Experimental and Numerical Thermal Analysis for Advanced Flip Chip Thermo-Compression Bonding via CMOS Micro-sensor Arrays and Finite Element Modelling

by
Depayne Athia

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement of the degree of
Masters of Applied Science
in
Mechanical and Mechatronics Engineering

Waterloo, Ontario, Canada, 2017

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Chapter 2 contains results that have been incorporated in a paper that was co-authored by Ari Laor, myself, Dr. Alireza Rezvani, Dr. Ivy Qin, Dr. Horst Clauberg and my supervisor, Dr. Michael Mayer. I co-wrote the code for post-processing analysis, assisted in experimental setup and procedure development, assisted in sensor chip calibration, and co-wrote the manuscript. For post-processing, I wrote all code as attached in the Appendices, with assistance from Dr. Mayer. Mr. Laor designed and calibrated the sensor chips, developed the experimental procedure, co-analyzed the results. Dr. Rezvani, Dr. Qin, Dr. Clauberg and Dr. Mayer helped develop the sensor chip, analyze the results, troubleshoot experimental setup issues, and the manuscript.

Chapter 3 consists of a paper that was co-authored by myself, Dr. Rezvani, Dr. Qin, Dr. Clauberg and my supervisor, Dr. Mayer. I wrote the methodology, created and developed the FE models, wrote the post-processing software and wrote the manuscript. Dr. Mayer helped develop the methodology, write the post-processing software and analyze the results. Dr. Rezvani, Dr. Qin and Dr. Clauberg assisted in developing the models, analyzing the results and with writing the paper.

Abstract

Thermo-compression bonding (TCB) relies on uniform thermal distribution during microelectronic packaging processes to ensure reliable interconnects are formed. During any TCB processes, the thermal application must uniformly distribute heat in order to produce robust, thoroughly bonded packages without being damaged due to thermo-mechanical effects. To better control and develop TCB processes, further insight through thermal analysis is required. Due to the form factors and complexity involved in TCB, it is difficult to accurately extract viable information such as temperature variation, lateral and vertical gradients, or interfacial bonding temperatures.

To extract real time in-situ temperature and force signals, a microsensor array was used to observe any thermo-mechanical features recorded during emulated TCB processes. Algorithms were developed to post-process the signals and produce quantifiable data. Finite element models were developed to verify the experimental thermal responses and subsequently post-analyze the numerical results. Models formed through hybridized contact resistance layers as well as surface contact models are also discussed.

Several features were identified and quantified: maximum heating rates, location of maximum lateral thermal gradients, internal joint thermal distributions, knee-region slope analysis and joint to joint thermal variation. The experimental responses in combination with numerical analyses show evidence that thermal applications during TCB is robust. Low thermal variation was found with respect to joint to joint temperatures. Chip design was found to heavily influence cooling on the periphery edges of the bump array. The sensor chip temperatures were to found to be about ≈ 6 °C lower than the extracted bump temperatures, signifying the use of microsensor arrays could be developed as accurate tools for thermal process control during TCB.

Acknowledgements

This work was funded by Kulicke & Soffa Industries. Special recognition is given to Dr. Alireza Rezvani, Dr. Ivy Qin, and Dr. Horst Clauberg for their invaluable contribution to the development of the work discussed in this thesis. I would also like to express my gratitude to Ari Laor, Dr. Erick (Xu) Di, and Dr. Michael Hook for their technical insight and support with all technical endeavors. Further appreciation is expressed towards my colleagues Gitanjali Shanbhag, Dr. Paola Russo and Emanuel Santos for their moral support and encouragement throughout this degree. I am gracious to have had Professor Mustafa Yavuz and Professor Richard Culham to provide invaluable feedback as readers during the drafting of this thesis.

Much appreciation is given to:

Neil Griffett, for helping setup the data acquisition system.
Mark Griffett, for always having the experimental setup and materials ready for the lab sessions every Monday and Friday, even at a moment's notice.
Matt Scott, for the training on the die bonder, plasma cleaner and for assisting with any other issues while conducting research at IQC.
Allison Walker, Karen Schooley and all other administrative and support staff in the MME department.

I would also like to express my overdue and undying appreciation towards the instructors and staff at Woburn Collegiate Institute. Special recognition goes to Jonathan Shulman, Judy Edwards, Tom Thompson, Rodica Iliescu, Todd Idenouye, Hanson Man, Rucsandra Seuleanu, Kosta Moliotsias, and Janice Campbell-Houston. The special roles each of you played have left me with lifelong lessons that have continued to resonate with me ever since parting from WCI.

Finally, I would like to thank my supervisor, Dr. Michael Mayer, for fostering and honing my abilities in areas ranging from technical aptitude to interpersonal skills. His patience, guidance and awe-inspiring ability to provide a multitude of solutions for any issue that arose during research made him an excellent supervisor that I will always be grateful for having. Thank you for providing me with this invaluable opportunity to grow as a researcher, an engineer, and as an individual.

Dedication

I dedicate this thesis to my parents, my family, my friends, and to all of those who believed in my dreams. I would not have made it this far on my mission without the love, support and encouragement from each and every one of you. Thank you for never losing hope nor giving up on me. This is also for those who never had a chance. Despite all obstacles, this process has been a privilege that I will never fail to acknowledge. Finally, I dedicate this to those who could not be with us today, to those who have faced adversity and hardship all throughout their lives, and to those, eternally fighting their inner demons. May this thesis give hope, encouragement, and substantiation that resilience and perseverance can conquer all.

Table of Contents

List of Figures

1 Introduction 1

2 Sensor Chip: Temperature Signal Analysis 26

List of Tables

List of Acronyms

B

BGA	Ball grid array

C

C2	Chip connect
C2W	Chip-to-wafer
C4	Controlled collapsed chip connection
CerDIP	Ceramic dual in-line package
CMOS	Complementary metal oxide semiconductor
CMY	Cooper-Mikic-Yovanovich
COB	Chip-on-board
CTE	Coefficient of thermal expansion

D

DAQ	Data acquisition system
DC	Direct current
DOE	Design of experiment

E

EFO	Electrical flame-off

F

fcTCB	Flip chip TCB
fcVFBGA	Flip chip very fine ball grid array
FE	Finite element

FEA	FE analysis
FEM	FE methods
FTRS	Final Third Region Slopes

H

HTS	High temperature solder
HTW	High temperature wire

I

IC	Integrated circuit
IEEE	Institute of Electrical and Electronics Engineers
I/O	Input/output or interconnection

J

JEDEC	Joint Electron Device Engineering Council

K

K&S	Kulicke and Soffa Industries
KP	Knee point

M

MFT	Maximum final temperatures
MHR	Maximum heating rates

N

NASA	National Aeronautics and Space Administration
NCF	Non-conductive film
NCP	Non-conductive paste

P

PC	Personal computer
PCB	Printed circuit board
PDE	Partial differential equation
PR	Photoresist
PTCO	Packaging test chip version O
PTCP	Packaging test chip version P
PWB	Printed wire board

R

R&D	Research and development
RMS	Root mean square
RTD	Resistive temperature detector

S

SEM	Scanning electron microscope
SMT	Solder mount technology

T

TCB	Thermo-compression bonding
TCR	Thermal contact resistance
TSV	Through-silicon-via

U

UBM	Under-bump metallurgy

W

W2W	Wafer-to-wafer

List of Symbols

S_t	amplified signal force sensitivity, in V/N
ΔV	measured excitation voltage, in V
F	force applied that caused the excitation voltage, in mN
S_N	normalized signal force sensitivity, in mV/V/N
t_0	initial start time, in s
$F_i(t)$	forces experienced at time t, by pad index i
i	pad index ranging from 1 to 64
$V_i(t_0)$	initial, reference voltage, in V
$V_i(t)$	excitation voltages experienced at time t
I	bias current applied, in mA
$R_{T,ref}$	calibrated resistance at a reference temperature, in Ω
$\alpha_{T,ref}$	temperature coefficient of resistance for a reference temperature, in 1/K or 1/ºC
ρ	density, in kg/m^3
m	mass, in kg
k	isotropic thermal conductivity, in W/m•K
h	heat transfer coefficient, in W/m^2•K
T_∞	temperature of the external fluid from a reasonable distance away from the surface, in K
$\boldsymbol{n}$	surface normal vector
C_p	specific heat or heat capacity at constant pressure, in J/kg•K
th	thickness of a surface element, in m
h_c	thermal contact constriction conductance, in W/m^2•K
h_g	thermal contact gap conductance, in W/m^2•K

m_{asp}	average RMS surface asperity slope
σ_{asp}	average RMS surface asperity height, in m
$k_{contact}$	sub-contrary or harmonic mean of thermal conductivity, in W/m•K
H_c	Vicker's microhardness, in Pa
p	contact pressure, in Pa
Δt_{steady}	time until ramp of bond head to ascend to the nominally set process temperature, in s
Δt_{ramp}	ramp duration of bond head press, in s
Δt_{press}	duration of bond head press, in s
Δt_{cool}	lift off/cooling time, in s

"...But tomorrow I see change
A chance to build anew
Built on spirit, intent of heart
and ideals based on truth

And tomorrow I wake with second wind
and strong because of pride
To know I fought with all my heart to keep my dream alive"
– Tupac Amaru Shakur

1 Introduction

1.1 Microelectronic Packaging

Microelectronic packaging provides the means to establish electrical connections between an integrated circuit (IC) and the remaining components along the packaging hierarchy. The three main levels are: primary levels connecting chips to packages; secondary levels connecting packages to PCBs; and tertiary levels connecting PCBs to PCBs or other higher-level assemblies (Fig. 1-1). Typical higher order packages beyond these three levels are consumer products such as laptops, PCs or smart phones. Primary level packaging is predominantly established by either wire bonding technologies or flip-chip technologies [2,3]. Each of these packages serve four main functions: signal distribution, power distribution, heat dissipation and mechanical support/protection. It must be able to distribute power and communicate signals to other components while sufficiently being cooled and

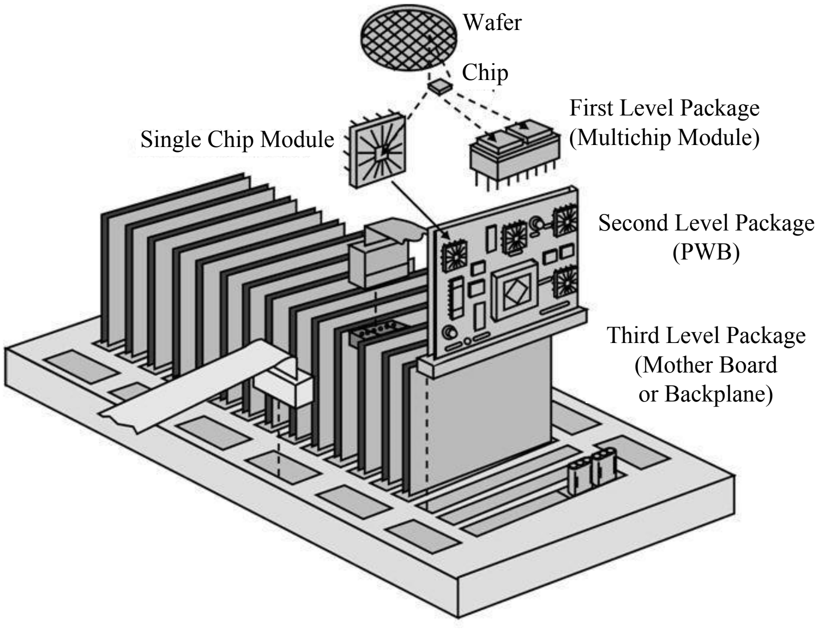


Fig. 1-1. Microelectronic packaging hierarchy. Reproduced from [1].

physically protected [3]. Therefore, it is crucial that this first packaging step is robust and reliable, regardless of its operating conditions or environment.

Package design considerations are driven by the ever-increasing industry needs of low cost, low power, small form factor and high performance packages. These requirements lead to progressively complicated structures that demand vigorous engineering methodologies to ensure they are electro-thermo-mechanically resilient.

1.1.1 Wire Bonding Technology

Wire bonding is a primary level solid state micro-welding technology, which establishes electrical connectivity through the use of metallic wires [2,3,4,5]. With trillions of wire bonds made each year, wire bonding is currently the most dominant primary level interconnection technology, representing about 85% of the total interconnections made as of 2014 [6, 7]. There are two main categorizations of wire bonding: ball-wedge bonding and wedge-wedge bonding. Phenomenal throughput is achieved with ball-wedge bonds, attributed to the faster process times in comparison to wedge-wedge bonds. A generalized flow diagram of a typical ball-wedge process is provided in Fig. 1-2. A computer-generated image of a wire-bond packaged chip is shown in Fig. 1-3. The first bond is typically made on the IC and the second to its package, substrate, or to a more recently common occurrence another IC.

Wire bonding technology remains dominant due to its process design producing flexible bonds in low temperature conditions with high yield rates [8]. Due to the rapid rise of gold prices following the recession of 2008, a significant market shift occurred leading to the expeditious development of wire bonding with Cu wires. Prices have been noted to be about 10x less for Cu wires compared to Au wires [9]. Mechanical and electrical properties of Cu such as high stiffness, tensile strength and electrical conductivity, as well as better process features, such as longer loop profiles and reduced wire thicknesses, further make it an appealing material choice over Au wires. This, of course, depends on the specific application of the package.

With an increased drive for reduced pitches and higher densities of interconnects as well as a need for the miniaturization of packages, wire bonding faces certain limitations

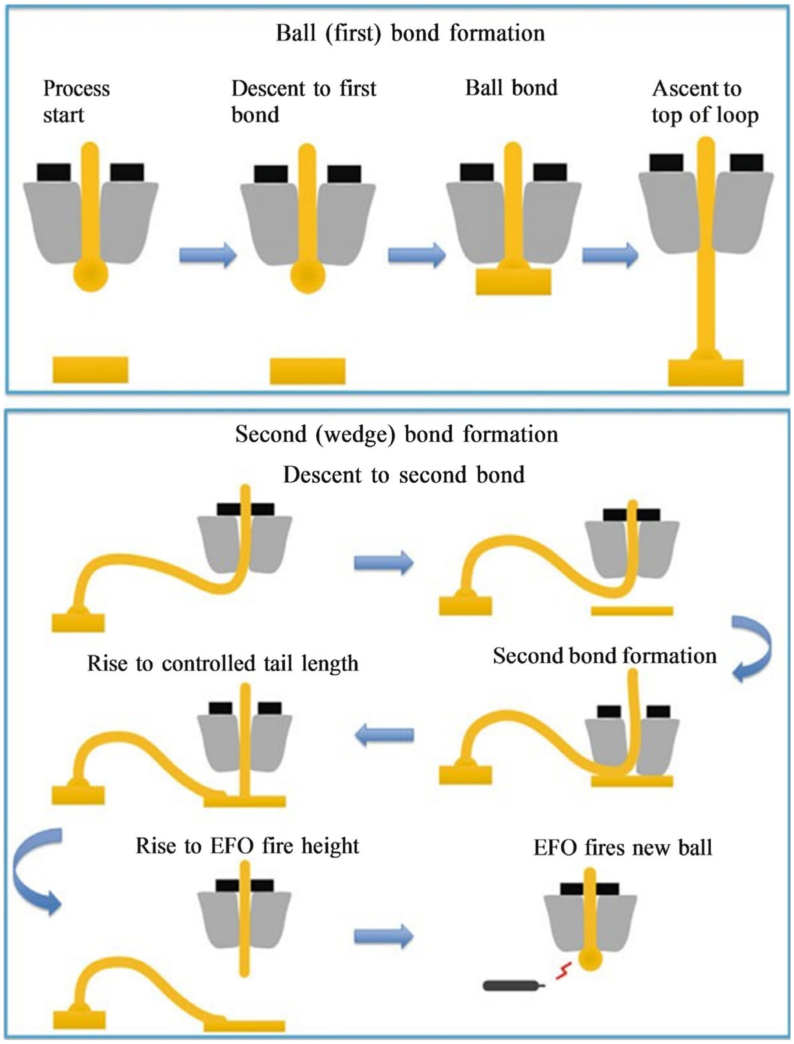


Fig. 1-2. Overview of the ball-wedge wire bonding process. Reproduced from [6].

Fig. 1-3. Three-dimensional graphical representation of Au wire bonds. Adapted from [5].

primarily due to geometric constraints, amongst other process-related and reliability issues [8]. Several innovations and creative techniques have been developed to follow industry trends such as stacked die wire-bonded packages, bond pad pitch reductions to ≈ 25 μm, ultra low loop height profiles and nano-insulated wire materials. However, significant redirection has shifted focus to flip chip thermo-compression bonding (fcTCB or TCB) as the more viable solution to these issues [10].

1.1.2 Flip Chip Technology

The denotation of flip chip comes from the step where the chip faces the substrate, whereas the opposite is the case in wire bonding. Flip chip technologies can be categorized into three main types: solder ball with reflow, thermo-sonic bonding and thermo-compression bonding. Controlled collapsed chip connection (C4) is the original solder ball with reflow technology, first introduced by IBM in 1964, is a pioneering primary level flip chip packaging technology that catered to demands of high input/outputs (I/Os) required by higher density applications [11]. Heat is applied through forms such as hot air, causing the solder to melt and forming the interconnection joint. Electrically insulative adhesive or underfills are then applied and cured [10,12]. It was able to provide high I/O counts by utilizing a ball grid array (BGA) of solder bumps which were bonded with lower pitches than wire bonding technologies at the time, while covering a fraction of the bonding area. It was originally used for processors in the IBM System/360 mainframe computer systems. Development of this technology and intellectual property related to it was kept protected within the domains of IBM until the 1990s [12]. Seizing the opportunity needed to advance development for their own interconnect solutions, major companies, such as AMD, Intel and HP, licensed IBM’s technology. A surge of research followed the introduction of solder-capped Cu pillar technologies leading to several concurrent advances in through-silicon-via (TSV) technology, fine pitch bonding, die thickness reductions and pure Cu-Cu interconnection technology [12].

In thermosonic bonding, the gridded array of bumps are electroplated onto the IC. The IC is then flipped such that the bumps are aligned downwards onto their respective substrate traces or pads. The bond head has an ultrasonic transducer applying transversal vibration to collectively form ultrasonic welds at all bump to substrate interfaces. This type is

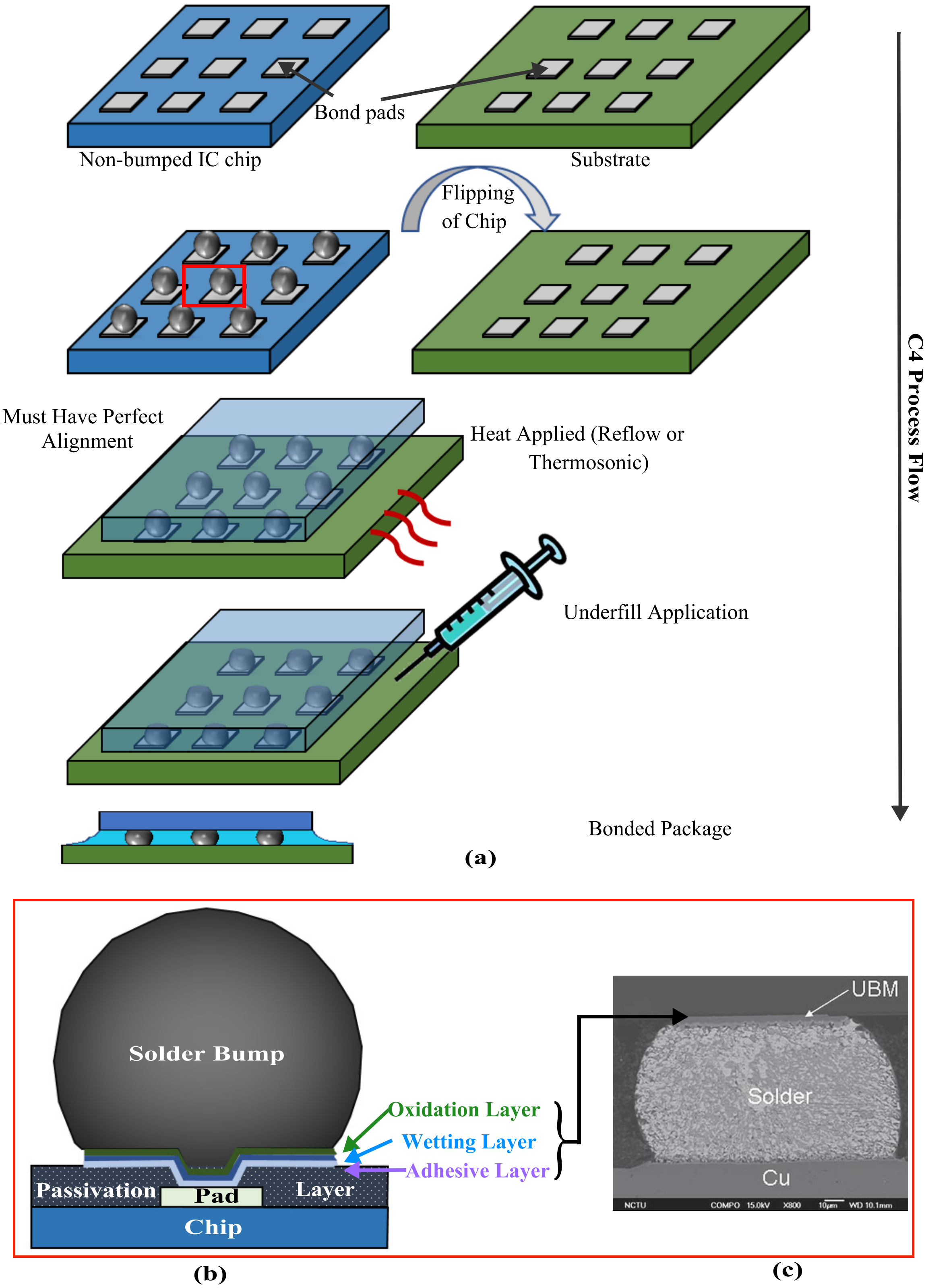


Fig. 1-4. (a) Typical C4 flip chip process; (b) under bump metallurgy (UBM) of a solder bump; (c) Cross sectional view of a bonded C4 joint [13].

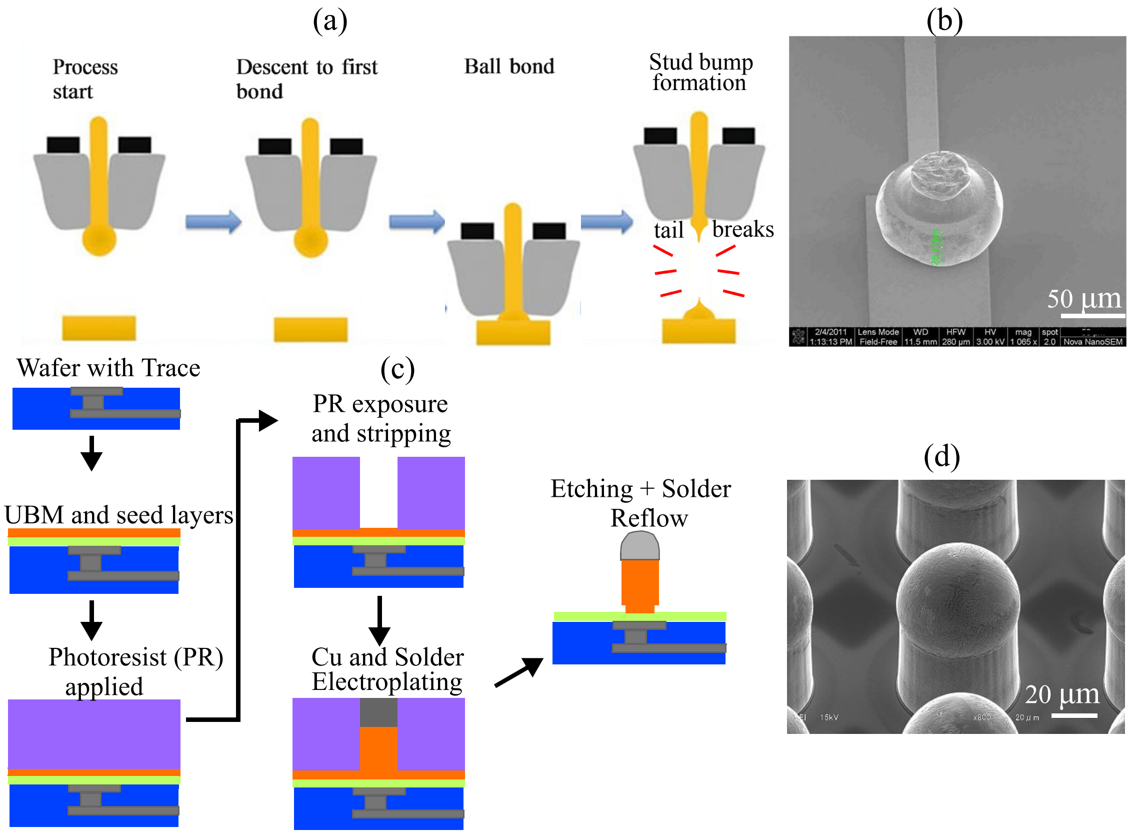


Fig. 1-5. (a) Stud bump formation process [6]; (b) SEM of a Au stud bump [14]; (c) Simplified process flow of solder capped Cu-pillar formation; (d) Cu-pillar array [15]

not used very often and limited to a relatively small number of I/Os. However, it allows for high throughputs.

