Thermal management of 3D integrated circuits with microchannel-based liquid cooling

Qian, Hanhua

2014

Qian, H. (2014). Thermal management of 3D integrated circuits with microchannel-based liquid cooling. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/61034>

<https://doi.org/10.32657/10356/61034>

Thermal Management of 3D Integrated Circuits with Microchannel-Based Liquid Cooling

QIAN HANHUA

School of Electrical and Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of
Doctor of Philosophy

2014

Acknowledgement

This thesis would not be achievable without the help from the following people.

First and foremost, I would like to thank my supervisor, Associate Professor Chang Chip Hong, for his continual support and patient guidance throughout my Ph.D. study in Nanyang Technological University. I appreciate the many discussions and valuable advices from him through which I was enlightened and motivated by his academic expertise and passion. My ideas were always more complete and well formulated after they were polished by him. I want to thank him for the effort on correcting all my manuscripts for journal and conference publications including this thesis. His high standard on professional writing has improved my writing skills a lot. Prof. Chang has a kind personality. The “Chip Family” that was formed by him, Mrs Chang, and all his students has provided a lot of fun and sentimental support especially for foreign students. Over the last four years, I have been deeply influenced by his positive attitude and wisdom of solving problems in both research and life, which will surely benefit me for the rest of my life.

I am also grateful to my co-supervisor, Associate Professor Vinod Achutavarrier Prasad. My first three year of PhD scholarship was funded by the three-year MOE AcRF Tier II project of which he and my supervisor are the principle investigator and co-investigator, respectively. Thanks to his special effort to convince the management of his school to continue my scholarship support in the fourth year after the MOE funded research project had concluded successfully.

I would like to express my special thanks to Assistant Professor Yu Hao who has led me to the research topic of three-dimensional integrated circuit, through a one-year collaboration. Without the help of his expertise in this area, I would not have accelerated the acquisition of the essential knowledge and gone so deep in this

topic. I also want to thank Assistant Professor Zhang Wei for providing me the Research Associate job position so that I remain in good financial state to finish the writing of this thesis.

I want to convey my gratitude to all the friends in the “Chip Family” for their cheerful support, especially Dr. Chen Jiajia, and Mathias Faust. Dr. Chen supervised my undergraduate final year project, and eventually recommended me to Professor Chang to continue with my Ph.D. study. Mathias has provided me a lot of help on technical problems that I encountered during my research.

I would like to thank my friends from the “Accelerated Program” class of the SM2 scholarship for their valuable friendship throughout the years in Singapore. They are Pan Hao, Lin Mengda, Li Henan, Lin Bo, Deng Guanglei, Chen Xi, Wang Hongru, Zhu Wenting, Chang Wanli and Zhao Rongrong.

I also wish to thank my friends, Chen Yupeng, Cui Jin, Liang Hao, Cui Yingnan, Jiang Lianlian and Wu Meiqing, from the same laboratory for their company and the fun activities they organized. I would like to specially thank the staff of the Centre for High Performance Embedded Systems, Mr. Jeremiah and Ms. Marilyn Yap, for the generous technical support whenever needed.

I would like to acknowledge the generosity of the Ministry of Education of Singapore and Nanyang Technological University for providing me the opportunity to study in Singapore with scholarships.

I would also like to express my gratitude to all the teachers back in my primary, secondary and high schools in China, who have led me to the hall of knowledge, inspired my interest on natural science, and helped me financially for my study so that I could fulfill my aspiration and the pursuit of a Ph.D. degree.

Finally, I would like to present this thesis as a gift to my beloved parents, who have raised me and sacrificed a lot to get me well educated.

Contents

List of Tables	v
List of Figures	vii
Abstract	xi
1 Introduction	1
1.1 Motivation	1
1.2 Research Objectives	7
1.3 Major Contributions	8
1.4 Organization of the thesis	9
2 Preliminaries and Literature Review	11
2.1 Overview of 3DIC Technology and Its Fabrication Processes	12
2.1.1 3DIC Fabrication Through Monolithic Approaches	12
2.1.2 3D Integration Through Wafer Bonding	16
2.2 The 3DIC Architecture Under Consideration and Its Associated Thermal Challenges	27
2.3 Thermal Management of 3DIC	30
2.3.1 Thermal Management from Power Perspective	30
2.3.2 Thermal Management by Advanced Cooling Techniques	49
2.4 Thermal Modeling of 3DIC	56
2.4.1 Thermal Model of Heatsink-Based Air-Cooled 3DICs	57
2.4.2 Thermal Model of Microfluidic Cooling in 3DICs	62
2.5 Summary	66

3	Compact Thermal Modeling of 3D Integrated Circuits	69
3.1	Introduction	69
3.2	Preliminaries on Thermodynamic Theories	71
3.2.1	Thermal Conduction in Solids	71
3.2.2	Thermal Convection through Heatsink	72
3.2.3	Convective Heat Transfer of Microchannels	72
3.3	Target Architectures of 3DIC	74
3.4	Compact Thermal Model of 3DIC	75
3.4.1	Grid Division of 3DIC	75
3.4.2	Thermal Conductance between Solid Cells considering Anisotropic TSV Effects	77
3.4.3	Convective Heat Transfer and Entrance Effect of Microchannels	80
3.4.4	Steady State Thermal Analysis	81
3.4.5	Software Implementation	82
3.5	Experiment Results	83
3.5.1	Heatsink-based Air-Cooled 3DIC	84
3.5.2	3DICs with Microfluidic Cooling by Microchannels	89
3.5.3	Parametric Study of Microfluidic Cooling for 3DIC	92
3.5.4	Influence of Granularity on the Modeling Accuracy	94
3.6	Summary	95
4	Power Aware Thermal Management for Microfluidic Cooled 3D Multi-Core Systems	97
4.1	Introduction	97
4.2	Problem Formulation of Runtime Thermal Management of 3D Multi- Core Systems	101
4.2.1	Typical 3D Multi-Core Cache-Processor System	101
4.2.2	Runtime Thermal Management Problem of 3D Multi-Core Systems	102
4.3	Power Aware Runtime Thermal Management Scheme of 3D Multi- Core Systems	104

4.3.1	Power Consumption Tracking for Cache and Microprocessors .	104
4.3.2	Real-time Cooling Demand Estimation	108
4.3.3	Runtime Thermal Management Scheme with Fine-Grained Flow Rate Control	111
4.4	Experimental Results and Discussion	113
4.4.1	Effectiveness of microfluidic cooling	114
4.4.2	Runtime thermal management with fine-grained flow rate con- trol	115
4.5	Summary	121
5	Microchannel Clustering, Splitting and Scaling for Thermal Bal- ancing of Liquid-Cooled 3DIC	123
5.1	Introduction	123
5.2	Microchannel Clustering and Flow Rate Allocation for Non-uniform Microfluidic Cooling	125
5.2.1	Clustering of Microchannels	126
5.2.2	Allocation of Flow Rate	130
5.2.3	Minimization of Cooling Effort under Thermal Constraints . .	130
5.2.4	Issues on Practical Implementation	132
5.3	Microchannel Splitting and Scaling Techniques for Low-overhead Ther- mal Balancing	133
5.3.1	Microchannel Splitting by Splitter Placement	134
5.3.2	Microchannel Scaling for Redistribution of Coolant	135
5.4	Validation of Proposed Thermal Balancing Techniques	137
5.4.1	Thermal Balancing by Channel Clustering	138
5.4.2	Thermal Balancing by Channel Splitting and Scaling	144
5.5	Summary	146
6	Conclusion and Future Work	149
6.1	Conclusion	149
6.2	Recommended Future Work	152

6.2.1	Investigation on Channel Splitter Placement for Simultaneous Thermal and Routing Optimization	152
6.2.2	3D Clock Tree Synthesis Considering Microchannel-Induced Thermal Gradient and Congestions	152

List of Tables

2.1	Precision Levels of Different Alignment Methods	25
2.2	Thermal Conductivity Values of Commonly Used Materials in 3DIC Fabrication	29
3.1	Experiment settings	84
3.2	Steady State Temperature (K) Comparison of COMSOL, Hotspot and Proposed Simulator	85
3.3	Runtime Comparison of COMSOL, Hotspot and Proposed Simulator for Air Cooling Simulation	86
3.4	Reduction of maximum temperature (K) due to TSV insertion	88
3.5	Comparison of maximum and minimum temperatures among the pro- posed thermal model, COMSOL and 3D-ICE, with the corresponding deviation (in %) from COMSOL shown in bracket	90
3.6	Comparison of average temperature of each row in Fig. 3.12	91
3.7	Average Runtime Comparison of COMSOL, 3D-ICE and Proposed Thermal Simulator for Microfluid Cooling Simulation	91
4.1	Parameters of 3D multi-core system with microfluidic channels	114
4.2	Cooling effect comparison between heatsink-based and microfluidic cooling by comparing the maximum and minimum temperatures and the number of nodes in each temperature range	115
5.1	Clustering results with different η values	140
5.2	Reallocation of cooling resources	140

5.3	Comparison of the maximum temperatures under uniform and clustered coolings	142
5.4	Maximum temperature after minimization of cooling effort	143
5.5	Comparison of cooling effort under thermal constraint	143
5.6	Comparison of maximum temperatures (T_{max}) and thermal gradient (ΔT), measured in <i>Kelvin</i> under different microchannel design techniques	146

List of Figures

1.1	Transistor density of main stream microprocessors with their corresponding manufacturing technology [5]	2
1.2	Throughput limitation due to interconnect scaling, adapted from [7] .	3
1.3	The increasing power density of Intel microprocessors, plotted based on [4, 18, 19]	6
2.1	Schematic of the STC RAM cell [24]	13
2.2	Schematic of the stacked inverter structure [25]	13
2.3	Fabrication process of a double-layer 3D circuit by laser recrystallization and SOI technology: (a) Deposition of oxide layer and planarization; (b) Deposition of amorphous silicon on the oxide layer; (c) Recrystallization by laser beams; (d) Planarization of recrystallized silicon and device fabrication; (e) Deposition of the second oxide layer; (f) Repetitive addition of recrystallized silicon layer and device fabrication.	14
2.4	The CUBIC process flow: (a) Formation of the W bump on MoSi_2/Al wires and attachment of backing substrate; (b) Wafer thinning from the backside; (c) Formation of through hole and the Au/In pool at the backside; (d) Bonding of the thin film and the bottom device layer; (e). Removal of the backing substrate and adhesives [33]. . . .	17
2.5	Typical fabrication process of TSV	19
2.6	Schematic of optical interconnection [49]	20
2.7	Schematic of capacitive coupled interconnection [52]	20
2.8	Schematic of inductive coupled interconnection	21

2.9	Schematic of different alignment methods	24
2.10	Schematic of the typical 3DIC structure with heatsink-based air cooling	28
2.11	Typical implementation of low-to-high level shifter [83]	32
2.12	The (a) balanced and (b) unbalanced logic implementations of $f = abcd$	38
2.13	3D Placement Flow	44
2.14	Schematic of microfluidic cooling structure for 3DICs	53
2.15	Microchannel fabrication process flow	54
2.16	One-dimensional resistive modeling of 3DIC	58
2.17	Modeling of TSV's thermal effect with and without considering the lateral thermal resistance	61
2.18	The six-resistor representation of a grid cell	63
3.1	Thermal Conduction between Solid Cells	71
3.2	Schematic of 3D integrated circuit with Heatsink and microfluidic coolings	75
3.3	Grid division of heat spreader showing thermal connections between the trapezoidal extra nodes and the inner cells.	76
3.4	Grid division of heatsink showing thermal connections among the trapezoidal extra nodes and the inner cells. The grid dimension of circuit layer is assumed to be 3×3 for this illustration.	76
3.5	Grid division of 3DIC with microchannel-based liquid cooling. A 2- layered structure is assumed for this illustration.	77
3.6	Thermal connections within a grid pile.	77
3.7	Geometry of grid cell with TSV	79
3.8	Nusselt number at different distances along the flow direction	81
3.9	Program flow chart	83
3.10	Runtime Scaling of Our Thermal Simulator and Hotspot	86
3.11	The three TSV test patterns	87
3.12	Comparison of top layer temperature profile against COMSOL at a heat flux of $150W/cm^2$ and $1ml/min$ flow rate per channel, coolant flows from north (upper) to south (lower) direction.	91

3.13	Top layer temperature map from 3D-ICE at a heat flux of $150W/cm^2$ and $1ml/min$ flow rate per channel, coolant flows from north (upper) to south (lower) direction.	92
3.14	Comparison of circuit temperature against COMSOL with varying microchannel width at a fixed microchannel pitch of $200\mu m$	93
3.15	Comparison of circuit temperature against COMSOL with varying microchannel pitch at a fixed microchannel width of $60\mu m$	94
3.16	More accurate minimum temperature prediction with fine-grained grid division along the flow direction.	95
4.1	(a) 3D multi-core and cache system <i>A</i> ; (b) 3D multi-core and cache system <i>B</i> . Both are with microfluidic cooling.	101
4.2	Heatsink-cooled 3D multi-core and cache system	102
4.3	Channel clustering with 3 and 6 clusters	112
4.4	Procedure of the proposed thermal management scheme	113
4.5	Effective heat removal of microfluidic cooling by increasing flow rate .	116
4.6	Total power consumption of Alpha-2 microprocessor running different benchmark applications	117
4.7	AR prediction of future power consumption with Kalman filter correction	117
4.8	Maximum temperature of the 3D testbench <i>A</i> with no runtime thermal management	118
4.9	Maximum circuit temperature of (a) - (c): system <i>A</i> and (d) - (f): system <i>B</i> under thermal management of different clustering granularity and triggered by real-time cooling demand.	120
4.10	Dynamic flow rate settings under different granularity of flow rate control	121
5.1	An illustration of non-uniform power distribution (W/cm^2) of 3DIC .	125
5.2	Grid division of 3D integrated circuit	127
5.3	Thermal resistances between the circuit and channel cells	133
5.4	Illustration on splitter insertion for channel splitting	134

5.5	Cooling demand of microchannels with different power distributions .	139
5.6	Temperature distribution comparison: thermal maps of the top layer for Case 1 under uniform flow rate cooling and reallocated clustered flow rate cooling with the same total flow rate, coolant flows from north (upper) to south (lower) direction.	141
5.7	Temperature distribution comparison: top layer thermal maps of Case 1 using the proposed minimization method and uniform flow rate minimization with the same thermal limit, coolant flows from north (upper) to south (lower) direction.	144
5.8	Comparison of the convective resistances of a channel and its sur- rounding circuit temperatures before and after channel splitting. . . .	145
5.9	Temperature distribution comparison: top layer thermal maps of Set A (a) under default microchannel setting and (b) with channel split- ting and scaling. Coolant flows from north (upper) to south (lower) direction.	147

Abstract

Three dimensional integrated circuits (3DICs) have attracted tremendous interest in recent years. The topological advantages of this emerging technology offers on one hand a gamut of new promising solutions to problems faced by today's 2D large-scale-integration (LSI) such as long propagation delay, high power consumption and limited I/O bandwidth. On the other hand, the stacking nature of 3DICs exacerbates the existing thermal problems due to the simultaneous increase of power density and thermal resistance. Effective heat removal has thus become a major concern for successful implementation of 3DIC. Conventional air cooling has been shown to be inadequate to maintain the low temperature of 3DIC chips. More advanced cooling techniques such as microchannel-based microfluidic cooling are hence proposed. In this thesis, thermal management of 3DICs with microchannel-based liquid cooling is exploited. A runtime thermal management scheme and microchannel design algorithms for efficient liquid cooling are proposed.

A finite-difference-method-based thermal simulator is first developed for 3DICs with both conventional air cooling and microchannel-based liquid cooling. This simulator considers the thermal effect of through-silicon-vias (TSVs) at fine-granularity by calculating the anisotropic equivalent thermal conductances of a solid grid cell if TSVs are inserted. Entrance effect of microchannels is also investigated for accurate modeling of microfluidic cooling. The proposed thermal simulator is verified against commercial multiphysics solver COMSOL and compared with Hotspot and 3D-ICE. Simulation results shows that for heatsink-fan-based air cooling, the proposed simulator is as accurate as Hotspot but runs much faster at moderate granularity. For microfluidic cooling, our proposed simulator is more accurate than 3D-ICE in its estimation of steady state temperature and thermal distribution.

3D Multi-Processor System-on-Chip (3D MPSoC) is an important application of 3DIC technology. Due to the high power density of the microprocessors, 3D multi-core systems are especially vulnerable to the die thermal run-way hazard. A real-time demand-based thermal management scheme is thus proposed to address this teething problem. A cooling demand predictor is used to foresee the needed cooling effort. The predictor adopts an autoregressive model to estimate the 3DIC's future power consumption based on historical power values, which is then fed into the thermal simulator to calculate the expected circuit temperature and hence the cooling demand. With that, the flow rate controller adjusts the flow rate accordingly to prevent potential thermal violations. A fine-grained flow rate control is deployed to better match the cooling effort with regional cooling demands. Simulation results show that the proposed thermal management scheme is capable of maintaining low temperature of the 3DIC with significant flow rate savings.

Without regarding to the fact of non-uniform power distribution of integrated circuits, existing microfluidic cooling with uniform cooling effort incurs large thermal gradient and wastes pump power. A microchannel clustering and flow rate allocation algorithm is hence proposed to implement a customized non-uniform cooling scheme. In this scheme, the microchannels are first divided into clusters where members of a channel cluster are characterized by their homogeneous cooling demand. An appropriate flow rate setting is then applied to each cluster based on the total flow rate and the maximum allowable temperature of the 3DIC. Such clustering and flow rate allocation process is guided by our proposed algorithm for an effective and efficient microfluidic cooling. Compared to the uniform flow rate cooling, the maximum temperature and thermal gradient have been reduced under the same total flow rate settings. On the other hand, for a specific peak temperature constraint, the required total flow rate has been largely reduced with the proposed non-uniform microfluidic cooling.

A novel splitting and scaling microchannel design technique is proposed for the thermal balancing of microfluidic-cooled 3DIC. By inserting appropriate number of channel splitters along various sections of microchannels, the convective thermal conductances at potential hot spots are effectively reduced. The disrupted coolant

allocation due to channel splitting are then redistributed by scaling the channel widths to match the coolant flow rates with the power distribution, which is usually non-uniform in practice. The proposed technique effectively reduces the thermal gradient and required total flow rate. Unlike existing thermal balancing method, the proposed technique only requires customized etching masks to allow different channel widths and planting of channel splitters, thereby incurring no noticeable overhead on the packaging and fluid networks.

Chapter 1

Introduction

1.1 Motivation

Since the inventions of the first transistor in 1947 [1] and the first integrated circuit in 1958 [2], the semiconductor industry has prospered for over half a century with continuous growth and performance improvements. From discrete transistor chips to today's high performance microprocessors and system-on-a-chip's (SoC's), electronic devices have evolved to be hundred thousand times faster and incorporated abundant functionalities. The main driving force behind this has been the technology scaling where transistors are continuously miniaturized to achieve faster switching speed and lower power consumption. Such transistor scaling has enabled increasingly larger scale of integration over the past decades. Fig. 1.1 depicts the transistor density of industry microprocessor benchmarks since 1971 with their corresponding manufacturing technology [3–5]. As projected by Moore's law [6], the transistor count doubles almost every two years. With more transistors integrated in a single chip, the slow off-chip wirings are replaced by fast and high-bandwidth on-chip interconnections which assures significant performance improvement and delivers additional functionalities.

Unfortunately, such growth is manifested to be unsustainable if it is solely dependent on technology scaling, due to the following problems. Firstly, the shrinking of the technology node effectively reduces the transistor area and latency but increases the delay of interconnections due to their smaller cross-sectional area. As

Figure 1.1: Transistor density of main stream microprocessors with their corresponding manufacturing technology [5]

shown in Fig. 1.2, even with the enhancement of better conductor and low- k materials, the wire delay predominates and becomes the bottleneck of system throughput when the feature size shrinks beyond $130nm$ [7]. In addition, the number of metal layers required for routing increases rapidly with larger size integration and is expected to reach the practical limit within a few generations [8]. New interconnection methodology is therefore in urgent need for the continual growth of integration size and circuit performance. Secondly, further miniaturization imposes many challenges on the yield of manufacturing. Severe transistor performance variations have been observed for today's nanometer-scale technology and expected to be worse for the future generation of technology nodes, due to both intrinsic and extrinsic reasons such as non-uniformity of dopant, line-edge roughness and lithography enhancement techniques [9]. The resulting functionality of the fabricated device may thus deviate much from designer's expectations. Such variation, together with the increased defect density [10], directly threatens the production yield. Stringent purity and precision controls are thus required for yield enhancement purpose, which will all add to the fabrication cost finally. Lastly, as the transistor size shrinks down to nanometer-scale, subthreshold leakage and gate oxide tunnelling current become much more pronounced than the older generation of electronic devices [11]. Both

types of leakage contribute a big portion to the circuit's total power consumption for today's nano-scale devices, which increases packaging cost, reduces reliability and creates thermal problems. The situation will only be worse with further device miniaturization.

Figure 1.2: Throughput limitation due to interconnect scaling, adapted from [7]

Mass production of microprocessors at $22nm$ technology has already been achieved in 2011 [12]. According to the ITRS's projection, the feature size will finally reach $10nm$ in the year of 2020 [13]. Shrinking beyond $10nm$ would then be extremely difficult, if not impossible. However, even if such target can be achieved through probably high implementation cost, the resulting performance gain may be marginal due to other constraints such as interconnections. It is clear that to sustain the chronicle performance improvement of integrated circuits, new technology and innovations are essential to solve or at least mitigate the aforementioned problems.

A lot of research effort has been put to emerging technologies such as carbon nanotubes [14], graphene [15] and quantum computing [16]. The progresses are encouraging but wide application of these advanced technologies is unlikely to happen in the near future. On the other hand, for economic reasons, it would be more practical and beneficial to rely on innovations to prolong the use of CMOS technology before other advanced technologies are ready for mass production. Under

such circumstances, three-dimensional integrated circuit (3DIC) has attracted enormous interest in both industry and academia in recent years. Different from the conventional single-layer approach, 3DIC expands the design space into the third dimension by stacking multiple layers of integrated circuits. Different layers of active devices communicate via fast vertical interconnections such as Through-Silicon-Vias (TSVs). Compared to conventional 2D circuits, electronic devices in 3DIC can be more densely packed within smaller footprint. Distances between devices are shortened and the wire delays are therefore reduced. Shorter wire length also implies lower power consumption due to less parasitics. In addition, the capability of heterogeneous integration in 3DIC further extends the yield of integrating multiple technologies within a chip.

The published results [22, 23, 177, 213–215, 255–257] have repeatedly shown the advantages of 3D integration over the conventional 2D technology. At system level, intensive research has been carried out on Multi-Processor System-on-Chip (MPSoC) and Network-on-Chip (NoC) [22, 23, 213, 255–257]. With 3D integration, MPSoCs and NoCs can be expanded over the third dimension, breaking the scalability limitations on 2D platform. The previously too-large-to-fabricate system can now be divided and fabricated on separate dies. Such divide-and-conquer approach enables the screening and matching process. Through process variation analysis, the die stacking process can be optimized to improve the final production yield [22, 23]. The vertical interconnects such as TSVs are able to provide large bandwidth for communication, thus can be a promising solution to mitigate the memory wall effect commonly seen on 2D core-cache systems [213]. With the advent of 3D electrical and optical routers, the development of NoCs in 3D environment is envisioned [255, 256]. Delay, power and throughput of 3D NoCs are much improved comparing with 2D designs [257]. At the circuit and device level, the results obtained are also encouraging. With the aid of 3DCacti [215], it is observed that 3D caches with two layers have already outperformed 2D caches by 20% in delay and 14% in energy. With more layers, more reduction in delay will be achieved with marginal gain in energy. Besides cache, processor design in 3D is also explored by many researchers. In [214], the authors summarized the state-of-the-art 3D proces-

sor designs and proposed several different approaches. Existing design issues such as power delivery, thermal, granularities of design approaches and cache stacking techniques were discussed. It was shown that by splitting functional unit blocks of an alpha processor in 3D, a 10.3% improvement on clock frequency and 8.7% overall performance improvement were achieved. In [177], Hao Hua et al. explored the design space of tier count and thermal via density for Fast Fourier Transform (FFT) and OpenRISC Platform-System-on-Chip (ORPSOC) circuits. 27% energy reduction and 20% delay reduction were demonstrated compared to the 2D alternatives.

The benefits of going 3D for circuit and system designs are obvious. However, while offering exciting opportunities, this new technology also brings about many new design challenges, among which the thermal problem is one of the major limiting factors. As pointed out by James Early, the total number of electronic elements that can be integrated within a system is constrained by the average power density [17] due to the thermal requirements. As shown in Fig. 1.3, the power density of electronic devices has been readily climbing up over the decades due to the shrinking of feature size and the increasing clock frequency [4, 18, 19]. Heat removal has been difficult even for 2D high performance large scale integration [239–243]. The stacking of circuits in 3DIC simultaneously increases power density and junction-to-ambient thermal resistances, thus significantly accelerates the heat accumulation. The induced thermal stress was notorious for being the main culprit of the aging and failure of electronic devices [268]. In time scale, large magnitude thermal cycles are always responsible for the plastic deformation of interconnections in IC chips and packages [20, 21]; and from the spatial point of view, the induced thermal gradient enlarges the divergence of device and interconnection performance, deviates the runtime parameters from design expectation, and results in performance degradation and malfunction. There is thus a strong need for effective and efficient cooling techniques to maintain a low and uniform circuit temperature for an optimal and reliable 3D implementation of integrated circuits.

Due to the wide availability and low implementation cost, traditional heatsink-fan-based air cooling was first applied to 3DICs. Soon it was revealed to be unsatisfactory on temperature control due to its high thermal resistance and lack

Figure 1.3: The increasing power density of Intel microprocessors, plotted based on [4, 18, 19]

of scalability [180]. A lot of research effort has then been spent on auxiliary techniques to supplement the air cooling. Some of the representative techniques are thermal-aware floorplan and placement, task scheduling and thermal TSVs [169, 170, 175, 176, 216, 217, 262]. Unfortunately, for 3DICs with high power density or more than two circuit layers, the long heat transfer path dominates and the temperature of circuits far from the heatsink can easily rise to an unacceptable level. The big temperature difference between active layers introduces non-negligible shear stress near the bonding surface, and may cause bonding failure or cracking of the wafers. Realizing the severity of this problem, recent research focus has been shifted to novel cooling solutions such as microchannel-based liquid cooling [179]. Attributed to the higher heat capacity and thermal conductivity of coolant, liquid cooling exhibits a much stronger heat removal ability than air cooling [179, 180, 187, 204, 247, 263]. In addition, by placing the microchannels in between the layers of 3D stacks, microfluidic cooling provides local thermal grounds to every circuit layer, thus is scalable with increasing number of circuit layers.

The physical properties of microchannel-based liquid cooling make it a promising solution to the thermal problems of 3DIC. Unlike heatsink-fan-based cooling where

the fan speed is adjustable for only system-level cooling control, the coolant speed in each microchannel of the microfluidic cooling can in principle be controlled separately to realize a fine-grained cooling effort adaptation. This important merit of microchannel-based liquid cooling enables us to devise a customized cooling scheme to achieve better cooling results and save the cooling power. Regrettably this has not been well investigated in the literature. Another characteristic of microfluidic cooling that has often been neglected is the built-up of thermal gradients due to non-uniform power density and rising coolant temperature along the flow direction. The entrance effect of microchannels aggravates the problem by dragging the inlet area temperature lower. Large thermal gradient is detrimental to normal circuit operation, complicates design constraints and may offset the benefits of microfluidic cooling if not properly handled.

To investigate the above mentioned problems, an accurate thermal modeling of 3DIC with microchannel-based liquid cooling is attempted. There are several options available for such thermal model. Unfortunately they either suffer from long runtime due to the complicated calculations involved [204–206], or produce inaccurate temperature results [207, 208]. This research work is hence motivated to model the thermal behavior of liquid-cooled 3DIC, and find new solutions to manage and optimize the thermal conditions of 3DIC using microchannel-based liquid cooling.

1.2 Research Objectives

The primary objective of this research is to solve the imperative thermal challenges faced by 3DIC with microchannel-based liquid cooling. To fulfill the objective, specific goals have been set in the research plan as follows.

1. To develop an accurate compact thermal model of 3DIC with both heatsink-fan-based air cooling and microchannel-based liquid cooling.
2. To investigate how microchannel-based liquid cooling responds to runtime uncertainties and then propose runtime thermal management scheme to achieve a more effective and efficient cooling.

3. To analyze the characteristics of microfluidic cooling and propose novel methodology for the design of microchannels to overcome the impact of non-uniform power density distribution and rising coolant temperature.

1.3 Major Contributions

The following contributions have been made to achieve the goals set out in this research.

1. A compact thermal simulator was developed for 3DICs with both heatsink-fan-based air cooling and microchannel-based liquid cooling. Finite difference method was adopted and the analogy between thermal and electric circuit was utilized. Compared with existing works, the proposed thermal model considers the effect of anisotropic thermal conductivity due to TSV insertion and the entrance effect of microchannels. The accuracy of this thermal model has been proven by verification against commercial multiphysics simulator. Peer comparisons with existing analytical thermal model further highlighted the accuracy and speed advantages of the developed thermal model. This thermal modeling work has been released online [231] as an open-source project for use by other researchers in this field.
2. A proactive thermal management scheme has been proposed for the runtime thermal control of 3D multi-core cache-processor system with microfluidic cooling. An autoregressive power predictor is built to estimate future power densities of the multi-core system. Based on this prediction, the expected circuit temperature is simulated and used as the real-time cooling demand. The coolant flow rate controller then uses this cooling demand information to adjust the flow rate accordingly. The proposed thermal management scheme was proven effective by applying it to a 4-layer multi-core system. The runtime temperature was shown to be kept well within thermal constraints and a significant flow rate saving was achieved.
3. An efficient microchannel clustering and flow rate allocation algorithm was

proposed to cater for the non-uniform power distribution of 3DIC. Instead of uniform flow rate allocation, the proposed algorithm results in several different flow rate settings for the 3D circuit or system. Each flow rate setting corresponds to a channel group where the microchannels have similar cooling demands. Compared with uniform flow rate cooling, such fine-grained adaptive flow rate control effectively suppressed the peak circuit temperature with the same total flow rate supply. On the other hand, significant flow rate saving was achieved when a temperature limit is set as the control target.

4. A microchannel splitting and scaling design methodology was proposed for the thermal balancing of 3DIC. Microfluidic cooling typically introduces large thermal gradient along the flow direction due to the rising coolant temperature. The non-uniform power distribution further added to the unbalance of 3DIC's thermal condition. By inserting channel splitters into microchannels, the thermal conductances between circuit and coolant are effectively reduced at the desired places. Combining with the channel scaling technique, which matches the coolant allocation with the non-uniform power distribution, a balanced cooling for 3DIC was achieved at a low overhead cost.

1.4 Organization of the thesis

This thesis is organized into six chapters. In Chapter 1, the motivation, objective and major contributions resulted from this research work are stated.

In Chapter 2, Preliminary knowledge on the general architecture and manufacturing processes of 3DIC are introduced. With that, existing thermal challenge for the adoption of the emerging 3DIC technology is briefly discussed, and the corresponding thermal management and modeling techniques in literature are reviewed.

From Chapter 3 to Chapter 5, the key contributions of this research work are presented in detail. In Chapter 3, the developed thermal simulator for 3DICs with both heatsink-fan-based air cooling and microchannel-based liquid cooling are described in detail. The basic heat transfer and thermodynamics theory are also reviewed. Different packaging designs of air cooling and liquid cooling are presented and the

division of 3DIC chips into grid cells for finite difference modeling is shown. Verification results against the commercial multiphysics simulator are presented to confirm the accuracy of the proposed thermal model. A peer comparison with existing compact thermal models is also performed to show the advantages of the simulator.

In Chapter 4, the proposed runtime thermal management scheme for 3D multi-core system is introduced. An autoregressive power estimator that were used to predict the future power consumption of 3DIC is first presented. This future power information is translated into cooling demand, which is later used by the flow rate controller to control the coolant supply. A fine-grained flow rate control for an efficient liquid cooling is then proposed and described in details.

In Chapter 5, two microchannel design techniques are proposed for the thermal balancing of 3DIC. For the microchannel clustering and flow rate allocation algorithm, the process of defining microchannel groups according to the cooling demands is described first, followed by the flow rate allocation which matches the coolant amount with the power distribution. After that, a more advanced microchannel splitting and scaling technique is proposed to mitigate the thermal gradients across and along the flow direction. In this part, the approaches to reduce the convective thermal conductance by inserting the channel splitters and allocate coolant among the microchannels by channel scaling are explained. For both microchannel design techniques, a case study on a 3D multi-core benchmark is used to show their effectiveness.

Finally, Chapter 6 summarizes the results achieved for this research work. Possible extensions and related topics worthy of further research are also identified and briefly discussed.

Chapter 2

Preliminaries and Literature Review

The technology of three dimensional integrated circuit (3DIC) offers numerous advantages over the conventional 2D circuits. It is regarded as a key enabling technology for further growth of semiconductor industry when integration density of 2D circuits becomes unsustainable by mere transistor scaling. As mentioned in previous chapter, due to the multilayer stacking approach, 3D integration introduces severe cooling challenge, especially for the inner circuit layers, which prevents a feasible 3D implementation of high performance (and usually high power) computing systems. To solve this problem, an accurate heat transfer analysis is needed in the first place. Such analysis requires detailed knowledge about the common architecture and composition of 3DIC chips. The background knowledge will be provided by an overview of 3DIC technology and its fabrication process in this chapter. Various proposed 3DIC fabrication approaches and their resulting architectures will also be introduced. In addition, key thermal modeling and thermal management approaches proposed in the literature will be reviewed in detail. Some of these approaches may not be 3DIC specific, but they can be extended to or potentially useful for 3D cases. Being the main cooling scheme focused in this thesis, the hardware configuration of microchannel-based microfluidic cooling will be presented in detail.

2.1 Overview of 3DIC Technology and Its Fabrication Processes

The topological advantage of 3DIC is clearly illustrated by a simple example. Suppose a 2D circuit or system consisting of N_e identical circuit elements, a quick estimation of the footprint A and total wirelength L yields:

$$A \propto N_e \quad \text{and} \quad L \propto \sqrt{N_e} \quad (2.1)$$

Due to the exponential increase of circuit complexity, the wire length and hence interconnect delay and power are expected to increase quickly and become the bottleneck of circuit performance. Interconnect delay and power consumption are the two critical challenges faced by today's large scale integration [213]. If there is a way to implement the same circuit in 3D with equal design freedom in all three dimensions, (2.1) can be reduced to:

$$A \propto N_e^{\frac{2}{3}} \quad \text{and} \quad L \propto \sqrt[3]{N_e} \quad (2.2)$$

In practical cases where the 3D circuit is realized by interconnecting a number of 2D circuits through vertical wires, the footprint and wire length will depend on the number of circuit layers N_l , as shown below.

$$A \propto \frac{N_e}{N_l} \quad \text{and} \quad L \propto \sqrt{\frac{N_e}{N_l}} \quad (2.3)$$

Clearly, 3D integration can substantially reduce the circuit footprint and wire length over its 2D counterpart, provided that a feasible 3D fabrication technology exists with low cost and high yield. Huge research and development investment has been expended on the 3D fabrication technology and many prototypes have been built over years for feasibility study and accurate device characterization. The interest and effort of the related research can be dated back to the 70s and 80s.

2.1.1 3DIC Fabrication Through Monolithic Approaches

Transistor-level 3D Structures

The benefit of the compactness of 3D integration was first demonstrated by transistor-level 3D structures during the 70s and 80s as exemplified by the stacked capacitor

(STC) random access memory (RAM) [24] and compact stacked inverter structure [25]. Fig. 2.1 shows the cross-sectional view of a STC RAM cell where a layer of polysilicon was deposited on top of the MOSFET to form the storage capacitor. Using this approach, the area of a RAM cell had been successfully reduced to one

Figure 2.1: Schematic of the STC RAM cell [24]

third of the conventional structure while having enough storage capacitance. In [25], a novel compact inverter structure was proposed, where the n-MOSFET was fabricated on top of the p-MOSFET, as shown in Fig. 2.2. Both works attempted to achieve a much higher packing density than the conventional 2D circuits by changing the micro-scale cell structure to achieve better result based on the available technology of the time.

Figure 2.2: Schematic of the stacked inverter structure [25]

Laser Recrystallization and SOI based 3D Integration

Later, it was foreseen that scaling would eventually reach its limit (although at that time transistor scaling was pessimistically predicted to reach its limit in the 90s), and further growth might have to be relied on 3D integration [26]. The research focus was soon shifted towards high level larger scale 3D integration, which relies on the capability of fabricating transistors on a layer of recrystallized silicon deposited on top of silicon-on-insulator (SOI) wafers [27, 28]. Fig. 2.3 illustrates a general process flow to fabricate a double-layer 3D circuit using the laser recrystallization and SOI technology. A layer of active device is first fabricated on a SOI wafer

Figure 2.3: Fabrication process of a double-layer 3D circuit by laser recrystallization and SOI technology: (a) Deposition of oxide layer and planarization; (b) Deposition of amorphous silicon on the oxide layer; (c) Recrystallization by laser beams; (d) Planarization of recrystallized silicon and device fabrication; (e) Deposition of the second oxide layer; (f) Repetitive addition of recrystallized silicon layer and device fabrication.

following the conventional procedure with metallization. A layer of oxide is then formed on top of the wafer, and its surface is refined through the planarization

process. After that, a thin film of amorphous silicon is deposited on the oxide layer, and recrystallized by laser beams. The second active layer is then fabricated on the thin film, and interlayer connection is realized through via holes. In principle, such fabrication process can be repeated to stack as many active layers as desired to implement multi-layer 3DICs.

A few functional chips had been built to verify the concept of 3DIC using this fabrication process. One example is a 256-bit RAM chip that was mapped into two active layers, where the bottom layer was used to fabricate the memory cells and the top layer accommodated I/O peripherals such as the sense amplifier and address decoders [29]. Another good example is a 10-bit linear image sensor reported by Mitsubishi [30], where the parallelism of image processing was conveniently fulfilled by the 3DIC architecture. The 3D image sensor comprises two layers. The top layer contains the photodiodes and the second layer contains the signal processing circuit. Comparing to conventional 2D assembly, 3D image sensor exhibits several obvious advantages: first, the photodetector and processing logic are two different types of circuit. Fabricating them in different layers provides the possibility to optimize them individually in terms of material and process; second, by stacking the photodetector and logic circuit, the footprint is greatly reduced and a compact packaging is resulted; third, the bandwidth between optical circuit and logic circuit is rather limited in 2D platform, which causes the difficulties in producing high resolution image sensors. In contrast, with 3D integration, the signal from each pixel of the image sensor can be directly passed to the downstream processing circuit through interlayer connections, resulting in high speed parallel computation. Its attractive high performance had led to the flourishing research on 3D integrated image sensor thereafter [31, 32].

The monolithic approach of fabricating transistors layer by layer on recrystallized silicon has enabled the pioneer exploitation of 3D integration. However, it has several drawbacks that prevented it from wide adoption beyond laboratories. One of the most important reasons is that the fabrication process is very complicated, and the complexity increases rapidly with the number of active layers. In addition, as a sequential approach, earlier fabricated active layers are involved in later fabrication

process every time when new active layers are added to the existing multi-layer circuit. A stringent thermal budget has to be imposed on the whole fabrication process in order to prevent damaging the circuit components on the bottom layers. Nevertheless, such thermal cycles will still degrade the performance of earlier fabricated devices, and reduce the yield of the final 3D chip.

