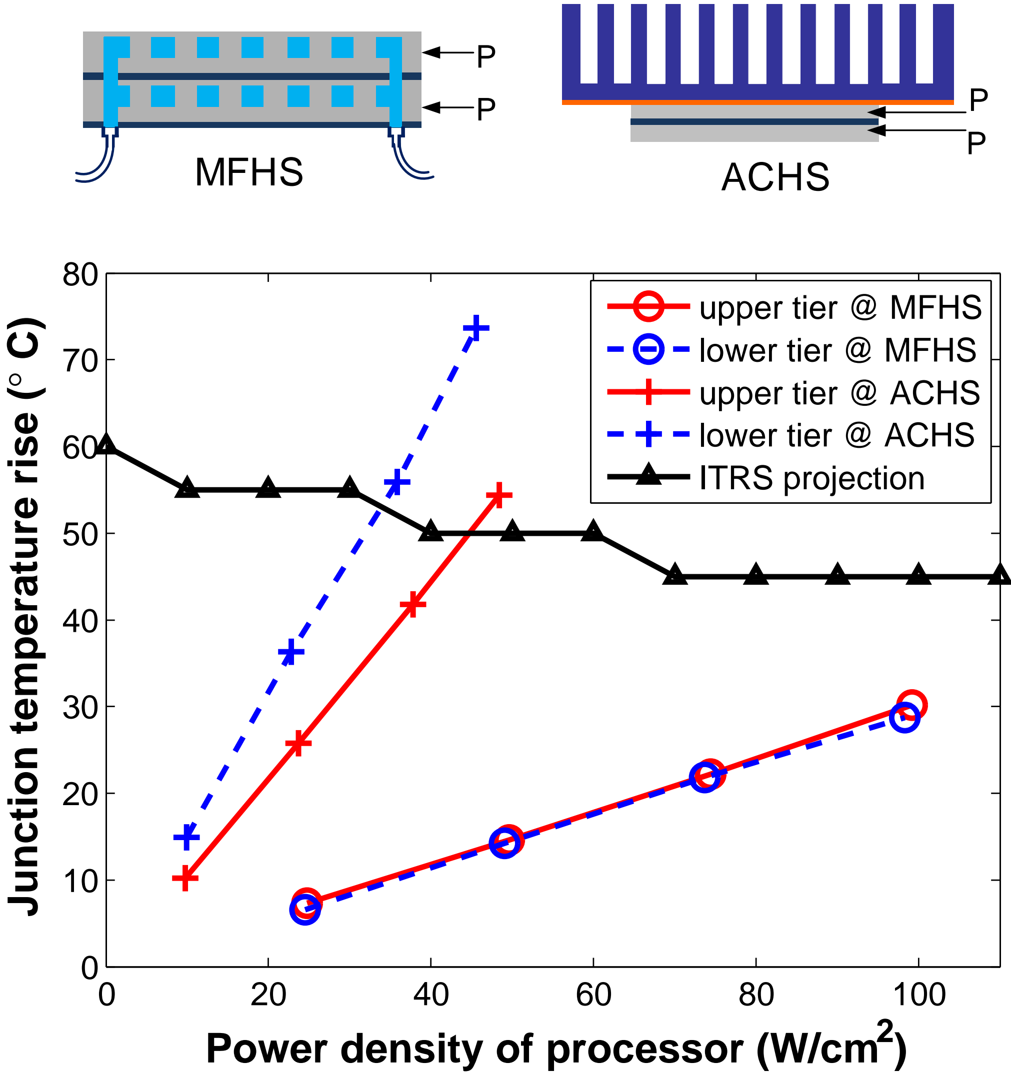


Figure 47: Junction temperature rise in a processor–processor stack under microfluidic heat sink and ACHS.

rise of more than 54 °C at 50 W/cm^2. The maximum junction temperature rise trend according to ITRS is also plotted in Figure 47 as a reference. The processor-on-processor stack cooled using a microfluidic heat sink can dissipate more than 100 W/cm^2 in each tier without reaching the ITRS projected maximum junction temperature.

4.3.2 Tier-Specific Flow Rates in ICs with Different Power Dissipations

In previous studies, the coolant is injected into the stack through one common inlet and is distributed into each tier. Thus, one cannot control the distribution (flow rate) of the coolant in each tier. However, in a realistic 3D stack with heterogeneous elements, one needs to control the coolant flow rate in each tier independently. For example, coolant may be supplied into the processor tier in a memory-on-processor stack, or coolants with different flow rates may be supplied to each tier in a two-processor stack with different workloads (and thus different power dissipations). For the first time, we propose and implement tier-specific interlayer microfluidic cooling in different tiers for heterogeneous 3D IC applications. This approach helps mitigate the vertical thermal gradient in a heterogeneous 3D stack, lowering thermomechanical stress as well as minimizing thermally induced variations in the stack [60]. Additionally, adjusting the flow rate according to the power dissipation saves pumping power by preventing over cooling of the low-power die.

A test case is evaluated using the existing 3D thermal testbed. In this test case, a 3D stack of two high-power tiers with different power densities is evaluated: 50 W/cm^2 (P1) and 100 W/cm^2 (P2). The tier-specific flow rate (and thus cooling) mechanism is implemented (Figure 7). As shown in Figure 48, when each tier in the stack is initially cooled under the same flow rate (Q1 = Q2 = 45 mL/min), the average junction temperature of P1 and P2 is 49.5 °C and 60.6 °C, respectively. Next, the flow rate of each tier is varied independently so that the junction temperature of the

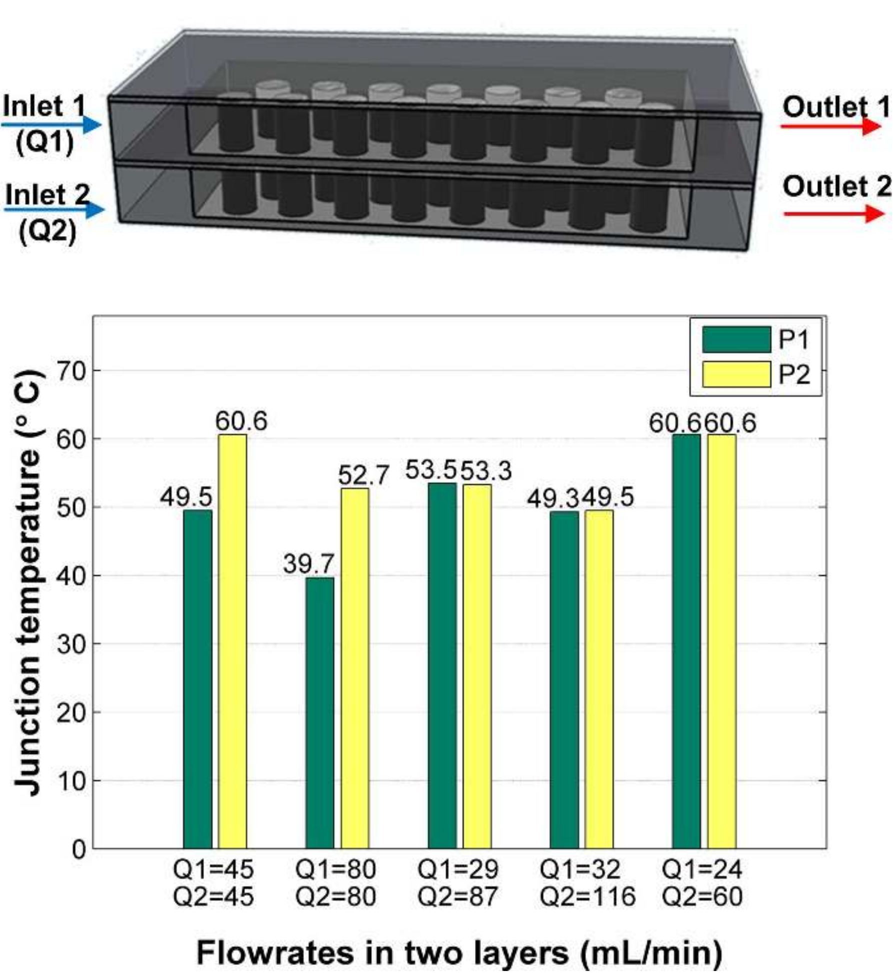


Figure 48: Junction temperature of the top layer (P1) and the bottom layer (P2) as a function of the flow rates.

two tiers is equalized at the higher and lower ends. For example, in the case where flow rates Q1 and Q2 are 32 mL/min and 116 mL/min, respectively, the junction temperature of the two tiers is equalized at approximately 49.5 °C. By mitigating the thermal gradient of the two tiers, thermomechanical stress and thermally induced variations are lowered. Additionally, when an operating temperature is specified, adjusting the flow rate according to the power dissipation saves pumping power by preventing over-cooling. Considering the conventional microfluidic delivery method (i.e., Figure 6) in which the flow rate in each tier has to be identical, the total flow rate is chosen based on the thermal needs of the tier with the highest power. The conventional method is emulated as the second set of flow rates in Figure 48. For example, for an operating temperature of 53 °C, Q1 and Q2 need to be 80 mL/min in order to maintain both tiers at a temperature lower than 53 °C. In our tier-specific cooling (the third set of flow rates in Figure 48), the needed Q1 and Q2 are 29 mL/min and 87 mL/min, respectively. The pressure drops at 29 mL/min, 80 mL/min, and 87 mL/min are measured to be 12 kPa, 60 kPa, and 67.9 kPa, respectively.

Pumping power is expressed in (27).

$$P_{pump} = Q \times \Delta P \tag{27}$$

where Q is the volumetric flow rate and ΔP is the pressure drop. As a result, using a tier-specific flow rate, the pumping power is reduced by 37.5 % relative to the conventional fluidic delivery method.

4.4 *Microfluidic Cooling in Multi-core Processor Stacking*

Multi-core processors have been adopted by main chip makers include Intel, IBM, and AMD since 2005 in servers, desktops, and laptop because multi-core processing improves performance while saving power [61]. As shown in Figure 49, multi-core processors outperform single-core processors when running SPECint2000 and SPECfp2000 benchmarks. According to Intel, "multicore chips' relative advantage will increase during the next few years" [61].

An image of an Intel Core i7 processor is shown in Figure 50. As can be seen, the four cores are placed side by side. The thermal testbed emulating the quad core processor is prepared and used for evaluating microfluidic cooling.

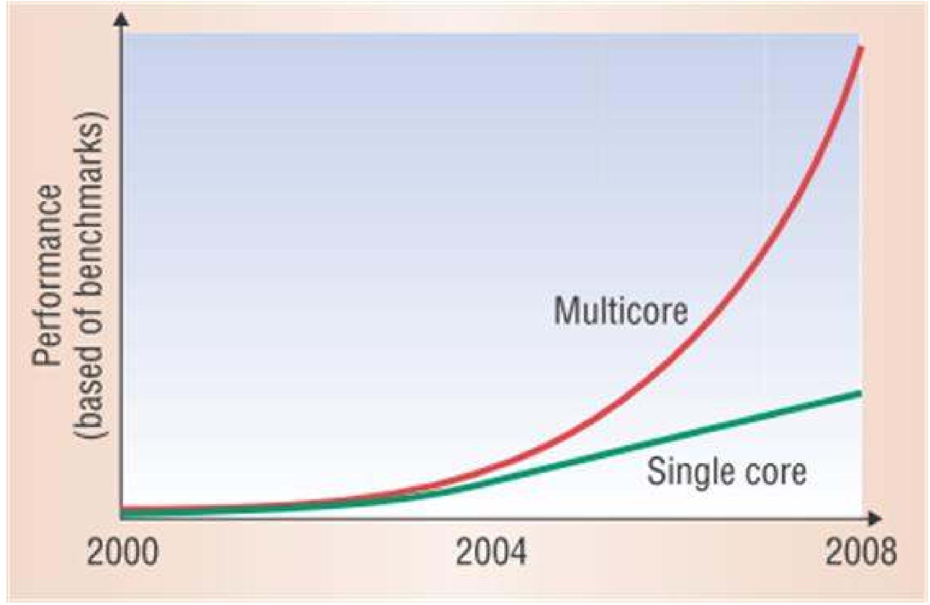


Figure 49: Performance of multi-core processor compared with single-core processor [61].

4.4.1 Preparation of the Thermal Testbed and Experimental Test Setup

In this section, two of the single-tier testbeds (prepared as shown in Figure 27) are stacked in parallel with TIM to perform 3D thermal testing. It is well known that coolant temperature increases as it passes through the microfluidic heat sink, and thus

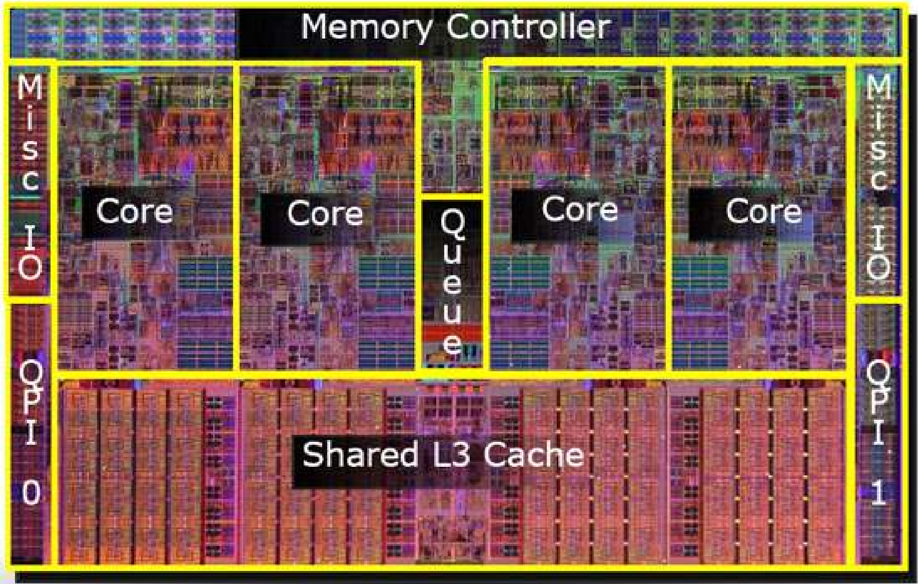


Figure 50: Intel Core i7 Processor.

the chip temperature will increase. To capture the lateral chip temperature gradient and to emulate the stacking of multi-core processors, four segmented Pt heaters are deposited along the flow direction. An image of the bonded two-tier testbed is shown in Figure 51 (a). The bottom chip is slightly larger than the top chip to facilitate fluidic port access. The four heaters are controlled independently to emulate the on and off of the quad cores. The dimensions of each heater are 0.22 cm $\times$ 1 cm with a spacing of 0.03 cm (Figure 51 (b)). The total heating area of each tier is 1 cm $\times$ 1 cm. Figure 52 shows a schematic of the assembled two-tier testbed. The experimental setup is similar to that shown in Figure 45 with the difference being the two tiers are stacked in parallel.

4.4.2 Lateral Thermal Gradient

To capture the lateral temperature increase as coolant flows from the inlet to the outlet, a single-tier measurement is performed. Figure 53 illustrates the temperature of each heater on the chip as the total chip power density ramps from 25 W/cm^2 to 100 W/cm^2. The DI water flow rate is 80 mL/min in all of the measurements. In

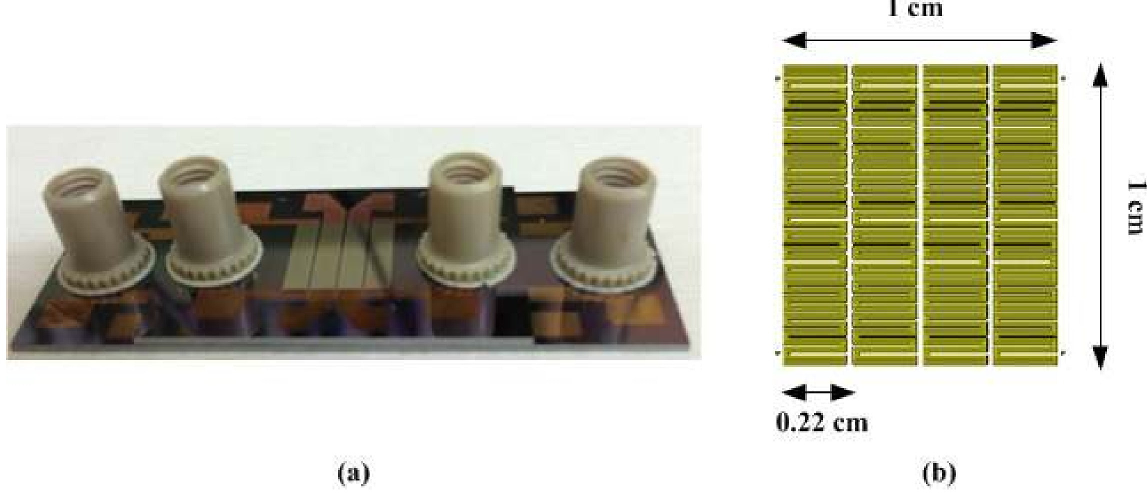


Figure 51: (a) Image of the bonded two-tier thermal testbed and (b) layout of the four heaters.

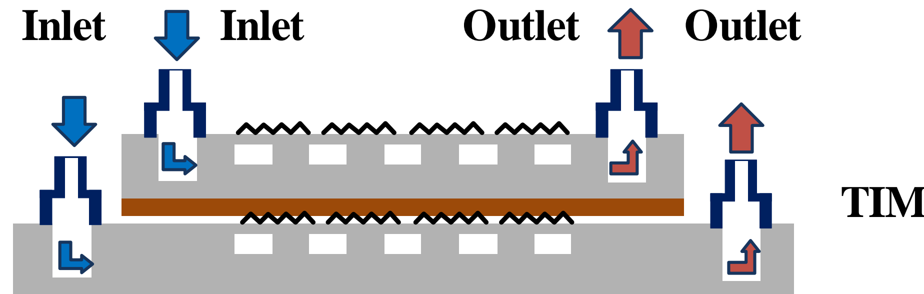


Figure 52: Schematic of the tier-specific fluidic delivery mechanism.

the high power density case (100 W/cm^2), the junction temperature of heater 4 (i.e., the heater closest to the outlet) increases by 33 °C while that of heater 1 increases by only 17 °C [62]. This result is expected since the coolant temperature increases as it flows from the inlet to the outlet and, thus, the chip junction temperature also increases. The chip design was simulated using ANSYS Fluent at a power density of 100 W/cm^2. Since the microfluidic heat sink structure is symmetric, only half of the micropin-fin array is modeled. Figure 54 and Figure 55 show the temperature maps of the base and coolant, respectively. Average junction temperatures are extracted from the simulation results and are also plotted in Figure 53. The difference between

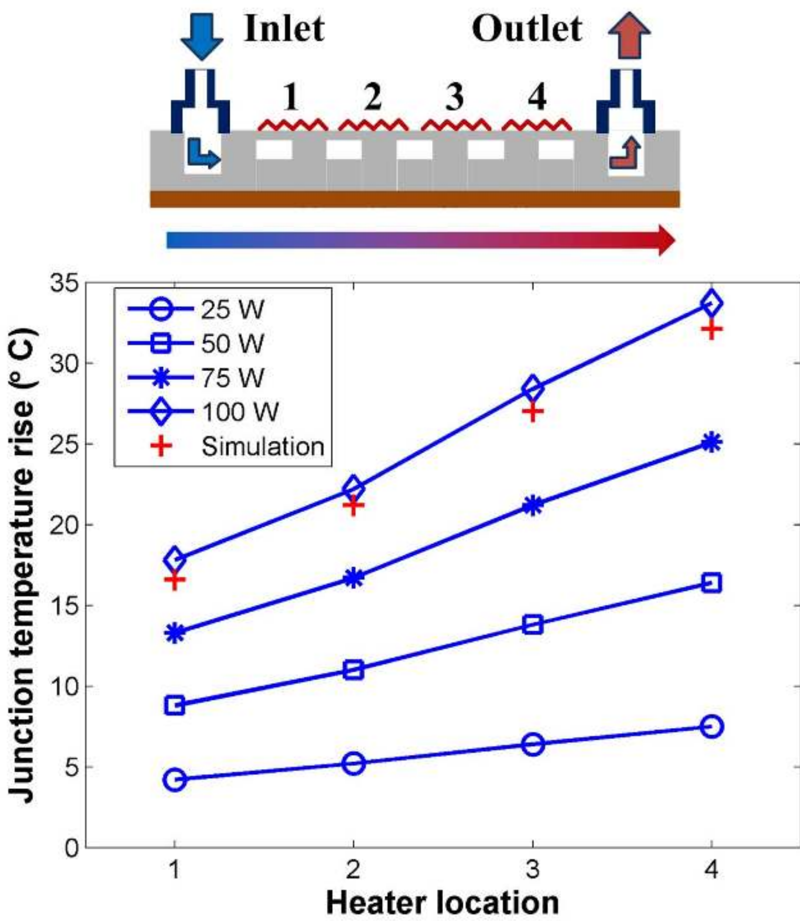


Figure 53: Junction temperature rise at different heater locations on the chip for different power dissipations. ANSYS simulation for 100 W case is also plotted for reference.

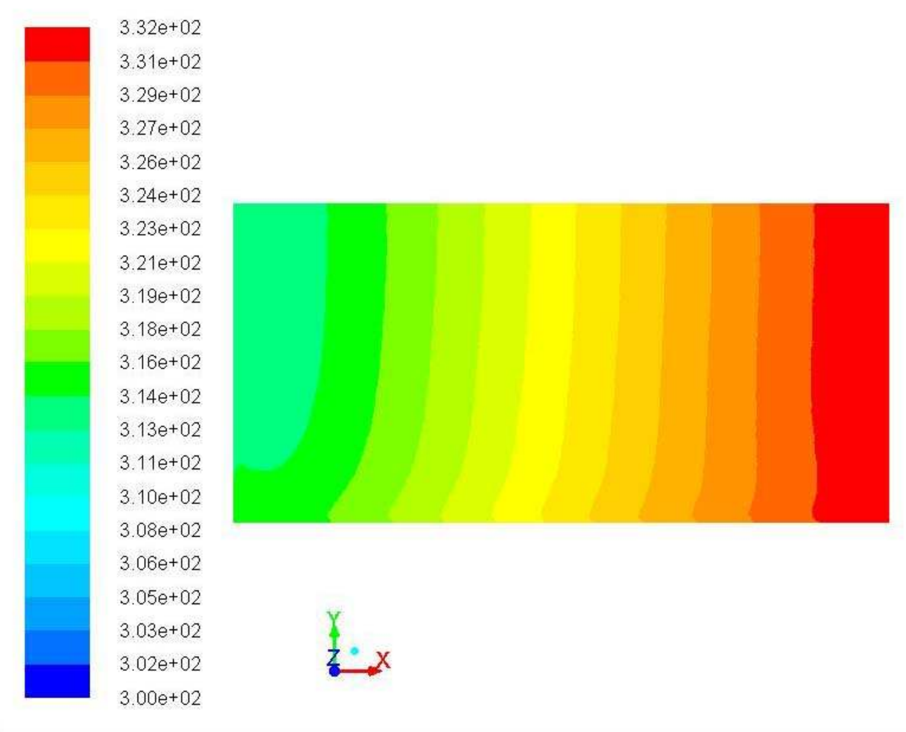


Figure 54: Base temperature map in ANSYS simulation while the chip dissipates 100 W/cm^2.

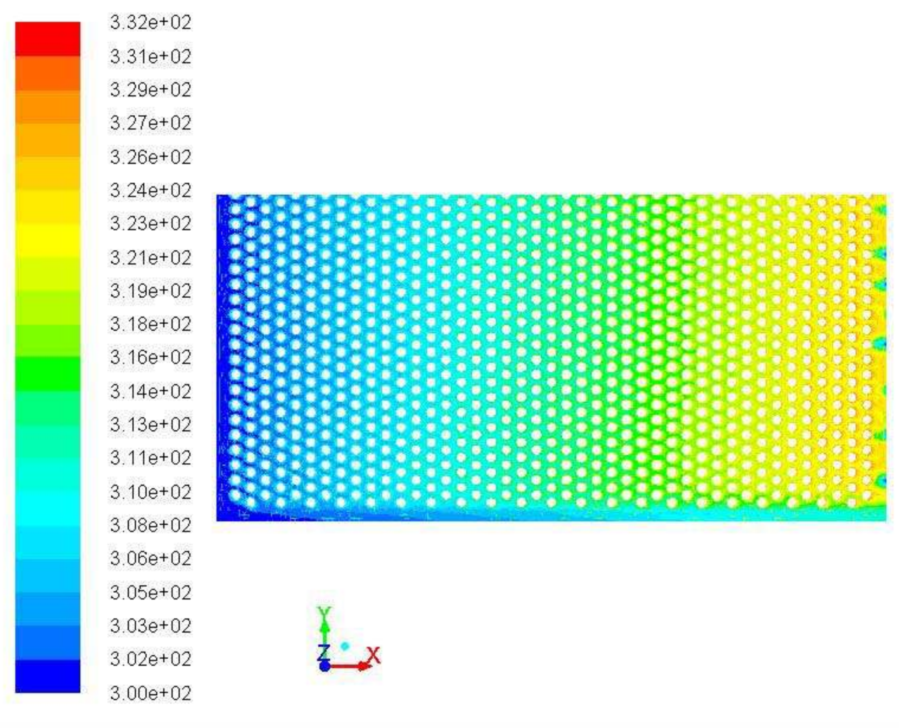


Figure 55: Water temperature map in ANSYS simulation while the chip dissipates 100 W/cm^2.

the experimental results and the simulations is less than 1.6 °C. More details of the ANSYS simulations will be included in Section 4.5. The lateral thermal gradient across the chip becomes exacerbated for higher power densities. One way to mitigate the thermal gradient is to increase the flow rate with the penalty of increased pressure drop and pumping power.

4.4.3 Electrical Implications Due to Lateral Thermal Gradient

The power consumption and performance of CMOS circuits depend on the operating temperature. In this section, a first order quantitative analysis of the leakage power is conducted based on the lateral thermal gradient experimental results. The power consumption of CMOS VLSI circuitry has three components as described in Eq. (28).

$$P_{tot} = P_{dynamic} + P_{static} + P_{sc} \tag{28}$$

where $P_{dynamic}$ represents the dynamic power, P_{static} represents the static power, and P_{sc} represents the short circuit power. The short circuit power is small compared to the other two terms and is neglected in this analysis.