1.1.2.1. Thermo-Compression Bonding

Thermo-compression bonding has been shown to be applied with materials other than just solder such as Au, Ag, and more currently of interest, solder-capped Cu pillar and Cu-Cu pillar bonding. Au and Ag TCB bonds are usually bonded in stud bump form. A comparison of the other two main forms of TCB bumps and their formation processes are shown in Fig. 1-5. TCB differs from thermosonic bonding through its heating mechanisms and application of forces. TCB also involves flipping the bumped/pillared chip, but applies a downwards pressure with concurrent heating to bond all I/Os at once (Fig. 1-6). In many cases, the underfill is applied prior to bonding, rather than after as seen in Fig. 1-4. The

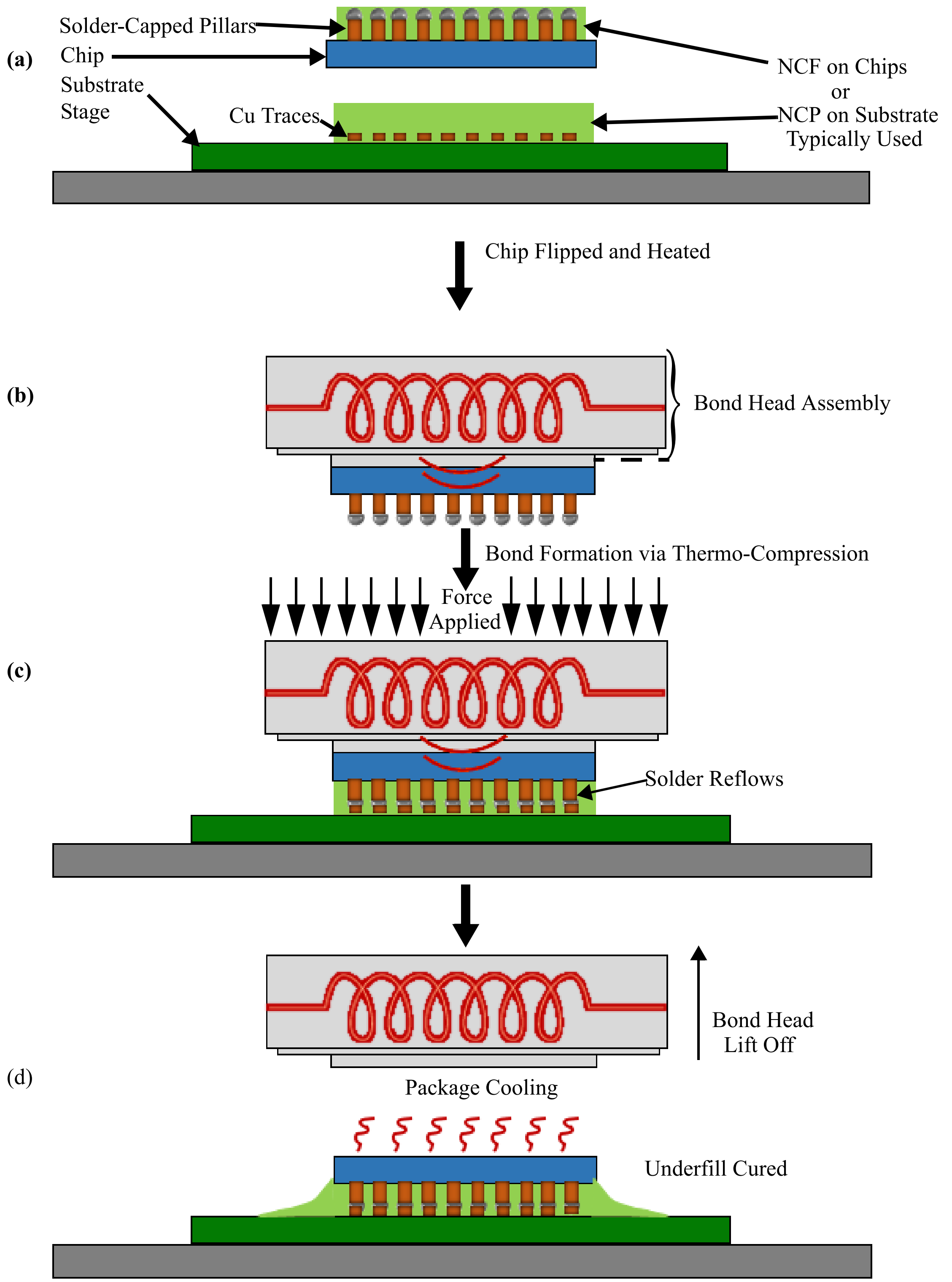


Fig. 1-6. Typical TCB process flow. (a) non-conductive film (NCF) is applied to the chip or non-conductive paste (NCP) is applied to the substrate; (c) chip is flipped and heated (c) bond head actuates downwards, curing the underfill and bonding the joints; (d) device packaged. Adapted from [16].

underfill improves reliability of flip chip packages through means such as mechanical reinforcement, induced stress relief and environmental protection. An Apple A10 processor bonded using TCB processes is shown in Fig.1-7.

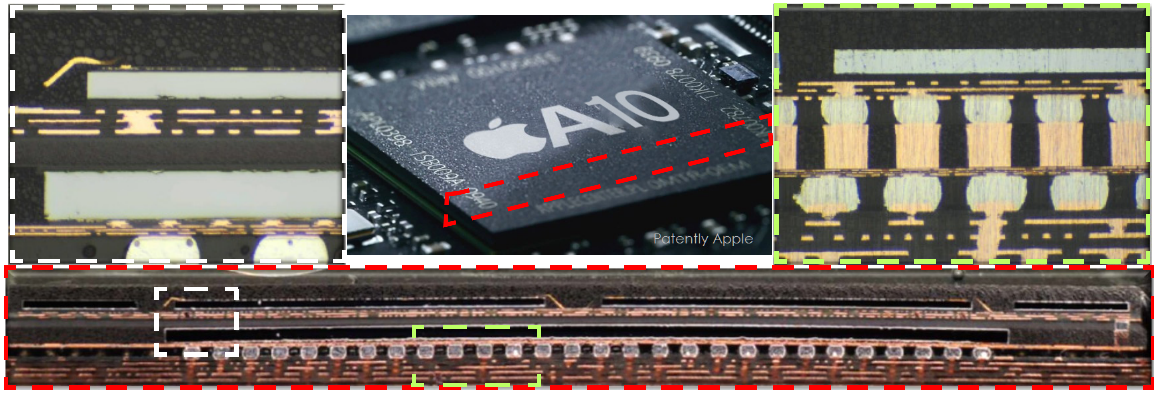


Fig. 1-7. An example product packaged via solder-capped Cu pillar TCB. The top-middle image indicates the location of the cross sectional cut of Apple’s A10 processor with magnified images on left, right and bottom. Reproduced from [17,18].

There are several advantages to using TCB over wire bonding and they are summarized in Table 1-1.

Due to the nature of the TCB process, there is great complexity associated with forming robust, reliable bonds. The concurrent application of heat and force onto a package with low pitched, high density pillar array(s) requires several process parameters to be optimized at once. Temperatures associated with TCB are typically beyond 200 °C ranging from 240 °C to 300 °C [19,20,21,22]. Forces are also crucial as under-compression can result incomplete bond formation and over-compression can result in issues such as solder leakage. There are a few important process-related features that could be detrimental to reliable bonding such as [23,24]:

1. Substrate and/or die warpage due to coefficient of thermal expansion (CTE) mismatch
2. Substrate and/or chip misalignment when bonding
3. Parallelism issues (tilt angle between normals of any of the package components)

4. Force variation during bonding (i.e. due to non-uniformity across pillar/bump heights)
5. Thermal variation during bonding (i.e. due to previous four factors)

Chip misalignment can result in incomplete joint formation due to reduction of thermally conductive pathways. If sufficient heat does not reach the bonding sites, the bonds may not have sufficient process temperatures to bond, resulting in incomplete or partial bonds. As the contact area between misaligned interconnects prior to bonding is reduced, there is a greater chance of Other issues related to TCB include underfill leakage, process compatibility issues with solder mount technology (SMT), cracking of dies during handling and last, but not least, bonding process windows that are currently too long for throughputs required by high volume applications.

Other variations or applications of TCB include Chip-on-Board (COB), Chip-to-Wafer (C2W) and Wafer-to-Wafer (W2W) bonding [25,26].

Table 1-1. General comparison of wire bonding and TCB technologies [10,23,24]

	Wire Bonding	**TCB**
Advantages	•Flexibility in positioning of I/Os and adaptive to different package types, orientations and die sizes •Very inexpensive (≈ \$0.0075 per bond) •Highly established understanding of reliability and concerns •Thoroughly understood process controls and infrastructure	•Space saving: I/Os contained within chip perimeter (area array) •More I/Os: several thousand common •Bonding: gang or collective; all I/Os bonded at once •Electrically sounder: better distribution of power and ground and reduced parasitic inductance •Significant higher improvements in speed/bandwidth applications •Better thermal management: typically, no encapsulation of chip •Shorter joints: I/O lengths comparable to diameters of wires
Disadvantages	•Number of I/Os limited to few 1000 max •Bonding: sequential; each I/O bonded one at a time	•Limited-to-nil flexibility in I/O position (accurate bonding placement crucial •Price: costly (≈ \$0.025 per bond) •Reliability and post-bonding testing methods currently expensive or destructive

1.1.2.2. Reliability/Quality Testing Methods

Wire bonding has well established bond quality tests conducted through destructive means such as shear testing (Fig. 1-8a), pull testing (Fig. 1-8b), or cross-sectioning (Fig 1-7). Reliability tests include humidity testing and high temperature storage (HTS) aging [27]. On the other hand, standard testing methods offered through organizations such as JEDEC, are still in development for flip chip TCB as process conditions can vary greatly. Considering the smaller form factors associated with TCB in contrast to wire bonding, as well as nature of the packages themselves having high densities of I/Os, it is difficult to effectively conduct quality and reliability testing. A solution to this could be through non-destructive testing methods. Non-destructive testing methods leave the original package intact without any exposure to external stresses or environmental factors, rendering them useful after testing. Non-destructive methods of evaluating post-process residual stress using Raman spectroscopy is discussed [29] and strain distribution using moiré interferometry in [12]. Other post-process methods are also used but are limited in their inability to provide real-time information of stress and thermal distribution during TCB processes. A technique of measuring temperatures using thermocouples have been demonstrated by Toray Engineering in [30], by Kulicke and Soffa Industries in [26], Daily et al., in [31], and Jeong et al. in [32]. While these are good for approximations of layer temperatures in a flip chip package, they are unable to provide a complete picture of the interfacial bonding temperatures. Geometric constraints such as very short die to die thicknesses, known as bond lines, limit the use of

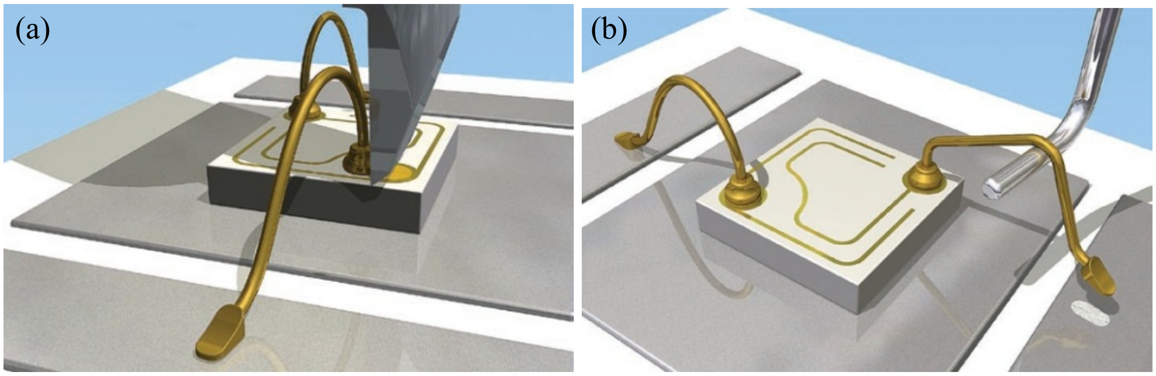


Fig. 1-8. Computerized representation of destructive (a) shear test and (b) pull test. Reproduced from [28].

thermocouples. All in all, these methods can be costly, time-intensive, inaccurate and can lead to misleading results. A possible alternative to these methods is the use of microsensors placed close to bonding sites to provide real time in-situ temperatures. A prototype microsensor array has been developed by a previous MASc. student in our group, Ari Laor, based on previous wire bonding microsensor designs as discussed in [2,23,33, 34,35]. Real time in-situ microsensors were also recently used by Bex et al. but is limited to three sensor measurements per test [36].

1.1.3 Future Trends of TCB

TCB has significantly developed over the past few decades shifting towards higher dimensional integration of IC chips. Packages are becoming denser with IC's having higher I/O counts through reduced pillar and pitch size reductions. It is common to find chip sizes greater than 10 mm × 10 mm, with sizes emerging in 2017 as large as 15 mm × 15 mm [37,38,39,40,41]. IC's are being stacked alongside and/or on top of multiple other IC's forming three-dimensional (3D) IC packages as seen in Fig. 1-9. It is predicted that by the year 2020, there could be up to 50,000 I/Os per IC [42]. With process throughput being a major bottle-neck of mass TCB adoption, chip stack packages produced via collective

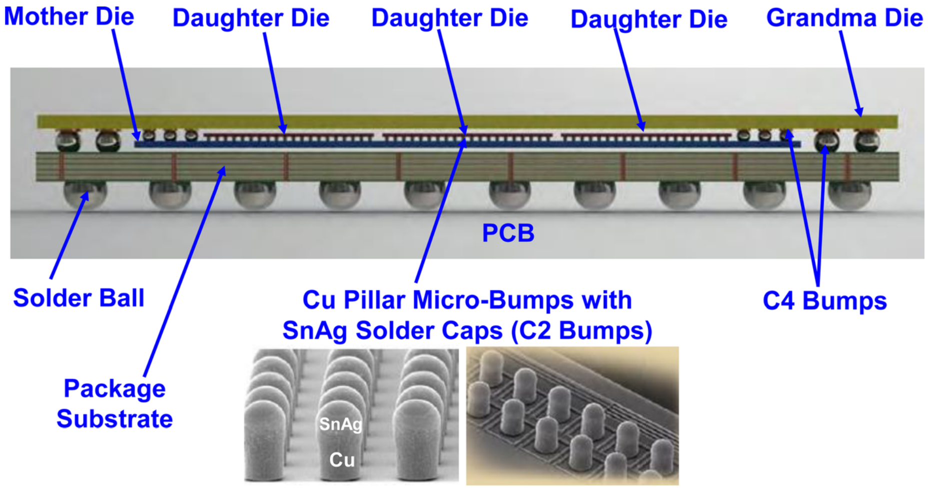


Fig. 1-9. 3D IC package fabricated through Amkor's Double-POSSOM™ TCB technology [42]. Solder-capped Cu pillars are typically denoted as C2 bumps.

bonding processes (Fig. 1-10) have gained interest. Through this form of TCB, ICs are first

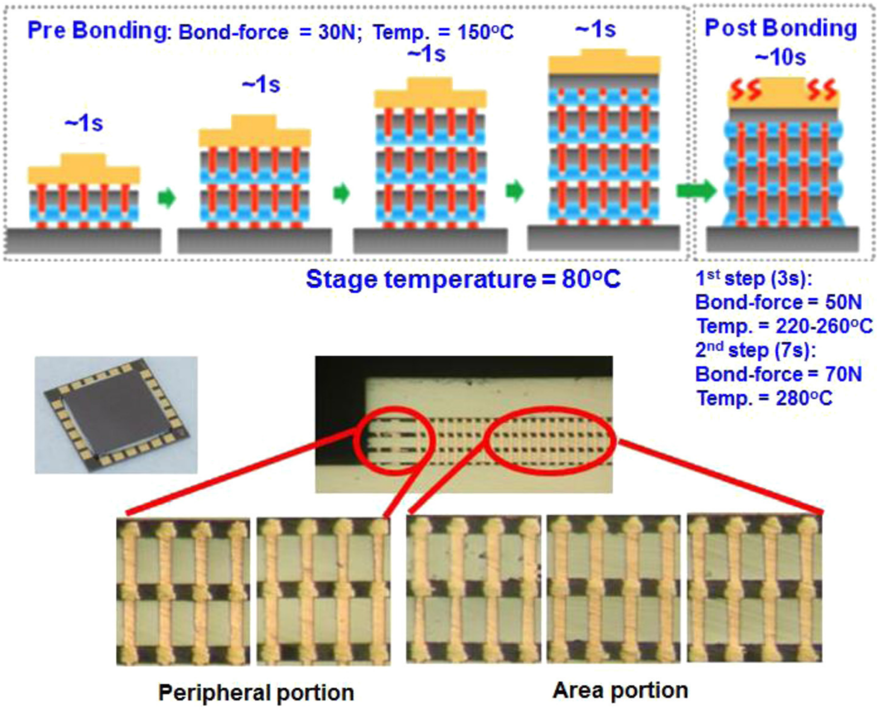


Fig. 1-10. Collective bonding process introduced by Toray in [30,43]. A four-layer die stack package has been produced through this process (middle-left). Image from [42].

stacked on top of each other using low forces and temperatures. Once completed, the entire stack is bonded at once, having the potential to significantly improve throughput. These 3D integrated packages are becoming more popular due to form factor advantages through reduced lateral area footprints, and electrical advantages via interconnection length reduction. Power savings, reduced parasitics and noise, and faster performance capabilities make this type of stacking a logical direction for the microelectronic packaging industry.

1.2 CMOS-Based Microsensor Array

A thorough description of theories, development, and operation of the microsensor array chip is provided in [23]. The microsensor array is a gridded array of stress and temperature sensors designed for a multitude of possible applications during TCB processes. The sensor chip was fabricated through a complementary metal oxide semiconductor (CMOS) technology, with both types of sensors deposited as n+ and p+ diffusions during a single-stage process. 95 μm square bond pads are placed in an 8 × 8 array with each bond pad localized by one of an X, Y or Z-orientated stress sensors and a resistive temperature detector (RTD) element. The sensor chip also has an on-board multiplexer, controlling the logic flow and pad switching sequence while reading measurements. The sensor chip was die-attached into a ceramic dual in-line package (CerDIP) and wire bonded using 25 μm 4N Au wire for electrical connectivity. For earlier stage experimentation, the bond pads were stud bumped with Au and coined.

1.2.1 Force Sensors

Each of the 64 stress sensors are arranged and connected in a Wheatstone bridge configuration in order to exploit the piezoresistive properties of the diffusions and measure stresses induced by force applications during TCB (Figs. 1-11a and 1-11b). Each Wheatstone bridge circuit (Fig. 1-11c) measures an excitation voltage caused by the compression or expansion of the piezoresistive force elements experiencing resistance changes. These resistance changes cause voltage changes which can then be converted to a force using a conversion factor extracted during calibration as follows:

$$S_t = \frac{\Delta V}{F} \tag{1-1}$$

where S_t is the amplified signal force sensitivity, ΔV is the measured excitation voltage in [V], and F, is the force applied in [N] that caused the excitation voltage. With the data acquisition system set to operate the sensor chip at a bias voltage of 3.3 V alongside a force signal amplifier of 250×, normalization must be implemented as:

$$S_N = \frac{S_t}{3.3V} \times \frac{1}{250} \tag{1-2}$$

The Z-stress sensor calibration procedure described in [23] resulted in S_t to equal 1.149 V/N ± 48.5 mV/N. This corresponds to a S_N of 1.39 mV/V/N ± 58 μV/V/N. Since the signals are captured on a transient basis, the time-dependence of the Z force experienced is calculated as:

$$F_i(t) = \frac{\Delta V_i(t)}{S_t} = \frac{V_i(t) - V_i(t_0)}{S_t} \qquad (1\text{-}3)$$

where i is the pad index ranging from 1 to 64, $F_i(t)$ and $V_i(t)$ are the forces and excitation voltages experienced at time t, respectively, and $V_i(t_0)$ is the initial, reference voltage when measurement begins. The ability of a sensor to detect X, Y or Z force changes, depends on the orientation of each Wheatstone bridge circuit and its applied dopant diffusion.

1.2.2 RTD Sensors

Top metal Al is wound around each bond pad in a serpentine shape (Figs. 1-11d and 1-11e) and wired in a Kelvin probe or four-wire measurement setup (Fig. 1-11f). As a bias DC current is applied to the terminals of each RTD element, resistance changes are measured as the temperatures experienced by the element are varied. These sensors provide absolute temperature measurements. While RTDs typically have the poorest response time of temperature measurement devices such as thermistors or thermocouples, the micron-sized RTDs used in our sensor chip exhibit virtually negligible response delays, in the range of micro-seconds. Due to the low mass of each micro-RTD sensor in the sensor chip, the localized temperature rises around the bonding sites can be readily captured, without a significant delay in response time. Therefore, these types of temperature sensors are better suited for our measurement purposes. Each sensor provides an average of the temperature experienced around the bond pad periphery within the perimeter of the RTD area. The RTDs were calibrated using customized minioven racks discussed in [44,45]. Starting with the relationship of the temperature coefficient of resistance α, and Ohm's law, $V = IR$, the formulation for the temperature experienced by each RTD at time t, is as follows:

$$\frac{V_i(T, t)}{I} = R_i(T, t) = R_i(T_{Ref}) \times [1 + \alpha_{Tref}(T_i(t) - T_{ref})] \qquad (1\text{-}4)$$

here $V_i(T,t)$ is the excitation voltage read for a pad index at a given time, I is the bias current applied, $R_i(T,t)$ is the resistance read, and $R_{T,ref}$ is the calibrated reference resistance at a predefined temperature. The time-dependant temperature can be then be re-formulated and simplified as:

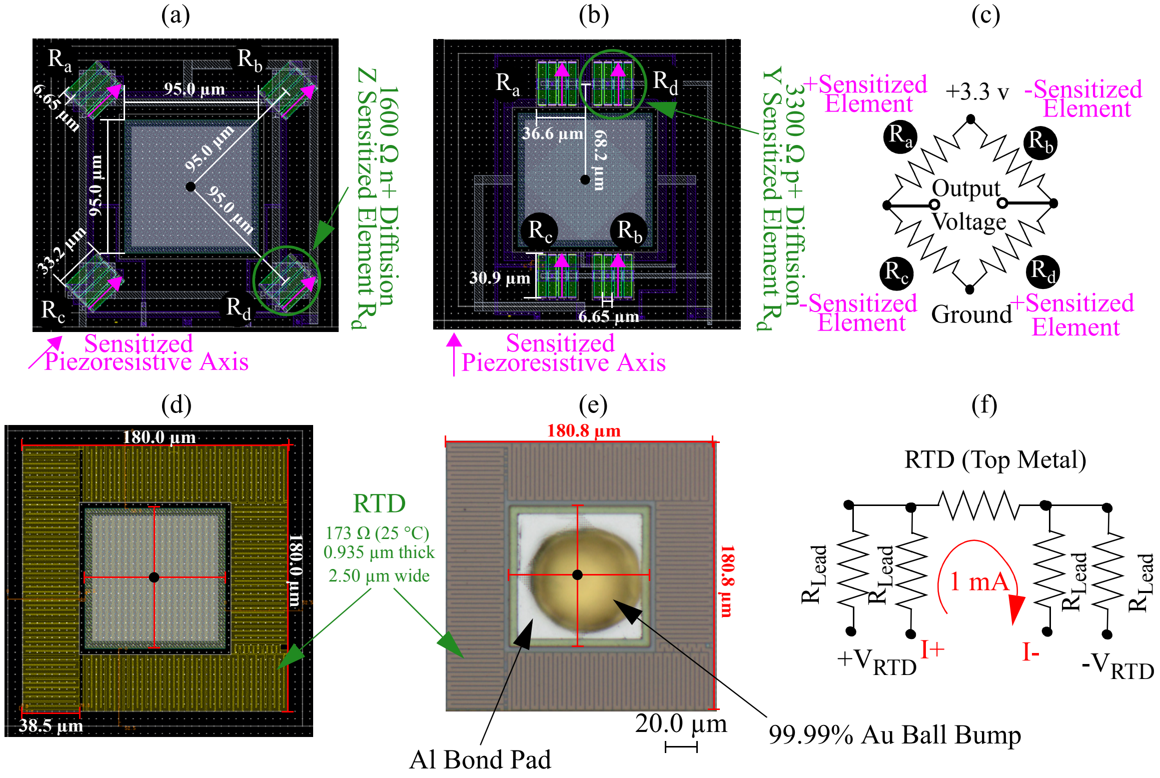


Fig. 1-11. (a) Z stress sensor orientation; (b) y-stress sensor orientation (x is 90° rotation-equivalent); (c) Wheatstone bridge circuit; (d) Design of RTD element around bond bad; (e) micrograph of RTD sensor localized around a bond pad; (f) four wire resistance measurement circuit. Reproduced from [34]

$$T_i(t) = \frac{1}{\alpha_{T_{ref}}}\left[\frac{R_i(T,t)}{R_i(T_{ref})} - 1\right] + T_{ref} \tag{1-5}$$

For a reference temperature of 50 °C, $\alpha_{50\ °\mathrm{C}}$ is 0.00277 1/°C and $R_{50\ °\mathrm{C}}$ is 186.94 Ω ± 0.89 Ω.

Upon assembly and characterization of sensor chips, TCB emulation events were conducted using a TCB bonder, available at the K&S Fort Washington R&D facility. Real-time responses were recorded. Thermo-mechanical features such as warpage and thermal expansion were identified [23]. Further thermal analysis was conducted, priming the results for a comparative finite element (FE) verification study.

1.3 Finite Element Methods

Finite element methods are first and foremost a means to numerically solve mathematical formulations describing the governing physics of a system of interest. These mathematical descriptions are usually constructed in the form of partial differential equations (PDEs), most of which are impossible to solve analytically [46]. They are tools crucial to the engineering design process for benefits such as cost savings, expeditious turnover rates and ease of access.

The foundation of original finite element methods (FEM) or analysis (FEA) can be attributed to Ritz for his work of using approximation methods to solve the mechanics of deformable solids, in 1909 [47]. Development continued with Courant in 1943, who expanded on Ritz's work, followed by many others such as Turner [48], Szmelter [49], and Clough [50], who was also the first to coin the term "finite element method". The abrupt development of "modern day" computers, alongside drive for FEM in civil and aircraft applications in the 1960s, led to vast developments in FEM. FEA software for structural analysis was first publicized and released in collaboration with National Aeronautics and Space Administration (NASA), aptly naming it NASTRAN [51,52]. About $701 million was saved by NASA between 1971 to 1984 from the sole introduction of this software [53]. FEM applications were eventually generalized to be applied to a variety of engineering problems involving heat transfer, fluid dynamics, and electromagnetics-related physics. Many software packages are now available for commercial and open access use including: ANSYS, ABAQUS, COMSOL, HyperMesh, LS-DYNA and COSMOS [47,54].