2.1.2 3D Integration Through Wafer Bonding

In view of the aforementioned disadvantages of monolithic approach, alternative wafer bonding based method was proposed to implement 3D integration. In 1990, a “Cumulatively bonded IC” (CUBIC) technology was proposed by NEC corporation as the first exemplary 3D integration technique through wafer bonding [33]. As shown in Fig. 2.4, the active devices are first fabricated on separate silicon wafers using the conventional semiconductor process. The wafers are then thinned, and bonded together to form the 3D stacked circuit. Compared to the laser recrystallization based monolithic approach, an obvious advantage of wafer bonding based methods is that the individual wafers to be stacked can be processed in parallel, which greatly improves the fabrication throughput. The independence of wafer processing also improves the yield as the quality of each device layer can be ensured before bonding. In addition, with such wafer bonding based approach, heterogeneous integration becomes feasible since the active layers are not forced to use the same semiconductor process and material set.

The wafer bonding based approach offers a promising solution to manufacturability and commercialization of 3DIC technology, and therefore attracted overwhelming attention and research interests from academia and industry. Many other bonding based 3DIC fabrication methods were thereafter proposed which share similar procedures, in the sense that the wafers are first thinned and then bonded together. Their differences reside in the fabrication details, namely the wafer thinning method, alignment tools, bonding techniques, and most importantly how the active layers are interconnected.

Figure 2.4: The CUBIC process flow: (a) Formation of the W bump on MoSi₂/Al wires and attachment of backing substrate; (b) Wafer thinning from the backside; (c) Formation of through hole and the Au/In pool at the backside; (d) Bonding of the thin film and the bottom device layer; (e). Removal of the backing substrate and adhesives [33].

Inter-layer Interconnection

Conceptually, 3DIC is a set of devices that is distributed in a 3D space and inter-connected to implement the specified functionality. The performance improvement of 3DIC is derived mainly from the wire reduction due to the vertical or inter-layer interconnections. For a worthwhile migration into 3D implementation, the inter-layer connectivity and bandwidth should at least be comparable to those of the conventional planar interconnections. To be more specific, the inter-layer interconnections should have sufficient density (allowable number of interconnects per unit area) and per interconnect bandwidth. They should also be reliable (considerably long lifetime and low bit error rate (BER)), cost-effective for mass production and power efficient. Various inter-layer interconnection methods have been proposed in the literature for wafer bonding based 3DIC fabrication. They can generally be

classified into two categories. One is direct wiring, of which the representatives are micro-bumps [33–35] and through silicon vias (TSVs) [36–38]. The other is contactless approaches, such as the optical [45–48, 50, 51] and capacitive/inductive coupling based [52] interconnections.

The direct inter-layer wiring method is a natural extension of the classical 2D metallization. It transmits signal or provides power/ground directly through conductive materials where metal is usually used. For example, in the CUBIC technology [33], the active layers are interconnected through the tungsten bumps formed on MoSi_2/Al wires and the Au/In pools deposited during the back-surface wiring. The wafers are precisely aligned so that the Au/In pools are just above the tungsten bumps. The aligned wafers are then heated and compressed to form the metallic bondings for high quality interconnection. Due to the low melting temperature ($250 \sim 350^\circ\text{C}$) of Au/In alloy, the thermal bonding process will not damage the fabricated devices on the active layers. Besides CUBIC technology, TSVs are usually used to bring signals from the active surface to the back of the silicon wafer without the formation of backside wiring [36–38]. As shown in Fig. 2.5, a representative way to manufacture TSVs is to first create high aspect ratio blind vias through reactive ion etching (RIE). After that, a layer of dielectric SiO_2 or Si_3N_4 liner is deposited on the silicon wall through chemical vapour deposition (CVD) for insulation purpose, followed by the deposition of a thin barrier and adhesion layer with Ti or TiN . A seed layer will usually be deposited for the best electrical property. The blind via is then fully filled with conductive materials such as copper or tungsten through electroplating [37] or CVD [39, 40]. Finally, the entire wafer will be thinned from the back to expose the blind vias so that they become through vias. The idea of using TSV for 3D integration was brought about in the 90s [41]. Through the past twenty years' development, this interconnection technology has been well defined and volume production is currently available as announced by companies such as IBM [42] and Tezzaron [43].

Besides direct inter-layer wiring, novel contactless interconnection techniques such as optical [45–48, 50, 51] and capacitive/inductive coupling [52, 54, 55] based interconnection methods were also explored. The optical interconnection method

Figure 2.5: Typical fabrication process of TSV

had been investigated in the 80s for intra-wafer high speed and low power data transmission [44]. Its application of inter-wafer communication was inspired by the fact that silicon wafers behave as transparent media for light waves with wavelength larger than the silicon's absorption cutoff wavelength of $1.1\mu m$ [45]. Several proof-of-concept prototypes were subsequently demonstrated, with various kinds of optoelectronic emitter/detectors [46, 47, 49], and with or without waveguides [48]. The basic idea of optical interconnection is shown schematically in Fig. 2.6.

The light emitter and detector are placed at two separate wafers. The light wave that carries signal is sent by the emitter, propagates through the silicon wafer with dielectrics and possible lenses. It is then captured by the detector to complete the data transmission. An obvious advantage of replacing vertical wiring with optical interconnection is its low power consumption. Compared to electrical wiring, a simple exemplary calculation shows nearly 400 times power consumption reduction for a $32 \times 32 \times 4$ multi-processor network system [50]. From fabrication point of

Figure 2.6: Schematic of optical interconnection [49]

view, the use of optical interconnection saves the effort of making through wafer vias. With no mechanical wire contact between layers, the reliability problems of inter-layer wires under thermal cycles are also eliminated. However, the density of optical interconnection is typically low compared to direct wiring, due to the bulky and non-scalable optoelectric devices [45]. The density mismatch between vertical interconnects and planar wiring worsens as scaling continues.

The capacitive coupled interconnection method was first presented in [52]. As shown in Fig. 2.7, electrical signals are sent to the transmitter first and then transmitted through the coupling capacitor, and collected by the receiver. Significant

Figure 2.7: Schematic of capacitive coupled interconnection [52]

reduction on delay and power was reported by comparing the SPICE simulation results to the sidewall-metallization method [53]. Unfortunately, there are several limitations associated with this method. For robust signal transmission, the distance between the coupling pads needs to be kept within $1 - 5\mu m$ which is difficult

to achieve. If backside metallization or thin film SOI technology is used as suggested by the authors, it may be more efficient to just apply direct inter-layer wiring instead of capacitive coupling. In addition, the transmitter, receiver and the coupling pad occupy considerable silicon areas. For instance, the coupling pad alone occupies $20 \times 20 \mu m^2$ in the authors' experiment. Considering the large quantity of vertical connections needed, this capacitive coupling based interconnection method would need further study for improvements.

The capacitive interconnects transmit signals through electric field coupling. Correspondingly, inductive interconnections achieve inter-layer communication through magnetic field coupling [54–57]. The general structure of inductive coupled links is as illustrated in Fig. 2.8, where a current change in the transmitter coil causes a voltage signal at the receiver coil, which is then interpreted as digital signals. This

Figure 2.8: Schematic of inductive coupled interconnection

approach exhibits several advantages over the capacitive coupling interconnection. First, inductive interconnection is capable of communicating through a longer distance than capacitive interconnection. Therefore, unlike capacitive interconnection which is normally used only for face-to-face-bonded 2-layer circuits, inductive coupled interconnects can accommodate multi-layer 3D stacks as shown in Fig. 2.8. Second, the inductive coupled interconnection is intrinsically more scalable than capacitive coupled interconnects. The latter is voltage driven, thus can only provide limited transmission power with constrained coupling plate area. While for current-driven inductive coupling, the transmission quality can be improved by increasing

the number of turns of the coil, which can be configured in multi-layer. As a result, inductive interconnect is more adaptive to technology scaling with the lowering of supply voltage. Higher bandwidth and lower power consumption are also observed for inductive links at smaller technology node due to reduced coil size and supply voltage. For example, Miura et. al. [54] reported a 1.25 *Gbps* inductive link at 0.35 μm technology with 4.56 *mW* power consumption at 3.3 *V* and an area of 0.025 mm^2 . When it is shrunk to 90 *nm* technology node, 1 *Tbps/mm*² inductive link with 1 *mW/channel* could be achieved.

Compared with direct wiring such as the TSV technology, inductive interconnection and other contactless approaches have the advantage of eliminating the low yield via fabrication process and ease of pre-bond testing for Known Good Dies (KGDs). However, the large size of these contactless technologies may limit their applications where dense interconnection is required. In contrast, the TSVs can reach the size of less than 1 μm in diameter [60], which is more practical and versatile. Thus, for the remaining of the thesis, TSV is presumed to be the default interconnection technology for 3DIC unless otherwise specified.

Wafer Thinning

Before stacking, the individual wafers carrying the active layers of 3DIC need to be thinned for two main reasons. From circuit performance point of view, wafer thinning is required for short, fast and reliable vertical interconnections. For example, the aspect ratio of TSVs is constrained by the etching process. Assuming a wafer thickness of 700 μm and TSV diameter to height ratio of 1 : 10, the diameter of the TSV would be 70 μm to etch through the wafer, which is too large for high density vertical interconnections. Long TSVs also imply long etching time required as well as additional parasitics and hence the extended latencies. As mentioned before, the TSVs are exposed by removing the bulk silicon from the back instead of etching through the wafer. For contactless interconnection methods, bringing the active layers into close proximity is more important for signal integrity. The second reason for wafer thinning is that compact packaging is highly desired in today's electronic devices. Mobile devices such as cell phones tend to incorporate as many functions

as possible in a small and thin form factor. Wafer thinning is thus needed to meet the stringent space budget.

The silicon wafers can be thinned through physical grinding/polishing or chemical etching. Grinding is a low cost thinning approach and usually provides fast thinning speed of around $5\mu\text{m}/\text{s}$ [58]. However, the induced stress often damages the wafer surface and creates defects. To remove the defect region, a further fine grinding [58], chemical mechanical polishing (CMP) [35] or dry etching [59] may be applied to the wafer for better surface quality. Before the thinning process, the wafer will be temporarily bonded to a glass or bulk wafer substrate for mechanical support. The wafers should in principle be as thin as possible for the best circuit performance. While in practical situations, they are usually thinned to no less than $50\mu\text{m}$ to prevent cracking. In some cases where SOI wafers are used, the thinning may continue until it reaches the buried oxide (BOX) layer, left with only the active layer [60]. The thinned wafer will remain attached to the supporting substrate for the following alignment and bonding processes.

Wafer Alignment

Precise wafer alignment is critical for a functional 3D circuit. The required precision level depends on the size of the inter-layer interconnections such as the diameter of TSVs. In other words, the achievable precision level directly defines the minimum size of vertical interconnection that can be used and hence the inter-layer bandwidth.

Micron or submicron level alignment precision is commonly required for the state-of-the-art inter-layer interconnection specifications. Several alignment techniques were proposed in literature which utilized either optical or infrared (IR) microscopy [61, 62]. Fig. 2.9(a) illustrates the general configuration of the IR microscopy alignment method. The alignment keys are registered on both wafers. An IR light source is located on one side and illuminated so that the keys can be observed and aligned through the optical objectives on the other side of the wafers. The alignment accuracy would depend on the IR wavelength, wafer thickness, planarity and material composition. For cases where submicron precision is needed, transparent supporting substrate is used to reduce IR signal degradation [60]. Fig.2.9 has also illustrated

other alignment methods relying on optical microscopy, namely (b) etched via hole method, (c) transparent substrate method, (d) backside alignment method, (e) intersubstrate alignment method, (f) SmartView alignment method and (g) 3DAlign method. Table 2.1 summarizes the precision level that can be achieved by all the alignment methods. Detailed review of these alignment methods can be found in [61] and [62].

Figure 2.9: Schematic of different alignment methods

Table 2.1: Precision Levels of Different Alignment Methods

Alignment Method	Precision Level
(a)	$\pm 1\mu m$
(b)	Hard to reach micron level precision
(c)	$\pm 5\mu m$
(d)	$\pm 5\mu m$
(e)	$\pm 2\mu m$
(f)	$\pm 0.25\mu m$
(g)	submicron precision

Wafer Bonding

After thinning and alignment, the active layers shall be bonded to form a complete and rigid 3D circuit stack. The bonding is expected to provide sufficient mechanical strength, be compatible with semiconductor process, and be achievable under low temperature so that the active devices are not damaged during the bonding process. There are at least four types of wafer bonding mechanisms proposed in the literature, namely metallic bonding, adhesive bonding, oxide bonding and surface activated bonding techniques.

The metallic bonding mechanism integrates the active layers both mechanically and electrically at the same time. It can be obtained through either thermal compression or eutectic bonding. The thermal compression method often refers to direct Cu-Cu bonding [63, 64]. The upper and lower wafers are first cleaned to remove the oxide and contamination. The wafer pair is then aligned, heated to around 400°C and compressed to stimulate the diffusion process. In the eutectic bonding approach, alloys with low melting point such as Au/In and Au/Sn are used [65]. During the bonding process, the alloy is heated and melted into liquid and then cooled down to form the alloy bonding.

The adhesive bonding method is a simple bonding approach that utilizes polymer adhesives to glue the thinned wafers together [66–68]. Polymer adhesive such as

benzocyclobutene (BCB) is first spin coated on the surfaces of one or both wafers followed by possible patterning. The two wafers are then aligned and compressed for intimate contact. The polymer adhesive is then hardened by heat or UV light to fix the bonding. Compared to other bonding methods, the adhesive bonding does not impose strict requirements on the flatness or cleanness of bonding surfaces. The bonding temperature can be as low as room temperature and up to 450°C. The adhesive bonding provides only mechanical integration. In many cases, it was used together with metallic bonding to fill in the tiny gaps between the metallic bonded wafers and enhance the bonding strength [35, 67]. Due to the big difference in the coefficients of thermal extension (CTE), there is concern on the bonding reliability under thermal cycles. The resulted bonding is not hermetic and thus not resilient to gases or moisture.

Oxide bonding was used in SOI-based 3DIC fabrication process [60, 69–71]. This bonding method requires careful bonding surface treatment and preparation. A thin film of SiO_2 is first deposited on the wafer by plasma enhanced chemical vapour deposition (PECVD) or low pressure CVD. The oxide is then polished by CMP to ensure a very flat surface with a roughness of $< 0.4nm$ RMS. After cleaning and activation by hot H_2O_2 solution, the oxide surface becomes rich of OH groups. Bringing the two wafers into intimate contact will result in weak Van Der Waals bonding between the OH groups, which is nonetheless strong enough for layer transferring purpose. Finally, by applying a heat of $> 150^\circ C$, the weak OH bonding will be converted to strong $Si - O - Si$ siloxane covalent bonds which can sustain the back-end-of-line (BEOL) processes. The advantage of oxide bonding is that it does not involve non-standard CMOS materials and it can sustain high temperature. Adopting this bonding technique can also help to achieve high alignment precision since the wafers are held in position by the Van DerVaals force and will not slide during the bonding process.

The surfaced activated bonding (SAB) technique, as the name suggested, activates the bonding surface through energetic ion beam bombardment in vacuum [72]. After the activation step, the wafer surface becomes very unstable and reactive so that once it is in contact with another activated surface, chemical bondings will be

formed spontaneously even at room temperature. Argon fast atom beam (Ar-FAB) is usually used as the energetic ion beam source. Commonly used CMOS materials have been shown bondable using SAB method such as $Al - Al$ [73], $Al - Si_3N_4$ [73], $Cu - Cu$ [74], $Si - Si$ [75, 77], and $Si - GaAs$ [76] direct bondings. In addition, bumpless $Cu - Cu$ interconnection has been successfully demonstrated using this bonding method [78, 79]. Similar to the oxide bonding method, SAB method does not introduce addition materials for bonding and the bonding is tight as atomic bondings are formed. However, it has difficulties to bond ionic materials such as glass, sapphire and SiO_2 [72]. More research is needed to make the technique mature enough for wide adoption.

2.2 The 3DIC Architecture Under Consideration and Its Associated Thermal Challenges

Three dimensional integrated circuits typically exhibit a stacked architecture with multiple active layers. Despite the differences in dimension, material and microstructures, such stacked architecture can be obtained by many different fabrication approaches, as discussed previously. In this thesis, the promising wafer bonding based fabrication method with TSV interconnections will be assumed unless otherwise stated, due to their obvious advantages on manufacturability and technical maturity.

As shown in Fig. 2.10, every active circuit layer comprises three sublayers, namely the back-end-of-line (BEOL) layer, the device layer and the substrate layer. The BEOL layer contains a number of metal wiring layers which are composed of metal and dielectrics. Its thermal property thus depends on the routing patterns and the percentage of each material in the composition. The device layer contains all the electronic components which are the main source of heat generation. The power consumption of non-radiative electronic circuits is a good estimation of heat generation since the electrical energy will be ideally converted into heat. Otherwise, for radiative cases such as Radio Frequency (RF) and optoelectronic circuits, a part of the electrical energy will be emitted as electromagnetic waves but the heat

Figure 2.10: Schematic of the typical 3DIC structure with heatsink-based air cooling

generation can still be estimated from the conversion efficiency. The substrate layer is the residue of the bulk silicon wafer after the wafer thinning process. It can be considered to be made of silicon and probably with TSVs that punch through. The structure and material in between the active layers are determined by the wafer bonding approach adopted. Fig. 2.10 illustrates the adhesive bonding method where a layer of polymer adhesives was used for gluing and also heat transfer.

As explained earlier, such 3D stacking structure enables significant performance improvement by bringing devices into close proximity. On the other hand, it also creates difficulties in removing the generated heat. Table 2.2 lists the thermal conductivities of some commonly used materials in 3DIC. It shows that the thermal conductivity can vary up to more than hundred times with different material types. Due to bonding and insulation needs, the dielectric materials are frequently used in 3DIC. They act as effective thermal barriers on the heat transfer paths. For example, a $1\mu\text{m}$ -thick BCB adhesive layer is equivalent to a $514\mu\text{m}$ -thick silicon from cooling point of view. Suppose a heatsink is attached to the top active layer for cooling, it would then be much more difficult to transfer the heat from the bottom

Table 2.2: Thermal Conductivity Values of Commonly Used Materials in 3DIC Fabrication

Material	Thermal Conductivity (W/(mK))
Metal	
<i>Al</i>	205
<i>Cu</i>	401
<i>W</i>	173
Semiconductor	
<i>Si</i>	149
<i>GaAs</i>	55
Dielectric	
<i>SiO₂</i>	1.4
<i>Si₃N₄</i>	30
BCB	0.29
Polyimide	0.52

layers through multiple wafer and adhesive layers and dissipate it into the ambient.

The power density of today's high performance computing devices can easily rise to $50W/cm^2$ on average and exceed $100W/cm^2$ at hot spots. Such high level of power density has reached the critical limit of thermal management of 2D circuits [239–243]. Stacking up the active layers effectively multiplies the heat generation per footprint, and makes the cooling of 3DIC even more difficult with the increased heat flux. In view of the aforementioned large thermal resistances, the 3DIC can easily run away from healthy operating temperature, and triggers thermal-induced fatigue and reliability problems [20, 21, 268].

2.3 Thermal Management of 3DIC

It is clear that proper thermal management is mandatory to prevent performance degradation and ensure reliable operation of 3D chips. Various thermal management schemes have been proposed. Some of them are extensions to the thermal management methods originally used for conventional 2D circuits, such as power reduction and redistribution techniques [81, 82, 97, 136, 140, 164, 186, 241]; Others are novel thermal optimization methods founded on the unique characteristics of the 3DIC architecture, such as the thermal TSVs [169–171, 173, 176] and microfluidic cooling techniques [179, 182, 185–191]. The thermal condition of a 3D chip is mainly determined by two factors. One is the amount and pattern of the heat generation, i.e., the power consumption. The other is how effectively such generated heat is carried away, i.e., the cooling methodology. Consequently, there are two possible thermal management approaches, one tackles the thermal problem from power perspective and the other tries to solve the problem by adopting advanced cooling techniques.

2.3.1 Thermal Management from Power Perspective

The electrical energy consumed by active elements acts as the heat source. Therefore manipulating power consumption has a direct effect on the circuit temperature of 3DICs. Power management problem of the conventional 2D circuits is already challenging enough and has attracted a lot of research interests driven by the substantial

power consumption of electronics and the demand for longer battery life of mobile devices. Since 3DICs are made of multiple layers of 2D circuits, most of the power management methods of 2D electronics can also be applied to 3DICs. A straight forward thought of thermal management is the downright adoption of existing power management techniques. For example, during the design phase, the transistors at non-critical paths can be downsized or operated at lower voltage to conserve power [80, 82, 176]. At runtime, techniques such as dynamic voltage and frequency scaling (DVFS) [86–91] and power gating [84] can be used to reduce the dynamic power. However, the aforementioned power reduction methods may lead to performance degradation or less than satisfactory thermal profile depending on the locality and extent of power savings. In most cases, it would be better to redistribute the circuit power adaptively instead of merely reducing the power consumption locally or statically to meet the thermal requirements. Such power redistribution can be performed spatially, such as thermal-aware floorplanning/placement [136, 140] and thermal-aware task allocation [160] for multi-core systems and network-on-chips (NoCs). On the other hand, temporal power redistribution techniques such as thermal-aware task scheduling [158] have also been investigated by many researchers to improve the 3DICs' thermal condition. In the following part, the power reduction and redistribution techniques will be discussed in more detail.

Power Reduction Techniques

The power consumption of electronic circuit can be classified into dynamic and static power consumptions. Dynamic power is consumed due to the switching operation of the active elements while static power is often caused by short circuit and leakage currents. For CMOS circuits, the dynamic power is formulated in (2.4), where α is the switching activity, C is the load capacitance, and f and V are the clock frequency and supply voltage respectively.

$$P_{dyn} = \alpha C f V^2 \quad (2.4)$$

Multiple Supply Voltage Design Approach Due to the quadratic contribution of supply voltage to the dynamic power, voltage scaling has been a very popular power reduction approach. Since lowering the supply voltage will increase the transistor delay and cause timing violation, multiple supply voltage (MSV) design approach has been proposed by many researchers [80, 82, 176]. The general idea is to provide a clustered power supply, where the gates in the critical paths are driven by normal voltage, and other gates with timing slacks are supplied with lower voltages. Such approach has been shown to be very effective in power reduction of large circuits. For example, by partitioning a MpactTM media processor chip into two subcircuits and driving them at 3.3V and 1.9V respectively, the chip's power consumption was reduced significantly by 47% at the expense of an area increase of 15% [80]. The area overhead is mainly caused by the level shifters that were required when signal travels across different voltage zones. Fig. 2.11 shows a typical implementation of a low-to-high level shifter. For signals travelling from high to low

Figure 2.11: Typical implementation of low-to-high level shifter [83]

voltage domain, inverter-based buffers can be used for the voltage transition. Besides the area overheads, the level shifters will also consume additional power and introduce timing delay especially when the clustered supply voltage is transiting

from low to high voltage. Transistors that were supplied with lower voltages will also have delays different from its standard value. Careful timing verification and power assessment are thus required to produce a functional and effective low power design. Fortunately, although the MSV approach may result in very complex power network design and timing analysis, they are essentially static and manageable with the aid of advanced circuit design and simulation tools.

The MSV approach for 2D circuits can be intuitively extended to 3D cases. The first attempt was presented in [81], where a grid-based supply voltage assignment method has been proposed. This method divides every tier of the 3DIC into regular grid cells. For each gate, the tendency of changing its supply voltage named *power_delay_sensitivity-slack* ($PDS - S$) product is defined. Such tendency is proportional to the power difference under different supply voltages and the available timing slack, and is inversely proportional to the extra delay induced by the voltage change. Based on such criteria, proper supply voltage is assigned for every grid cell and the $PDS - S$ values are incrementally updated by considering the thermal effects throughout the process. A 17.59% reduction in power and 4.15°C reduction in temperature have been reported. This work was later refined in [82] to take proximity and level shifter budget into consideration. The voltage assignment strategy was also changed to be more greedy by first assigning all grids with lower V_{dd} and then rescuing the sites with timing violations. As a result, the reduction in power and temperature have been improved to 33.5% and 26.34°C, respectively.

It can be seen that MSV technique is effective in power and hence thermal reduction of 3DIC. In addition, the 3D architecture may provide opportunity to overcome some of the dilemmas in 2D MSV designs. For example, for a feasible power network design, the gates assigned with the same supply voltage are forced to be grouped into a limited number of voltage domains. Doing so elongates the interconnections between the gates at different voltage domains, and increases the propagation delays. In 2D designs, the MSV technique is normally applied in coarse granularity and then routing optimization is performed to minimize the wiring across voltage domains. Such constraint may be relaxed in 3DIC, since if the number of interconnections between two voltage domains exceeds certain threshold, the two domains

can be put into adjacent layers and connected through TSVs with minimum delay overhead. Alternatively, the TSVs can also be used for power delivery to mitigate routing congestions and stimulate new voltage island design algorithms. All these possibilities have great potentials worthy of more rigorous investigation on large-scale 3D circuits. The works presented in [81] and [82] used small circuits (the ISCAS89 benchmarks) for experimentation, thus the aforementioned design problem was not obvious. In practice, much larger circuits such as multi-core systems, system-on-a-chips (SoCs), and network-on-chips (NoCs) are the usual targets for 3D implementations. They are expected to benefit more from the MSV technique, with the development of TSV-based 3D design automation algorithms and tools.

Dynamic Voltage and Frequency Scaling Meanwhile, for those large circuits or systems, dynamic voltage and frequency scaling (DVFS) is also frequently used to tune the circuit performance according to the workload conditions [241]. Dynamic voltage scaling is a natural extension of the static MSV scheme. As the electronic circuit evolves to include more components and functions, it is realized that not all parts of the circuit need to be operated at full power and at all time. When a subcircuit or subsystem is found to be performing non-time-critical tasks, its supply voltage and associated frequency can be temporarily lowered to reduce the power consumption. If a system component is found to be idle, techniques such as power gating [84] and clock gating [85] can be applied to put the circuit into standby mode.

The DVFS method is often addressed on microprocessor-based platforms, due to their wide application, well-defined microarchitecture, and feasibility of algorithm implementation with the aid of operating system (OS) [86–91]. For simplicity, it is commonly assumed that power is inversely proportional to f^3 , and the runtime is proportional to $1/f$. Early works [86–88] tend to consider central processing unit (CPU) alone, where task scheduling was proposed to exploit the available slack of each time slot defined by the scheduler of the OS. The voltage and frequency were adjusted accordingly so that every task was finished just on time. Later, it was observed that whenever a cache miss occurred, the required data had to be fetched from the external memory while the CPU was waiting. As such, CPU-bound and

memory-bound tasks were identified and the CPU clock frequency (and voltage) can be lowered depending on the amount of off-chip activities, without too much performance degradation [89–91]. In recent years, multi-core system has emerged as a viable solution to the problem of power and performance limitations faced by the uni-core architectures. With multiple number of cores, the DVFS strategy can be even more flexible by selecting suitable voltage/frequency of the individual cores according to their running applications or threads. The problem of maximizing throughput under a given power and thermal budget have been studied by some researchers [93, 95, 100]. In all cases, it is critical to obtain the information on the amount and composition of the future workload for correct voltage and frequency assignment. It may not be practical to assume a priori knowledge about future tasks since those information are usually unavailable. Offline techniques such as compiler-assisted profiling and system call insertion [94] were proposed. However, a detailed analysis and modification of the programs were needed and a priori knowledge about the workload was actually implied. A more practical approach would be to use embedded hardware such as performance monitoring unit (PMU) to profile the workload. This method was adopted by several published works [89–91], where the number of executed instructions and memory accesses were read from the PMU to compute the off-chip and on-chip activity ratio. With this kind of history tracking, the future workload can then be predicted with mathematical models such as regression [91], moving average [92] or even simple heritage [90].

As the power level of high performance processors hits the maximum cooling ability of existing packaging technology, thermal budget soon became an important design and runtime constraint. There are many works trying to address the dynamic thermal management (DTM) problem with the aid of voltage and frequency scaling technique [95–100, 241]. In [95], Muraliet. al. proposed a convex optimization approach to obtain optimal voltage/frequency assignment for best performance under power and thermal constraint, where a 2-phase iterative process was adopted to tackle the non-convexity of the DTM problem. In [241], a closed-loop DVFS controller was presented, where the voltage/frequency assignment was determined based on the measurements from the thermal sensors; Zhang and Chatha [96] stud-

ied a stochastic DTM problem, where the applications have uncertain execution time and the performance was optimized to satisfy a low probability of peak temperature violation. Most of those existing DVFS algorithms can be extended to 3D cases. For example, Sun et. al. [97] proposed a temperature optimization algorithm called “3D-Wave” for three-dimensional multiprocessor system-on-chip (3D-MPSoC), where DVFS was used to adjust the processor power according to the slack distribution of scheduled tasks. A thermal management framework for chip-multiprocessor (CMP) was proposed in [98]. By considering the heterogeneous cooling efficiency, power-thermal budget for each core was determined and appropriate DVFS was applied. Zhao et al. [99] formulated the DVFS of 3D-MPSoC into a polynomial programming problem. Besides maximizing the system performance while meeting the thermal constraint, it was also observed that there is greater potential for 3D-MPSoC to gain more performance improvement than its 2D counterpart from DVFS. This is mainly due to the heterogeneous cooling received by heatsink-cooled 3D stack, as opposed to the flat cooling condition of single-layered multi-core systems. In [100], a concept called “temperature slack” was proposed which was defined as the difference between the instantaneous temperature and the maximum allowable temperature. With this concept, the power budget was aggressively assigned to improve the system throughput.

DVFS has been reported as an effective power reduction approach. However, the limitations associated with it were also revealed as technology evolves and more stringent power and thermal constraint are imposed. First, the effectiveness of DVFS is highly dependent on the nature of the workload. For instance, a 70% average energy reduction with 12% performance loss was reported in [91] for memory-bound applications, whereas for CPU-bound applications, the energy saving drops to 15 – 60% subject to 5 – 20% performance degradation. Second, as illustrated by the previous example, a non-negligible performance loss is usually caused by the DVFS approach, which introduces a trade-off between energy and delay. Third, the benefit brought by DVFS is expected to diminish with shrinking transistor size and technology improvement [101]. As supply voltage of CMOS circuit decreasing from 5V, 3.3V to 1V and below from one technology node to the next, the dynamic voltage range

that can be selected by DVFS becomes much narrower now comparing to the days when DVFS was just proposed. In [102] and [103], the technology-driven limitations of DVFS techniques were studied. It was pointed out that the dynamic range of supply voltage greatly influences the effectiveness of DVFS, and the larger process variation induced by advanced technology node will significantly reduce the controllability of DVFS schemes. In addition, with smaller transistors, much larger on-chip caches can now be accommodated, which reduces the cache miss rate and hence the off-chip activities. In the worst case, even when the data need to be fetched from external memory, a much shorter memory access time is now required especially if the core-cache implementation is three-dimensional. Finally, as the circuit complexity increases, the overhead of DVFS that was induced by the extra power network, level shifters and control mechanism will also be exaggerated. The overheads can be mitigated by using simplified or coarse-grained DVFS strategies but that will limit their effectiveness. Therefore, DVFS alone may not be sufficient to manage the power and hence thermal conditions of future 3DIC architectures. The incorporation of other thermal management approaches is needed to ensure the thermal integrity of 3D stacks.

Switching Activity Reduction, Transistor Sizing and Interconnect Power Reduction

The MSV and DVFS techniques focused on reducing the dynamic power through adjusting the voltage and frequency levels. From (2.4), the other two factors that affecting dynamic power are switching activity and capacitive loadings. The switching activity refers to the $0 \rightarrow 1$ and $1 \rightarrow 0$ logic level transitions at the gate outputs from which the circuit function is realized. However, there are unwanted switching activities which cause unnecessary power dissipation. “Glitch” is one such switching activity caused by unbalanced data paths that lead to skew arrival time of gate inputs. The implementation of $f = abcd$ in Fig. 2.12(a) results in unequal arrival time of the inputs to the *AND* gates *and2* and *and3*. Spurious switching occurs at the outputs of *and2* and *and3* before the final result has settled. The glitches can be reduced by an alternative implementation shown in Fig. 2.12(b) owing to the more balanced signal arrival time at the gate inputs.

Figure 2.12: The (a) balanced and (b) unbalanced logic implementations of $f = abcd$

Another commonly encountered case of excessive or redundant switching activity is when the output from a gate or functional block is not used. 26.6% of dynamic power reduction was reported in [104] by gating off the control signals of the bus driver when the data bus is not used in an integer processing unit of a 64-bit microprocessor. Besides equalizing data path delays and eliminating redundant bus switching, another design time approach is to use logic gates with low switching activity through logic manipulation [105–107]. It was found that inverter has the greatest switching activity. Eliminating input inverters were thus proposed in [105, 107]. For example, instead of implementing $f = \bar{x} + \bar{y} + z$ using inverters and *OR* gates, it can be implemented as $f = \overline{xy} + z$ with *NAND* gates. More than 10% of switching activity reduction was reported by this approach in [107].

The capacitive loading of CMOS circuits is closely related to the size of transistors. Larger transistors have faster switching speed and higher driving capability, at the cost of larger area and input capacitance. From the point of view of dynamic power reduction, the transistor size should be minimized without violating the timing requirement [108]. However, a different conclusion can be drawn if conflicting factors such as the short circuit power is considered. The short circuit power of

CMOS circuit is caused by the short on-period of both PMOS and NMOS during the rise and fall times of a transition. Decreasing the transistor size would increase the rise and fall times, thus increasing the short circuit power. As such, a convex relationship between the transistor size and power consumption was presented in [109]. It was concluded that there is an optimal transistor size for power instead of the previously assumed linear correlation. An optimization algorithm that resizes the transistors to meet the timing specification begins with such optimal transistor sizing for power was then proposed.

Capacitive loadings are also contributed by the interconnections. As the wiring delay improved much slower than that of transistors, the weighted contribution of interconnections to delay and power is becoming much more significant in the next technology node of the International Technology Roadmap for Semiconductors (ITRS) [8]. There is a compelling need to reduce interconnection power for future low power design especially for circuits with heavy input/output (I/O) activities. One way to reduce the interconnection power is to reduce the wire length. This is in line with the wire delay reduction and concerns all floorplanning and placement algorithms [126, 127, 151]. New interconnect technology and novel circuit architectures are the long-term solution for revolutionary improvement of interconnection performance. Some of the new interconnect technology under study are TSV [36, 37], optical interconnect [44], and carbon nanotubes [14]. These new interconnections typically show shorter delay, larger bandwidth and lower power consumption over the classical wiring. It is worth noting that all these new interconnection technologies can be adopted in 3DIC, making it an ideal circuit architecture for reducing the interconnection power.

Leakage Reduction The leakage power has drawn a lot of attention in recent years due to its increase dominance of overall power dissipation as transistor shrinks. The leakage current of a transistor is contributed by many different mechanisms such as reverse-biased p-n junction, gate tunnelling and subthreshold leakage [110], of which the major factor in current CMOS technology is the subthreshold leakage. It refers to the phenomenon that the transistor is not completely turned off when

the gate voltage V_{gs} drops below the threshold V_{th} . The leakage current can be calculated by (2.5), where the thermal voltage $V_t = kT/q$. From this equation, it can be seen that $I_{leakage}$ is exponentially dependent on V_{th} . V_{th} reduces with the transistor sizes, which accounts for the alarming standby power dissipation in the advanced sub-micron digital circuits and memory arrays. The associated short channel effects (SCEs) such as drain induced barrier lowering (DIBL) reduce V_{th} further and make the situation worse. $I_{leakage}$ is also exponentially correlated with circuit temperature, which forms a strong electrical-thermal positive feedback loop. Thermal run-away may be resulted if the leakage and circuit temperature are not properly controlled.

$$I_{leakage} = \mu_0 C_{ox} \frac{W}{L} e^{b(V_{dd}-V_{dd0})} V_t^2 (1 - e^{\frac{V_{dd}}{V_t}}) e^{\frac{-|V_{th}|-V_{off}}{nV_t}} \quad (2.5)$$

Existing leakage reduction techniques tackle the problem by device fabrication and circuit design approaches. During fabrication, SCEs can be improved by reducing the gate oxide thickness, junction depth and depletion depth [110]. However, the resulting heavy channel doping degrades the transistor driving capability. Retrograde doping and halo doping were thus proposed to provide further control on the channel doping profile to improve the transistor's on-state characteristic while providing the leakage control [111–113]. Recently, multi-gate devices have become a popular solution for leakage control. Novel structures such as FinFET, Double Gate, Tri-Gate, Pi-Gate, and "Gate All Around" MOSFET have been proposed [114], where the Tri-Gate transistors have already been used in Intel's 22-nm microprocessors [115]. Due to greater control over the channel, multi-gate transistors have good SCE suppression without the need for heavy channel doping, which results in better channel mobility and smaller gate leakage current.

Stacking effect is a frequently exploited phenomenon in circuit design approaches. A smaller leakage current is observed in stacked transistors when one or more of them are at off-state [110]. The leakage power depends on the circuit inputs. It was shown that with different input vectors, the power consumption of a *NAND* gate can vary up to 10 times [118]. During normal operation, it is difficult to constrain the input vector for low leakage. When a circuit or subcircuit is in standby mode, the Min-

imum Leakage Vector (MLV) can be used as inputs to the circuit, as proposed in [116]. Another way to take advantage of the stacking effect is to add control transistors into selected gates [117]. This will increase the number of "OFF" transistors in the stack and reduce the leakage current. Since $I_{leakage}$ is highly dependent on V_{th} , multiple-threshold-voltage design is another commonly used leakage reduction technique. For example, transistors with high threshold voltage can be used at non-critical paths to reduce leakage while low threshold transistors are used at critical paths without violating timing constraint [119]. Using body effect, the threshold voltage of the transistor can even be changed dynamically, according to the operating states [120]. During active mode, the source-body bias can be lowered to speed up the switching of transistor; during standby, the bias can be increased to reduce the leakage. A more sophisticated dynamic V_{th} scaling (DVTS) is made possible by scaling the threshold voltage according to the workload conditions [121, 122], similar to DVFS. According to (2.5), supply voltage is also a contributing factor. Thus, the aforementioned MSV, DVFS and power gating methods can also be used for leakage reduction, by adding leakage power as a triggering parameter into their control mechanisms [123–125].

Power Redistribution Techniques

The thermal condition of 3DIC does not solely depend on the magnitude of power but also its distribution. Due to the heterogeneous functionalities and workload conditions of various components, the power density of electronic circuits is typically non-uniform. As the chip size increases, such heterogeneity will cause large temperature difference between the hot and cool regions on the chip, adding complexity and challenges to thermal management. This is particularly true for 3DICs, as they usually have a more complex and heterogeneous integration than 2DICs. The circuit layers are often thinned to tens of microns to minimize the lateral heat spreading effect. As a result, a properly distributed power density is required at both the design stage and runtime to meet the thermal constraints of 3D circuits.

Thermal-aware Floorplanning and Placement During design phase, the power consumption of functional blocks can be estimated, through either analytical models [252] or simulation with proper input vectors. By associating the power values with circuit blocks, the power distribution problem can be modeled as the thermal constraints of physical design in the floorplanning and placement algorithms.

The primary goal of floorplanning was to pack the functional blocks or system modules in an efficient way to achieve small packing area, short wire length and low power consumption. In conventional 2D designs, this problem was often formulated as a packing problem of rectangular shapes [126–131]. Depending on whether the floorplanning are based on slicing or non-slicing methods, different representations were used, including binary slicing trees [126] and Polish expressions [127] for slicing methods; and corner block list (CBL) [128], Q-sequence [130], bounded-sliceline grid (BSG) [132] and TCG/TCG-S structures [133, 134] etc. for non-slicing methods. The floorplanning problem has been known to be NP-hard. Using stochastic combinatorial optimization techniques such as simulated annealing (SA) method [135] to find the optimal or near optimal solution has a long runtime when the problem size is large. In view of this, more stable and scalable analytical method such as the force-directed approach [136] was proposed.