The leakage power represents the power consumption due to leakage current (I_{leak}). The main contributor to leakage current is the subthreshold current [63]. Subthreshold current is the current conduction between source and drain when the transistor is in the subthreshold region. This current has historically been very small in the off state. However, owing to the ongoing voltage scaling with transistor size scaling, the threshold voltage has become small enough that the subthreshold current becomes significant. The leakage power is strongly dependent on the chip temperature. Figure 56 shows the percentage of leakage power and dynamic power at different chip temperatures for an Intel chip fabricated using 100 nm technology [64]. Eq. (29) depicts the subthreshold current I_{sub} as a function of temperature.

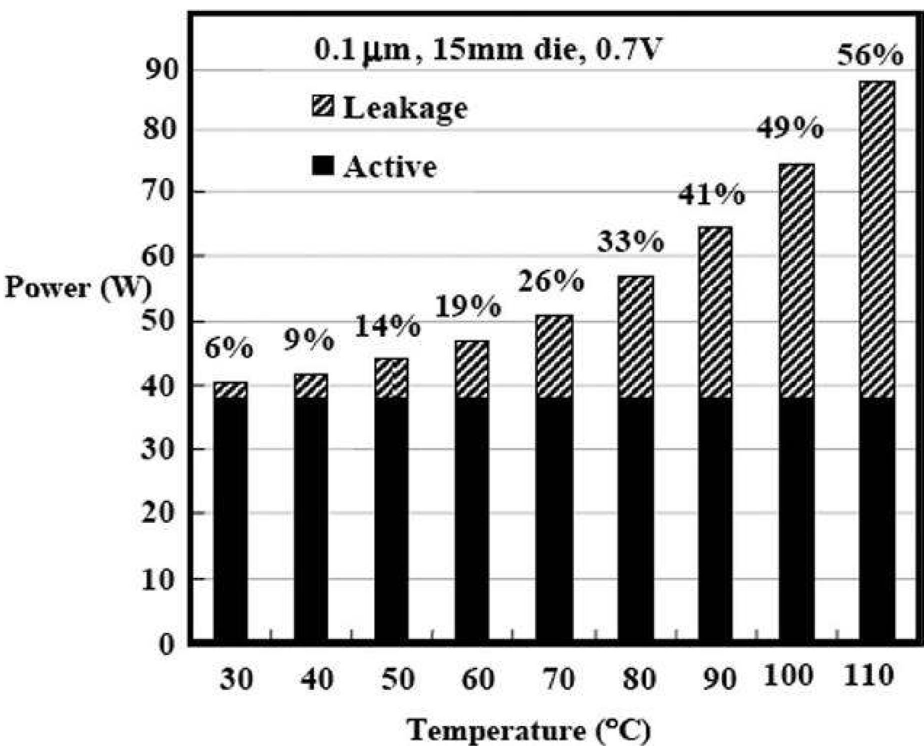


Figure 56: Increase in leakage power as a function of chip temperature for a Intel 15 mm die with 100 nm technology [64].

$$I_{sub} = \frac{W_{eff}}{L_{eff}} \mu(T) C_{ox} (V_T)^2 e^{(\frac{V_{GS}-V_{TH}}{nV_T})} \tag{29}$$

where the three temperature dependent terms are:

$$\mu(T) = \mu(T_0)(T/T_0)^{\alpha} \tag{30}$$

$$V_T = \frac{kT}{q} \tag{31}$$

$$V_{TH} = V_{TH0} - m(T - T_0) \tag{32}$$

W_{eff} and L_{eff} are effective width and length of the transistor; C_{ox} is the gate capacitance of a single transistor; $\mu(T)$ is the carrier mobility, which also depends on temperature; V_T is the thermal voltage derived by k (Boltzmann constant), temperature (T), and the charge of an electron (q); V_{GS} and V_{TH} are gate-source current and threshold voltage; α is a constant, and has a typical value of 1.5; m is the temperature coefficient of V_{TH} and has a typical value of 0.2 mV today. Using this compact physical model, the leakage current (normalized to the leakage current in room temperature) as a function of temperature is plotted in Figure 57. Without dynamic control of the supply voltage, leakage power is proportional to leakage current.

Based on the leakage current model proposed in [63] and assuming a constant supply voltage, the leakage power is calculated for a single tier with uniform power. Junction temperature and the normalized leakage power of the four cores are listed in Table 53 for a uniform power density of 100 W/cm^2. The leakage power of heater 4 (nearest to outlet) increases by 42.8 %, compared to heater 1 (nearest to inlet).

This phenomenon becomes a greater issue for larger sized chips. Intel has unveiled

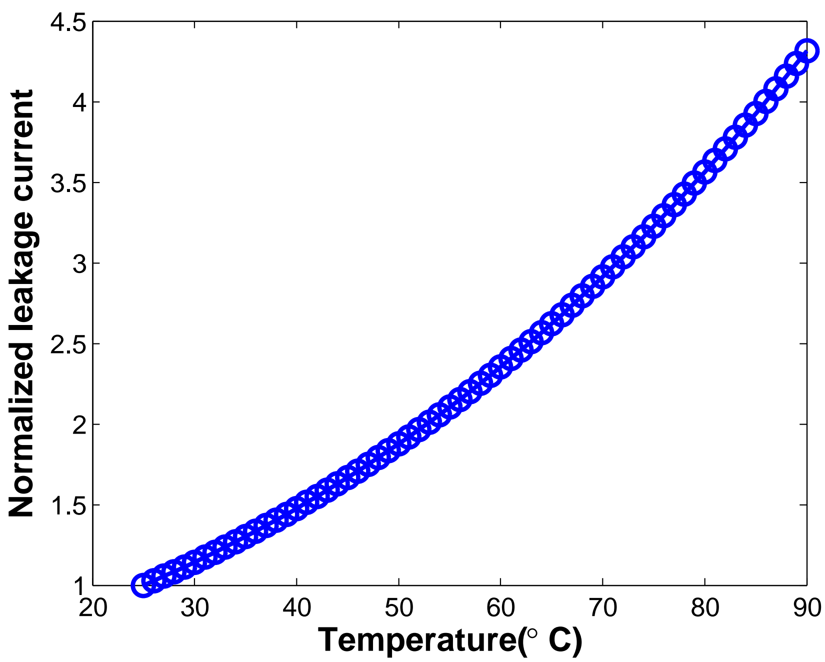


Figure 57: Normalized leakage current as a function of temperature.

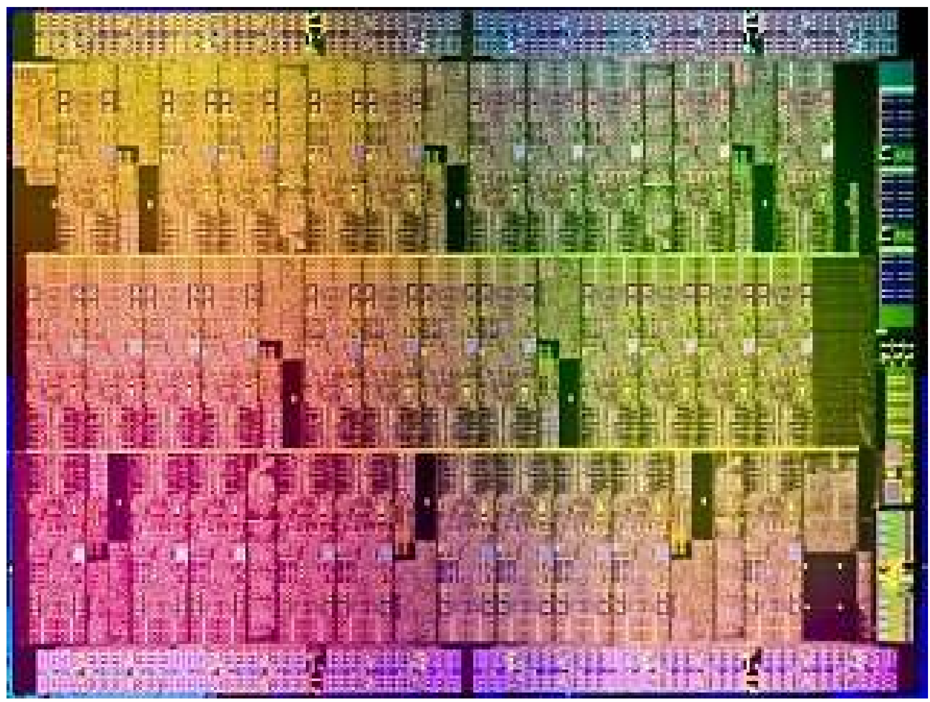


Figure 58: A diagraph of Intel's Knights Landing CPU, consisting of up to 72 x86 cores for exascale supercomputing.

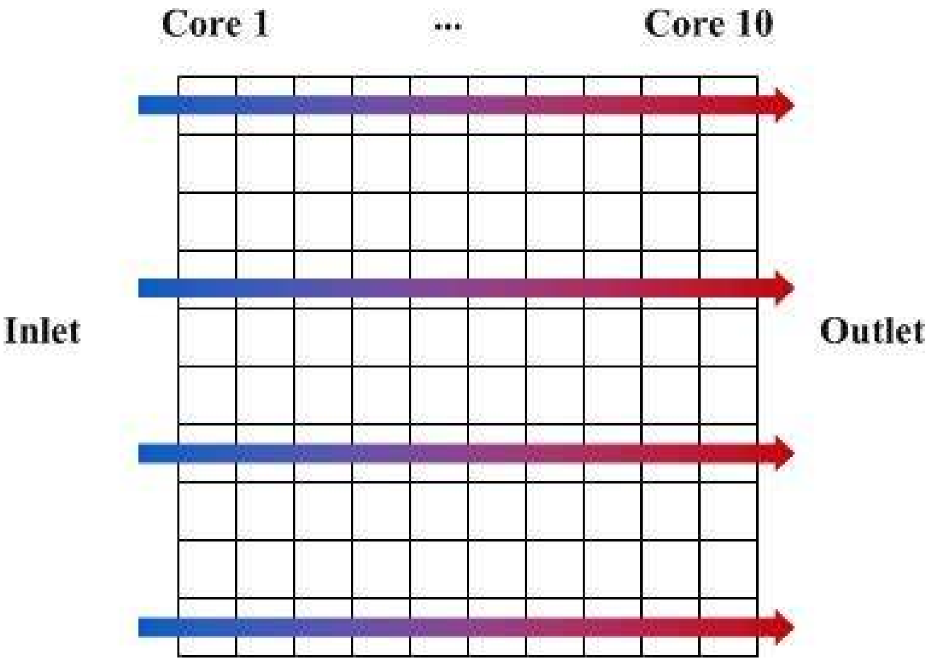


Figure 59: A diagraph illustration of the assumed 100-core CPU.

Table 5: Electrical implications due to lateral thermal gradient

	core 1	core 2	core 3	core 4	core 10
Junction temperature (°C)	36.9	41.3	47.5	52.8	85
Normalized P_{leak}	1.4	1.5	1.8	2	3.9

its plan to build an up-to-72-core CPU for exascale supercomputers. The sever, named Knights Landing, is slated to appear on the market in 2015 (Figure 58). Leakage current is analyzed here based on an assumed 100-core processor (Figure 59) comprised of a 10 × 10 core array covering a 2.5 cm × 2.5 cm chip. Since the testbed only has four heaters, the data for core 10 is extrapolated based on the measurements. The leakage power of core 10 is 2.8 times that of core 1.

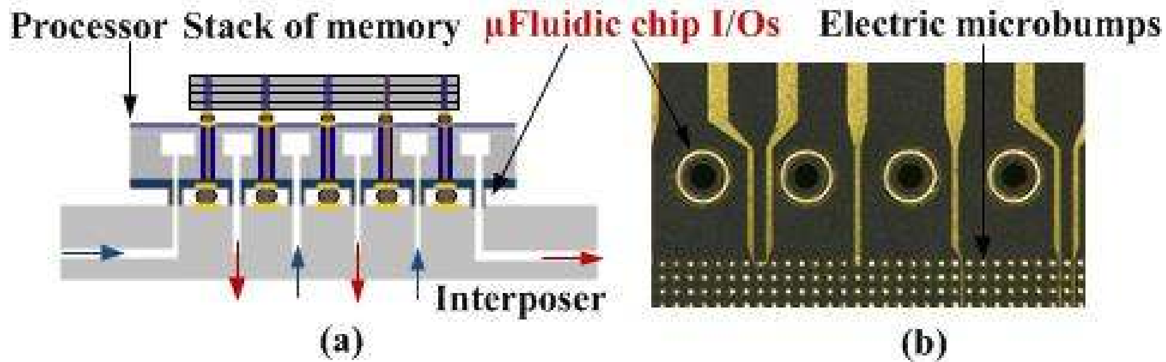


Figure 60: (a) Prototype of 3D stack with microfluidic chip I/Os for localized coolant delivery and (b) solder based microfluidic chip I/Os and electric microbumps.

4.4.4 Localized Coolant Delivery Method to Mitigate Lateral Thermal Gradient

One way to mitigate the thermal gradient is to increase the flow rate. But in doing so, the pressure drop and the pumping power will increase. In order to allow all the cores to work symmetrically, a localized coolant delivery method is proposed (Figure 60(a)). As can be seen, each core is associated with its own inlet and outlet so that each core can benefit from fresh coolant. The key technology is the microscale fluidic chip I/Os that is based on SnPb solder Figure 60(b). SEMs of the microfluidic I/Os with an outer diameter of 210 μm, an inner diameter of 150 μm, and a height of 12 μm are shown. The microfluidic chip I/Os have been experimentally shown to withstand a pressure drop of 100 kPa without leakage for 3 hours. Polymer-based microscale I/Os have been explored in [52]. SEMs of polymer pipe and polymer socket are shown in Figure 61 [52]. One big advantage of the solder-based microfluidic chip I/Os is that it can be fabricated with the electrical I/Os at the same steps. In addition, during the flip-chip bonding to form the electrical connectivities, the fluidic connectivity can also be formed. Also shown in the figure are the electrical microbumps with a density of 40,000 /cm^2 (microbump pitch of 50 μm), which is critical for power delivery and high-bandwidth off-chip signaling.

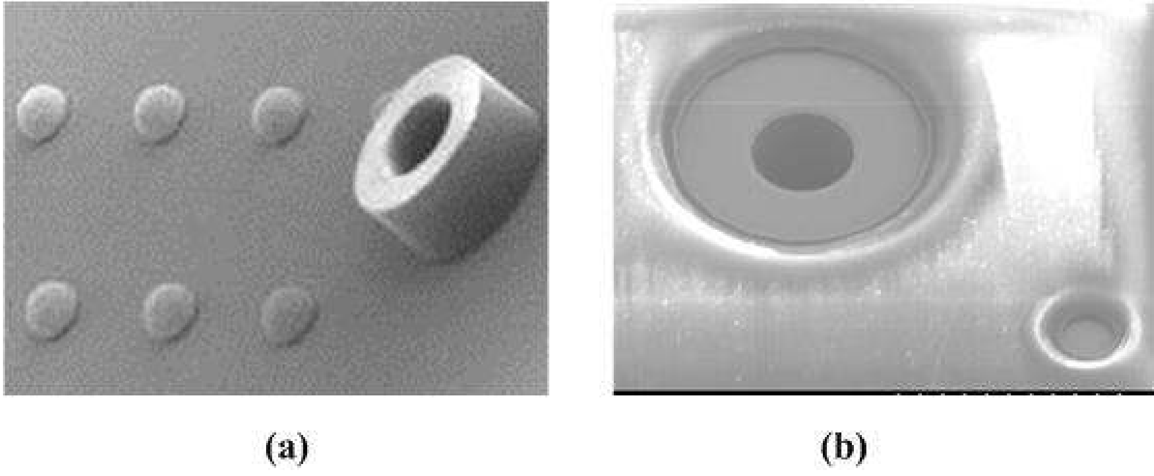


Figure 61: (a) A polymer pipe and (b) a polymer socket for fluidic delivery.

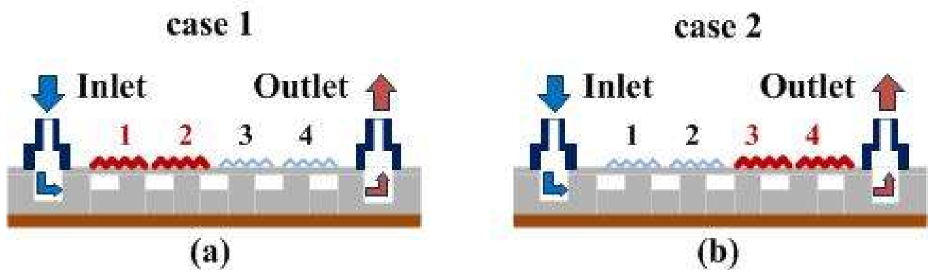


Figure 62: Evaluation of microfluidic cooling in chips with nonuniform power dissipation: (a) case 1 where heater 1 and 2 dissipate 100 W/cm^2 and heater 3 and 4 are off and (b) case 2 where heater 3 and 4 dissipate 100 W/cm^2 and heater 1 and 2 are off.

4.4.5 Microfluidic Cooling Under Nonuniform Power Dissipation

In a multicore CPU, there are cases where only some of the cores are active while the rest are idle. In this subsection, two test cases are measured to emulate these conditions. The two test cases are shown in Figure 62. The junction temperature rise of the four heaters in the two cases is plotted in Figure 63. In case 1, the two heaters (1 and 2) near the inlet are turned on while heaters 3 and 4 are off. The power density of heater 1 and 2 is 100 W/cm^2. The junction temperature rise of heaters 1 to 4 is 16.8 °C, 21.1 °C, 12.4 °C, and 10.6 °C, respectively. Heater 2 will be the hottest in this case. Heaters 1 and 2 follow a comparable trend, as shown in the

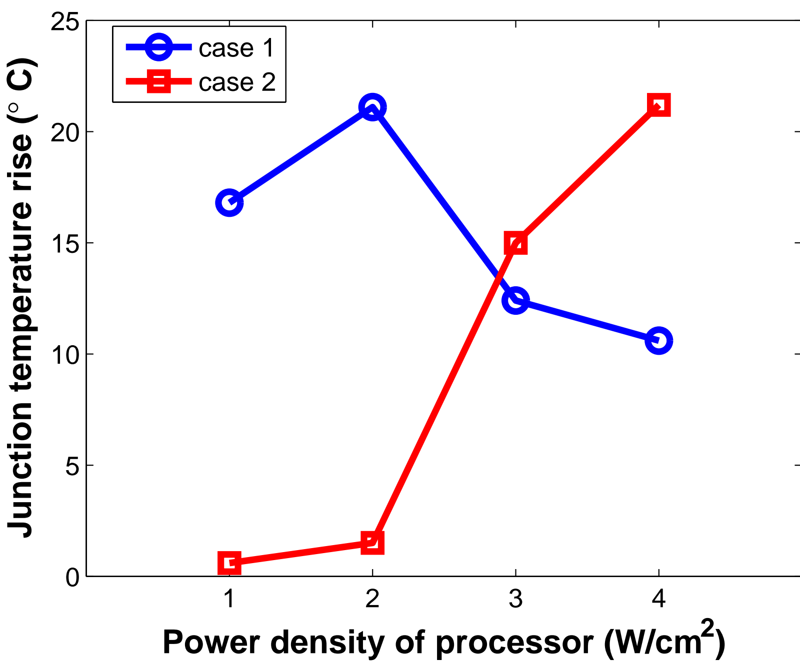


Figure 63: Junction temperature rise of heater 1 to 4 under the two different test cases shown in Figure 62

uniform power case (Figure 53). Even though heaters 3 and 4 are idle, the junction temperature still increases under the influence of the already warmed fluid. In case 2, the two heaters (3 and 4) that are near the outlet are turned on while heaters 1 and 2 are off. The junction temperature rise of heaters 1 to 4 is 0.6°C, 1.5 °C, 15 °C, 21.2 °C, respectively. Since the fluid is still cool when it flows across heater 1 and 2, the junction temperature of these two heaters barely increases. However, heater 2's temperature is higher than that of heater 1 because the heat generated in heater 3 spreads through silicon. But this spreading effect is minimal compared to the thermal coupling due to the fluid. From these two test cases, we understand that the fluid can introduce a thermal coupling effect between the cores. The cores should be placed in different locations on the chip depending on the applications. The other method to minimize the thermal coupling between cores is to implement the localized coolant-delivery mechanism.

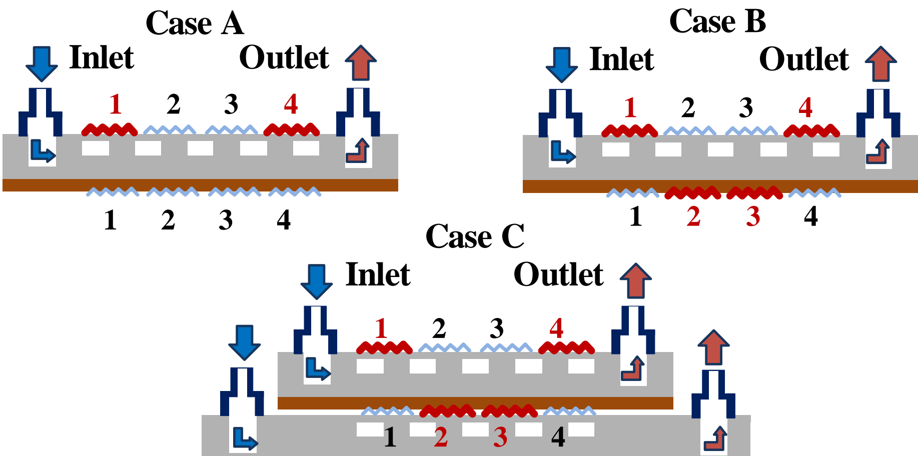


Figure 64: Vertical thermal coupling test cases. (Case A) Heaters 1 and 4 in upper tier are powered. (Case B) Heaters 1 and 4 in upper tier and heaters 2 and 3 in lower tier are powered. (Case C) Heaters 1 and 4 in upper tier and heaters 2 and 3 in lower tier are powered with DI water pumped into both tiers.

4.4.6 Vertical Thermal Coupling

Vertical thermal coupling between two tiers with embedded microfluidic heat sinks is investigated next [65]. In Case A and B (Figure 64), DI water is only pumped into the top tier such that the two tiers share the same microfluidic heat sink. In Case A, heaters 1 and 4 of the top tier are each powered up to 25 W. In Case B, heaters 2 and 3 in the lower tier are each powered up to 25 W in addition to the heaters in the upper tier. Once the heaters in the lower tier are turned on, as shown in Figure 65, the junction temperature of heaters 1, 2, 3, and 4 in the upper tier is elevated by 3.3 °C, 6.8 °C, 9.7 °C, and 10.1 °C, respectively. In Case C, the power dissipation profile in the two tiers is the same as that in Case B. The difference is that DI water is pumped into both tiers. Clearly, the temperature of the upper tier in Case A and Case C overlap indicating the impact of the lower tier is minimal. In Case C, embedding a microfluidic heat sink in the bottom tier provides a heat flow path with a lower thermal resistance. This would greatly diminish the heat transfer to the upper tier. In Figure 66, the temperature of the lower tier in the three cases

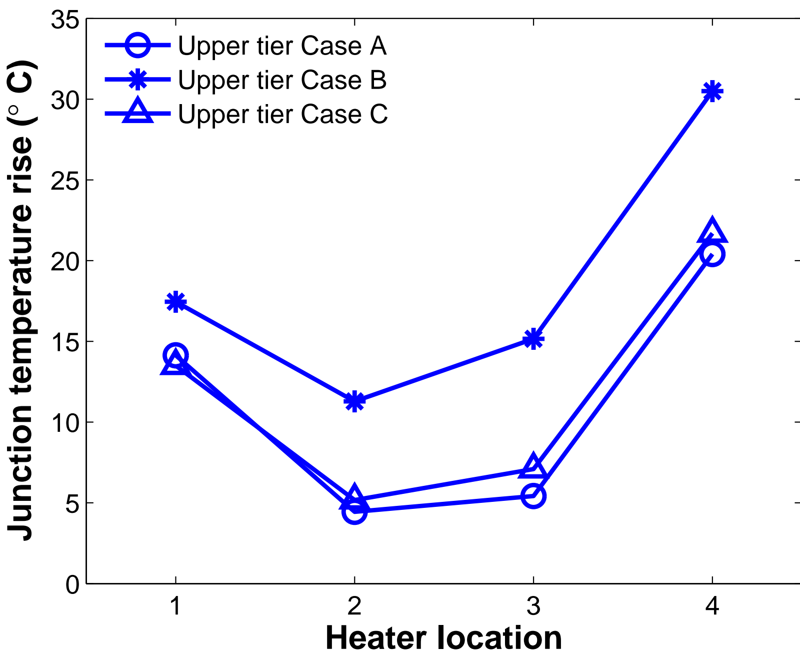


Figure 65: The junction temperature increase of the upper tier at different heater locations on the chip for the three cases.

is plotted. For Case A, the lower tier is idle. However, because of the temperature increase of the coolant, the temperature of heaters 1, 2, 3 and 4 of the lower tier is elevated by 4.9 °C, 5.7 °C, 7.8 °C, and 9.7 °C, respectively. Vertical thermal coupling may cause idle tiers to get warmer, leading to unwanted leakage power [63]. To reduce the vertical thermal coupling between tiers in microfluidic cooling, each tier can have its own microfluidic heat sink (Case C) instead of sharing one heat sink (Case B).