1.3.1 Model Formulation Process

The process of conducting FEA for a given physical phenomena can be divided into 5 main steps: physical problem definition, mathematical modelling assumptions, FE model parameters, error analysis and post-processing of results. Additional steps include analysis refining, mathematical model improvement, re-assessment of physical problem, model design improvement or optimization and refinement of mesh or FE model parameters [47,51,52,54]. A process flow of these 10 steps is provided in Fig. 1-12. The most important

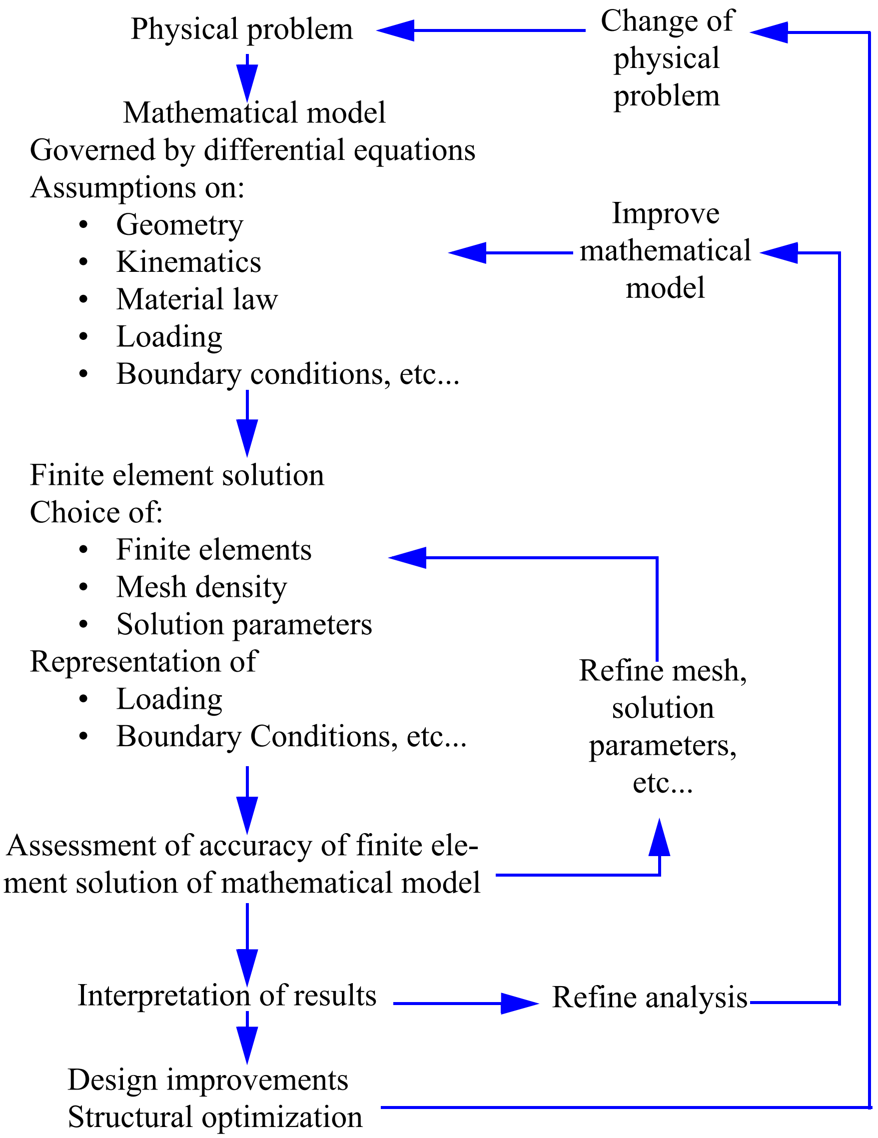


Fig. 1-12. General FEA process used in engineering analysis. Reproduced from [51].

step, prior to any actual modelling, is the problem definition. This must be thoroughly understood to ensure all possible factors are accounted for. This could include identifying the appropriate physics involved, any coupled-interactions between these physics, physical constraints, materials of components, and geometric specifications. Due to the complexity of many structures, several assumptions may need to be made to simplify mathematical modelling. Typically, one or more of the following simplifications are made:

- Partial component modelling (most important/relevant components only) [2,7,8,16,51 54]
- Symmetry modelling (fractional slice of full model used) [2, 54]
- Geometric simplification (minor components merged into larger domains) [55,56,57]
- Material property simplification (temperature-dependant to room temperature values) [2,16,33,58]

Once the problem has been clearly defined and any important mathematical assumptions have been made, the parameters can be inputted into the FEA software. Here, the geometry is constructed, material parameters are inputted and assigned to components, boundary conditions are implemented, study conditions are applied and finally, mesh is applied. The mesh is crucial for accuracy, convergence and is an important factor in computation time. Meshing refers to the discretization of components into several elements with nodes. This is where the numerical methods, built on those developed by Ritz et al., play a role in the solution of the governing PDEs. Solutions to the mathematical equations governing the physics of the system are numerically approximated through linear combinations of basis functions. These equations are applied to each element. The approximations to these equations are made by combining the values assigned to each variable at each node with the basis functions applied to each element. Therefore, the greater the number of nodes, the closer the approximations get to the actual solution. The software assembles the localized basis functions at each element and iteratively solves them until a converged, continuous solution is resolved. Several errors could arise in this calculation and therefore, special care should be taken when analyzing post-processed results. The aforementioned additional steps are conducted if the accuracy of the results is not within acceptable limits.

1.3.2 Heat Transfer Analysis

Heat transfer is the motion of thermal energy caused by thermal gradients within or between media. These temperature differences result in three main modes of heat transfer:

1. Conduction - heat transfer within a medium; thermal energy is transferred through collisions or vibrations in fluids and solids
2. Convection - heat transfer between a surface and a fluid in motion

3. Radiation - heat transfer via electromagnetic waves (photons produced by scattering, emission or absorption)

In TCB, the dominant mode of heat transfer is conduction. Conduction is fundamental to the process as heat from the bond head propagates throughout the solid components to the joints, in order to form bonds. Ambient (air) convection plays a role in package and bond head cooling as heat reaches the surfaces of the components and should be considered in all TCB heat transfer analyses [12, 32].

1.3.3 Thermal Contact Resistance

Two solid objects in thermal contact with each other typically experience a temperature drop across their interfacial boundary. As heat is conducted from one object to the other, resistance is experienced at the boundary, resulting in a drop in heat flow, and correspondingly, a drop in temperature. There are several factors that be attributed to affecting thermal conductance such as surface topography, material microhardness, contact pressure and interstitial fluid properties [59,60,61,62]. An example of resistance could be due to pockets of interstitial fluid that are found between the surfaces of each boundary as thermal contact is made (Fig. 1-13). These pockets can exhibit poor thermal conductivity and can reduce heat flow across their interfacial boundary. On the microscopic level, every surface, regardless of how smooth or “planar” it may optically appear, can have peaks and troughs

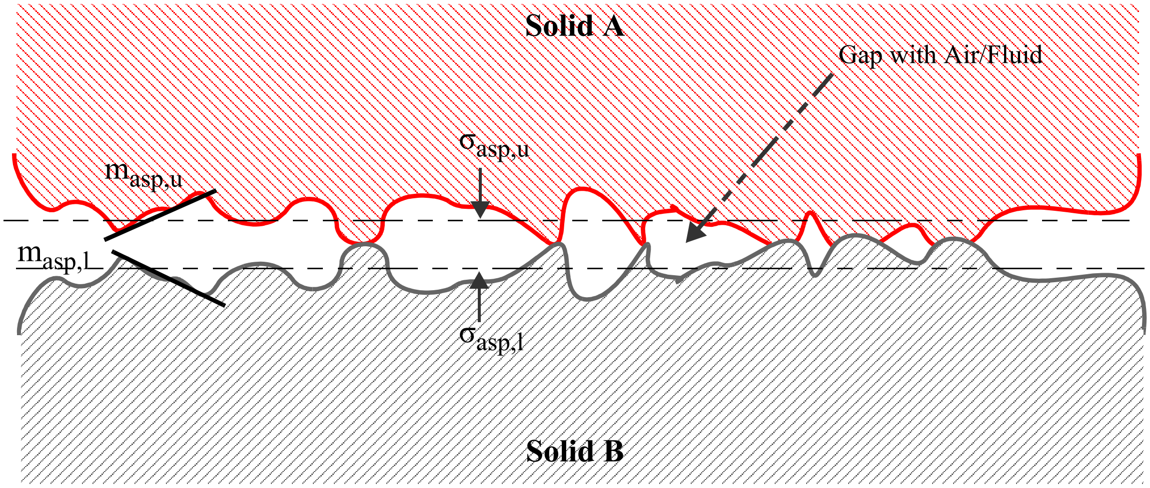


Fig. 1-13. Solid A in contact with Solid B. The peaks and troughs of each solid are called surface asperities. “m” and “σ” are the root mean square (RMS) slopes and average heights of the asperities, accordingly. “u” and “l” denote upper and lower boundaries.

called surface asperities. The study of these types of surfaces is called surface microtopography. Ignoring the effects of thermal friction, the mathematical relation to describe the heat flux across the upper boundary is

$$-\boldsymbol{n}_l \cdot (-k_l \nabla T_l) = -h(T_u - T_l) \tag{1-6}$$

and,

$$-\boldsymbol{n}_u \bullet (-k_u \nabla T_u) = -h(T_l - T_u) \tag{1-7}$$

.for the lower boundary. The thermal conductance at the interface *h,* is the summation of three forms of conductances: the constriction conductance, the gap conductance and radiative conductance. Conductance describes the quantity of heat transferred through a surface of area *A*, with thickness *th,* for a given unit of time assuming there is a temperature difference between the two objects. For our purposes, the key terms are the constriction conductance, h_c and gap conductance, h_g. These are the conductances that corresponds to conductive flow between asperity-asperity contacts, and the conductive flow across the air gaps between thermal contacts. The most used and widely accepted formulation for thermal contact conductance considering the surface roughness model is the Cooper-Mikic-Yovanovich (CMY) correlation,

$$h_c = 1.25 k_{contact} \frac{m_{asp}}{\sigma_{asp}} \left(\frac{p}{H_c}\right)^{0.95} \tag{1-8}$$

which includes the sub-contrary mean of thermal conductivity as $k_{contact}$ in W/m•K, contact pressure *p,* in Pa, and Vicker's microhardness $H_{c,}$ in Pa [62,63,64].

With respect to TCB, thermal contact is typically made between the ICs and their substrates, during bonding processes. As the chip is lowered and pressed down onto the receiving traces or pads, poor conductive heat flow could arise in the earliest stages of bonding. This is especially true for solder-capped Cu pillar processes where good thermal contact is required for reflow. Thermal contact resistances could also be observed between other components, such as the bond head assembly or between other non-adhered components in the package. These resistances could consequently slow down processes and lead

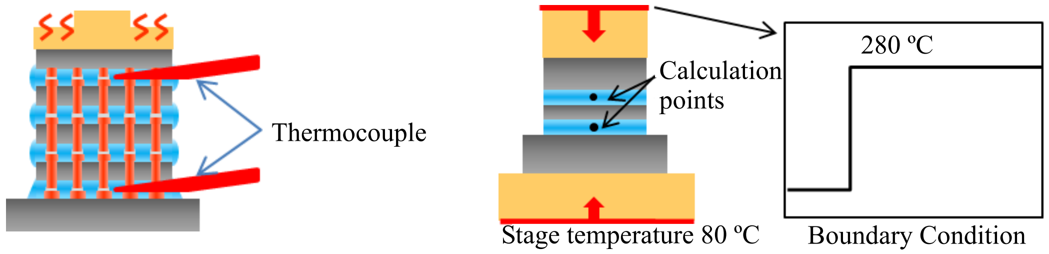


Fig. 1-14. Thermocouple attachment locations in experiment (left) and simplified model with two layer stack used in simulation model (right) [30].

to reduced throughput. Consequently, thermal contact resistances are important to implement in FEA of TCB processes to observe their effects and understand their role in thermal processes.

1.3.4 Current FE Thermal Analyses in Literature

1.3.4.1. Two-Dimensional Transient FE Model

Thermal modelling analysis is not new to the field of microelectronic packaging. Nor is it new to the area of flip chip TCB either [66, 67]. Asahi et al. [30] modelled a four-layer die stack package for collective bonding TCB with thermocouples attached to NCF layers using transient thermal analysis. They employed numerical methods to measure the simulated temperatures at the locations of the thermocouples in their test package (Fig. 1-14). To extract more from their model, they reduced the thermal conductivity of the stage from 1.7 W/m•K to 0.3 W/m•K, making it significantly more insulative. The result was that the simulated temperature signals between the two bondlines reduced the temperature difference between the two layers, reducing vertical thermal gradients within the stack. While this paper is unique for its insight on the effects of stage conductivity, it does not provide a complete thermal analysis. It is limited in its model by the several assumptions made: the analysis is two-dimensional, several components are missing such as pillars and vias, lateral thermal distributions are not presented and the boundary condition on the heater is unrealistic (step function). While these assumptions are acceptable for this specific analysis, many important features are not implemented in this model. This model was extended in 2016 to observe the effects of die thicknesses and air cooling on lateral thermal distributions across the substrate, but did not further address any other limitations [65].

1.3.4.2. Three-Dimensional Transient FE Model

Another interesting study which addresses some of these neglections was conducted by Jeong, Choi et al. [32]. They used experimental and FE methods optimize process bonding temperatures. Using a flip chip very fine ball grid array (fcVFBGA) test vehicle with a 11 mm × 11 mm × 100 µm chip attached as a test vehicle, temperatures were measured in the NCP layer close to periphery pillar locations indicated in Fig. 1-15 Once

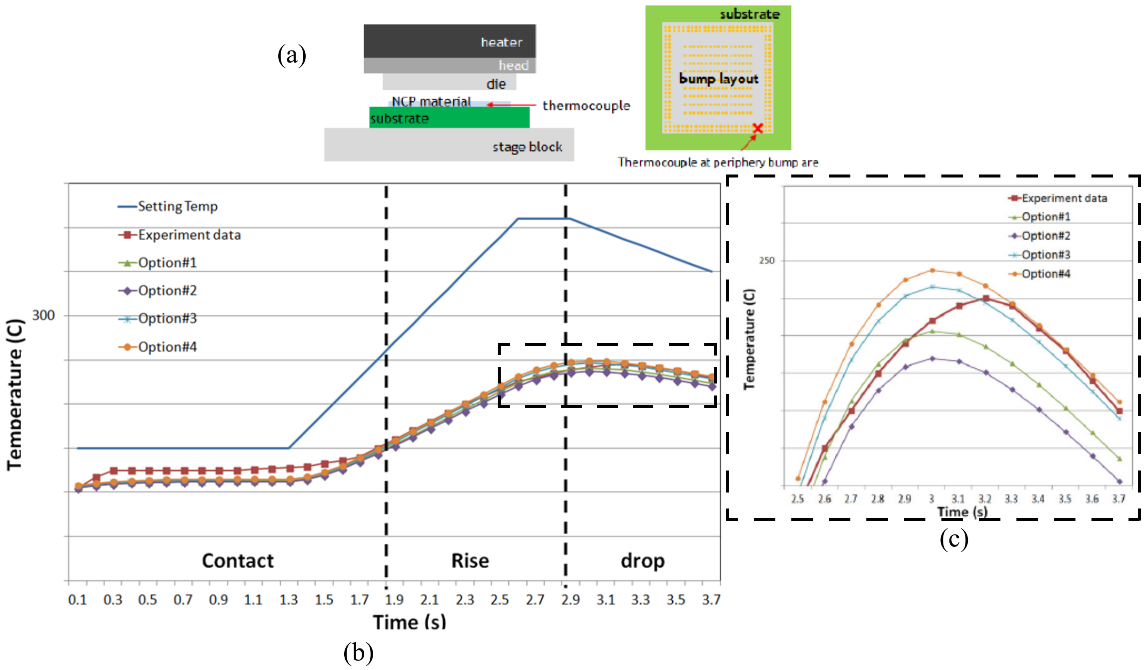


Fig. 1-15. (a) The experimental setup and location of thermocouple attached (top); (b) Nominal heater, experimental and simulated thermal responses; (c) magnified responses. Options #1-3 are the results of the varied homogenized layer trials tested in the simulated. For example, option # 1 homogenizes the pillars, solder and NCP material as one layer with effective material properties. Option # 4 is a complete model with all components and materials modelled independently [32].

again, the use of thermocouple measurements in the bondline layer was used to approximate bonding temperatures. FE models were developed to verify these response using variations of component simplifications denoted options #1-4. These models are significant because they are three-dimensional, include various configurations of simulation geometries ranging from over-simplified to thoroughly complete and show reasonably good match with experimental responses. They are an improvement over the previous model discussed, drawing a more complete thermal analysis of bonding in TCB. They conclude that thermal control of die, bondline and substrate components are key to optimized bonding processes. Once again, these models are limited due to lack of insight on lateral thermal distributions, effects of thermal contact resistances and most importantly, pillar temperature variation. Anisotropic material properties for the substrate, which are important due to their significantly different in-plane thermal conductivities, are also assumed negligible. The experimental measurements were only taken at one location and this thermocouple was embedded in the NCP at the die periphery.

1.3.4.3. Three-Dimensional Transient Thermal Model with Contact Resistance

The study of the greatest interest and relevance to this thesis was recently published by Bex. et al. [36]. One of the goals of this paper was to measure interfacial bonding temperatures and compare this to the results of FE models to better understand heat transfer mechanisms during TCB reflow processes. Using RTDs in a similar manner to those mentioned previously, in-situ thermal measurements were measured during experimental TCB processes. A total of three RTDs were situated at the locations indicated in Figs. 1-16a, 1-16b and 1-16c. Two main types of experimental TCB procedures were conducted to extract process temperatures: one which measured chip and substrate temperatures (Fig. 1-16b) and another which measured lateral temperature distribution across the substrate (Fig. 1-16c). The test vehicles also had about 128 microbumps that were bonded during these experimental TCB processes. The 3D thermal transient model responses deviated from the experimental responses in terms of differences in temperatures measured. Nonetheless, the models provided further insight on thermal variation due to I/O orientations and counts as well as vertical and lateral distribution across substrate and chips. Areas with no I/Os resulted in colder regions than those with I/Os due to heat sink effects caused by the Si. A temperature difference between centers and corners of the PTCP chips is also attributed to this effect. Thermal contact resistance was accounted for between bottom heater and bottom chuck. It is unclear how the formulation was implemented, but it is believed that predefined default settings were used in the FEA software, MSC Mark™.

Of the three FEA types discussed, this is the most relevant to this thesis. The work of this thesis intends to help close the gap to form a more complete understanding of thermal flow in TCB processes by adding on to studies such as these in a thorough, methodological manner.

1.4 Goal of this Thesis

In general, the goal is to better understand thermal distribution and heat transfer mechanisms within TCB processes. Post-processing thermal responses measured via the micro-sensor array during TCB emulations will be combined with the use of FE models. This combination is to be used to observe and identify any factors affecting uniform heat flow

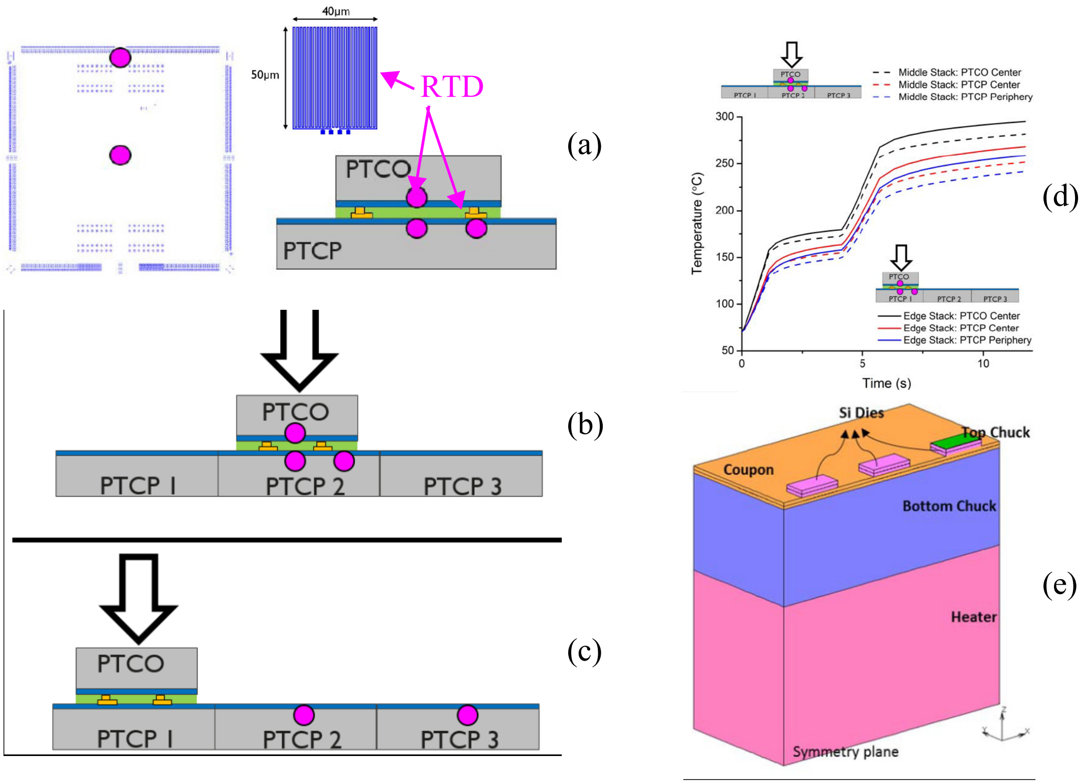


Fig. 1-16. (a) The locations of the three RTD sensors used (magenta circles); (b) the first and (c) second experimental TCB procedure conducted; (d) Simulated thermal responses from FE model of experimental (b); (e) FE model design [36]. PTCO is Packaging Test Chip Version O and PTCP is version P [36].

during TCB processes as well as suggest any recommendations to improve thermal uniformity.

In particular, the objectives of this thesis are:

- To develop a FE model that fits the experimental data
- To identify model parameters or components that significantly affect thermal uniformity or rate of heat flow
- To develop an advanced FE model based on surface roughness thermal contact resistance theories
- To quantify any thermal variation data and analyze the causes
- To learn which design options engineers have to improve TCB for stacks of large chips with many bumps
- To provide a process for numerical-experimental analyses that can be extended for package and process-specific applications

2 Sensor Chip: Temperature Signal Analysis

This Chapter provides an overview of the experimental procedure conducted, and provides the results extracted from post-processing of the thermal measurements. It is included to show the pre-FEA work from which values were derived and later used in the development of FE models. It is important to illustrate and clearly define the steps taken in order to effectively reduce the physical problem into more understandable domains and ensure accuracy of the models. Moreover, the inclusion of these steps is intended to provide readers the necessary background to replicate the FE models, in hopes of enhancing their FEA abilities, and expediting further development of TCB processes. All post-processing of experimental data are conducted in MATLAB®.

The research reported in this Chapter has been published in parts in *Microelectronic Reliability* [34] and in the *Proceedings of 67th IEEE Electronics Components and Technology Conference* [16].

2.1 TCB Experimental Setup

The experimental procedure of focus in this study is an emulated TCB process conducted using a K&S TCB bonder. Since bonding is not a focus of this study and did not occur in our experimental procedure, we are only partially replicating the TCB process. Furthermore, the IC (sensor chip) with the bumps is attached to the substrate facing upwards, whereas it would typically be the flipped-chip bonded downwards in the actual process. Therefore, this is an inverted TCB emulation. Nonetheless, the procedure can provide representative measurements of in-situ temperatures experienced during similar TCB processes, particularly in the case of Au stud bump bonding.