For most 2D circuits, the heatsink on top of the chip provides an almost uniform cooling to the entire chip, due to the relatively homogeneous chip material and single layer structure. A uniformly distributed power profile keeps the circuit temperatures low. Thermal-aware floorplanning algorithms add circuit temperature into the cost function during the optimization process. For example, the peak steady-state temperature is used as the thermal indicator in the SA optimization engine in [139], where the temperature value was obtained from the compact thermal model Hotspot [222]. Similar strategies may be applied to 3D cases but the following differences need to be considered. First, for the heatsink-cooled 3D stack, different circuit layers have different cooling efficiencies. More power-hungry units are recommended to be placed in layers that are closer to the heatsink [140]. Second, due to the passivation between circuit layers, inter-layer vertical connections such as TSVs behave as fast heat transfer conduit, which causes varying lateral cooling

efficiency. High power-consuming circuit blocks are better placed at locations with more TSVs [141].

Most of the existing 2D floorplan representations can be extended to the 3D designs. If only 2D blocks are involved, a quick solution is to assign the blocks to the layers according to some predefined rules and objectives before applying the 2D floorplanning method to each layer. Flat packing [142], where the blocks can be freely arranged across layers during the process, requires higher computational complexity due to the much bigger solution space. Many 3D components, such as 3D caches [215], 3D arithmetic units [143–145], 3D register files [146] and 3D instruction schedulers [147], have been developed in recent years and many more are forthcoming. The floorplan of these 3D components cannot be directly accomplished with conventional 2D floorplanning methods. 3D cubic packing methods, such as 3D slicing tree [148], 3D CBL [149], sequence triple and sequence quintuple [150], were developed to tackle this problem. The basic concept is similar to the 2D methods except the incorporation of the placement constraints and freedoms from the third dimension. An inherent complexity associated with 3D cubic packing is that most 3D components and modules can have multiple configurations for different design purposes. For example, a 3D cache can have variable layers, bitline and wordline lengths and bank sizes [215]. Thus the shapes and sizes of the 3D cubes can vary during the floorplanning process. This flexibility provides more options in the selection of suitable configuration for optimal floorplanning. Both stochastic and analytical optimization techniques have been proposed for 3D floorplanning [136, 140]. In [140], three floorplanning algorithms with different speed-accuracy trade-off were proposed by adopting the Combined Bucket and 2D Array (CBA) representation, and integrating a compact resistive thermal model with the SA-based floorplan engine. In [136], the force-directed approach is made thermal-aware by adding a “thermal force” to push the 3D blocks away from the hot spots, with careful power and thermal modeling. Juan et. al. [137] exploited the die-restacking capability of 3DIC. A learning-based regression thermal model [138] was adopted to predict the system maximum temperature and the thermal contribution of each tier, which was then used to determine the stacking order with the larger-contribution

tiers placed nearer to the heatsink.

After floorplanning, placement is performed to arrange the cells of a circuit in the given area or volume so that the total wirelength, power and delay will be minimized without overlapping cells or routing congestion. The placement problem is also NP-complete. It would be impractical to produce a legalized and optimized final placement in a single step. Most proposed methods divide the placement task into three stages, as shown in Figure 2.13. The global placement in the first stage is most critical in 3D placement. The objective of this stage is to produce a coarsely optimized placement with limited overlaps. Various placement engines such as partition-based approaches [151, 152], flat placement techniques [153, 154] and transformation methods [155] have been proposed. After the global placement,

Figure 2.13: 3D Placement Flow

legalization will be applied to generate a non-overlapping placement by fine-tuning the cell positions. The legalization process comprises two steps: coarse legalization and detailed legalization. The coarse legalization uses cell shifting techniques to make the cell distribution more even. The 3D circuit is divided into grid of bins. The bins with high cell density will be expanded and those with low cell density will be shrunk. The detailed legalization uses techniques such as DAG-based method and Tetris-Style method [155] to further legalize the result of coarse legalization in fine granularity. In the detailed placement, circuit cells will be swapped locally to refine the objective function. Existing 2D placers can be used if the swapping is constrained to the same layer. Detailed placement methods catered for 3D swapping has also been proposed, such as the Relaxed Conflict Net (RCN) graph based technique [156].

Thermal-aware placement of 3DIC is similar to thermal-aware floorplanning, in the sense that the thermal distribution is also optimized by introducing an additional thermal cost term in the objective functions for refinement iteration [151, 154]. The thermal terms will be related to the net information and placement variables so

that analytical methods can be applied to minimize the maximum temperature. A force-directed approach was proposed in [153], which added a thermal repulsive force to push the cells away from the high temperature cells to achieve an even thermal distribution. The placement performances of various transformation-based techniques were investigated using a compact thermal model in [155]. In [151], a net weighting technique was adopted. By adding *thermal resistance reduction nets*, high power units were pulled towards the heatsink for better thermal results. At the placement stage, TSV is a key component that distinguishes 3D placement from 2D placement. The insertion of TSVs adds fast heat conduction paths to alleviate the heat accumulation. Thus, it is proposed to leverage existing signal TSVs. For optimal thermal condition, it is proved in [141] that the power distribution should be proportional to the percentage area of TSVs in a bin. Other works [169–171, 173, 176] that place TSVs to address the thermal issues will be reviewed under the subject of thermal TSVs in 2.3.2.

Thermal-aware Task Allocation and Scheduling Thermal-aware floorplanning and placement provide a good start point to meet the thermal requirements by arranging the circuit blocks based on their anticipated power consumption. These methods are usually based on averaged power consumption values while the actual runtime power may vary due to different input excitations and runtime power reduction techniques used. On the other hand, for circuits with homogeneous computing units, e.g., multi-core systems, there is little room for floorplanning and placement optimization since these identical cores have similar nominal power rating. However, during operation, the real power consumption of the cores may vary over a wide range. Runtime power redistribution mechanisms are needed to prevent possible thermal violations caused by such power variation. As the location of circuit blocks have already been fixed, the power can only be redistributed by arranging the inputs of every circuit block. Such arrangement can be executed in two ways. One is the spatial arrangement which assigns the tasks to different circuit blocks. The other is the temporal arrangement which determines the order of the task sequences of the computational units. The former is usually referred to as task allocation and

the latter as task scheduling. They were initially studied to maximize the system throughput or energy efficiency [157]. Soon it was found that meeting the thermal constraint is an equally important optimization criterion for cooling cost reduction and reliability.

The task allocation and scheduling problem was usually studied on processor-based system, especially multi-core systems and Network-on-Chips (NoCs). The problem was often formulated as a selection and assignment problem of tasks from a global queue or task pool to individual cores or Intellectual Property (IP) blocks to generate their local task queues, subject to power-performance trade-off and thermal constraints. In the degenerated case where a single core structure is present, only task scheduling would be considered. A complete thermal-aware task allocation and scheduling should contain three components, namely the prior knowledge of future tasks, a system temperature monitoring mechanism, and the allocation and scheduling algorithm.

A general rule that is obeyed by most allocation and scheduling proposals is to assign high power tasks to low temperature cores [160] and interleave high power and low power tasks [158] to mitigate the spatial and temporal activity intensity. In this case, the power consumption and thermal effect induced by a task need to be known a priori. One way to gain such knowledge is to profile the thermal impact of tasks offline with thermal sensors, as proposed in [159–161]. During runtime, the tasks were said to be recognizable through metadata comparison, and the corresponding pre-stored thermal profiles was invoked to facilitate decision making. For unrecognized tasks, its metadata and thermal trace will be recorded as a new entry to catch similar future tasks. The other method is to track the power of tasks online based on their instructions per cycle (IPC) ratings. It was observed that the power consumption of a microprocessor is closely related to the IPC count [162, 163]. The task power can be estimated based on the readings from the build-in performance counter of the processors. The obtained power values can then be used to predict the power of similar future tasks. For its feasibility and ease of implementation, this method was favored by most relevant works [158, 164, 167].

For thermal-related decision making, there must be a way to monitor present

(and past) circuit temperature(s). A simple and straight forward way is to read the temperature directly from the build-in thermal sensors of the processors [159–161]. Most modern microprocessor has hardware-embedded thermal sensors. The overhead caused by reading and storing the collected data is trivial but the silicon area overhead of thermal sensors cannot be neglected. On the other hand, the reliability and usefulness of the temperature readings are highly dependent on the quantity and locations of the sensors. Due to the heterogeneity of computational units, the power density and the circuit temperature are usually non-uniform. To provide a high resolution temperature map, a high number of thermal sensors is required which incurs a high area overhead. The maximum circuit temperature is often the designer’s top concern. It normally appears at the high power density regions. To obtain the accurate maximum reading, the thermal sensors will be cramped at such critical region which may cause congestions and degrade the performance of the design. Alternatively, software-sensing methods, which estimate the circuit temperature by accurate thermal models, were adopted by [165, 167]. Thermal models typically provide moderately accurate temperature prediction, but its resolution is much higher and customizable. The underlying thermal formulation can also provide guidance for task allocation and scheduling, and predict future temperature if the future task powers are known. Due to runtime requirement, fast compact thermal models such as Hotspot [222] were often used. This temperature monitoring method was also adopted in my research work due to its versatility. More details about existing thermal models will be reviewed in Section 2.4 of this chapter.

Knowing the power consumption of tasks and circuit temperature, the only problem left is how to arrange the tasks based on these information. Powell et al. [166] proposed a simple strategy called “heat-and-run” to maximize the system throughput by assigning and scheduling tasks that fully utilizes the system resources. Thread migration from a overheated core to a cooler core was performed only if a thermal violation was observed. The weakness of this method is that the cores may be overheated very quickly and frequent thread migration will degrade performance. Coskun et al. [160] proposed an “adaptive-random” task allocation algorithm, which assigns tasks according to a probability value associated with every core. The prob-

ability value was updated at every time interval based on the processor temperature history collected from the thermal sensors. A cooler processor will have a higher probability to get job assignment. This method is a more sophisticated application of the simple idea of assigning tasks to cooler processor. Since the thermal impact of individual tasks was not considered, “adaptive-random” method alone could not ensure thermal integrity. Podobas and Brorsson [158] proposed a task interleaving scheme, where multiple global queue were kept and sorted according to their power consumption. Tasks with appropriate power ratings were then selected based on the current temperature. This work adopted a simplified RC model for thermal simulation. A mixed-integer linear programming (MILP) approach was proposed in [168], which fully described the thermal-aware allocation and scheduling problem. However, since such problem is NP-complete, the MILP method cannot deal with large-sized problem. To ensure thermal integrity, Yang et al. [167] proposed a thermal-aware scheduling algorithm that simulates the circuit temperature with every candidate task based on the predicted power. The task that induced the lowest circuit temperature was chosen.

All the aforementioned algorithms were proposed for 2D circuits. Nevertheless, they can all be used for 3DICs with simple modifications. The only difference between 2D and 3D scheduling is the thermal behavior of 3DIC stacking structure. Circuit layers near the heatsink are generally cooler and the thermal conditions of adjacent circuit layers are strongly correlated. This difference has been exploited for 3DIC thermal-aware allocation and scheduling algorithms. The “adaptive-random” method was extended to a 3D MPSoC task allocation algorithm “Adapt3D” in [161]. In the extended version, the probability values were updated with a thermal index value taking into consideration the heterogeneous cooling received by different circuit layers. Zhou et al. [164] proposed the concept of “super core” and “super task”, which grouped vertically correlated cores and assigned the tasks to the super cores to distribute the power evenly. Furthermore, high power tasks were assigned to layers near the heatsink to utilize their higher cooling efficiency. In [165], the thermal simulation efficiency was improved by incrementally updating the temperature. The empty cores with high-power tasks were filled for thermal simulation to account for

the inter-core thermal influences. In a nut shell, 2D algorithms for task allocation and scheduling can be adapted for 3DICs by incorporating 3D-specific properties and accurate thermal modeling.

2.3.2 Thermal Management by Advanced Cooling Techniques

Power reduction and redistribution techniques are most effective if there is enough slack in the power budget or the original power distribution is obviously uneven. These circuit design and runtime management techniques maximize the cooling efficiency of the available cheap cooling solution, e.g., heatsink-based air cooling. Unfortunately, as power density increases, the circuit temperature may still rise beyond the safety threshold even with a reduced and optimally distributed power consumption. This is very likely to occur in 3DICs due to their stacked power density and large thermal resistances. Applying aggressive power reduction such as heavily scaling supply voltage and frequency to meet the thermal requirement will scarify performance which is not suitable for tasks with hard deadlines. The total energy consumption would also be increased due to longer execution time. The capability of the most advanced devices will be undermined by running it at a slower speed than they could most of the time just to avoid the peak operating temperature. Consequently, more advanced cooling techniques are needed when power reduction and redistribution techniques are insufficient to maintain the circuit temperature. This is analogous to maintain the water level of a swimming pool. If the inlet has reached its maximum tunable range while the water level (temperature) is still increasing, a larger outlet is needed to expel water (heat) at a higher rate of flow. Various cooling solutions leveraging the 3DIC's stacking structure have been proposed, which can be divided into two main categories. One way is to improve the current heatsink-based air cooling method with novel and economical heat conduction structures, such as the insertion of thermal-TSVs (T-TSVs) [169–171, 173, 176]. The other way is to use active cooling techniques with a higher cooling capability, such as microfluidic cooling methods [179, 182, 185–191]. In this section, T-TSV enhanced air cooling and microfluidic cooling techniques will be reviewed.

Thermal-TSV Enhanced Heatsink-based Air Cooling

The large thermal resistance between the circuit layers and heatsink is one major reason that made 3DICs prone to the thermal problem. Such large thermal resistance is caused by the stacking structure of 3DICs, especially when multiple dielectric layers are present along the heat transfer path. Without changing the 3D stack architecture, the large thermal resistance can be reduced by inserting Thermal-TSVs (T-TSVs). TSVs are commonly used as interlayer connections in 3DICs. As introduced in Section 2.1.2, a TSV is often made of metal and surrounded with a thin layer of dielectric material. Table 2.2 shows that the thermal conductivity of the used metal is up to 1.7 times better than silicon and over hundreds times better than dielectrics. Therefore, as TSVs punching through silicon wafers for electrical interconnection, a fast heat transfer path is also provided at the same time. The number and locations of power and signal TSVs are often constrained by timing and power supply quality requirements, and are usually not optimized for thermal purposes. Therefore, dummy TSVs that are electrically isolated can be inserted merely for the interest of thermal management. These TSVs are called T-TSVs [169–171, 173, 176].

Although T-TSVs can effectively improve the local thermal conductivity, they are commonly seen as blockages during placement and routing. Such conflict between thermal and area introduces a trade-off where a minimum number of T-TSVs is required to achieve the thermal targets. To solve this problem, a complete knowledge about the heat transfer behavior within the 3D architecture is needed so that T-TSVs can be inserted at the most critical position. In [169] and [170], the T-TSV placement was integrated into the multilevel routing process. The T-TSV minimization problem was first formulated as a nonlinear programming (NLP) problem. Two heuristics are then proposed, namely VPPT [169] and m-ADVP [170], to optimize the number of T-TSVs with thermal constraints. VPPT stands for “via planning proportional to temperature”. It is an intuitive algorithm to put more T-TSVs at hot locations. m-ADVP stands for “multilevel alternating direction via planning” T-TSV planning algorithm. It considers vertical T-TSV insertion and horizontal T-TSV insertion separately. The vertical T-TSV was planned through a convex

optimization, and the horizontal T-TSV was placed according to the heat flows to propagate the heat flux from the hot spots to the cooling regions. Due to a more precise formulation with thermal models, m-ADVP was shown to be able to reduce 68% of TSVs as compared to VPPT algorithm. These two works were further improved recently in [171] by considering the thermal related leakage and delay during the T-TSV insertion process. The T-TSVs were placed based on a weighted model, where the vertical T-TSVs were placed according to the power density and the horizontal T-TSVs were placed according to a delay and leakage weighted heat flow index, constrained by the available TSV capacity. It was shown that leakage and delay does have a strong impact on the final T-TSV distribution. A leakage and delay aware T-TSV is therefore needed for practical designs.

The methods presented in [169–171] are flat T-TSV placement approaches. Goplen and Sapatnekar [172], [173] proposed an alternative T-TSV algorithm with predefined structure. The T-TSVs were allowed to be inserted only at the specified thermal via regions. By changing the density of the T-TSVs, the thermal conductivity of the thermal via regions could be adjusted accordingly. The main purpose of creating such thermal via regions is to eliminate the randomness of routing congestions created by T-TSVs. Since the T-TSVs were fixed in such a framework, the only problem left was to determine the TSV density of each thermal via region. Finite element analysis (FEA) was adopted in [172, 173] to model the thermal condition of 3D structure. To meet the desired thermal requirement, the thermal conductivity needed for each thermal via region was calculated. The TSV density of each thermal via region was then obtained. The FEA and TSV density calculations were iterated until the convergence was reached.

Since not only T-TSVs but also all other TSVs, such as power and signal TSVs, conduct heat as well, there are works that attempted to utilize the exiting TSVs so that the T-TSVs added are minimized. Cong et al. [141] proposed to use signal TSVs only by arranging the TSVs in proportion to the power density. A 34% peak temperature reduction was obtained with a slight wirelength overhead. In [174, 176], Yu et al. managed to co-optimize the power/group and thermal TSVs so that both power and thermal integrity were addressed and the number of non-signal TSVs

was minimized. A macromodel was built and reduced to formulate the optimization problem. The problem was solved by a sensitivity based iterative optimization process. Transient temperature was proposed for thermal monitoring instead of steady state temperature to avoid the over-estimation of thermal requirements [175]. A 34% reduction of the non-signal TSVs was reported by simultaneously optimizing the power and thermal TSVs.

Although significant temperature reduction was claimed by most T-TSV related works, it was found that such large reduction may be due to the inappropriate boundary condition, which assumed an isothermal top or bottom layer with room temperature for their heatsink model [169–173]. With proper thermal modeling, the improvement by inserting T-TSVs was less impressive [177]. In addition, as TSV diameter and pitch shrink, its dielectric liner starts to show thermal blockage effects for the lateral heat flow [178]. While T-TSVs may be added to address some local thermal emergencies, with current material composition, it may not be an ideal solution for the thermal problems faced by 3DICs, considering its induced congestion, limited thermal improvement and non-scalability as the number of circuit layers increases.

Microfluidic Cooling for Thermal Management of 3DIC

The conventional planar cooling techniques are no longer suitable for 3D circuits since they are not scalable on the third dimension. Irrespective of the surface of the 3D stack to which the heatsink is attached, some circuit modules will always be far from the heatsink and experience a high thermal resistance. Hence, liquid cooling that utilizes interlayer cavities was proposed. By forcing the liquid coolant to flow through the cavities, the generated heat is absorbed and carried away from both the upper and lower circuit layers. The number of cavity layers increases along with the number of circuit layers, thus provides local thermal ground to every circuit component. The interlayer cavity can have different shapes, including microchannel [179], pin and fin structures [191], where microchannels were adopted and modeled by most works due to its regularity and implementation feasibility.

Microchannel-based microfluidic cooling for electronic chips was first proposed

in [179], where the microchannels were fabricated at the back of silicon substrate and water was used to cool a single-layered chip. Maximum thermal resistance of around $0.1K/W$ was obtained and a maximum heat dissipation of $790W/cm^2$ was demonstrated. In comparison, the junction-to-ambient resistance of the state-of-the-art air-cooled heat sink is about $0.6K/W$. Heat dissipation of over $100W/cm^2$ with the air cooling solution poses a problem [180]. Inspired by the impressive and scalable cooling potential, many recent research works [182, 185–190] have turned to microfluidic cooling method for solving the cooling problem of 3DICs. The general scheme of microfluidic cooling method is shown in Fig. 2.14 [181, 182]. The cold

Figure 2.14: Schematic of microfluidic cooling structure for 3DICs

coolant is fed through interlayer microchannels of the 3D stack with an external micropump. After absorbing heat from the 3D circuit, the hot coolant is extracted from the outlet and sent to the external heat exchanger to “unload” the absorbed heat. The coolant is then pumped into the 3D stack again to form the complete liquid cooling loop. The philosophy behind such an approach is to move the thermal problem from sites where simple and cheap solutions are either not permissible or

ineffective due to the space and architecture constraints to places free from these constraints. The microchannels can be fabricated by following the steps shown in Fig. 2.15, where an array of deep trenches is first etched on the silicon substrate and then covered by another piece of wafer or polymer plate through bonding [181, 182]. In this thesis, the direct wafer bonding method [75, 77] is assumed, which avoids the use of non-CMOS process materials.

Figure 2.15: Microchannel fabrication process flow

Although microfluidic cooling possesses remarkable heat removal capability, it introduces additional implementation costs and overheads such as micropumps, fabrication of the fluid network and the fluid I/O components [182]. It is imperative to optimize the cooling performance while minimizing the implementation overheads. The heat absorbing capability of the microchannel array is often characterized by the heat transfer coefficient (HTC), which is related to the channel geometry and fluid properties such as its thermal conductivity, viscosity and flow rate. Several works have attempted to maximize the microchannel's HTC at the design stage by optimizing the channel geometry [183, 184]. On the other hand, the power consumed by the cooling package at runtime adds to the operational cost. The cooling power of micropumps is formulated in (2.6), where ΔP denotes the pressure drop along the microchannel and Q denotes the flow rate.

$$P_{pump} = \Delta P \cdot Q \quad (2.6)$$

Since the microchannel geometry is fixed during runtime, the cooling power could be solely determined and tunable by the flow rate. Increasing the flow rate will cost more cooling power but obtain better cooling result. The runtime thermal management solutions aim to minimize the cooling power under thermal constraint. In [185], Shi and Srivastava proposed a non-uniform microchannel design technique

that optimizes the channel number by iteratively removing the channel that incurs the lowest thermal penalty. A DTM technique was then applied to adaptively change the flow rate according to the Kalman-filter-corrected thermal sensor readings. A series of runtime thermal management schemes that combines microfluidic cooling with DVFS and task scheduling techniques have also been proposed [186–189]. In [186], reactive and proactive load balancing approaches under microfluidic cooling with both fixed and variable flow rate settings were demonstrated. An autoregressive moving average (ARMA) time series model was used to forecast the future temperature. In [187], a look-up-table-based flow rate control mechanism was proposed. A weighted load balancing scheme was adopted this time so that the hot processors execute less tasks. 30% cooling energy reduction was achieved compared to the worst-case design. In [188], a fuzzy DTM controller was proposed to account for the processor location (with respect to microchannel, i.e., up-stream or down-stream). Different DVFS settings were decided based on the processor location and temperature. A 63% reduction in cooling energy was obtained. A hierarchical DTM policy was proposed in [189]. Under this policy, the thermal management task was shared by the global and local controllers. The global controller sets the flow rate based on the temperature measurements and partitions the workload of each circuit layer with a local controller. The local controller then implements DVFS strategies based on the workload and temperature requirements. A further 50% average cooling energy reduction was reported, as compared to previous proposals. An intra-layer thermal balance within 10°C was also reported. In general, the cooling power can be greatly reduced if the flow rate setting matches well with the 3DIC’s cooling demand. However, due to the non-uniform circuit power distribution, it would be difficult to match the cooling effort and demand by a uniform flow rate setting. The idea of non-uniform flow rate distribution presented in this thesis was motivated by this observation.

In most microfluid-based DTM schemes, the peak temperature and thermal gradient ¹ were normally selected to be the evaluation criteria. Interestingly, although

¹this term is used to express the range of temperature variation of the circuit, i.e. $\Delta T = T_{max} - T_{min}$ over a period of time. As the common time denominator has often been omitted

microfluid-cooled 3DIC usually exhibits a low peak temperature, the thermal gradient remains large and difficult to be reduced [188, 189]. This is due to the fact that the coolant temperature continue to rise as it absorbs heat along the microchannel. This phenomenon was noted by Sabry et al. [190] and they proposed a channel modulation technique. By varying the channel width along the flow direction, the circuit temperature rises due to convective thermal resistance can be controlled and thus offset part of the coolant thermal building effect. Although theoretically correct, the proposal is impractical since it requires independent liquid network for every microchannel. Therefore, a more feasible solution to the large thermal gradient is needed to take advantage of the strong heat removal capability of microfluidic cooling.

2.4 Thermal Modeling of 3DIC

From the review of various design-time and runtime strategies for effective thermal management of 3DIC in the previous section, fast and accurate thermal modeling is always needed to offer quick thermal assessment at the design stage or temperature estimation and prediction during runtime. For the solid-state circuit, the generated heat is mostly transferred through conduction within the chip. Such heat conduction can be described by the Fourier heat diffusion equation (2.7) under Cartesian coordinate [223], where $k(x, y, z, T)$ is the temperature-dependent thermal conductivity of the material, $T(x, y, z, t)$ is the circuit temperature, and $q(x, y, z, t)$ is the outgoing heat flux. In practice, the weak temperature dependence of k is often ignored for simplicity. The error due to its omission is negligible.

$$\nabla(k(x, y, z, T)\nabla T(x, y, z, t)) = -q(x, y, z, t) \quad (2.7)$$

This partial differential equation (PDE) can be solved by methods such as finite element analysis (FEA), finite difference methods (FDM) and Green function based approaches. The FEA-based approach is often adopted by commercial simulation tools such as ANSYS [218] and COMSOL [219] due to its high accuracy. However, conveniently, it refers more to the difference rather than gradient.

the large number of meshes created often results in significant memory usage and long computation time for interactive thermal-aware design and runtime thermal management. FDM and Green function based approaches are thus more frequently used in the literature and the FEA-based tools are usually reserved for verification purpose. Thermal modeling of 2D chips has been well studied especially for high performance circuits such as MPSoCs [192, 222]. Compared to the conventional 2D design, the problem size of 3DIC is much larger with heterogeneous material composition due to the bonding and TSVs. The use of microfluidic cooling techniques also invalidates existing heatsink-based air cooling models and adaptation of fluid thermodynamic theories is needed for correct modeling. In this section, thermal modeling approaches for 3DICs with air cooling and microfluidic cooling will be reviewed.

2.4.1 Thermal Model of Heatsink-Based Air-Cooled 3DICs

In heatsink-based air-cooled 3DICs, a heatsink is attached to the top or bottom of the 3D stack to dissipate heat into the ambient through convection. The convection efficiency is often characterized by the convective resistance which can be found in the heatsink's specification. Within the 3D stack, the main heat transfer path is in the vertical direction where the generated heat propagates through multiple circuit layers until the heatsink is reached.

FDM-based Approaches

Finite difference methods based on thermal-electrical analogy² are often adopted to solve the heat transfer problem. Assuming that the 3D stacks are composed of multiple homogeneous layers, the simplest approach [193, 194] would be to assume a constant ambient temperature as the thermal ground and represent each layer with a thermal resistance as shown in Fig. 2.16. The resistances for such simple model can even be hand-calculated with an average temperature value obtained for each layer. This one-dimensional thermal analysis can provide instant temperature estimation

²where temperature corresponds to voltage, heat flow corresponds to current, thermal resistance corresponds to the ohmic resistance, and heat capacity corresponds to capacitors.

for 3DIC but non-trivial error is expected if there exists intralayer heterogeneous material composition or large lateral power variations.

Figure 2.16: One-dimensional resistive modeling of 3DIC

Hotspot [222] is a more sophisticated FDM-based thermal model that has been widely used for thermal modeling of microprocessor and general circuits. In this model, the chip is divided into multiple grid cells. Each grid cell is regarded as a node and represented by six thermal resistors (and a thermal capacitor if transient analysis is required). The thermal network is then converted into an equivalent RC circuit. For each node, the following ordinary differential equation can be established.

$$C_i \frac{\delta T_i}{\delta t} + \sum_{j \in i's \text{ neighbors}} ((T_i - T_j)/R_{ij}) = 0 \quad (2.8)$$

(2.8) can be solved by Euler's method or Runge-Kutta method. The latter is adopted in Hotspot for transient analysis so that the system temperature is updated after every elapsed time interval. For steady state analysis, $\frac{\delta T_i}{\delta t} = 0$ is substituted into (2.8) and the thermal problem can be generalized and simplified as

$$GT = P \quad (2.9)$$

where G is thermal conductance matrix. T and P are the temperature and power arrays, respectively. Typically, G is a highly sparse matrix and (2.9) can be solved

by existing sparse solvers such as SuperLU or KLU. In Hotspot, the steady state problem is solved by a recursive multigrid method. The existing grid resolution is first lowered to obtain a rough temperature estimation based on the total power and lumped thermal resistance. More accurate results are then achieved by iteratively doubling the grid row and column resolutions to obtain the new convergent steady state results. This recursive approach has the advantage of low run time complexity, but the grid resolution is limited to a power-of-two integer, i.e., 2^n . Hotspot was initially proposed for thermal evaluation of 2D circuits. From version 2.0, it has also included modules for 3DIC thermal modeling. However, its 3DIC modeling ignores the intra-layer heterogeneity, thus may introduce non-negligible error.

Due to its simplicity, FDM-based thermal modeling approach is frequently used [170, 192–195]. Similar to Hotspot, most methods convert the thermal modeling problem into a resistive circuit simulation problem. This approach can generally offer moderate accuracy and reduced execution time compared with FEA methods. However, when the grid resolution is high, the execution time can be long with a large amount of memory used.

Power Blurring Method

An alternative approach to avoid this problem is the Green function based method such as power blurring (PB) [196–198]. The PB method treats the power map as an image, and the method is named after the observation that the temperature distribution appears like a blurred version of this image. From signal processing point of view, the power distribution is analogous to the input and the 3D stack is analogous to a linear system. The output (temperature) can thus be obtained by the convolution of the power map with the system impulse response, which was often referred to as thermal masks in the literature. As such, image processing techniques such as discrete cosine transform (DCT) can be applied to speed up the simulation process [199]. The thermal masks can be obtained by applying a point heat source in the chip and collect the resultant steady state temperature profile. At this stage, the steady state temperature needs to be obtained by using other thermal models or through physical measurements. Since the accuracy of the

PB method is strongly dependent on the quality of the thermal masks, accurate FEA-based thermal simulators are usually adopted. Long simulation time and high resource requirement are expected. However, such kind of simulation needs to be run only once for a fixed circuit architecture. After that, very efficient and fast thermal simulation can be achieved.

The PB method was initially proposed for 2D circuits [200, 201], and later extended to model 3DICs [196]. For 3DIC thermal modeling, multiple thermal masks need to be prepared to account for the inter-layer thermal coupling effects [197]. Due to multiple circuit layers, the problem size is usually large. Thermal mask collection with accurate thermal models may hence be run time inefficient especially when a high resolution is required. In [197], such problem was solved by generating thermal response of every homogeneous area separately. Since the homogeneous area is smaller than the full chip, error will be introduced in the obtained thermal response but can be corrected by removing the offset and adding on the “far” responses. With carefully generated thermal masks and error compensation techniques [197, 198], PB method can usually achieve accuracy close to FEA-based methods, and can reduce the execution time by six orders of magnitude. As long as the circuit infrastructure remains unchanged, there is no need to rebuild the mesh or resistive network for new simulations. It is thus well suited for tasks such as thermal-aware floorplanning and placement and runtime thermal tracking with pre-stored thermal masks. The main weakness of PB method is that it cannot be stand-alone and must rely on other thermal models for thermal mask preparation. In addition, the long run time (hours to days) of multiple high resolution FEA simulation for thermal mask generation makes this method inappropriate when the circuit contains reconfigurable components such as a fan with multiple speed settings.

Thermal Effect of TSVs

In accurate 2D thermal models, the substrate, active and back-end-of-line (BEOL) sublayers have already been identified and modeled separately. The 3D extension is in principle merely an incremental update by adding more circuit layers provided that all the layers are homogeneous. Unfortunately, due to the existence of interlayer

interconnections, specifically the TSVs, the intralayer homogeneity is annihilated and a straightforward extension of existing 2D thermal models is inadequate.

Since both heat transfer path and thermal improvement by TSV are mainly in the vertical direction, the simplest way to include TSV is by adding lumped vertical thermal resistances in parallel with the corresponding thermal resistors if FDM is adopted [193, 194]. This simple modification tends to over-estimate the vertical heat conduction. In more accurate thermal models, the lateral thermal resistance, taking into consideration the dielectric liners, induced by TSV is also added [178, 195]. Fig.2.17 illustrates these two different resistive approaches. The vertical resistance is generally proportional to the length of TSV, and inversely proportional to the TSV's cross-sectional area. The lateral resistance is proportional to the diameter of the TSV, and strongly dependent on the thickness of the liner.

Figure 2.17: Modeling of TSV's thermal effect with and without considering the lateral thermal resistance

TSV inclusion in FEA-based thermal models has also been investigated. If the mesh network has a high granularity, the TSV itself will be divided into smaller homogeneous elements. In this case, no extra action will be required except to

provide the correct thermal conductivity settings for the metal and dielectric regions [202]. However, this will require a very high mesh granularity and a long simulation time especially for thin TSVs with diameters less than $5\mu m$. As a compromise, the equivalent thermal conductivity of the TSV regions along all three dimensions is calculated. The equivalent thermal conductivity can be calculated as proposed in [172] as follows:

$$\eta = \frac{A_{tsv}}{A_{region}} \quad (2.10)$$

$$k_z = \eta k_{tsv} + (1 - \eta) k_z^{region} \quad (2.11)$$

$$k_x = k_y = (1 - \sqrt{\eta}) k_{lateral}^{region} + \frac{\sqrt{\eta}}{\frac{1 - \sqrt{\eta}}{k_{lateral}^{region}} + \frac{\sqrt{\eta}}{k_{tsv}}} \quad (2.12)$$

Note that A_{tsv} is the total area of the TSVs in the TSV region. Although the dielectric liner was not explicitly considered in [172], it can be included by using equivalent TSV thermal conductivities.

Homogeneity has been a prerequisite of Green function based thermal models. Only then can (2.7) be degenerated to a Poisson equation and the solution be represented by Green functions [203]. However, with some modification, the Green function based method can still be used for thermal modeling of 3DICs with TSVs. Oh et al. [203] proposed a virtual power source (VPS) method. It converts the TSV-induced heterogeneous problem into a homogeneous problem by adding VPS at the TSV locations. The acceleration methods such as DCT and FFT can be adopted to speed up the simulation. Ziabari and Shakouri [198] extended the PB method for use on 3DICs with TSVs. In addition to the normal silicon thermal masks, they proposed the viathermal masks to count for the TSVs' effect. The power map is divided into mesh elements, and each mesh element is convoluted with its corresponding thermal masks. This way superposition principle can be applied to model the 3DICs with TSVs.

2.4.2 Thermal Model of Microfluidic Cooling in 3DICs

In microfluid-cooled 3DICs, the liquid coolant flows through the interlayer microchannels to absorb and carry the heat away from the 3D stack. The generated heat is transferred to the coolant through convection at the microchannel walls.

Unlike heatsink-based air cooling where the convection between air and heatsink is usually characterized by the convective resistance, microfluidic cooling model necessitates basic thermodynamics. Since the microfluidic model usually involves variable coolant flow rate, Green function based methods are no longer suitable. Therefore, most of the proposed microfluidic thermal models have adopted the FDM-based approaches.

Microfluid-cooled 3DICs are often modeled by assuming adiabatic boundary conditions due to thermally isolated packaging material. The generated heat is assumed to be dissipated only through the microchannels to simplify the problem. Mixed cooling models are always possible by adding more thermal grounds with proper convective resistances. The 3D stack can be modeled in two parts. One part is the solid part, which can be modeled by grid division and representing every grid cell by thermal resistors as shown in Fig. 2.18. The other is the coolant part, which

Figure 2.18: The six-resistor representation of a grid cell

contains flowing liquid that needs to be modeled by fluid thermodynamics. The thermal networks of both parts are then connected to form a complete thermal circuit and solved. Since modeling of the heat conduction in solid is mature and less prone to errors, the accuracy of microfluid-cooled 3DIC will depend heavily on the

correct modeling of heat convection, conduction and mass transfer in the fluid part.

Microfluidic cooling for electronic circuits was first proposed in [179]. The authors derived the formulation of microchannels' thermal resistance by following conventional thermodynamic theory and provided a very coarse-grained convective resistance estimation (using one thermal resistance to represent all microchannels). Koo et al. [204] studied the problem in finer granularity by dividing both the solid and fluid parts into grid cells. Energy conservation equations for all the grid cells were formulated in a finite difference style. The equations were then solved by numerical simulation based on the assumed initial values until convergence. In [205], the thermal formulation in [204] was enhanced to allow non-uniform heat flux. Microfluidic cooling with different cooling modes (single and two phases) and channel designs (single and dual pass) were also demonstrated and compared. An interesting concept called "thermal wake" was proposed in [206]. This notion stems from the fact that the temperature of downstream cells is affected by that of the upstream cells due to coolant heating effect. Numerical simulations are needed beforehand to extract the coefficients of the thermal wake function. The thermal wake effect can then be represented by the thermal resistance between coolant cells. A complete thermal resistive network can then be built and solved by matrix solvers. This method may not be practical since once the flow rate changes, new thermal wake coefficients need to be generated by long numerical simulations.

Recently, a thermal model called 3D-ICE [207] was proposed which can model both heatsink-based air cooling and fluidic cooling for 3DICs. It adopts the classical thermal resistive method which divides the 3D stack into regular grid cells including the coolants. The thermal resistances between solid cells and coolant cells are calculated as $r = 1/(hA)$, where the heat transfer coefficient $h = k \cdot Nu/D_h$ and Nu is determined by an empirical formula provided in [211]. The coolant cells are inter-related by a "voltage controlled current source" formulation to count for the rise of coolant temperature. The produced resistive thermal network is then solved by SuperLU. The simulation result was claimed to be accurate according to the verification against commercial tools, but the granularity of grid cell partitioning is too fine to be efficient for reasonably simple circuits and the problem explodes

in size for large circuits. This problem was later addressed in [208] by model reduction using 2-resistor-model. Significant reduction in CPU time was reported. In addition, simulation acceleration by mesh number reduction was studied in both [209] and [210]. Xie and Swaminathan [209] adopted the non-conformal domain decomposition method, which performed grid division at different granularities in the respective homogeneous domains. The continuity across domains was assured by the coupling matrices. Fourmigue et al. [210] adopted a non-uniform grid division approach. It enlarges the grid cells as long as the homogeneity is conserved. The grid will then be split if it is not aligned with the neighbors to ensure a consistent interface. The splitting operator technique is then applied to calculate the temperature change in three stages, which provides an ideal parallelism for multi-thread accelerations.

Despite the good modeling frameworks, most of the abovementioned thermal models fail to address the accuracy of the fluid coolant modeling, which is an essential prerequisite to achieve reliable simulation results. Two weaknesses have been identified. One is the heat transfer coefficient of the microchannel, which were calculated based on empirical formula targeted for macro-scale ducts [211]. A survey done by Sobhan and Garimella [228] on microchannel thermodynamics shows different thermal behaviors of microfluids from fluids in macro-scale. Therefore, the wrong formulation taken by [207–210] may cause non-trivial errors. The other inappropriate assumption is the microfluid is fully developed but in fact they are not. As the coolant enters the microchannel, the entrance effect caused by the developing thermal and velocity boundary layers will provide significantly higher heat transfer coefficient so that the circuit temperatures are low and close to the input coolant temperature at the inlet region. In this sense, most of the previously proposed thermal models have underestimated the thermal gradient of microfluid-cooled 3DICs³. Therefore, it is imperative to have accurate microfluidic thermal models based on correct microfluidic thermodynamics.