4.5 Validation through ANSYS Simulations

To better understand the flow distribution and heat transfer at different regions of the chip, ANSYS simulations were performed. The following simulations were done using ANSYS Fluent. The simulations were done for a single layer microfluidic cooled chip with uniform power dissipation. Since we are most interested in the flow characteristics below the chip, only the center of the testbed is simulated. The guiding

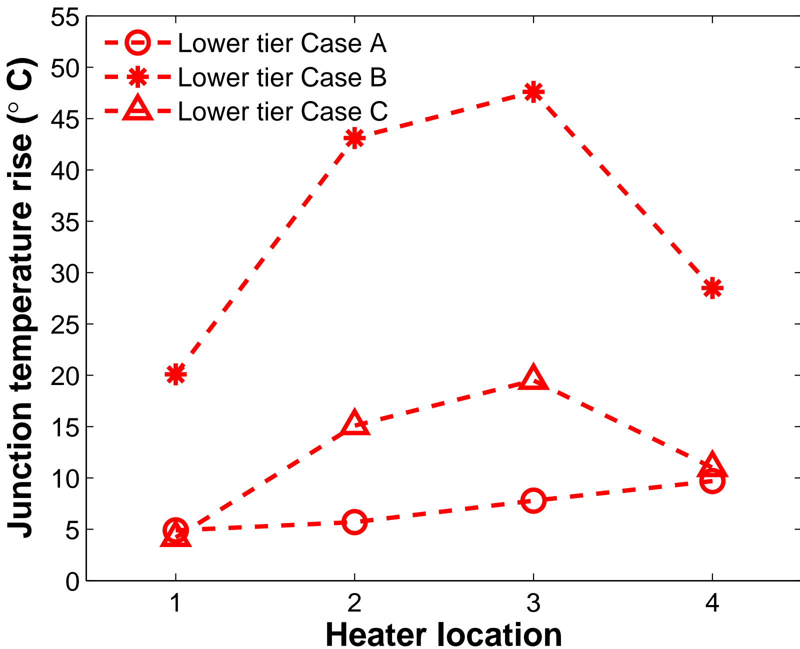
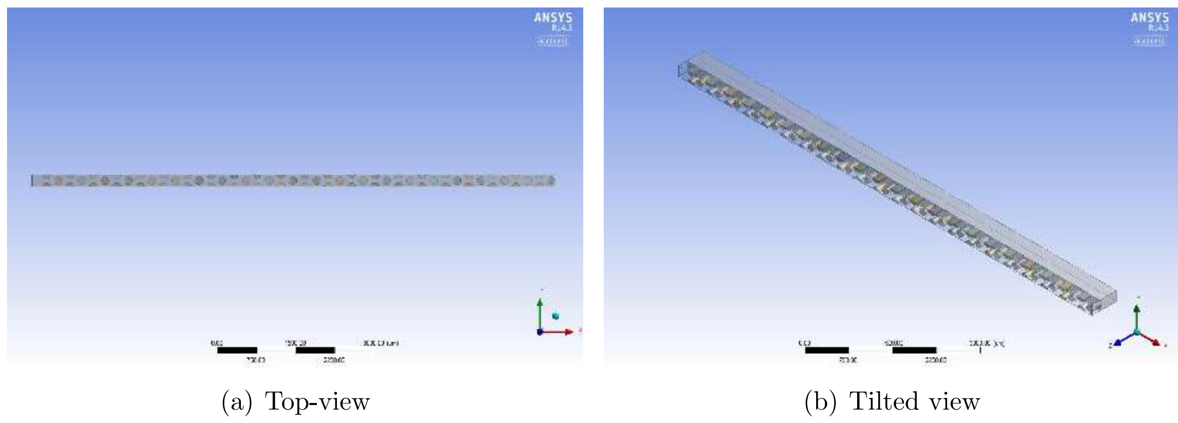


Figure 66: The junction temperature increase of the lower tier at different heater locations on the chip for the three cases.

channels and the nanoports are ignored in the simulations.

4.5.1 Initial Simulation of Repeatable Cell Rows Assuming Even Flow Distribution

At first, to reduce the computation time, a cell row is created, as shown in Figure 67. Fluid flows from inlet (left) to outlet (right). The boundary planes at top and bottom are set to be symmetric to indicate that this row is a part of a larger array. The entire micropin-fin array can be considered to consist of 43 cell rows. Assuming the flow distributes evenly across the 43 cell rows, the flow rate for one cell row can be calculated. The geometries and materials are selected to most closely represent the testbed. The thickness of the base is 400 μm. The heat sink has the exact same geometries as described in Chapter III. In the real sample, there is a 2 μm thick silicon dioxide layer at the bottom the base. Adding a thin layer in a model is generally not ideal and will create many more elements since the element in the thin layer is small.



(a) Top-view

(b) Tilted view

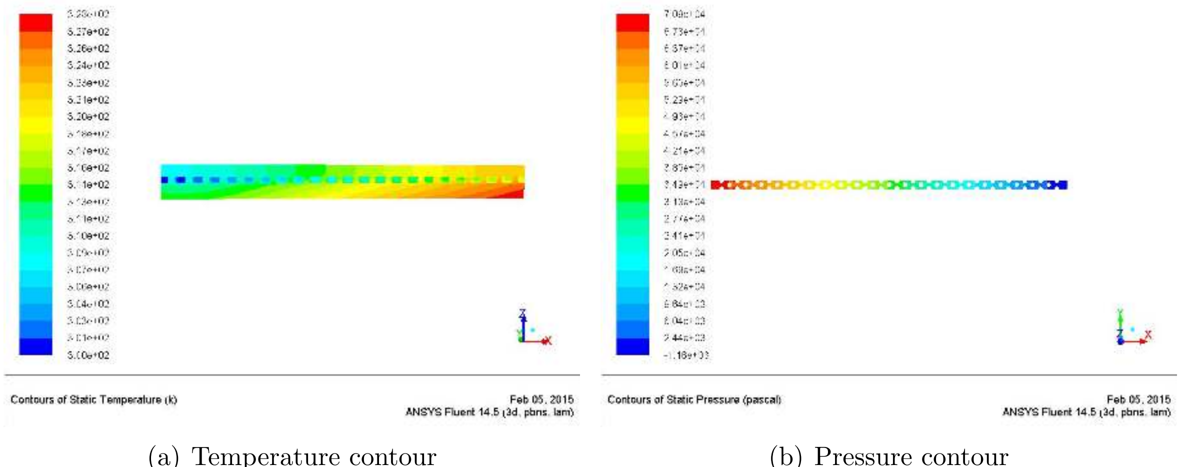
Figure 67: Created cell row in ANSYS to simulate MPFHS in a single-layer chip.

Table 6: Summary of the simulation results vs. experimental results

	T_rise1 °C	T_rise2 °C	T_rise3 °C	T_rise4 °C	ΔP (kPa)
Experiments	17.8	22.2	28.4	33.7	60
Initial Simulation	14.4	18	22.5	26.5	70.9
Modified simulation	16.1	21.1	27.4	32.8	41.7

Therefore, the silicon dioxide layer is considered as a shell that conducts heat at the bottom of the base in the simulation through the shell conduction option in Fluent.

Figure 68 shows the temperature profile and pressure profile after the initial simulation. The average junction temperature rise for heaters 1 to 4 is 14.4 °C, 18.0 °C, 22.5 °C, and 26.5 °C, respectively. Compared with the experiments, we notice this is significantly lower than the measured increase in junction temperature. The experimental data and initial simulation data are included in the first and second rows of Table 6. The pressure drop in the initial simulation is 70.9 kPa which is also higher than the experiments. By analyzing the data, we conclude that the simulated fluid velocity is higher than the experiment. The assumption that the flow distributes evenly across the array may not be valid. To understand the flow distribution in the array, a half array is modeled in ANSYS. Since the array is symmetric, the other half of the array is not modeled.



(a) Temperature contour

(b) Pressure contour

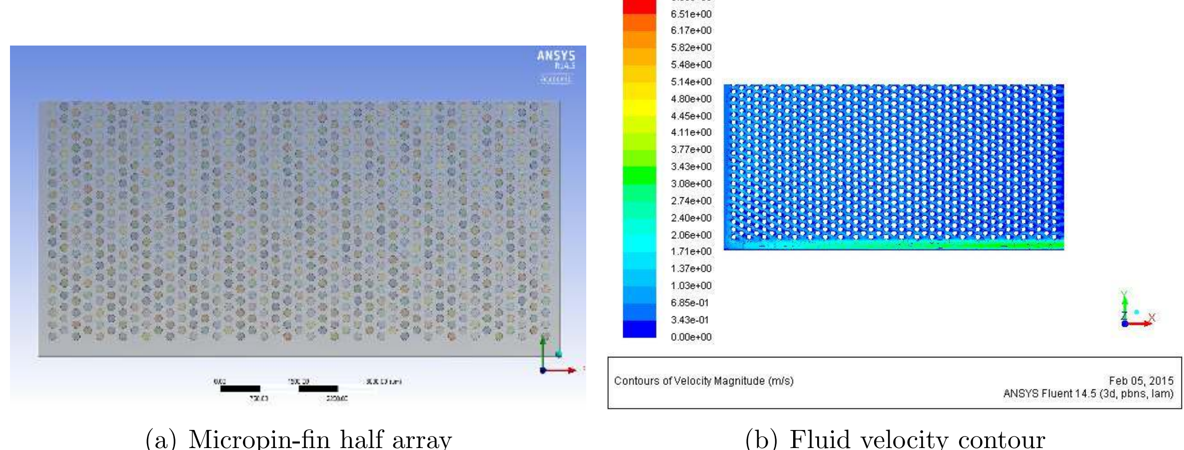
Figure 68: The temperature and pressure contour profiles after initial simulation

4.5.2 Adjusted Simulation of Repeatable Cell Rows

In the half-array simulation, the dimensions of the array are exactly the same as in the testbed. To be noted, there is a 272.5 μm gap between the micropin-fin array and the edge (shown in Figure 69(a)). The velocity contour is plotted in Figure 69(b). As seen, the fluid velocity is much higher near the edge. The maximum velocity in the micropin-fin array is 1.3 m/s while that near the edge it is 3.0 m/s. The key conclusions from the half-array simulations are:

- The flow across the heat sink is not evenly distributed; fluid tends to flow near the edge because the friction factor is lower.
- The cooling capability is diminished since fluid flows around the micropin-fin array.
- In future heat sink designs, the gap between the micropin-fin array and the edge should be as small as possible to force the fluid to flow across the micropin-fin array.

Taking the flow wasted through the gap into consideration, the maximum velocity in the micropin-fin region is calculated to be 1.3 m/s. This velocity is used to adjust



(a) Micropin-fin half array

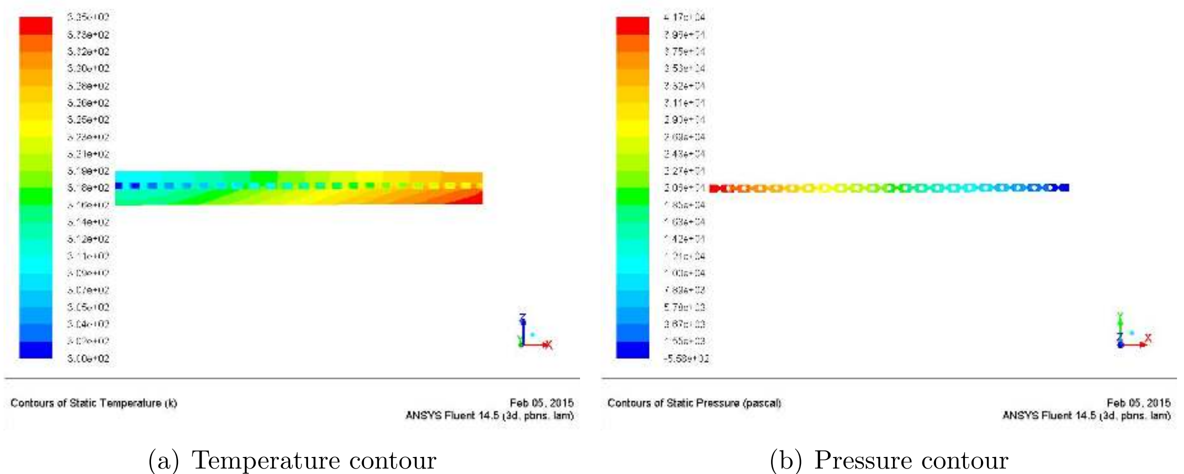
(b) Fluid velocity contour

Figure 69: The created cell row in ANSYS to simulate MPFHS sink in a single-layer chip.

the initial simulation using cell row and the results are included in Figure 70.

The junction temperature increase in this case is listed in the third row of Table 6. By comparing the new simulation results with the experimental data, the temperature difference is approximately 1 °C. As for the pressure drop, the simulated pressure drop across the micropin-fin array is 41.7 kPa while the total pressure drop in the experiment is 60.0 kPa. In the experiment, the differential pressure gauge is connected to the inlet and outlet tubes. Therefore, the total pressure drop consists of the pressure drop across the micropin-fin array, the pressure drop across the guiding channels, and the pressure drop due to the transition from tubes to the guiding channels. It is expected that the total pressure drop to be higher.

Although simulation of the entire array or half of the array can give us the closest result to the experiment, the computational time is long (several hours). By adjusting the velocity of the inlet, the one-row simulation provides results that are reasonably close. Owing to the reduction of element number from $3\text{x}10^6$ to $3\text{x}10^5$, the computation time reduces to several minutes. But we should note that one-row simulation is only good if the heat sink is designed to have a small gap between the array and



(a) Temperature contour

(b) Pressure contour

Figure 70: The temperature and pressure contour profiles from the adjusted simulation.

edge.

4.6 Conclusion

In this chapter, a microfluidic heat sink that has the same geometries as described in Chapter III is implemented in a two-tier thermal testbed. Memory-on-processor and processor-on-processor are emulated using the testbed. In both cases, microfluidic cooling outperforms air cooling. In addition, a tier-specific cooling mechanism that allows tailoring the flow rate according to the power dissipation of each tier is implemented. This method is shown to be able to minimize the thermal gradient between tiers and thus minimize the thermal–mechanical stress. Pumping power is also reduced by preventing overcooling of the low-power chip. At the end, microfluidic cooling is evaluated in a multi-core chip. The lateral thermal coupling is observed because of the warmed fluid. A lateral thermal gradient caused leakage power increase is analyzed. Vertical thermal coupling is also emulated. To reduce the vertical thermal coupling, each high-power tier should have its own microfluidic heat sink.

CHAPTER V

THERMAL ISOLATION FOR HETEROGENEOUS 3D ICS

5.1 Introduction

The previous chapters addressed heat removal needs in a 3D stack. However, in some applications where high-power dice (e.g. logic dice) are stacked along with low-power and temperature-sensitive components (memory or silicon nanophotonic dice, for example), thermal management will not only require effective cooling, but may also require effective thermal isolation to 'protect' the temperature-sensitive components from the time-varying power dissipation of other chips in the stack. By placing such tiers next to each other, the thermal coupling between them will be significant, leading to possibly undesirable junction temperature variation in the temperature-sensitive tier as a result of the high-power chips.

Silicon photonic based interconnects are emerging as an alternative to electrical interconnects for high-bandwidth and low-power consumption. Stacking a silicon photonic die with logic and memory has been explored in the literature [25], [26]. The temperature sensitivity of the optical elements presents challenges for integration; for example, a microring modulator with 5 μm diameter is reported to have a wavelength drift of 0.11 nm/°C in [27]. A temperature change of 13.5 °C will result in a complete passband mismatch between transmitter–receiver pairs in 64-channel wavelength-division multiplexing (WDM). Local thermal tuning and control circuitry are commonly used to maintain the microring resonators at a constant temperature. The projected energy-per-bit for a photonic link is 300 fJ/bit at 15 Gbps in 2015 according to [66]. However, currently, the demonstrated tuning power is approximately 164 fJ/bit at 10 Gbps [66]. Thus, the thermal tuning power is a significant portion

of the total power budget and may become more significant as high temperature variations occur in adjacent high-power chips.

In applications involving stacking memory on logic, SRAM and DRAM stacking on a processor has been widely explored [23], [24]. It is shown in [23] that the temperature of SRAM increases by 30-to-40 °C because of the heating from the processor tier. The leakage power of the SRAM increases by approximately two times because of the increased temperature. Even worse, the average cache access time also increases by 50 ps, leading to an approximately 28% performance degradation [23].

In today's approach to 3D IC stacks, tiers are bonded using microbumps along with underfill, which is applied between tiers to alleviate the thermomechanical stress on the solder microbumps, as shown in Figure 1. However, the thermal conductivity of underfill is around 0.4 W/mK–1.3 W/mK. This will introduce a small thermal resistance between the two tiers and cause thermal coupling between the tiers. To thermally decouple the tiers in 3D ICs, we propose to integrate an air gap and thermally degraded mechanically flexible interconnects (MFIs) to replace both the microbumps and the underfill. The proposed 3D IC approach is shown in Figure 71. Unlike rigid solder microbumps, MFIs can deform elastically under stress, which helps maintain the electrical connectivity between tiers. Thanks to this phenomenon, MFIs can help eliminate the underfill and thus reduce the thermal coupling between tiers. The thermally degraded MFIs are a type of MFIs that are specially designed to have a large thermal resistance and small electrical parasitics.

Local microcavities beneath the temperature-sensitive components have been explored in [67], [68] seeking to reduce the thermal coupling from the surrounding devices. A local undercut microcavity is created beneath the resonator and is shown to reduce the tuning power by an order of magnitude in [68]. However, to our knowledge, little effort has been made to investigate the thermal isolation between the low-power chips from the high-power chips in a 3D stack. Adopting the chip-scale air/vacuum

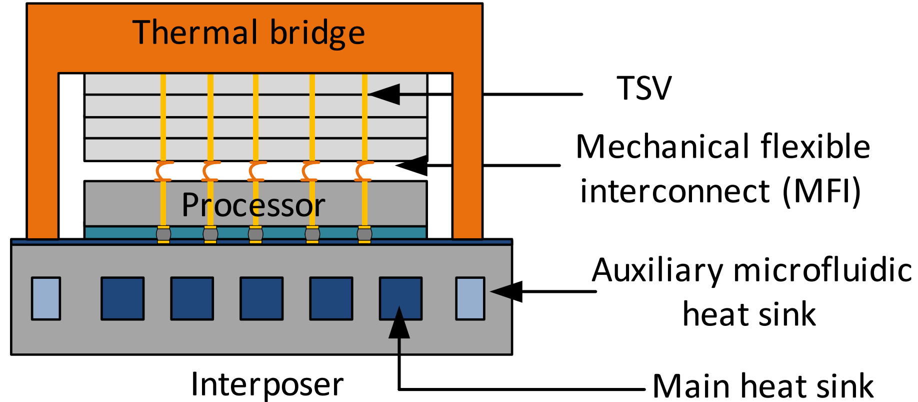


Figure 71: Prototype shows the proposed thermal isolation technology that replaces microbumps and underfill with air gap and thermally degraded MFIs.

cavity and MFIs in the low-power chip brings about smaller temperature variations. Moreover, the local thermal isolation method based on an undercut microcavity can be deployed in addition to our proposed concept in order to ensure a constant local temperature. Thus, by combining our chip-scale thermal isolation technology with 3D stack architectures, new opportunities for improved heterogeneous system integration and miniaturization become possible.

5.2 Resistance Network Modeling

Stacking of high-power dice (e.g. processor) along with low-power and temperature-sensitive dice presents a number of challenges. The time-dependent temperature variation in the processor tier, which is workload dependent, will be directly coupled to nearby stacked chips. In current 3D IC approaches, it is common to use an adhesive with high thermal conductivity between stacked chips to ensure the thermal resistance between each tier is as small as possible. This method helps remove the heat from within the stack to the top most portion of the stack where an air-cooled heat sink is attached. However, the thermally conductive adhesive will also enhance the thermal cross-talk between the processor tier and other dice in the stack that are temperature-sensitive (memory and silicon nanophotonics, for example) leading to temperature

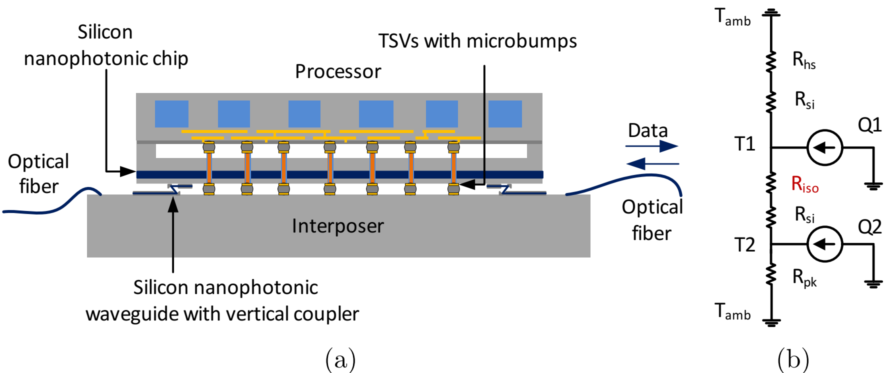


Figure 72: (a) A 3D stack of processor and silicon nanophotonic chips with hybrid thermal management: within-tier microfluidic cooling in processor and air/vacuum cavity to thermally isolate the silicon nanophotonic chip. (b) The corresponding thermal resistance network.

variations. In our proposed vision, thermal isolation technology is used to enable the stacking of a high-power die and low-power temperature-sensitive dice.

An example illustration of a logic-silicon nanophotonic stack using within-tier cooling and air gap thermal isolation is shown in Figure 72(a). The figure illustrates a high-power processor tier with an embedded microfluidic heat sink stacked above a low-power silicon nanophotonic chip with air/vacuum cavities formed between the two dice to provide thermal isolation. The power of the silicon nanophotonic chip is dissipated through the silicon interposer.

To understand the thermal benefits of air gap isolation, we begin with a simplified compact physical model that neglects the impact of interconnects within the air/vacuum cavity. Using a 2D thermal resistance network model, as shown in Figure 72(b), the junction temperature of the bottom tier as a function of the power dissipation in the logic chip is plotted in Figure 73. When the power density of the processor tier increases from 50 W/cm^2 to 100 W/cm^2, the processor temperature increases from 46 °C to 68 °C and thus, yields a slope of 0.44 °C/W [69]. This trend is similar with and without thermal isolation. Without any thermal isolation, the temperature of the bottom tier follows the same trend. With a 5 μm thick air cavity,

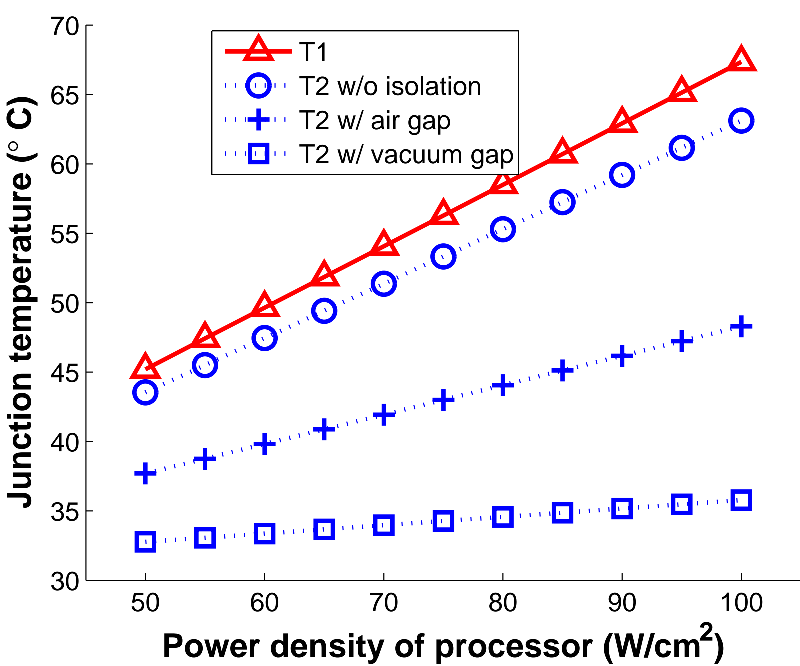


Figure 73: The junction temperature increase of the upper tier at different heater locations on the chip for the three cases. T1 and T2 denote the temperature of the high-power and low-power tiers, respectively.

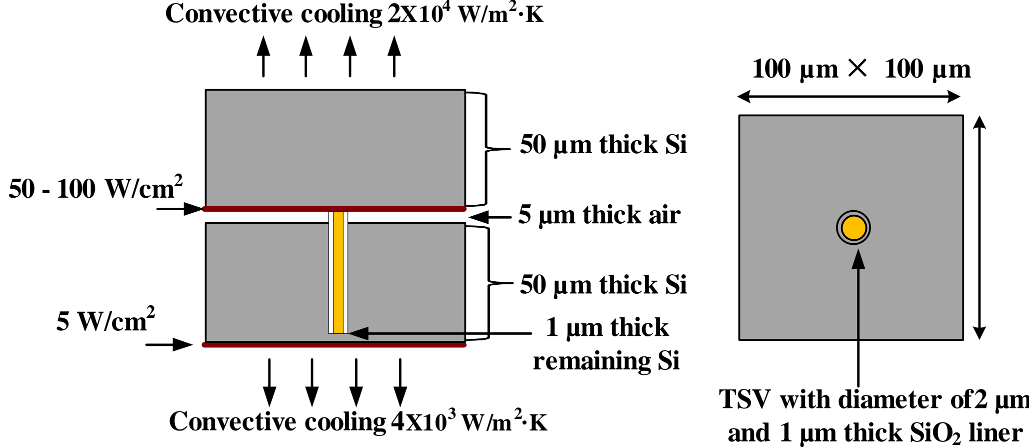


Figure 74: Illustration of the cross-sectional (left) and top (right) view of the structures simulated in ANSYS to represent TSVs through air cavity.

the temperature of the bottom tier increases with a smaller slope from (41 °C to 54 °C and thus yields a slope of 0.26 °C/W). If vacuum is created between the two tiers, the temperature of the bottom tier only increases by 4 °C (yielding a slope of 0.08 °C/W).