The full experimental procedure is illustrated in Fig. 2-1. During this procedure, a square prime Si die was used distribute uniform pressure amongst the bumps. The TCB

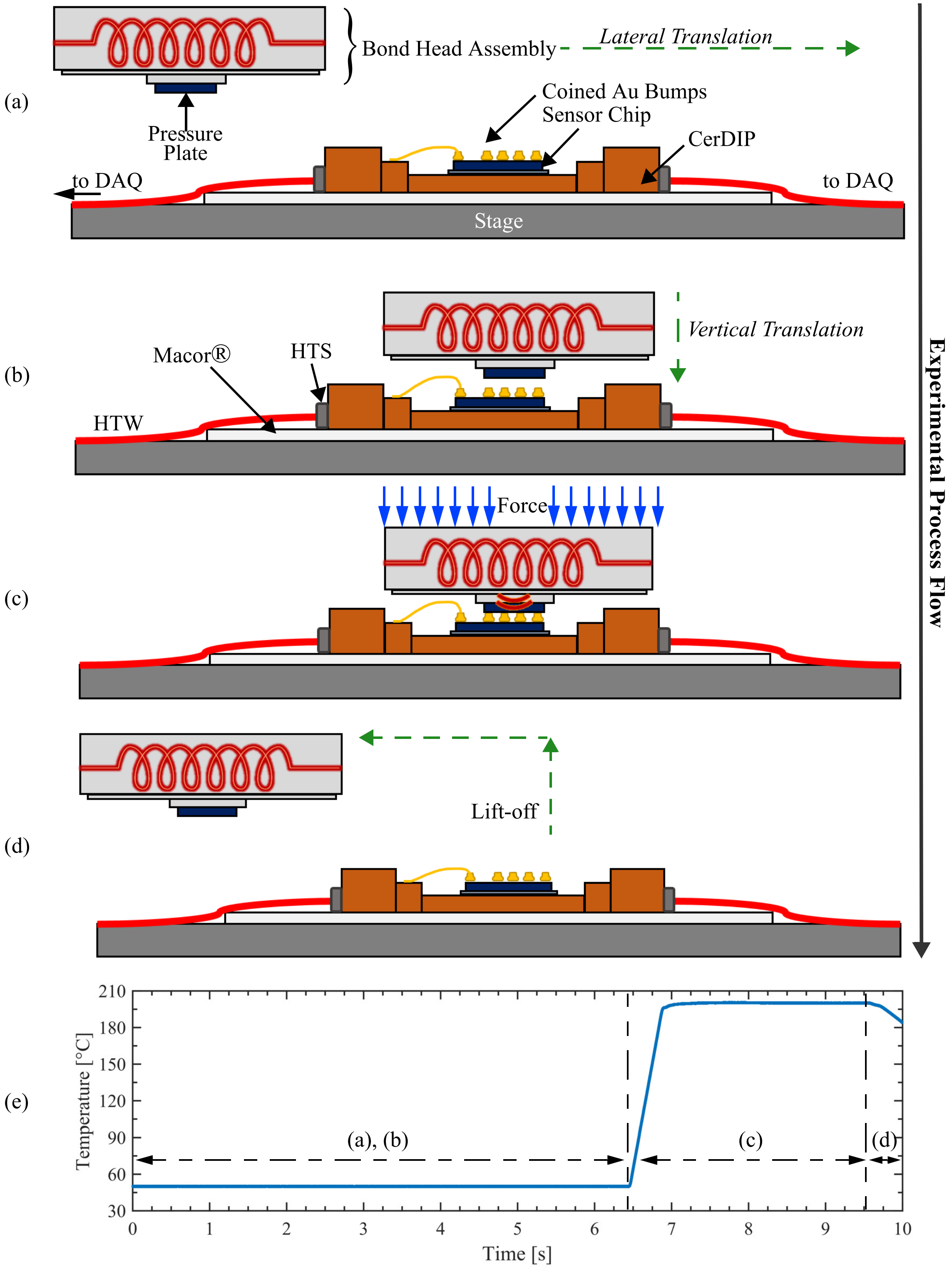


Fig. 2-1. (a) Bond head picks up the pressure plate and all components are heated to 50 °C. Measurement begins; (b) bond head aligns the pressure plate with the bumps and moves downward; (c) a force and temperature ramp is applied and held steady; (d) the bond head lifts off, cools and measurement ends; (e) bond head thermal profile as measured. HTS and HTW denote high temperature wire and solder, respectively. DAQ denotes the data acquisition system.

bonder picked up and redundantly flipped this die, preheating it to a steady temperature of 50 °C. The substrate assembly was also held steady at 50 °C and the stage remains at this temperature throughout the duration of the trial. The DAQ powered the sensor chip and began recording the measurements. After the pressure plate was aligned with the bump array area using the bonder's vision system, the bond head applied a nominal ramp of 350 °C/s, to reach a steady "bonding" temperature of 200 °C from 50 °C. While, this thermal ramp was applied, a 80 N force profile was also applied. These process conditions were held for up to 3 s prior to lift off. A three-factor factorial study was conducted by varying forces applied, bond head heating rates and nominal "bonding" temperatures. This resulted in twelve total trials conducted. However, several process issues, such as excessive tilt and measurement faults, resulted in responses inadequate for thermal transient analysis. For example, excessive tilt lead to non-compliant contact where many bumps either lacked contact or experienced non-uniform force distribution. Consequently, temperature signals from these types of trials were insufficient and inaccurate for this particular study. Therefore, only the responses from the final trial are considered in the remainder of this thesis. Due to unforeseen constraints, these trials were unable to be repeated at the K&S, Fort Washington R&D facility.

The raw RTD signals were converted into temperatures using Eqn. 1-5, via the algorithm in Appendix A. The measured RTD responses from the last trial conducted could be divided into four main regions of interest as indicated in Fig. 2-2.

These four regions/points are denoted the following:

1. MFT - Maximum Final Temperatures
2. MHR - Maximum Heating Rates
3. FTRS - Final Third Region Slopes
4. KP - Knee Point

Additional regions are indicated for the bond head profile including the time until ramp $\Delta t_{steady,}$ ramp duration $\Delta t_{ramp,}$ duration of bond head press $\Delta t_{press,}$ and lift off/cooling time Δt_{cool}.

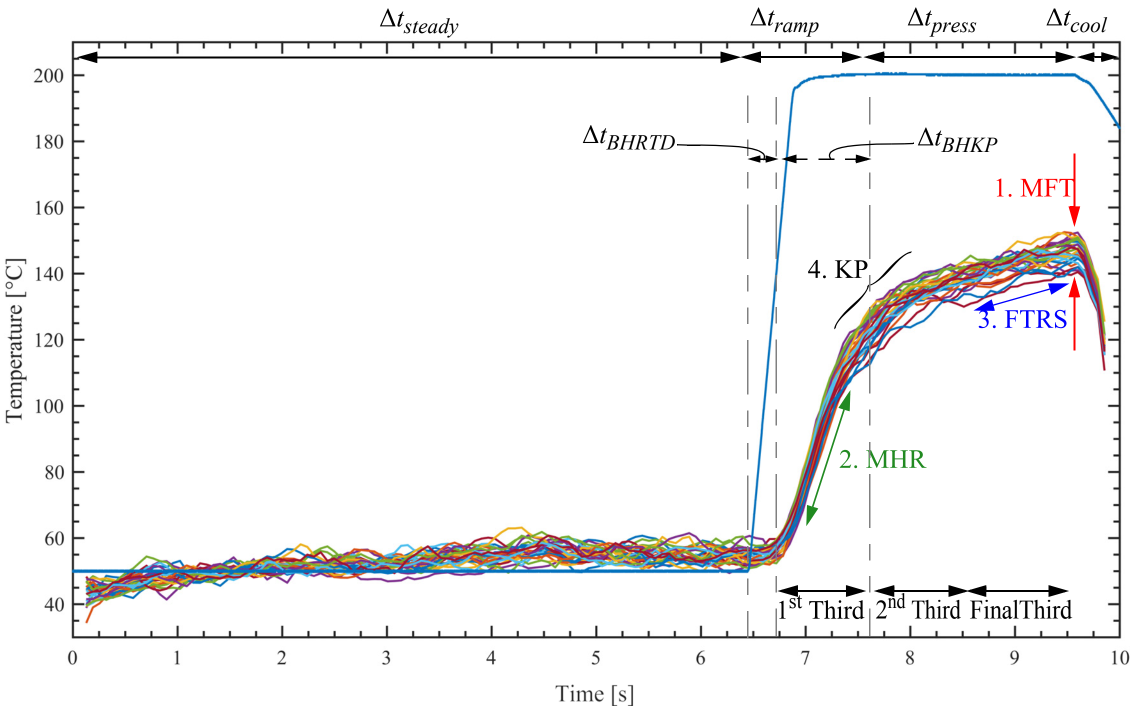


Fig. 2-2. Measured experimental RTD signals and measured bond head profile. The thirds indicated on the bottom right are arbitrary divisions of the thermal response from ramping to the MFT region.

2.2 Simple Extractions: MFT, MHR and FTRS

The MFT is simply the mean of the highest temperatures experienced by each RTD sensor. Ideally, this temperature should be as close to the nominal temperature as possible. If processes can be optimized to reduce the temperature difference between set process temperatures and IC temperatures, the set temperatures can possibly be reduced. In turn, this would improve throughput and reduce power consumption in TCB processes. The MHR is the average of the maximum thermal rates experienced during ramping (1^{st} Third). The average MHR of all signals can be defined as:

$$MHR \approx \frac{\left(\sum_{i=1}^{n} \left.\frac{\mathrm{d}T}{\mathrm{d}t}\right|_{i,\,max} \right)}{n} \tag{2-1}$$

As seen in Eqn. 1-13, the heating rate is inversely proportional to the specific heat capacity C_p, and proportional to the thermal conductivity k. The rapid heating of the RTD sensors can be attributed to the high conductivities and low heat capacities of the bond head components, pressure plate and Au bumps. Finally, the FTRS is the average of the linearly interpolated slopes of each signal between 8.5 s and 9.5 s. The generalized code for rate calculations is in Appendix A. Table 2-1 summarizes the results from these calculations. The values for the temporal divisions in Fig. 2-2 are extracted as follows:

1. $\Delta t_{steady} = 6.45 \text{ s}$
2. $\Delta t_{ramp} = (7.34–6.45) = 0.89 \text{ s}$
3. $\Delta t_{press} = (9.59–7.34) = 2.25 \text{ s}$
4. $\Delta t_{cool} = (9.99 - 7.34) = 0.40 \text{ s}$

Table 2-1. Summary of experimental response quantifications

Response Feature	Mean	Standard Deviation	Min	Max	Range
MFT [°C]	147.47	3.20	140.54	152.54	12.0
MFT Times [s]	9.49	0.17	9.09	9.6	0.51
MHR [°C/s]	92.06	6.91	80.82	108.1	27.27
MHR Times [s]	7.06	0.03	6.98	7.1	0.12
FTRS [°C/s]	7.94	2.55	2.66	12.76	10.1

These compose the ≈ 10 s total recording time. The RTD measurements and bond head profile are also analyzed to observe thermal delays and features that could be relevant to TCB processes. The measured (calculated) bond head heating rate is 348 °C/s. The thermal delay between the start of the bond head ramp and the start of the RTD ramp Δt_{BHRTD} is about 230 ms, indicating it took about a quarter of second before heat from the bond head reached the sensor chip. This time could be an important parameter in TCB processes, particularly for stacked die packages as throughput can be improved by reducing the required time to reach process temperatures. Typically, faster higher ramping rates, reduced die thicknesses and material improvements are being implemented to reduce this time.

2.3 Knee Point Region Extraction

An additional region of interest is the knee point (KP) region. The KP is defined as the point where the rapid primary ramping region slows into a secondary, steadier-ramping region. Physically, this region is due to thermal saturation of heat from the bond head as the heat is rapidly drawn down through the sensor chip and substrate assembly. Initially, the low heat capacities and high thermal conductivities of the bond head, pressure plate, sensor chip, and Au bumps lead to swift temperature rises. As this heat is further drawn down into the substrate assembly with components that have very low thermal conductivities and high heat capacities, the heating rate slows. Thus, a sluggish rise in temperature is experienced after this transition zone or knee point. The generalized knee point algorithm for a curve is provided in Appendix B.

The procedure to calculate the knee point temperature and time starts by selecting a starting point with a time situated reasonably before the KP region. This point initiates the process by forming the first linear interpolation with the point to the right of it. Concurrently, a second linear interpolation is made with all of the points on the right of the first point up to the end of the curve. Iteratively, one data point on the right is added to the first linear interpolation while one on the outmost-left is removed from the second interpolation. This is looped until an intersection occurs between the two slopes where the intersection point defines the KP. An example of an intersection found during this procedure is illustrated in Fig. 2-3. Table 2-2 summarizes the result of the KP extractions.

Table 2-2. Summary of KP calculations.

KP Feature	Mean	Standard Deviation	Min	Max	Range
Knee Temperature [°C]	122.15	3.63	113.43	129.30	15.87
Knee Temperature Times [s]	7.60	0.09	7.45	7.79	0.34
Δt_{BHKP} [s]	1.14 s	0.09	0.99	1.34	0.34

The temperatures of the RTDs reached 59.3% of the mean MFT at the mean MHR point, and 83% of the mean MFT at the mean KP. This tells us that 1.14 s after the bond head

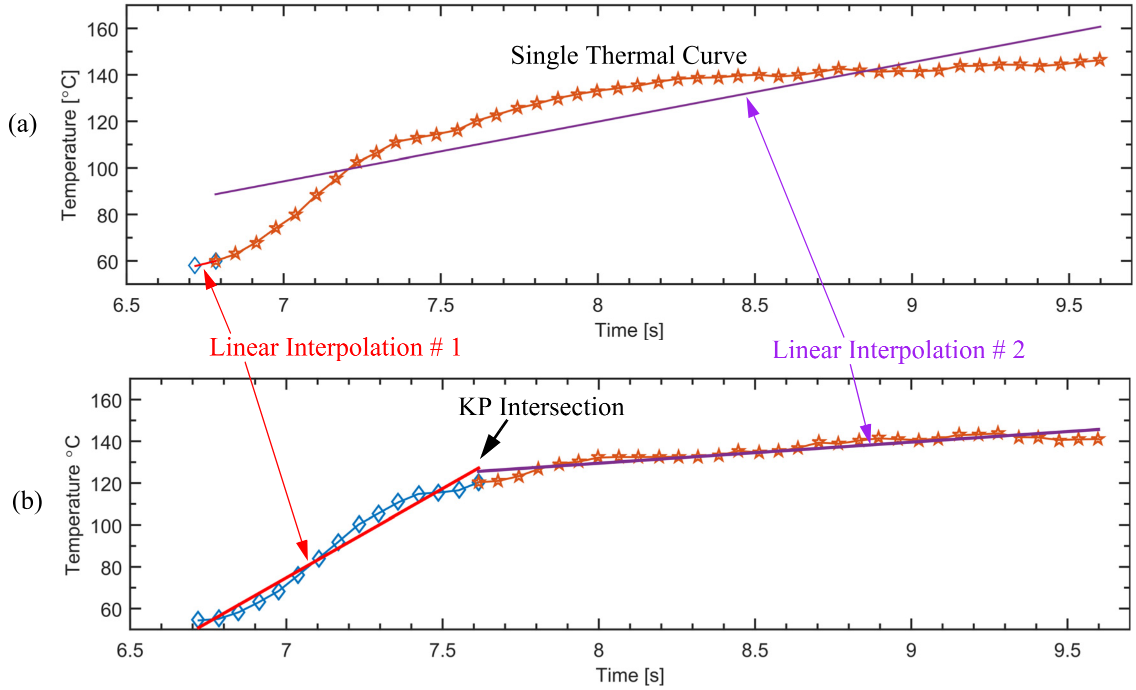


Fig. 2-3. KP algorithm example plots; (a) The first two linear interpolations made; (b) The intersection point. Note: this algorithm continues until all points have been considered, even if the intersection has been reached.

starts ramping to 200 °C, the sensor chip is heavily saturated with heat energy. It also tells us that most of the heating occurs during this ramp, as the temperature of the sensor chip only rises a further 17% in the last 2 s of the press. The KP is important because it characterizes the heating curve by providing an approximation of when a heat energy saturation point is reached and the corresponding temperature at this point. Processes can possibly be optimized to reduce the time until the KP is reached while maximizing the temperature reached at this point. This would produce a thermal profile that would more closely resemble the shape of the nominal bond head profile. Hence, improvements can be made to the throughput of TCB processes.

2.4 Thermal Gradient Analysis

Thermal gradients describe the change of temperature over a distance, written in its differential form along the *x*-direction, in Eqn. 1-8. For this case, the thermal gradient magnitudes across the sensor chip surface are considered to determine the locations and times

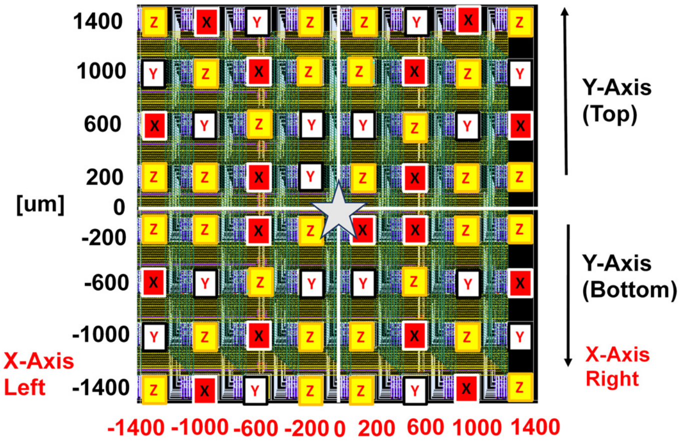


Fig. 2-4. Arbitrary division system overlaid on the bump array area. The X, Y and Z's denote the locations of types of each force sensor at that location. The star signifies the origin.

of the greatest thermal gradient magnitudes. The goal is to measure, observe and understand what causes the temperature spread or range amongst the sensors.

The code implemented to determine the gradients and produce the following visuals is in Appendix C. The sensor chip's bump array area is arbitrarily divided, producing an XY Cartesian plane where each sensor assumes a coordinate position *(x,y)* as seen in Fig. 2-4. Bilinear interpolation techniques are then used to calculate the temperatures within the predefined grid and then differentiated over distance to extract the thermal gradients in each x and y directions. The magnitude of each gradient vector, assuming no changes in elevation or height z, can be simply calculated as:

$$|\nabla T(x, y, t)| = \sqrt{\left(\frac{\partial T}{\partial x}\right)^2 + \left(\frac{\partial T}{\partial y}\right)^2} \tag{2-2}$$

Figure. 2-5 captures thermal contours and gradients prior to ramping, at the MHR region, at the maximum gradient experienced time, and at the MFT region. The top-right corner of each quiver gradient plot indicates the location of the greatest thermal gradient magnitude experienced. During prior experimental trials, the Au bump in this location became overly deformed, remaining colder due to lack of contact with the pressure plate. Therefore, this location always experienced the highest thermal gradient magnitude.

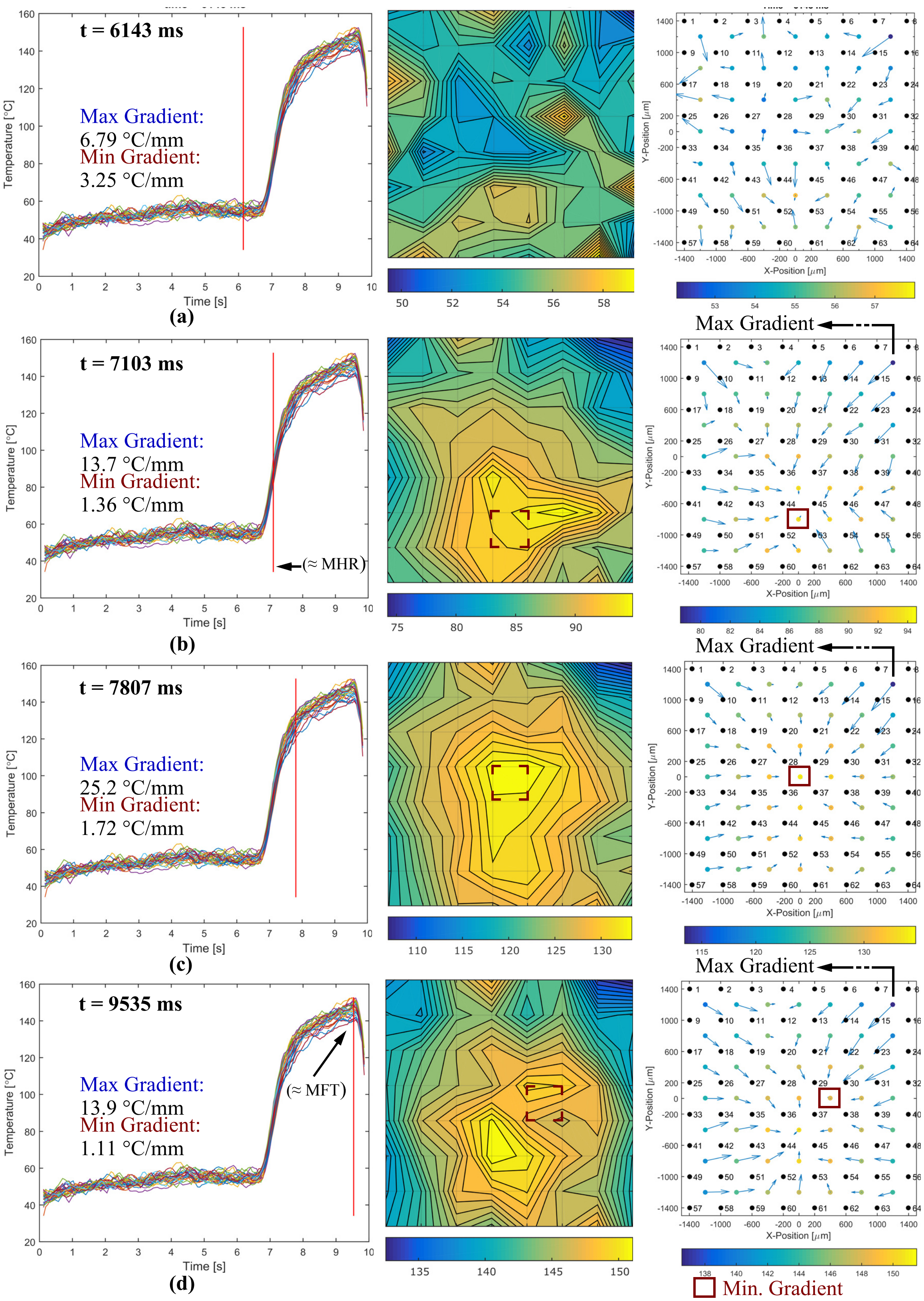


Fig. 2-5. Progression of thermal distribution at key locations (a) before ramping; (b) at the MHR time; (c) at the time of greatest gradient experienced; (d) end of press near MFT time. The middle column contains plots of contours at the times indicated in the first column. The third column contains the gradient quiver plots.

The left-most and right-most regions of the contour and quiver plots show lower temperatures while exhibiting relatively large thermal gradients. The left-most regions experienced the following highest maximum gradients, measured up to 15.2 °C/mm. This is due to the Si in the sensor chip around the bump array periphery drawing heat away from the RTDs. The time of the maximum thermal gradient magnitude experienced is about 200 ms from the mean KP point, 800 ms from the mean MHR time, and about 1.36 s from the beginning of the bond head ramp. This time and thermal feature is particularly important when coupled with mechanical effects due to thermal expansion. Thermal gradients result in non-uniform strains which can further lead to reliability issues if these effects manifest as cracks or voids. The ranges of RTD temperatures and maximum thermal gradient magnitudes are shown in Fig. 2-6.

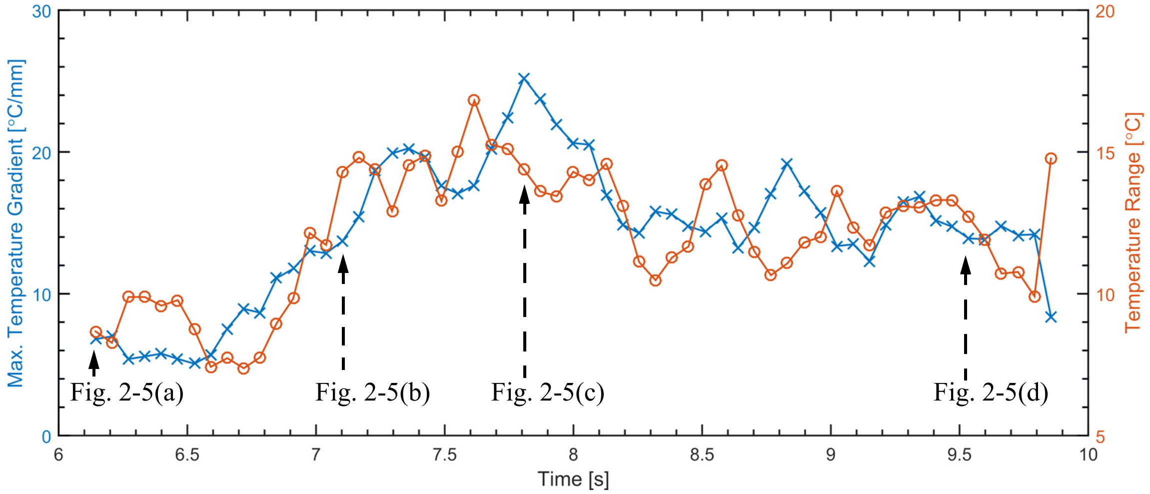


Fig. 2-6. Transient curves of the thermal gradient magnitudes and temperature ranges between 6.1 s and 9.6 s.

2.5 Summary

The experimental thermal results of a temperature microsensor array subjected to TCB emulation process is post-processed to analyze sensor surface temperatures, heating rates, knee point regions, thermal gradients and bond head profile characterization. A mean thermal delay of 230 ms is calculated between the start of the bond head ramp and the start of the RTD ramps. The maximum final temperatures experienced is calculated to be 147.47

± 3.2 °C, about 53.5 °C lower than the set process temperature. The maximum heating rates of the RTDs is 26.3% of the nominal heating rate, with a mean of 92.06 ± 6.91 °C/s, occurring at 7.06 ± 0.03 s. At this time, 59.3% of the maximum final temperatures were reached. The knee point, which occurs at 7.6 ± 0.09 s or about 500 ms after the maximum heating rates, indicates that 83% of the maximum final temperatures is reached after about 1 s of heating due to saturation of thermal energy. A maximum thermal gradient magnitude of 25.2 °C/mm is calculated to occur about 200 ms after the mean knee point. Throughout the duration of the trial, the periphery RTD regions experienced greater thermal gradients due to the sensor chip acting as a heat sink, effectively reducing the local temperatures around each bump. This heat sink effect also causes the spread in the temperature signals.

3 Finite Element Modelling

This Chapter reports on FE modelling strategies, formulation, implementation and post-processing of thermal transient 3D models. FE models are developed to investigate the effects of material property parameters and component inclusions on thermal responses. An iterative design of experiment (DOE) procedure to match FE models to experimental results, presented in Chapter 2, is discussed. The effects of heater thickness and air content in thermal contact resistance layers, on thermal responses, are investigated for the simple thermal models. The effects of surface asperity slopes and mean heights are investigated for the complex thermal model. Finally, both lateral and vertical thermal distributions determined from the models, are presented. The following FE modelling work, including all pre-processing, meshing, and post-processing was conducted in COMSOL Multiphysics 5.0®. The computer used to conduct the simulations has the following specifications:

1. Processor: Intel® Core™ i7-4770K CPU
2. Clock Speed: 3.50 GHz
3. Installed Memory: 32 GB (31.7 usable)
4. System type: 64-bit Operation System (Windows 7 Enterprise)

The research reported in this Chapter has been published in parts in the *Proceedings of 67th IEEE Electronics Components and Technology Conference* [16] and *Microelectronics Reliability* [34].

3.1 Strategy of Modelling

A sequential modelling process is employed to abide to the typical FEA process procedure illustrated in Fig. 1-12. The key experimental responses of comparison are provided in Chapter 2. Using these responses, the goal is to match FE models to have good agreement in terms of MFT, MHR, FTRS, and/or the KP. As discussed in Chapter 1, it is important to first simplify the physical problem in order to simplify the mathematical models and hence, the computational complexity. To address this issue, the FE models can be classified under one of two domains, each with a specific set of assumptions: simple thermal and complex thermal. The chief difference between the two is the method of thermal contact resistance (TCR) implementation between components in mechanical contact. The remainder of the modelling process, in addition to model parameters such as material properties and boundary conditions, is identical for both domains. The simple thermal TCR implementation is discussed in Section 3.2.6, and Section 3.10 for the complex thermal TCR.