³Supportive evidences can be found in the experimental results of Chapter 3

2.5 Summary

In this chapter, we have reviewed the architecture and fabrication of 3DIC as well as the thermal management and modeling techniques. In the early days, recrystallization-based sequential circuit fabrication was the pioneering 3DIC technology but soon this was replaced by more feasible wafer stacking approaches. By thinning and bonding pre-fabricated wafers or chips, and connecting different circuit layers with fast vertical links such as TSVs, the fabrication process is greatly simplified with increased yield. After decades of fundamental research, the manufacturability of 3DIC is less of an issue. The research focus has been shifted towards the design and design automation in order to fully exploit the benefits of 3D integration.

One of the teething design problems of stacking high performance circuits is the heat dissipation due to high power density and large thermal resistance. Originating from 2D thermal management strategies, techniques such as MSV, DVFS, power gating and clock gating were proposed to reduce the power dissipation. On the other hand, power redistribution techniques such as thermal-aware floorplanning and placement at design time and thermal-aware task allocation and scheduling at runtime were also proven to be effective in solving the heat dissipation problem in 3DICs. However, once the power level rises beyond the cooling capability of existing cooling techniques, both static and runtime power manipulation methods are insufficient and better cooling solutions are needed. Adopting new cooling method is usually costly, thus thermal-TSVs were proposed to augment existing cooling methods by improving the local thermal conductivity. For very high-speed and power hungry 3DICs, the effectiveness of passive cooling method is still limited. Microfluidic cooling method emerged as a promising active cooling solution due to its high thermal conductivity and scalability. Various channel design and runtime control schemes were proposed to optimize its cooling result while minimizing its overhead.

Except for cases where circuit temperature is purely monitored by build-in thermal sensors, a fast and accurate thermal model is almost always needed. For heatsink-based air-cooled 3DICs, mature thermal models are available except that special attention on the thermal effects of TSVs needs to be paid. Thermal modeling

of microfluid-cooled 3DICs, however, is a new problem. A few FDM-based thermal models have been proposed but most of them have either omitted or misinterpreted the thermodynamic behavior unique to microchannels. Since an accurate thermal model is a prerequisite for effective thermal management, in the following chapter, an accurate thermal model that considers both the anisotropic TSV thermal effect and microchannel entrance effect will be introduced. It is used for the simulation of the research works presented in the remaining chapters.

Chapter 3

Compact Thermal Modeling of 3D Integrated Circuits

3.1 Introduction

Leveraging on the economies of scale of semiconductor processes, the emerging technology of 3D Integrated Circuit (3DIC) has provided a gamut of promising solutions for higher density large-scale-integration (LSI) [212]. By expanding the design space into the third dimension, 3D integration significantly reduces interconnect latency, power consumption and footprint compared with the 2D equivalent circuits [213–215]. However, stacking of multiple circuit layers simultaneously increases the power density and junction-to-ambient thermal resistance, which makes effective cooling even more challenging.

Proper thermal management is hence mandatory to avoid thermal-induced aging, performance degradation and even malfunctioning. Various solutions have been proposed in the literature, including the relocation and redistribution of heat sources and heat paths through thermal-aware floorplanning and scheduling [216, 217], exploitation of novel active and passive cooling methods such as microfluidic cooling [179] and thermal through silicon vias (TSVs) [169, 170, 175]. For any of these approaches or their hybrid combinations [187] to be effective, an accurate and computationally efficient thermal model is almost always needed to adapt and control the circuit temperature under a given power profile and the cooling requirements.

Numerical multiphysics simulators such as ANSYS [218] and COMSOL [219] are general and accurate but their long simulation cycle makes design phase and runtime thermal management inefficient. As a trade-off, several analytical models were proposed to provide rough estimations to speed up the simulation. Early tools such as the well-accepted Hotspot [222] simulates both transient and steady state temperature of air-cooled 2D/3D circuits. Results from Hotspot were shown to be accurate for 2D circuits. However, its modeling of 3D architecture was too simplistic. As microfluidic cooling was proposed as a scalable cooling solution with better cooling effect than heatsink-fan-based air cooling for high performance 3D systems, various thermal models [205–208] were subsequently proposed for liquid-cooled structure, among which the latest released model is 3D-ICE [207, 208]. As 3D-ICE ignores the entrance effect [208, 224] and adopts the conventional parametric correlation for macro-scale pipes [211], it tends to underestimate the liquid-cooling-induced thermal gradient. Owing to its irregularity and profiling difficulty, thermal effects of TSVs in 3DIC have not been properly investigated in previous thermal models. In practical implementation, TSVs are likely to concentrate in small regions [226] as I/O interfaces. A detailed analysis on the thermal influence of TSV’s location and size is thus necessary for an accurate plot of the thermal map of 3D systems.

In this chapter, a fast and accurate steady state thermal simulator for TSV-based 3DIC will be presented. The proposed simulator models both heatsink-based air cooling and microchannel-based liquid cooling as these are two major cooling methods found in literature. For both cooling settings, the target system is divided into small control volumes called grid cells. Thermal conductance matrix is then built and solved by sparse matrix solver KLU [230]. For air-cooled systems, our model shows significant speedup compared with steady state function of Hotspot. While for microfluidic-cooled systems, our model reports more accurate simulation than 3D-ICE due to the adoption of microchannel-specific correlations and consideration of entrance effect. Furthermore, this model analyzes the thermal effect of TSVs by their exact positions and sizes. It delineates the vertical heat spreading phenomenon of TSVs in greater detail to avoid the incorrect prediction of maximum and minimum temperatures.

This chapter is organized as follows. Section 3.2 reviews the thermodynamic theories used in the thermal model. Section 3.3 describes typical architectures of 3DICs with either heatsink-based air cooling or microfluidic cooling. In Section 3.4, our proposed thermal model will be presented in detail, with emphasis on the anisotropic thermal effect of TSV and entrance effect of microchannels. The simulation results are compared and verified in Section 3.5. Section 3.6 provides a brief summary of the chapter.

3.2 Preliminaries on Thermodynamic Theories

In ordinary electronic circuits, heat transfer occurs mainly in the form of conduction and convection. Irrespective of the architectural differences of different cooling methods, the heat transfer pattern is at large similar, where the heat generated by the circuits conducts through the material and is finally removed by convection. In this section, the theories of thermodynamics used in the development of our thermal model will be briefly reviewed.

3.2.1 Thermal Conduction in Solids

Figure 3.1: Thermal Conduction between Solid Cells

Heat conduction in solids obeys Fourier's law [223], which describes heat flux as follows.

$$\vec{q} = -k\Delta T \quad (3.1)$$

where $\vec{q}$, k and ΔT denote the heat flux density, thermal conductivity and temperature gradient, respectively. Under the finite difference approximation, the solid

is divided into small control volumes called grid cells, as shown in Fig. 3.1. A one-dimensional integration of (3.1) yields

$$q_{ij} = g_{ij}(T_i - T_j) \quad (3.2)$$

where the thermal conductance g_{ij} between two solid grid cells i and j is defined as

$$g_{ij} = kA/L \quad (3.3)$$

where A is the cross-sectional area where the heat flux passes through and L is the length of heat transfer path. An analogy between the thermal and electrical circuit networks is therefore established with temperature, heat flux and thermal conductance corresponding to potential, current and electrical conductance, respectively.

3.2.2 Thermal Convection through Heatsink

Convective heat transfer between heatsink and air was well studied in the literature [220, 221]. A sophisticated modeling of either natural or forced convection between heatsink and air involves the size and shape of heatsink, as well as the properties of air flow. To be concise, all influencing factors are usually consolidated into a convective resistance [222] formulated as follows:

$$R_{conv} = 1/(h * A_{conv}) \quad (3.4)$$

where h is the average heat transfer coefficient and A_{conv} is the cross-sectional area of interface where the convection takes place. In our thermal model, the convective resistances of heatsinks are considered as one of the configuration inputs in accordance to Hotspot [222]. Tabulated values of convective resistances for heatsink can be found from the manufacturer datasheet. Using the thermal-electric analogy, the ambient node can be integrated into the chip's thermal network through convective resistance. The temperature at each node can then be obtained by solving the thermal circuit problem.

3.2.3 Convective Heat Transfer of Microchannels

The concept of convective resistance and (3.4) are also applicable to quantify the heat transfer between channel walls and coolant for liquid cooling with microchan-

nels. Unlike heatsink, pre-calculated convective resistances are usually not available. Therefore, sophisticated models for obtaining R_{conv} or more specifically, the heat transfer coefficient h , are needed. The heat transfer problem has been well-studied for ducts and pipes in macro-scale [211, 224, 225]. Later research on microchannel thermodynamics [228] shows that microfluids may exhibit significant difference in thermal behavior compared with fluids in macro-scale. In that sense, the adoption of conventional macroscale thermal formulation by existing thermal models of microfluidic-cooled 3DIC [205, 207, 208] may introduce non-trivial errors.

The research on microfluidic cooling has been active since the first proposal by Tuckerman and Pease [179] on embedding microchannels at the back of silicon substrate for effective cooling of high performance circuits. Experiment setups in the literature were diverse in microchannels shapes and coolants. In our study, the microchannels are set to be in rectangular shape and water is used as the coolant due to its high heat capacity and wide availability. Reynolds number Re and Prandtl number Pr are two frequently used dimensionless quantities that help describe the dynamic properties of fluids and are defined as follows:

$$Re = \frac{\rho \cdot Q \cdot D_h}{\eta \cdot A_{ch}} \quad (3.5)$$

$$Pr = \frac{c_f \cdot \eta}{k_f} \quad (3.6)$$

where Reynolds number Re describes the ratio of inertia force versus viscous force, which depends on both the fluid property and geometric parameters of the conduit; Prandtl number Pr denotes the ratio of momentum diffusivity to thermal diffusivity which only depends on the intrinsic properties of the fluid; ρ , Q , η , c_f and k_f denote the fluid's density, flow rate, dynamic viscosity, specific heat and thermal conductivity, respectively; A_{ch} denotes the crosssectional area of the conduit, D_h denotes the conduit's hydraulic diameter, and is given by (3.7) for rectangular channels, where P_{ch} denotes the perimeter of the microchannel.

$$D_h = \frac{4 \cdot A_{ch}}{P_{ch}} \quad (3.7)$$

At normal flow rate, water has a Reynold number $Re \ll 2100$, which implies a laminar flow in the microchannel [223]. Therefore, the problem is cast as the

determination of heat transfer coefficient of laminar water flow in rectangular microchannels. It is worth noting that such specification does not limit our modeling method from applying to other 3D systems with different cooling setups but serves only to ease the initial development of the thermal model.

For a rectangular microchannel, the heat transfer coefficient h can be formulated as

$$h = \frac{k_f Nu}{D_h} \quad (3.8)$$

where Nusselt number Nu denotes the ratio of heat transfer contributed by convection and conduction, and is usually obtained by empirical correlations with Re , Pr and channel dimension in the literature. According to the criteria of accuracy and similarity of cooling setup, the correlation proposed by Peng and Peterson [229] is referred and adapted by model fitting against COMSOL in our thermal model.

$$Nu_{PP} = 3.67 \cdot \left(\frac{W_{ch}}{W_{ch} + H_{ch}} \right)^{3.33} \left(\frac{W_{ch}}{H_{ch}} \right)^{0.67} Re^{0.62} Pr^{1/3} \quad (3.9)$$

where H_{ch} and W_{ch} denotes the height and width of the microchannels, respectively.

3.3 Target Architectures of 3DIC

In 3DIC, multiple layers of circuit are stacked with interlayer connections. A circuit layer typically comprises of Back End of Line (BEOL), device and thinned substrate sublayers. Depending on the cooling demand, different cooling methods are applied. The schematics in Fig. 3.2 show the cross-sectional views of 3DICs with heatsink cooling for low power circuits and microfluidic cooling for high performance 3D stacks. For heatsink-cooled 3D circuits, thermal interface material (TIM) is smeared in between circuit layers for better thermal connectivities. The generated heat is then dissipated into ambient through the spreader and heatsink usually with the aid of a fan. For microfluidic-cooled 3D stacks, microchannels are etched at the back of dies before bonding. Those microchannels are assumed to be uniformly distributed with identical channel width and height. The channel walls are thus of the same width except possibly for the walls at the boundaries. Liquid coolant flows through the microchannels and carries away the generated heat. Without loss of generality,

Figure 3.2: Schematic of 3D integrated circuit with Heatsink and microfluidic coolings

it is assumed that the coolant reservoir is large enough so that the inlet coolant is always at the ambient temperature, keeping in mind that the ambient temperature setting can always be modified to simulate for non-ideal cases where a higher inlet temperature is observed.

3.4 Compact Thermal Model of 3DIC

3.4.1 Grid Division of 3DIC

To achieve a balance between accuracy and speed on estimating the steady state temperature of 3DICs, finite difference approximation is adopted, where the 3D stack under study is divided into regular grid cells according to the user-defined row-column resolution.

For the heatsink-cooled 3DIC, the vertical resolution is by default down to the sublayers, i.e, the BEOL, device, substrate and TIM layers. The heat spreader and heatsink are specially treated due to their larger sizes than the die part. Fig. 3.3 and 3.4 shows the grid division for heat spreader and heatsink as adopted from Hotspot [222], where the innermost parts follow the grid division of the circuit layers and twelve extra nodes are created on the periphery.

For microfluidic-cooled 3DIC, the default vertical resolution for circuit layer part is lowered to reduce problem size by treating the BEOL, device and substrate layers as a whole. The microchannel layer are naturally separated by the microchannels and

Figure 3.3: Grid division of heat spreader showing thermal connections between the trapezoidal extra nodes and the inner cells.

Figure 3.4: Grid division of heatsink showing thermal connections among the trapezoidal extra nodes and the inner cells. The grid dimension of circuit layer is assumed to be 3×3 for this illustration.

walls, and are further divided by following the lateral grid division of circuit layers. Fig. 3.5 illustrates the grid division of a 2-layer 3DIC where the grid dimension for circuit layer is 3×3 . The thermal connections within a grid pile are shown in Fig. 3.6, where the microchannel absorbs the heat generated from both the circuit and channel cells.

Figure 3.5: Grid division of 3DIC with microchannel-based liquid cooling. A 2-layered structure is assumed for this illustration.

Figure 3.6: Thermal connections within a grid pile.

3.4.2 Thermal Conductance between Solid Cells considering Anisotropic TSV Effects

In a 3D circuit, thousands of TSVs may co-exist to deliver the power and signal. These TSVs are normally filled with metal which has a higher thermal conductivity

than silicon. Previous works on thermal-TSVs [169, 170, 175] suggest that they can have a significant influence on the steady state temperature of 3DICs. Such influence is due to the change of thermal conductivity in the heat transfer path when a TSV is inserted and is highly dependent on the TSV's size and position. In the following discussion, all TSVs are assumed to have squarish cross-section to simplify the model, with the understanding that round TSVs can always be projected to the squarish ones with appropriate sizes.

For two solid grid cells without TSV, the thermal conductance between them is formulated as in (3.3). Approximations are made for the trapezoidal extra nodes of heat spreader and heatsink as in Hotspot [222] due to their different shapes. For heat spreader, the thermal resistances from the centers of the extra nodes to their inner edges are calculated as follows.

$$R_{sp,x} = \frac{W_{sp} - W_{chip}}{k_{sp} \cdot (H_{sp} + 3H_{chip}) \cdot t_{sp}}, \quad R_{sp,y} = \frac{H_{sp} - H_{chip}}{k_{sp} \cdot (W_{sp} + 3W_{chip}) \cdot t_{sp}} \quad (3.10)$$

where k_{sp} and t_{sp} denote the thermal conductivity and thickness of the spreader, respectively. Similarly, for heatsink, the thermal resistances from the centers of the extra nodes to their edges are approximated as follows.

$$\begin{aligned} R_{hs,ci,x} &= \frac{W_{sp} - W_{chip}}{k_{hs} \cdot (H_{sp} + 3H_{chip}) \cdot t_{hs}}, & R_{hs,ci,y} &= \frac{H_{sp} - H_{chip}}{k_{hs} \cdot (W_{sp} + 3W_{chip}) \cdot t_{hs}} \\ R_{hs,co,x} &= \frac{W_{sp} - W_{chip}}{k_{hs} \cdot (3H_{sp} + H_{chip}) \cdot t_{hs}}, & R_{hs,co,y} &= \frac{H_{sp} - H_{chip}}{k_{hs} \cdot (3W_{sp} + W_{chip}) \cdot t_{hs}} \\ R_{hs,outer,x} &= \frac{W_{hs} - W_{sp}}{k_{hs} \cdot (H_{hs} + 3H_{sp}) \cdot t_{hs}}, & R_{hs,outer,y} &= \frac{H_{hs} - H_{sp}}{k_{hs} \cdot (W_{hs} + 3W_{sp}) \cdot t_{hs}} \end{aligned} \quad (3.11)$$

where k_{hs} and t_{hs} denote the thermal conductivity and thickness of the heatsink, respectively. When a TSV is inserted as shown in Fig. 3.7(a), the equivalent x , y , and z direction thermal conductances of the grid cell can be computed as:

$$\begin{aligned} g_x &= \left(1 + \frac{wh(\beta-1)}{WH(1+(1-w/W)(\beta-1))}\right)g_{x0} \\ g_y &= \left(1 + \frac{wh(\beta-1)}{WH(1+(1-h/H)(\beta-1))}\right)g_{y0} \\ g_z &= \left(1 + \frac{wh(\beta-1)}{WH}\right)g_{z0} \end{aligned} \quad (3.12)$$

where $g_{x0} = \frac{K_{cell}HL}{W}$, $g_{y0} = \frac{K_{cell}WL}{H}$ and $g_{z0} = \frac{K_{cell}WH}{L}$ are the thermal conductances of the grid cell without TSV and L is the depth of the grid cell. $\beta = K_{TSV}/K_{cell}$ is the ratio of the thermal conductivities between TSV and grid cell. Rectangular

TSV is considered in general since a TSV may fall on the boundaries of grid cells such that only a part of it is inside the grid cell. This set of equations reveals that thermal conduction improvement due to TSV insertion is anisotropic, which is more significant vertically than laterally. Note that superposition principle is applicable if multiple TSVs are inserted into the same grid cell.

Figure 3.7: Geometry of grid cell with TSV

In real implementation, TSVs are normally surrounded by a thin layer of less thermally conductive dielectrical liner for insulation purpose, as shown in Fig. 3.7(b). To account for the thermal isolation of dielectrics, the equivalent thermal conductivity of TSV with liner is defined as follows.

$$\begin{aligned} K_{TSV,xy} &= \left(1 + \frac{d_m^2(\gamma-1)}{(d_m+2d_l)(d_m+2d_l\gamma)}\right) K_{liner} \\ K_{TSV,z} &= \left(1 + \frac{d_m^2(\gamma-1)}{(d_m+2d_l)^2}\right) K_{liner} \end{aligned} \quad (3.13)$$

where d_m is the side length of the metal core, d_l is the liner thickness and $\gamma = K_{metal}/K_{liner}$ is the ratio of the thermal conductivity between metal and liner.

To calculate the thermal conductances between all the solid grid cells with TSV effects, thermal conductances without TSV are first obtained according to (3.3). The sizes and positions of TSVs are then imported to determine their overlaps with all grid cells. Based on the equivalent thermal conductivity of every TSV, their contributions to thermal conductance are then added up to the overlapped solid grid cells according to (3.12). The thermal conductance between two adjacent solid grid cells can then be obtained as half the sum of their thermal conductances.

3.4.3 Convective Heat Transfer and Entrance Effect of Microchannels

For heatsink-based air cooling, convection exists between the heatsink and air flow. The tabulated thermal resistance value on the manufacturer sheet is meant for the entire heatsink. It needs to be distributed among the heatsink cells according to their respective area percentages as follows.

$$g_{hs_i,amb} = \frac{A_{hs_i}}{A_{hs}} \cdot g_{hs,amb} \quad (3.14)$$

where $g_{hs_i,amb}$ represents the thermal conductance between the i^{th} heatsink cell and the ambient; A_{hs_i} denotes the area of the i^{th} heatsink cell; A_{hs} denotes the area of the entire heatsink; and $g_{hs,amb}$ represents the thermal conductance between the entire heatsink and ambient.

It was noted that the heat transfer rate is significantly higher near the inlet of microchannels due to the developing thermal boundary layer [208, 224], which was ignored by most of the previous works. From (3.4) and (3.8), it is clear that to obtain accurate convective conductance value at the entrance region, the critical step is to evaluate the Nusselt number Nu correctly, since other parameters are all well defined. For macro-scale calculation of Nu considering the entrance effect, the following correlation between Nu and Reynolds number Re and Prandtl number Pr has been proposed by Seider and Tate [224] in the 30s.

$$Nu_{ST} = Nu_0 (Re \cdot Pr)^{1/3} \left(\frac{D_h}{l} \right)^\eta \quad (3.15)$$

where l denotes the distance away from the channel inlet, D_h denotes the hydraulic diameter of the conduit. Nu_0 and η are two experimentally determined coefficients, with values 1.86 and 1/3 in the original proposal. To consider the entrance effect in our thermal model, an aggregated Nusselt number is proposed as follows by combining (3.9) and (3.15).

$$Nu = Nu_{PP} + Nu_{ST} \quad (3.16)$$

However, to fit the old entrance region correlation into the new micro-scale problem, the values of coefficients Nu_0 and η have been revised to 30 and 1.67 respectively

by model-fitting against COMSOL [219]. A plot of Nusselt number is shown in Fig. 3.8. The microfluid exhibits high heat transfer coefficient initially at the entrance region and decays asymptotically to a constant as the thermal boundary profile is fully developed.

Figure 3.8: Nusselt number at different distances along the flow direction

3.4.4 Steady State Thermal Analysis

At steady state, all the grid cells are under thermal equilibrium and the following thermal balance equation holds for every grid cell.

$$\sum_{j \in \text{grid cells}, j \neq i} (T_i - T_{i,j}) * g_{i,j} = P_i \quad (3.17)$$

where T_i and $T_{i,adj}$ denote the temperatures of the i^{th} grid cell and its adjacent grid cell, respectively. $g_{i,adj}$ represents the thermal conductance between the two cells, and P_i is the heat generated by the i^{th} grid cell. The matrix equation of thermal equilibrium can be solved by the sparse solver KLU [230] with the provided power consumption.

For microfluidic cooling, the channel temperature is also held constant where the heat is transferred from upstream to downstream in the form of massive flow. Thus the following equation is satisfied by the liquid channel cells under thermal

equilibrium.

$$\sum_{j \in \{walls\}} g_{ch,i,j}(T_{ch,i} - T_j) = (T_{ch,i} - T_{ch,i-1})QC_V \quad (3.18)$$

where $T_{ch,i}$ and $T_{ch,i-1}$ indicate the temperatures of the i^{th} channel cell and its upstream cell. Q and C_V are the coolant's flow rate and volumetric heat capacity, respectively.

Based on (3.17) and (3.18), the following matrix equation can be set up irrespective of the cooling technique.

$$GT = P \quad (3.19)$$

where T is a column array for the unknown steady state temperature of each grid cell. P is the heat generation column array, where the entries are equal to the power consumption if the node is not directly connected to ambient. Otherwise, for heatsink cells of air cooling,

$$P_{hs_i} = g_{hs_i,amb}T_{amb} \quad (3.20)$$

and for the inlet channel cells of microchannel-based liquid cooling,

$$P_{inlet_i} = Q_iC_VT_{amb} \quad (3.21)$$

where T_{amb} is the ambient temperature. G is an $N \times N$ coefficient matrix for N total number of grid cells. Note that the thermal conductance is zero for two cells that are not in contact physically. Thus G is a sparse matrix with the number of non-zero elements linearly proportional to N . This sparse matrix equation can be efficiently solved by existing sparse solvers such as KLU [230] to obtain the steady state temperature of every grid cell.

3.4.5 Software Implementation

The proposed thermal model has been implemented in C program. The basic program flow is shown in Fig. 3.9. The thermal model accepts a layer configuration file to set up the physical description of the 3DIC, including detailed sizing, cooling method, TSV setting and power consumption. After that, the model calculates the thermal conductances by considering the TSV effects. The G matrix is then built

Figure 3.9: Program flow chart

by completing the column pointer array, row index array and numerical value array required by KLU. Finally, the P array is created and the steady state temperature is obtained by solving (3.19). In our program, the time spent on thermal conductance calculation is linearly dependent on the number of TSVs, which is an integer constant of limited value for a given circuit. The construction of G and P both have a time complexity of $O(m)$, where m denotes the number of non-zero entries of the matrix. The KLU algorithm has a time complexity of $O(N + m)$, where N is the total number of grid cells. Since m is linearly proportional to N , our proposed thermal model has a time complexity of $O(N)$, which scales linearly with the total number of grid cells. Our proposed thermal model are available for downloading via [231].

3.5 Experiment Results

In this section, we verify the accuracy of the proposed thermal simulator and compare our results with existing works. For both cooling methods, a two-layer 3D stack is used for the test. The size of the chip is set to $1.8 \times 1.8 \text{ mm}^2$ with $12 \mu\text{m}$ BEOL layers and $2 \mu\text{m}$ device layers. The thickness of the bulk layer is $48 \mu\text{m}$ for the top tier and $148 \mu\text{m}$ for the bottom tier. Ambient temperature is assumed to

be $300K$. The default setup of the cooling package is as listed in Table 3.1, unless otherwise specified. In the absence of physical prototypes, our thermal simulator is

Table 3.1: Experiment settings

Heatsink Cooling Setting	
TIM thickness	$20 \mu m$
TIM thermal conductivity	$4 W/(mK)$
Spreader thermal conductivity	$400 W/(mK)$
Heatsink thermal conductivity	$400 W/(mK)$
Heatsink Convective conductivity	$10 W/K$
Microfluidic Cooling Setting	
Microchannel Size (μm)	100×100
Channel Wall Width	$100 \mu m$
Default Flow Rate	$5.4 - 9 ml/min$
Coolant Type	Deionized Water

validated against the accurate commercial multiphysics solver COMSOL [219]. A simple 3×3 mosaic floorplan is adopted for the two circuit layers with customizable power densities. All the experiments are performed on a Linux server with Intel Xeon 3.47 GHz processor and 50 GB of memory.

3.5.1 Heatsink-based Air-Cooled 3DIC

For heatsink-cooled 3DIC, we compare our results with COMSOL and Hotspot 5.0. A set of uniform power density at typical magnitude is assigned to the device layers of 3DIC. The maximum and minimum steady state temperatures from the three simulators are reported in Table 3.2, where power density is measured in W/cm^2 .

In this experiment, a 16×16 grid division on each layer is used for Hotspot and our simulator. The mesh in COMSOL is set at the finest granularity to produce the most accurate result. Both our model and Hotspot adopted the same equation (3.3) to calculate the thermal conductance between solid cells, and resulted in the same

Table 3.2: Steady State Temperature (K) Comparison of COMSOL, Hotspot and Proposed Simulator

Power Density / Temperature		COMSOL	Hotspot	Ours	Error
50 W/cm^2	Max	313.72	312.83	312.83	0.3%
	Min	308.16	307.41	307.40	0.2%
100 W/cm^2	Max	327.45	325.66	325.66	0.5%
	Min	316.33	314.81	314.81	0.5%
150 W/cm^2	Max	341.17	338.49	338.49	0.8%
	Min	324.49	322.22	322.22	0.7%
200 W/cm^2	Max	354.90	351.32	351.32	1.0%
	Min	332.66	329.62	329.62	0.9%
250 W/cm^2	Max	368.62	364.15	364.15	1.2%
	Min	340.82	337.03	337.02	1.1%

thermal circuit for heatsink-cooled 3DICs. Thus, it is not surprising that our thermal simulator and Hotspot produce almost identical results although different solvers are adopted. Compared with COMSOL, the deviation of our thermal simulator is in the range of 0.2 – 1.2% due to coarser vertical granularity which was set as the total number of sublayers. Finer vertical granularity is possible by further dividing the sublayers. However, the tradeoff between accuracy and runtime shall be considered to determine a suitable grid resolution. The deviation of temperatures increases monotonically with the increase of heat flux, because increasing the heat flux amplifies the differences in the thermal circuits.

Table 3.3 lists the runtime of the three simulators in *seconds* for this experiment set. Only the speedup against Hotspot is reported in this table because it is unfair to compare the speedup against COMSOL, which has a larger problem size due to much finer granularity. Up to $21\times$ speedup is achieved for this test case. Furthermore, Fig. 3.10 depicts the runtime scaling of both simulators at a moderate power density of $150W/cm^2$. The five problem sizes correspond to 8×8 , 16×16 , 32×32 , 64×64 and

Table 3.3: Runtime Comparison of COMSOL, Hotspot and Proposed Simulator for Air Cooling Simulation

Model	Power Density				
	50 W/cm^2	100 W/cm^2	150 W/cm^2	200 W/cm^2	250 W/cm^2
COMSOL	42 s	42 s	42 s	42 s	42 s
Hotspot	0.763 s	1.019 s	1.187 s	1.31 s	1.41 s
Ours	0.068 s	0.068 s	0.059 s	0.061 s	0.066 s
Speedup	11x	15x	20x	21x	21x

128×128 lateral grid dimensions respectively, with vertical grid resolution equal to the total number of sublayers which is 16.

Figure 3.10: Runtime Scaling of Our Thermal Simulator and Hotspot

It can be seen that hotspot performs better when the grid resolution rose beyond 64×64 for this particular case. This is due to the different time complexity of the algorithms adopted by Hotspot and our model. In Hotspot, a recursive multigrid approach was used to calculate the steady state temperature, where coarse level temperature was first estimated and then interpolated to finer granularity with iterative refinement for each granularity transition. For example, for the 8×8

case, Hotspot first estimates the steady state temperature at the coarsest level, i.e. 1×1 , based on the lumped power consumption and thermal resistances. The obtained temperature values are then interpolated bilinearly to the resolution of 2×2 , and iteratively refined based on the power consumption information at 2×2 resolution to reduce the estimation error. This interpolation and refinement process will repeat until the finest resolution of 8×8 is reached. Therefore, the algorithm adopted by Hotspot has a time complexity of $O(C(p) \log_2 n)$, where n denotes the problem size, $C(p)$ denotes a power related parameter due to the iteration needed by the refinement at each stage. From Table 3.3 it is clear that larger power density will induce a large $C(p)$. In contrast, our model adopts a direct sparse matrix solver to calculate the steady state temperature, which has a time complexity of $O(n)$. Thus at small problem size, the dominating $C(p)$ causes Hotspot to have a longer runtime. However, as the problem size increases, the effect of problem size scaling dominates as shown in Fig. 3.10. Moreover, the grid dimension in Hotspot is restricted to powers of two for the convenience of its recursive solving method, while our model has no restriction on the grid dimension which is important for modeling microfluid-cooled 3DICs.

Figure 3.11: The three TSV test patterns

We then examine the accuracy of TSV modeling by inserting three different patterns (Fig. 3.11) of TSVs into the circuit. For convenience, all TSVs in COMSOL are assumed to have identical size, and punch through the substrate of the top tier to the device layer of the bottom tier. Si_3N_4 is assumed to be the liner material and copper is used as the filling metal. Table 3.4 lists the reduction of maximum

Table 3.4: Reduction of maximum temperature (K) due to TSV insertion

Model	Power Density				
	50 W/cm^2	100 W/cm^2	150 W/cm^2	200 W/cm^2	250 W/cm^2
Pattern A					
COMSOL	0.20	0.40	0.60	0.80	1.00
Ours	0.31	0.63	0.94	1.26	1.57
Error	0.11	0.23	0.34	0.46	0.57
Pattern B					
COMSOL	0.14	0.29	0.44	0.59	0.73
Ours	0.18	0.36	0.53	0.71	0.89
Error	0.04	0.07	0.09	0.12	0.16
Pattern C					
COMSOL	0.12	0.24	0.36	0.48	0.60
Ours	0.16	0.31	0.47	0.63	0.78
Error	0.04	0.07	0.11	0.15	0.18

temperature compared with the baseline without TSVs. It shows the estimation error varies with different TSV patterns, where the largest deviation appears on Pattern *A*. Consider the base temperature of 300K, our TSV thermal model is accurate with the largest deviation of only 0.57K.

3.5.2 3DICs with Microfluidic Cooling by Microchannels

To verify the accuracy of the proposed thermal model on microfluidic cooling, the simulation results are first compared with COMSOL and 3D-ICE v2.1. The same 2-layer 3D circuit is used except that it is now cooled by the microfluidic channels between two circuit layers instead of heatsink. The same set of typical power densities is applied to the device layers with per channel coolant flow rate varying from 0.6 to 1ml/min. A lateral grid dimension of 9×9 is adopted for our thermal model. The default microchannel settings are first applied for the case study. The maximum and minimum temperatures for the top device layer are reported in Table 3.5 with their corresponding percentage errors enclosed in brackets. It shows that our proposed thermal model is very accurate with a percentage error of less than 3.8% for this test case. In contrast, 3D-ICE exhibits larger error especially for the minimum temperature due to the ignorance of entrance effect, which causes the thermal gradient to be severely under-estimated.

As shown in Fig. 3.12, a sample set of the top layer thermal profiles from our model and COMSOL are plotted and compared. Due to a uniform heat flux and uniform coolant distribution, the temperature varies only along the channel direction. Both graphs are hence divided into 9 rows and the average temperature of each row is compared in Table 3.6 in ascending order of temperature. A maximum average temperature deviation of 1.17% is observed. In contrast, the top layer thermal profile obtained from 3D-ICE, as shown in Fig. 3.13, presents a diagonal shift of temperature distribution and deviates significantly from both COMSOL's and ours with a much higher minimum temperature.

Table 3.7 lists the average runtime required for the steady state temperature calculation by the three simulators. Speedups against COMSOL are significant for both 3D-ICE and our simulator where ours has the shortest runtime. However,

Table 3.5: Comparison of maximum and minimum temperatures among the proposed thermal model, COMSOL and 3D-ICE, with the corresponding deviation (in %) from COMSOL shown in bracket

Model	250 W/cm^2	200 W/cm^2	150 W/cm^2	100 W/cm^2	50 W/cm^2
1.0 ml/min per channel					
COMSOL	341.2	332.9	324.6	316.3	308.1
	303.4	302.7	302.0	301.3	300.7
3D-ICE	387.0(13.4)	369.6(11.0)	352.2(8.5)	334.8(5.8)	317.4(3.0)
	371.0(22.3)	356.8(17.9)	342.6(13.4)	328.4(9.0)	314.2(4.5)
Ours	344.3(0.9)	335.4(0.8)	326.6(0.6)	317.7(0.4)	308.9(0.3)
	309.9(2.1)	307.9(1.7)	305.9(1.3)	304.0(0.9)	302.0(0.4)
0.8 ml/min per channel					
COMSOL	349.5	339.4	329.5	319.6	309.7
	304.0	303.2	302.4	301.6	300.8
3D-ICE	391.8(12.1)	373.4(10.0)	355.1(7.8)	336.7(5.4)	318.4(2.8)
	373.3(22.8)	358.7(18.3)	344.0(13.8)	329.3(9.2)	314.7(4.6)
Ours	352.2(0.8)	341.7(0.7)	331.3(0.5)	320.9(0.4)	310.4(0.2)
	312.2(2.7)	309.8(2.2)	307.3(1.6)	304.9(1.1)	302.4(0.5)
0.6 ml/min per channel					
COMSOL	361.4	348.9	336.4	324.2	312.2
	305.0	304.0	303.0	302.0	301.0
3D-ICE	399.8(10.6)	379.9(8.9)	359.9(7.0)	339.9(4.8)	320.0(2.5)
	377.3(23.7)	361.8(19.0)	346.4(14.3)	330.9(9.6)	315.5(4.8)
Ours	364.7(0.9)	351.8(0.8)	338.8(0.7)	325.9(0.5)	312.9(0.2)
	316.5(3.8)	313.2(3.0)	309.9(2.3)	306.6(1.5)	303.3(0.8)

considering the problem size of 1197 for 3D-ICE and 405 for our simulator, the computational efficiencies of the two simulators are similar.

Figure 3.12: Comparison of top layer temperature profile against COMSOL at a heat flux of $150W/cm^2$ and $1ml/min$ flow rate per channel, coolant flows from north (upper) to south (lower) direction.

Table 3.6: Comparison of average temperature of each row in Fig. 3.12

Row Number								
1	2	3	4	5	6	7	8	9
COMSOL (K)								
304.4	308.5	311.2	313.8	316.3	318.7	320.9	322.8	324.0
Ours (K)								
305.9	310.0	313.9	317.2	320.0	322.4	324.3	325.7	326.6
Error (%)								
0.49	0.49	0.87	1.08	1.17	1.16	1.06	0.90	0.80

Table 3.7: Average Runtime Comparison of COMSOL, 3D-ICE and Proposed Thermal Simulator for Microfluid Cooling Simulation

Simulator	COMSOL	3D-ICE	Ours
Runtime	150 s	0.016 s	0.006 s

Figure 3.13: Top layer temperature map from 3D-ICE at a heat flux of $150 W/cm^2$ and $1 ml/min$ flow rate per channel, coolant flows from north (upper) to south (lower) direction.

In microfluid-cooled 3DICs, the TSVs are normally inserted through the channel walls to avoid the microchannels. They do not provide fast heat conduction paths to the microfluids but only increase the lateral thermal conductivity of solid grid cells by a limited amount. As such, the thermal effect of TSV is less pronounced compared with that in heatsink-cooled 3DICs. For example, a reduction in maximum temperature of merely $0.28 K$ is reported by our thermal model for the test bench at a power density of $250 W/cm^2$ and flow rate of $1 ml/min$ per channel with TSV pattern C .

3.5.3 Parametric Study of Microfluidic Cooling for 3DIC

In this section, we investigate the performance of microfluidic cooling and the accuracy of the proposed thermal model under different cooling settings. The 2-layer 3D circuit is again used as the cooling object at a fixed medium power density of $150 W/cm^2$. The accuracy of our model with different flow rates has already been shown by Table 3.5. Therefore, the total flow rate will be fixed at $9 ml/min$ as before while the microchannel geometry settings including the microchannel width and pitch are varied. Due to similar temperature patterns produced as those in Fig. 3.12, only the maximum and minimum temperatures obtained from the proposed model and COMSOL will be recorded for comparison. Note that 3D-ICE is not

included in this round of study due to its less accurate simulation results proven earlier.

Fig. 3.14 presents the results obtained from both the proposed model and COMSOL when the channel width changes from $40\ \mu m$ to $160\ \mu m$ while the channel pitch is fixed at $200\ \mu m$. The grid resolution is set as 9×9 . It can be seen that

Figure 3.14: Comparison of circuit temperature against COMSOL with varying microchannel width at a fixed microchannel pitch of $200\ \mu m$

at fixed pitch (and hence the total number of microchannels is fixed), the peak circuit temperature rises significantly when the microchannel width decreases below the magnitude of channel height, due to the reduced heat transfer coefficient and contact area. Our model has predicted the maximum circuit temperature well with a maximum percentage error of 2.5% and average error of 0.1%. However, similar to the results presented in Table 3.5, the prediction of minimum temperature shows a larger percentage error of 1.4% due to the limited resolution.

Fig. 3.15 depicts the trend of the circuit temperature when the microchannel pitch varies from $100\ \mu m$ to $200\ \mu m$. It can be seen that although the flow rate per channel (hence the heat transfer coefficient) decreases with increasing number of microchannels, it is compensated by the increasing contact area between the circuit and the coolant, leading to slight reduction in the peak circuit temperature. Our model shows very accurate simulation for this case with an average percentage error of 0.2%. Likewise, the reported minimum circuit temperature shows an average

Figure 3.15: Comparison of circuit temperature against COMSOL with varying microchannel pitch at a fixed microchannel width of $60 \mu\text{m}$

percentage error of 1.7%. With the above results, it can be concluded that our proposed model remains accurate and is applicable to microfluidic cooling of 3DIC with different cooling settings.

3.5.4 Influence of Granularity on the Modeling Accuracy

There is always a trade-off between the simulation speed and accuracy. Finer granularity usually results in more accurate results at the expense of runtime. In the case of thermal simulation of 3DICs, such trade-off can be optimized by adopting higher resolution along the primary heat transfer direction, which is the direction towards heatsink for air cooling and the coolant flow direction for microfluidic cooling.