5.2.1 TSVs' Impact on Thermal Isolation

ANSYS simulations were performed to analyze the impact of routing TSVs through the air/vacuum cavity. The TSVs are partially embedded in the bottom chip and partially exposed in the air/vacuum cavity (as shown in Figure 74). Because of the heat conduction through the TSVs, the thermal coupling between the two tiers will increase. The results are plotted in Figure 75. TSVs are assumed to have a diameter of 2 μm and a silicon dioxide liner of 1 μm on a 100 μm $\times$ 100 μm pitch. The temperature variation of the bottom tier is more obvious with TSVs. The bottom tier temperature increases by 16.5 °C (yielding a slope of 0.33 °C/W) when the TSVs are present and 13 °C without the TSVs. The results are expected since TSVs, which are formed using copper, have good thermal conductivity, and thus cause undesired thermal coupling between the two tiers. One solution to this is to decrease the TSV

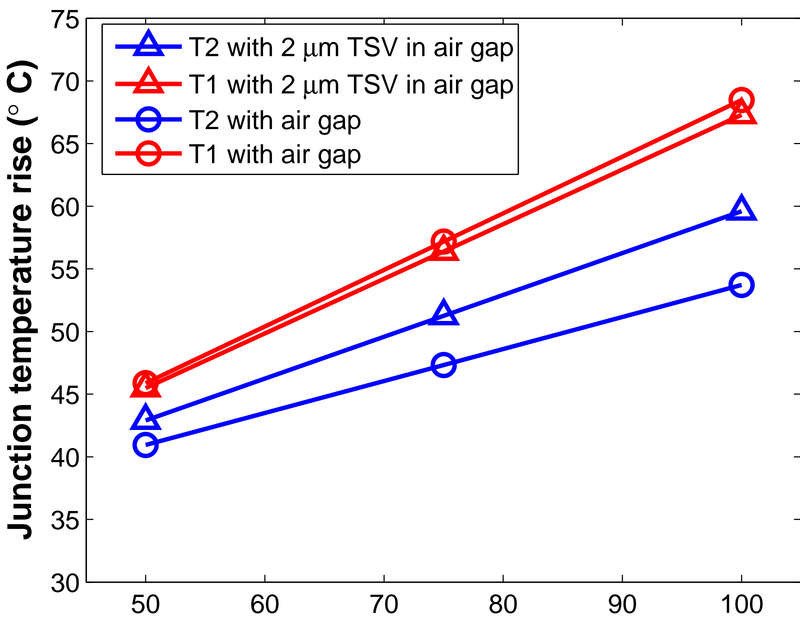
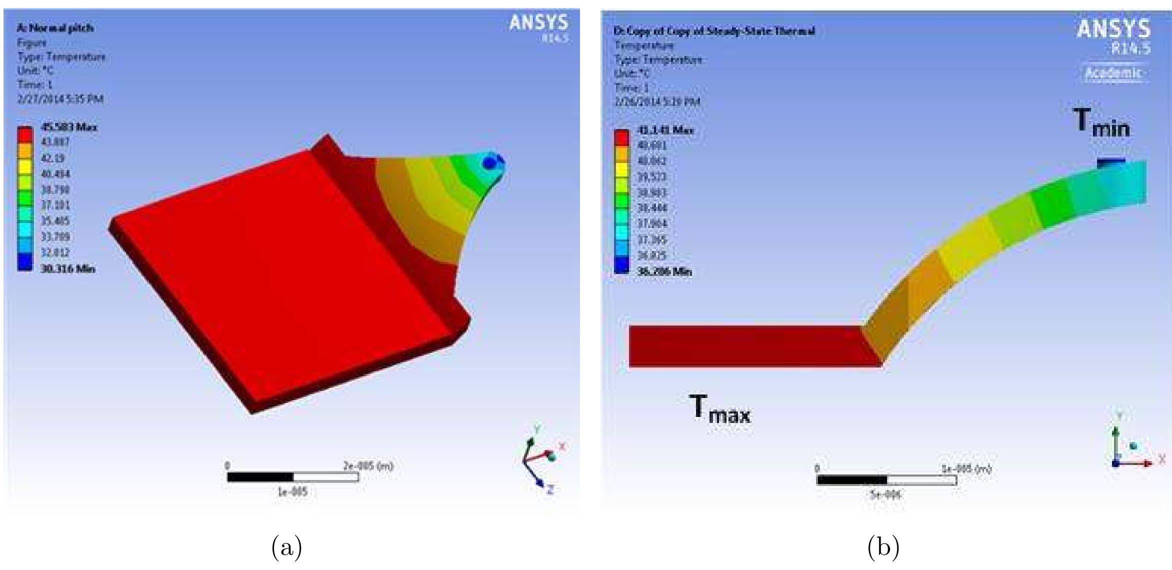


Figure 75: Temperature of both tiers in the simulated structure shown in Figure 74. T1 and T2 denote the temperature of the high-power and low-power tiers, respectively.

diameter, which will also lower parasitics.

5.2.2 MFI Thermal Resistance

The MFI designs have been simulated in ANSYS in order to understand their equivalent thermal resistance (Figure 76). The parameters varied in the simulation include MFI thickness and contact area between the MFI tip and the landing pad. Power is applied at the bottom of the MFI and heat is removed from the top. The highest temperature (T_{max}) appears at the bottom of the MFI while the lowest temperature (T_{min}) appears at the top of the MFI (Figure 76(b)). By measuring the temperature gradient from top to bottom, we can calculate the thermal resistance. Table 7 summarizes the thermal resistances of the MFIs by varying MFI thickness and contact area (contact area is influenced by the force applied between the two chips). As shown in Table 7, the thermal resistance of a single MFI ranges from 1×10^4 to 2×10^4 K/W, depending on the thickness and contact area.



(a) (b)

Figure 76: (a) The MFI structure created in ANSYS and (b) the corresponding thermal profile in a static thermal simulation.

Table 7: ANSYS simulated thermal resistance of a single MFI with various designs

MFI Thickness (μm)	D_{cont} [1](μm)	R_{th} ($\times 10^4$ °C/W)
3	4	0.8
3	2	1.12
2	4	1.25
2	2	1.7

[1] Assuming a circular contact region between MFI tip and the landing pad with a diameter of D_{cont}.

The 2D resistance network model is used to evaluate a stack with high-power and low-power tiers with different types and numbers of interconnects. The thermal resistance of the MFI is chosen to be 1×10^4 K/W in the model. The high-power tier dissipates 100 W/cm^2 and the low-power tier dissipates 5 W/cm^2. The assumed cooling methods for the bottom and top are microfluidic cooling and thermal bridge, respectively. An effective heat transfer coefficient is calculated based on the equivalent thermal resistance of these two cooling methods. At the bottom, an equivalent thermal resistance of 0.25 K/W is obtained from our measurement. At the top, the equivalent thermal resistance of the thermal bridge is obtained from ANSYS simulations, as discussed in Section 5.3.1. Using Eq. (33), the heat transfer coefficient of the top and bottom can be calculated.

$$h = \frac{1}{R \cdot A} \tag{33}$$

where A is the chip area, and equals to 1 cm^2. The obtained heat transfer coefficient is 4×10^4 W/m^{2}K and 8×10^3 W/m^{2}K at the bottom and the top of the stack, respectively. The gap between the tiers is 10 μm in both cases.

The different interconnect scenarios simulated include (a) air gap with uniformly distributed MFIs and (b) conventional microbumps with underfill (Figure 77). In case (b), the temperature of the high-power and low-power tiers remains constant regardless of the number of microbumps. The reason is that underfill is assumed to have a thermal conductivity of 0.9 W/mK, which essentially dominates the heat conduction between the two tiers. Moreover, because heat conduction occurs through the underfill, it is observed that the temperature of the two tiers is very close (44.3 °C for both tiers). Meanwhile, in case (a), it is shown that the temperature difference between the high-power and low-power tiers varies as a function of the number of interconnects. This is mainly due to the low thermal conductivity of air. For example, for the MFI

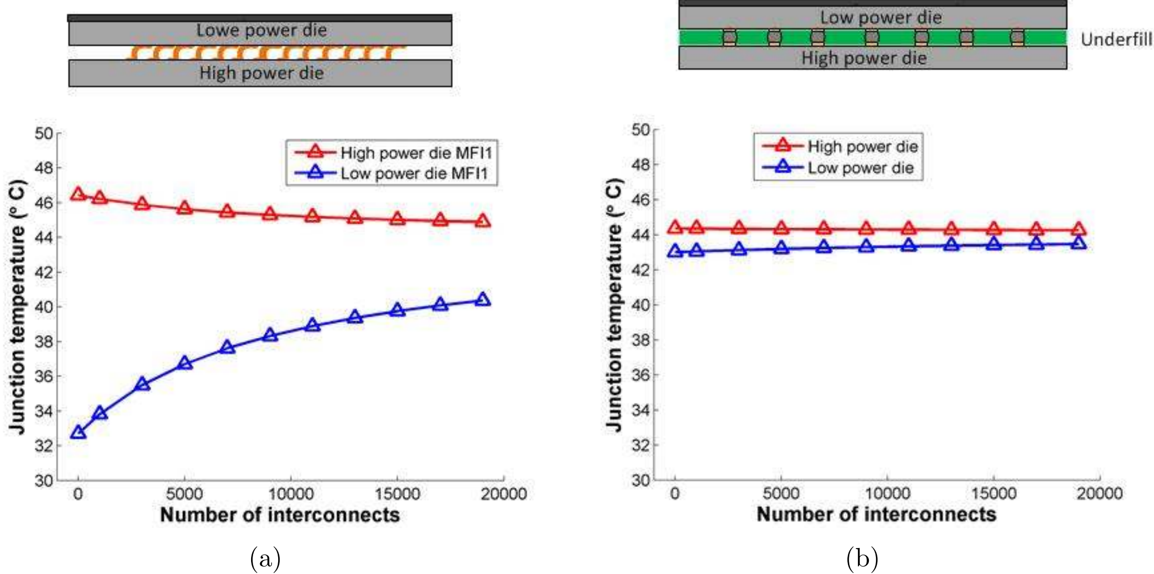


Figure 77: Temperature of the high-power and low-power die with different interconnects: (a) Uniform MFIs within air gap and (b) microbumps and underfill.

in air case, when there are 3,000 MFIs, the low-power tier has a temperature of 35.5 °C, which is 10 °C lower than that of the high-power tier. This demonstrates that the MFIs and air gap can effectively isolate the low-power tier from the high-power tier.

5.3 *Finite Difference Modeling of the Proposed Stack with Thermal Isolation Technology*

The resistance network model is used to demonstrates the thermal isolation concept in Section 5.2. However, it is only suitable for structures with uniform power dissipation and uniform interconnects. Therefore, a finite difference modeling scheme is developed for the proposed architecture, as shown in Figure 71. Because of the complicated geometries of the MFIs, TSVs are used as an alternative in order to simplify the modeling. Equation (34) describes the general heat transfer at steady state.

$$\nabla(K(x,y,z)\nabla(T(x,y,z)) = -P(x,y,z) \tag{34}$$

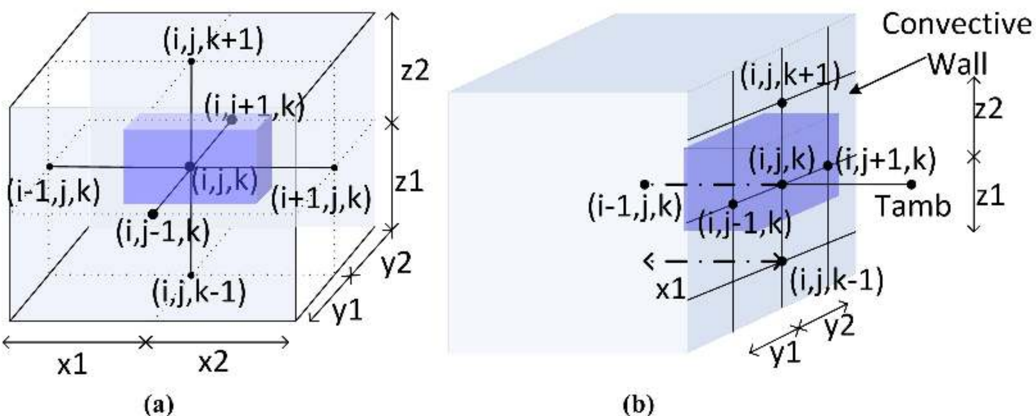


Figure 78: Finite difference scheme: (a) general points inside the stack and (b) boundary points in the face of the stack [70].

where $K(x, y, z)$ and $T(x, y, z)$ are the thermal conductivity and temperature, respectively, while $P(x, y, z)$ represents the power consumption.

By meshing a stack into 3D grids, we can employ the six-node first-order approximation to equation (34), as shown in Figure 78(a). Then a finite-difference scheme for equation (34) is built:

$$\begin{aligned} &\frac{T_{i,j,k} - T_{i-1,j,k}}{\frac{x_1}{k_x l_y l_z}} + \frac{T_{i,j,k} - T_{i+1,j,k}}{\frac{x_2}{k_x l_y l_z}} + \frac{T_{i,j,k} - T_{i,j-1,k}}{\frac{y_1}{k_y l_x l_z}} + \\ &\frac{T_{i,j,k} - T_{i,j+1,k}}{\frac{y_2}{k_y l_x l_z}} + \frac{T_{i,j,k} - T_{i,j,k-1}}{\frac{z_1}{k_z l_x l_z}} + \frac{T_{i,j,k} - T_{i,j,k+1}}{\frac{z_2}{k_z l_x l_z}} = P_{total} \end{aligned} \tag{35}$$

where $l_x = (x_1 + x_2)/2$, $l_y = (y_1 + y_2)/2$, $l_z = (z_1 + z_2)/2$; P_{total} is the power consumption in the shaded rectangle. When solving equation (35), the boundary conditions should be added. Usually we use a convective boundary for the stack. Here we derive the finite difference scheme for the nodes at the boundaries, as shown in Figure 78(b). A convective boundary equation is:

$$K\frac{\partial T}{\partial \vec{n}}|_{boudary} = -h(T - T_{amb}) \tag{36}$$

where h is the heat transfer coefficient for the convection. T_{amb} is the ambient temperature. According to Figure 78(b), the scheme at node (i, j, k) is:

$$\begin{aligned} \frac{T_{i,j,k} - T_{i-1,j,k}}{\frac{x_1}{k_x l_y l_z}} + \frac{T_{i,j,k} - T_{amb}}{\frac{1}{h l_y l_z}} + \frac{T_{i,j,k} - T_{i,j-1,k}}{\frac{2y_1}{k_y l_x l_z}} + \\ \frac{T_{i,j,k} - T_{i,j+1,k}}{\frac{2y_2}{k_y l_x l_z}} + \frac{T_{i,j,k} - T_{i,j,k-1}}{\frac{2z_1}{k_z l_x l_z}} + \frac{T_{i,j,k} - T_{i,j,k+1}}{\frac{2z_2}{k_z l_x l_z}} = P_{total} \end{aligned} \tag{37}$$

In each mesh, there is only one type of material. This meshing strategy may increase the number of mesh nodes, but improves the modeling accuracy. When the geometry and material details are provided, we can build the equation in the following matrix form:

$$Ax = P \tag{38}$$

where A is the coefficient matrix, P is the power consumption vector, and x is the unknown temperature vector. The model was implemented using Matlab and was used to analyze the proposed architecture.

5.3.1 Thermal Bridge

An air/vacuum gap was introduced previously in order to decrease the thermal coupling between the high-power tier and the low-power tier, thus providing a measure of protection to the low-power tier. However, another cooling path needs to be provided to the thermally isolated tier. Taking a memory-on-processor stack (as shown

in Figure 71) as an example, the heat generated in the memory tier encounters a large thermal resistance because of the air gap when it travels downwards. Without a cooling path from the top, the junction temperature of the memory tier may be high. We have developed a novel concept to resolve this issue: the idea is to attach a 'thermal bridge' on top of the memory tier and interconnects it to a microfluidic cooled interposer. This thermal bridge can be made of copper and, thus, exhibit a small thermal resistance. We envision integrating two independent microfluidic heat sinks in the interposer in which the main microfluidic heat sink is used for cooling the processor while the auxiliary microfluidic heat sink is dedicated to the cooling of the memory tier.

The structure shown in Figure 79 is simulated using ANSYS. The dimensions of the memory die are 1 cm $\times$ 1 cm $\times$ 50 μm and the dimensions of the interposer are 1.5 cm $\times$ 1.5 cm $\times$ 200 μm. The copper thermal bridge is attached to the top of the memory tier using a 10 μm thick thermal interface material (TIM) with a thermal conductivity of 6 W/mK. The top surface of the copper thermal bridge is the same size as the interposer and has a thickness of 500 μm. The two support structures have a width of 2 mm and a height of 115 μm. A 10 μm TIM is assumed at the bottom to connect the thermal bridge to the interposer. In the simulation, the memory tier dissipates 5 W. The objective is to understand the thermal resistance of the thermal bridge (including TIM)). Figure 80 shows the temperature map of the memory tier. The highest temperature of the memory tier is 28.2 °C and appears at the middle of the memory tier. For reference, the ambient temperature is set to 22 °C. Thus, the temperature gradient from the memory to ambient is 6.2 °C, yielding a total thermal resistance of 1.24 °C/W. Thus, we assume a heat transfer coefficient of 8,000 W/m^{2}K (due to the thermal bridge) at the top.

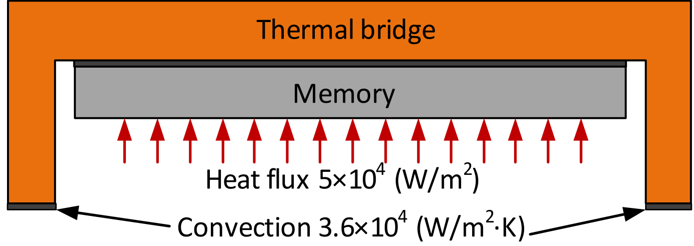


Figure 79: Thermal bridge on top of a memory tier simulated in ANSYS.

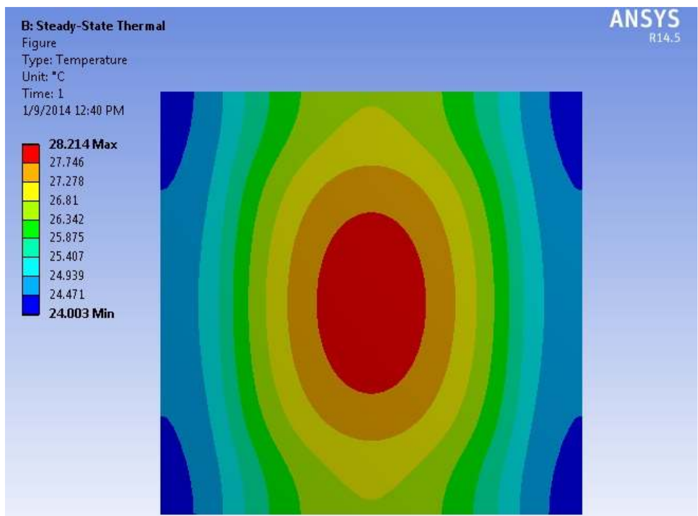


Figure 80: Memory tier temperature map for the calculation of the thermal resistance of the thermal bridge.

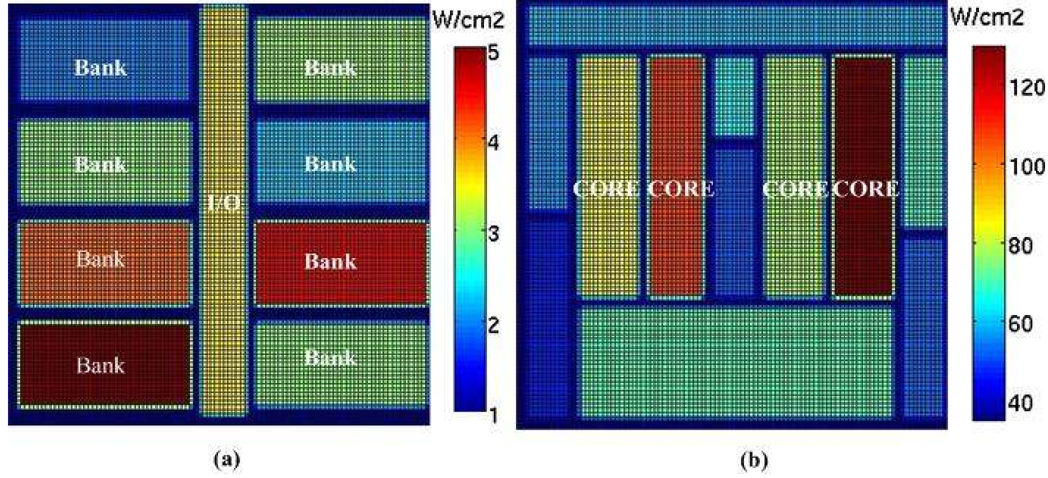


Figure 81: Power maps of the memory and processor tiers used in the finite difference modeling.

5.3.2 Uniformly Distributed TSVs vs. Clustered TSVs

In current approaches to 3D DRAM stacks and Wide I/O technologies, the TSVs are usually clustered instead of uniformly distributed. As a result of TSV clustering in the center of the stack, the thermal coupling increases in this region. Fortunately, the active memory cells are normally located away from the TSV cluster and, thus, the thermal coupling from the high-power tier to the active devices in the low-power tier is reduced. In this section, finite-difference modeling is used to understand the thermal coupling effect using uniformly distributed and clustered TSVs. The power maps of the memory tier and the processor tier are included in Figure 81.

The area containing the clustered TSVs has a high thermal conductivity and, thus, strong vertical thermal coupling. Fortunately, since the TSV cluster is away from the active memory cell (circuits within the yellow box, as shown in Figure 82(a)), the temperature of the memory cell is lower than the middle region of the chip.

In the simulations, the TSVs are only located in the center. The cluster is assumed to be 1 mm $\times$ 5 mm and contains 49 $\times$ 100 TSVs. The cluster is identified by the red rectangle in Figure 82(a). On the other hand, a uniformly distributed TSV array

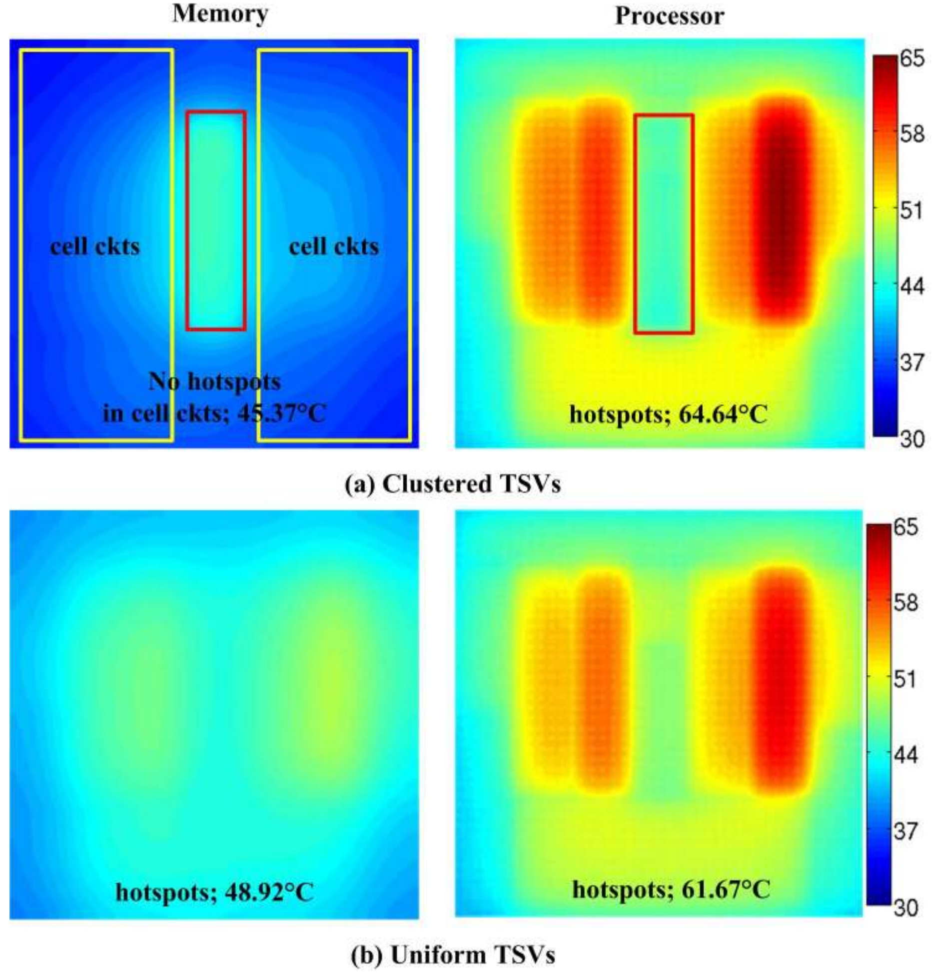


Figure 82: Power maps of the memory and processor tiers in (a) the clustered TSV case and (b) the uniform TSV case.

case is also simulated (4,900 TSVs). The results are shown in Figure 82(b).

In the clustered TSVs case, the maximum temperature of the whole DRAM die drops only by 3.55 °C compared to the uniform TSV case. However the maximum temperature of the cell array circuits is only 42.27 °C, which is a drop of 6.65 °C. By clustering the TSVs far from the memory cells, the most thermally-sensitive portion of the die is effectively isolated from the high-power die. Thus, we conclude that clustering TSVs can limit the localized thermal coupling in 3D stacks.

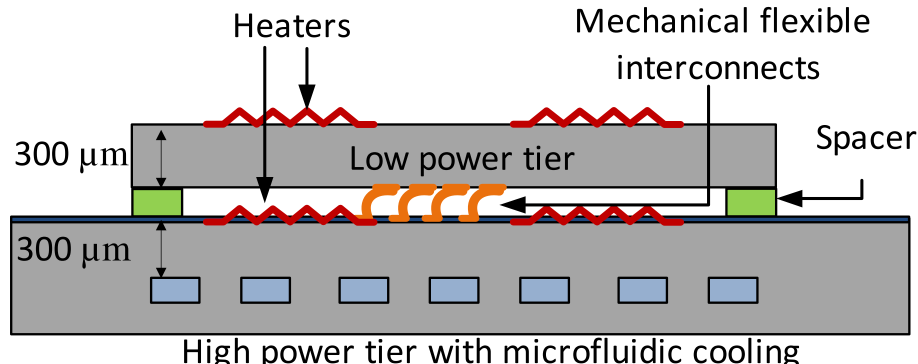


Figure 83: Schematic of the designed testbed for evaluation of the proposed thermal isolation technologies.