With respect to the remainder of the modelling plan, Fig. 3-1 illustrates the sequence of steps and summarizes the assumptions made in the development of FE models, for both domains. The iterative DOE procedure is described in Section 3.3. The physical problem is resolved such that all components are assumed stationary and unified. The bond head assembly is assumed to be initially in contact with the Au bumps on the sensor chip, and any modelling of bond head translation is not included. During the preliminary mathematical modelling stage, a factorial variation of model geometries are simulated to understand the effects that each component has on the thermal responses. An overview of this

Table 3-1. Factorial models to study effects of component additions. A total of 22 models resulted from this study ranging from M0 to M6f. The × denotes the components included for each model [16].

Model # / Components	**M0**	**M1**	**M2a**	**M2b**	**M3a**	**.**	**.**	**.**	**.**	**M6f**
Base Components: Bond Head Assembly, Air Gap, Sensor Chip	×	×	×	×	×	.	.	.	.	×
Air Holes in Bond Head and Place Tool		×		×		.	.	.	.	×
Bumps			×	×		.	.	.	.	×
TCR Layers (Bond Head/Place Tool and Place Tool/Pressure Plate)					×	.	.		.	×
Bumps + TCR Layers (Pressure Plate/Bumps)							.	.	.	×
Substrate									.	×
Die Attach										×

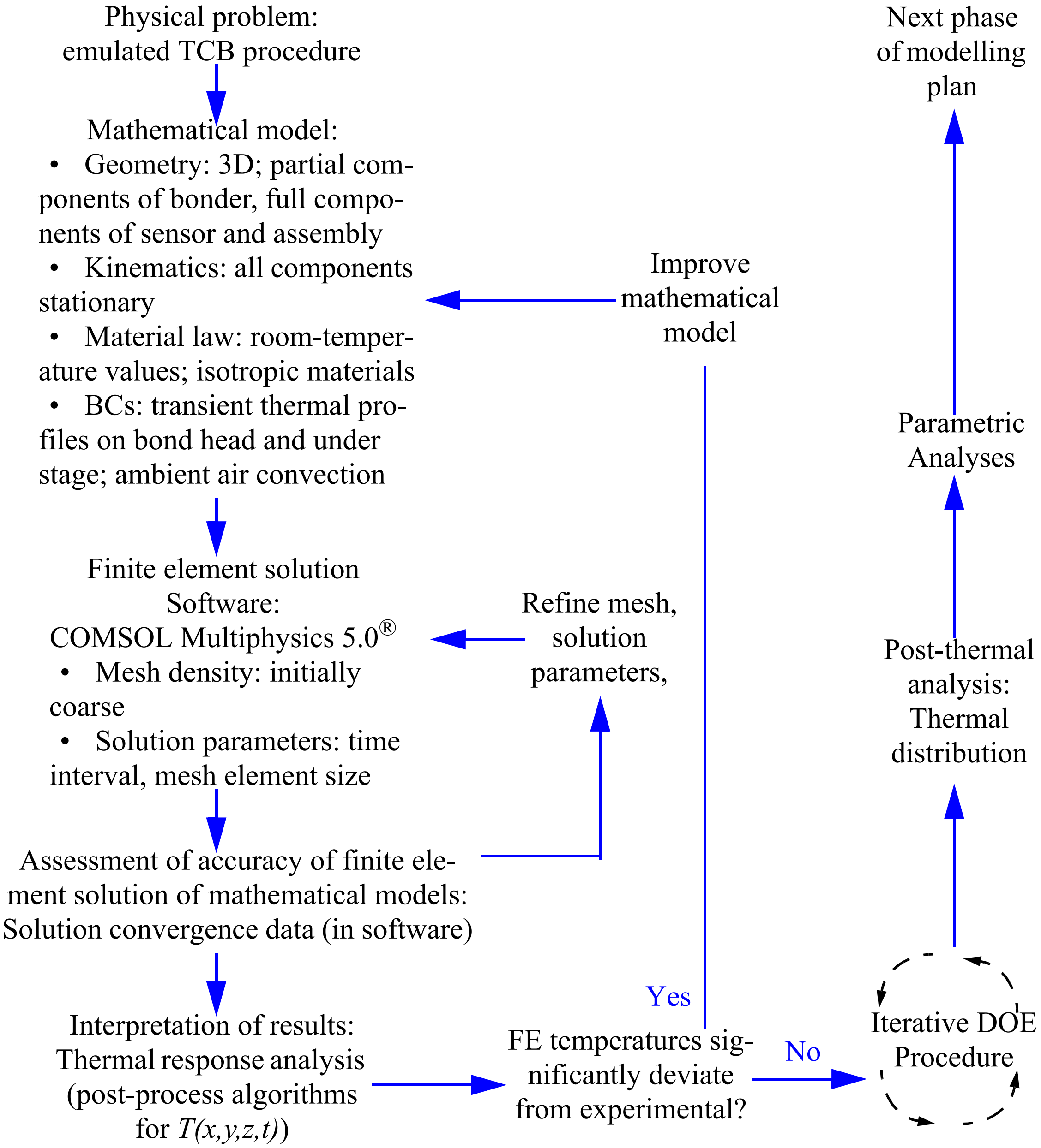


Fig. 3-1. FEA development modelling plan and thermal analysis procedure used for both simple thermal and complex model FE models [16].

factorial study is summarized in Table 3-1. Starting from the base model M0, components are added sequentially, isolating the effects of each variation on any of the four response features of interest (MHR, MFT, etc...). Important findings from this study is discussed

Table 3-2. Geometry of FE models [16].

Division	Component	Dimensions (X × Y× Z) (L × W × H)	Units
Bond Heat Assembly	Air Hole (Heater)	Radius = 0.6; Height = 5.48^a/8^b	mm
	Heater Block	25 × 23 × 5.48^a/8^b	mm
	Place Tool (Upper)	22 × 22 × 0.3	mm
	Place Tool (Lower)	3.5 × 3.5 × 0.5	mm
	Air Holes (Place Tool)	Radius = 0.2; Height = 0.8	mm
	Pressure Plate	4.68 × 3.31 × 0.5	mm
Intermediate Layer	Air Gap	4.68 × 3.31 × 0.04	mm
Sensor Chip/Substrate Assembly	Sensor Chip	4.68 × 3.31 × 0.5	mm
	Die Attach	4.88 × 3.41 × 0.03	mm
	CerDIP	1.4 × 0.595 × 0.05	in
	Macor® Sheet	2 × 2 × 1/16	in
	Stage	310 × 70 × 7.59	mm

[a]Heater block thickness used in Model A
[b]Heater block thickness used in Model B
(Discussed in Section 3.4)

throughout Section 3.2.

3.2 Model Initialization

The remainder of the mathematical model is discussed in this section. The Heat Transfer Module, part of the COMSOL Multiphysics software package, is used.

3.2.1 Geometries

All modeled geometries are three-dimensional to more accurately gauge the physical model. The geometries of the components were extracted from their datasheets, or through manual measurements via optical micrographs and scanning electron microscope (SEM) images. The geometries of the true boundaries of the sensor chip are extracted from Fig. 3-2a. The measured pitches between the bond pads and sensors accurately agree to those designed in the layout for the CMOS process. Thus, this layout is programmed as the geometry for the sensor chip and data extraction nodes. The dimensions of the epoxy are approximated and assumed to extend equally from each edge of the sensor chip in the *x*-*y* plane. The features of the digital logic region that occupy the arbitrarily assigned West end

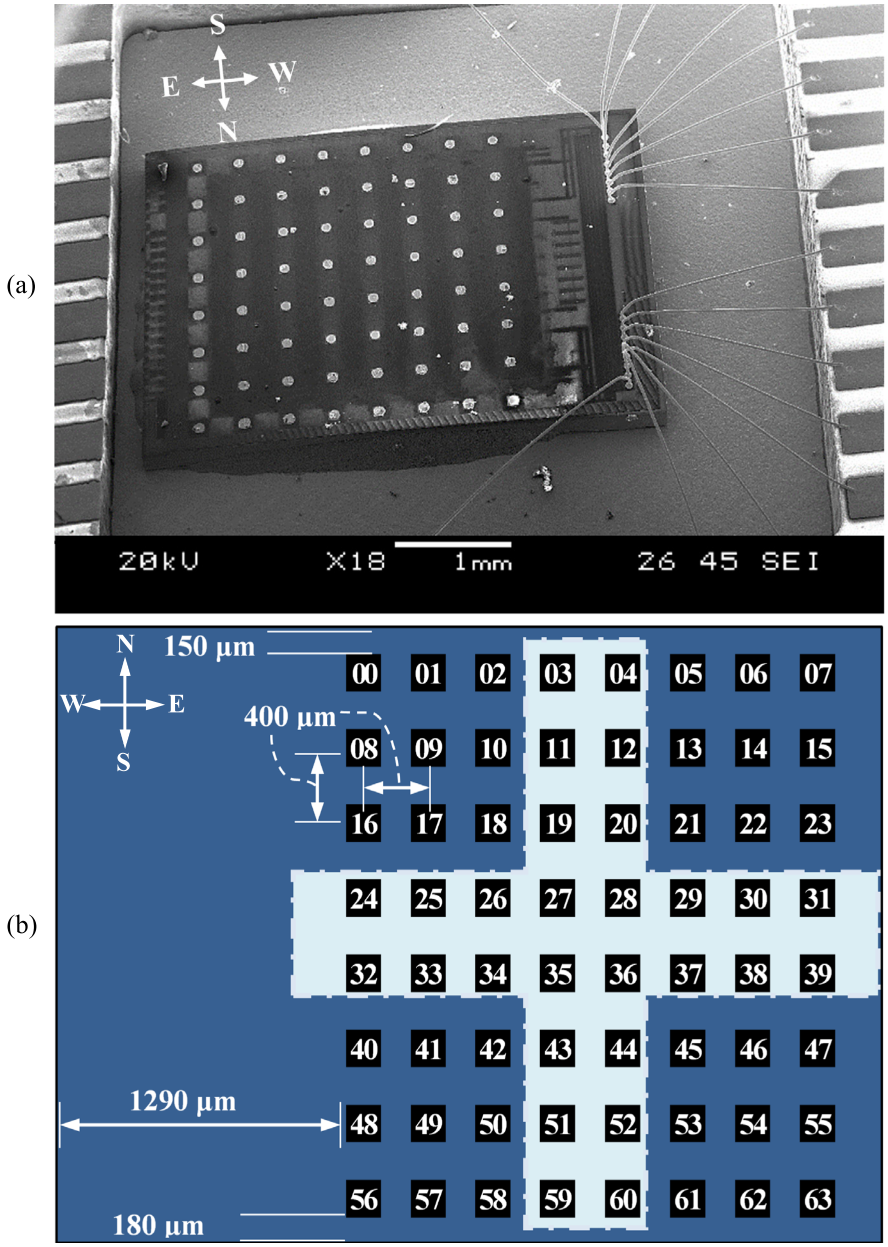


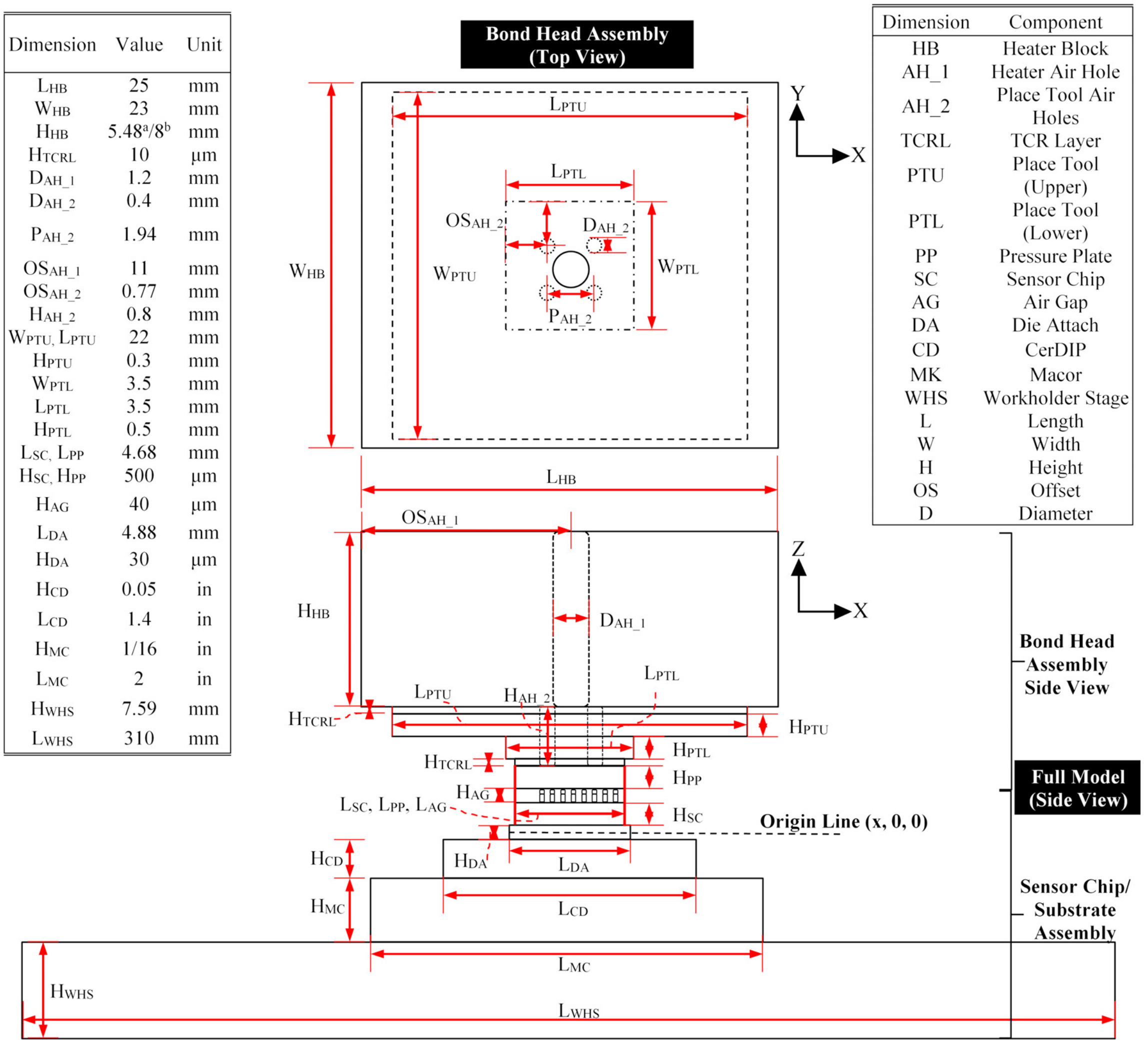
Fig. 3-2. (a) SEM image of the sensor chip die attached and wire bonded inside a CerDIP cavity; (b) layout and values of sensor chip components as implemented in the FE models [16].

of the chip are excluded from the model and are assumed to be part of the sensor chip volume. The wire bonds and bond pads are also excluded from the model, for simplification purposes. The bump geometries are approximated by taking the average diameters and heights from optical images, such as the micrograph in Fig. 1-11e. Due to the proprietary design of the bond head, only partial geometries are available. The remainder of the geometries are currently confidential.

These sources provide sufficient quantitative data to form the complete geometry illustrated in Fig. 3-3. An arbitrary division to reference components is made by splitting the entire geometric model into two main divisions: components above the sensor chip and components below. The sensor chip, Au bumps, chip substrate, die attach, electrical insulation layer, and a stage comprise the substrate assembly. The pressure plate, a place tool with four air holes, and a heater block with one air hole comprise the bond head assembly. The air holes are modelled to represent the vacuum holes in the true bond head. An intermediate air gap layer is modelled between the sensor chip, pressure plate, and between the bumps. The dimensions of each component are summarized in Table 3-2. The positions of component features are also provided in Fig. 3-3.

3.2.2 Material Parameters

Materials subjected to thermal gradients, such as those involved in TCB processes, experience changes in material properties. Thus, each of the three thermal material properties of interest; density, heat capacity and thermal conductivity, are temperature-dependant properties. However, modelling temperature-dependant effects increase computation times as they introduce additional complexity to the mathematical formulations. For simplicity, it is more convenient to use the room-temperature values during earlier stage FE development. Table 3-3 summarizes the materials used in each model and the inputted values for



Dimension	Value	Unit
L_{HB}	25	mm
W_{HB}	23	mm
H_{HB}	$5.48^{a}/8^{b}$	mm
H_{TCRL}	10	µm
D_{AH_1}	1.2	mm
D_{AH_2}	0.4	mm
P_{AH_2}	1.94	mm
OS_{AH_1}	11	mm
OS_{AH_2}	0.77	mm
H_{AH_2}	0.8	mm
W_{PTU}, L_{PTU}	22	mm
H_{PTU}	0.3	mm
W_{PTL}	3.5	mm
L_{PTL}	3.5	mm
H_{PTL}	0.5	mm
L_{SC}, L_{PP}	4.68	mm
H_{SC}, H_{PP}	500	µm
H_{AG}	40	µm
L_{DA}	4.88	mm
H_{DA}	30	µm
H_{CD}	0.05	in
L_{CD}	1.4	in
H_{MC}	1/16	in
L_{MC}	2	in
H_{WHS}	7.59	mm
L_{WHS}	310	mm

Dimension	Component
HB	Heater Block
AH_1	Heater Air Hole
AH_2	Place Tool Air Holes
TCRL	TCR Layer
PTU	Place Tool (Upper)
PTL	Place Tool (Lower)
PP	Pressure Plate
SC	Sensor Chip
AG	Air Gap
DA	Die Attach
CD	CerDIP
MK	Macor
WHS	Workholder Stage
L	Length
W	Width
H	Height
OS	Offset
D	Diameter

Fig. 3-3. Partial views of the complete geometric layout used in the FE models [16].

each property, excluding the TCR layers.

Table 3-3. Thermal material property values (@ 20 °C) used in the FE models.

Components	Material	ρ	k	Cp	Reference
Air Hole (All) Air Gap	Air	1.204	0.026	1005.42	×
Heater Block Place Tool (Upper) Place Tool (Lower)	AlN	3300	180	719.47	×
Pressure Plate Sensor Chip	Si	2330	152.9	702.2	[33]
Bumps	Au	19290	319	129	[33]
Die Attach	Ag-filled Epoxy	35	2.5	240	[68]
CerDIP	Al_2O_3	3900	27	900	×
Macor® Sheet	Macor®	2520	1.6	600	[33]
Stage	Invar	8080	13.86	477.7	×

× - Comsol built-in values

3.2.3 Boundary Conditions

There are two main boundary conditions (BCs) applied to the model. A transient thermal profile, nearly identical to that in Fig. 2-2, is applied to the top surface of the heater block. A constant, steady temperature of 55 °C is applied to the bottom surface of the stage. 5 °C is added to both BCs in order to better match the experimentally measured steady state temperatures of ≈55 °C. The initial temperatures of all components are set to 55 °C on the assumption that they have reached steady state temperatures prior to ramping. The heater thermal profile is implemented by formulating a ramp function in COMSOL, which holds a constant temperature of 55 °C for 6.45 s, until ramping to 200 °C at a rate of 350 °C/s. This is then held until 9.59 s, where it is then cooled. Ambient air convection was applied to all other surfaces of the model components, with heat transfer coefficients *h*, of 5 W/m^{2}K

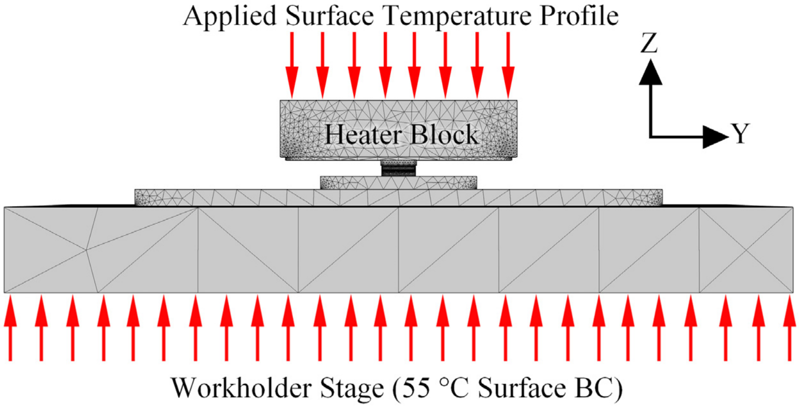


Fig. 3-4. Boundary condition locations [16].

and 10 W/m^{2}K, and external temperatures T_∞, set to room temperature, 20 °C. This boundary condition is later removed during modelling due to the negligible effects its inclusion has on the simulated thermal responses. The locations of the two boundary conditions are indicated in Fig. 3-4.

3.2.4 Meshing

A total of ≈ 2.2 million tetrahedron-shaped mesh elements are applied to the models through the available meshing feature in COMSOL. The majority of the model has rectangular-based components, tetrahedron-shaped elements are the default option in COMSOL. Therefore, this is the most time-effective and convenient option to apply. Overviews of the model are provided in Fig. 3-5.

3.2.5 Data Extraction from Model

The 64 RTD sensors are modelled to capture the average temperature experienced within the confines of the area covered by each serpentine winding. Replications of these RTD sensors are made by applying data nodes on the sensor chip surface, through the boundary probe option in COMSOL. These surface data nodes are implemented around each bump, covering the area as replicated from CMOS designs and confirmed from micrographs. These nodes do not have an associate thickness, and use surface mesh element nodes to calculate average temperatures. For each time interval solved during computation,

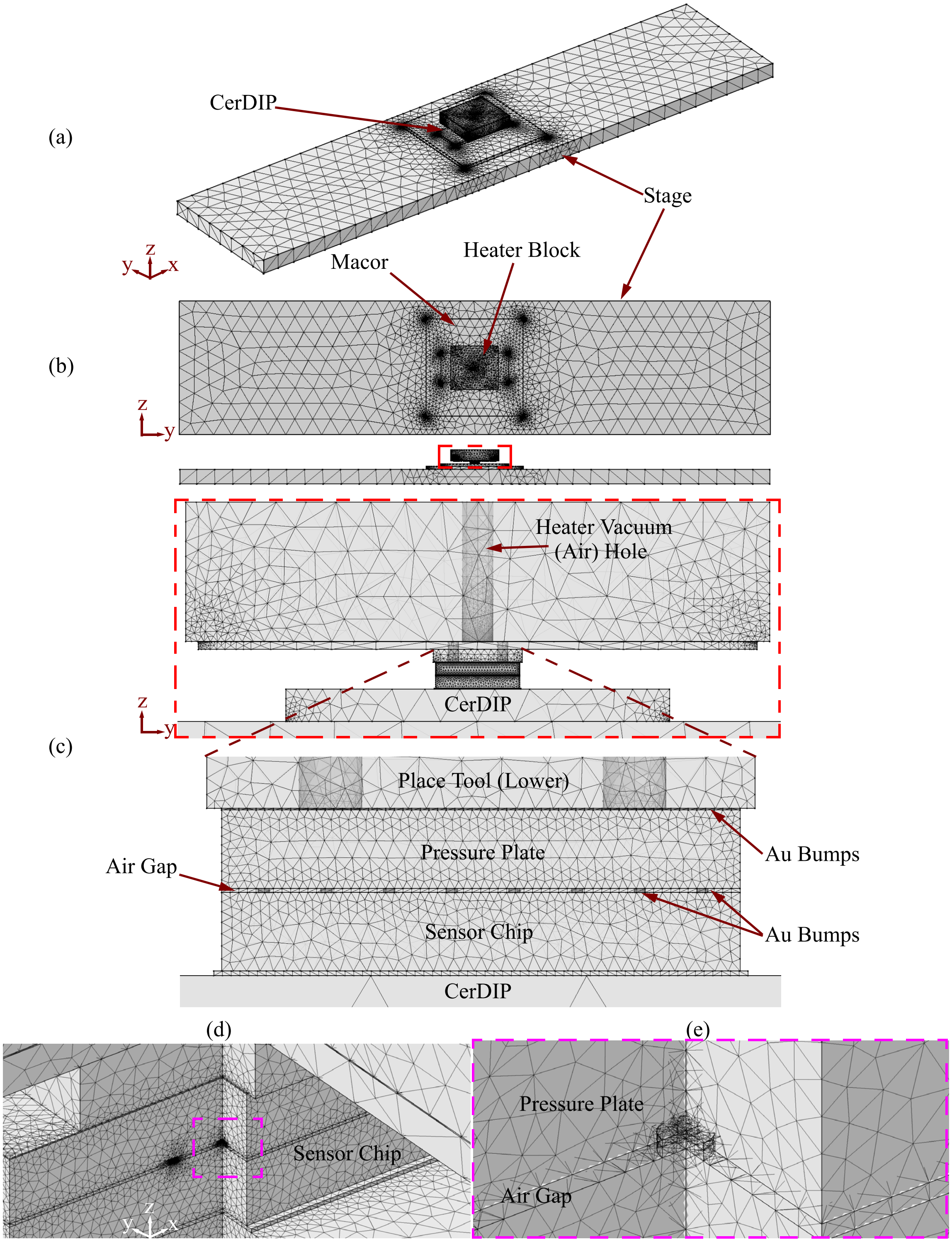


Fig. 3-5. (a) Orthographic view of the complete meshed model; (b) top view; (c) translucent partial side view; (d) 1/4 section cut-out at pressure plate - sensor chip interface; (e) magnified view of (d).

an average of the temperature within each node is calculated and stored for post-processing. In our models, the time resolution is 15 ms. Ultimately, these nodes are used to extract the numerical data from which the simulated temperatures are derived from. Fig. 3-6 shows one of the 64 surface data nodes used in the models.