Take microfluidic cooling as an example, from the results presented above, it is noted that a larger deviation of minimum circuit temperature against COMSOL is always observed. Such discrepancy is mainly contributed by the finite and uniform grid division adopted by our thermal model. Note that the minimum temperature always appears at the inlet area where the heat transfer coefficient varies significantly along the flow direction. COMSOL adopts an adaptive mesh construction method so that very fine-grained grid division is performed at critical places such as the inlet area. Thus it is able to capture the very high heat exchange rate right after the inlet.

Our model adopts uniform grid division method for compact thermal modeling, which under-estimates the inlet area heat transfer at low grid resolution and reports the average temperature of a much larger grid cell as the minimum temperature. As shown in Fig. 3.16, the prediction of minimum temperature becomes more accurate when finer grid division along the flow direction, where COMSOL reports a minimum temperature of 305 K, is adopted. In contrast, grid dimension in other directions has

Figure 3.16: More accurate minimum temperature prediction with fine-grained grid division along the flow direction.

no significant influence on the simulation results when power distribution is uniform. For example, with a power density of 250 W/cm^2 and flow rate of 0.6 ml/min per channel, lowering the grid resolution across the flow direction from 9 to 1 results in a minimum temperature of 317.1 K, which increases the deviation by only 0.6 K. However, for 3DICs with non-uniform power density, the granularity setting should follow the power density resolution to have an accurate temperature plot of the chip.

3.6 Summary

In this chapter, a fast and accurate compact steady state thermal simulator has been presented for both heatsink-cooled and microfluidic-cooled 3DICs. This simulator is capable of estimating the thermal effect of TSVs in fine granularity. It also considers entrance effect of microchannels based on the most appropriate thermodynamics for

microfluidic cooling. This thermal simulator has been implemented in C. It adopts a finite difference approach for fast simulation with customizable accuracy. The time complexity of the proposed simulator has been proven to be $O(N)$, where N denotes the total number of grid cells. A 2-layer 3DIC with uniform power density has been used as the testbench to verify the accuracy of the proposed thermal model. The simulation results were shown to be in good concordance with those obtained from commercial software COMSOL, with a maximum error of 1.2% for heatsink cooling and a maximum error of 3.1% for microfluidic cooling. Such error can be reduced with proper granularity selections along the primary heat transfer paths. Comparing with Hotspot for heatsink cooling at moderate grid dimension, our thermal model generated identically accurate results in a much shorter runtime. Comparing with 3D-ICE for microfluidic cooling, our thermal model has a similar computational efficiency but is more accurate in the estimations of temperature and temperature distributions. The results obtained from our thermal model also conclude that the thermal effect due to TSV has been marginalized under microfluidic cooling. However, it is still possible to add special thermal TSVs to provide heat conduction paths directly from potential hotspots to the microfluidic coolant.

The developed thermal simulator has provided a feasible way to predict the thermal behavior of 3DICs under different cooling techniques. In the following chapters, thermal management schemes and efficient microchannel design techniques will be proposed and this simulator will be used to validate all the proposals. We will focus on 3DIC's thermal problems such as runtime workload fluctuation, non-uniform power distribution and large lateral thermal gradients. Solutions that exploit microchannel-based liquid cooling will be explored to effectively solve the problems at reasonably low overhead costs.

Chapter 4

Power Aware Thermal Management for Microfluidic Cooled 3D Multi-Core Systems

4.1 Introduction

The integrated circuit (IC) design industry has witnessed a paradigm shift of computing systems to multiprocessor system-on-chip (MPSoC) [232–234] because it offers the system-level improvement of performance and flexibility while keeping the implementation cost low. On the other hand, stemming from the highly parallel and heterogeneous nature of its underlying hardware architecture, the design complexity of MPSoC has also increased tremendously. This poses many challenges when mapping an application to the architecture, thereby motivates the development of an MPSoC design platform to facilitate the mapping by hiding the physical complexity and exposing the application parallelism. In order to manage the complexity, a design platform is needed to guarantee that an application can complete its tasks within strict timing deadlines without violating physical constraints such as bandwidth, power and temperature. With 3D integration techniques [142, 235, 236, 238] emerged as an attractive alternative to alleviate the performance bottleneck of MP-SoC in 2D integration, high-throughput applications can now be executed on MP-SoC with a broader design space for exploration. Nevertheless, the aforementioned

challenges for MPSoC still remain in 3D.

The utilization of the short vertical interconnection in 3D will certainly provide a promising solution to break down the memory wall observed in 2D for the traditional multi-core cache-processor system, which is a commonly used MPSoC architecture. However, due to the high integration density and long heat-transfer path, the increased circuit temperature of 3DIC causes more severe performance degradation and reliability problems than in 2D. This effect might offset the benefit gained from the 3D integration. Therefore, a 3D multi-core cache-processor system needs to have an effective thermal management to maintain the temperature at a normal level.

Previous dynamic thermal managements [158, 160, 239–245] in 2D MPSoC are mainly based on power reduction and redistribution techniques such as dynamic voltage-frequency scaling (DVFS) and task scheduling. According to (4.1), the power consumption of CMOS circuit is linearly proportional to frequency and quadratically increases with the supply voltage.

$$P_{CMOS} = \alpha C f V^2 \quad (4.1)$$

Therefore, dynamic scaling of voltage and frequency through dedicated hardware can effectively reduce the power consumption. However, such power reduction is subjected to slower switching speed and hence performance penalty. Careful timing analysis is also required at design stage to avoid timing violations when the voltage is scaled. In comparison, the software-based task scheduling are usually claimed to have less overheads. For a single processor, task scheduling optimizes the temporal task allocation so that when the processor approaches the thermal threshold, techniques such as fetch throttling [245] and task interleaving [158] can be applied to lower the processor temperature. While for multi-core systems, spatial reallocation becomes possible where tasks can be migrated from hot to cool processors with no significant performance loss [160]. The effectiveness of task scheduling is highly dependent on the nature of instructions which can be CPU-intensive or memory-intensive and thus not assured. In addition, due to the much higher power density than that of 2D MPSoC, both DVFS and task scheduling would be helpful but

may be insufficient to maintain the 3D MPSoC at a safe temperature with existing heatsink-based air cooling.

Besides reducing the heat generation, the other angle of view to solve the thermal problem of 3D MPSoCs is enhancing the cooling solution. The state of art heatsink-based air cooling provides a junction-to-ambient thermal resistance of around 0.6 K/W [180] which is insufficient, and this cooling method is not scalable as heat dissipation for circuit layers far from the heatsink remains difficult. Many of the prior works address this problem by utilizing the through-silicon-vias (TSVs) during the design stage [140, 175, 176, 246]. By inserting thermal TSVs (T-TSVs) at potential hot spots to provide fast heat conduction path, temperature reduction is expected. However, TSVs are large and occupy significant silicon area. T-TSV competes with other circuit devices for the precious silicon area therefore the quantity of T-TSVs that can be inserted is limited. Furthermore, this method is static and requires prior knowledge about the target applications to determine the potential hot spots. Once implemented, the T-TSVs cannot be adjusted to adapt to runtime dynamic workload changes.

Recently, it was found that microfluidic channel array [187, 204, 247] is more effective in removing heat in 3D integrated circuit (3DIC) by providing local thermal ground. The thermal resistance can be as low as 0.09 K/W [179]. However, implementation of microfluidic cooling causes overheads such as extra manufacture procedures, increase of stack height (chip thickness) and additional cooling power. Therefore good thermal management scheme is needed to keep the overheads low while achieving the best cooling effect. An exemplary work is presented in [187]. It adjusts the flow rate uniformly according to the predicted future maximum temperature, which depends on the historical maximum temperatures measured by temperature sensors. This kind of thermal management is too coarse to maintain a safe operating temperature for today's high performance circuits which are prone to diversified thermal hot spots. The uniformly changing flow rate may cause overcooling at the low power regions and hence waste cooling power. In addition, the implementation of temperature sensors is expensive in 3D which implies a high cost of this thermal management scheme. A well designed thermal management scheme

is desirable to adjust the cooling effort according to the real-time workload. It should maintain the chip temperature well below the threshold while keeping the implementation overhead low.

In this chapter, a novel scheme of runtime temperature management for 3D multi-core cache-processor system will be presented. Under this scheme, the system temperature will be maintained at a desired level by adjusting the fluidic flow rate according to the real-time cooling demand. Such a real-time cooling demand is determined by the future power consumption and thermal model of 3DIC with microfluidic channels. Auto Regressive (AR) predictor [250] is adopted to predict the future power consumptions of on-chip components and the prediction is further corrected by Kalman filtering [251] to remove the error contributed by system variation and control. As such, the expected steady state temperature under existing cooling setting, that is defined as the real-time cooling demand, can be estimated from the predicted future power consumption. With such an efficient determination of cooling demand at thermal-time-constant scale, the flow rate is adjusted at runtime to prevent the system from running out of the thermal threshold or over-cooling. Moreover, the concept of channel clusters is also introduced for the first time to provide a more efficient fine-grained control of flow rates for disparate groups of channels to reduce the cooling effort required. Note that this work assumes a software-sensing based temperature evaluation. A smaller implementation cost is expected when compared to the use of temperature sensors. Nevertheless, the similar thermal-management policy can still be deployed for temperature-sensor based scenario. In order to better maintain the chip temperature at reduced cooling cost, it is also possible to apply this thermal management scheme for the conventional power reduction circuit design techniques such as dynamic voltage and frequency scaling and load scheduling.

The remaining part of this chapter is organized as follows. In Section 2, the targeted cache-core system with microfluidic cooling is presented followed by the formulation of runtime temperature management problem. In Section 3, our proposed thermal management scheme is detailed with: (1) power sensing, prediction and correction for the 3D multi-core system; (2) real-time cooling demand estimation;

and (3) fine-grained flow rate allocation. The experimental results are presented in Section 4 followed by a chapter summary in Section 5.

4.2 Problem Formulation of Runtime Thermal Management of 3D Multi-Core Systems

4.2.1 Typical 3D Multi-Core Cache-Processor System

3D multi-core and cache system in the literature were usually implemented in logic-on-logic or logic-on-memory configurations. Correspondingly, two different 3D multi-core and cache structures will be studied for the runtime thermal management problem. Fig. 4.1(a) illustrates the multi-core and cache system *A*, which consists of stacked layers with two cores and caches on each layer. Each core communicates with the cache on the upper or lower layer through TSVs. Cores and caches on adjacent layers are arranged in a complementary manner to avoid highly concentrated heat generations. Each cache is divided into four banks, and one core can access one bank of caches at one time. System *B* is shown in Fig. 4.1(b), where four alpha cores are placed on the top layer, with three cache layers stacked underneath.

Figure 4.1: (a) 3D multi-core and cache system *A*; (b) 3D multi-core and cache system *B*. Both are with microfluidic cooling.

Similar to the previous 3D multi-core performance studies [142, 238–242, 244], the alpha-2 architecture for each computing core is assumed and set-associative

Figure 4.2: Heatsink-cooled 3D multi-core and cache system

SRAM structure is adopted for L2-cache. Conventional 3D designs [140, 175, 176, 246] assume heatsink-based air cooling, as shown in Fig. 4.2. The heatsink placed at the top of the device layers acts as the primary heat removal path to the ambient air. As a result, this architecture has a long heat removal path problem especially for the layers at the bottom. The more effective microfluidic cooling with microchannels is shown in Fig. 4.1. The microfluidic channels are sandwiched in between circuit layers, which provide local thermal grounds to the ambient and dramatically shorten the heat-removal path for each active device layer. This microfluidic cooling scheme is adopted as the cooling solution for our thermal management scheme. The cooling effect of heatsink is also investigated for the purpose of comparison.

4.2.2 Runtime Thermal Management Problem of 3D Multi-Core Systems

Given the 3D multi-core cache-processor system with microfluidic channels, our objective is to develop a thermal management scheme that can actively control the flow rate of microfluidic channels to maintain the system temperature within certain temperature range. For any thermal control system, its essential components should include sensing system, processing unit and actuator, where sensing system collects real-time thermal information of the 3D system; processing unit then analyzes the

situation and sends out appropriate adjustment commands to the actuator for flow rate control.

Most of the previous works [239–244] assume on-chip thermal sensors for direct temperature measurement. This method is straight forward but has its limitation. A microprocessor typically has less than ten thermal sensors due to their area overheads. Such limited number of thermal sensors are incapable of providing a detailed temperature map down to the microarchitecture level. The maximum and minimum temperatures based on their measurements may deviate from the actual thermal extremes of the 3D system and thus mislead the decision making process. To overcome such limitation and reduce implementation cost of thermal sensors, a software-based remote sensing technique is used in our work, where the steady state circuit temperature is simulated based on workload predictions.

The predicted future circuit temperature acts as the cooling demand of the 3D multi-core system. With this information, the processing unit then decides on how much coolant should be supplied to meet such demand and maintain the circuit temperature. Coskun et al. [187] proposed a possible solution by using look-up tables to correlate flow rate settings with maximum temperatures. Despite the potential error on the maximum temperatures reported by a single thermal sensor per processor as assumed in their work, it is not reliable to decide the flow rate based on a single maximum temperature. 3D systems that have the same maximum temperature may have different power consumptions and temperature distributions therefore require different coolant supply. In addition, this look-up table approach requires extensive experiments or simulations to determine the correct flow rate and temperature pair for a given 3D MPSoC. Therefore, although this method appears feasible and low cost for a particular 3D chip, it implies a lot of preparation effort and would be complicated for wide adoption by different 3D systems. In our thermal management scheme, the required flow rate setting is determined through approximation based on the microfluidic cooling characteristics. This approximation method is analytical and can be used by different 3D systems without any preparation.

One important feature in our 3D MPSoC thermal management is the granularity of flow rate control. Since each functional unit in the system works under dif-

ferent conditions, the power and temperature would be non-uniformly distributed across different layers. Therefore, different from previous works [187, 204, 247], non-uniform flow rates are allowed in our system and they can be dynamically controlled to remove the heat in a more effective manner. In addition, fine-grained flow rate control allows graceful degradation instead of catastrophic failure when one microchannel gets clogged. However, fine-grained control at every channel is not realistic for the physical implementation due to the need of many pumps and complicated fluid network. Hence in this paper, a more practical approach using channel clusters is employed. The microchannels can be divided into a number of groups and the channels in the same group will have the same flow rate. Each cluster can be powered by one micro-pump and its flow rate is controllable through the pump. Experiments show that the use of channel clusters can efficiently reduce the hotspots with significant flow rate savings.

4.3 Power Aware Runtime Thermal Management Scheme of 3D Multi-Core Systems

In this section, the proposed power aware runtime thermal management scheme will be presented in detail. Instead of fully relying on the thermal sensors for temperature measurement, a software-sensing method is adopted to predict the future circuit temperature through simulation. The power consumption information is thus a prerequisite as input for the thermal simulations.

4.3.1 Power Consumption Tracking for Cache and Micro-processors

During real-time operation, the system workload is not static but dynamically changing over time. Modern power management techniques such as clock gating can further introduce additional dynamic behavior for the system power trace. Any thermal management scheme based on the static demand would be insufficient to follow the actual system behavior and may lose control on system temperature after some time.

To overcome this problem, an online monitoring mechanism is needed to keep track of the caches and microprocessor's power consumptions. This power monitoring system should be able to provide accurate power information at microarchitecture level for a meaningful high-resolution thermal simulation. Hardware-based measuring technique is available by inserting current sensors [171, 248] into the power delivery network. However, similar to thermal sensors, the expensive implementation cost limits their number and hence the ability to provide detailed power values. Therefore, this hardware approach is only suitable for homogeneous systems with coarse-grained monitoring. In our work, we aim at more practical cases where the 3D multi-core system has diversified power values due to the microarchitecture-level heterogeneity of functionality and workload. One way to achieve such level of power sensing is to correlate the readings from the performance monitoring unit (PMU) with power consumption [89–91], as mentioned in Chapter 2. However, to provide simulation inputs for our case study, a software-based power sensing method is adopted.

Software-based Power Sensing of Microprocessor and Cache

For a given 3D multi-core system, its power consumption depends on the running applications. From the scheduler of the operating system, the task queue can be obtained for each microprocessor. The power consumption of the processors can thus be simulated using existing architecture-level power simulators such as Wattch [252]. Besides microprocessor, the power consumption of caches can be estimated by Cacti [253]. Both dynamic power and leakage power are accounted for each cache bank being accessed, while the standby cache banks are assumed to consume only leakage power. An evenly distributed accessing rate for memory cells is assumed so that the dynamic power distribution of caches would be even but the leakage power will depend on the temperature of individual cache bank. The BSIM3 model is adopted to account for the thermal dependency of leakage power, where the unit channel-width leakage current is formulated as follows:

$$I_{leakage} = \mu_0 \times C_{ox} \times \frac{W}{L} \times e^{b(V_{dd}-V_{dd0})} \times V_t^2 \times (1 - e^{\frac{V_{dd}}{V_t}}) \times e^{\frac{-|V_{th}|-V_{off}}{n \times V_t}} \quad (4.2)$$

where the thermal voltage $V_t = kT/q$ and threshold voltage V_{th} are both functions of temperature.

For validation purpose, a power trace library for different categories of workloads is pre-stored to emulate the power profile at real-time. Benchmarks from SPEC2000 [254] are applied to alpha-2 cores and the corresponding power information are obtained from Wattch and Cacti. Note that the aforementioned power is usually estimated at nano-second scale, known as the cycle-accurate electrical power but the temperature management is usually performed at milli-second scale. Thus, the over-sampled power information has to be down-sampled to thermal constant time scale (milli-second scale). The thermal-power, P_T is thus defined by the runtime average of the electrical power of both dynamic and leakage power as follows:

$$P_T = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} (P_{dyn} + P_{leak}) dt \quad (4.3)$$

After conversion to thermal power, the entire power history is stored into power trace files. The prepared power trace file is then input line by line to the thermal simulator as if the system power consumption and thermal condition are periodically sampled. The software sensing period is thus defined as $t_{sense} = t_2 - t_1$. To update the cache power according to the real-time temperature distribution, Cacti is integrated into our thermal simulator to perform the updating task whenever the temperature profile has changed. To reduce the implementation cost, this update is only performed once without the iterative feedback loop during our implementation.

Prediction and Correction of Future Workload

To proactively maintain the thermal condition of 3D multi-core systems, the near future power consumption information is needed to foresee potential thermal contingency. As the source of heat generation, the accuracy of power prediction has a strong influence on the effectiveness of any proactive thermal management scheme. Autoregressive, i.e., AR(p) model [250] is a commonly used prediction method to correlate time series data. The basic idea of an AR(p) model is to find the best polynomial fitting of a time varying series $f(n)$, which is the system power trace in our problem. The p coefficients, $c_1, \dots, c_p$, and p previous thermal power values,

$f(n-1), f(n-2), \dots, f(n-p)$, are then used to predict the future power state by

$$f_p(n) = c_1 f(n-1) + c_2 f(n-2) + \dots + c_p f(n-p) + \varepsilon(0, \sigma_p) \quad (4.4)$$

where $\varepsilon(0, \sigma_p)$ denotes a random error with Gaussian distribution (zero mean and variance σ_p).

The AR(p) model can be applied to predict power in the next sensing period and hence the near-future cooling demand. If the electrical power of core and cache are deterministic, the system power trace can be exactly determined or predicted. However, in real-time applications, the workload can vary dynamically, the leakage current and power can fluctuate stochastically, and hence both the electrical and thermal power can have uncertain variability or error. To have an accurate and timely temperature regulation, a relatively accurate estimation of the future power and temperature are necessary to filter this variability. This can be resolved through a controller with prediction-and-correction by treating the uncertainty as a random process variable of a probability distribution.

Kalman filter [251] is used for the correction of prediction error in our work. The purpose of Kalman filter is to adjust the estimated result $f_p(n)$ based on an additional physical measurement according to its noise level. Such measured result $f_m(n)$ refers to the software-sensed power in our proposed scheme, which is first obtained from the cycle-accurate electrical power simulation and then further averaged to the thermal power at the scale of the thermal time constant. Assume that such a software-sensed physical measurement has a noise with variance σ_m , and the measurement result is $f_m(n)$, then the Kalman gain can be obtained by

$$K = \frac{\sigma_p}{\sigma_m + \sigma_p} \quad (4.5)$$

Then, the corrected estimation can be calculated as

$$f_e(n) = f_p(n) + K \bullet (f_m(n) - f_p(n)) \quad (4.6)$$

After this correction, a more accurate thermal power prediction is obtained. Such prediction and correction processes are invoked only when required by the thermal simulator for determining the cooling demand. The frequency of invocation is specified by the user-defined control period $t_{control}$. Every time when a period of $t_{control}$

has elapsed, the thermal controller requests for a corrected power prediction and estimates a real-time cooling demand as described in the following section.

4.3.2 Real-time Cooling Demand Estimation

Steady State Thermal Modeling of 3DIC with Microfluidic Cooling

With the predicted future workload and the hardware specifications of the 3D multi-core system, the steady state temperature can be simulated by following the thermodynamic theories presented in Chapter 3. Assuming that there are N grid cells in the 3D system and an additional node “ a ” for the ambient, knowing the thermal conductance between nodes, the following equation can be established for any cell, i ($i = 0, \dots, N$).

$$\sum_{j=0}^{N-1,a} g_{ij}T_i - \sum_{j=0}^{N-1} g_{ij}T_j = P_i + g_{ia}T_a \quad (4.7)$$

where g_{ij} denotes the thermal conductance between nodes i and j . As a result, N equations can be set up for N cells and a steady state matrix equation is obtained as

$$G_{steady}T_{steady} = P \quad (4.8)$$

The matrix T_{steady} stands for the steady-state temperature profile and can be analyzed for decision making. In our thermal management scheme, an upper temperature limit T_{upper} and a lower temperature limit T_{lower} are specified as the triggering thresholds. The flow rate will be increased if the maximum temperature T_{max} is higher than T_{upper} or reduced if T_{max} is lower than T_{lower} to avoid over-cooling and reduce runtime thermal cycle. Such requirement of flow rate adjustment is thus defined as the real-time demand. Here the steady-state temperature is chosen to be the judging criteria instead of the instantaneous temperature as in previous works [187] due to the following considerations:

1. Instantaneous temperature indicates only current thermal state and does not tell whether the temperature is rising or decreasing. To obtain such information, a history of the instantaneous temperature needs to be kept, thus requiring many more thermal simulations and increasing implementation cost.

2. Unlike DVFS and task scheduling techniques, thermal management through microfluidic cooling typically have a longer response time for the pump to adjust the flow rate. Steady state temperature provides a larger prediction step size, thus giving the control system more time to respond.
3. Consider a scenario when a high power application is loaded such that the power consumption is suddenly increased, with instantaneous temperature, several flow rate adjustments may be needed to finally match the flow rate with cooling demand. Alternatively, by steady state temperature prediction, the flow rate needs to be adjusted once only to satisfy the cooling need.
4. Simulation of instantaneous temperature depends on previous results. Such dependency will cause accumulation of errors. The estimated results may gradually stray away from the actual circuit temperature. In contrast, deviation of steady state simulation is kept within one simulation and does not propagate to the next and subsequent simulations.

Flow Rate Adjustment Approximation Based on Maximum Temperature

Due to the complexity of 3D system's thermal behavior, to determine the amount of flow rate adjustment to meet the cooling demand is not trivial. As explained earlier, the method used in [187] is inappropriate due to its inaccuracy and hassle of preparation needed. With the aid of accurate thermal modeling, a straightforward solution is the trial-and-error approach, whereby an iterative loop of flow rate adjustment and assessment by thermal simulation is performed until a matching flow rate is found. The amount of flow rate adjustment can be related to the difference between target temperature and current maximum temperature as suggested by [276]. Due to the monotonic relationship between circuit temperature and flow rate, the needed flow rate can definitely be obtained by this method. However, this approach may be too slow for runtime management because of the time taken to reach convergence in this iterative thermal simulation.

An approximation method is therefore proposed to solve this problem. For any

circuit grid cell, its temperature can be formulated as follows.

$$T_{circuit} = T_{ch} + (R_{conv} + R_{cond})q \quad (4.9)$$

where T_{ch} is the temperature of the channel cell in contact, R_{conv} and R_{cond} denote the convective and conductive resistances along the heat transfer path. q denotes the heat flux from the circuit cell to the channel cell. Assume that the channel cell is in the n^{th} grid row, its temperature rise would depend on the total amount of heat it absorbed from the surroundings, as approximated by (4.10).

$$T_{ch} = \frac{\sum_{i=0}^{n-1} q_i}{\rho_f Q C_f} + T_a \quad (4.10)$$

where ρ_f , Q , and C_f denote the density, flow rate and heat capacity of the coolant. The heat absorbed by each section of the channel q_i is estimated as follows.

$$q_i = \alpha P_{above} + \beta P_{below} \quad (4.11)$$

where α and β are factors to account for the partial absorption of circuit cell power dissipation by the channel cell.

Changing flow rate Q influences both T_{ch} and R_{conv} . However, due to the large contribution of T_{ch} to the magnitude of $T_{circuit}$, and T_{ch} 's superlinear dependency on Q , the change in T_{ch} greatly dominates the $T_{circuit}$'s change when the flow rate is adjusted. Thus the control of circuit temperature can be achieved through controlling the coolant temperature. Based on the above observations, the following flow rate adjustment algorithm is proposed. To simplify the calculation of the new flow rate, the influence of Q on R_{conv} is ignored, which will cause a small deviation in the resulting temperature. This turns out to be a blessing in disguise. If the maximum temperature is larger than the upper thermal limit, the flow rate will be increased until the new maximum temperature drops slightly below T_{upper} due to this small deviation. On the other hand, if the 3D system is over-cooled, the deviation would cause the flow rate to be reduced more and save more coolant. Due the dominating effect of T_{ch} , this deviation will be limited.

Algorithm 4.1 Flow Rate Adjustment Algorithm

Function Adjust():

Input: Steady state temperature and 3D system specification

Output: Adjusted flow rate value matching the cooling demand

Find the max temperature T_{max} and its location

Record the channel coolant temperature T_{ch} in contact with the hottest spot

if $T_{max} > T_{upper}$ **then**

$$\delta T = T_{max} - T_{upper}$$

else
if $T_{max} < T_{lower}$ **then**

$$\delta T = T_{max} - T_{lower}$$

end if
end if

Expected new channel temperature: $T'_{ch} = T_{ch} - \delta T$

Adjust flow rate to $Q' = \frac{T_{ch} - T_a}{T'_{ch} - T_a} Q$

4.3.3 Runtime Thermal Management Scheme with Fine-Grained Flow Rate Control

Adaptive cooling of microprocessors was commonly seen in modern computers with air cooling where the fan speed was adjusted according to the processor temperature. The resulted change of cooling effort applies to the entire chip regardless of the location and range of the triggering hot spot. In practical situations, the triggering power rise is usually localized at several hot spots and not globally spread. Increasing the cooling effort of entire chip may result in over-cooling at low power areas thereby creating large spatial thermal gradient and wasting cooling power. Such problem does not have a good solution under the conventional heatsink-fan-based air cooling due to physical constraints. On the contrary, the isolation of coolant between microchannels enables a fine-grained microfluidic cooling, which has not been fully utilized in [187] due to their uniform flow rate cooling.

In our proposed thermal management scheme, different flow rate settings are

Figure 4.3: Channel clustering with 3 and 6 clusters

applied to the microchannels. The ultimate fine-grained control where each microchannel has its own flow rate setting would be too costly to implement. Instead, a trade-off method called channel clustering is proposed, where the microchannels are divided into groups and members in one group share one common flow rate setting. Each channel cluster is in charge of cooling a part of the chip defined as a cooling region. In this way, when a thermal contingency happens, the cooling region where it located can be identified and only the flow rate of the corresponding channel group needs to be adjusted according to Algorithm 4.1. In our experiment, the microchannels are divided into 3 and 6 clusters, as shown in Fig. 4.3, considering design feasibility of the fluid network. With a fine-grained flow rate control, the effectiveness of thermal management by active cooling is largely increased when compared to the uniform flow rate control. As the non-uniform flow rate control is not applied to each channel, the implementation, for example, by providing each layer with a micro-pump is practical and economical. With such a fine-grained flow rate control, a more uniform temperature distribution along with a reduction in the total required flow rate is also expected.

Fig. 4.4 summarizes the general flow of our proposed runtime thermal management scheme. It reacts once for every control period in the scale of thermal time constant. Given the latest sampled power values in the last control period, the system power trace at the next sampling time is first predicted and then the related error is corrected to ensure its accuracy. The corrected value of the system power

Figure 4.4: Procedure of the proposed thermal management scheme

trace is used to simulate the expected steady state temperature of the 3D system. Based on the temperature information, the real-time cooling demand can be determined and a fine-grained flow rate control is finally achieved with channel clustering technique.

4.4 Experimental Results and Discussion

The proposed runtime thermal management scheme is implemented in C/C++ including: (1) thermal model of 3DIC with microfluidic cooling; (2) update of cache leakage power based on Cacti [253]; (3) AR prediction and Kalman filter correction for the future workload; (4) real-time cooling demand calculation; and (5) fine-grained flow rate control with channel clustering. The details of the thermal model have been presented in Chapter 3. The simulations are run on an Intel Dual-Core Server with 3 *GHz* CPU and 8 *GB* RAM. Table 4.1 summarizes the common design parameters of the two 3D multi-core testbenches used in the following experiments. Other assumptions and experimental settings are listed below:

- For the input applications and power estimation, SPEC2000 applications are

Table 4.1: Parameters of 3D multi-core system with microfluidic channels

Parameter	Value
Chip Size	$32mm \times 32mm$
BEOL Thickness	$12\mu m$
Source Thickness	$2\mu m$
Bulk Thickness	$48\mu m$
Microfluidic Channel Height	$100\mu m$
Microfluidic Channel Width	$100\mu m$
Microfluidic Channel wall width	$100\mu m$
Total Number of Microchannels	480

applied as input to Wattch [252] at its default settings. The average thermal-power over 10 million clock cycles is recorded. Hence, $T_{sense} = 3\ ms$ assuming a clock period of $0.3\ ns$.

- For the 3D multi-core cache-processor design, the alpha-2 core architecture is assumed with a 2-way set-associative (L2) cache, which consists of 4 banks and 32 address bits.
- Water is used as the coolant. The ambient temperature T_a is set to be $300\ K$. The upper temperature threshold is specified as $358.15\ K$ ($85^\circ C$), and the lower limit is set to be $354.15\ K$ ($81^\circ C$).
- Third order AR prediction is used when implementing the runtime temperature management. In addition, 10% estimation error is assumed for the power trace generated by the power simulators.

4.4.1 Effectiveness of microfluidic cooling

The effectiveness of cooling by microfluidic channels is first examined by its comparison with conventional heatsink-based cooling using 3D structure A. A static power consumption is used at this time by averaging the power trace generated

Table 4.2: Cooling effect comparison between heatsink-based and microfluidic cooling by comparing the maximum and minimum temperatures and the number of nodes in each temperature range

Temperature Range	No. of Nodes in Each Range			
	300-315	315-330	330-345	345-360
Heatsink				
324.12-348.96	0	3692	398	6
Microfluid (80 <i>ml/min</i>)				
301.34-349.00	790	1705	1598	0
Microfluid (240 <i>ml/min</i>)				
300.26-326.10	3725	371	0	0

when loading *gcc* to all the cores. Default settings in Hotspot [222] is adopted for the heatsink-cooled case, and the coolant flow rate is set to be uniform for microfluidic cooling. The temperature distribution of the 3D system is recorded by counting the number of grid nodes within each temperature range.

As shown in Table 4.2, at the small flow rate of 80 *ml/min*, microfluidic cooling has already achieved similar cooling result as the heatsink-based cooling method. At higher flow rate, the circuit temperature is effectively suppressed. Fig. 4.5 plots the trend of maximum circuit temperature when increasing the total flow rate. It can be seen that increasing the flow rate can effectively cool down the chip and remove hotspots. The temperature reduction reaches diminishing marginal returns when the flow rate continues to increase, as implied by (4.10).

4.4.2 Runtime thermal management with fine-grained flow rate control

In this section, the proposed runtime thermal management scheme will be demonstrated using the aforementioned 3D multi-core testbenches.

Figure 4.5: Effective heat removal of microfluidic cooling by increasing flow rate

Power Sensing, Prediction and Correction

The microarchitecture-level power consumption of the processors running different applications is first pre-determined by Wattch and stored in the power trace file. Fig. 4.6 shows the dynamic power consumption of a single microprocessor when running four different SPEC2000 benchmarks, namely *gcc*, *fma*, *bzip* and *eon*. To enrich the application pool, five pseudo benchmark power traces *2Xgcc*, *2Xfma*, *2Xbzip* and *2Xeon* are also created by doubling the original power traces to represent high-power applications. The microprocessors in the 3D system are randomly loaded with these eight benchmarks and the simulation of thermal management runs for 6000 lines of thermal power traces to emulate a period of practical operation. The power traces in Fig. 4.6 exhibit significant temporal variations and are very different from each other. Both temporal and spatial thermal strikes are hence expected when the tasks are loaded onto the 3D system. To catch these potential real-time thermal strikes, the system needs to be regularly monitored with proper control action applied. On the other hand, the control action should not be unnecessarily frequent to avoid high thermal management cost. In this experiment, system probing is engaged for every 200 samples of thermal power trace, i.e. at 0.6s interval. Based on the 200 historical power values, future workloads are predicted by an AR(3) predictor and further corrected by a Kalman filter. Fig. 4.7 plots the 29 predictions and corrections along

Figure 4.6: Total power consumption of Alpha-2 microprocessor running different benchmark applications

the 6000 inputs of power traces for system *A*. The difference in the predicted total

Figure 4.7: AR prediction of future power consumption with Kalman filter correction
power after the Kalman filter correction is discernible. Such corrected prediction

to mitigate possible deviations and system uncertainties is necessary for a more accurate control. From Fig. 4.7, several power spikes can be detected. Whether these spikes will cause thermal violations depends on the cooling settings at their individual arrival time.

Fine-Grained Flow Rate Control According to Realtime Cooling Demand

Fig. 4.8 tracks the maximum temperature of the 3D testbench A if no thermal management is applied and the coolant flow rate is fixed at $100\text{ml}/\text{min}$. It is clear

Figure 4.8: Maximum temperature of the 3D testbench A with no runtime thermal management

that proper thermal management is mandatory, without which the 3D system will experience severe thermal cycle with maximum temperature up to 401K .

The proposed thermal management scheme is implemented on both systems A and B as follows. At the beginning of the simulation, the 3D systems were assumed to be at steady state running *gcc* benchmark with a coolant flow rate of $80\text{ml}/\text{min}$, where system A has a maximum temperature of 349K and system B has a maximum temperature of 331K . The benchmarks are then loaded onto the system. For every 0.6s elapsed, the future power consumption is predicted and corrected. The thermal model developed in Chapter 3 is then invoked to simulate the

steady state temperature profile at current flow rate setting. The resulting maximum temperatures are interpreted as the real-time cooling demand. Appropriate coolant flow rate settings are then decided for every channel cluster according to Algorithm 4.1. The microchannels are divided into 3 and 6 clusters as shown in Fig. 4.3. For the purpose of comparison, thermal management with uniform flow rate cooling is also implemented. Note that due to different clustering and hence different flow rate settings, the real-time cooling demand at every control point may be different across the three cases even if the power profiles are the same. Fig. 4.9 depicts the real-time cooling demand and maximum circuit temperature under thermal control. It can be seen that for any granularity of flow rate control, the proposed thermal management scheme always meets the real-time demand and maintains the circuit temperature within the upper and lower constraints.

However, the total flow rates required to achieve the same cooling objective are different for 3D systems with different clustering settings of microchannels. As shown in Fig. 4.10, the total flow rate (summation of all channel flow rates) decreases dramatically with fine-grained flow rate control. Comparing with uniform flow rate cooling, the 3-channel-cluster setting has saved on average 8.6% flow rate for system *A* and 2.4% flow rate for system *B*; for the case of 6-channel cluster, a significant 27.4% flow rate saving is achieved for system *A* and 8.8% for system *B*. It can be seen that due to the lower and static power profile, the cache-dominated system *B* requires a lower coolant flow rate that is also much less dynamic than that of system *A*. The almost uniform power density of the cache layers of system *B* directly leads to the smaller flow rate savings. It is clear that with finer clustering granularity, more flow rate savings can be achieved. The internal mechanism accounted for this result can be explained as follows. With the fine-grained flow rate control, the flow rates have been optimally customized to local hotspots and unnecessary flow rate increments can be avoided. When there are more clusters, there is a better chance to encounter regions that are totally cooled and hence the corresponding flow rates can be decreased. More importantly, the total flow rate directly reflects the pumping power needed. Hence our fine-grained flow rate control has a much lower pumping power overhead. The fine-grained flow rate control provides an additional design

Figure 4.9: Maximum circuit temperature of (a) - (c): system A and (d) - (f): system B under thermal management of different clustering granularity and triggered by real-time cooling demand.

freedom to balance the pump power and control circuit overheads.

(a) Flow rate settings of system *A* (b) Flow rate settings of system *B*

Figure 4.10: Dynamic flow rate settings under different granularity of flow rate control

4.5 Summary

This chapter has introduced a dynamic thermal management scheme to maintain chip temperature for 3D multi-core cache-processor systems deployed with microfluidic channels. The proposed thermal management scheme implements a closed-loop control system that is driven by real-time demand, and endowed with prediction-

and-correction capability and fine-grained adaptation. A combination of power and thermal simulators is proposed to provide detailed circuit temperature information and reduce the cost of thermal sensor implantations. The AR prediction and Kalman filter correction of future workloads forecast potential power spikes and allow the cooling system to react proactively to achieve thermal stability. With the aid of the thermal model, the real-time cooling demand is determined and the flow rate is allocated in fine granularity to achieve the best cooling result while reducing the cooling cost. Experiment results show that with the proposed thermal management scheme, the system temperature is suppressed well under a reliable threshold. The total flow rate, which is a direct reflection of the pumping power, has been reduced by up to 27.4% with a fine grained flow rate control.

An important innovation of the proposed thermal management scheme is the fine-grained microfluidic cooling using the concept of channel clusters. The channel clustering method used in this work is intuitive and lacks theoretical guidance thus the optimality is not guaranteed. From experiment results it can be seen that the way the channels are clustered significantly affects the effectiveness of the fine-grained liquid cooling. There is thus a need to develop a channel clustering algorithm for the best cooling results. In addition, the comparison between microfluidic cooling and conventional heatsink-based air cooling shows clear advantage of the former due to its high heat transfer rate and scalability. However, it is also worth noting that microfluidic cooling typically exhibits large spatial thermal gradients, as shown by the max-min temperatures in Table 4.2. A good channel clustering method is one possible solution to naturally mitigate the thermal gradient by allocating coolant according to the needs at different locations on the chip. Therefore, in the following chapter, we will focus on the development of a clustering algorithm for fine-grained microfluidic cooling, and investigate possible solutions to the thermal balancing problem of microfluidic-cooled 3D integrated circuits.

Chapter 5

Microchannel Clustering, Splitting and Scaling for Thermal Balancing of Liquid-Cooled 3DIC

5.1 Introduction

Despite the promising cooling potential and scalability reported by most literature, microchannel-based microfluidic cooling tends to introduce large spatial thermal gradients. Such large thermal gradient is mainly contributed by the non-uniform power distribution and the rising coolant temperature along the flow direction. In practical situations, the power density may vary significantly across the chip. To meet cooling requirements, the coolant flow rate is usually set at high level according to the worst case, thus over-cooling the low power region and inducing high cooling costs. Furthermore, as the coolant flows through the microchannels, heat is absorbed and the coolant temperature increases. The rise of coolant temperature directly increases the circuit temperature along the flow direction, thereby determining the basic temperature pattern of microfluidic-cooled circuits. At high flow rate settings, coolant temperature rise is small and power-induced thermal gradient is more pronounced. On the other hand, at low flow rate settings, the circuit temperature follows closely the basic pattern with variations due to non-uniform power distribution.