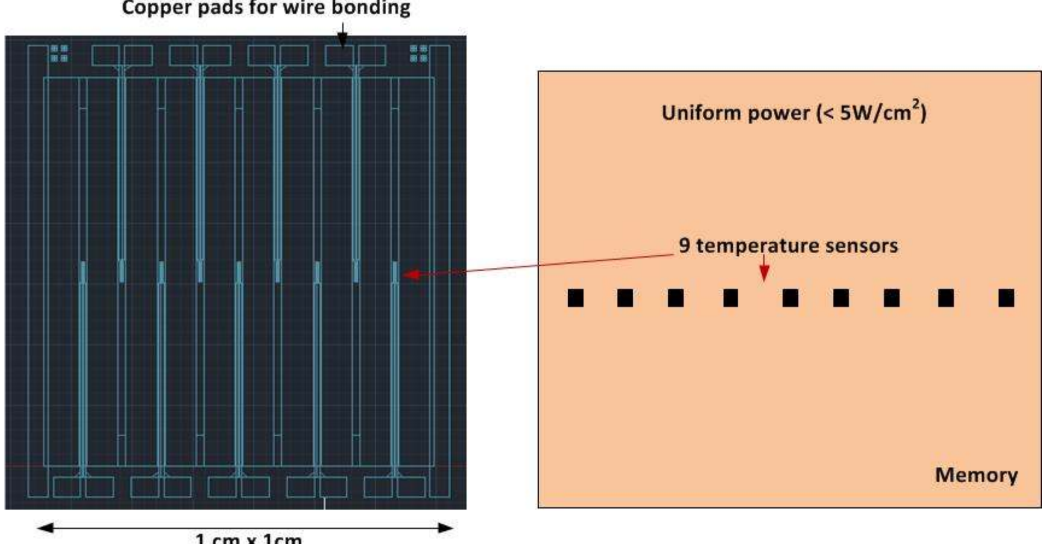


Figure 84: Layout (left) and schematic (right) of the power map designs of the top tier (low-power tier).

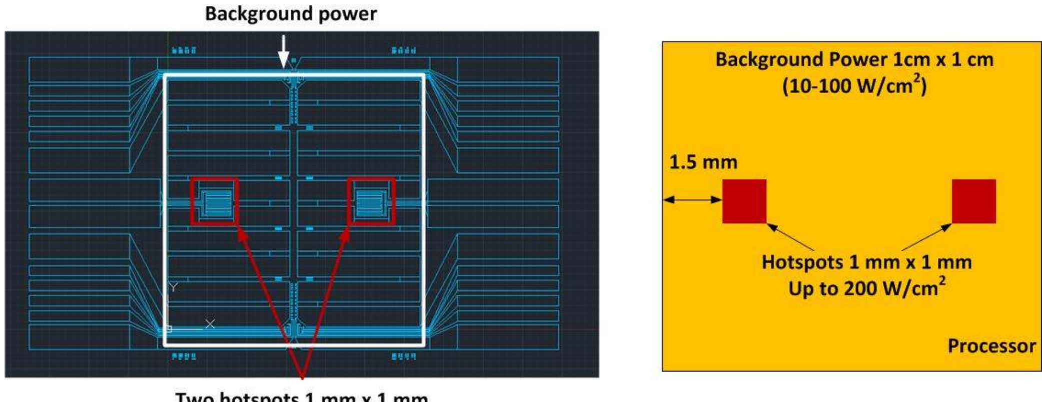
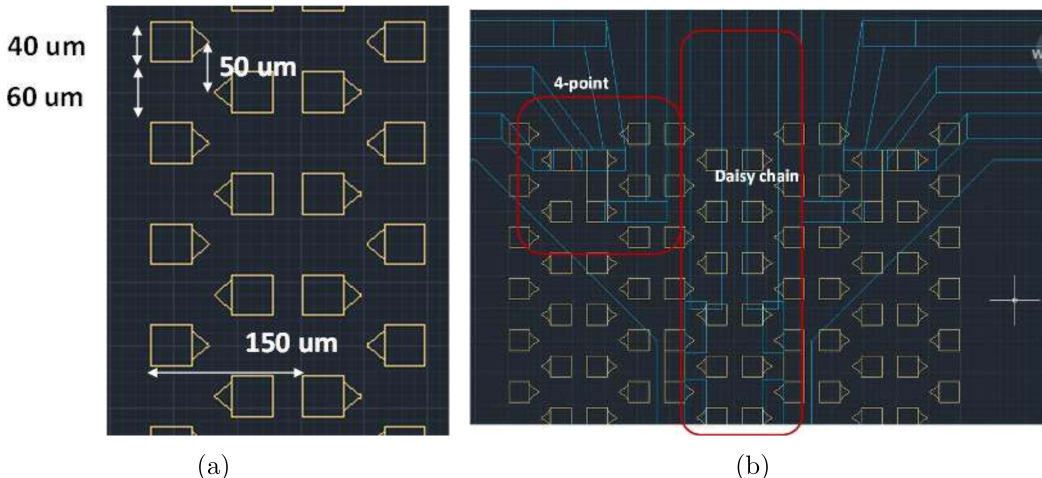


Figure 85: Layout (left) and schematic (right) of the power map designs of the bottom tier (high-power tier).

5.4 Design of the Testbed

Guided by the previous modeling and analysis, the thermal testbed is designed. A schematic illustration is shown in Figure 83. The testbed consists of a low-power tier and a high-power tier to emulate a heterogeneous 3D stack. The testbed is designed to emulate the proposed architecture shown in Figure 71. The microfluidic heat sink is integrated in the high-power tier (bottom tier). MFIs are used as interconnects between the two tiers and are designed to be clustered only in the middle of the die. Figure 84 shows the power map and temperature sensor designs for the low-power tier. The low-power tier dissipates a uniform power $\leq$ 5 W. A spiral heater spreads uniformly in a 1 cm $\times$ 1 cm area. Nine temperature sensors are inserted along the middle of the chip in order to measure temperature along the length of the chip (in other words, these temperature sensors are designed to capture the temperature gradient across the die). Since the MFIs are clustered in the middle region, the thermal coupling between the tiers is expected to be nonuniform across the chip; in particular from the center to the edges.

Figure 85 shows the power map and the schematic illustration of the high-power tier. The chip area is 1 cm $\times$ 1 cm. There are two hotspots on the chip and each



(a) (b)

Figure 86: Layout of (a) the MFI array and (b) the connections for daisy-chain resistance and four-point resistance measurements.

measures 1 mm × 1 mm, as shown in Figure 85. The hotspots are located 1.5 mm away from the edges.

The two chips are interconnected with an array of NiW MFIs. The MFI array design is shown in Figure 86(a). In this design, there are 12 columns by 100 rows, yielding a total number of 1,200 MFIs. This number is chosen based on the Wide I/Os specifications [71]. The MFI design, as shown in Figure 86(a), has a lateral pitch of 75 μm and a vertical pitch of 100 μm. The entire MFI array is 9,940 μm by 870 μm. Four-point resistance measurements and daisy-chain resistance measurements of 38 MFIs will be performed to verify electrical connectivity. Figure 86(b) shows the layout design of the four-point resistance measurement and daisy-chain resistance measurement.

5.5 Testbed Fabrication and Test Setup

5.5.1 Testbed fabrication

The process flows for the two tiers (low-power and high-power) are discussed in this section. For the low-power tier, the process begins with a double side-polished, 300 μm thick Si wafer (Figure 87). The bottom side of the wafer has 0.2 μm thick Si_3N_4

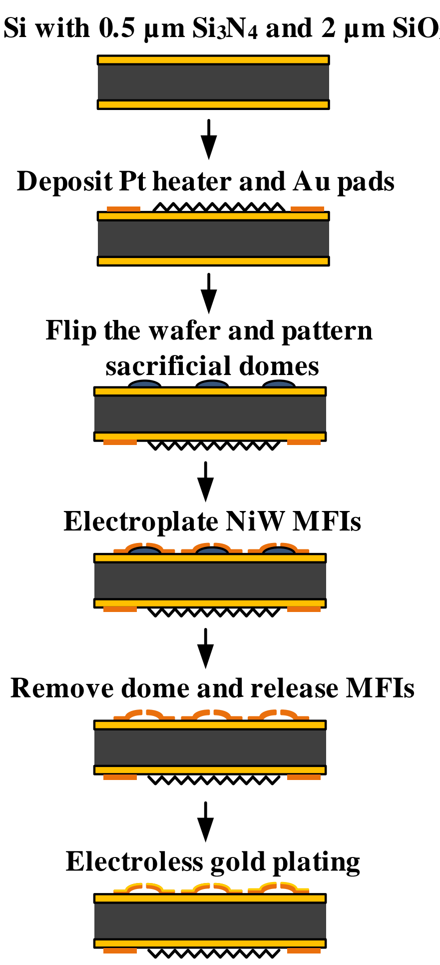


Figure 87: Process flow for the low-power tier.

for MFIs formation, and the top side has 2 μm thick SiO_2 for heater and temperature sensor formation. Next, the 0.2 μm thick Pt-based heaters/RTD are formed on the top side using lift-off. Figure 88 shows the nine RTDs and the RTD pads on the top die. Each RTD has dimensions of 500 μm x 88 μm and yields a resistance of approximately 200 Ω. The next step is to deposit 0.5 μm thick gold pads above the RTD pads. The gold pads facilitate wire-bonding, which is needed during testbed assembly. The sample is next flipped over for MFI-related processes. SPR220 is spun and patterned to form sacrificial squares. The squares then undergo a reflow process to form a dome structure with a height of 20 μm [72]. The wafer is then placed in a NiW electroplating solution to electroplate the MFIs to a thickness of 4.5 μm. After removing the sacrificial polymer dome beneath the MFIs, the MFIs become freestanding. Figure 89 shows images of the fabricated MFIs. Figure 89(a) shows the NiW MFIs electroplated on top of the sacrificial dome while Figure 89(b) shows an array of freestanding MFIs after dome removal. In Figure 89(b), the microscope image is focused on the top of the MFI and thus the anchor of the MFI is out of focus. The last fabrication step for the low-power tier is to passivate the MFI surface with gold by electroless plating. The gold passivation prevents NiW from oxidizing and also provides a lower electrical contact resistance. Figure 89(c) and 89(d) show images of the gold-passivated MFIs. The final height of the MFIs is 25 μm.

The process steps involved in the fabrication of the high-power tier are shown in Figure 90. The process starts with a double side-polished 500 μm thick Si wafer. Since the micropin-fins will be etched in this wafer, the wafer is chosen to be thicker to provide enough mechanical stability. The next step involves the deposition of a 2 μm thick SiO_2 layer on the top side. Next, 0.2 μm thick Pt heaters/RTDs and 0.5 μm thick gold pads are patterned using two lift-off steps. Figure 91 illustrates one of the hotspots on the high-power tier. The wafer is next flipped over and 200 μm-deep micropin-fins are etched using a standard Bosch etching process. Fluidic vias are

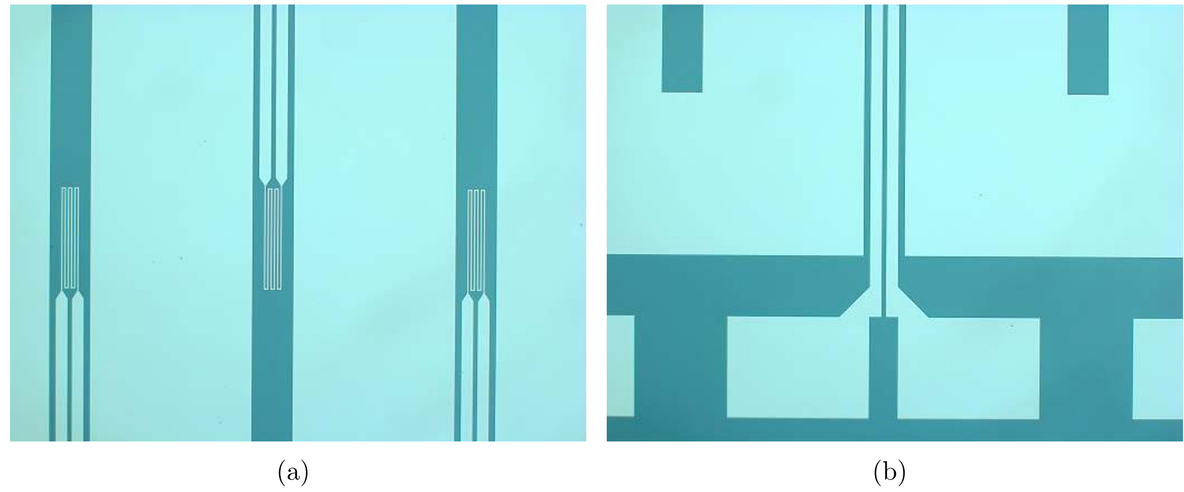
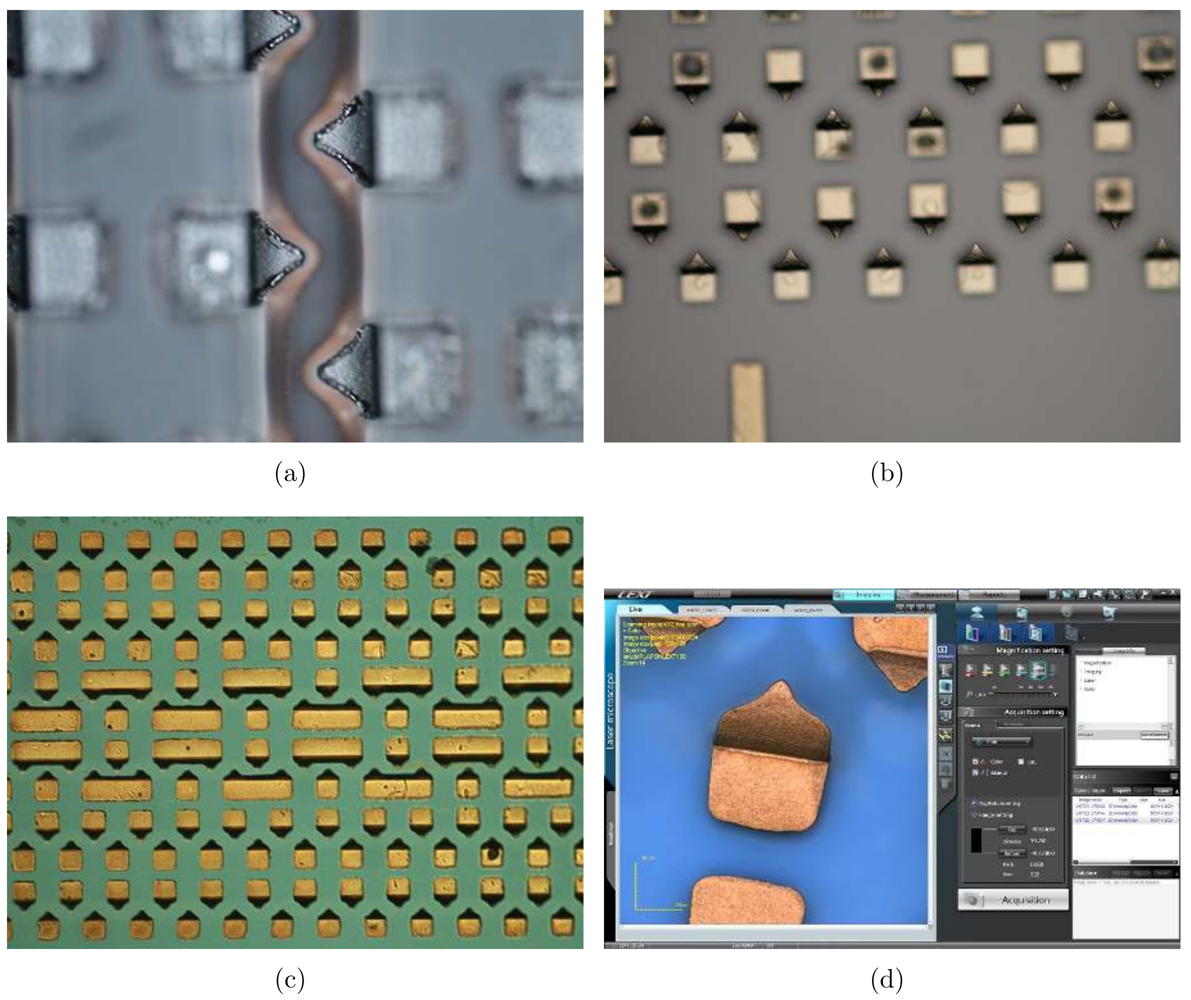


Figure 88: Images of (a) parts of the RTD array and (b) the pad of the RTD.

etched on a second wafer that is 300 μm thick and serves as a cover. The two wafers are then bonded using Si–Si fusion bonding and undergo an annealing process at 400 °C. The final step for preparing the high-power tier is to deposit polymer pillars on the heater side. The pillars serve as spacers to ensure a gap of greater than 10 μm.

5.5.2 Assembly

The wafer is next diced into individual dice, as shown in Figure 92. Figure 92(a) shows the diced bottom die (high-power die) before assembly. The background heater has dimensions of 1 cm × 1 cm and can be powered up to 100 W/cm^2. The two hotspots are 1 mm × 1 mm each and can be powered up to 200 W/cm^2. The temperature is measured using the heater/RTD and will then be compared with the temperature of the low-power tier. Daisy-chain resistance measurements and four-point resistance measurements of MFIs will also be performed to confirm assembly yield. The low-power and high-power tiers are then assembled using a Finetech sub-micron flip-chip bonder. Figure 93(a) is an image taken during flip chip bonding. After aligning the two tiers, the alignment head is placed in contact with the stack and applies a force of 15 N to bond the stack; while force is applied, epoxy is applied to the four corners



(a) (b) (c) (d)

Figure 89: Images of (a) the MFIs electroplated on top of the polymer dome, (b) the free standing MFIs after dome removal, (c) the MFI array with gold passivation, and (d) a single MFI with gold passiavtion.

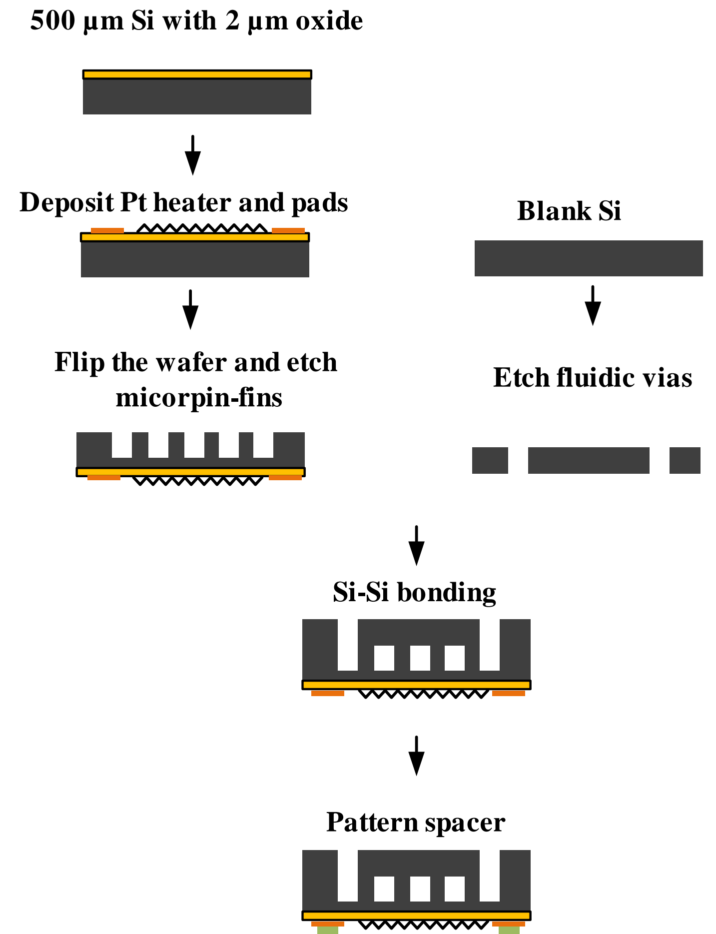


Figure 90: Process flow for the high-power tier.

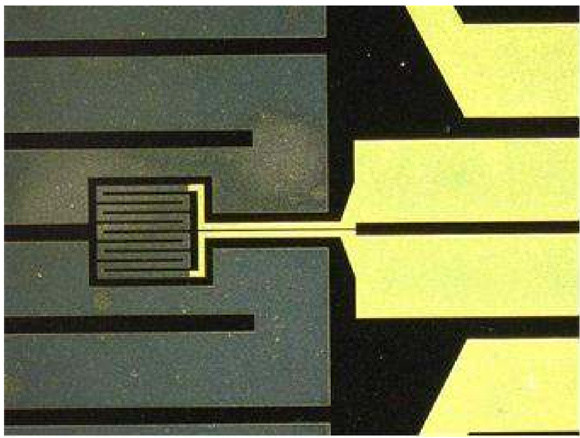
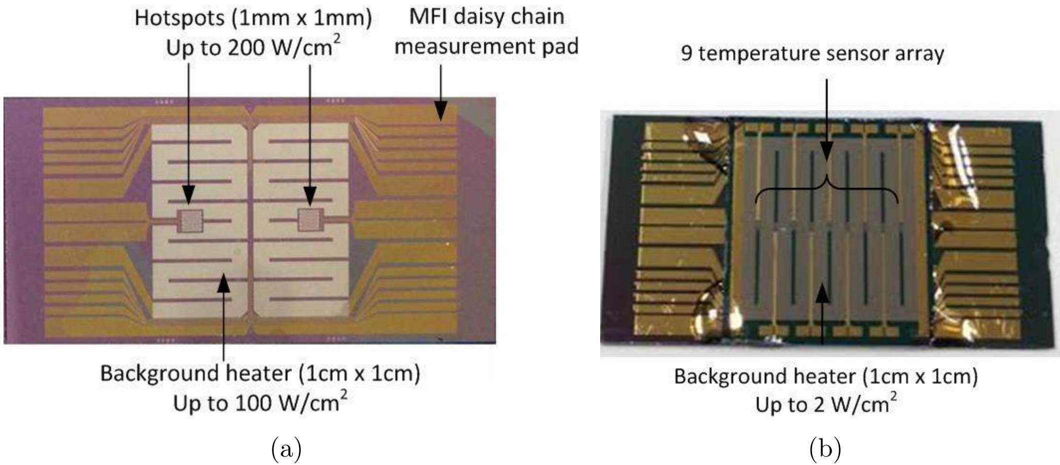


Figure 91: Optical image of one hotspot on the high-power tier.



(a) (b)

Figure 92: Images of (a) the bottom tier after dicing and (b) the assembled two-tier testbed.

to hold the position of the tiers. The force is released once the epoxy is dry. Figure 93(b) shows the alignment between the two tiers. Figure 94(a) shows an X-ray image of the bonded sample. The region within the black square is magnified and shown in Figure 94(b). The micropin-fins, the four-point resistance measurement structures, and the daisy-chain resistance measurement structures can be seen in Figure 94(b). Figure 92(b) shows a stack where the top die is bonded on the bottom die through flip-chip bonding.

5.5.3 Thermal and Electrical Test Setup

The microfluidic test setup is shown in Figure 95. Micropin-fins are only etched in the high-power tier (bottom tier), which means that the fluid is only flowing beneath the high-power tier. The top tier is bonded to the bottom tier through MFIs that are located in the center region. The stack is then bonded to a pre-designed PCB for testing (Figure 96(a)). Nanoports are attached to the bottom of the sample, as shown in Figure 96(b). An Agilent data logger is used to source current into the on-chip heater/RTDs on both tiers. The data logger is used to measure the resistance of the RTD on the top and bottom tiers and to extrapolate the junction temperatures using

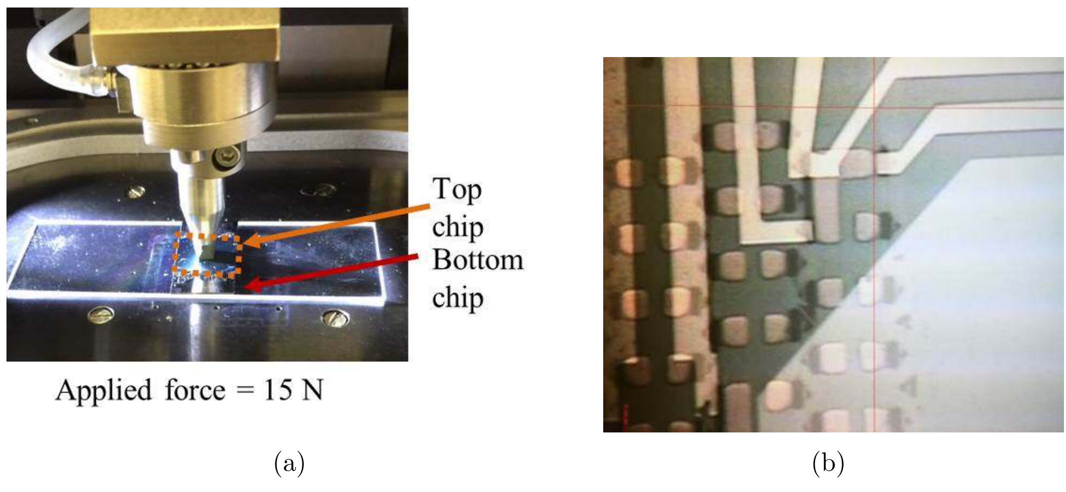


Figure 93: (a) Flip chip bonding assembly and (b) the alignment between the two tiers.

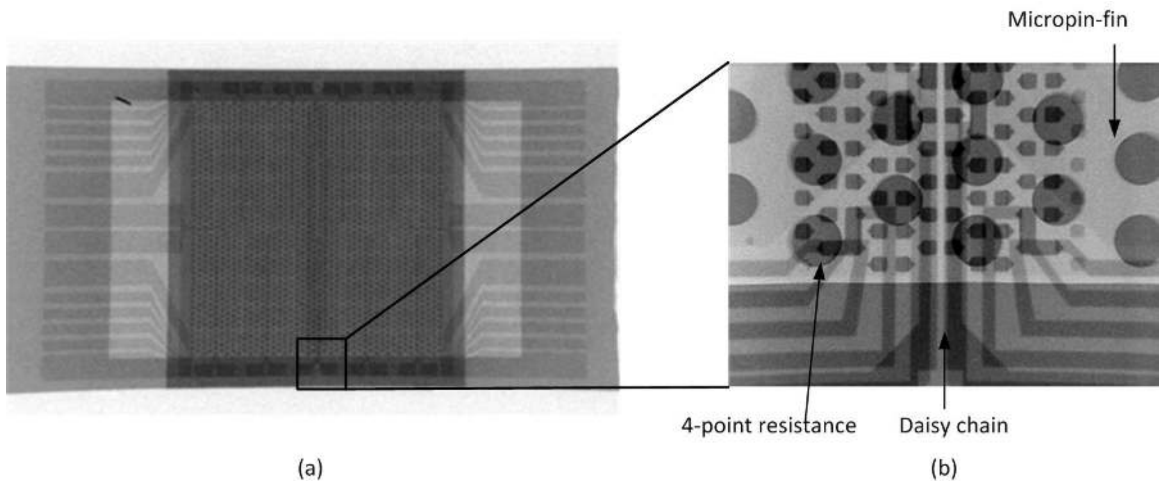


Figure 94: X-ray of (a) overall view of the boned chip and (b) a magnified view.