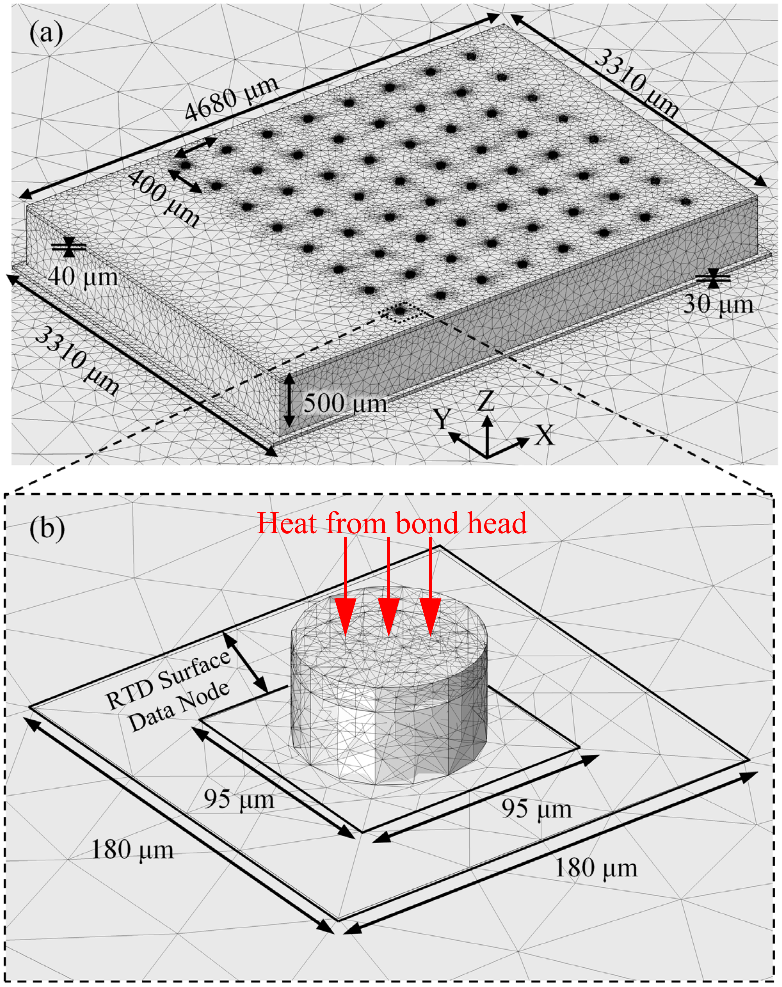


Fig. 3-6. (a) Magnified view of sensor chip geometry; (b) Magnified view of bump, TCR layer and RTD surface data node. The air gap is shown in (a), but not shown in (b). Both images are to scale [16].

3.2.6 Thermal Contact Resistance Layers

Unlike the TCR formulations from theory in Section 1.3, the simple thermal model uses volumetric boundary layers between two components in mechanical contact, to simulate thermal resistance. These layers incorporate the effects of air content between components by applying hybridized material properties of air and the material of the component below the contact boundary. While these boundary layers neglect the effects of contact pressure in the thermal model, as well as add more elements to the mesh density, they simplify mathematical formulations in the earlier stages of FEA. They do this by reducing the number of input parameters involved in multi-dimensional contact models.

A total of three TCR layers are added: one between the heater block and place tool (P1); a second between the place tool and pressure plate (P2); and the third between the pressure plate and bumps. Each boundary or TCR layer is 10 μm thick and assumes the length and width of the component contact pair with the lowest surface area. For example, the P1 layer assumes a width and length equal to the width and length of the upper part of the place tool. There are total of 64 P3 layers: one for each bump and pressure plate interface. An example of a P3 TCR layer is seen in Figs. 3-6b and 3-7.TCR layers between the CerDIP and the Macor sheet, and between the Macor sheet and stage, are excluded from models as they have shown to have negligible effects on simulated sensor signals. This is due to the proximity and relative sizes these components have with respect to the sensor chip. They are farther away from the bond head and are fairly insulative, thus having little effect on the temperatures in the sensor chip. Furthermore, the thickness ratio of the TCR layer to the substrate assembly components is significantly low.

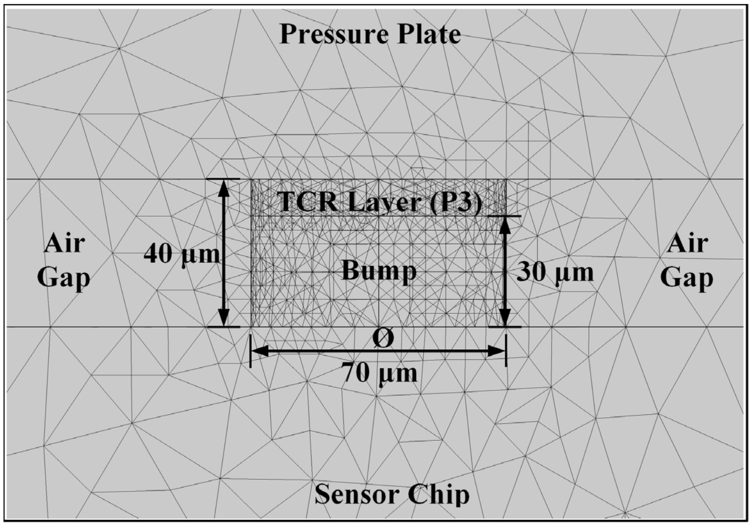


Fig. 3-7. Cross-section view of the “bonding interface” between pressure plate and sensor chip bump [16].

To implement hybridized material properties, weighted averages of the thermal properties are applied for each TCR layer. These averages are implemented as:

$$\rho = P \cdot \rho(air) + (100 - P) \cdot \rho(material) \tag{3-1}$$

$$k = P \cdot k(air) + (100 - P) \cdot k(material) \tag{3-2}$$

$$Cp = P \cdot Cp(air) + (100 - P) \cdot Cp(material) \tag{3-3}$$

where the three contact materials of interest are AlN (P1), Si (P2) and Au (P3), and *P* is the percentage or fraction of air content in the TCR layer. Trial studies independently varying the percentages of air in each layer conclude that each layer more or less has the same effects on simulated thermal responses. Therefore, to minimize the number of model parameters further, the percentages of air content for P1, P2, and P3 are all held constant such that a change in *P* affects all three TCR layers and their corresponding material properties, respectively. Therefore, this consolidated TCR model describes the average thermal contact quality.

3.3 Model Fitting Process

Based on the findings from preliminary modelling trials, such as those conducted in Table. 3-1, three of four experimental responses have been found be effectively matched varied using model parameters. Despite testing several different fitting parameters via multiple regression analysis, only two partially orthogonal input parameters have been found to be directly proportional to the MHR or FTRS, and MFT regions. Since the KP is dependant on all three of three regions, it is determined that it is not a suitable response to be matched via parameter variation. Since there are only two input parameters that correlate to the three responses, only two regions can be intentionally matched at once. Hence, two models named “Model A” and “Model B” have been developed: Model A best matches the MHR and MFT regions/values and Model B best matches the FTRS and MFT regions/values. The heater block thickness and the air content percentage in the TCR layers are the two orthogonal input parameters. The heater block thickness introduces a thermal delay in the sensor signals as the BC heat source is on top of the heater block surface. Thus, the rate of heat transfer from the heater block to the sensor chip surface is directly proportional to the thickness of the heater block. As mentioned earlier, only partial geometries and dimensions were provided for the bond head heater. Therefore, the heater block thickness makes sense as the most appropriate parameter to be varied to fit the FE models to experimental responses. As the thermal conductivity of the heater block is not dependant on its geometry, and is physically defined as the rate that heat energy can pass through a material (Eqn. 1-9), the MFT does not get affected. On the other hand, air content in TCR layer introduce thermal resistance due to air's very low thermal conductivity and high heat capacity. As the thickness of these TCR layers are relatively small compared to the other components, the MHR is negligibly affected for percentages less than ~97%. However, the high heat capacity, which is inversely proportional to the thermal gradient (Eqn. 1-13), strongly affects the MFT. As heat passes from the bond head down through the components, the TCR layers absorb heat and subsequently pass less heat energy down to the bumps. This resistance results in temperature drops during heating and thus, a reduced MFT within the given process time frame.

With two input parameters and two responses that can be matched at once, an iterative 2 × 2 DOE procedure can be conducted, as illustrated in Fig. 3-8. In this procedure, two trial heater thickness values and two trial air content percentage values are inputted into the simulation model, one combination at a time. The simulated sensor signals from each trial are stored as ".csv" files and then loaded processed by the post-processing algorithms. These algorithms produce the MHR, MFT, FTRS, and the remainder of the relevant thermal data. After the four initial combinations are simulated, target response curves are plotted on contour plot interpolations of the data in order to observe whether an intersection of the two targets occur. If an intersection does occur, the intersection point values, which are the heater thickness and air content percentage, are inputted into the model. If the simulated response values from this intersection point trial are within 2% of the experimental target values, then a suitable match is found. If not, the DOE is refined such that 2 more combinations are conducted, forming a 2 × 3 partial DOE or 5 more combinations to form a 3 × 3 trial DOE. This iterative process continues until the simulated responses are within acceptable limits. If the intersection points do not converge to values within acceptable limits after the DOE refining process, a new 2 × 2 DOE is initialized using a broader range of input values. Alternatively, a justifiable change to the model formulation itself is made

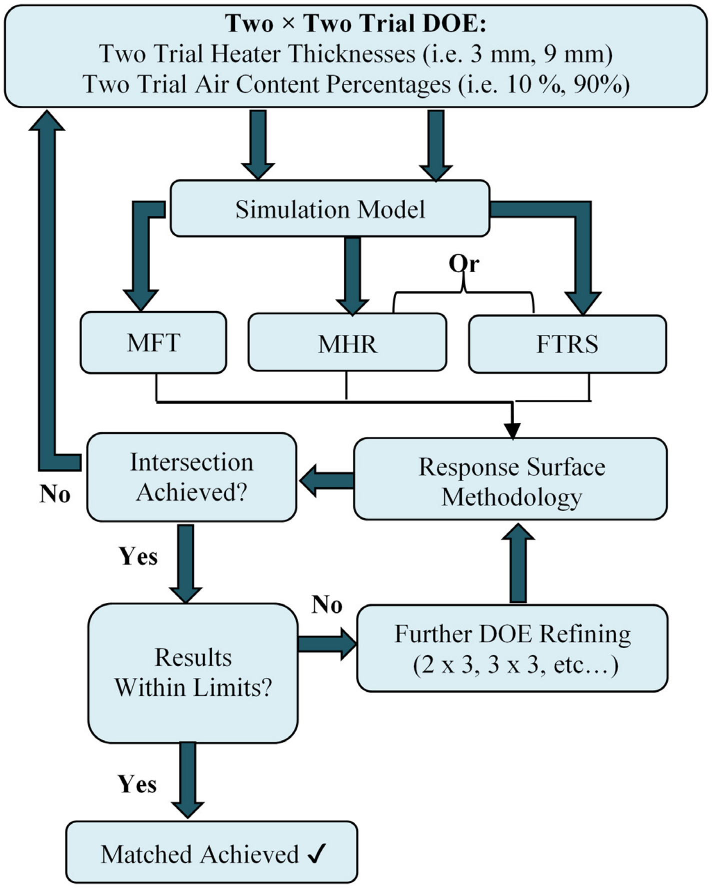


Fig. 3-8. Flowchart of the iterative 2 × 2 DOE procedure conducted to match FE model responses to experimental responses [16].

through either problem redefinition, geometric changes, boundary condition changes or material property value changes. This process is repeated for both Model A and Model B with the target values of each corresponding to their target responses.

3.4 Matched Models

3.4.1 Model A

The goal of Model A is to produce simulated sensor responses that best match the MHR and MFT values. Through the iterative DOE procedure, the intersected heater thickness and air content percentage values are 5.48 mm and 91.7%, respectively. The intersection point on the response contour plot can be seen in Fig. 3-9. The values from the simulated MHR and MFT regions differ from the experimental values by 0.09% and

0.04%, respectively, indicating a very good numerical match is found for these two values. The FTRS, however, deviates 41.6% below the experimental value. The KP temperature is 5.3% below the experimental mean and within 1% of the mean experimental KP time. Nonetheless, the signal overlap in Fig. 3-11 confirms a good optical and numerical match is achieved. The gap identified is discussed in Section 3.5.

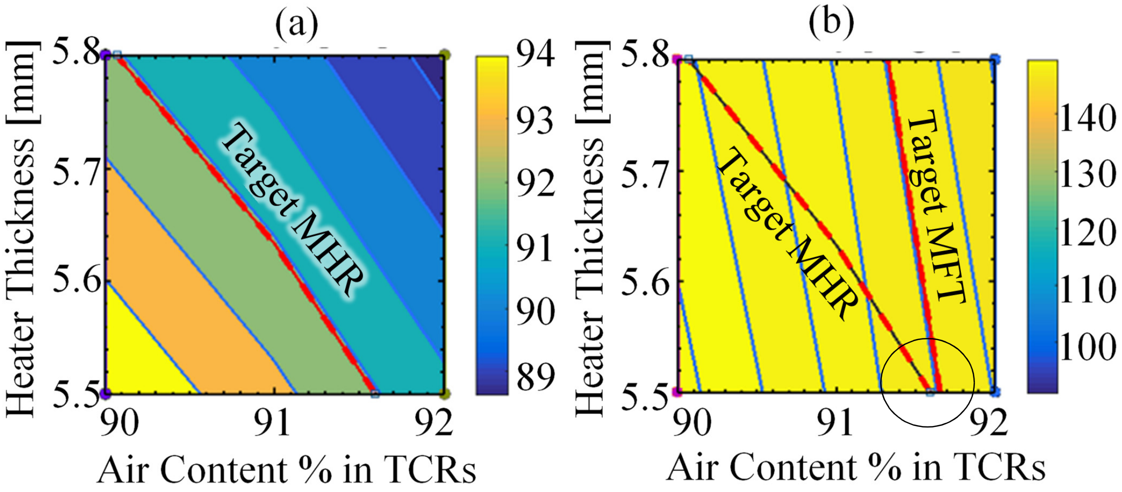


Fig. 3-9. (a) Contour of the MHR values of the four trial parameters. The target MHR value can be reached by selecting any heater thickness and value on the red dashed line; (b) The response contour for the MFT values of the same four trial parameters. The intersection point is slightly below the *x*-axis in this image [16].

3.4.2 Model B

Model B's target response regions and values are matched with intersection values of 8.0 mm and 91.5% for the heater thickness and air content percentage, respectively. This presents a negligible difference in air content percentage compared to Model A, but a significant difference of 45.9% in heater thickness. This is because the heater block thickness significantly affects the MHR and FTRS, as does the air content percentage, but to a lesser extent. This is why the air content percentage remains relatively constant between the two models, while the heater thickness is varied. While Model B's MFT and FTRS values differ from the experimental values by 0.04% and 1.3% respectively, the signal overlaps indicate Model B to be a weaker match overall. As seen in Fig. 3-10, Model B appears to be an opti-

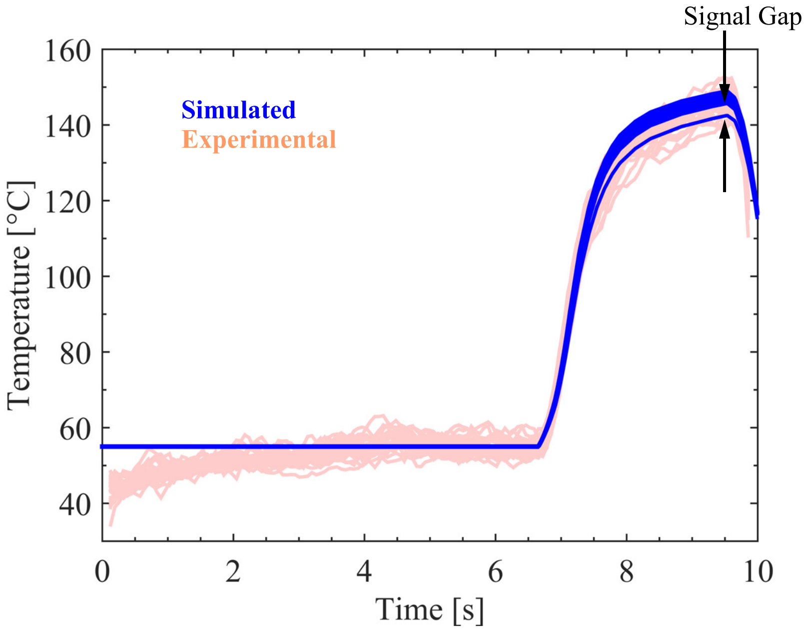


Fig. 3-11. Signal overlap of Model A's simulated response [16].

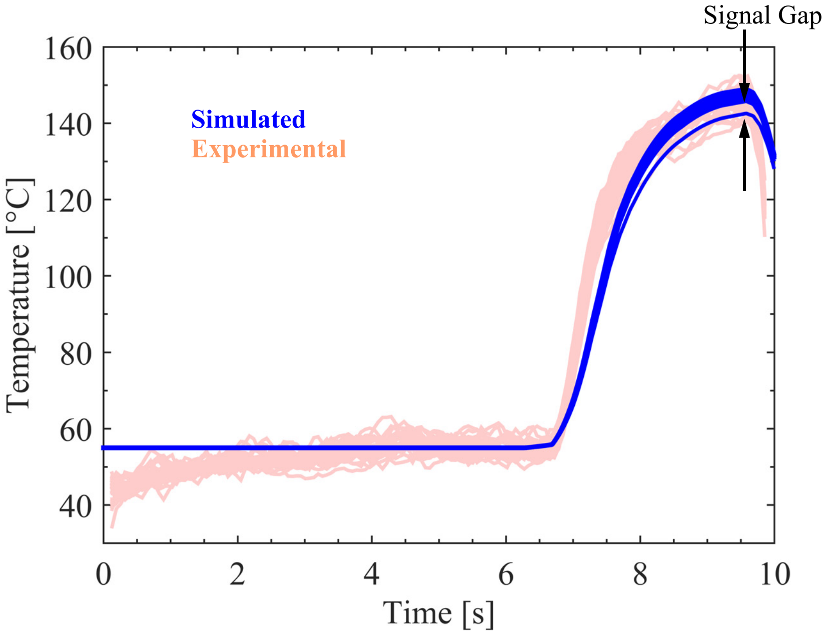


Fig. 3-10. Signal overlap of Model B's simulated response [16].

cally weaker fit than Model A in terms of signal overlap differences. This is primarily due to the 22.5% lower average MHR of Model B compared to the experimental mean, as evident in its shallower slope.

3.5 Lateral Surface Thermal Distribution

With matched FE models fitted based on reasonable assumptions about the experimental conditions, further analysis can be conducted to better understand heat flow during TCB. The first interesting feature of the simulated model results is the signal gap identified in Figs. 3-11 and 3-10. This gap insinuates that some of the RTD sensor locations experience a significant deviation from the mean of the RTD temperatures. Using post-processing analysis in the FE software itself, Fig. 3-12 is produced, illustrating the source of this gap. Backed by numerical and graphical evidence, the signal spread and gap is physically caused by the area on the left of the bump array, below the air gap. The relatively large volume of silicon mass acts as a heat sink, drawing heat away from the left-most region of the bump array. Consequently, the signal gap arises from the left-most bumps experiencing the most cooling due to the heat sink effect. The remaining peripheries of the bump array also experience this heat sink effect, but to a more moderate extent. By the end of the heating period, Δt_{press}, the sensor chip experienced localized maximum temperatures up to ≈155 °C at the center of the bump array, and temperatures as low as ≈129 °C to the left of the bump array.

The range of temperatures amongst the RTDs, as well as the maximum thermal gradients experienced over time, is shown in Fig. 3-13. The maximum thermal gradient experienced occurs at almost the same time measured from experimental responses, even though its magnitude is a fraction of the experimental maximum magnitude. The simulated maximum thermal gradient of Model A is 7.42 °C/mm which is about half of the second highest experimental thermal gradient magnitude of ≈15 °C/mm. The locations of the maximum gradients experienced occur at the same bottom left and top left corners identified from in the experimental data. Similarly, Model B's maximum thermal gradient magnitude experienced is ≈7.03 °C/mm. However, due to the slower primary heating ramp of the RTD sensors, this magnitude is reached at ≈8.34 s. This magnitude is reached within 0.3 s of the knee point. The timing coincidence is particularly interesting as the findings suggest the

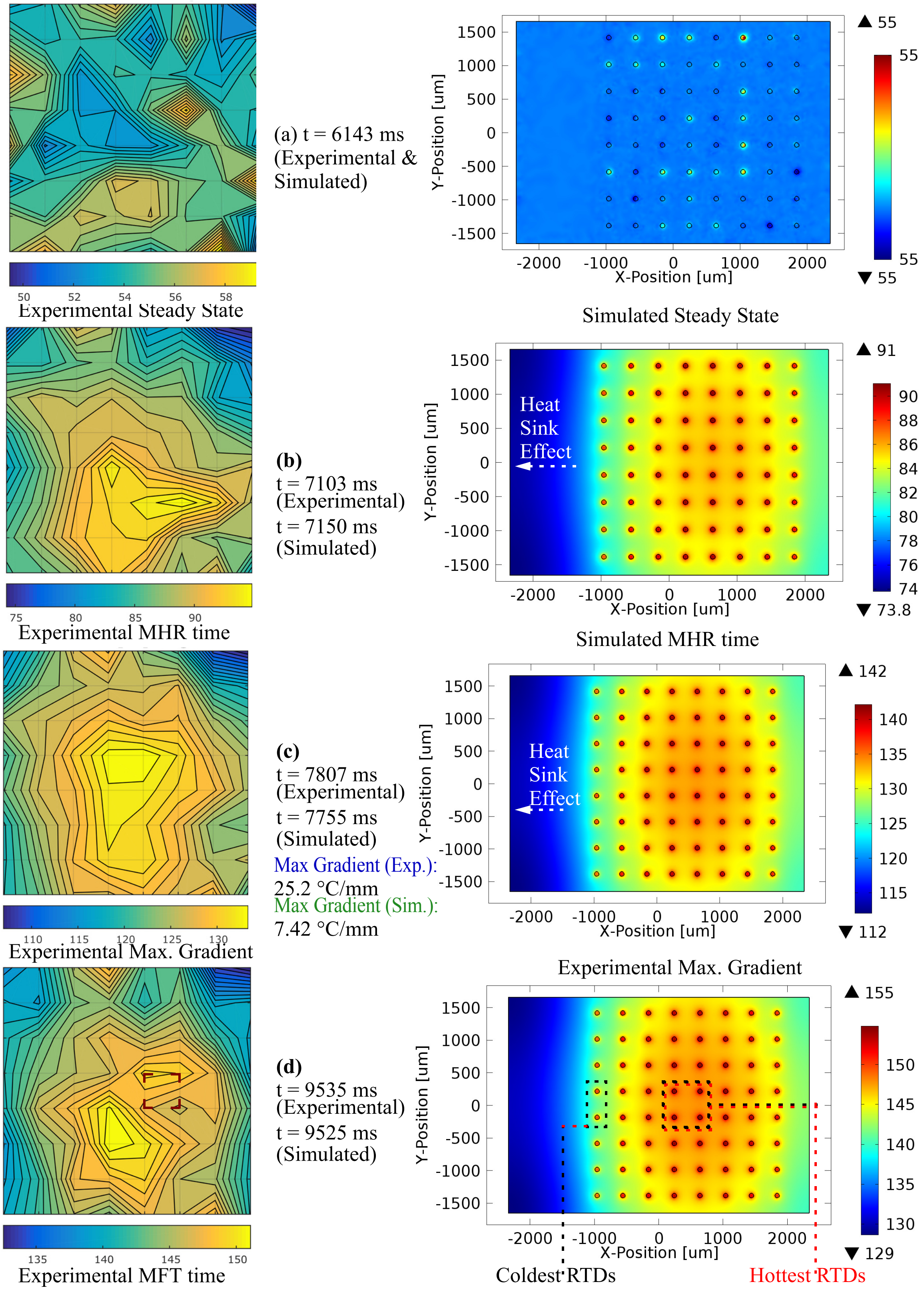


Fig. 3-12. Comparison of experimental and simulated thermal distributions across the sensor chip surface. (a) Steady state; (b) average MHR time; (c) time of maximum thermal gradient; (d) average MFT time. The interpolated experimental and FE-based contour plots exhibit similar results and verify the heat sink effect.

greatest thermal non-uniformity occurs approximately after the KP region of the heating curve.

The large thermal gradients can pose as problematic as it can lead to thermo-mechanically induced strains, possibly affecting the quality and/or reliability of bonds made. However, any conclusions about the effects of these gradients, with respect to these assumptions, is premature and would require thermo-mechanical analysis.

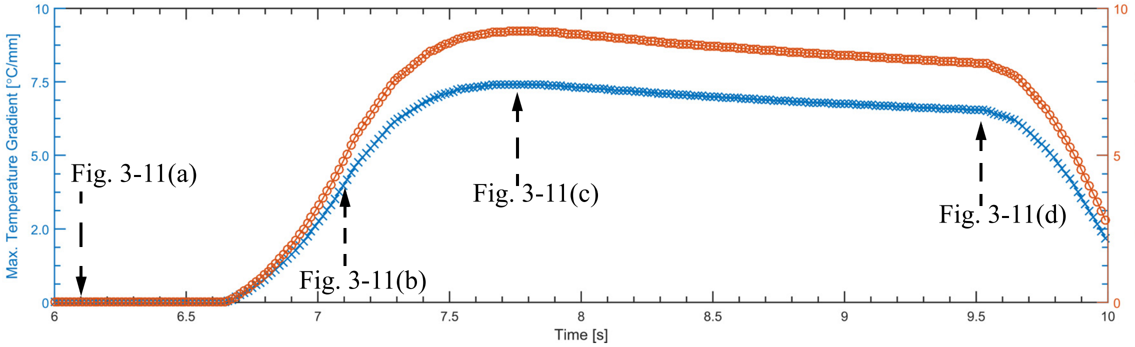


Fig. 3-13. Transient curves of the thermal gradient magnitudes and temperature ranges between 6.0 s and 10.0 s for Model A.

3.6 Joint to Joint Thermal Variation

The variation between the temperatures of each bump is calculated by first extracting the volumetric averages of the bumps. With the use of FEA software, this is an easy feat to accomplish. The variation can be described as the range or temperature difference between the highest bump temperature and the lowest bump temperature. These averages are taken at 9.53 s, slightly before the bond head cools, to account the effects of data smoothing algorithms on the number of useful data points, caused by smoothing algorithms. The comparisons of thermal responses for both models is summarized in Table 3-4. It is further clear that Model A and B exhibit nearly identical responses.