The stress built up by the large thermal gradient is detrimental to reliable circuit operation. It complicates circuit modeling and also increases design constraints. To solve this problem, a non-uniform channel design scheme was proposed in [185], where microchannels are placed only at necessary hot spots to match with the non-uniform power density. Significant reduction of pump power was reported. Another proposal is the channel modulation technique by [190], which aims to mitigate the effect of rising coolant temperature. By varying the channel width along the flow direction, the convective resistance is reduced at hot spots and hence partially offsets the rising coolant temperature. However, the narrowed microchannels would require extraordinary high pressure drop to deliver enough amount of coolant. Independent coolant supply for every microchannel is also needed to implement this channel modulation technique, which may not be pragmatic even if it is technically possible.

In this chapter, we investigate on thermal balancing techniques for liquid-cooled 3DICs. A detailed analysis on microfluidic cooling for 3DICs by different clustered cooling efforts will be first presented. A channel clustering and flow rate allocation algorithm is proposed to provide a customized cooling according to the circuit power distribution. After that, a combination of novel microchannel splitting and scaling techniques is presented. Different from the clustering method, this splitting and scaling approach aims to match the cooling effort with the heat generation, and at the same time mitigate the effect of rising coolant temperature.

The remaining of this chapter is organized as follows. In Section 5.2, the clustering and flow rate allocation algorithm is presented in detail. A guideline for achieving minimum total flow rate under the maximum temperature constraint is also introduced. In Section 5.3, the channel splitting and scaling algorithm is proposed. We will explain how the splitters are inserted to reduce convective resistance and how the channels are scaled to regulate the flow rates. Experiment results will be presented in Section 5.4, where the proposed algorithms are implemented on a 3D multi-core testbench to observe their effectiveness. Finally, the chapter will be summarized in Section 5.5.

5.2 Microchannel Clustering and Flow Rate Allocation for Non-uniform Microfluidic Cooling

Depending on the circuit function and floorplanning, the power consumption of 3D stack is usually distributed unevenly on each layer. Fig. 5.1 shows the power density map of a multi-core circuit layer, where the red colored cells represent high power density areas as opposed to the blue regions where the power densities are relatively low. If uniform cooling effort is applied to the 3D stack, to keep the temperatures of potential hot spots under threshold, a large flow rate is needed. This will result in wasted energy for over-cooling of the low power regions, which creates large lateral thermal gradients and shortens the device's lifespan. In contrast, if the cooling effort can be customized to adapt to the appropriate cooling demand variation caused by the uneven power distribution, efficient microfluidic cooling with better steady state results is envisioned.

Figure 5.1: An illustration of non-uniform power distribution (W/cm^2) of 3DIC

The scalability and high heat removal ability of microfluidic cooling have been well documented in the literature [179, 180, 187, 204, 247, 263]. One feature that has escaped the attention is the flow rates of the microchannels can in principle be controlled individually to provide a customized cooling since they are independent

of each other. The ultimate control of cooling effort, i.e., the microchannel flow rate, requires one flow rate setting per channel, which leads to large overheads and is usually unnecessary. Instead, a limited number of flow rate settings can be assigned to different channels. The microchannels that share the same setting are said to belong to the same channel cluster. In Chapter 4, a thermal management scheme has been presented based on clustered microfluidic cooling. Although an impressive saving of total flow rate over the uniform flow rate cooling was reported, the channels were grouped intuitively and the results could actually be undermined by poorly grouped clusters. An algorithm that guides the clustering of channels and appropriate flow rate assignment, if correctly designed, can produce better cooling result for this notion of microfluidic cooling.

5.2.1 Clustering of Microchannels

The heat transfer problem of 3DIC can be analyzed by dividing the 3D chip into regular grid cells as presented in Chapter 3. The heat exchange paths of any silicon cell in the 3D grid can be classified into two categories. One is referred to as the silicon to channel heat transfer, which contributes to microfluidic cooling. The other is referred to as the silicon to silicon heat transfer, which only helps to spread the heat passively over the entire chip. The former can be controlled directly by varying the flow rate while the latter depends only on the nearby thermal gradient due to the fixed silicon to silicon thermal conductance for a fixed architecture. However, the silicon to channel heat transfer is predominant due to the larger circuit-coolant temperature differences and highly conductive liquid cooling. An even temperature profile, as part of the goal of our proposed cooling method, will further reduce the silicon to silicon heat transfer to result in nearly zero thermal gradient across the flow direction. In the steady state, the heat generated by each silicon cell needs to be absorbed locally by the neighboring microchannels. Consequently, the cooling effort of microchannels should match the power distribution of the silicon cells. Hence, the power consumption level to be coped by each channel makes a good criterion to divide the channel clusters.

As shown in Fig. 5.2, any microchannel (denoted by the channel number l_i ,

Figure 5.2: Grid division of 3D integrated circuit

$i = 0, 1, \dots, M - 1$) is in contact with two columns of silicon cells that generate heat, one above and one below the channel. We define the two columns of silicon cells as the cooling region of channel l_i , denoted by

$$CR(l_i) = \{\text{silicon cell } j \mid j \text{ is in contact with channel } l_i\} \quad (5.1)$$

Let the power consumption of silicon cell j be P_{c_j} . It is possible to define a cooling demand for each channel as a weighted sum of power dissipation based on the silicon cells in the region, i.e., $P_{sum,l_i} = \sum_{j \in CR(l_i)} (\alpha_j \cdot P_{c_j})$, where

$$\alpha_j = \begin{cases} 0.5 & \text{if cell } j \text{ is on an inner layer} \\ 1 & \text{if cell } j \text{ is on the top or bottom layer} \end{cases}$$

The factor α accounts for the advantage of cells situated at the inner layers that have twice the number of channels for cooling compared to those at the top or bottom layer. These cooling demands are then sorted in ascending order of magnitude into an array $P = [p_0, p_1, \dots, p_{M-1}]$, where $p_i < p_j$ for $i < j$ and $p_i = P_{sum,l_i}$ for $i, l_i \in [0, M - 1]$. The channel indexes corresponding to the sorted power values in P can be retrieved from an order array, $L = \{l_0, l_1, \dots, l_{M-1}\}$, where $l_i = j$ if $p_i = P_{sum,j}$.

The next step would be the segmentation of the sorted array P to divide the M channels into N ($N < M$) clusters. Since channels in one cluster will share the same flow rate setting, the standard deviation of their P_{sum,l_i} values should be as low as possible to achieve an uniform cooling under the same flow rate. This can be readily

resolved by finding the natural separations, which are the largest differences between every pairs of adjacent elements of the array P . An array $D = [d_0, d_1, \dots, d_{M-2}]$ can be obtained by computing the difference $d_i = p_{i+1} - p_i$ for $i = 0, 1, \dots, M - 2$. The demarcations between clusters can then be found by searching for the $(N - 1)$ largest elements in D .

The number of clusters, N can create subtle trade-off between conflicting factors such as the thermal profile, total flow rate and overhead for a given power trace. The dilemma is that there is no easy way to predetermine an ideal N . Therefore, a divide and conquer approach is adopted in this work. By searching for the largest difference in D , the elements in the sorted array P are divided into two clusters. If the largest difference is d_j , then all channels with indexes l_i for $i \in [0, j]$ will be grouped into one cluster and the remaining channels with indexes l_i for $i \in [j + 1, M - 1]$ will be grouped into another cluster. Let $r_i = p_{M-1} - p_i$ be the moderated cooling demand of the i^{th} channel, and μ_R and σ_R be the mean and standard deviation, respectively of the modulated power values of all channels within the same cluster. If $\sigma_R > \eta \cdot \mu_R$, where η is a parametric fraction, the cluster will be further divided into two clusters by the same approach. This process iterates until every clusters have their σ_R less than $\eta \cdot \mu_R$. The pseudo code of the clustering algorithm is shown in Algorithm 5.1.

The reason for using a biased r_i value instead of the p_i value directly for the computation of mean and standard deviation is to relax the uniformity criterion for clusters with lower power consumption while forcing a tighter homogeneity in power consumption for high power regions. This is because high power regions often have low population with high variation in power values. If the clustering is not carried out with enough granularity, severe mismatch between the cooling effort and cooling demand may occur in these regions and lead to potential thermal management failure. The control parameter η provides the desire trade-off between cooling efficiency and overheads. If η is set too high, the clustering may not be optimal; and if η is set too low, excessive clustering may incur heavy overheads. In our simulation, the number of clusters is constrained to between two and six.

Algorithm 5.1 A divide and conquer algorithm for channel clustering

Function clustering(L, η).

Input: L : Set of channels to be clustered; η : cluster control fraction.

Output: C : Set of channels in a cluster;

P_{c_i} : Power consumption of cell i ;

$CR(l_i)$: Cells in contact with channel $l_i \forall i \in [0, M - 1]$

$M = n(L)$: Number of channels in L ;

for $l_i = 0$ to $M - 1$ step 1 **do**

$$P_{sum, l_i} = \sum_{j \in CR(l_i)} (\alpha P_{c_j});$$

end for

Sort P_{sum, l_i} : $P = \{p_i\}_{i=0}^{M-1}$, where $p_i \leq p_{i-1} \forall i \in [0, M - 2]$;

Order $L = \{l_i\}_{i=0}^{M-1}$: $l_i \leftarrow j$ if $p_i = P_{sum, j}$;

Calculate $R = \{p_{M-1} - p_i\}_{i=0}^{M-1}$;

for $i = 0$ to $M - 2$ step 1 **do**

$$d_i = p_{i+1} - p_i;$$

end for

$$D = \{d_i\}_{i=0}^{M-2};$$

$$t = \arg\{max_j d_j\};$$

$$C1 = \{l_i\}_{i=0}^t; \quad C2 = \{l_i\}_{i=t+1}^{M-1};$$

Calculate μ_{R1} and $\sigma_{R1} \forall l_i \in C1$;

if $\sigma_{R1} < \eta \cdot \mu_{R1}$ **then**

return C1;

else

return clustering($C1, \eta$);

end if

Calculate μ_{R2} and $\sigma_{R2} \forall l_i \in C2$;

if $\sigma_{R2} < \eta \cdot \mu_{R2}$ **then**

return C2;

else

return clustering($C2, \eta$);

end if

5.2.2 Allocation of Flow Rate

The power profile of a 3D chip often exhibits a Pareto distribution, i.e., there is a large number of low power areas and a few patches of high power points. As a result, the channel clustering algorithm will normally produce some clusters with many microchannels of low cooling demand and a few clusters with much less microchannels that require high cooling effort. If the cooling resource is constrained by the total available flow rate, the basic principle of its reallocation is just to move the flow rate saved from low-cooling-demand clusters to where it is needed the most. From the system point of view, the amount of heat removed is $(T_{out} - T_{in}) \cdot Q \cdot SpHeat_f$. If the flow rate is ideally adjusted, uniform temperature distribution can be achieved whereby the inlet and outlet temperature differences of microchannels are similar. In that case, the flow rate should be proportional to the amount of heat removed, which is equal to the steady state power dissipation. Quantitatively, the average cooling demand $\bar{P}(C_i)$ of any cluster C_i can be obtained by calculating the mean of P_{sum,l_j} for all channels $l_j \in C_i$. Due to the predominant silicon to channel heat transfer, the following approximation is used to apportion the total flow rate to each cluster for the best cooling effect.

$$Q(C_0) : Q(C_1) : \dots : Q(C_{N-1}) \approx \bar{P}(C_0) : \bar{P}(C_1) : \dots : \bar{P}(C_{N-1}) \quad (5.2)$$

Thus, the flow rate per channel for the i^{th} cluster is set to be

$$Q(C_i) = Q_{total} \cdot \frac{\bar{P}(C_i)}{\sum_{j=0}^{N-1} n(C_j) \bar{P}(C_j)} \quad (5.3)$$

where Q_{total} denotes the total available flow rate and $n(C_j)$ denotes the number of channels in Cluster C_j .

5.2.3 Minimization of Cooling Effort under Thermal Constraints

In most thermal management scheme, a maximum allowable temperature is set as the control target instead of a fixed total flow rate [187, 206]. As clustered microfluidic cooling is capable of producing an evenly distributed temperature profile, the

cooling effort can be reduced and the total flow rate can be minimized effectively. In our proposed scheme, this is accomplished by incorporating the thermal modeling metric into the clustering algorithm. The procedure is described as follows:

- Step 1: Perform channel clustering according to the power information.
- Step 2: Simulate the steady state temperature of the 3D stack.
- Step 3: Record the current maximum steady state temperature of silicon cells as T_{max} .
- Step 4: Determine scaling factor

$$\varphi = \frac{\overline{T_{max}(Ch)} - T_a}{\overline{T_{max}(Ch)} + (T_{th} - T_{max}) \cdot (1 - \frac{g_{Si,y} + g_{Si,wall}}{nr \cdot g_{Si,Ch}}) - T_a} \quad (5.4)$$

where $\overline{T_{max}(Ch)}$ denotes the average temperature of the channel sections that are in contact with the silicon cells of maximum temperature, and T_{th} denotes the temperature threshold. $g_{Si,y}$, $g_{Si,wall}$ and $g_{Si,Ch}$ denote the thermal conductances from silicon circuit cell to silicon circuit cell along the channel direction, from silicon circuit to channel wall cell, and from silicon circuit cell to channel cell, respectively.

- Step 5: Update the flow rate settings to $Q_{new} = \varphi \cdot Q_{old}$.

The scale factor φ provides a rough guide on the minimization of cooling effort. When flow rate is minimized, the maximum circuit temperature will approach the threshold. The temperature of channel cells in contact with the hottest circuit cell will also rise accordingly to remove the same amount of heat at low flow rate. Therefore, the flow rate is inversely proportional to the inlet-outlet temperature difference. In Equation (5.4), the temperature rise of channel cells in contact with the hottest spot is estimated as a scaled temperature rise of the hottest spot. The use of thermal conductance at high flow rate tends to overestimate the channel temperature while the inclusion of circuit cell to channel wall flux tends to underestimate the channel temperature. This kind of offset makes the estimation relatively accurate.

As such, the total flow rate is reduced by $(1 - \varphi)$. However, since the overall steady state temperature of 3D stack is raised, the thermal gradient will be amplified as well. A positive effect of such amplification is that higher thermal gradient induces a bigger heat flux among the silicon cells to swiftly spread the temperature away from the hot spots. This has added a safety margin spontaneously into our minimization method to ensure that the maximum temperature will not exceed the thermal limit.

5.2.4 Issues on Practical Implementation

Our proposed clustered microfluidic cooling for 3DIC requires coolant supply at multiple flow rates simultaneously. Intuitively, an easy way of implementation is to provide each cluster with an independent pump and manifold system. However, doing so would incur excessive overhead on design complexity, space and power consumption. A more feasible way is to supply the coolant for one 3DIC or several 3D stacks with only one micropump, while using valves to achieve different flow rate settings as suggested in [264]. Micropumps such as piezo-electric micropump [266], and miniature valves [267] are widely available in the market. In such case, the overhead is only the increased complexity of manifold system and additional valves for multiple flow rate settings. With a restricted number of clusters, this overhead is justifiable by the longer lifespan, stable operation and possible speedup of 3DIC in the long run.

As the clustering of channels will be implemented in hardware, once decided, it will be almost impossible to change the channel groupings during runtime. Therefore, the power profile needs to be chosen such that the resulting clustering addresses most of the applications executed on the 3DIC. A reasonable solution is to consider typical applications executed by the 3DIC. Fortunately, the optimal distribution of microchannels will vary much less than the power variations for different applications. Thus, the minor mismatch between cooling and power dissipation would not be catastrophic during runtime. This problem will be discussed further in Section 5.4.

5.3 Microchannel Splitting and Scaling Techniques for Low-overhead Thermal Balancing

As shown in Fig. 5.3, the temperature rise of circuit cells above ambient can be represented by

$$\delta T_{circuit} = \delta T_{channel} + q(R_{cond} + R_{conv}) \quad (5.5)$$

At steady state, the temperature rise of a channel fluid cell is proportional to the

Figure 5.3: Thermal resistances between the circuit and channel cells

total amount of heat absorbed since it enters the channel, which can be formulated as

$$\delta T_{channel} = \frac{\sum_{inlet}^{current\ location} q_i}{\rho_f Q C_f} \quad (5.6)$$

The above formula clearly indicates that the amount of channel temperature rise depends on the power density of the region through which the coolant flows, and can be adjusted by varying the flow rate Q . From the regional view, the spatial variation of power density makes the temperature rise of microchannels different from each other. On the other hand, the power consumption of a particular circuit cell determines the local heat flux q to be transferred to the coolant, which creates further temperature variations among circuit cells with different power densities. Therefore, to reduce the thermal gradient of microfluidic-cooled 3DIC, the microchannels need to be designed to partially offset (if not completely) the effect of power density variation and the disparity in coolant temperature distribution, without increasing the flow rates.

5.3.1 Microchannel Splitting by Splitter Placement

From (5.5), it is evident that to reduce the thermal gradient, R_{conv} needs to be made small at downstream regions and also hot spots that have high q 's. A novel microchannel splitting technique is proposed here to effectively adjust R_{conv} . Let the 3D stack be partitioned into $m \times n$ grids for thermal analysis, such that the microchannels are divided into m sections with equal section length $L_{ch_{ij}}$. As shown in Fig. 5.4, by depositing etching mask on the substrate to preserve n_{sp} parallel thin silicon splitters in any microchannel section during the etching process, the original wide channel can be divided into $n_{sp} + 1$ subchannels. As such, the hydraulic diameter D_h of the subchannels becomes smaller and the contact area A between the circuit cell and the channel fluid cell is substantially increased.

$$g_{conv} = hA \text{ and } h = k_f Nu / D_h \quad (5.7)$$

According to (5.7), R_{conv} can thus be effectively reduced by the placement of splitters at the desired sections of microchannels, and the reduction rate can be regulated by the number of splitters n_{sp} .

Figure 5.4: Illustration on splitter insertion for channel splitting

The maximum number of splitters $N_{sp,max}$ that can be placed in a microchannel section is constrained by the user-defined minimum channel width $W_{ch,min}$ and the

splitter width W_{sp} . To maximally balance the circuit temperature, $N_{sp,max}$ is set at the hottest sections. The number of splitters of other channel sections are then determined according to their heat generation relative to those sections. For implementation feasibility, the subchannels shall be wide enough to prevent high pressure drop. The splitters are deposited evenly inside the microchannel, with enough thickness for physical reliability and allowing the through silicon vias (TSVs) to punch through. In our case studies, we set $W_{ch,min} = 20\mu m$ and $W_{sp} = 10\mu m$ to provide passages for signal TSVs with diameters around $5\mu m$. The length of the splitters is set equal to the channel section length, except for those at transition regions where consecutive channel sections have different n_{sp} 's. As illustrated in Fig. 5.4, when there is a change in the number of splitters from Ch_{ij} to $Ch_{i(j+1)}$, the length of these splitters will be reduced to leave a gap between them so as to avoid uneven share of flow among the subchannels in $Ch_{i(j+1)}$. In our experiments, the splitters at transition regions are shortened by 10%. Algorithm 1 shows the proposed heuristic to optimize the placement of splitters for thermal balancing.

5.3.2 Microchannel Scaling for Redistribution of Coolant

Appropriate placement of channel splitters effectively reduces convective resistances to offset excessive temperature rise at potential hot spots. However, reducing channel width may significantly increase the pressure drop, as indicated by the following equation [269],

$$\Delta P = \frac{4}{3}\pi^2\rho vL\frac{(H_{ch}^2 + W_{ch}^2)}{H_{ch}^3 \cdot W_{ch}^3}Q \quad (5.8)$$

where ΔP denotes the pressure drop, ρ and v denotes the density and dynamic viscosity of the fluid, Q denotes the flow rate, L , H_{ch} and W_{ch} denote the length, height and width of the microchannel respectively. For most microfluid-cooled 3DIC, the coolant is supplied by only one pump. The fluid network is designed such that all the microchannels have the same pressure drop. As a result, coolant flow rate will be reduced from channels that have more splitters and added to channels with less splitters, which may contradict with the cooling needs. A sophisticated fluid network that has separate supply loop for each microchannel is infeasible with

Algorithm 5.2 Heuristic for optimal channel splitting

Function Splitting():

Input: Power Distribution, Cooling Setting

Output: Splitter quantity n_{sp} for every channel section Ch_{ij} : j^{th} section counting from the inlet of channel Ch_i $q_{ij,above}$: heat flux from circuit above Ch_{ij} $q_{ij,below}$: heat flux from circuit below Ch_{ij} **for** every channel section Ch_{ij} **do** $qsum_{ij} = qsum_{i(j-1)} + q_{ij,above} + q_{ij,below}$ where $qsum_{i0} = q_{i0,above} + q_{i0,below}$ **end for**Find $qsum_{max}$: maximum of $qsum_{ij}$ $N_{sp,max} = round\{W_{ch}/(W_{ch,min} + W_{sp})\}$ **for** every channel section Ch_{ij} **do** $n_{sp,ij} = round(N_{sp,max} \cdot qsum_{ij}/qsum_{max})$ **end for**

existing technology. Instead, a channel scaling technique is proposed to redistribute the coolants after channel splitting. Algorithm 5.3 shows the proposed channel scaling heuristic to match the power distribution.

Initially, all microchannels are assumed to be equally wide and uniformly distributed. After the channel splitting, more splitters are placed in channels passing through high power areas and their flow rates have been reduced substantially. These highly split microchannels will be expanded to accommodate more coolant while those unnecessarily wide channels will be shrunk to save coolant. The expansion of a channel is constrained by the minimum channel wall width $W_{wall,min}$, and the shrinking is limited by $W_{ch,min}$. Boundary values shall be used when there is a violation during channel scaling. Fortunately, changing channel width has a superlinear effect on the channel's flow rate as evinced by (5.8). By setting the initial channel width at medium level, and choosing a proper reference channel to limit

Algorithm 5.3 Heuristic for optimal channel scaling

Function Scaling():

Input: Power Distribution, Cooling Setting

Output: Adjusted channel width W_{ch} of every microchannel

 m : grid row number

Sort $qsum_{i(m-1)}$ for all i in ascending order

 $qsum_{mid}$: median of $qsum_{i(m-1)}$
 mid : index of $qsum_{mid}$

Define $f(i) = 1 / \sum_{j=0}^m \frac{(H_{ch,i}^2 + W_{ch,ij}^2)}{H_{ch,i}^3 W_{ch,ij}^3 \cdot (n_{sp,ij} + 1)}$
for every microchannel Ch_i **do**

 Solve $f(i) = f(mid)qsum_{i(m-1)}/qsum_{mid}$ for $W_{ch,i}$
end for

the extent of scaling, violation of constraints can usually be avoided to achieve a good result. For compatibility with bigger TSVs, $W_{wall,min} = 20\mu m$ is set in our experiments.

5.4 Validation of Proposed Thermal Balancing Techniques

In this section, the proposed thermal balancing techniques will be validated on a 4-layer multi-core testbench. The 3D stack integrates 16 Alpha-2 microprocessors with four of them on each layer. SPEC2000 applications *bzip*, *eon*, *fma*, *gcc* and *gzip* [254] are used as input to the processors and the micro-architecture-level power information is obtained from simulation by Wattch [252]. The 3DIC has a footprint of $32 \times 32mm^2$. The BEOL, Source and Substrate of the circuit layers are $12\mu m$, $2\mu m$ and $48\mu m$ in thickness, respectively. The microchannel has a height of $100\mu m$ and the default channel and channel wall width are both set to be $100\mu m$. All the experiments are based on simulations on a Linux server with the ease and flexibility to analyze the cooling performances for different parameters.

5.4.1 Thermal Balancing by Channel Clustering

To show the effect of the clustering algorithm, four rounds of random allocation of the benchmarks to cores and one round of average power assignment are performed to generate five different power distributions for the case study. For all the case studies, the steady state power density varies between $2.98W/cm^2$ to $68.95W/cm^2$ across the entire 3D stack.

Clustering of Microchannels

As described in Section 5.2, the total heat generation to be handled by each microchannel in the five study cases is determined. In Cases 1-4, the five benchmarks are randomly distributed to the 16 cores, while in Case 5, all the cores are assigned with an average power consumption which is the mean value of the five benchmarks. The last case is presented to simulate the practical case where clustering is performed with an average power consumption to cater for various runtime scenarios. The sorted power values in ascending order of magnitude are charted in Fig. 5.5. To be concise, each dot in the chart actually represents the power level of a channel bundle of five channels sharing the same cooling region.

The clustering results at different η values are summarized in Table 5.1, where the number of channels in each cluster is listed in ascending order of cooling demand. As the value of η decreases, the deviation in the total power dissipations among channels in the cluster reduces. It can be seen that more clusters are defined with a stringent variation control with coarse clustering when η is relaxed. To have the best cooling effect while not imposing excessive overhead, η is set to be 20% for Cases 1-4 and 10% for Case 5.

Uniform Cooling Effort v.s. Clustered Cooling Effort

With the above clustering results, the fixed total flow rate of $288ml/min$ are re-allocated to each cluster, as shown in Table 5.2, where the flow rate per channel allocated to each cluster is listed in ascending order of cooling demand. Here the upper limit for flow rate per channel is specified as $1.5ml/min$ corresponding to

Figure 5.5: Cooling demand of microchannels with different power distributions

Table 5.1: Clustering results with different η values

Case No.	η	N	Number of channels per cluster
Case 1	20%	3	300, 60, 120
	10%	5	300, 60, 20, 60, 40
	5%	7	280, 20, 60, 20, 60, 20, 20
Case 2	20%	3	360, 100, 20
	10%	6	300, 60, 20, 50, 30, 20
	5%	9	40, 200, 60, 40, 20, 20, 50, 30, 20
Case 3	20%	3	360, 100, 20
	10%	6	300, 60, 20, 50, 30, 20
	5%	9	40, 200, 60, 40, 20, 20, 50, 30, 20
Case 4	20%	3	360, 100, 20
	10%	6	200, 100, 60, 40, 60, 20
	5%	10	200, 100, 40, 20, 20, 10, 25, 25, 20
Case 5	20%	2	72, 24
	10%	4	300, 60, 40, 80
	5%	6	100, 300, 60, 40, 20, 60

about 90KPa pressure drop according to the method of calculation in [265].

Table 5.2: Reallocation of cooling resources

Case No.	N	Flow Rate per channel (ml/min)
Case 1	3	0.348, 0.798, 1.131
Case 2	3	0.417, 1.126, 1.255
Case 3	3	0.417, 1.128, 1.250
Case 4	3	0.412, 1.139, 1.281
Case 5	4	0.403, 0.696, 0.996, 1.068

Fig. 5.6 depicts the top layer's steady state thermal maps in Case 1 under both

uniform and non-uniform flow rate coolings by channel clustering. It is clear that with clustered microfluidic cooling, the overall temperature of the 3D chip is much lower. Numerical results for all the study cases are summarized in Table 5.3. Note that for Cases 1-4, the power distribution and channel clustering are ideally matched as the same power profile is used for both clustering and thermal simulation. While for Case 5, power profiles of Cases 1-4 are used for thermal simulation after clustering based on average power consumption. Only the maximum temperature is recorded because the minimum temperature is always close to $300K$ at the inlet area. From the table, it is clear that clustered cooling method is effective in reducing the peak temperature under both ideal and practical cases. On average, a drop of $6.22^{\circ}C$ in the maximum temperature is observed.

Figure 5.6: Temperature distribution comparison: thermal maps of the top layer for Case 1 under uniform flow rate cooling and reallocated clustered flow rate cooling with the same total flow rate, coolant flows from north (upper) to south (lower) direction.

Minimization of Cooling Effort under Thermal Constraint

Based on the steady state simulation results, the total cooling effort can be further minimized by the proposed minimization steps presented in Section 5.2.3. By finding

Table 5.3: Comparison of the maximum temperatures under uniform and clustered coolings

Case No.	Uniform(K)	Clustered(Case 5)(K)	δT (K)
Case 1	333.62	326.40(327.48)	6.68
Case 2	332.02	326.64(326.02)	5.69
Case 3	333.63	327.45(327.50)	6.16
Case 4	333.94	327.49(327.68)	6.36

the hottest point of the 3D stack under clustered microfluidic cooling and its nearby channel temperatures, the scaling factors φ are calculated and the steady state thermal conditions are re-simulated after reducing every channel's flow rate by $(1 - \varphi)$. The maximum temperatures after minimization are reported in Table 5.4. It can be seen that the minimization method accurately predicts the flow rate demand under certain thermal constraint with a suitable safety margin in both ideal and practical cases. For the purpose of comparison, uniform flow rate thermal management is also implemented by a recursive decrease of flow rate until the maximum temperature of the 3D stack approaches the maximum temperature limit of $358.15K(85^\circ C)$. For a fair comparison, the maximum temperature after minimization is further pushed towards the thermal limit by reducing the flow rate proportionally. Some key findings are summarized in Table 5.5, where the total flow rate is measured in ml/min . It is shown that with clustered cooling scheme, the total flow rate can be saved up to 21.8% under the same thermal constraint. Note that the cooling power of micropumps can be formulated as $P_{cool} = \Delta P Q$ where ΔP denotes the pressure drop along the microchannel and Q denote the flow rate. Together with (5.8), $P_{cool} \propto Q^2$ can be derived. The power savings can thus be obtained by summing up the power consumption of every microchannel and then comparing with the uniform cooling case. A maximum of 22.8% cooling power saving has been achieved for our case studies. It can be seen that the cooling effort (power) has been effectively distributed to the places where it is needed most by controlling the flow rate distribution.

In addition, Fig. 5.7 compares the top layer thermal distribution of two mini-

Table 5.4: Maximum temperature after minimization of cooling effort

Case No.	φ (Case 5)	Max Temperature(Case 5)(K)
Case 1	0.35 (0.38)	354.06 (352.84)
Case 2	0.37 (0.36)	352.52 (352.16)
Case 3	0.36 (0.38)	354.30 (352.72)
Case 4	0.34 (0.36)	356.00 (353.99)

Table 5.5: Comparison of cooling effort under thermal constraint

Case No.		Case 1	Case 2	Case 3	Case 4
$T_{max}(K)$	Uniform	357.28	357.73	357.27	358.01
	Cluster	357.85	357.43	358.01	358.10
	Case 5	357.69	358.40	357.55	357.62
Q_{total}	Uniform	119.95	112.58	120.01	117.72
	Cluster	91.93	94.70	94.97	93.42
	Case 5	95.69	90.24	96.02	95.03
Flow Rate Saving		21.8%	17.9%	20.4%	19.96%
Power Saving		22.8%	16.0%	21.3%	19.89%

mization methods for Case 1. With the uniform cooling scheme, most of the chip area is over-cooled to keep the center hot spot within threshold. Over cooling wastes unnecessary cooling powers as it does not help to reduce the thermal gradient. The over-cooled regions may exhibit disparate electrical properties, and create unexpected timing mismatch between the circuits at hot spot. Note that although the thermal map of our proposed method exhibits high temperature in the downstream region, none of the spots exceeds the thermal limit. Furthermore, with uniform flow rate, there is no clue on how to adjust the flow rate to maintain the maximum temperature just below the thermal limit (hence it was accomplished by recursive trial and error). In contrast, our proposed method provides a logical way to scale flow rates according to the maximum allowable temperature. This is useful for runtime

thermal management controller.

Figure 5.7: Temperature distribution comparison: top layer thermal maps of Case 1 using the proposed minimization method and uniform flow rate minimization with the same thermal limit, coolant flows from north (upper) to south (lower) direction.

5.4.2 Thermal Balancing by Channel Splitting and Scaling

In this part of the experiment, the total coolant flow rate is fixed at $288\text{ml}/\text{min}$ to maintain the circuit temperature.

Thermal Effect of Channel Splitting

To avoid disturbing the flow rate distribution while examining the effect of channel splitting, a special power distribution is assumed where the top and bottom circuit layers have uniform power densities of $12.5\text{W}/\text{cm}^2$ and inner layers have twice the power densities. In this way, the splitting scheme will be the same for every channel and the flow rate will remain uniformly distributed. Fig. 5.8 compares the convective resistances of a channel before and after channel splitting, together with the change of circuit temperature along the channel.

It can be seen that both the convective resistances and circuit temperature have been effectively reduced by channel splitting. In fact, reductions of 7.11K in the

Figure 5.8: Comparison of the convective resistances of a channel and its surrounding circuit temperatures before and after channel splitting.

maximum temperature and $7.12K$ in the thermal gradient (difference between the maximum and minimum circuit temperatures across the entire 3D stack) are observed for this case.

Thermal Gradient Reduction with Channel Splitting and Scaling

To investigate the effectiveness of our proposed techniques under different power conditions, four sets of practical power distribution are generated by feeding different applications to the microprocessors. Three different microchannel design strategies, namely splitting only, scaling only and combined splitting and scaling, are then applied. The resulting maximum temperatures and thermal gradients are shown in Table 5.6.

As expected, due to the redistribution of coolants, splitting the channel alone worsens the results. By applying merely channel scaling, the peak temperature and temperature gradient are on average reduced by $8.28K$ and $8.77K$, respectively. However, a further reduction of $4.56K$ and $4.42K$ from the maximum temperature and thermal gradient respectively can be achieved if scaling is applied after splitting. This can be explained by (5.5) and (5.6) that $\delta T_{channel}$ is reduced due to the matching of flow rate with heat generation by channel scaling. On top of that, $\delta T_{circuit}$

Table 5.6: Comparison of maximum temperatures (T_{max}) and thermal gradient (ΔT), measured in *Kelvin* under different microchannel design techniques

Power Set\Strategy		Default	Splitting	Scaling	Combined
Set A	T_{max}	341.08	352.29	332.79	328.46
	ΔT	40.25	51.63	31.51	27.40
Set B	T_{max}	339.11	350.60	331.61	327.12
	ΔT	38.25	49.92	30.33	25.96
Set C	T_{max}	341.09	352.15	332.68	328.27
	ΔT	40.21	51.47	31.33	27.10
Set D	T_{max}	342.34	352.78	333.44	328.41
	ΔT	41.46	52.11	31.94	26.96
Average Reduction			-11.05	8.28	12.84
			-11.24	8.77	13.19

is further suppressed by channel splitting through the reduction of convective resistances. As $\delta T_{channel}$ dominates the temperature rise in 3DIC, channel scaling usually contributes more than channel splitting to the temperature reduction. Fig. 5.9 compares the top layer temperature distributions under default condition and combined channel design strategy for Power Set A. Several hotspots appear in the default case. In contrast, the temperature distribution after splitting and scaling is much flatter with an overall lower temperature rise.

5.5 Summary

This chapter has proposed two novel techniques to reduce the large thermal gradient that was commonly seen on microfluid-cooled 3DICs. The first technique is the channel clustering method, where the microchannels are divided into groups called channel clusters based on the cooling demand, and appropriate flow rate settings are then assigned to every cluster. A channel clustering algorithm has been developed

Figure 5.9: Temperature distribution comparison: top layer thermal maps of Set A (a) under default microchannel setting and (b) with channel splitting and scaling. Coolant flows from north (upper) to south (lower) direction.

to guide such grouping and flow rate allocation process for optimal results. Experiment results show that this technique effectively reduces the peak temperature and thermal gradients. In addition, the proposed cooling effort minimization technique has saved up to 21.8% of total flow rate and 22.8% of cooling power compared to the uniform flow rate cooling without exceeding the pre-specified thermal limit of the 3D stacks.

The second technique is an augmented channel splitting and scaling technique for microchannel design. This method reduces the thermal gradients from two perspectives. One is to re-allocate the coolant resource to match with the power distribution by scaling the channel width. The other is to mitigate the temperature rise along coolant flow direction by adding channel splitters to reduce the convective resistances. Experiment results from the 4-layer 3D multi-core testbench have corroborated our expectations with an average thermal gradient reduction of $13.19K$. The reduced thermal gradient can potentially improve circuit performance and reliability from several aspects. From design aspect, large thermal gradient imposes significant constraint on the component selection and placement process. Extensive effort needs to be spent on the thermal simulation and timing verification. With

reduced thermal gradient, the design space is directly expanded with more options thus simplifies the design process and may lead to better solutions. From operation aspect, smaller thermal gradient causes only a limited thermal-induced device characteristic variation, thus reduces the chance of thermal-induced timing violation and malfunction. Finally, as the failure mechanisms such as fatigue and plastic deformation are closely related to the thermal cycles, a prolonged life span of 3DIC chips can be expected with reduced thermal gradients.

The two proposed techniques shared one common underlying optimization principle that the flow rate should be allocated to match with the cooling demand distribution. The channel cluster method adopted such optimization actively with complicated coolant delivery system including micro-pump, valves and customized fluid network. In contrast, the flow rate redistribution is achieved passively by the channel splitting and scaling method. Only customized etching masks are needed to insert the splitters and allow the variable microchannel width. Therefore, compared to the channel clustering technique and previous works [190] in literature, the channel splitting and scaling method incurs no noticeable overheads and is more practical for implementation.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

Much of the dramatic improvement in the integrated circuit performances over the last several decades has been attributed to the increased levels of integration. Device scaling enables more devices to be packed into a single die and therefore more functionality to be offered at a lower cost. Due to subthreshold leakage, short channel effects, gate induced drain leakage, gate tunneling current, time-dependent dielectric breakdown and hot carrier effects, it is predicted that the limit of device scaling will soon be reached. Furthermore, scaling device alone without simultaneously scaling the interconnects that provide for communications between devices will have a severe impact on the overall circuit delay. By expanding circuit integration in the vertical direction, three dimensional integrated circuit (3DIC) offers a commercially viable solution to overcome barriers in interconnect scaling, thereby sustaining the continued higher performance with increased device density, greater flexibility in routing the signals, power and clock based on the same CMOS technology. The fast vertical links such as TSVs are promising to mitigate bandwidth-induced throughput limitations such as the “memory wall” effect. The fabrication process of bonding separately processed wafers has also enabled the heterogeneous integration of disparate technologies on the same chip to reduce the form factor. However, despite all the potential benefits, 3DICs are prone to thermal induced hazards. Due to the stacking structure, 3DIC typically exhibits high power density and thermal re-

sistance which impose challenges on maintaining the circuit operation under safe temperature. New thermal management schemes at design and runtime need to be proposed to guarantee the safe and reliable operation of 3DIC chip.

To provide thermal evaluation and prediction of 3DICs, fast and accurate thermal modeling is almost always needed for any thermal-aware design or runtime algorithms. A compact thermal model has been proposed in this thesis which is capable of simulating the steady state temperature of 3DICs with both heatsink-based air cooling and microfluidic cooling through microchannels. Unlike state-of-the-art thermal simulators, the proposed model considers anisotropic thermal effect of TSVs and entrance effect of microchannels, which produces simulation results closer to the real physical model. The proposed thermal model has adopted the finite difference approach by building the equivalent thermal resistive networks. Sparse matrix solvers is invoked to solve the circuit problem to achieve the best computational efficiency. The model has been verified against commercial simulator COMSOL [219] and the state-of-the-art compact thermal models, Hotspot [222] and 3D-ICE [207]. It was shown that for the case of air-cooled designs, the proposed model produces similarly accurate results as Hotspot. While for cases of microfluid-cooled designs, the proposed model has achieved better accuracy than 3D-ICE in terms of the magnitude and distribution of temperatures.