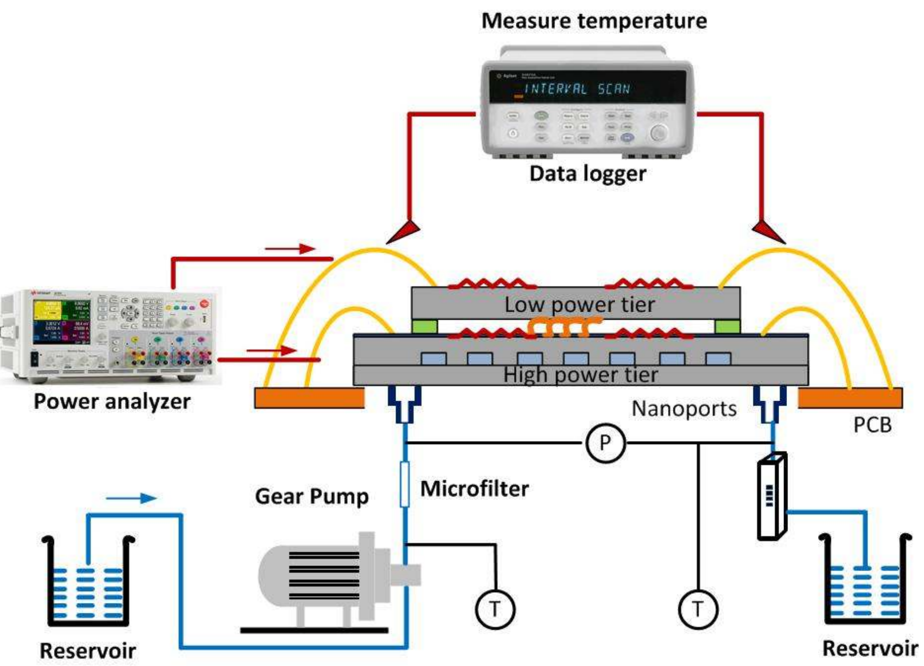


Figure 95: Microfluidic test setup to evaluate the thermal isolation technologies.

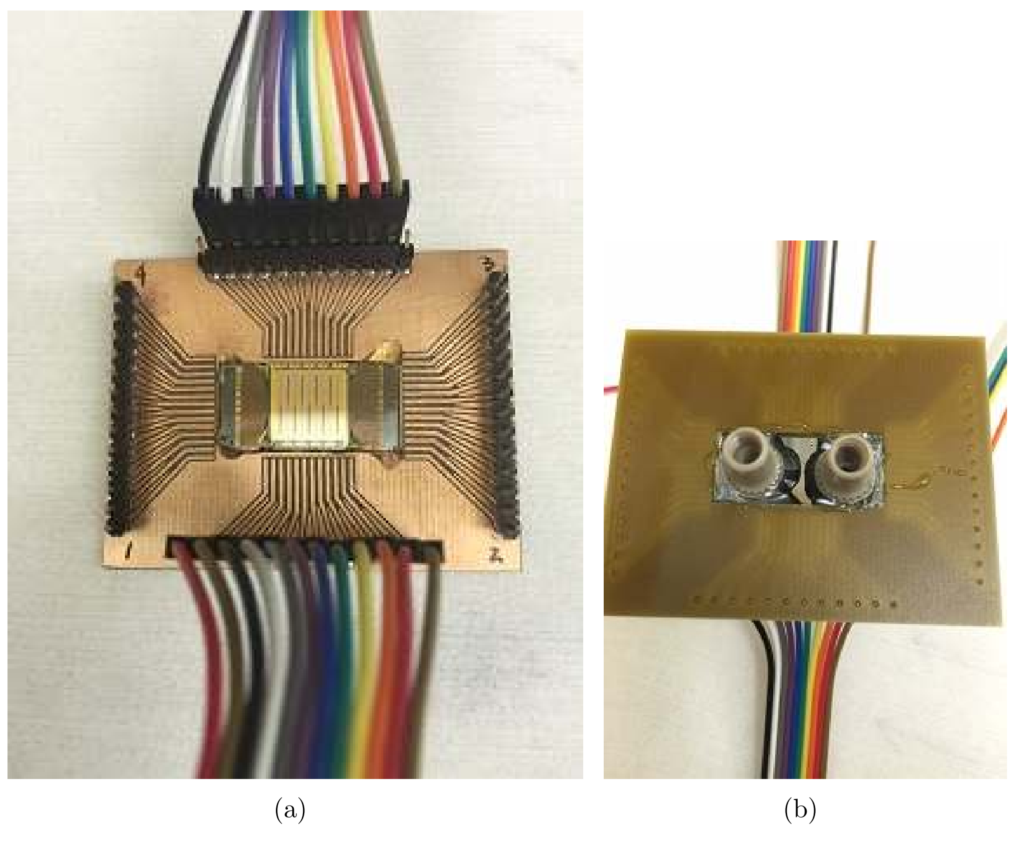
Eq. (16) in Section 3.3.

Figure 97 illustrates the test setup for the four-point resistance measurement of a single MFI. The measured resistance consists of the electrical resistance of a single MFI and contact resistance between the MFI and the gold pad.

5.6 Thermal and Electrical Experimental Results

To demonstrate the thermal isolation concept, several test cases are emulated as shown in Table 8 and Table 9. In all of the test cases, the inlet DI water temperature is 19.5 °C ± 0.5 °C. The room temperature is 22.5 °C ± 0.5 °C.

In the following sections, the test cases are compared in pairs to better understand the impact of thermal isolation. T_1 to T_9 represent the nine temperature sensors on the top die, as shown in Figure 84. They correspond to location 1 to location 9 on the chip, respectively. In the bottom tier, T_{bg1} and T_{ht1} represent the average of the left background region and the hotspot region, respectively; T_{bg2} and T_{ht2} represent



(a) (b)

Figure 96: (a) Top and (b) bottom view of the stack assembled to a PCB board using wire bonding.

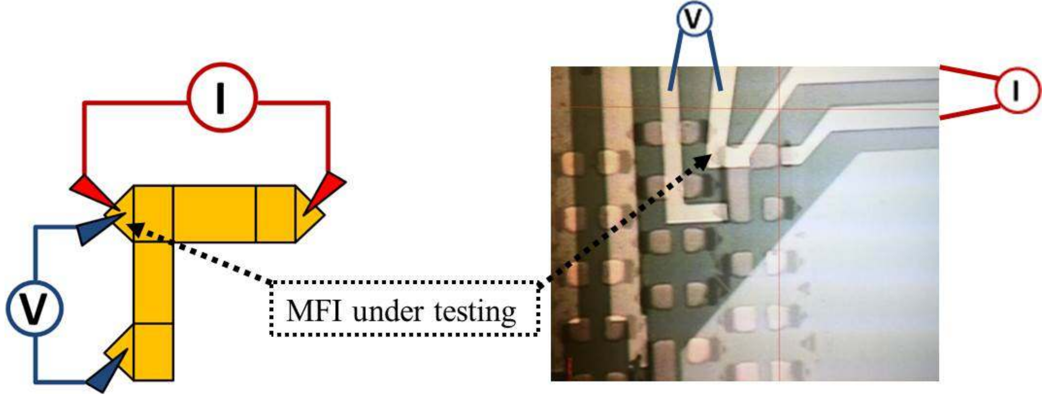


Figure 97: Four-point resistance measurement of MFI

Table 8: Summary of the temperature of the top and bottom tiers under different scenarios (Part I)

	Case A	Case B	Case C	Case D
P_{top}[1](W)	0.5	0.5	0.5	0.5
P_{bg1}[2](W)	4.8	4.8	4.8	4.8
P_{ht1}[3](W)	0.1	0.5	1.5	1.0
P_{bg2}[4](W)	4.8	4.8	4.8	4.8
P_{ht2}[5](W)	0.1	0.5	1.5	1.0
T_1 (°C)	20.9	20.8	21.1	21.0
T_2 (°C)	21.4	21.4	21.7	21.5
T_3 (°C)	21.8	21.8	22.1	21.9
T_4 (°C)	22.1	22.1	22.4	22.2
T_5 (°C)	22.5	22.6	22.6	22.5
T_6 (°C)	22.6	22.6	22.9	22.7
T_7 (°C)	22.7	22.7	23.0	22.8
T_8 (°C)	22.7	22.7	23.1	22.9
T_9 (°C)	23.0	22.8	23.1	22.8
T_{bg1}[6](°C)	21.5	21.5	22.0	21.7
T_{ht1}[7](°C)	22.1	24.5	31.4	27.8
T_{bg2}[8](°C)	22.8	23.0	23.6	23.2
T_{ht2}[9](°C)	22.8	25.6	33.0	29.2

[1] Power of the top tier.
[2] Background power of the left side of the bottom tier (inlet side).
[3] Hotspot power of the left side of the bottom tier (inlet side).
[4] Background power of the right side of the bottom tier (outlet side).
[5] Hotspot power of the right side of the bottom tier (outlet side).
[6] Background temperature of the left side of the bottom tier (inlet side).
[7] Background temperature of the left side of the bottom tier (inlet side).
[8] Background temperature of the right side of the bottom tier (outlet side).
[9] Background temperature of the right side of the bottom tier (outlet side).

the average of the right background region and the hotspot region, respectively. The junction temperature at the center of the chip is computed as the average of T_{bg1} and T_{bg2}.

5.6.1 Thermal Testing I: Powering the high-power tier

In a heterogeneous 3D stack consisting of a low-power die bonded on top of a high-power die using microbumps and underfill, thermal coupling is expected to result in an increase of the temperature in the low-power die when the high-power tier is powered. In this subsection, this scenario is emulated using the thermal isolation

Table 9: Summary of the temperature of the top and bottom tiers under different scenarios (Part II)

	Case E	Case F	Case G	Case H	Case I
P_{top}[1](W)	0.5	0.5	0.5	0.5	0.5
P_{bg1}[2](W)	14.2	14.2	14.2	0	0
P_{ht1}[3](W)	1.0	1.5	1.5	1.5	2.0
P_{bg2}[4](W)	14.3	14.3	0	0	0
P_{ht2}[5](W)	1.0	1.5	1.5	1.5	2.0
T_1 (°C)	24.0	24.2	23.3	19.6	19.7
T_2 (°C)	24.9	25.1	24.1	20.0	20.2
T_3 (°C)	25.6	25.8	24.4	20.2	20.4
T_4 (°C)	26.2	26.4	24.6	20.3	20.5
T_5 (°C)	30.0	29.8	24.7	20.4	20.6
T_6 (°C)	27.2	27.4	24.8	20.6	20.8
T_7 (°C)	27.6	27.8	24.8	20.6	20.8
T_8 (°C)	27.9	28.1	24.8	20.6	20.8
T_9 (°C)	28.3	29.2	24.5	20.4	20.6
T_{bg1}[6](°C)	28.6	28.8	28.6	19.3	19.2
T_{ht1}[7](°C)	31.6	35.1	35.0	29.3	32.8
T_{bg2}[8](°C)	32.4	32.8	23.4	19.8	19.8
T_{ht2}[9](°C)	36.4	40.1	33.9	29.3	33.0

[1] Power of the top tier.
[2] Background power of the left side of the bottom tier (inlet side).
[3] Hotspot power of the left side of the bottom tier (inlet side).
[4] Background power of the right side of the bottom tier (outlet side).
[5] Hotspot power of the right side of the bottom tier (outlet side).
[6] Background temperature of the left side of the bottom tier (inlet side).
[7] Background temperature of the left side of the bottom tier (inlet side).
[8] Background temperature of the right side of the bottom tier (outlet side).
[9] Background temperature of the right side of the bottom tier (outlet side).

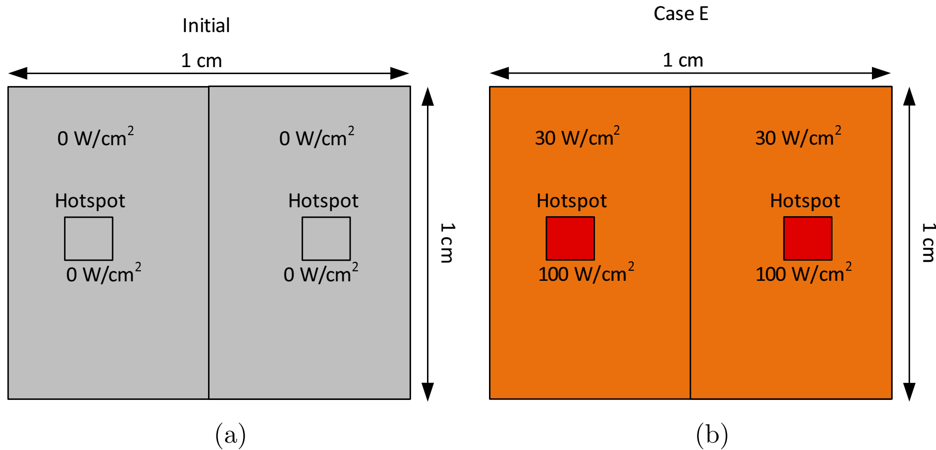


Figure 98: (a) Initial case when the high-power tier dissipates 0 W and (b) Case E in Table 9 where the background power density is 30 W/cm^2 and the hotspot power density is 100 W/cm^2.

testbed.

The power maps of the high-power tier are shown in Figure 98. The low-power tier dissipates 0.5 W in all the evaluated cases. In the initial case, the high-power tier does not dissipate any power. In Case E (as listed in Table 9), the background power density is 30 W/cm^2 while the hotspots dissipate 100 W/cm^2. The junction temperature across the two tiers is plotted in Figure 99. In the initial case, the temperature of both tiers is close to the inlet water temperature. When the bottom tier is powered and dissipates a background power density of 30 W/cm^2 and a hotspot power density of 100 W/cm^2 (power map shown in Figure 98(b)), the temperature of the bottom tier increases at all locations. The temperature of the left and right sides of the background heater increases to 28.6 °C and 32.4 °C, respectively. The temperature of the left and right side hotspots increases to 31.6 °C and 36.4 °C, respectively. However, the average temperature of the upper tier increases to 26.9 °C. The temperature of the upper tier follows the same temperature trend of the bottom tier. However, owing to the thermal isolation technology, the temperature increase is not as high as the bottom tier. An interesting point to be noted is that

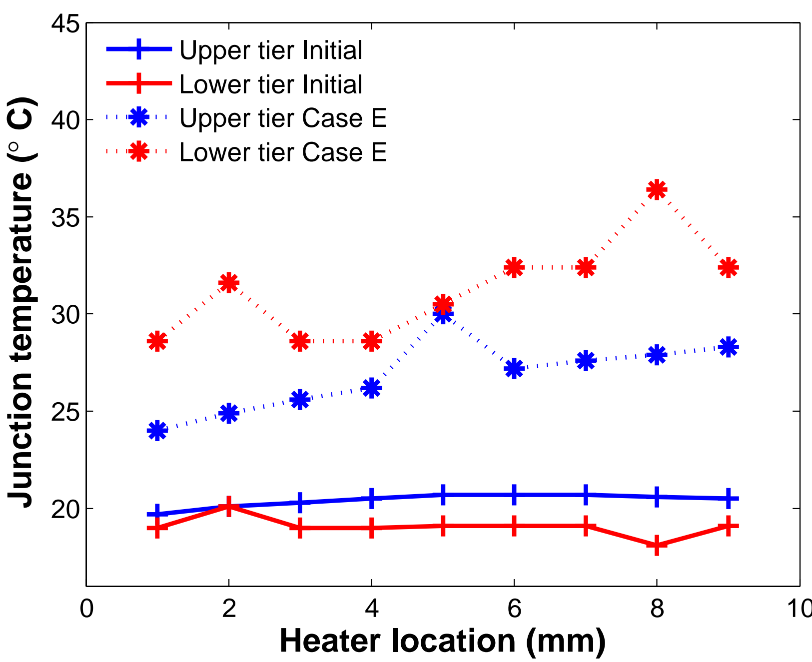


Figure 99: Junction temperature fluctuation before and after the high-power tier is powered.

the highest temperature of the upper tier is located at the center of the die. The temperature of both tiers is very close at the center. This effect can be attributed to the dense MFI array that is clustered in the middle and, thus, creates a good thermal path. This phenomenon provides confidence in having good contact between the tiers using MFIs. Another point to be noted is that the temperature of the upper die gradually increases from inlet to outlet. One reason is that it follows the same temperature trend of the bottom tier. The other reason is introduced from the actual testbed. In the testbed, epoxy is used at the four corners to securely bond the upper die to the bottom die, and thus heat can be conducted through the epoxy. When the temperature of the coolant becomes elevated at the outlet, it also impacts the temperature of the upper die at the outlet. Therefore, the temperature at location 9 is higher than that at location 1. This effect induced by the epoxy is taken into account when we compare the measurements to the finite difference modeling in Section 5.7.

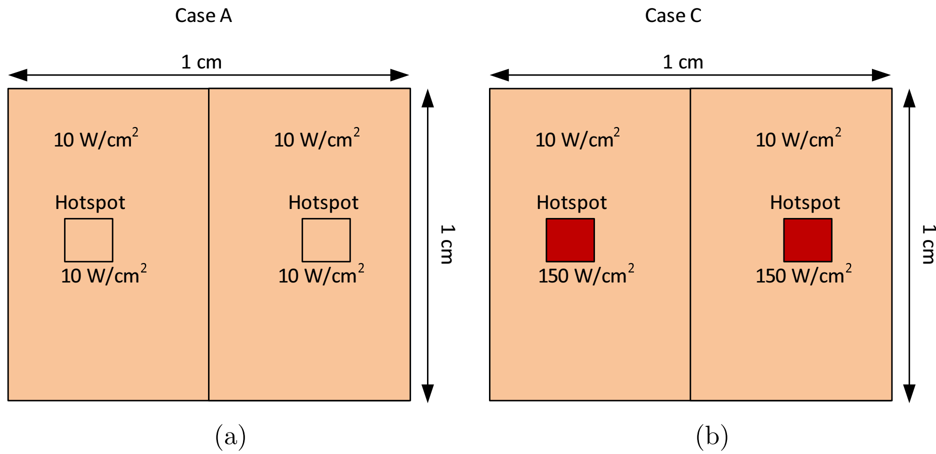


Figure 100: (a) Uniform power density of 10 W/cm^2 in the bottom tier (Case A) and (b) background power of 10 W/cm^2 plus two hotspots each dissipates 150 W/cm^2 (Case C).

5.6.2 Thermal Testing II: Minimize the Hotspot Coupling

Hotspot cooling is a critical issue for today's high performance computers. By s-tacking a low-power die with a high-power die, the hotspots can also occur in the low-power die because of the thermal coupling. This presents a number of challenges for temperature-sensitive low-power dice. Therefore, the cases where hotspots occur in the high-power tier are emulated in this subsection.

The power maps of the simulated cases are illustrated in Figure 100. In Figure 100(a), the bottom tier dissipates 10 W/cm^2 across the chip. The junction temperature for each location on both tiers is plotted in Figure 101 (Case A). Next (Case C), the power density of the two hotspots increases to 150 W/cm^2 while the background power remains unchanged (Figure 100(b)). The corresponding temperature of each chip is plotted in Figure 101 (Case C). In Case A, the temperature is relatively flat indicating uniform temperature without hotspots. When the hotspot region dissipates more power, one obvious observation is that there are two peak temperatures that occur in the bottom die. This is expected because of the large power density

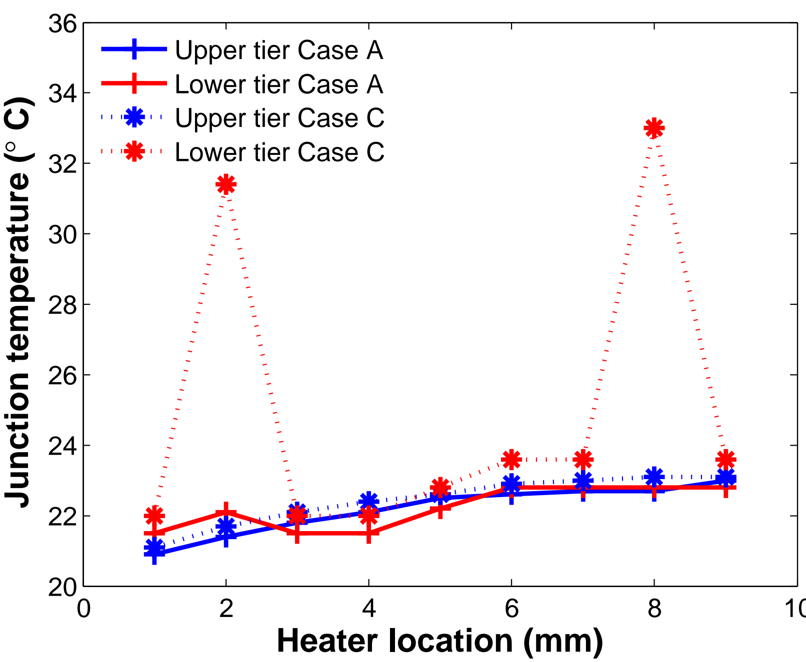


Figure 101: Junction temperature fluctuation of top and bottom tiers in Case H and Case I in Table 9.

of the hotspot region. The two peak temperatures are 31.4 °C and 33.0 °C, respectively. However, also in Case C, there are no obvious hotspots in the upper tier. The temperature of the upper tier gradually increases from 21.1 °C to 23.1 °C. This demonstrates that the proposed thermal isolation concept effectively minimizes the hotspot coupling between the vertical tiers.

To illustrate this point, two extreme cases are emulated where only the hotspot regions are powered while the background dissipates no power. In addition, the power density of the hotspot increases to 200 W/cm^2. The two power maps are illustrated in Figure 102(a) and Figure 102(b). The corresponding temperature in the two cases is plotted in Figure 103. In these two cases, the temperature of the bottom tier is close to room temperature except for the two hotspots where the temperatures are 29.3 °C and 33 °C for Case H and Case I, respectively. Even though the power density of the hotspot is high, the total power is low since each hotspot is only 1 mm × 1 mm.

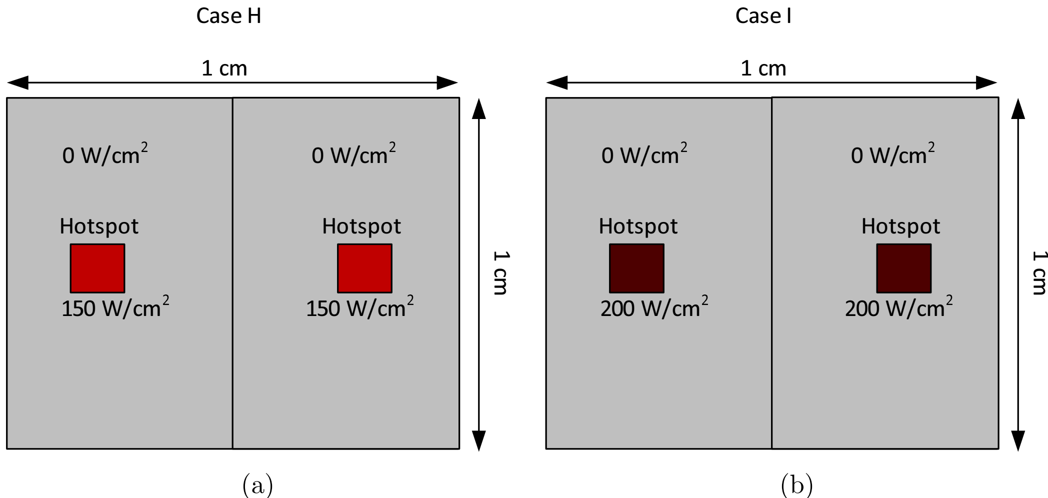


Figure 102: (a) Zero background power with two hotspots each dissipates 150 W/cm^2 (Case H) and (b) zero background power with two hotspots each dissipates 200 W/cm^2 (Case I).

The fluid temperature barely increases after flowing through the hotspot. Thus, the hotspots near the inlet and outlet have the same temperature. In Cases H and I, the temperature at location 2 in the upper tier is 20 °C and 20.2 °C, respectively. However, the temperature at location 2 in the bottom tier is 29.3 °C and 32.8 °C, respectively. The maximum junction temperature difference is 12.6 °C. For reference, in Figure 99, when the bottom chip is not dissipating any power, the junction temperature at location 2 in the upper tier is also 20 °C. The temperature barely changes in the upper tier after the hotspot power increases. This demonstrates that the thermal isolation technology with MFIs has greatly decreased the hotspot coupling between tiers.

5.6.3 Thermal Testing III: Bottom Tier Power Increases

In the previous subsection, we demonstrated that the thermal isolation technology can prevent vertical coupling and thus 'protect' the low-power tier from the hotspots in the high-power tier. In this subsection, the bottom tier dissipates an elevated power density of 30 W/cm^2 in addition to the hotspots. In the two cases (Cases E

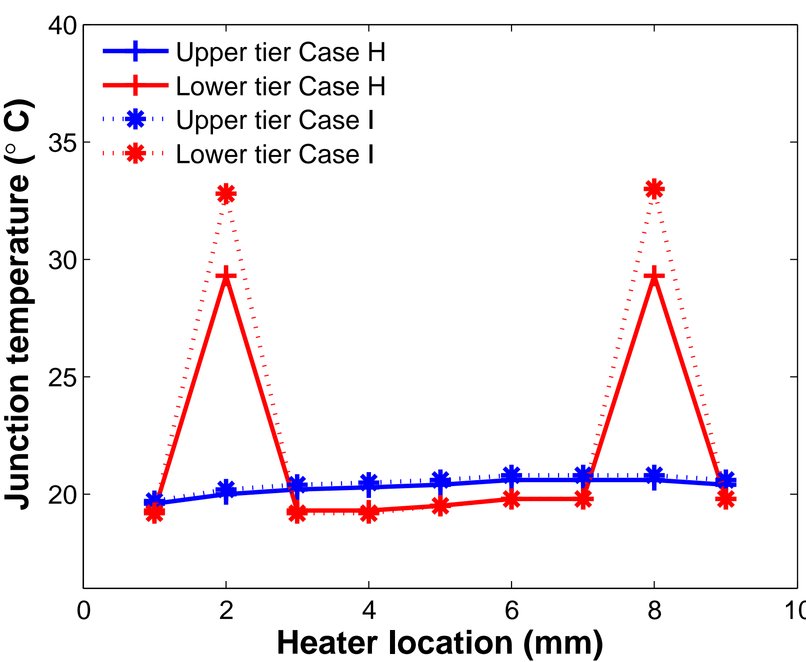


Figure 103: Junction temperature fluctuation of the top and bottom tiers in Case H and Case I in Table 9.

and H listed in Table 9), the two hotspots dissipate 100 W/cm^2 and 150 W/cm^2, respectively. The corresponding temperature of the two tiers in the two cases is plotted in Figure 104.