The corner bumps at pads 00, 07, 56 and 63 experience the lowest bump temperatures of ≈151 °C. The left two corner bumps (00 and 56) are ≈1.5 °C lower in temperature than the right two bumps (07 and 63). However, the left two pads are ≈3.3 °C lower in temperature than the right two. These differences can be explained by the heat sink effect. With

Table 3-4. Summary of Model A and B's bump and RTD temperature variation [16].

		Mean	**St. Dev**	**Min.**	**Max.**	**Range**
Model A °C	Bump	154.63	1.55	150.63	156.3	5.67
	RTD	147.33	1.68	142.47	148.98	6.51
	ΔT	7.3	0.13	8.16	7.32	0.84
Model B °C	Bump	154.60	1.56	150.56	156.2	5.66
	RTD	147.41	1.68	142.55	149.07	6.52
	ΔT	7.19	0.12	8.01	7.13	0.86

no bumps in the West region of the sensor chip, there are a lack of thermally conductive pathways for the heat to travel to the sensor chip, resulting in cooler surface temperatures. Due to the lack of conductive pathways, the bottom surface of the pressure plate remains hotter, explaining the greater temperature gradient across the air gap on the left of the sensor chip. The cooler Si mass in the West region draws more heat down through the left-most bumps to compensate for this effect, but the surface remains cooler. This is why the difference between the bumps are lower than the difference between the RTDs.

The central bumps at pads 27, 28, 35, and 36 experience the greatest temperatures, up to ≈156 °C. These bumps have the greatest number of neighboring interconnects and the lowest area of periphery Si around them. Thus, these bumps experience the greatest thermal uniformity, posing the least concern in TCB processes.

Despite wide thermal ranges across the sensor chip, both models indicate robustness in thermal application during TCB. The temperature range amongst the bumps at 9.53 s is ≈5.7 °C. This low range indicates that thermal distribution for this specific setup can be acceptably uniform. Temperature differences of ≈7.3 °C between bumps and RTD data nodes indicate that microsensor arrays can be accurate thermal measurement tools in TCB processes.

3.7 Internal Joint Thermal Variation

Due to the strong similarities of Model A and Model B, the following results presented are mainly extracted from Model A, to reduce redundancy. Findings from the

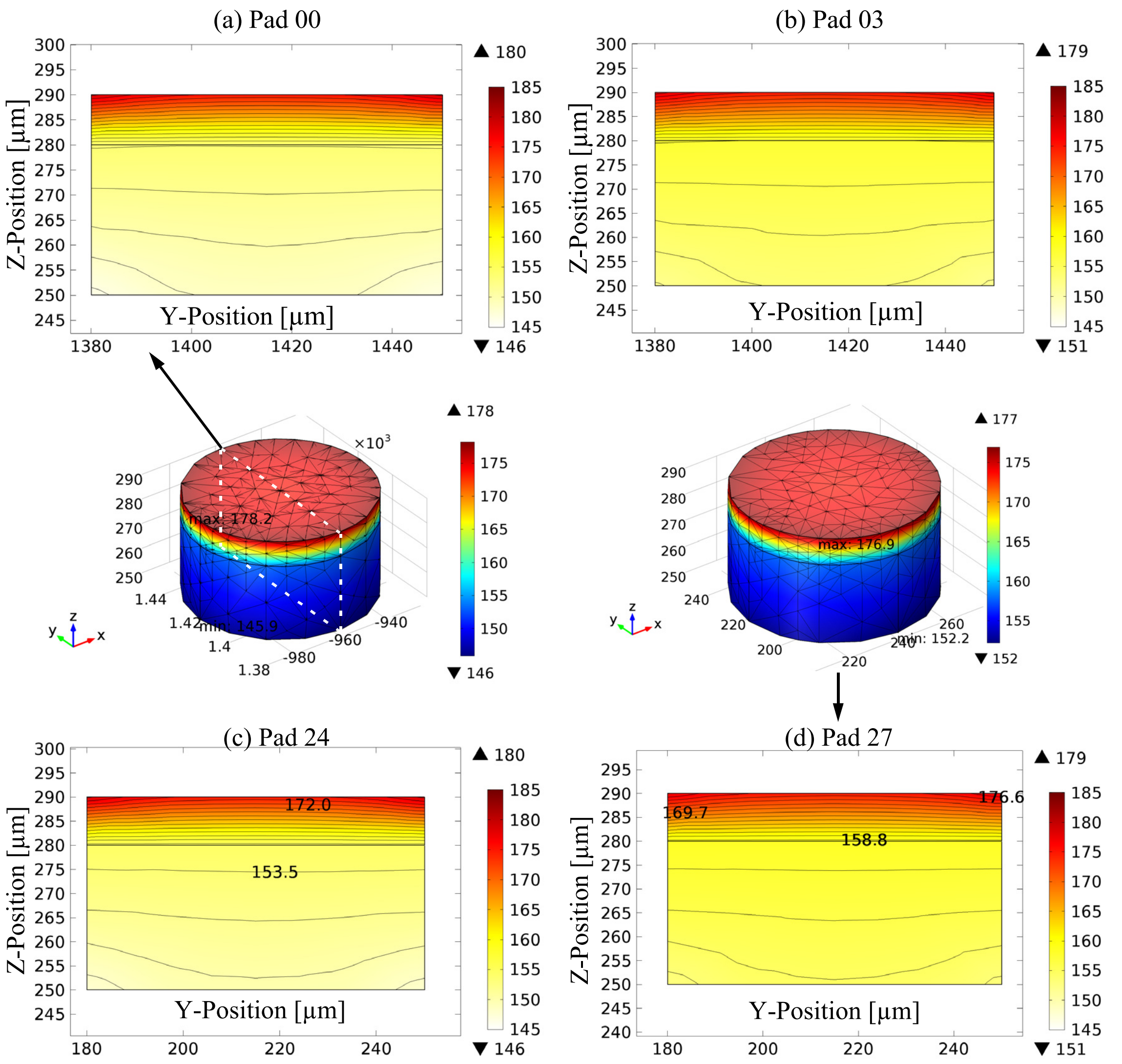


Fig. 3-14. Z-X Contour plots of bumps from (a) pad 00 (b) pad 03 (c) pad 24 and (d) pad 27. The thermal distribution within the bumps at pads 00 and 27 are shown. Of these four, pad 27 has the bump with the highest temperature, and pad

models indicate that bumps not only experience thermal variation depending on location, but they also experience temperature drops within the confines of their geometry, as seen in Fig. 3-14. As expected, the bumps are hotter on their top surface, than on their bottom. The average temperature of the top surfaces of all bumps is ≈156.6 °C and ≈151.8 °C on the bottom. This results in an average temperature drop of ≈4.8 °C within the bump volumes. As each bump has a P3 TCR layer above it, a temperature drop also occurs from the top of each TCR layer on the bottom surface of the pressure plate, to the top of each bump surface. The average temperature on the top of each P3 TCR layer is ≈175.7 °C, which is an average temperature drop of ≈19.1 °C within the 10 µm thick layer. This is due to the high air content of ≈91.7% (91.5% in Model B) of each layer. This percentage equates to a thermal conductivity of ≈26.5 W/m/K and heat capacity of ≈932.7 J/kg/K in each P3 layer. This explains the high thermal resistance exhibited by the layer. As air content percentages are reduced, thermal contact conductance improves and higher temperatures are experienced by the bumps. This is discussed in further detail in Section 3.9.

These internal variations could once again present problems in the thermo-mechanical domain. Thermal gradients within the bumps could possibly manifest as voids, defects or cracks if the magnitudes of induced stresses are significant enough. This could present as a potential source of reliability issues and could be studied to understand whether it has any implications during bonding.

3.8 Sensitivity Study I: Effects of Heater Block Thickness

A sensitivity analysis of the effects of the heater block thickness on simulated temperatures, is conducted. As seen with the differences between Model A and Model B, the heater block thickness affects the rate at the sensor chip heats up and consequently, the rate at which RTDs do. Using Model A's heater thickness as the base value, the thickness is varied by ±25% and the results of this study is plotted in Fig. 3-15

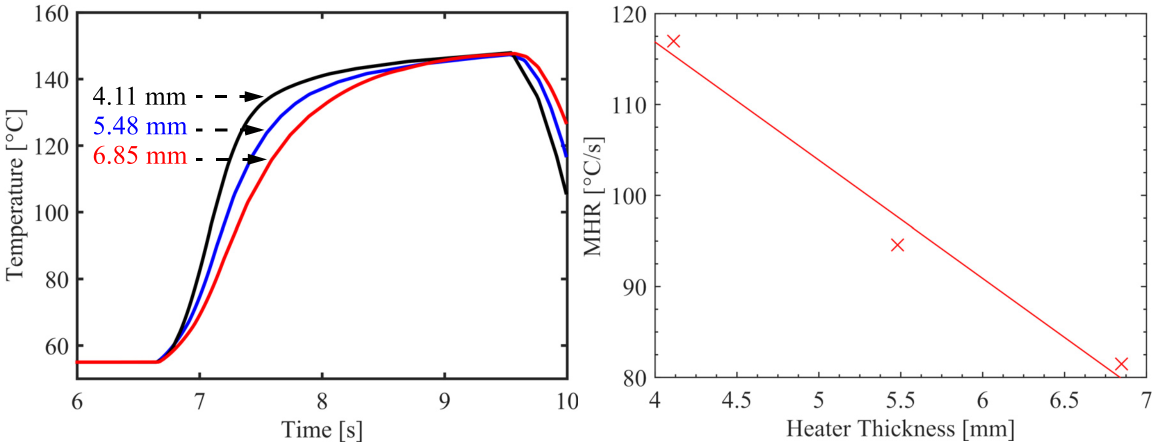


Fig. 3-15. Heater block thickness variation and its effect on mean MHRs. The full signals for the three trials are averaged and the mean RTD temperature is plotted for each heater block thickness. Air content percentage held at 91.7%. All other parameters held constant. Only the last four seconds shown for emphasis. Heater block thickness is linearly related to the average MHR. [16]

As the heater block thickness is increased, there is a greater mass through which the heat from the bond head BC must propagate through, resulting in a delayed transference of heat to the sensor chip.

3.9 Sensitivity Study II: Effects of Air Content in TCR Layers

A similar sensitivity analysis is conducted for the percentage of air content in the TCR layers. Using Model A as a reference once again, the air content is varied by adding or removing 5% to the matched air content value of 91.7%. The responses from this analysis is shown in Fig. 3-16. It is visibly apparent that this input parameter most greatly affects the MFT, while maintaining fairly constant MHRs and FTRSs. At air content percentages higher than ~97%, the trends significantly deviate from these presented, because the TCR layer material properties essentially converge to those of air.

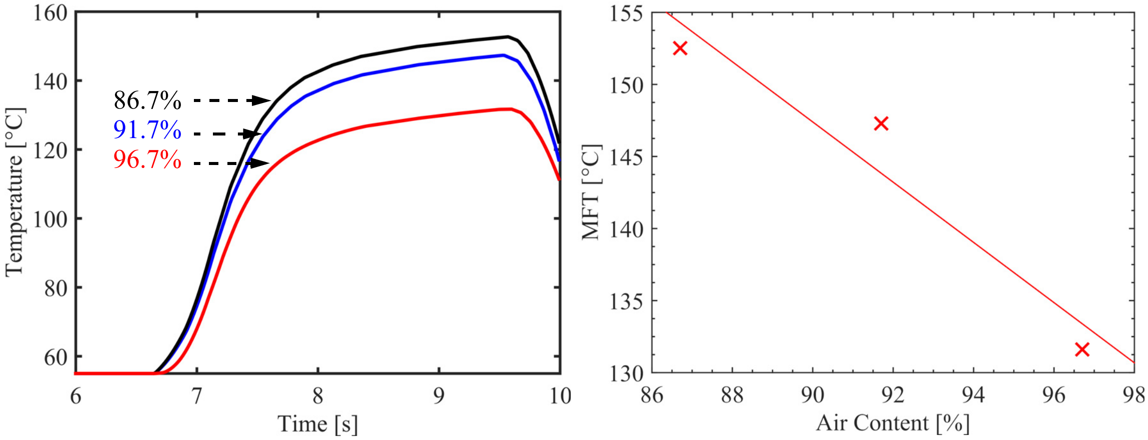


Fig. 3-16. Air content percentage variation and its effect on mean MFTs. [16]

3.10 Complex Thermal Model: CMY-Implementation

The simple complex model uses hybridized boundary blocks inserted between two contact components. To convert the simple complex model to the complex thermal model, these blocks are first removed and geometric adjustments are made. By removing these layers, air content percentages are no longer applicable and a new system of fitting the model to experimental data is required. To shift dependence to theoretical contact parameters, the heater block thickness is also disregarded as a fitting variation parameter.

3.10.1 Modelling Methodology Adjustments

To systematically approach this problem, the thermal conductance between layers is considered. Equation 1-8, which provides the conductance correlation, asserts that heat flux across the boundaries of two solids is proportional to the average RMS slope and contact pressure at the interface, and inversely proportional to the RMS height and microhardness. Given that the microhardness is a material property and contact pressure can be assumed to be the pressure applied during the emulated TCB procedure, only two parameters remain. Hypothetically, these parameters should provide similar effects on the simulated thermal responses as the heater block and air content did for the simple thermal models. This assertion is backed by the illustration in Fig. 3-17. The CMY correlation assumes that the surface asperities, of the material with the lower microhardness, plastically deform. Since these plastic deformations are not stored in the governing equations, the same thermal conductance is experienced, despite whether cyclic mechanical loading is applied or not. As the RMS slope of the asperities is linearly proportional to the constriction conductance, an increase of the slope leads to an increased conductance. Conversely, an increased height leads to a reduced thermal conductance. Asides from the asperity-asperity conductances, the slopes and heights are factors in the amount and volumes of air gaps between the two contact components. These air gaps introduce thermal gradients, just as the hybridized TCR layers do. Therefore, it is expected that these two parameters have similar effects on both MFTs and MHRs regions, with limited orthogonality between the two.

3.10.2 TCR Layer Application

To implement this thermal contact formulation in COMSOL, the “Thermal Contact” option under the Heat Transfer in Solids module is created for each type of layer. The boundaries of contact are defined by selecting the surfaces of interest, identified in Fig 3-18. The remaining contact parameters, listed in Table. 3-5, are also entered under the

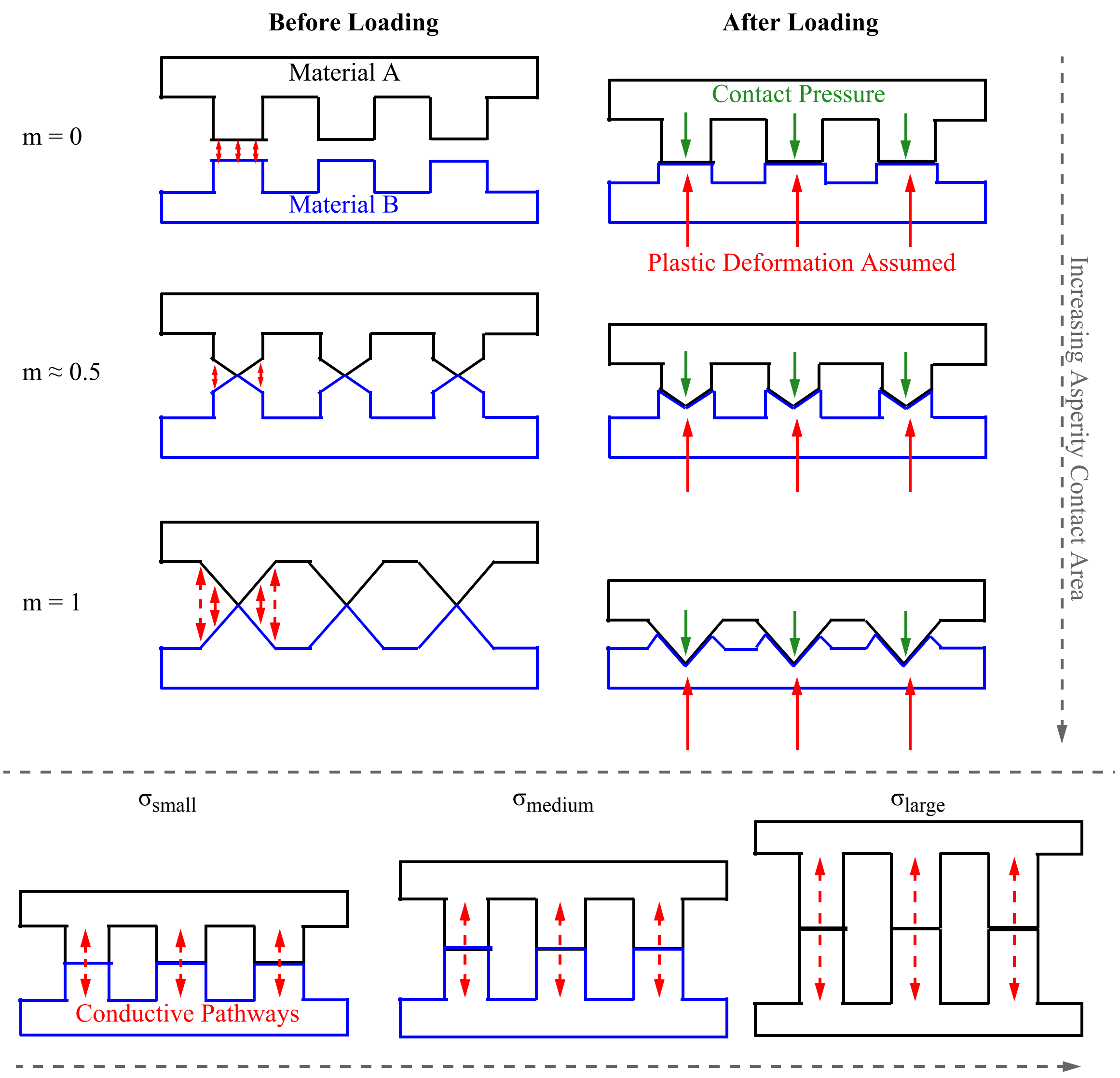


Fig. 3-17. Top: RMS asperity slope variations; Bottom: RMS asperity height variations. Material A has a microhardness higher than Material B. Thus, the mathematical model assumes the asperities of Material B to plastically deform under mechanical contact. Note: these illustrations are simplified and exaggerated for clarity.

“Thermal Contact” option. The remaining other model parameters are held constant. The heater block thickness is also held constant at 5.48 mm.

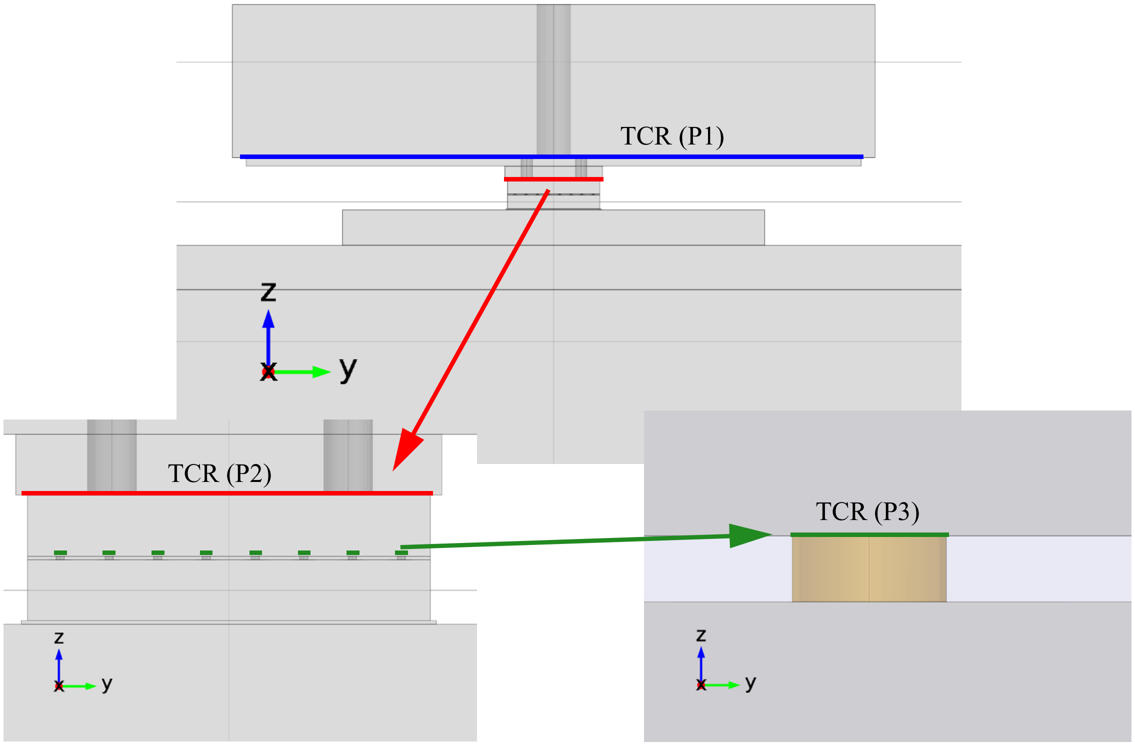


Fig. 3-18. Locations of the three TCR boundaries defined under each "Thermal Contact" option.

Table 3-5. Summary of contact model parameter values for each TCR layer.

Contact Model Parameter	TCR P1	TCR P2	TCR P3
Contact Pressure (80 N Equivalent)	165.3 [kPa]	5.16 [MPa]	324.8 [MPa]
Vickers Microhardness	11.49 [GPa]	11.27 [GPa]	882 [MPa]
Reference	[69]	[70]	[71]

3.10.3 Fitting Procedure and Results

These two parameters are tested by inputting four trial values into the iterative DOE process, as discussed in Section 3.3. A contour response is seen in Fig. 3-19. This matched model can be denoted as Model C. The signal overlap of Model C's matched response over experimental responses is seen in Fig. 3-20. This model's results appear to be a decent

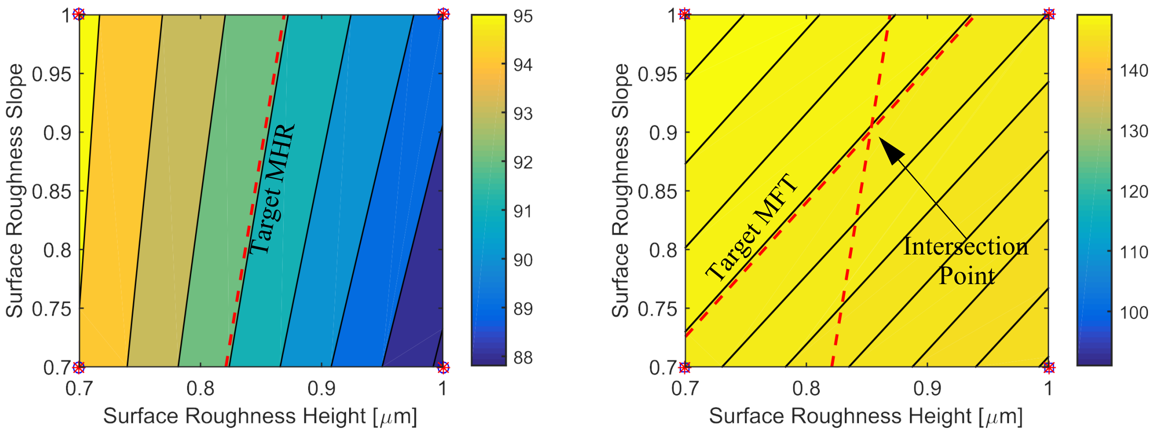


Fig. 3-19. Contour plots from the iterative DOE process used to fit the complex thermal model.

match to the experimental, but not as optically apparent as Model A's results. The signal gap identified in Figs. 3-11 and 3-10 is practically non-existent in Model C's temperature spread. Although Model C's thermal contour slice on the sensor chip surface is nearly identical to that of Models A and B (Fig. 3-12), the left periphery bumps in Model C are slightly hotter by ≈2 °C. As summarized in Table 3-6, the numerical data proves the range of bump

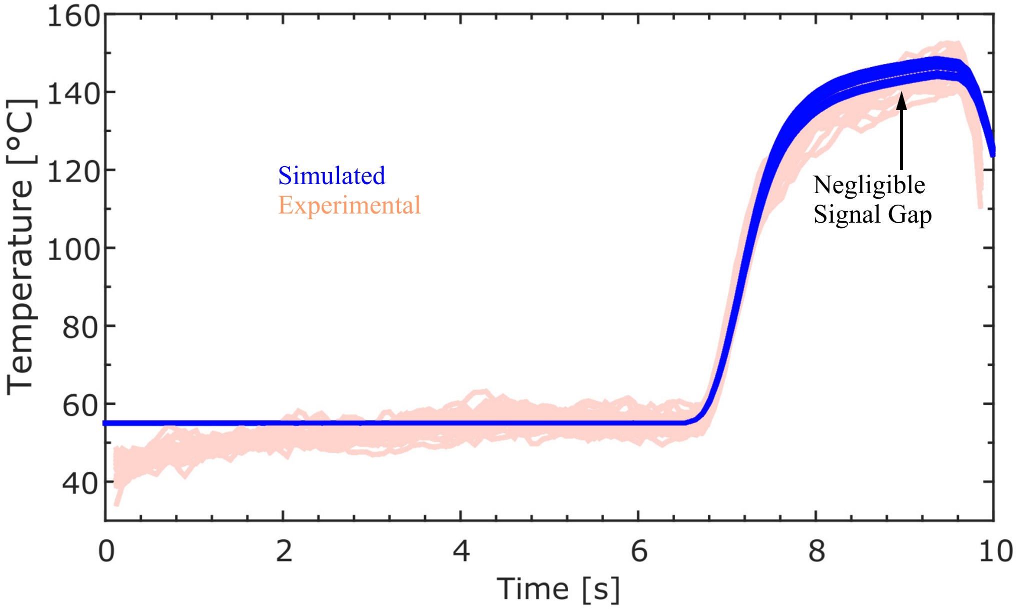


Fig. 3-20. Signal overlap of Model C's simulated response. Note: the signal gap, seen clearly in Models A and B, is not as definitive here

temperatures in Model C are lower than those of Models A and B by about ≈1.2 °C. Interestingly, the RTD thermal range is only ≈0.24 °C higher than the bump range, whereas it is ≈0.85 °C higher than the bump range in Models A and B. While the temperature difference between simulated RTD temperatures and extracted bump temperatures of Model C are slightly higher than those of Models A, the range is still less than 8 °C. Although, this does not explain the lack of a visibly wide signal gap, it is believed that greater conductance of the P3 bump layers have improved thermal distribution on the sensor chip surface and amongst the bumps. This conductance increase is due to the inclusion of mechanical loading. Furthermore, the reduction of the air gap thickness by 10 μm, results in a slightly reduced thermal drop and thus, the surface temperatures on the sensor chip are closer to the temperatures on the bottom surface of the pressure plate. This explains the reduced spread amongst the simulated signals and the lack of the signal gap.