With the help of the developed thermal model, runtime thermal management problem of 3D multi-processor system-on-chip (MPSoC) has been investigated. A proactive thermal management scheme has been proposed to adaptively adjust the coolant flow rate according to the real-time cooling demands. The complete scheme consists of three parts, namely power sensing, cooling demand prediction and flow rate adjustment. To reduce the implantation cost of thermal sensors, runtime system profile has been obtained by software-sensing methods using the architectural level power simulators, Wattch [252] and Cacti [253]. The historical power values are used to predict the future power consumption by assuming an autoregressive correlation and then further corrected by Kalman filter. Based on the predicted future workload, the proposed thermal model is invoked to estimate the future system temperature so that potential thermal spikes could be anticipated. The predicted

future system temperature profile is then used as the cooling demand to guide the flow rate adjustment process. A fine-grained flow rate adjustment has been proposed to provide a customized cooling through intuitively clustered microchannels. Experiment results showed that the proposed thermal management scheme effectively maintained the system temperature under the maximum threshold, and a 27.4% flow rate reduction was achieved by fine-grained flow rate control in comparison with the uniform flow rate distribution.

Upon analyzing abundant simulation results, it was found that microfluidic cooling typically introduces large lateral thermal gradients. This problem was mainly caused by the rising coolant temperature along the flow direction and highly unbalanced circuit power distribution. Two design time strategies have been proposed to tackle this problem. The first strategy matches the cooling effort and spatially non-uniform cooling demand using the concept of “channel clusters”. A channel cluster is defined as a group of microchannels that dissipate similar amount of heat. The coolant flow rate is distributed to the channel clusters according to their respective cooling needs. It was shown that only a small number of channel clusters are needed to effectively reduce the peak temperature. The total flow rate was also reduced by up to 21.8% under the same peak temperature constraint. The second proposed strategy is a set of microchannel design algorithms called channel splitting and scaling techniques. The channel splitting technique aims to decrease the local convective resistance of microchannels by inserting channel splitters. With more splitters at the downstream, the coolant temperature rise can be partially offset. The channel scaling technique then matches the flow rate distribution with the power distribution at the granularity of each channel by adjusting the channel width. Experiment results proved the effectiveness of the proposed techniques and the thermal gradient was reduced by 13.19°C on average. One key advantage of the proposed channel splitting and scaling techniques over the channel modulation method [190] is their implementation feasibility. Only customized masks are needed during microchannel etching process, which incur negligible overheads.

6.2 Recommended Future Work

Based on existing PhD research work, the following relevant topics are suggested for further exploration.

6.2.1 Investigation on Channel Splitter Placement for Simultaneous Thermal and Routing Optimization

In microfluid-cooled 3DICs, the microchannels are seen as blockages where TSVs are not allowed to punch through. Instead, TSVs have to be routed through the channel walls which restrict their placement. The final routing result may be sub-optimal due to such blockage effect. The channel splitters may be utilized to mitigate this problem. In Chapter 5, channel splitters have been proposed to be inserted at microchannel sections where the convective resistance needs to be reduced. In fact, the channel splitters can be used not only for thermal control purpose but also to ease the placement of TSVs. By inserting the splitters at the desirable locations, passages for TSVs can be provided. However, since thermal and TSV routing optimization are not necessarily correlated, they may produce conflicting optimal splitter locations. Therefore, a trade-off needs to be established which is worth further investigation.

6.2.2 3D Clock Tree Synthesis Considering Microchannel-Induced Thermal Gradient and Congestions

Clock tree is an essential routing network in synchronous circuits. The primary design objective of clock tree is to deliver clock signals to all the clock sinks simultaneously or with a bounded skew, while minimizing the routing wire length and hence the power consumption. Previous works have extended algorithms such as 3D-MMM [270] and NN-3D [271] for the topology generation and DME [272] algorithm for 3D clock tree embedding. Thermal [270] and testability [273] issues have been discussed and fault-tolerant designs [274] were also presented. However, none of them addressed 3D clock tree synthesis problem for microfluid-cooled 3DICs. The microfluidic cooling imposes two unique challenges to 3D clock tree synthesis,

namely the large lateral thermal gradient and routing congestions due to the microchannels. Fault-tolerant design strategies are needed to address these problems for robust clock tree synthesis.

Author's Publications

Journal Publications

- [275] H. Qian, C. H. Chang and H. Yu, "An efficient channel clustering and flow rate allocation algorithm for non-uniform microfluidic cooling of 3D integrated circuits", *Integration, the VLSI Journal*, vol. 46, no. 1, pp. 57-68, Dec. 2011.
- [276] H. Qian, X. Huang, H. Yu and C. H. Chang, "Cyber-physical thermal management of 3D multi-core cache processor system with microfluidic cooling", *ASP Journal of Low Power Electronics*, vol. 7, no. 1, pp. 110-121, Feb. 2011.

Conference Publications

- [277] H. Qian and C. H. Chang, "Microchannel splitting and scaling for thermal balancing of liquid-cooled 3DIC", *Proc. Int. Symp. Circuits and Syst.*, Beijing, PRC, 19-23 May 2013, pp. 801-804.
- [278] H. Qian, H. Liang, C. H. Chang, W. Zhang and H. Yu, "Thermal simulator of 3D-IC with modeling of anisotropic TSV conductance and microchannel entrance effects", *Proc. Asia and South Pacific Design Autom. Conf.*, Yokohama, Japan, 22-25 Jan. 2013, pp. 485-490.
- [279] H. Qian, X. Huang, H. Yu and C. H. Chang, "Real-time thermal management of 3D multi-core system with fine-grained cooling control", *Proc. IEEE Int. 3D System Integration Conf.*, Munich, Germany, 16-18 Nov. 2010, pp. 1-6.
- [280] J. Chen, C. H. Chang and H. Qian, "New power index model for switching

power analysis from adder graph of FIR filter", *Proc. Int. Symp. Circuits and Syst.*, Taipei, Taiwan, 24-27 May 2009, pp. 2197-2200.

Bibliography

- [1] American Physical Society, *This Month in Physics History–November 17 - December 23, 1947: Invention of the First Transistor*, Available: <http://www.aps.org/publications/apsnews/200011/history.cfm>.
- [2] J. S. Kilby, “Invention of the integrated circuit ”, *IEEE Trans. on Electron Devices*, vol. 23, no. 7, pp. 648-654, July 1976.
- [3] M. R. Betker, J. S. Fernando and S. P. Whalen, “The history of the microprocessor”, *Bell Labs Technical Journal*, Lucent Technologies Inc., Autumn 1997.
- [4] Intel[®] Corp.: Microprocessor Quick Reference Guide, Available: <http://www.intel.com/pressroom/kits/quickreffam.htm>.
- [5] Wikipedia: Transistor count, Available: http://en.wikipedia.org/wiki/Transistor_count.
- [6] G. E. Moore, “Cramming more components onto integrated circuits”, *Proc. IEEE*, vol. 86, no. 1, pp. 82-85, Jan. 1998.
- [7] S.C. Sun, “Process technologies for advanced metallization and interconnect systems”, *Proc. Int. Electron Device Meeting*, Washington, DC, 7-10 Dec. 1997, pp. 765-768.
- [8] M. T. Bohr, “Interconnect scaling-the real limiter to high performance ULSI”, *Proc. Int. Electron Device Meeting*, Washington, DC, 10-13 Dec. 1995, pp. 241-244.
- [9] S. Saxena, C. Hess, H. Karbasi, A. Rossoni, S. Tonello, P. McNamara, S. Lucherini, S. Minehane, C. Dolainsky and M. Quarantelli, “Variation in transistor performance and leakage in nanometer-scale technologies”, *IEEE Trans. on Electron Devices*, vol. 55, no. 1, pp. 131-144, Jan. 2008.
- [10] G. S. May and C. J. Spanos, *Fundamentals of semiconductor manufacturing and process control*, John Wiley & Sons, Inc., Hoboken, New Jersey, 2006.

- [11] K. K. Kim, Y-B Kim, N. Park and M. Choi, "Leakage minimization technique for nanoscale CMOS VLSI", *IEEE Design & Test of Computers*, vol. 24, no. 4, pp. 322-330, July-Aug. 2007.
 - [12] W. Gruener, "Intel to sell ivy bridge late in Q4 2011", Available: <http://www.tomshardware.com/news/ivy-bridge-processor-release-22nm-3d-transistor,13753.html>, 19 Oct. 2011.
 - [13] ITRS: International Technology Roadmap for Semiconductors, Available: <http://www.itrs.net>, 2011 Edition.
 - [14] M. F. L. D. Volder, S. H. Tawfick, R. H. Baughman and A. J. Hart, "Carbon nanotubes: present and future commercial applications", *Science*, vol. 339, no. 6119, pp. 535-539, Feb. 2013.
 - [15] A. K. Geim and K. S. Novoselov, "The rise of graphene", *Nature Materials*, vol. 6, no. 3, pp. 183-191, 2007.
 - [16] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information*, Cambridge University Press, 23 Oct. 2000.
 - [17] J. M. Early, "Speed, power and component density in multielement high-speed logic systems", *Proc. IEEE Int. Solid-State Circuits Conf.*, Philadelphia, PA, 10-12 Feb 1960, pp. 78-79.
 - [18] H. Goto: Intel CPU die-size and microarchitecture, Available: http://pc.watch.impress.co.jp/docs/2008/0402/kaigai01_07.pdf, 2008.
 - [19] Wikipedia: List of CPU power dissipation figures, Available: http://en.wikipedia.org/wiki/List_of_CPU_power_dissipation_figures.
 - [20] M. Huang, Z. Suo and Q. Ma, "Plastic ratcheting induced cracks in thin film structures", *J. Mech. and Phy. of Solids*, vol. 50, no. 5, pp. 1079-1098, May 2002.
 - [21] B. Z. Hong and L. G. Burrell, "Modeling thermally induced viscoplastic deformation and low cycle fatigue of CBGA solder joints in a surface mount package", *IEEE Trans. on Components, Packaging, and Manufacturing Tech., Part A*, vol. 20, no. 3, pp. 280-285, Sept. 1997.
 - [22] S. Garg and D. Marculescu, "System-level process variability analysis and mitigation for 3D MPSoCs", *Proc. Design, Automation and Test in Europe Conf.*, Nice, France, 20-24 Apr. 2009, pp. 604-609.
-

- [23] S. Garg and D. Marculescu, "3D GCP - An analytical model for the impact of process variations on the critical path delay distribution of 3D ICs", *Proc. IEEE Int. Symp. Quality Electronic Design*, San Jose, CA, 16-18 Mar. 2009, pp. 147-155.
 - [24] M. Koyanagi, H. Sunami, N. Hashimoto and M. Ashikawa, "Novel high density, stacked capacitor MOS RAM", *Proc. Int. Electron Devices Meeting*, Washington, DC, 4-6 Dec. 1978, pp. 348-351.
 - [25] J. F. Gibbons and K. F. Lee, "One-gate-wide CMOS inverter on laser-recrystallized polysilicon", *IEEE Electron Device Lett.*, vol. EDL-1, no. 6, pp. 117-118, June 1980.
 - [26] Y. Akasaka, "Three-dimensional IC trends", *Proc. IEEE*, vol. 74, no. 12, pp. 1703-1714, Dec. 1986.
 - [27] Y. Akasaka, T. Nishimura and H. Nakata, "High-speed CMOS devices fabricated on laser-recrystallized polycrystalline silicon island", *Proc. Symp. VLSI Technol.*, Maui, HI, 13-15 Sept. 1983, pp.48-49.
 - [28] J. P. Colinge, E. Demoulin, D. Bensahel and G. Arvert, "Use of selective annealing for growing very large grain silicon on insulator films", *Appl. Phys. Lett.*, vol. 41, no. 4, pp.346-347, Aug. 1982.
 - [29] Y. Inoue, K. Sugahara, S. Kusunoki, M. Nakaya, T. Nishimura, Y. Horiba, Y. Akasaka and H. Nakata, "A three-dimensional static RAM", *IEEE Electron Device Lett.*, vol. 7, no. 5, pp. 327-329, May 1986.
 - [30] S. Hirose, T. Nishimura, K. Sugahara, S. Kusunoki, Y. Akasaka and N. Tsubouchi, "A 10-bit linear image sensor fabricated in double-active layers", *Proc. Symp. VLSI Technol.*, Kobe, Japan, 14-16 May 1985, pp. 34-35.
 - [31] K. Kioi, T. Shinozaki, S. Toyoyama, K. Shirakawa, K. Ohtake and S. Tsuchimoto, "Design and implementation of a 3D-LSI image sensing processor", *IEEE J. Solid-State Circuits*, vol. 27, no. 8, pp. 1130-1140, 1992.
 - [32] H. Kurino, K. W. Lee, T. Nakamura, K. Sakuma, N. Miyakawa, H. Shimazutsu, K. Kim and M. Koyanagi, "Intelligent image sensor chip with three dimensions structure", *Proc. IEEE Int. Electron Devices Meeting*, Washington, DC, 5-8 Dec. 1999, pp.879-882.
 - [33] Y. Hayashi, S. Wada, K. Kajiyana, K. Oyama, R. Koh, S. Takahashi and T. Kunio, "Fabrication of three-dimensional IC using 'cumulatively bonded IC' (CUBIC) technology", *Proc. Symp. VLSI Technol.*, Honolulu, HI, 4-7 June 1990, pp.95-96.
-

- [34] R. Williams and O. Marsh, "Future WSI technology: stacked monolithic WSI", *IEEE Trans. Compon., Hybrids Mfg. Tech.*, vol. 16, no. 7, pp. 610-615, 1993.
 - [35] M. Koyanagi, H. Kurino, K. W. Lee; K. Sakuma, N. Miyakawa and H. Itani, "Future system-on-silicon LSI chips", *IEEE Micro*, vol. 18, no. 4, pp. 17-22, Jul./Aug. 1998.
 - [36] S. Spiesshoefer, Z. Rahman, G. Vangara, S. Polamreddy, S. Burkett, and L. Schaper, "Process integration for through-silicon vias", *J. Vac. Sci. Technol. A: Vacuum, Surfaces, and Films*, vol. 23, no. 4, pp. 824-829, Jul. 2005.
 - [37] L. W. Schaper , S. L. Burkett , S. Spiesshoefer , G. V. Vangara , Z. Rahman and S. Polamreddy, "Architectural implications and process development of 3-D VLSI Z-axis interconnects using through silicon Via", *IEEE Trans. Adv. Packag.*, vol. 28, no. 3, pp.356-366, Aug. 2005.
 - [38] K. Takahashi, M. Hoshino, H. Yonemura, M. Tomisaka, M. Sunohara, M. Tanioka, T. Sato, K. Kojima, and H. Terao, "Development of advanced 3-D chip stacking technology with ultra-fine interconnection", *Proc. Electronic Components and Technol. Conf.*, Orlando, FL, 29 May - 1 Jun. 2001, pp. 541- 546.
 - [39] F. Liu, R. R. Yu, A. M. Young, J. P. Doyle, X. Wang, L. Shi, K.-N. Chen, X. Li, D. A. Dipaola, D. Brown, C. T. Ryan, J. A. Hagan, K. H. Wong, M. Lu, X. Gu, N. R. Klymko, E. D. Perfecto, A. G. Merryman, K. A. Kelly, S. Purushothaman, S. J. Koester, R. Wisnieff and W. Haensch, "A 300-mm wafer-level three-dimensional integration scheme using tungsten through-silicon via and hybrid Cu-adhesive bonding", *Proc. Int. Electron Devices Meeting*, San Francisco, CA, 15-17 Dec. 2008, pp. 1-4.
 - [40] G. Parès, N. Bresson, S. Minoret, V. Lapras, P. Brianceau, J. F. Lugand, R. Anciant, N. Sillon, "Through Silicon Via technology using tungsten metallization", *Proc. IEEE Int. Conf. IC Design & Technol.*, Kaohsiung, Taiwan, 2-4 May 2011, pp. 1-4.
 - [41] D. Bollmann et. al., "Three dimensional metallization for vertically integrated circuits", *Abstracts Booklet European Workshop Materials for Advanced Metallization*, Villard de Lans, France, 16-19 March 1997, pp. 94-98.
 - [42] IBM Thomas J. Watson Research Center, Available: <http://www.research.ibm.com/labs/watson/index.shtml>.
 - [43] Tezzaron® Semiconductor: Available: <http://www.tezzaron.com/>.
 - [44] J. W. Goodman, F. J. Leonberger, S. Y. Kung and R. A. Athale, "Optical interconnections for VLSI systems", *Proc. IEEE*, vol. 72, no. 7, pp. 850-866, Jul. 1984.
-

- [45] L. A. Hornak and S. K. Tewksbury, "On the feasibility of through-wafer optical interconnects for hybrid wafer-scale-integrated architectures", *IEEE Trans. Electron Devices*, vol. ED-34, no. 7, pp. 1557-1563, Jul. 1987.
 - [46] G. W. Turner, C. K. Chen, B.-Y. Tsaur and A. M. Waxman, "Through-wafer optical communication using monolithic InGaAs-on-Si LED's and monolithic PtSi-Si Schottky-barrier detectors", *IEEE Photon. Technol. Lett.*, vol. 3, no. 8, pp. 761-763, Aug. 1991.
 - [47] K. H. Calhoun, C. B. Camperi-Ginestet and N. M. Jokerst, "Vertical optical communication through stacked silicon wafers using hybrid monolithic thin film InGaAsP emitters and detectors", *IEEE Photon. Technol. Lett.*, vol. 5, no. 2, pp. 254-257, Feb. 1993.
 - [48] A. D. Norte, A. E. Willner, W. Shieh and A. R. Tanguay, "Multiple- layer optical interconnections using through-wafer hollow-dielectric-waveguide vias", *IEEE Photon. Technol. Lett.*, vol. 6, no. 7, pp. 851-854, Jul. 1994.
 - [49] N. M. Jokerst, C. Camperi-Ginestet, B. Buchanan, S. Wilkinson and M. A. Brooke, "Communication through stacked silicon circuitry using integrated thin film InP-based emitters and detectors", *IEEE Photon. Technol. Lett.*, vol. 7, no. 9, pp. 1028-1030, Sept. 1995.
 - [50] M. Brooke, N. M. Jokerst, S. W. Bond and D. S. Wills, "A comparison of CMOS chip to chip interconnections: Vertical optical through-Si links versus high density electrical interconnect", *Proc. IEEE Annu. Meeting Lasers and Electro-Optics Society*, Orlando, FL, 1-4 Dec 1998, pp. 186-187.
 - [51] S.W. Bond, O. Vendier, M. Lee, S. Jung, M. Vrazel, A. Lopez-Lagunas, S. Chai, G. Dagnall, M. Brooke, N. M. Jokerst, D. S. Wills and A. Brown, "A three-layer 3D silicon system using through-Si vertical optical interconnections and Si CMOS hybrid building blocks", *IEEE J. Selected Topics in Quantum Electronics*, vol. 5, no. 2, pp. 276-286, Mar./Apr. 1999.
 - [52] S. A. Kühn, M. Kleiner, R. Thewes and W. Weber, "Vertical signal transmission in three-dimensional integrated circuits by capacitive coupling", *Proc. IEEE Int. Symp. Circuits and Syst.*, Seattle, WA, 30 Apr-3 May 1995, pp.37-40.
 - [53] S. F. Al-sarawi, D. Abbott and P. D. Franzon, "A review of 3-D packaging technology", *IEEE Trans. Compon. Packag. Manuf. Technol. B*, vol. 21, no. 1, pp. 2-14, Feb 1998.
 - [54] N. Miura, D. Mizoguchi, T. Sakurai and T. Kuroda, "Analysis and design of inductive coupling and transceiver circuit for inductive inter-chip wireless superconnect", *IEEE J. Solid-State Circuits*, vol. 40, no. 4, pp. 829-837, Apr. 2005.
-

- [55] N. Miura, D. Mizoguchi, M. Inoue, K. Niitsu, Y. Nakagawa, M. Tago, M. Fukaishi, T. Sakurai and T. Kuroda, "A 1Tb/s 3W inductive-coupling transceiver for inter-chip clock and data link", *Dig. Tech. Papers IEEE Int. Solid-State Circuits Conf.*, San Francisco, CA, 6-9 Feb. 2006, pp. 1676-1685.
 - [56] M. Saen, K. Osada, Y. Okuma, K. Niitsu, Y. Shimazaki, Y. Sugimori, Y. Kohama, K. Kasuga, I. Nonomura, N. Irie, T. Hattori, A. Hasegawa and T. Kuroda, "3-D system integration of processor and multi-stacked SRAMs using inductive-coupling link", *IEEE J. Solid-State Circuits*, vol. 45, no. 4, Apr. 2010.
 - [57] S. Han and D. D. Wentzloff, "Performance improvement of resonant inductive coupling for wireless 3D IC interconnect", *Proc. IEEE Int. Symp. Antennas and Propagation Society*, Toronto, ON, 11-17 Jul. 2010, pp. 1-4.
 - [58] J. U. Knickerbocker, P. S. Andry, B. Dang, R. R. Horton, M. J. Interrante, C. S. Patel, R. J. Polastre, K. Sakuma, R. Sirdeshmukh, E. J. Sprogis, S. M. Sri-Jayantha, A. M. Stephens, A. W. Topol, C. K. Tsang, B. C. Webb and S. L. Wright, "Three-dimensional silicon integration", *IBM J. Res. Develop.*, vol. 52, no. 6, pp.553-569, Nov. 2008.
 - [59] M. Sunohara, T. Fujii, M. Hoshino, H. Yonemura, M. Tomisaka and K. Takahashi, "Development of wafer thinning and double-sided bumping technologies for the three-dimensional stacked LSI", *Proc. Electron. Compon. and Technol. Conf.*, San Diego, CA, 31 May 2002, pp. 238-245.
 - [60] A. W. Topol et. al., "Enabling SOI-based assembly technology for three-dimensional (3d) integrated circuits (ICs)", *Proc. IEEE Electron Devices Meeting*, Washington, DC, 5 Dec. 2005, pp.352-355.
 - [61] A. R. Mirza, "One micron precision, wafer-level aligned bonding for interconnect, MEMS and packaging applications", *Proc. Electron. Compon. & Technol. Conf.*, Las Vegas, NV, 21-24 May 2000, pp. 676-680.
 - [62] S. H. Lee, K.-N. Chen and J. J.-Q. Lu, "Wafer-to-wafer alignment for three-dimensional integration: a review", *J. Microelectromech. Syst.*, vol. 20, no. 4, pp. 885-898, Aug. 2011.
 - [63] P. R. Morrow, C.-M. Park, S. Ramanathan, M. J. Koblinsky and M. Harmes, "Three-dimensional wafer stacking via CuCu bonding integrated with 65-nm strained-Si/low-k CMOS technology", *IEEE Electron Device Lett.*, vol. 27, no. 5, pp. 335-337, May 2006.
 - [64] B. Swinnen, W. Ruythooren, P. De Moor, L. Bogaerts, L. Carbonell, K. De Munck, B. Eyckens, S. Stoukatch, D. S. Tezcan, Z. Tókei, J. Vaes, J. Van Aelst, and E. Beyne, "3D
-

- integration by Cu-Cu thermo-compression bonding of extremely thinned bulk-Si die containing 10 m pitch through-Si vias”, *Proc. Int. Electron Devices Meeting*, San Francisco, CA, 11-13 Dec. 2006, pp. 1-4.
- [65] Y. Tomita, M. Tago, Y. Nemoto and K. Takahashi, “Copper bump bonding with electroless metal cap on 3 dimensional stacked structure”, *Proc. Electron. Packag. Technol. Conf.*, Singapore, 5-7 Dec. 2000, pp. 286-291.
- [66] S. Pozder, J.-Q. Lu, Y. Kwon, S. Zollner, J. Yu, J. J. McMahon, T. S. Cale, K. Yu and R. J. Gutmann, “Back-end compatibility of bonding and thinning processes for a wafer-level 3D interconnect technology platform”, *Proc. IEEE Int. Interconnect Technol. Conf.*, Burlingame, CA, 7-9 June 2004, pp. 102-104.
- [67] M. Umemoto, K. Tanida, Y. Tomita, T. Takahashi and K. Takahashi, “Non-metallurgical bonding technology with super-narrow gap for 3D stacked LSI”, *Proc. Electron. Packag. Technol. Conf.*, Singapore, 10-12 Dec. 2002, pp. 285-288.
- [68] J. Q. Lu, Y. Kwon, G. Rajagopalan, M. Gupta, J. McMahon, K. W. Lee, R. P. Kraft, J. F. McDonald, T. S. Cale, R. J. Gutmann, B. Xu, E. Eisenbraun, J. Castracane and A. Kaloyeros, “A wafer-scale 3D IC technology platform using dielectric bonding glues and copper damascene patterned inter-wafer interconnects”, *Proc. IEEE Int. Interconnect Technol. Conf.*, Burlingame, CA, 5 Jun. 2002, pp. 78-80.
- [69] F. Niklaus, G. Stemme, J.-Q. Lu and R. J. Gutmann, “Adhesive wafer bonding”, *J. Applied Physics*, vol. 99, no. 3, pp. 031101-031101-28, Feb. 2006.
- [70] K. Warner, J. Burns, C. Keast, R. Kunz, D. Lennon, A. Loomis, W. Mowers and D. Yost, “Low-Temperature oxide-bonded three-dimensional integrated circuits”, *Proc. IEEE Int. SOI Conf.*, Williamsburg, VA, 7-10 Oct. 2002, pp. 123-125.
- [71] C. S. Tan, K. N. Chen, A. Fan and R. Reif, “A back-to-face silicon layer stacking for three-dimensional integration”, *Proc. IEEE Int. SOI Conf.*, Honolulu, HI, 3-6 Oct. 2005, pp. 87-89.
- [72] T. Suga and R. Kondoh, “Low temperature bonding for 3D interconnects”, *Proc. IEEE Int. 3D Syst. Integration Conf.*, Osaka, Japan, 31 Jan.-2 Feb. 2012, pp. 1-4.
- [73] T. Suga, Y. Takahashi, H. Takagi, B. Gibbesch and G. Elssner, “Structure of Al-Al and Al- Si_3N_4 interfaces bonded at room temperature by means of the surface activation method”, *Acta Metallurgica et Materialia*, vol. 40, supplement, pp. S133-S137, 1992.
-

- [74] T. Suga, T. Fujiwaka and G. Sasaki, "Surface activated bonding and its application on micro-bonding at room temperature", *Proc. ISHM European Hybrid Microelectronics Conf.*, Nice, France, 2-4 Jun. 1993, pp. 314-321.
 - [75] H. Takagi, K. Kikuchi, R. Maeda, T. R. Chung and T. Suga, "Surface activated bonding of silicon wafers at room temperature", *Appl. Phys. Lett.*, vol. 68, no. 16, pp. 2222-2224, 1996.
 - [76] T. R. Chung, N. Hosoda, T. Suga and H. Takagi, "1.3 μm InGaAs/InP lasers on GaAs substrate fabricated by the surface activated wafer bonding method of room temperature", *Appl. Phys. Lett.*, vol. 72, no. 13, pp. 1565-1566, 1998.
 - [77] H. Takagi, R. Maeda, N. Hosoda and T. Suga, "Wafer-scale spontaneous bonding of silicon wafers by argon-beam surface activation at room temperature", *Sensors and Actuators A: Physical, Volume*, vol. 105, no. 1, pp. 98102, Jun. 2003.
 - [78] A. Shigetou, T. Itoh, M. Matsuo, N. Hayasaka, K. Okumura and T. Suga, "Bumpless interconnect through ultrafine Cu electrodes by means of surface-activated bonding (SAB) method", *IEEE Trans. Adv. Packag.*, vol. 29, no. 2, pp. 218-226, May 2006.
 - [79] A. Shigetou, T. Itoh, K. Sawada and T. Suga, "Bumpless Interconnect of 6-m-pitch Cu electrodes at room temperature", *IEEE Trans. Adv. Packag.*, vol. 31, no. 3, pp. 473-478, Aug. 2008.
 - [80] M. Igarashi, K. Usami, K. Nogami, F. Minam, Y. Kawasaki, T. Aoki, M. Takano, C. Mizuno, T. Ishikawa, M. Kanazawa, S. Sonoda, M. Ichida and N. Hatanaka, "A low-power design method using multiple supply voltages", *Proc. Int. Symp. Low Power Electronics and Design*, Monterey, CA, 18-20 Aug. 1997, pp. 36-41.
 - [81] S.-A. Yu, P.-Y. Huang and Y.-M. Lee, "A multiple supply voltage based power reduction method in 3-D ICs considering process variations and thermal effects", *Proc. Asia and South Pacific Design Autom. Conf.*, Yokohama, Japan, 19-22 Jan. 2009, pp. 55-60.
 - [82] S.-H. Whi and Y.-M. Lee, "Supply voltage assignment for power reduction in 3D ICs considering thermal effect and level shifter budget", *Proc. Int. Symp. VLSI Design, Automation and Test*, Hsinchu, Taiwan, 25-28 Apr. 2011, pp. 1-4.
 - [83] M. Keating, D. Flynn, R. Aitken, A. Gibbons and K. Shi, *Low Power Methodology Manual – For System-on-Chip Design*, Springer US, 2007.
 - [84] Y. Shin, J. Seomun, K.-M. Choi and T. Sakurai, "Power gating: Circuits, design methodologies, and best practice for standard-cell VLSI designs", *ACM Trans. Design Automation of Electron. Syst.*, vol. 15, no. 4, pp. 28:1-28:37, Sept. 2010.
-

- [85] J. Kathuria, M. Ayoubkhan and A. Noor, “A review of clock gating techniques”, *MIT Int. J. Electron. and Commun. Eng.*, vol. 1, no. 2, pp 106-114, Aug 2011.
 - [86] T. Ishihara and H. Yasuura, “Voltage scheduling problem for dynamically variable voltage processors”, *Proc. Int. Symp. on Low Power Electronics and Design*, Monterey, CA, 10-12 Aug. 1998, pp.197-202.
 - [87] G. Quan and X. Hu, “Minimum energy fixed-priority scheduling for variable voltage processors”, *Proc. Design Automation and Test in Europe*, Paris, France, 04-08 Mar. 2002, pp.782-787.
 - [88] I. Hong, G. Qu, M. Potkonjak, and M.B. Srivastava, “Synthesis techniques for low-power hard real-time systems on variable voltage processor”, *Proc. IEEE Real-Time Systems Symp.*, Madrid, Spain, Dec. 1998, pp.178-187.
 - [89] K. Choi, R. Soma and M. Pedram, “Dynamic voltage and frequency scaling based on workload decomposition”, *Int. Symp. Low Power Electronics and Design*, Newport Beach, CA, 11 Aug. 2004, pp.174-179.
 - [90] K. Choi, W. Lee, R. Soma and M. Pedram, “Dynamic voltage and frequency scaling under a precise energy model considering variable and fixed components of the system power dissipation”, *Proc. IEEE/ACM Int. Conf. Computer Aided Design*, San Jose, CA, 7-11 Nov. 2004, pp. 29-34.
 - [91] K. Choi, R. Soma and M. Pedram, “Fine-grained dynamic voltage and frequency scaling for precise energy and performance tradeoff based on the ratio of off-chip access to on-chip computation times”, *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 24, no. 1, pp. 18-28, Jan. 2005.
 - [92] A. K. Coskun, T. S. Rosing and K. C. Gross, “Proactive temperature balancing for low cost thermal management in MPSoCs”, *Proc. IEEE/ACM Int. Conf. Computer Aided Design*, San Jose, CA, 10-13 Nov. 2008, pp. 250-257.
 - [93] J. Lee and N. S. Kim, “Optimizing throughput of power- and thermal-constrained multicore processors using DVFS and per-core power-gating”, *ACM/IEEE Design Automation Conf.*, San Francisco, CA, 26-31 July 2009, pp. 47-50.
 - [94] C.-H. Hsu and U. Kremer, “Compiler-directed dynamic voltage scaling for memory-bound applications”, Dept. Comp. Sci., Rutgers Univ., NJ, Tech. Rep. DCS-TR-498, 2002.
 - [95] S. Murali, A. Mutapcic, D. Atienza, R. Gupta, S. Boyd and G. D. Micheli, “Temperature-aware processor frequency assignment for MPSoCs using convex optimization”,
-

- IEEE/ACM/IFIP Int. Conf. Hardware/Software Codesign and System Synthesis*, Salzburg, Austria, 30 Sept.-03 Oct. 2007, pp. 111 - 116.
 - [96] S. Zhang and K. S. Chatha, "System-level thermal aware design of applications with uncertain execution time", *Proc. IEEE/ACM Int. Conf. Computer Aided Design*, San Jose, CA, 10-13 Nov. 2008, pp. 242-249.
 - [97] C. Sun, L. Shang and R. P. Dick, "Three-dimensional multiprocessor system-on-chip thermal optimization", *IEEE/ACM/IFIP Int. Conf. Hardware/Software Codesign and System Synthesis*, Salzburg, Austria, 30 Sept.-3 Oct. 2007, pp. 117-122.
 - [98] C. Zhu, Z. Gu, L. Shang, R. P. Dick and R. Joseph, "Three-dimensional chip-multiprocessor run-time thermal management", *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 27, no. 8, pp. 1479 - 1492, Aug. 2008.
 - [99] G. Zhao, H.-K. Kwan, C.-U. Lei and N. Wong, "Processor frequency assignment in three-dimensional MPSoCs under thermal constraints by polynomial programming", *IEEE Asia Pacific Conf. Circuits and Syst.*, Macao, PRC, 30 Nov.-3 Dec. 2008, pp. 1668-1671.
 - [100] K. Kang, J. Kim, S. Yoo and C.-M. Kyung, "Runtime power panagement of 3-D multi-core architectures under peak power and temperature constraints", *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 30, no. 6, pp. 905-918, Jun. 2011.
 - [101] E. L. Sueur and G. Heiser, "Dynamic voltage and frequency scaling: the laws of diminishing returns", *Proc. Int. Conf. Power Aware Computing and Syst.*, Vancouver, BC, Canada, 2010, pp. 1-8.
 - [102] S. Garg, D. Marculescu and R. Marculescu, "Technology-driven limits on runtime power management algorithms for multiprocessor Systems-on-Chip", *ACM J. Emerging Techno. Computing Syst.*, vol. 8, no. 4, art. 28, Oct. 2012.
 - [103] S. Garg, D. Marculescu and R. Marculescu, "Fundamental limits on run-time power management algorithms for MPSoCs", in *Design technologies for green and sustainable computing systems*, P. P. Pande, A. Ganguly and K. Chakrabarty (Eds.), Springer Verlag, 2013.
 - [104] H. Kapadia, L. Benini and G. D. Micheli, "Reducing switching activity on datapath buses with control-signal gating", *IEEE J. Solid-State Circuits*, vol. 34, no. 3, pp. 405-413, Mar. 1999.
 - [105] I. Brzozowski and A. Kos, "Minimization of power consumption in digital integrated circuits by reduction of switching activity", *Proc. Euromicro Conf.*, Milan, Italy, 8-10 Sept. 1999, pp. 376-380. September 1999.
-

- [106] R. V. Menon, S. Chennupati, N. K. Samala, D. Radhakrishnan and B. Izadi, "Power optimized combinational logic design", *Proc. Int. Conf. Embedded Syst. and App.*, Jun. 2003, pp. 223-227.
 - [107] R. V. Menon, S. Chennupati, N. K. Samala, D. Radhakrishnan and B. Izadi, "Switching activity minimization in combinational logic design", *Proc. Int. Conf. Embedded Syst. and App.*, Las Vegas, NV, 21-24 Jun. 2004, pp. 47-53.
 - [108] J. P. Fishburn and A. E. Dunlop, "TILOS: a posynomial programming approach to transistor sizing", *The Best of ICCAD, Part IV*, Springer, US, pp. 295-302, 2003.
 - [109] M. Borah, R. M. Owens and M. J. Irwin, "Transistor sizing for low power CMOS circuits", *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 15, no. 6, pp. 665-671, Jun. 1996.
 - [110] K. Roy, S. Mukhopadhyay and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits", *Proc. IEEE*, vol. 91, no. 2, pp. 305-327, Feb. 2003.
 - [111] S. Venkatesan, J. W. Lutze, C. Lage and W. J. Taylor, "Device drive current degradation observed with retrograde channel profiles", *Proc. Int. Electron Devices Meeting*, Washington, DC, 10-13 Dec. 1995, pp. 419-422.
 - [112] J. Jacobs and D. Antoniadis, "Channel profile engineering for MOSFETs with 100 nm channel lengths", *IEEE Trans. Electron Devices*, vol. 42, no. 5, pp. 870-875, May 1995.
 - [113] W. Yeh and J. Chou, "Optimum halo structure for sub-0.1 μ m CMOSFETs", *IEEE Trans. Electron Devices*, vol. 48, no. 10, pp. 2357-2362, Oct. 2001.
 - [114] A. DasGupta, "Multiple gate MOSFETs: The road to the future", *Int. Workshop Physics of Semicon. Devices*, Mumbai, India, 16-20 Dec. 2007, pp. 96-101.
 - [115] Intel@22nm Technology. Available: <http://www.intel.com/content/www/us/en/silicon-innovations/intel-22nm-technology.html>.
 - [116] J. P. Halter and F. N. Najm, "A gate-level leakage power reduction method for ultra-low-power CMOS circuits", *Proc. IEEE Custom Integrated Circuits Conf.*, Santa Clara, CA, 5-8 May 1997, pp. 475-478.
 - [117] H. Rahman and C. Chakrabarti, "A leakage estimation and reduction technique for scaled CMOS logic circuits considering gate-leakage", *Proc. IEEE Int. Symp. Circuits and Syst.*, Vancouver, BC, 23-26 May 2004, pp. 297-300.
-

- [118] A. Abdollahi, F. Fallah and M. Pedram, "Leakage current reduction in CMOS VLSI circuits by input vector control", *IEEE Trans. Very Large Scale Integration Syst.*, vol. 12, no. 2, pp. 140-154, Feb. 2004.
 - [119] L. Wei, Z. Chen, M. Johnson, K. Roy, Y. Ye, and V. De, "Design and optimization of dual threshold circuits for low voltage low power applications", *IEEE Trans. Very Large Scale Integration Syst.*, vol. 7, no. 1, pp. 16-24, Mar. 1999.
 - [120] F. Assaderaghi, D. Sinitsky, S. Parke, J. Bokor, P. K. Ko, and C. Hu, "A dynamic threshold voltage MOSFET(DTMOS) for ultra-low voltage operation", *Proc. IEEE Int. Electron Devices Meeting*, San Francisco, CA, 11-14 Dec. 1994, pp. 809-812.
 - [121] S. Lee and T. Sakurai, "Run-time voltage hopping for low-power real-time systems", *Proc. IEEE/ACM Design Automation Conf.*, Los Angeles, CA, 5-9 Jun. 2000, pp. 806-809.
 - [122] C. H. Kim and K. Roy, "Dynamic V scaling scheme for active leakage power reduction", *Proc. Design, Automation and Test in Europe Conf.*, Paris, France, 4-8 Mar. 2002, pp. 163-167.
 - [123] A. J. Bhavnagarwala, B. L. Austin, K. A. Bowman, and J. D. Meindl, "A minimum total power methodology for projecting limits on CMOS GSI", *IEEE Trans. Very Large Scale Integration Syst.*, vol. 8, no.3, pp. 235-251, Jun. 2000.
 - [124] S. Tyagi et al., "A 130 nm generation logic technology featuring 70 nm transistors, dual Vt transistors and 6 layers of Cu interconnects", *Proc. Int. Electron Devices Meeting*, San Francisco, CA, 10-13 Dec. 2000, pp. 567-570.
 - [125] M. Takahashi et al., "A 60-mw MPEG4 video codec using clustered voltage scaling with variable supply-voltage scheme", *IEEE J. Solid-State Circuits*, vol. 33, no. 11, pp. 1772-1780, Nov. 1998.
 - [126] R. H. J. M. Otten, "Automatic floorplan design", *Proc. IEEE/ACM Design Automation Conf.*, Las Vegas, NV, 14-16 Jun. 1982, pp. 261-267.
 - [127] D. F. Wong and C. L. Liu, "A new algorithm for floorplan design", *Proc. IEEE/ACM Design Automation Conf.*, Las Vegas, NV, 29 Jun.-2 Jul. 1986, pp. 101-107.
 - [128] X. Hong, et al., "Corner block list: an effective and efficient topological representation of nonslicing floorplan", *Proc. IEEE/ACM Int. Conf. Computer-Aided Design*, San Jose, CA, 5-9 Nov. 2000, pp. 8-12.
-