At location 2 and location 8, where the hotspots are located, the temperature difference between the top and bottom tier is large. For example, the temperature difference between the two tiers at location 8 in Case H is 12 °C. However, at the other locations, the temperature difference is less than that of the hotspot. For example, the temperature of location 9 in the upper and bottom tiers is 32.8 °C and 29.2 °C, respectively. The thermal isolation effect is weakened in this case since the coolant temperature increases as it absorbs heat from the bottom tier. The higher the power the bottom tier dissipates, the warmer the coolant is. The elevated coolant temperature in turn causes the temperature to increase in the upper tier. This effect is expected when the two tiers share the same microfluidic heat sink especially for

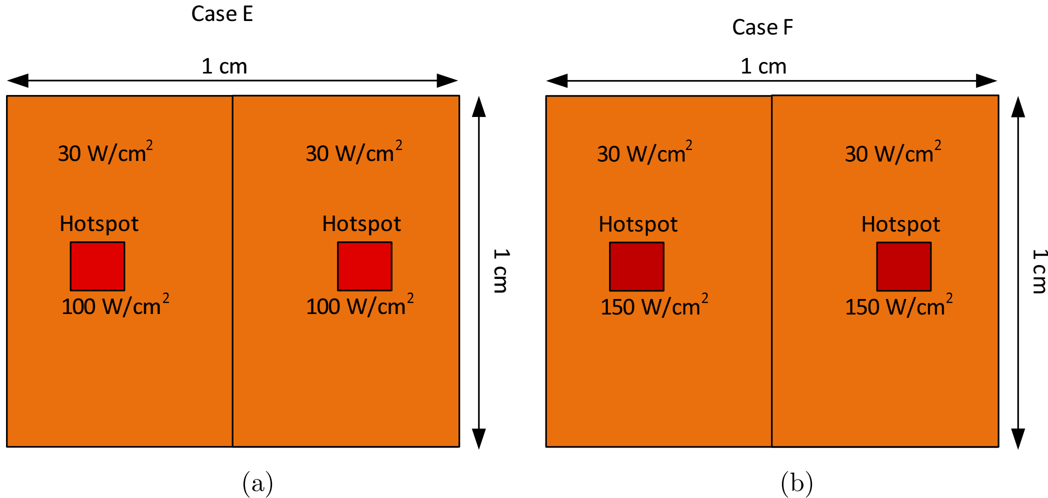


Figure 104: (a) Background power of 30 W/cm^2 plus two hotspots each dissipates 100 W/cm^2 (Case E) and (b) background power of 30 W/cm^2 plus two hotspots each dissipates 150 W/cm^2 (Case F).

the locations near the outlet. On the other hand, the measured temperature in the bottom tier is an average of half of the die. The actual temperature of location 9 should be higher than the average temperature. One method to eliminate the impact of warm coolant in the upper tier is to allocate an independent microfluidic heat sink to it.

Another interesting observation is that the temperature is highest at the center of the upper tier and similar to that of the bottom tier. The reason is that the MFIs are densely clustered in the middle. This is a desirable result and complies with the Wide I/O technology. In the Wide I/O technology, all TSVs are located in a rectangular array in the middle of the chip. There are no active devices in the middle region. Although the middle region is warmer because of the heat conduction through the dense MFI array, the device region is actually cooler. Compared to evenly distributed TSVs, our layout can reduce the temperature of the active regions by sacrificing the less critical middle region (the region for I/Os).

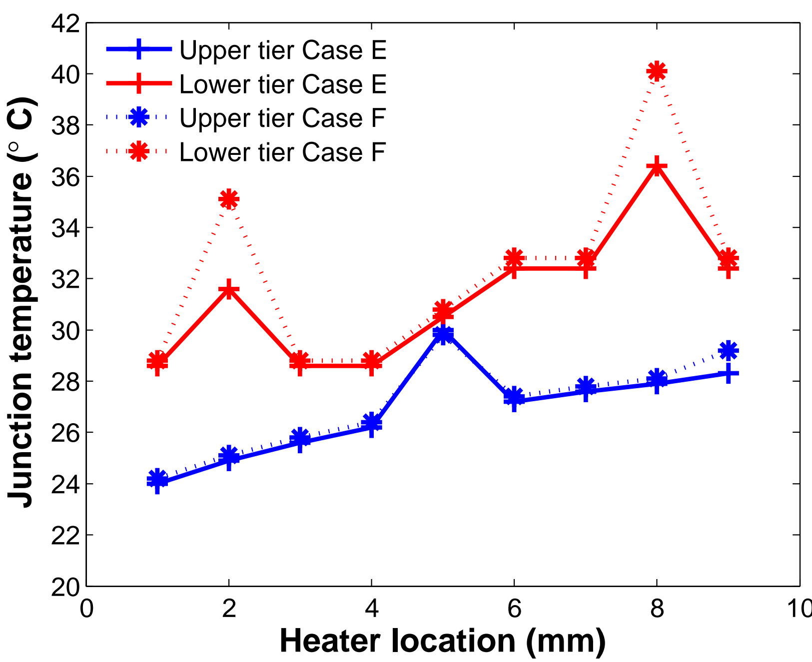


Figure 105: Junction temperature fluctuation of the top and bottom tiers in Case E and Case F in Table 9

5.6.4 Electrical Testing of MFIs

In order to demonstrate the electrical connectivity between the two tiers after bonding, two sets of electrical resistance measurements are performed. The four-point resistance measurements are done at four locations on the sample. The measured electrical resistance is 46.49 mΩ. The measured resistance consists of the resistance of the MFI, part of the landing pad, and the contact resistance. The daisy chain resistance of 38 MFIs is also measured during the thermal measurements. At room temperature, the resistance of the daisy chain including the leading wires is 19.55 Ω. When temperature of the bottom chip increases, the daisy-chain resistance also increases. The highest measured resistance during all thermal testing is 19.77 Ω. The daisy-chain resistance provides confidence that all the electrical contacts remain throughout thermal testing.

Table 10: Parameters used in the finite difference model

	Conductivity	Thickness
	(W/mK)	(μm)
Memory die	149	300
Underfill layer	0.9	25
Air gap	2.4E-2	25
Processor die	149	300
Micro-bump	60	25
Copper	400	N/A
MFIs	10	25
SiO_2	1.38	2.5

5.7 Validation by Finite Difference Modeling

The experimental results are used to validate our finite-difference modeling. The assumptions of the boundary conditions used in the model are made based on measurement results. For microfluidic cooling in the bottom tier, the convective heat transfer coefficient is assumed to be 5.2 $\times 10^4$ W/m^{2}K. For the cooling of the top tier, a heat transfer coefficient of 1.3 $\times 10^4$ W/m^{2}K is assumed. This convective boundary condition is only applied on the edges of the top chip. The calculation is based on the initial case (discussed in Section 5.6.1). In this case, the top tier dissipates 0.5 W while the bottom tier dissipates 0 W. The equivalent thermal resistance is calculated to be 3.4 K/W for the top chip. All the parameters and boundary conditions used in the model are included in Table 10 and Table 11. Two examples of the measured and modeled results are listed in Table 12. In all the simulated cases, the error is within 2 °C.

In Section 5.6, the measured temperatures of the bottom tier are average temperatures. Using finite-difference modeling, the localized temperatures can be obtained. Using the validated model, we re-plot the top and bottom junction temperature in Cases E and F, as shown in Figure 106. This allows us to directly compare the temperatures at the exact same locations in the two tiers. This figure can be compared with the measured results shown in Figure 104. The temperature difference between

Table 11: Boundary conditions assumed in the finite difference model

Stack	Boundary face	Heat Transfer Coefficient (W/Km^2)
Conventional microbump stack	MFHS (bottom)	52000
	Others(near adiabatic)	5
Our proposed stack	Top	13000
	MFHS (bottom)	52000
	Side (near adiabatic)	5

Table 12: Comparison between the measured and modeled junction temperatures

	Measured Case E	Modeled Case E	Measured Case F	Modeled Case F
P_{top}[1](W)	0.5	0.5	0.5	0.5
P_{bg1}[2](W)	14.2	14.2	14.2	14.2
P_{ht1}[3](W)	1.0	1.0	1.5	1.5
P_{bg2}[4](W)	14.3	14.3	14.3	14.3
P_{ht2}[5](W)	1.0	1.0	1.5	1.5
T_1 (°C)	24.0	23.3	24.2	23.4
T_2 (°C)	24.9	24.2	25.1	24.4
T_3 (°C)	25.6	25.6	25.8	25.8
T_4 (°C)	26.2	27.5	26.4	27.7
T_5 (°C)	30.0	28.2	29.8	28.5
T_6 (°C)	27.2	28.3	27.4	28.5
T_7 (°C)	27.6	27.8	27.8	28.0
T_8 (°C)	27.9	27.8	28.1	28.0
T_9 (°C)	28.3	28.0	29.2	28.2
T_{bg1}[6](°C)	28.6	28.6	28.8	28.9
T_{ht1}[7](°C)	31.6	32.7	35.1	35.6
T_{bg2}[8](°C)	32.4	32.5	32.8	32.9
T_{ht2}[9](°C)	36.4	37.1	40.1	40.2

[1] Power of the top tier.
[2] Background power of the left side of the bottom tier (inlet side).
[3] Hotspot power of the left side of the bottom tier (inlet side).
[4] Background power of the left side of the bottom tier (outlet side).
[5] Hotspot power of the left side of the bottom tier (outlet side).
[6] Background temperature of the left side of the bottom tier (inlet side).
[7] Background temperature of the left side of the bottom tier (inlet side).
[8] Background temperature of the right side of the bottom tier (outlet side).
[9] Background temperature of the right side of the bottom tier (outlet side).

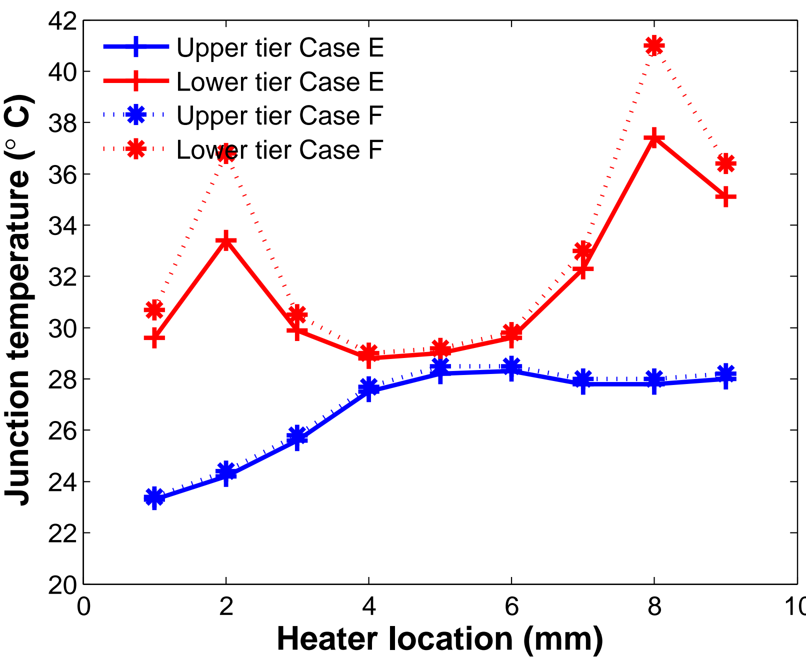


Figure 106: Junction temperature in Case E and Case F (as listed in Table 9 using the finite-difference model. This figure can be compared with the measured results shown in Figure 105.

the two tiers at location 1 and 9 are much larger than that those at locations 4, 5, and 6. The reason is that heat conduction occurs through the array of MFIs and, thus, leads to enhanced thermal coupling in the center region of the dice.

To benchmark with conventional 3D integration scenarios, the same stack is modeled with microbumps and underfill, as shown in Figure 107. For a fair comparison, the same number of MFIs and microbumps is assumed. Table 10 and Table 11 list all the parameters used in the model. The power map in Case F (as listed in Table 9) is used in this simulation. In this case, the background power density is 30 W/cm^2 while the hotspots dissipate 150 W/cm^2. In the case without thermal isolation, we can see that the temperature of the upper tier follows the trend of the lower tier. In most locations, the temperature is similar in both tiers. At the hotspot near the outlet, the temperature of the upper tier and the lower tier is 35.9 °C and 38.7 °C,

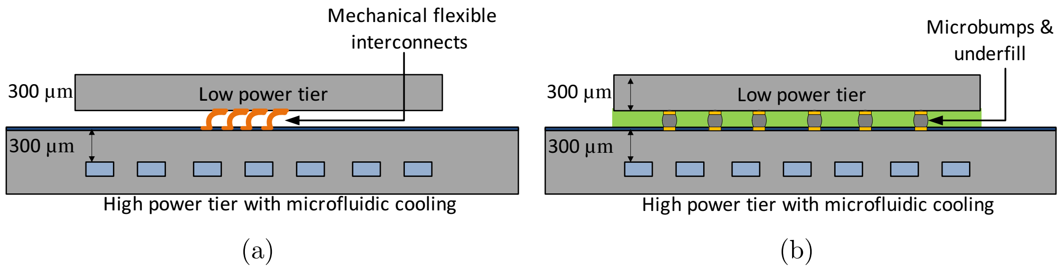


Figure 107: The modeled heterogeneous stack with (a) MFI and air cavity and (b) microbumps and underfill.

respectively. For the case with thermal isolation, the temperature of the upper tier and the lower tier is 28 °C and 41 °C, respectively. The thermal isolation technology is shown to reduce the upper-tier temperature by 8 °C at location 8 and, thus, yields a temperature reduction of 19.5 %, while the bottom-tier temperature increases by 2 °C. This is because the upper tier helps to spread the heat and thus lower the hotspot temperature of the bottom tier. At locations without hotspots such as location 9, thermal isolation is also observed. In the case without thermal isolation, the temperature of the upper tier and the lower tier is 35.7 °C and 35.8 °C, respectively, while with thermal isolation, the temperature of the upper tier drops to 28.2 °C without causing the temperature of the lower tier to increase.

The implication from the analysis in Section 5.6.3 is that allocating an independent microfluidic heat sink to the low-power die may further decouple it from the high-power die. Therefore, the proposed concept with thermal bridge and independent microfluidic heat sink is modeled (as shown in Figure 109) and benchmarked with the conventional microbump and underfill approach. In the ideal thermal isolation case, the temperature of the high-power and low-power tier at location 8 is 40.8 °C and 23.0 °C, respectively. While in the conventional bonding scenario, the temperature of the high-power and low-power tier is 38.7 °C and 35.9 °C, respectively. A temperature reduction of 35.9 % is achieved in the low-power tier by implementing the MFIs and independent microfluidic heat sink.

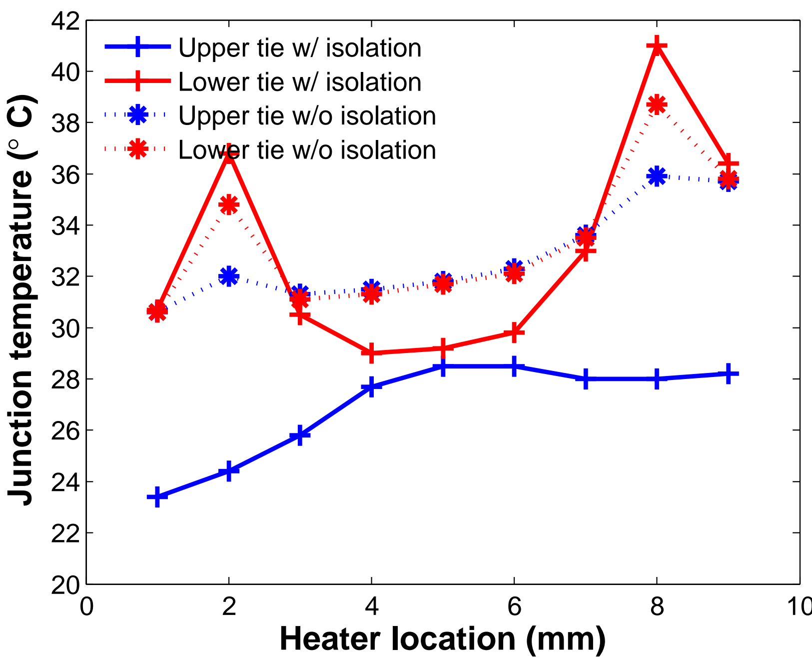


Figure 108: Junction temperature in both tiers with and without the thermal isolation. In the case without thermal isolation, microbumps and underfill are integrated between the tiers.

5.8 Conclusion

For heterogenous 3D integration including high-power dice and low-power dice (e.g., memory and nanophotonics), thermal coupling is a critical issue. The proposed thermal isolation technology features an air/vacuum cavity between the tiers, MFIs as the interconnects, and a thermal bridge for the isolated chips. To demonstrate the thermal isolation technology, a two-tier testbed with heterogeneous elements is designed, fabricated, and tested. Various thermal test cases are evaluated. The proposed technology effectively decouples the two tiers thermally. One case shows that the proposed technology effectively prevents hotspots on the high-power chip from coupling to the low-power tier. For example, the low-power tier remains at 20.8 °C while the hotspot in the high-power tier reaches 32.8 °C. For elevated power density on the bottom tier, the top tier also becomes warmer because of the temperature increase of the coolant

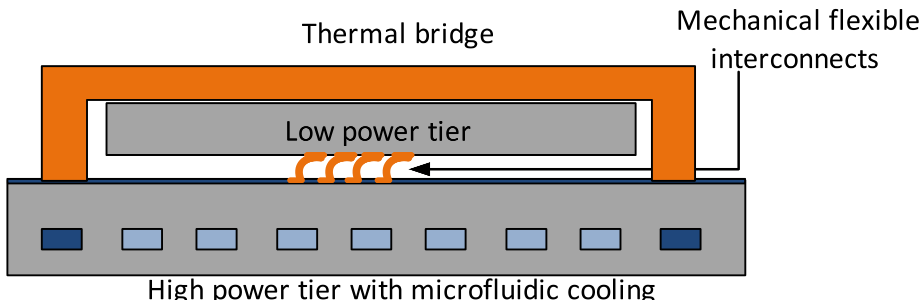


Figure 109: A heterogeneous 3D stack with MFIs and independent microfluidic heat sink for the low-power die.

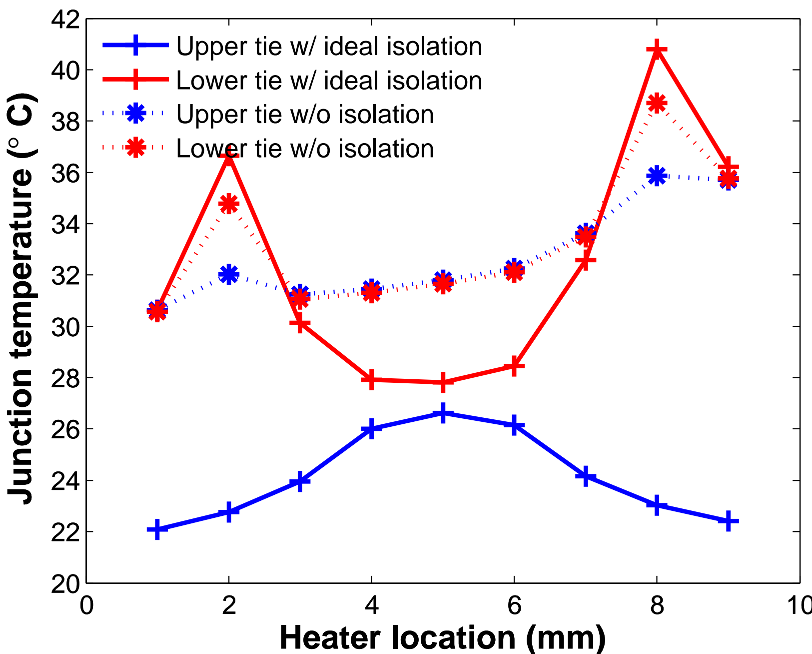


Figure 110: Benchmark the ideal thermal isolation technology with conventional 3D stacking approach.

at the outlet. Independent microfluidic heat sinks can solve this issue.

Four-point resistance measurements of the MFIs is performed along with a resistance of an MFI daisy chain. The measured resistance demonstrates the electrical connectivity between the two tiers at all time during the thermal measurements.

Finite-difference modeling is used to validate the experimental results. All errors are within 2 °C. The thermal isolation technology is also benchmarked with a stack that contains conventional microbumps and underfill. The temperature of the upper tier (low-power tier) is reduced by 12.9 °C by implementing our proposed thermal isolation technology.

CHAPTER VI

SUMMARY AND FUTURE WORK

6.1 Summary of the Presented Work

The objective of this research is to propose and implement a hybrid thermal solution combining effective cooling solutions that scale with the number of dice in the stack and effective thermal isolation solutions to 'protect' the low-power tiers from the high-power tiers in the stack.

6.1.1 Advanced Microfluidic Cooling Solution for 3D ICs Containing High Power Chips

The thermal challenge is one of the primary issues for 3D ICs, especially for stacks containing multiple high-power chips. Because of the increased power density, 3D IC stacks go beyond the cooling capability of conventional air cooling. This motivates our work on effective within-tier microfluidic cooling research. On the other hand, TSVs are key enablers of 3D ICs and will need to be co-integrated with microfluidic heat sinks. Therefore, designing a microfluidic heat sink without considering TSV fabrication compatibility and TSV parasitics greatly diminishes the advantages of 3D ICs.

In Chapter II, thermal electrical modeling are developed in order to capture the trade-off between microfluidic heat sinks and TSVs. Based on the trade-off analysis, the height of microfluidic heat sinks is most critical to the TSV electrical parasitics. By setting a target thermal resistance and TSV parasitics, a MPFHS with a diameter of 150 μm, a pitch of 225 μm, and a height of 200 μm is designed. Novel liner and heterogeneous TSV integration are also proposed to further lower the TSV parasitics. In addition, TSV aspect ratios need to be scaled up in order to minimize the parasitics

and ensure high TSV density. The analysis motivates our work in Chapter III.

In Chapter III, a MPFHS that is compatible with TSV technology is designed, fabricated, and thermally tested. This is a solution that addresses the cooling needs of 3D ICs while accounting for TSV fabrication compatibility and electrical performance (minimizing TSV parasitics). In the test case, a staggered MPFHS is shown to provide a thermal resistance as low as 0.269 K·cm^2 /W at a flow rate of 70 mL/min for a heat sink height of 200 μm. In addition, this result is benchmarked against a state-of-art air-cooled heat sink. Based on the experimental data, microfluidic cooling provides a lower chip junction temperature with a much smaller heat sink volume compared to air cooling. Finally, in order to demonstrate the compatibility with TSVs, high aspect ratio (18:1) TSVs are integrated in MPFHS. The four-point resistance of the TSVs is found to be 36.5 $\pm$ 1.5 mΩ.

In Chapter IV, the microfluidic cooling is implemented in a 3D stack. The microfluidic heat sink has the same geometries as described in Chapter III. Memory-on-processor and processor-on-processor are emulated using the testbed. In both cases, microfluidic cooling outperforms air cooling. In addition, a tier-specific cooling mechanism that allows tailoring the flow rate according to the power dissipation of each tier is implemented. This method is shown to be able to minimize the thermal gradient between tiers and thus minimize the thermal-mechanical stress. Pumping power is also reduced by 37.5% by preventing overcooling of the low-power chip. In the end, microfluidic cooling is evaluated in a multi-core chip. The lateral thermal coupling is observed to be due to the warmed fluid. A lateral thermal gradient-caused leakage power increase is analyzed. Vertical thermal coupling is also emulated. To reduce the vertical thermal coupling, each high-power tier should have its own microfluidic heat sink.

6.1.2 Advanced Thermal Isolation Technology for Heterogeneous 3D ICs

For heterogeneous 3D integration including high-power dice and low-power dice (e.g., memory and nanophotonics), thermal coupling is a critical issue. Using an air cavity between the tiers is proposed to mitigate the vertical thermal coupling. Mechanically flexible interconnects are also integrated between the vertical tiers for power and signaling.

In Chapter V, the thermal isolation technology with air gap is firstly modeled and compared with conventional microbump technology. A two-dimensional resistance network model is developed to analyze the stack with air cavity. When the power density of the processor tier increases from 50 W/cm^2 to 100 W/cm^2, with a 5 μm thick air cavity, the temperature of the bottom tier increases by 13 °C, while the temperature increment will be 22 °C without air cavity [69]. However, when interconnects are taken into consideration, the thermal isolation effect is weakened. To solve this problem, we propose to place all the interconnects in the middle region. By locating the active devices away from the middle, the thermal impact can be further reduced.

Guided by the analysis obtained from finite-difference modeling, we have designed a two-tier heterogeneous testbed where MFIs are clustered in the middle region. The testbed is fabricated and tested. It is shown that the proposed technology effectively decouples the two tiers thermally. For example, the low-power tier remains at 20.8 °C while the temperature of the hotspot in the high-power tier is 32.8 °C.

The results are also simulated using the finite-difference modeling and the error is shown to be less than 2 °C. We also simulate the same heterogeneous stack with conventional microbumps and underfill. Significant temperature reduction in the low-power tier is shown in all cases.

6.2 Future Work

The opportunities for advancing the technologies in this dissertation will be discussed in the following sections. Firstly, opportunities to advance microfluidic cooling are described. Secondly, opportunities to advance thermal isolations are describe. Lastly, opportunities to advance the thermal-electrical analysis are discussed.

6.2.1 Explore a System with Interposer Cooling

As part of the future research, a microfluidic cooled interposer can be studied. For the applications where only one high-power tier is included in a stack (Figure 111), integrating the microfluidic heat sink in the interposer may be efficient enough. There are three main benefits of integrating microfluidic cooling in the interposer:

1. This method does not use the precious on-chip resources. There is no need of on-chip fluidic I/Os which consume the on-chip surface area. With a within-tier microfluidic heat sink, TSVs can only be routed through the areas with no flow path. Integrating microfluidic heat sinks in the interposer will eliminate these constraints for TSV placement.