- [129] E. F. Y. Young, C. C. N. Chu and Z. C. Shen, “Twin binary sequences: a nonredundant representation for general nonslicing floorplan”, *IEEE Trans. Computer-Aided Design*, vol 22, no. 4, pp. 457-469, Apr. 2003.
 - [130] C. Zhuang, Y. Kajitani, K. Sakanushi and L. Jin, “An enhanced Q-Sequence augmented with empty-room-insertion and parenthesis trees”, *Proc. Design, Automation and Test in Europe Conf.*, Paris, France, 4-8 Mar. 2002, pp. 61-68.
 - [131] H. Murata, K. Fujiyoshi, S. Nakatake and Y. Yajitani, “Rectangle packing based module placement”, *Proc. IEEE/ACM Int. Conf. Computer-Aided Design*, San Jose, CA, 5-9 Nov. 1995, pp. 472-479.
 - [132] S. Nakatake, K. Fujiyoshi, H. Murata, and Y. Yajitani, “Module placement on BSG-structure and IC layout applications”, *Proc. IEEE/ACM Int. Conf. Computer-Aided Design*, San Jose, CA, 10-14 Nov. 1996, pp. 484-491.
 - [133] J. M. Lin and Y. W. Chang, “TCG: A transitive closure graph-based representation for nonslicing floorplans”, *Proc. IEEE/ACM Design Automation Conf.*, Las Vegas, NV, 18-22 Jun. 2001, pp. 764-769.
 - [134] J. M. Lin and Y. W. Chang, “TCG-S: Orthogonal coupling of P^* -admissible representations for general floorplans”, *Proc. IEEE/ACM Design Automation Conf.*, New Orleans, LA, 10-14 Jun. 2002, pp. 842-847.
 - [135] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi Jr, “Optimization by simulated annealing”, *Science*, vol. 220, no. 4598, pp. 671-680, May 1983.
 - [136] P. Zhou, Y. Ma, Z. Li, R. P. Dick, L. Shang, H. Zhou, X. Hong, and Q. Zhou, “3D-STAF: scalable temperature and leakage aware floorplanning for three-dimensional integrated circuits”, *Proc. IEEE/ACM Int. Conf. Computer-Aided Design*, San Jose, CA, 4-8 Nov. 2007, pp. 590-597.
 - [137] D.-C. Juan, S. Garg and D. Marculescu, “Impact of manufacturing process variations on performance and thermal characteristics of 3D ICs: Emerging challenges and new solutions”, *Proc. Int. Symp. Circuits and Syst.*, Beijing, PRC, 19-23 May 2013, pp. 541-544.
 - [138] D.-C. Juan, S. Garg and D. Marculescu, “Statistical thermal evaluation and mitigation techniques for 3D chip-multiprocessors in the presence of process variations”, *Proc. Design, Automation and Test in Europe Conf.*, Grenoble, France, 14-18 Mar. 2011, pp. 1-6.
 - [139] K. Sankaranarayanan, S. Velusamy, M. Stan and K. Skadron, “A case for thermal-aware floorplanning at the microarchitectural level”, *J. Instruction-Level Parallelism*, vol. 7, pp. 1-16, 2005.
-

- [140] J. Cong, J. Wei and Y. Zhang, "A thermal-driven floorplanning algorithm for 3D ICs", *Proc. IEEE/ACM Int. Conf. Computer-Aided Design*, San Jose, CA, 7-11 Nov. 2004, pp. 306-313.
 - [141] J. Cong, G. Luo and Y. Shi, "Thermal-aware cell and through-silicon-via co-placement for 3D ICs", *Proc. IEEE/ACM Design Automation Conf.*, New York, NY, 5-9 Jun. 2011, pp. 670-675.
 - [142] Z. Li et al., "Hierarchical 3-D floorplanning algorithm for wirelength optimization", *IEEE Trans. on Circuits and Sys. I: Regular Papers*, vol. 53, no. 12, pp. 2637-2646, Dec. 2006.
 - [143] K. Puttaswamy and G. H. Loh, "The impact of 3-dimensional integration on the design of arithmetic units", *Proc. IEEE Int. Symp. Circuits and Syst.*, Island of Kos, Greece, 21-24 May 2006, pp. 49514954.
 - [144] J. Ouyang, G. Sun, Y. Chen, et al. "Arithmetic unit design using 180nm TSV-based 3D stacking technology", *Proc. IEEE Int. Conf. 3D Sys. Integration*, San Francisco, CA, 28-30 Sept. 2009, pp. 1-4.
 - [145] R. Egawa, J. Tada, H. Kobayashi and G. Goto, "Evaluation of fine grain 3-D integrated arithmetic units", *Proc. of IEEE Int. Conf. on 3D Sys. Integration*, San Francisco, CA, 28-30 Sept. 2009, pp. 1-8.
 - [146] M. Tremblay, B. Joy, and K. Shin, "A three dimensional register file for superscalar processors", *Proc. Hawaii Int. Conf. System Sciences*, Wailea, HI, 3-6 Jan. 1995, pp. 191-201.
 - [147] K. Puttaswamy and G. Loh, "Dynamic instruction schedulers in a 3-dimensional integration technology", *Proc. ACM/IEEE Great Lakes Symp. VLSI*, Philadelphia, PA, 1 May, 2006 pp. 153-158.
 - [148] L. Cheng, L. Deng, and M. D. Wong, "Floorplanning for 3D VLSI design", *Proc. IEEE/ACM Asia South Pacific Design Autom. Conf.*, Shanghai, PRC, 18-21 Jan. 2005, pp. 405-411.
 - [149] Y. Ma, X. Hong, S. Dong and C. K.Cheng, "3D CBL: an efficient algorithm for general 3-dimensional packing problems", *Proc. Midwest Symp. Circuits and Syst.*, Covington, KY, 7-10 Aug. 2005, pp. 1079-1082.
 - [150] H. Yamazaki, K. Sakanushi, S. Nakatake, and Y. Kajitani, "The 3D-packing by meta data structure and packing heuristics", *IEICE Trans. Fundam. Electron., Commun. and Computer Sci.*, vol. E82-A, no. 4, pp. 639-645, Apr. 2000.
-

- [151] B. Goplen and S. Sapatnekar, "Placement of 3D ICs with thermal and interlayer via considerations", *Proc. Design Automation Conf.*, San Diego, CA, 4-8 Jun. 2007, pp. 626-631, 2007.
 - [152] K. Balakrishnan, V. Nanda, S. Easwar, and S. K. Lim, "Wire congestion and thermal aware 3D global placement", *Proc. IEEE/ACM Asia South Pacific Design Autom. Conf.*, Shanghai, PRC, 18-21 Jan. 2005, pp. 1131-1134.
 - [153] B. Goplen and S. Sapatnekar, "Efficient thermal placement of standard cells in 3D ICs using a force directed approach", *Proc. IEEE/ACM Int. Conf. Computer-Aided Design*, San Jose, CA, 9-13 Nov. 2003, pp. 86-89.
 - [154] H. Yan, Q. Zhou, and X. Hong, "Thermal aware placement in 3D ICs using quadratic uniformity modeling approach", *Integration, the VLSI Journal*, vol. 42, no. 2, pp. 175-180, Feb. 2009.
 - [155] J. Cong, G. Luo, J. Wei, and Y. Zhang, "Thermal-aware 3D IC placement via Transformation", *Proc. IEEE/ACM Asia South Pacific Design Autom. Conf.*, Yokohama, Japan, 23-26 Jan. 2007, pp. 780-785.
 - [156] C.-C. Chang and J. Cong, "An efficient approach to multilayer layer assignment with an application to via minimization", *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 18, no. 5, pp. 608-620, May 1999.
 - [157] H. El-Rewini, T. G. Lewis and H. H. Ali, *Task scheduling in parallel and distributed systems*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1994.
 - [158] A. Podobas and M. Brorsson, "Architecture-aware task-scheduling: a thermal approach", *Future Architectural Support for Parallel Programming*, San Jose, CA, 2011.
 - [159] D. C. Vanderster, A. Baniassadi and N. J. Dimopoulos, "Exploiting task temperature profiling in temperature-aware task scheduling for computational clusters", *Proc. Asia-Pacific Conf. Advances in Computer Syst. Architecture*, Seoul, Korea, 2007, pp. 175-185.
 - [160] A. K. Coskun, T. S. Rosing and K. Whisnant, "Temperature aware task scheduling in MPSoCs", *Proc. Design, Automation and Test in Europe Conf.*, Nice, France, 16-20 Apr. 2007, pp. 1-6.
 - [161] A. K. Coskun, José, L. Ayala, D. Atienza, T. S. Rosing and Y. Leblebici, "Dynamic thermal management in 3D multicore architectures", *Proc. Design, Automation and Test in Europe Conf.*, Nice, France, 20-24 Apr. 2009, pp. 1410-1415.
-

- [162] T. Li and L. K. John, "Run-time modeling and estimation of operating system power consumption", *Proc. ACM SIGMETRICS Int. Conf. Measurement and Modeling of Computer Syst.*, San Diego, CA, 10-14 Jun. 2003, pp. 160-171.
 - [163] C. Isci and M. Martonosi, "Runtime power monitoring in high-end processors: methodology and empirical data", San Diego, CA, 3-5 Dec. 2003, pp. 93-105.
 - [164] X. Zhou, J. Yang, Y. Xu, Y. Zhang and J. Zhao, "Thermal-aware task scheduling for 3D multicore processors", *IEEE Trans. Parallel and Distributed Syst.*, vol. 21, no. 1, pp. 60-71, Jan. 2010.
 - [165] C.-L. Lung, Y.-L. Ho, D.-M. Kwai and S.-C. Chang, "Thermal-aware on-line task allocation for 3D multi-core processor throughput optimization", *Proc. Design, Automation and Test in Europe Conf.*, Grenoble, France, 14-18 Mar. 2011, pp. 1-6.
 - [166] M. D. Powell, M. Gomaa and T. N. Vijaykumar, "Heat-and-Run: leveraging SMT and CMP to manage power density through the operating system", *Proc. Int. Conf. Architectural Support For Programming Languages and Operating Systems*, Boston, MA, 7-13 Oct. 2004, pp. 260-270.
 - [167] J. Yang, X. Zhou, M. Chrobak, Y. Zhang and L. Jin, "Dynamic thermal management through task scheduling", *Proc. IEEE Int. Symp. Performance Analysis of Syst. and Software*, Austin, TX, 20-22 Apr. 2008, pp. 191-201.
 - [168] T. Chantem, X. S. Hu and R. P. Dick, "Temperature-aware scheduling and assignment for hard real-time applications on MPSoCs", *IEEE Trans. Very Large Scale Integration Syst.*, vol. 19, no. 10, pp. 1884-1897, Oct. 2011.
 - [169] J. Cong and Y. Zhang, "Thermal-driven multilevel routing for 3-D ICs", *Proc. Asia and South Pacific Design Autom. Conf.*, Shanghai, PRC, 18-21 Jan. 2005, pp. 121-126.
 - [170] J. Cong and Y. Zhang, "Thermal via planning for 3-D ICs," *Proc. IEEE/ACM Int. Conf. Computer-Aided Design*, San Jose, CA, 6-10 Nov. 2005, pp. 744-751.
 - [171] K. Wang, Y. Ma, S. Dong, Y. Wang, X. Hong and J. Cong, "Rethinking thermal via planning with timing-power-temperature dependence for 3D ICs", *Proc. Asia and South Pacific Design Autom. Conf.*, Yokohama, Japan, 25-28 Jan. 2011, pp. 261-266.
 - [172] B. Goplen and S. S. Sapatnekar, "Thermal via placement in 3D ICs", *Proc. Int. Symp. Physical Design*, San Francisco, CA, 3-6 Apr. 2005, pp. 167-174.
-

- [173] B. Goplen and S. S. Sapatnekar, "Placement of thermal vias in 3-D ICs using various thermal objectives", *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 25, no. 4, pp. 692-709, Apr. 2006.
 - [174] H. Yu, J. Ho and L. He, "Simultaneous Power and Thermal Integrity Driven Via Stapling in 3D ICs", *Proc. IEEE/ACM Int. Conf. Computer-Aided Design*, San Jose, CA, 5-9 Nov. 2006, pp. 802-808.
 - [175] H. Yu, Y. Shi, L. He and T. Karnik, "Thermal via allocation for 3D ICs considering temporally and spatially variant thermal power," *IEEE Trans. on VLSI Syst.*, vol. 16, pp. 1609-1619, Dec. 2008.
 - [176] H. Yu, J. Ho and L. He, "Allocating power ground vias in 3D ICs for simultaneous power and thermal integrity". *ACM Trans. on Design Automation of Electronic Systems*, vol. 41, no. 3, pp. 41:1-41:31, 2009.
 - [177] H. Hua, C. Mineo, K. Schoenfliess, A. Sule, S. Melamed, R. Jenkal and W. Rhett Davis, "Exploring compromises among timing, power and temperature in three-dimensional integrated circuits", *Proc. Design Auto. Conf.*, San Francisco, CA, 24-28 Jul. 2006, pp. 997-1002.
 - [178] Y. Chen, E. Kursun, D. Motschman, C. Johnson and Y. Xie, "Analysis and mitigation of lateral thermal blockage effect of through-silicon-via in 3D IC designs", *Proc. Int. Symp. Low Power Electronics and Design*, Fukuoka, Japan, 1-3 Aug. 2011, pp. 397-402.
 - [179] D. B. Tuckerman and R. F. W. Pease, "High-performance heat sinking for VLSI", *IEEE Electron Device Lett.*, vol. EDL-2, no. 5, pp. 126-129, May 1981.
 - [180] M. S. Bakir, B. Dang, and J. D. Meindl, "Revolutionary NanoSilicon Ancillary Technologies for Ultimate-Performance Gigascale Systems", *Proc. Custom Integrated Circuits Conf.*, San Jose, CA, 16-19 Sept. 2007, pp. 421-428.
 - [181] G. Y. Tang, S. P. Tan, N. Khan, D. Pinjala, J. H. Lau, A. B. Yu, K. Vaidyanathan and K. C. Toh, "Integrated Liquid Cooling Systems for 3-D Stacked TSV Modules", *IEEE Trans. Compon. Packag. Technol.*, vol. 33, no. 1, pp. 184-195, Mar. 2010.
 - [182] C. R. King, Jr., D. Sekar, M. S. Bakir, B. Dang, J. Pikarsky and J. D. Meindl, "3D stacking of chips with electrical and microfluidic I/O interconnects", *Proc. Electron. Compon. Technol. Conf.*, Lake Buena Vista, FL, 27-30 May 2008, pp. 1-7.
 - [183] S. Ndao, Y. Peles and M.K. Jensen, "Multi-objective thermal design optimization and comparative analysis of electronics cooling technologies", *Int. J. Heat and Mass Transfer*, vol. 52, pp. 4317-4326, Sept. 2009.
-

- [184] D. Liu and S.V. Garimella, "Analysis and optimization of the thermal performance of microchannel heat sinks", *Int. J. Numerical Methods for Heat & Fluid Flow*, vol. 15, no. 1, pp. 7-26, 2005.
 - [185] B. Shi and A. Srivastava, "Cooling of 3D-IC using non-uniform micro-channels and sensor based dynamic thermal management", *Proc. Annual Allerton Conf. Communication, Control, and Computing*, Monticello, IL, 28-30 Sept. 2011, pp. 1400-1407.
 - [186] A. K. Coskun, J. L. Ayala, D. Atienza and T. S. Rosing, "Modeling and dynamic management of 3D multicore systems with liquid cooling", *Proc. Annual IFIP/IEEE Int. Conf. Very Large Scale Integration (VLSI-SoC)*, Florianopolis, Brazil, 12-14 Oct. 2009, pp. 60-65.
 - [187] A. K. Coskun, D. Atienza, T. S. Rosing, T. Brunschwiler, and B. Michel, "Energy-efficient variable-flow liquid cooling in 3D stacked architectures", *Proc. Design, Automation and Test in Europe Conf.*, Dresden, Germany, 8-12 Mar. 2010, pp. 111-116.
 - [188] M. M. Sabry, A. K. Coskun, D. Atienza, T. Š. Rosing and T. Brunschwiler, "Energy-efficient multiobjective thermal control for liquid-cooled 3-D stacked architectures", *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 30, no. 12, pp. 1883-1896, Dec. 2011.
 - [189] F. Zanini, M. M. Sabry, D. Atienza and G. D. Micheli, "Hierarchical thermal management policy for high-performance 3D systems with liquid cooling", *IEEE J. Emerging and Selected Topics in Circuits and Syst.*, vol. 1, no. 2, pp. 88-101, Jun. 2011.
 - [190] M. M. Sabry, A. Sridhar, and D. Atienza, "Thermal balancing of liquid-cooled 3D-MPSoCs using channel modulation", *Proc. IEEE/ACM Design Automation and Test in Europe Conf.*, Dresden, Germany, 12-16 Mar. 2012, pp. 599-604.
 - [191] Y. Zhang, C. R. King Jr., J. Zaveri, Y. J. Kim, V. Sahu, Y. Joshi and M. S. Bakir., "Coupled electrical and thermal 3D IC centric microfluidic heat sink design and technology", *Proc. Electron. Compon. Technol. Conf.*, Lake Buena Vista, FL, 31 May-3 Jun. 2011, pp. 2037-2044.
 - [192] T.-Y. Wang and C. C. Chen, "3-D Thermal-ADI: a linear-time chip level transient thermal simulator", *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 21, no. 12, pp. 1434-1445, Dec. 2002.
 - [193] T.-Y. Chiang, S. Souri, C. O. Chui and K. Saraswat, "Thermal analysis of heterogeneous 3D ICs with various integration scenarios", *Proc. IEEE Int. Electron Devices Meeting*, Washington, DC, 2-5 Dec. 2001, pp. 31.2.131.2.4.
-

- [194] A. Jain, R. E. Jones, R. Chatterjee and S. Pozder, “Analytical and numerical modeling of the thermal performance of three-dimensional integrated circuits”, *IEEE Trans. Compon. Packag. Technol.*, vol. 33, no. 1, pp. 56-63, Mar. 2010.
 - [195] H. Xu, V. F. Pavlidis and G. D. Micheli, “Analytical heat transfer model for thermal through-silicon vias”, *Proc. Design Automation and Test in Europe*, Grenoble, France, 14-18 Mar. 2011, pp. 1-6.
 - [196] J.-H. Park, A. Shakouri and S.-M. Kang, “Fast thermal analysis of vertically integrated circuits (3-D ICs) using power blurring method”, *Proc. Pacific Rim/ASME Int. Electron. Packag. Tech. Conf.*, San Francisco, CA, 19-23 Jul. 2009, pp. 701-707.
 - [197] S. Melamed, T. Thorolfsson, T. R. Harris, S. Priyadarshi, P. Franzon, M. B. Steer and W. R. Davis, “Junction-level thermal analysis of 3-D integrated circuits using high definition power blurring”, *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 31, no. 5, pp. 676-689, May 2012.
 - [198] A. Ziabari and A. Shakouri, “Fast thermal simulations of vertically integrated circuits (3D ICs) including thermal vias”, *Proc. IEEE Intersociety Conf. Thermal and Thermomechanical Phenomena Electronic Syst.*, San Diego, CA, 30 May-1 Jun. 2012, pp. 588-596.
 - [199] Y. Zhan and S. S. Sapatnekar, “Fast computation of the temperature distribution in VLSI chips using the discrete cosine transform and table look-up”, *Proc. Asia and South Pacific Design Autom. Conf.*, Shanghai, PRC, 18-21 Jan. 2005, pp. 8792.
 - [200] T. Kemper, Y. Zhang, Z. Bian and A. Shakouri, “Ultrafast temperature profile calculation in IC chips”, *Proc. Int. Workshop Thermal Investigations ICs*, Nice, France, 27-29 Sept. 2006, pp. 1-5.
 - [201] V. M. Heriz, J.-H. Park, T. Kemper, S.-M. Kang and A. Shakouri, “Method of images for the fast calculation of temperature distributions in packaged VLSI chips”, *Proc. Int. Workshop Thermal Investigation ICs Syst.*, Budapest, Hungary, 17-19 Sept. 2007, pp. 1825.
 - [202] S. Nakanekar, A. Kaisare and S. Tonapi, “Optimal thermal characterization of a stacked die package with TSV technology”, *Proc. IEEE Intersociety Conf. Thermal and Thermomechanical Phenomena Electronic Syst.*, San Diego, CA, 30 May-1 Jun. 2012, pp. 130-136.
 - [203] D. Oh, C. C. P. Chen and Y. H. Hu, “Efficient thermal simulation for 3-D IC with thermal through-silicon vias”, *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 31, no. 11, pp. 1767-1771, Nov. 2012.
-

- [204] J. M. Koo, S. J. Im, L. Jiang, and K. E. Goodson, "Integrated microchannel cooling for three-dimensional electronic circuit architectures", *J. of Heat Transfer (Trans. of the ASME)*, vol. 127, no. 1, pp. 49-58, 2005.
 - [205] Y. J. Kim et. al., "Thermal characterization of interlayer microfluidic cooling of three-dimensional integrated circuits with nonuniform heat flux," *ASME J. Heat Transfer*, vol. 132, no. 4, pp. 1-9, Apr. 2010.
 - [206] H. Mizunuma, C.-L. Yang and Y.-C. Lu, "Thermal modeling for 3D-ICs with integrated microchannel cooling," *Proc. 2009 IEEE/ACM Int. Conf. CAD*, San Jose, CA, 2-5 Nov. 2009, pp. 256-263.
 - [207] A. Sridhar et. al., "3D-ICE: Fast compact transient thermal modeling for 3D ICs with inter-tier liquid cooling," *Proc. 2010 IEEE/ACM Int. Conf. CAD*, San Jose, CA, 7-11 Nov. 2010, pp. 463-470.
 - [208] A. Sridhar et. al., "Compact transient thermal model for 3D ICs with liquid cooling via enhanced heat transfer cavity geometries," *16th Int. Workshop on Thermal Investigations of ICs and Syst. (THERMINIC)*, Barcelona, Spain, Oct. 2010, pp.1-6.
 - [209] J. Xie and M. Swaminathan, "3D transient thermal solver using non-conformal domain decomposition approach", *Proc. IEEE/ACM Int. Conf. on Computer-Aided Design*, San Jose, CA, 5-8 Nov. 2012, pp. 333-340.
 - [210] A. Fourmigue, G. Beltrame, G. Nicolescu and E. M. Aboulhamid, "A linear-time approach for the transient thermal simulation of liquid-cooled 3D ICs", *Proc. Int. Conf. Hardware/Software Codesign and Syst. Synthesis*, Taipei, Taiwan, 9-14 Oct. 2011, pp. 197-205.
 - [211] R. Shah and A. London, *Laminar Flow Forced Convection in Ducts*. New York: Academic Press, 1978.
 - [212] F.P. Carson, Y. C. Kim, and I. S. Yoon, "3-D Stacked Package Technology and Trends," *Proc. IEEE*, vol. 97, pp. 31-42, Jan. 2009.
 - [213] P. Jacob et al. , "Mitigating memory wall effects in high-clock-rate and multicore CMOS 3-D processor memory stacks," *Proc. IEEE.*, vol. 97, pp. 108-122, Jan. 2009.
 - [214] G. H. Loh, Y. Xie and B. Black, "Processor design in 3D die-stacking technologies," *IEEE Micro*, vol. 27, no. 3, pp. 31-48, May 2007.
 - [215] Y. F. Tsai, Y. Xie, N. Vijaykrishnan and M. J. Irwin, "Three-dimensional cache design exploration using 3DCacti," *Proc. IEEE Int. Conf. on Comp. Design (ICCD)*, San Jose, CA, 2-5 Oct. 2005, pp. 519-524.
-

- [216] P. Ghosal, H. Rahaman and P. Dasgupta, "Thermal Aware Placement in 3D ICs," *Proc. Int. Conf. on Advances in Recent Technologies in Comm. and Computing (ARTCom)*, Kottayam, India, 16-17 Oct. 2010, pp. 66-70.
 - [217] H. Wang, Y. Fu, T. Liu and J. Wang "Thermal management via task scheduling for 3D NoC based multi-processor," *Proc. Int. SoC Design Conf. (ISOCC)*, Seoul, Korea, 22-23 Nov. 2010, pp. 440-444.
 - [218] ANSYS Multiphysics. Available: <http://www.ansys.com>.
 - [219] COMSOL Multiphysics. Available: <http://www.comsol.com>.
 - [220] W. Elenbaas, "Heat dissipation of parallel plates by free convection", *Physica*, vol. 9, no. 1, pp. 1-28, 1942.
 - [221] P. Teertstra, M. M. Yovanovich and J. R. Culham, "Analytical Forced Convection Modeling of Plate Fin Heat Sinks", *Journal of Electronics Manufacturing*, Vol. 10, No. 4, pp. 253-261, 2000.
 - [222] W. Huang, M. R. Stan, K. Skadron, K. Sankaranarayanan, S. Ghosh, and S. Velusamy. "Compact Thermal Modeling for Temperature-Aware Design," *Proc. Design Automation Conf.*, San Diego, CA, 7-11 July 2004, pp. 878-883.
 - [223] D. A. Kaminski and M. K. Jensen, *Introduction to Thermal and Fluids Engineering*, John Wiley & Sons. Inc, 2005.
 - [224] E. N. Seider and G. E. Tate, "Heat Transfer and Pressure Drop of Liquids in Tubes," *Industrial & Engineering Chemistry*, vol. 28, no. 12, pp. 1429-1435, Dec. 1936.
 - [225] W. M. Kays, and M.E. Crawford, *Convective Heat and Mass Transfer*, 3rd ed., McGraw-Hill, New York, 1993.
 - [226] M.B. Healy et. al., "Design and analysis of 3D-MAPS: A many-core 3D processor with stacked memory," *Proc. IEEE Custom Integrated Circuits Conf. (CICC)*, San Jose, CA, 19-22 Sept. 2010, pp. 1-4.
 - [227] S. G. Kandlikar, "A General Correlation for Saturated Two-Phase Flow Boiling Heat Transfer Inside Horizontal and Vertical Tubes," *ASME J. Heat Transfer*, vol. 112, no. 1, pp. 219-228, 1990.
 - [228] C. B. Sobhan and S. V. Garimella, "A comparative analysis of studies on heat transfer and fluid flow in microchannel," *Microscale Thermophysical Engineering*, vol. 5, no. 4, pp. 293-311, Apr. 2001.
-

- [229] X. F. Peng and G. P. Peterson, "Convective heat transfer and flow friction for water flow in microchannel structures," *Int. J. Heat Mass Transfer*, vol. 39, no. 12, pp. 2599-2608. Aug. 1996.
 - [230] T. A. Davis and E. Palamadai, KLU: sparse LU factorization algorithm for circuit simulation. Available: <http://www.cise.ufl.edu/research/sparse/klu>.
 - [231] 3D-ACME. Available: <http://www.3dacme.allalla.com>.
 - [232] J. Friedrich, B. McCredie, N. James, B. Huott, B. Curran, E. Fluhr, G. Mittal, E. Chan, Y. Chan, D. Plass, Sam Chu, Hung Le, L. Clark, J. Ripley, S. Taylor, J. DiLullo and M. Lanzerotti, "Design of the power 6 microprocessor", *Proc. IEEE Int. Solid State Circuits Conf.*, San Francisco, CA, 11-15 Feb. 2007, pp. 9697.
 - [233] U.M. Nawathe, M. Hassan, L. Warriner, K. Yen, B. Upputuri, D. Greenhill, A. Kumar and H. Park, "An 8-core 64-thread 64b power-efficient SPARC SoC", *Proc. IEEE Int. Solid State Circuits Conf.*, San Francisco, CA, 11-15 Feb. 2007, pp. 108590.
 - [234] B. Rogers, A. Krishna, G. Bell, K. Vu, X. W. Jiang, and S. Yan, "Scaling the bandwidth wall: Challenges in and avenues for CMP scaling", *Proc. IEEE Int. Symp. on Computer Architecture*, Austin, TX, 20-24 Jun. 2009, pp. 371382.
 - [235] K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat, "3D ICs: A novel chip design for improving deep submicron interconnect performance and systems-on-chip integration". *Proc. IEEE*, vol. 89, no. 5, pp. 602-633, 2001.
 - [236] B. Black, M. Annavaram, E. Brekelbaum, J. DeVale, G. Loh, L. Jiang, D. McCauley, P. Morrow, D. Nelson, D. Pantuso, P. Reed, J. Rupley, S. Shankar, J. Shen, and C. Webb, "Die stacking (3D) microarchitecture", *Proc. IEEE/ACM Int. Symp. on Microarchitecture*, Orlando, FL, 9-13 Dec. 2006, pp. 469479.
 - [237] F. Li, C. Nicopoulos, T. Richardson, Y. Xie, N. Vijaykrishnan, and M. Kandemir, "Design and management of 3D chip multiprocessors using network-in-memory", *Proc. IEEE/ACM Int. Sym. on Computer Architecture*, Boston, MA, 17-21 Jun. 2006, pp. 130141.
 - [238] H. B. Sun, J. B. Liu, R. S. Anigundi, N. N. Zheng, J. Lu, K. Rose, and T. Zhang, "3D DRAM design and application to 3D multicore systems". *IEEE Design and Test of Computers*, vol. 26, no. 5, pp. 36-47, 2009.
 - [239] D. Brooks, and M. Martonosi, "Dynamic thermal management for high-performance microprocessors", *Proc. Int. Symp. on High-Performance Computing Architecture*, Monterrey, Mexico, 19-24 Jan. 2001, pp. 171182.
-

- [240] W. P. Liao, L. He, and K. M. Lepak, "Temperature and supply voltage aware performance and power modeling at microarchitecture level". *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 24, no. 7, pp. 1042-1053, 2005.
 - [241] J. Donald and M. Martonosi, "Techniques for multi-core thermal management: Classification and new exploration", *Proc. IEEE Int. Sym. of Computer Architecture*, Boston, MA, 17-21 Jun. 2006, pp. 7888.
 - [242] A. Kumar, L. Shang, L.-S. Peh, and N. K. Jha, "System-level dynamic thermal management for high-performance microprocessors". *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, vol. 27, no. 1 pp. 96-108, 2008.
 - [243] T. Ebi, M. Abdullah, A. Faruque, and J. Henkel, "TAPE: Thermal-aware agent-based power economy for multi/many-core architectures", *Proc. IEEE/ACM Int. Conf. on Computer-Aided Design*, San Jose, CA, 2-5 Nov. 2009, pp. 302309.
 - [244] D. Li, S. X.-D. Tan, E. H. Pacheco, and M. Tirumala, "Parameterized architecture-level dynamic thermal models for multicore microprocessors". *ACM Trans. on Design Automation Electrical System*, vol. 15, no. 2, pp. 16:1-16:22, 2010.
 - [245] H. Sanchez, B. Kuttanna, T. Olson, M. Alexander, G. Gerosa, R. Philip and J. Alvarez, "Thermal management system for high performance PowerPC microprocessors", *Proc. IEEE Computer Society Int. Conf.*, San Jose, CA, 23-26 Feb. 1997, pp. 325-330.
 - [246] D. H. Kim, K. Athikulwongse, and S. K. Lim, "A study of through-silicon-via impact on the 3D stacked IC layout", *Proc. IEEE/ACM Int. Conf. on Computer-Aided Design*, San Jose, CA, 2-5 Nov. 2009, pp. 674-680.
 - [247] B. B. Dang, P. J. Joseph, M. S. Bakir, T. Spencer, P. Kohl, and J. Meindl, "Wafer-level microfluidic cooling interconnects for GSI", *Proc. IEEE Int. Interconnect Technology Conf.*, Burlingame, CA, 6-8 June 2005, pp. 180-182.
 - [248] S. Xu, V. Venkatraman and W. Burleson, "Energy-aware differential current sensing for global on-chip interconnects", *Proc. IEEE Int. Midwest Symp. on Circuits and Syst.*, San Juan, PR, 6-9 Aug. 2006, pp. 718-722.
 - [249] T. Wang, D. Chen and R. Geiger, "Multi-site on-chip current sensor for electromigration monitoring", *Proc. IEEE Int. Midwest Symp. on Circuits and Syst.*, Seoul, Korea, 7-10 Aug. 2011, pp. 1-4.
 - [250] J. D. Hamilton, *Time Series Analysis*, Princeton Press, 1994.
-

- [251] R. E. Kalman, "A new approach to linear filtering and prediction problems". *J. of Basic Engineering (Trans. of the ASME)* vol. 82, Series D, pp. 35-45, 1960.
 - [252] D. Brooks, V. Tiwari and M. Martonosi, "Wattch: a framework for architectural-level power analysis and optimizations", *Proc. Int. Symp. Computer Architecture*, vol. 28, no. 2, Vancouver, BC, Canada, 14 June 2000, pp. 83-94.
 - [253] Cacti: An integrated cache and memory access time, cycle time, area, leakage, and dynamic power model, Available: <http://www.hpl.hp.com/research/cacti/>.
 - [254] SPEC2000: Standard Performance Evaluation Corporation, Available: <http://www.spec.org/cpu2000/>.
 - [255] L. P. Carloni, P. Pande and Y. Xie, "Networks-on-chip in emerging interconnect paradigms: advantages and challenges", *Proc. 3rd ACM/IEEE Inter. Sym. on Networks-on-Chip*, pp. 93-102, San Diego, California, 10-13 May 2009.
 - [256] Y. Ye et al., "3D optical networks-on-chip (NoC) for multiprocessor systems-on-chip (MP-SoC)", *Proc. IEEE Int. Conf. on 3D Sys. Integration (3DIC)*, pp. 1-6, San Francisco, California, 28-30 Sept. 2009.
 - [257] B. Feero and P. Pande, "Networks-on-chip in a three-dimensional environment: A performance evaluation", *IEEE Trans. on Computers*, vol. 58, no. 1, pp. 32-45, Jan. 2009.
 - [258] J. A. Burns et al., "A wafer-scale 3-D circuit integration technology", *IEEE Trans. on Elec. Dev.*, vol. 53, no. 10, pp. 2507-2516, Oct. 2006.
 - [259] A.P. Karmarkar, X. Xu and V. Moroz, "Performanace and reliability analysis of 3D-integration structures employing Through Silicon Via (TSV)", *Proc. 2009 IEEE Int. Reliability Physics Symp.*, pp. 682-687, Montreal, Canada, 26-30 April 2009.
 - [260] E. Culurciello and A. G. Andreou, "Capacitive coupling of data and power for 3D silicon-on-insulator VLSI", *Proc. IEEE Int. Symp. Circuits and Syst.*, Kobe, Japan, 23-26 May 2005, pp. 4142-4145,.
 - [261] M. Scandiuazzo et al., "3D system on chip memory interface based on modeled capacitive coupling interconnections", *Proc. 2010 IEEE Int. Conf. on 3D Systems Integration Conference (3DIC)*, pp.1-4, Munich, Germany, 16-18 Nov. 2010.
 - [262] T. Zhang, Y. Zhang and S. Sapatnekar, "Temperature-aware routing in 3-D ICs", *Proc. IEEE Asia South Pacific Design Autom. Conf.*, pp. 309-314, Yokohama, Japan, 24-27 Jan. 2006.
-

- [263] Y. J. Lee, R. Goel, and S. K. Lim, "Multi-functional interconnect co-optimization for fast and reliable 3D stacked ICs", *Proc. 2009 IEEE/ACM Int. Conf. on Computer-Aided Design*, pp. 645-651, San Jose, CA, 2-5 Nov. 2009.
 - [264] M. M. Sabry, A. K. Coskun and David Atienza, "Fuzzy control for enforcing energy efficiency in high-performance 3D systems", *Proc. IEEE/ACM Int. Conf. Computer-Aided Desig*, pp. 642-648, San Jose, California, 7-11 Nov. 2010.
 - [265] H. Oertel et. al., *Prandtl-Essentials of Fluid Mechanics*, Springer, 2009.
 - [266] CurieJet Piezoelectric Micropump: <http://www.microjet.com.tw/>.
 - [267] ASCO miniature valves: <http://www.ascovalve.com/>.
 - [268] M. Ohring, *Reliability and failure of electronic materials and devices*, Academic Press, 1998.
 - [269] M. Bahrami, M. M. Yovanovich and J. R. Culham, "Pressure drop of fully-developed, laminar flow in microchannels of arbitrary cross-section", *J. of Fluids Engineering*, vol. 128, no. 5, pp. 1036-1044, Sept. 2006.
 - [270] J. Minz, X. Zhao, and S. K. Lim, "Buffered clock tree synthesis for 3D ICs under thermal variations", *Proc. Asia and South Pacific Design Autom. Conf.*, Seoul, Korea, 21-24 Mar. 2008, pp. 504-509.
 - [271] T. Y. Kim and T. Kim, "Clock Tree synthesis for TSV-based 3D IC designs", *ACM Trans. Design Automation Electron. Syst.*, vol. 16, no. 4, pp. 48:1-48:21, Oct. 2011.
 - [272] K. D. Boese and A. B. Kahng, "Zero-skew clock routing trees with minimum wirelength", *Proc. IEEE Int. Conf. ASIC*, Rochester, NY, 21-25 Sept. 1992, pp. 17-21.
 - [273] X. Zhao, D. Lewis, H.-H. S. Lee and S. K. Lim, "Pre-bond testable low-power clock tree design for 3D stacked ICs", *Proc. Int. Conf. Computer-Aided Design*, San Jose, CA, 2-5 Nov. 2009, pp. 184-190.
 - [274] C.-L. Lung, Y.-S. Su, S.-H. Huang, Y. Shi and S.-C. Chang, "Fault-tolerant 3D clock network", *Proc. ACM/EDAC/IEEE Design Automation Conf.*, New York, NY, 5-9 Jun. 2011, pp. 645-651.
 - [275] H. Qian, C. H. Chang and H. Yu, "An efficient channel clustering and flow rate allocation algorithm for non-uniform microfluidic cooling of 3D integrated circuits", *Integration, the VLSI Journal*, vol. 46, no. 1, pp. 57-68, Dec. 2011.
-

- [276] H. Qian, X. Huang, H. Yu and C. H. Chang, "Cyber-physical thermal management of 3D multi-core cache processor system with microfluidic cooling", *ASP Journal of Low Power Electronics*, vol. 7, no. 1, pp. 110-121, Feb. 2011.
 - [277] H. Qian and C. H. Chang, "Microchannel splitting and scaling for thermal balancing of liquid-cooled 3DIC", *Proc. Int. Symp. Circuits and Syst.*, Beijing, PRC, 19-23 May 2013, pp. 801-804.
 - [278] H. Qian, H. Liang, C. H. Chang, W. Zhang and H. Yu, "Thermal simulator of 3D-IC with modeling of anisotropic TSV conductance and microchannel entrance effects", *Proc. Asia and South Pacific Design Autom. Conf.*, Yokohama, Japan, 22-25 Jan. 2013, pp. 485-490.
 - [279] H. Qian, X. Huang, H. Yu and C. H. Chang, "Real-time thermal management of 3D multi-core system with fine-grained cooling control", *Proc. IEEE Int. 3D System Integration Conf.*, Munich, Germany, 16-18 Nov. 2010, pp. 1-6.
 - [280] J. Chen, C. H. Chang and H. Qian, "New power index model for switching power analysis from adder graph of FIR filter", *Proc. Int. Symp. Circuits and Syst.*, Taipei, Taiwan, 24-27 May 2009, pp. 2197-2200.
-