2. Integrating the microfluidic cooling in the chip will increase chip thickness and thus result in longer TSVs with high parasitics (as discussed in Chapter II). Integrating the microfluidic heat sink in the interposer allows the chip to be as thin as possible and thus reduce TSV parasitics and increase TSV density.

3. The fabrication of the interposer can be done separately with the fabrication of the CMOS chip; this allows more process flexibility and eases the fabrication constraints of the microfluidic heat sink.

However, the chip is bonded on the interposer through microbumps (Figure 111). Because of the increased conductive thermal resistance across the microbumps and

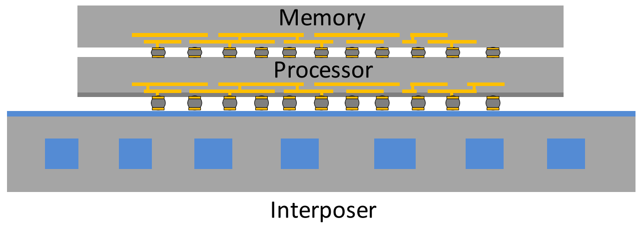


Figure 111: Illustration of a 3D stack with a microfluidic cooled interposer.

the on-chip interlayer dielectric (ILD), the cooling capability is not as good as within-tier microfluidic cooling. MPFHSs with the same design can be implemented in the interposer. The cooling capability (thermal resistance as a function of the flow rate) should be characterized and compared with within-tier microfluidic cooling.

6.2.2 Advancing the Thermal Isolation Technology

The proposed thermal isolation technology was described in Chapter V (Figure 71). The novelties in the proposed architecture not only include the vertical thermal isolation technology, but also include other novel concepts that are described as follows:

1. Interposer-level multi-optimized microfluidic heat sink
2. Thermal bridge that interconnects the memory tier to the interposer-level microfluidic heat sink

The future research opportunities in the proposed architecture are discussed in the following sections.

6.2.2.1 Interposer-level Multi-optimized Microfluidic Heat Sink

Independent microfluidic heat sinks that are dedicated to cool different dice can be integrated in one interposer. Figure 112 shows the top view of an example of the interposer with independent microfluidic heat sinks. In the example shown, the auxiliary heat sinks (shown in light blue) are dedicated to the low-power tiers while the main heat sink (shown in dark blue) in the center is dedicated to high-power tiers.

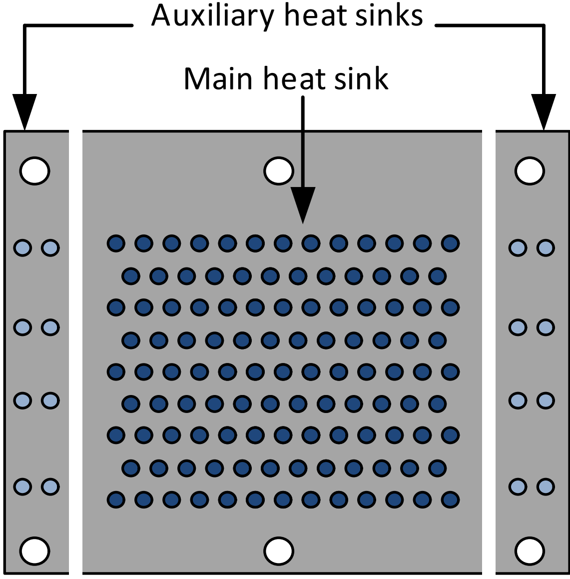


Figure 112: Illustration of multi-optimized microfluidic heat sinks.

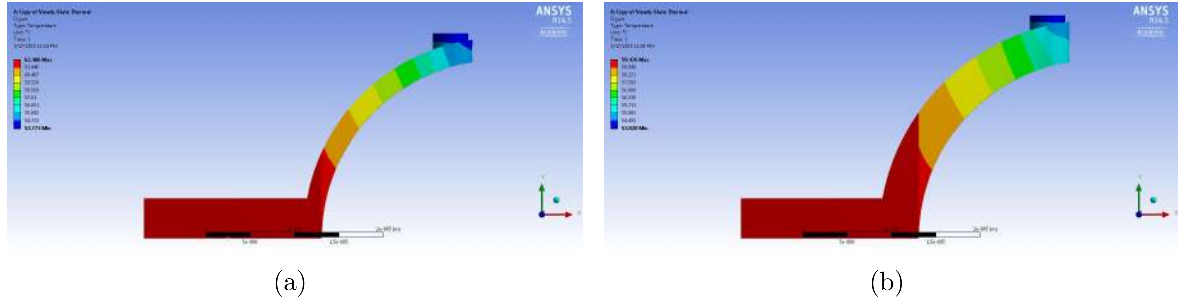
Because of the different power loads, the heat sinks can have different designs and different flow rates. A denser micropin-fin design can be adopted in the main heat sink, compared to the auxiliary heat sink.

6.2.2.2 Thermal Bridge Demonstration

The thermal bridge that we have modeled is made of copper and is attached to the interposer through TIM. Copper has a much higher CTE than Si and may create stress on the silicon die. Therefore, a thermal bridge made of Si can be a promising alternative. Although the thermal conductivity of Si is one third of that of copper, Si introduces no CTE mismatch issue and the fabrication is CMOS-compatible. The fabricated testbed uses epoxy to hold the top chip in place. The proposed thermal bridge needs to be manufactured and characterized in the testbed.

6.2.2.3 Electrical Thermal Co-optimization of MFIs

The MFIs in the present work are designed to have small pitch and low thermal resistance. However, the electrical resistance is not taken into consideration. If the



(a) (b)

Figure 113: (a) MFI with a thickness of 2 μm and (b) MFI with a thickness of 4.5 μm.

MFIs are designed thicker, the electrical resistance can be smaller. But thicker MFIs are more thermally conductive and, thus, degrade the thermal isolation. MFIs with a height of 20 μm are simulated to show the trade-offs between thermal resistance and electrical resistance. MFIs with thickness of 2 μm and 4.5 μm are simulated, as shown in Figure 113. The electrical resistance of the 2 μm thick MFI and 4 μm thick MFI are 64.4 mΩ and 41.9 mΩ, respectively. The thermal resistance of the 2 μm thick MFI and 4 μm thick MFI are 10.8 $\times 10^3$ K/W and 7.05 $\times 10^3$ K/W, respectively. More analysis of MFIs with different thickness, shapes, and materials should be done for different applications.

6.2.2.4 Thermal Isolation Using Vacuum

As discussed in Section 5.2, vacuum provides superior thermal isolation than air. Methods to create reliable and sustainable vacuum in microelectronics needs to be developed.

6.2.3 System Performance Implications

Power consumption and system throughput are temperature dependent. It is shown that by reducing the chip temperature from 88 °C to 47 °C, the total power of a high-performance chip decreases from 102 W to 83 W for the same operating frequency [58]. The present electrical analysis includes TSV capacitance. The electrical modeling can

be extended to include bandwidth density, system throughput, energy-per-bit, etc. With the extracted thermal data from the microfluidic cooling and thermal isolation technologies, electrical system performance can be analyzed. Currently, the trade-offs between TSV capacitance and thermal resistance are analyzed. In the future, the lateral interconnects can also be included in the trade-off analysis.

In the applications involving silicon photonics, a model that captures the energy dissipation of the photonic link should be developed. The model should analyze the energy dissipation with and without the thermal isolation technology.

REFERENCES

[1] J. Jeddeloh and B. Keeth, “Hybrid memory cube new dram architecture increases density and performance,” in *Proc. Symposium on VLSI Technology (VLSIT)*, 2012, pp. 87–88.

[2] U. Kang, H.-J. Chung, S. Heo, D.-H. Park, H. Lee, J.-H. Kim, S.-H. Ahn, S.-H. Cha, J. Ahn, D. Kwon, J.-W. Lee, H.-S. Joo, W.-S. Kim, D. H. Jang, N. S. Kim, J.-H. Choi, T.-G. Chung, J.-H. Yoo, J.-S. Choi, C. Kim, and Y.-H. Jun, “8 Gb 3-D DDR3 DRAM using through-silicon-via technology,” *IEEE J. of Solid-State Circuits*, vol. 45, no. 1, pp. 111–119, 2010.

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

[4] Semiconductor Industry Association. International technology roadmap of semiconductors. Semiconductor Industry Association. [Online]. Available: http://public.itrs.net

[5] S.-C. Lin and K. Banerjee, “Cool chips: Opportunities and implications for power and thermal management,” *IEEE Trans. Electron Devices*, vol. 55, no. 1, pp. 245–255, 2008.

[6] D. Tuckerman and R. F. W. Pease, “High-performance heat sinking for VLSI,” *IEEE Electron Device Lett.*, vol. 2, no. 5, pp. 126–129, May 1981.

[7] M. Bakir, C. King, D. Sekar, H. Thacker, B. Dang, G. Huang, A. Naeemi, and J. Meindl, “3D heterogeneous integrated systems: Liquid cooling, power delivery, and implementation,” in *Proc. IEEE Custom Integrated Circuits Conference*, 2008, pp. 663–670.

[8] 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, Jan. 2009.

[9] J.-M. Koo, S. Im, L. Jiang, and K. E. Goodson, “Integrated microchannel cooling for three-dimensional electronic circuit architectures,” *J. Heat Transfer*, vol. 127, no. 1, pp. 49–58, 2005.

[10] M. S. Bakir and J. D. Meindl, *Integrated Interconnect Technologies for 3D Nanoelectronic Systems.* Boston: Artech House, 2008.

[11] L. Jiang, J. Mikkelsen, J.-M. Koo, D. Huber, S. Yao, L. Zhang, P. Zhou, J. G. Maveety, R. Prasher, J. G. Santiago *et al.*, “Closed-loop electroosmotic microchannel cooling system for VLSI circuits,” *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 25, no. 3, pp. 347–355, 2002.

[12] D. Liu and S. V. Garimella, “Analysis and optimization of the thermal performance of microchannel heat sinks,” *International J. of Numerical Methods for Heat and Fluid Flow*, vol. 15, no. 1, pp. 7–26, 2005.

[13] Y. Peles, A. Kosar, C. Mishra, C. Kuo, and B. Schneider, “Forced convective heat transfer across a pin fin micro heat sink,” *International J. Heat Mass Transfer*, vol. 48, pp. 3615–3627, Aug. 2005.

[14] T. Brunschwiler, S. Paredes, U. Drechsler, B. Michel, W. Cesar, G. Toral, Y. Temiz, and Y. Leblebici, “Validation of the porous-medium approach to model interlayer-cooled 3D-chip stacks,” in *Proc. IEEE International Conference on 3D System Integration*, 2009, pp. 1–10.

[15] N. Khan, H. Yu, T. S. Pin, S. W. Ho, N. Su, W. Y. Hnin, V. Kripesh, Pinjala, J. Lau, and T. K. Chuan, “3D packaging with through ilicon via (TSV) for electrical and fluidic interconnections,” in *Proc. IEEE Electronic Components and Technology Conference*, 2009, pp. 1153–1158.

[16] L. Zheng, Y. Zhang, and M. Bakir, “Design, fabrication and assembly of a novel electrical and microfluidic I/Os for 3-D chip stack and silicon interposer,” in *Proc. IEEE Electronic Components and Technology Conference (ECTC)*, May 2013, pp. 2243–2248.

[17] B. Shi and A. Srivastava, “TSV-constrained micro-channel infrastructure design for cooling stacked 3D-ICs,” in *Proc. ACM international symposium on International Symposium on Physical Design*, 2012, pp. 113–118.

[18] H. B. Chang, H. Y. Chen, P. C. Kuo, C. H. Chien, E. Liao, T. C. Lin, T. S. Wei, Y. C. Lin, Y. H. Chen, K. F. Yang, H. A. Teng, W. C. Tsai, Y. C. Tseng, S. Y. Chen, C. C. Hsieh, M. F. Chen, Y. H. Liu, T. J. Wu, S. Hou, W. C. Chiou, S. P. Jeng, and C. H. Yu, “High-aspect ratio through silicon via (TSV) technology,” in *Proc. Symposium on VLSI Technology (VLSIT)*, 2012, pp. 173–174.

[19] Y. Civale, S. Armini, H. Philipsen, A. Redolfi, D. Velenis, K. Croes, N. Heylen, Z. El-Mekki, K. Vandersmissen, G. Beyer, B. Swinnen, and E. Beyne, “Enhanced barrier seed metallization for integration of high-density high aspect-ratio copper-filled 3D through-silicon via interconnects,” in *Proc. IEEE Electronic Components and Technology Conference (ECTC)*, 2012, pp. 822–826.

[20] A. Yu, J. Lau, S. W. Ho, A. Kumar, W. Y. Hnin, W. S. Lee, M. C. Jong, V. Sekhar, V. Kripesh, D. Pinjala, S. Chen, C.-F. Chan, C.-C. Chao, C.-H. Chiu, C.-M. Huang, and C. Chen, “Fabrication of high aspect ratio TSV and assembly

with fine-pitch low-cost solder microbump for Si interposer technology with high-density interconnects," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. 1, no. 9, pp. 1336–1344, Sept. 2011.

[21] R. Ayoub, R. Nath, and T. Rosing, "Joint energy thermal and cooling management for memory and CPU subsystems in servers," in *Proc. IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2012, pp. 1–12.

[22] I. Paul, S. Manne, M. Arora, W. L. Bircher, and S. Yalamanchili, "Cooperative boosting: needy versus greedy power management," in *Proc. 40th Annual International Symposium on Computer Architecture*, 2013, pp. 285–296.

[23] S. Chatterjee, M. Cho, R. Rao, and S. Mukhopadhyay, "Impact of die-to-die thermal coupling on the electrical characteristics of 3D stacked SRAM cache," in *Proc. IEEE Semiconductor Thermal Measurement and Management Symposium (SEMI-THERM)*, 2012, pp. 14–19.

[24] H. Oprins, V. O. Cherman, B. Vandevelde, G. Van der Plas, P. Marchal, and E. Beyne, "Numerical and experimental characterization of the thermal behavior of a packaged dram-on-logic stack," in *Proc. IEEE Electronic Components and Technology Conference (ECTC)*, 2012, pp. 1081–1088.

[25] D. Brunina, D. Liu, and K. Bergman, "An energy-efficient optically connected memory module for hybrid packet- and circuit-switched optical networks," *IEEE J. of Selected Topics in Quantum Electronics*, vol. 19, no. 2, pp. 3 700 407–3 700 407, 2013.

[26] Z. Li, M. Mohamed, X. Chen, E. Dudley, K. Meng, L. Shang, A. Mickelson, R. Joseph, M. Vachharajani, B. Schwartz, and Y. Sun, "Reliability modeling and management of nanophotonic on-chip networks," *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, vol. 20, no. 1, pp. 98–111, 2012.

[27] S. Manipatruni, R. K. Dokania, B. Schmidt, N. Sherwood-Droz, C. B. Poitras, A. B. Apsel, and M. Lipson, "Wide temperature range operation of micrometer-scale silicon electro-optic modulators," *Optics letters*, vol. 33, no. 19, pp. 2185–2187, 2008.

[28] S. Hu, Y. Hoe, H. Li, D. Zhao, J. Shi, Y. Han, K. H. Teo, Y. Z. Xiong, J. He, X. Zhang, M. Je, and M. Madihian, "A thermal isolation technique using through-silicon vias for three-dimensional ics," *IEEE Trans. Electron Devices*, vol. 60, no. 3, pp. 1282–1287, 2013.

[29] P. Dong, W. Qian, H. Liang, R. Shafiiha, N.-N. Feng, D. Feng, X. Zheng, A. V. Krishnamoorthy, and M. Asghari, "Low power and compact reconfigurable multiplexing devices based on silicon microring resonators," *Opt. Express*, vol. 18, no. 10, pp. 9852–9858, May 2010.

[30] R. Havemann and J. Hutchby, "High-performance interconnects: an integration overview," *Proc. IEEE*, vol. 89, no. 5, pp. 586–601, May 2001.

[31] M. Bohr, "Interconnect scaling-the real limiter to high performance ULSI," in *Proc. International Electron Devices Meeting*, 1995, pp. 241–244.

[32] J. Meindl, J. Davis, P. Zarkesh-Ha, C. Patel, K. Martin, and P. Kohl, "Interconnect opportunities for gigascale integration," *IBM Journal of Research and Development*, vol. 46, no. 2.3, pp. 245–263, Mar. 2002.

[33] S. Borkar, N. Jouppi, and P. Stenstrom, "Microprocessors in the era of terascale integration," in *Proc. Design, Automation Test in Europe Conference Exhibition*, 2007, pp. 1–6.

[34] S. Souri, K. Banerjee, A. Mehrotra, and K. Saraswat, "Multiple Si layer ICs: motivation, performance analysis, and design implications," in *Proc. Design Automation Conference*, 2000, pp. 213–220.

[35] K. Banerjee, S. Souri, P. Kapur, and K. Saraswat, "3-D ICs: a novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration," *Proc. IEEE*, vol. 89, no. 5, pp. 602–633, May 2001.

[36] E. Colgan, B. Furman, M. Gaynes, W. Graham, N. LaBianca, J. Magerlein, R. Polastre, M.-B. Rothwell, R. Bezama, R. Choudhary, K. Marston, H. Toy, J. Wakil, J. Zitz, and R. Schmidt, "A practical implementation of silicon microchannel coolers for high power chips," *IEEE Trans. on Components and Packaging Technologies*, vol. 30, no. 2, pp. 218–225, June 2007.

[37] S. G. Kandlikar, "High flux heat removal with microchannelsa roadmap of challenges and opportunities," *Heat Transfer Engineering*, vol. 26, no. 8, pp. 5–14, 2005.

[38] T. Lin and S. G. Kandlikar, "An experimental investigation of structured roughness effect on heat transfer during single-phase liquid flow at microscale," *J. Heat Transfer*, vol. 34, Oct 2012.

[39] S. Isaacs, Y. J. Kim, A. McNamara, Y. Joshi, Y. Zhang, and M. Bakir, "Two-phase flow and heat transfer in pin-fin enhanced micro-gaps," in *Proc. IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*, 2012, pp. 1084–1089.

[40] S. Isaacs, Y. Joshi, Y. J. Kim, Y. Zhang, and M. Bakir, "Two-phase flow and heat transfer in pin-fin enhanced micro-gaps with non-uniform heating," in *Proc. International Conference on Micro and Nanoscale Heat and Mass Transfer*, 2013.

[41] B. Agostini, J. R. Thome, M. Fabbri, B. Michel, D. Calmi, and U. Kloter, "High heat flux flow boiling in silicon multi-microchannels ?part i: Heat transfer characteristics of refrigerant {R236fa}," *International Journal of Heat and Mass Transfer*, vol. 51, no. 21-22, pp. 5400 – 5414, 2008.

[42] V. Sahu, Y. Joshi, and A. Fedorov, "Hybrid solid state/fluidic cooling for hotspot removal," in *Proc. IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, May 2008, pp. 626–631.

[43] A. A. Zhukauskas, *Heat Transfer from Tubes in Cross Flow, Advances in Heat Transfer*. New York: Academic Press, 1972.

[44] B. E. Short, P. E. Raad, and D. C. Price, "Performance of pin fin coldwalls constructed of cast aluminum: Part I: Friction factor correlations," *J. Thermophys. Heat Transfer*, vol. 16, pp. 389–396, Aug. 2002.

[45] Y. Civale, M. Gonzalez, D. Tezcan, Y. Travaly, P. Soussan, and E. Beyne, "A novel concept for ultra-low capacitance via-last TSV," in *Proc. IEEE International 3D Systems Integration Conference (3DIC)*, Nov. 2010, pp. 1–4.

[46] S. Ndao, Y. Peles, and M. K. Jensen, "Multi-objective thermal design optimization and comparative analysis of electronics cooling technologies," *International J. of Heat and Mass Transfer*, vol. 52, no. 19, pp. 4317 – 4326, Sept. 2009.

[47] A. Kosar, C. Mishra, and Y. Peles, "Laminar flow across a bank of low aspect ratio micro pin fins," *Journal of Fluids Engineering*, vol. 127, no. 3, pp. 419–430, Jul. 2005.

[48] P. Thadesar and M. Bakir, "Novel photo-defined polymer-enhanced through-silicon vias for silicon interposers," *IEEE Trans. Compon., Packag., and Manuf. Technol.*, vol. 3, no. 7, pp. 1130–1137, July 2013.

[49] C. Karnfelt, C. Tegnander, J. Rudnicki, J. Starski, and A. Emrich, "Investigation of parylene-C on the performance of millimeter-wave circuits," *IEEE Trans. Microw. Theory Tech.*, vol. 54, no. 8, pp. 3417–3425, Aug. 2006.

[50] Y. Zhang, C. King, J. Zaveri, Y. J. Kim, V. Sahu, Y. Joshi, and M. Bakir, "Coupled electrical and thermal 3D IC centric microfluidic heat sink design and technology," in *Proc. IEEE Electronic Components and Technology Conference (ECTC)*, 2011, pp. 2037–2044.

[51] A. Jourdain, S. Stoukatch, P. De Moor, and W. Ruythooren, "Simultaneous Cu-Cu and compliant dielectric bonding for 3D stacking of ICs," in *Proc. IEEE International Interconnect Technology Conference*, Jun. 2007, pp. 207–209.

[52] C. King, D. Sekar, M. Bakir, B. Dang, J. Pikarsky, and J. Meindl, "3D stacking of chips with electrical and microfluidic I/O interconnects," in *Proc. IEEE Electronic Components and Technology Conference*, May 2008, pp. 1–7.

[53] L. Yu, F. E. Tay, G. Xu, B. Chen, M. Avram, and C. Iliescu, "Adhesive bonding with SU-8 at wafer level for microfluidic devices," in *J. of Physics: Conference Series*, vol. 34, no. 1, 2006, p. 776.

[54] M. Shimbo, K. Furukawa, K. Fukuda, and K. Tanzawa, "Silicon-to-silicon direct bonding method," *J. of Applied Physics*, vol. 60, no. 8, pp. 2987–2989, 1986.

[55] A. Del Campo and C. Greiner, "SU-8: a photoresist for high-aspect-ratio and 3D submicron lithography," *J. of Micromechanics and Microengineering*, vol. 17, no. 6, p. R81, 2007.

[56] Z. Xue and H. Qiu, "Integrating micromachined fast response temperature sensor array in a glass microchannel," *Sensors and Actuators A: Physical*, vol. 122, no. 2, pp. 189 – 195, 2005.

[57] A. Dembla, Y. Zhang, and M. S. Bakir, "Fine pitch TSV integration in silicon micropin-fin heat sinks for 3D ICs," in *Proc. IEEE International Interconnect Technology Conference (IITC)*, 2012, pp. 1–3.

[58] D. Sekar, C. King, B. Dang, T. Spencer, H. Thacker, P. Joseph, M. Bakir, and J. Meindl, "A 3D-IC technology with integrated microchannel cooling," in *Proc. IEEE International Interconnect Technology Conference*, Jun. 2008, pp. 13–15.

[59] 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. Compon., Packag., Manuf. Technol.*, vol. PP, no. 99, pp. 1–1, 2013.

[60] Y. Zhang and M. Bakir, "Independent interlayer microfluidic cooling for heterogeneous 3D IC applications," *Electronics Letters*, vol. 49, no. 6, pp. 404–406, 2013.

[61] D. Geer, "Chip makers turn to multicore processors," *Computer*, vol. 38, no. 5, pp. 11–13, May 2005.

[62] Y. Zhang, L. Zheng, and M. S. Bakir, "Tier-independent microfluidic cooling for heterogeneous 3D ICs with nonuniform power dissipation," in *Proc. IEEE International Interconnect Technology Conference (IITC)*, 2013, pp. 1–3.

[63] K. Shakeri and J. Meindl, "Temperature variable supply voltage for power reduction," in *Proc. IEEE Computer Society Annual Symposium*, 2002, pp. 64–67.

[64] M. Pedram and S. Nazarian, "Thermal modeling, analysis, and management in vlsi circuits: Principles and methods," *Proceedings of the IEEE*, vol. 94, no. 8, pp. 1487–1501, Aug 2006.

[65] Y. Zhang, L. Zheng, and M. Bakir, "3-D stacked tier-specific microfluidic cooling for heterogeneous 3-D ICs," *IEEE Trans. Compon., Packag., Manuf. Technol.*, vol. PP, no. 99, pp. 1–1, 2013.

[66] R. Ho, P. Amberg, E. Chang, P. Koka, J. Lexau, G. Li, F. Liu, H. Schwetman, I. Shubin, H. Thacker, X. Zheng, J. Cunningham, and A. Krishnamoorthy, "Silicon photonic interconnects for large-scale computer systems," *IEEE Micro*, vol. 33, no. 1, pp. 68–78, Jan 2013.

[67] C. Holzwarth, J. Orcutt, H. Li, M. Popovic, V. Stojanovic, J. Hoyt, R. Ram, and H. I. Smith, "Localized substrate removal technique enabling strong-confinement microphotonics in bulk si cmos processes," in *Proc. Conference on Laser and Electro-Optics*, May 2008, pp. 1–2.

[68] P. Dong, W. Qian, H. Liang, R. Shafiiha, D. Feng, G. Li, J. E. Cunningham, A. V. Krishnamoorthy, and M. Asghari, "Thermally tunable silicon racetrack resonators with ultralow tuning power," *Opt. Express*, vol. 18, no. 19, pp. 20 298–20 304, Sep 2010.

[69] Y. Zhang, H. Oh, and M. Bakir, "Within-tier cooling and thermal isolation technologies for heterogeneous 3D ICs," in *Proc. IEEE International Conference on 3D System Integration (3DIC)*, 2013, pp. 1–6.

[70] J. Xie and M. Swaminathan, "Electrical-thermal co-simulation of 3d integrated systems with micro-fluidic cooling and joule heating effects," *IEEE Trans. Compon., Packag., and Manuf. Technol.*, vol. 1, no. 2, pp. 234–246, Feb 2011.

[71] JEDEC. Jedec standard for wide i/o single data rate. JEDEC. [Online]. Available: http://www.jedec.org/standards-documents/docs/jesd229

[72] C. Zhang, H. S. Yang, and M. Bakir, "Highly elastic gold passivated mechanically flexible interconnects," *IEEE Trans. Compon., Packag., and Manuf. Technol.*, vol. 3, no. 10, pp. 1632–1639, Oct 2013.