Consistency between the simple thermal and complex thermal models are further observed when comparing their internal thermal bump distributions. The average surface temperature of the top surfaces of the bumps is ≈157.9 °C, and ≈152.5 °C on the bottom. This represents a drop of about 5.4 °C, which is about the same as the average internal drop of Model A. The complex thermal model provides affirmative results on the findings of the simple thermal model.

Table 3-6. Summary of Model C's bump and RTD temperature variation

		Mean	**St. Dev**	**Min.**	**Max.**	**Range**
Model C [°C]	Bump	155.17	1.24	152.18	156.65	4.47
	RTD	147.24	1.35	143.91	148.62	4.71
	ΔT	7.93	0.11	8.27	8.03	0.24

3.10.4 Sensitivity Study III: Effects of Asperity Heights

The results of a sensitivity analysis, varying the matched average RMS asperity height of 0.85 μm by ± 25%, is plotted in Fig. 3-21. For thoroughness, average RMS asperity heights of 2.00 µm and 5.00 µm are also tested, to observe if trends continue as expected. As predicted, the asperity heights affect both MFT and MHR regions without any of the orthogonality seen with air content percentages and heater block thickness variations.

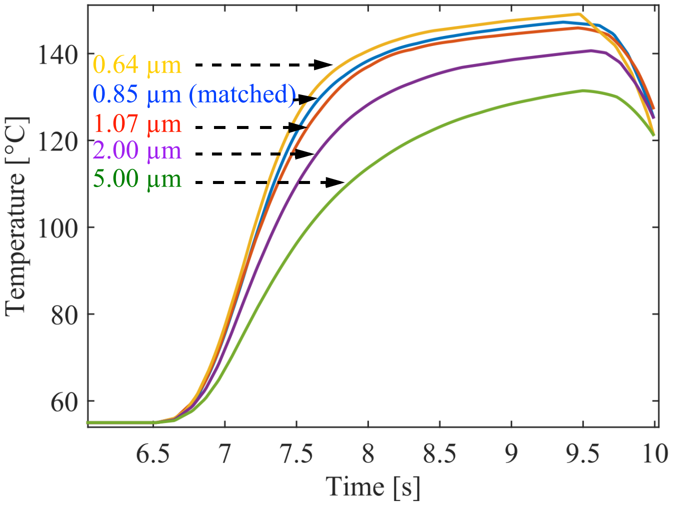


Fig. 3-21. RMS asperity slope variation and its effect on simulated signal responses. Note: the mean of all signals produces one averaged signal for each trial as shown. The average RMS slope is held at 0.9 for all trials.

3.10.5 Sensitivity Study IV: Effects of Asperity Slopes

Varying the average asperity RMS slopes results in similar responses to simulated temperatures to those seen with the average RMS heights. The limitation with slopes is that they can only range between 0 and 1. Therefore, its effects on response parameters is limited, and requires the variation of RMS heights to fit the model to experimental results. As seen in Fig. 3-22, slopes greater than 0.5 bear fairly minimal effects on the simulated responses. As the slope approaches 0, the MHR and MFT region become shallower and lower, respectively.

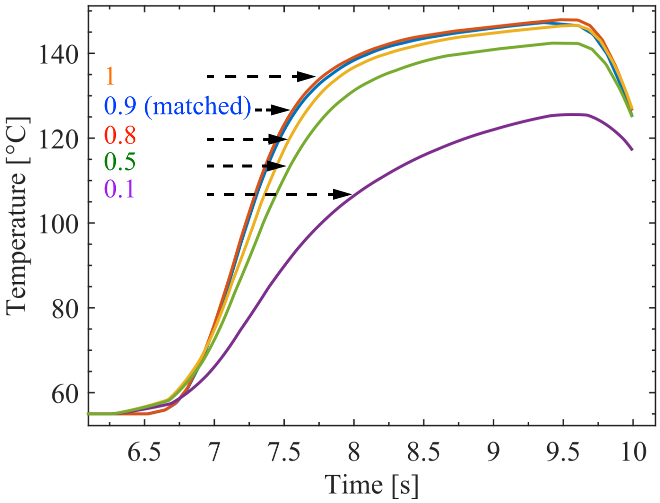


Fig. 3-22. Average RMS slope variation and its effect on simulated signal responses. The average RMS height is held at 0.85 μm for all trials.

3.10.6 Sensitivity Study V: Effects of Contact Pressure

To further confirm proper implementation of the model formulation, the effects of contact pressure on the signal responses are also examined. As the contact pressure is proportional to the constriction conductances, an increase of this pressure results should result in faster heating and smaller thermal gradients. Figure 3-23 confirms this proportionality. Physically, an increased contact pressure results in greater plastic deformation of the surface asperities. This results in better thermal contact and hence, reduced thermal resistance and hence, a better thermal contact.

3.11 Summary

A unique FEA procedure for fitting numerical models to experimental responses is presented. Two simple thermal models are fitted to match two of three average MFT, MHR or FTRS experimental response values, producing simulated signals closely resembling the experimental signals. Both models exhibit nearly identical thermal responses in terms of thermal distribution and heat flow. Model B differs mainly in terms of the rate of heating of the sensor chip. Heating rates are mainly affected by the distance of the heater block boundary condition via the heater block thickness parameter. Vertical thermal gradients are

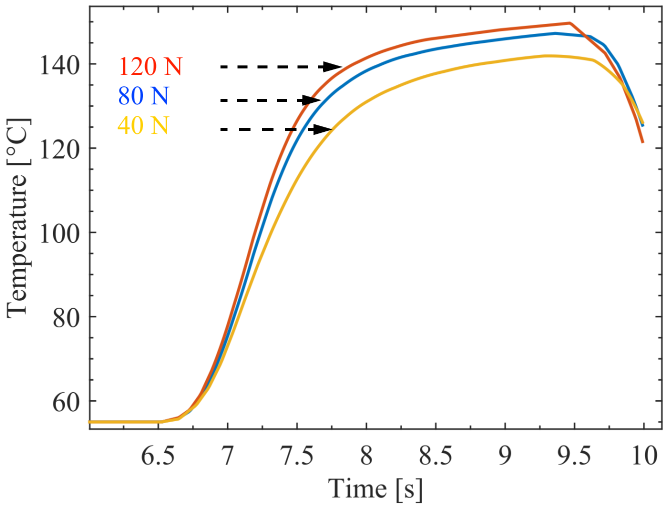


Fig. 3-23. Contact pressure variation and its effect on simulated signal responses. The average RMS height and slope is held at 0.85 μm and 0.9 for all trials, respectively.

most greatly affected by the air content percentages in the simple thermal models. The periphery bumps always experience the least heat due to cooling, heat sink effects by the sensor chip. The center of the bump array experiences the least thermal variation amongst neighboring bumps as well as on the sensor chip surface. The development and results of Model C, a complex thermal model, are also presented. Model C is fit to experimental measurements using average RMS surface asperity heights and slopes. The findings of this model align closely with those from the simple thermal models. With bump temperature ranges representing about 3.4% of the mean MFT values, the thermal variation amongst bumps can be considered low, which is ideal for TCB processes.

4 Conclusions and Outlook

An experimental thermo-compression bonding procedure has been conducted and *in-situ* temperatures have been successfully captured and analyzed. The temperature signals are characterized by four main regions: maximum heating rates, temperatures reached at the end of the press, the steady rise heating rate in the final third of the heating ramp, and lastly, knee points. The RTDs heated fastest, immediately after the bond head applied heat, at a rate up to 108.1 ºC/s, and slowed to an average of 8 ºC/s, nearing the end of the press cycle. Heat energy saturating throughout the sensor chip and substrate components slowed the heating rates from the earlier rapid rises, producing a characteristic KP region in their temperature curves. Thermal gradient analysis across the sensor chip surface determined the greatest thermal gradient magnitudes to occur at the bump array peripheries due to heat sink effects. The peak magnitudes occur immediately after the KP region, which could be further investigated to understand if these thermal gradients result in thermo-mechanical stresses, and if so, how detrimental they are on TCB processes and bond reliability. The corner bumps experience the greatest thermal gradient magnitudes across the sensor chip surfaces as they have the least number of neighbouring conductive pathways. The spread in the temperatures is mainly attributed to these heat sink effects as heat is drawn away from the hot bumps into the cooler, unoccupied silicon mass. Finite element models have been developed in both simple thermal and complex thermal domains, using a formal FEA methodology and a unique, iterative DOE procedure to fit models to experimental responses. For simple thermal models, hybridized material properties are implemented between components in contact to simulate thermal contact resistance. Complex thermal models are developed using surface roughness parameters, based on the Cooper-Mikic-Yovanovich correlation. Both modelling domains have produced considerably successful models; simulated temperature responses show great agreement with experimental responses. In fact, simplified models developed during earlier stages of modelling produced very similar response curves to the experimental response curves, prior to any TCR layer implementation. Furthermore, very similar responses in terms of lateral thermal distribution, bump to bump temperature variation, and internal thermal bump variation have been extracted, analyzed, and observed between both modelling domains.

Thermal variation of both the bumps and sensor chip are attributed to the heat sink effect as the colder, non-bumped region draws heat away from the bump array region. This issue is specific to chip design and can be mediated through more uniform interconnect distribution, which is typically what is found in commercial TCB packages. Greater numbers of interconnections should also result in more uniform lateral thermal distribution and a reduction in interconnect variation. Finally, pitch reductions between interconnects should also result in desired reductions. Parametric analysis of models revealed the following results:

1. Thermal delays from bond head to bumps are sensitive to bond head geometries — thinner heater element blocks can increase heating rates due to the reduction in mass that needs to be heated, and the thickness through which the heat must propagate through to achieve thermal equilibrium.
2. Air content fractions in hybridized TCR layers strongly affect vertical thermal gradients and the bump temperatures — an increase of the percentage of air's highly insulative thermal properties resulted in thermal gradients forming within the hybridized layers, averaging a temperature drop of about 19 ºC from the top of the pressure plate to the top of each bump. The insulative properties also slow the rates at which the heat travels to the bumps, as the TCR layers are all above the bumps, and below the main heat source.
3. Surface asperity heights and slopes both equally affect heating rates of the bumps, and vertical thermal gradients — their linear and inversely linear proportionality to constriction conductance in the CMY correlation result in similar effects on the rates at which bumps heat. These heights and slopes simulate the effects increasing the volumes of insulative microscopic air gaps between two surfaces in thermal contact.

4. Higher contact pressure can improve thermal contact — due to the nature of the CMY correlation, to which it models plastically deformable surface asperities, an increased contact pressure increases the temperature experienced by the bumps.

It is clear that thermal contact resistance plays a significant role in thermal processes, affecting the rate at which heat flows to interconnects and the amount of heat that reaches them. Surface finish quality and higher loading pressures could improve thermal conductance and thus, throughput in TCB processes. Moreover, thermal contact resistance is undeniably the most influential parameter for thermal flow modelling in the FE models presented. The success is limited by the fact that the model was not able to reproduce responses that fit both the MHR and FTRS slopes at the same time. It is currently unknown why the agreement was not nearly as perfect as it could be, and it is possible that another physical mechanism, asides from the tested implementation of ambient air convection, is responsible for "shaping" the characteristic temperature curves. Better agreement between experimental and simulated temperatures can also be possibly made by fine tuning TCR layer values between components in contact, under both the simple thermal and complex thermal domains. This could further improve the accuracy of the simulated temperatures from the FE models, such as at the interconnect regions and throughout the substrate components.

Several other limitations were also discovered in the FE modelling process. Due to the high aspect ratios of the bond head and substrate component geometries with respect to the bump geometries, mesh elements are increasingly difficult to apply. Further refining of the bumps and surrounding regions, required for mesh quality improvement, results in an even greater number of mesh elements which in turn requires significantly greater computational power. This issue amplifies for cases involving several thousand bumps per IC die as well as with cases modelling stacked die configurations. In terms of FEA software, there are perhaps other modelling packages available that could better resolve these issues. Without simplification of models, via means such as anisotropically conductive material homogenization, this problem is currently better suited for higher performance computers, or for parallel computing setups, rather than standalone desktop computers.

Nonetheless, this work has proved useful for not only providing quantitative data from a specific TCB process, but for also shedding light on the dynamics of how heat traverses, depending on the type of TCB setup. Preliminary work from an FE model of a four-layer stacked die setup (Fig. 4-1) has shown that periphery bumps get heated faster than central bumps. This is an interesting result that opposingly contradicts the dynamics of heat flow in the sensor chip FE model configuration where the central bumps were heated first. Further examination revealed that the periphery bumps were heated faster than the central bumps in the stacked die configuration because of the place tool's geometry. Heat around the edges of the upper place tool region travels through the edges of the low place tool fastest, resulted in heat being drawn faster around the edges of the die stack. Consequently, this heat leads to a faster rise in periphery bump temperatures, rather than central temperatures.

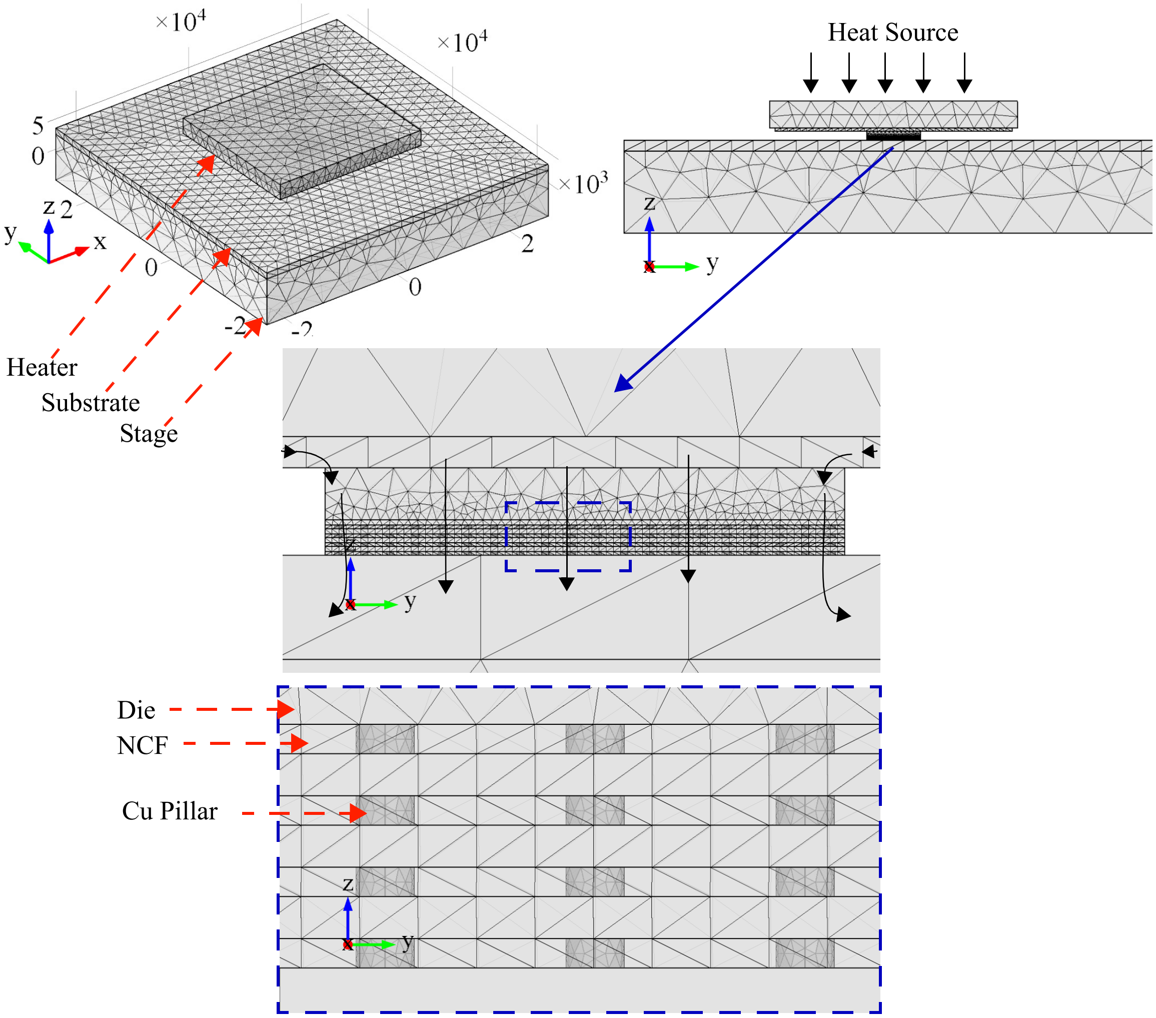


Fig. 4-1. Coarsely meshed translucent FE model sample of a stacked die configuration. Units in µm

The low temperature range of about 7ºC between the RTD temperatures and simulated bump temperatures, suggests the use of microsensor arrays to be potentially highly accurate *in-situ* measurement tools, valuable for further understanding of TCB processes. They can also be viable as tools for process feedback control purposes, once a full setup has been thoroughly developed, and possibly made portable. Experimental *in-situ* sensors used in tandem with FE models have been demonstrated to provide very unique insight on temperatures and thermal features during TCB processes.

4.1 Future Direction

More work can be done to better calibrate the models to the experimental responses, in order to better match all regions. Possible ideas for model refinement include:

1. Surface roughness parameters or thermal conductance measurements extracted through experimental means.
2. Thermo-mechanical transient analysis to include effects of warpage, thermal expansion and bump compression. Thermo-mechanical effects, such as warpage and thermal expansion, may result in higher variations between joints due to possibly reduced contact areas. However, it could also result in even lower variations as mechanical loading can improve thermal processes.
3. Temperature-dependent material properties for completeness.

With respect to the direction of this research, stacked die models for collective bonding processes currently present the greatest thermal challenge of interest. Some of these dimensions, that should be considered in optimizing thermal flow during stacked die TCB processes are:

- Interconnects: sizes, positioning, pitch, periphery vs. area array, C2 vs. C4
- ICs: thicknesses, number of die layers, through-silicon vias (TSVs)
- Underfill: thermal conductivity, thermal expansion
- Substrates: anisotropy, vias, layering, surround ICs

This type of TCB setup presents several challenges due to the multi-dimensional nature of the heat transfer problem. The models and modelling methodology presented can be extended to study thermal processes in multi-layer die configurations.

Using the experimental and numerical methods presented, the effects of each of these components on overall process parameters can be a valuable tool for packaging engineers. Optimizing packages for efficient heat flow during TCB can improve throughput of processes, improve joint reliability, and effectively reduce cost. With the increasingly complex and diverse TCB configurations in development today, it could be highly advantageous to have FE models that could be tuned and adjusted to package design parameters, in order to effectively approximate thermal flow during prototyping and testing.

Bibliography

[1] "Images and Visuals: Microsystem products." *CPMT Images and Visuals.* N.p., n.d. Web. 23 June 2017.

[2] Schwizer, Jürg, Michael Mayer, and Oliver Brand. *Force sensors for microelectronic packaging applications.* Springer Science & Business Media, 2006.

[3] Kwak, Hocheol, and Todd Hubing. *An overview of advanced electronic packaging technology.* Technical Report: CVEL 007-001, 2007.

[4] Harman, George G. *Wire bonding in microelectronics.* McGraw-Hill, 2010.

[5] Shelleman, Eric. *Cost savings example: Elimination of top pareto scrap code, Bond Off Pad (BOP)*. N.p., 30 July 2012. Web. 26 June 2017. <https://ericshelleman.wordpress.com/work/>.

[6] Chauhan, P.S., Choubey, A., Zhong, Z. and Pecht, M.G., 2014. Copper wire bonding. In *Copper Wire Bonding* (pp. 1-9). Springer New York.

[7] Rezvanigilkolaee, Alireza. "Experimental and numerical study of the mechanical aspects of the stitch bonding process in microelectronic wire bonding." (2014).

[8] Satish Shah, Aashish. "Mechanical and tribological aspects of microelectronic wire bonding." (2010)

[9] Qin, I., Yauw, O., Shah, A., Xu, H., Chylak, B. and Wong, N., 2016, March. Advances in wire bonding to lower package cost and improve capability. In *Semiconductor Technology International Conference (CSTIC), 2016 China* (pp. 1-3). IEEE.

[10] Lu, Daniel, and C. P. Wong, eds. *Materials for Advanced Packaging.* Springer, 2016.

[11] DeHaven, Keith, and Joel Dietz. "Controlled collapse chip connection (C4)-an enabling technology." *Electronic Components and Technology Conference, 1994. Proceedings., 44th.* IEEE, 1994.

[12] Tong, Ho-Ming, Yi-Shao Lai, and C. P. Wong, eds. *Advanced flip chip packaging. Springer US*, 2013.

[13] Liang, S.W., Chang, Y.W., Chen, C., Liu, Y.C., Chen, K.H. and Lin, S.H., 2006. Geometrical effect of bump resistance for flip-chip solder joints: Finite-element

modeling and experimental results. *Journal of electronic materials, 35*(8), pp.1647-1654.

[14] Neher, C., Lander, R.L., Moskaleva, A., Pasner, J., Tripathi, M. and Woods, M., 2012. Further developments in gold-stud bump bonding. *Journal of Instrumentation,* 7(02), p.C02005.

[15] Chemicals, Dipsol. "Sn-Ag alloy plating process for bump." *Products.* N.p., n.d. Web. 01 July 2017. <http://en.dipsol-jp.com/products/close-up/sn-ag-alloy-plating-process-for-bump-high-speed-bath>.

[16] Athia, D., Rezvani, A., Clauberg, H., Qin, I. and Mayer, M., 2017, May. Numerical simulations of joint-to-joint temperature variation during thermo-compression bonding. *In Electronic Components and Technology Conference (ECTC), 2017 IEEE 67th* (pp. 1906-1915). IEEE.

[17] Garrou, Phil. "Insights from leading edge." *Solid State Technology.* N.p., n.d. Web. 29 June 2017. <http://electroiq.com/insights-from-leading-edge/2017/01/>.

[18] Purcher, Jack. "Samsung Enters 'State of Emergency' as Apple Chooses TSMC as Exclusive A10 Supplier." *Patently Apple*. N.p., 11 Feb. 2016. Web. 29 June 2017. <http://www.patentlyapple.com/patently-apple/2016/02/samsung-enters-state-of-emergency-as-apple-chooses-tsmc-as-exclusive-a10-supplier.html>.

[19] Swinnen, B., Ruythooren, W., De Moor, P., Bogaerts, L., Carbonell, L., De Munck, K., Eyckens, B., Stoukatch, S., Tezcan, D.S., Tokei, Z. and Vaes, J., 2006, December. 3D integration by Cu-Cu thermo-compression bonding of extremely thinned bulk-Si die containing 10 μm pitch through-Si vias. I*n Electron Devices Meeting, 2006. IEDM'06. International* (pp. 1-4). IEEE.

[20] Nonaka, T., Kobayashi, Y., Asahi, N., Niizeki, S., Fujimaru, K., Arai, Y., Takegami, T., Miyamoto, Y., Nimura, M. and Niwa, H., 2014, May. High throughput thermal compression NCF bonding. In *Electronic Components and Technology Conference (ECTC), 2014 IEEE 64th* (pp. 913-918). IEEE.

[21] Li, M., Tian, D., Cheung, Y., Yang, L. and Lau, J.H., 2015, May. A high throughput and reliable thermal compression bonding process for advanced Interconnections. *In Electronic Components and Technology Conference (ECTC), 2015 IEEE 65th* (pp. 603-608). IEEE.

[22] Xie, L., Wickramanayaka, S., Chong, S.C., Sekhar, V.N., Ismeal, D. and Ye, Y.L., 2016, May. 6um pitch high density Cu-Cu bonding for 3D IC stacking. *In Electronic Components and Technology Conference (ECTC), 2016 IEEE 66th* (pp. 2126-2133). IEEE.

[23] Laor, Ariel. *A novel 8x8 CMOS sensor array for thermal compression bonding with in-situ XYZ force and temperature measurement.* MS thesis. University of Waterloo, 2016.

[24] Lai, Zonghe, and Johan Liu. “The nordic electronics packaging guideline.” *Internet: http://extra. ivf. se/ngl/A-WireBonding/ChapterA. htm (2000)*

[25] Clauberg, H., Marte, A., Yang, Y., Eder, J., Colosimo, T., Buergi, D., Rezvani, A. and Chylak, B., 2015, May. high productivity thermocompression flip chip bonding. *In Electronic Components and Technology Conference (ECTC), 2015 IEEE 65th* (pp. 22-29). IEEE.

[26] Lim, A.B., Rezvani, A., Bacay, R.D., Colosimo, T., Yauw, O., Clauberg, H. and Chylak, B., 2016, November. High throughput thermo-compression bonding with pre-applied underfill for 3D memory applications. *In Electronics Packaging Technology Conference (EPTC), 2016 IEEE 18th* (pp. 427-434). IEEE.

[27] Ulrich, Richard K., and William D. Brown, eds. *Advanced electronic packaging.* Hoboken, NJ: Wiley, 2006.

[28] Dage, Nordson. “Bond Testing of Wire Bonds within LEDs.” *AZoM.com.* N.p., 28 Jan. 2015. Web. 29 June 2017. <http://www.azom.com/article.aspx?ArticleID=11678>.

[29] Nonaka, T., Sugie, R., Suzuki, A. and Ito, M., 2015, May. Non destructive stress evaluation by Raman spectroscopy of flip chip thin die on organic substrate assembled by TCB. *In Electronic Components and Technology Conference (ECTC), 2015 IEEE 65th* (pp. 1148-1153). IEEE.

[30] Asahi, N., Miyamoto, Y., Nimura, M., Mizutani, Y. and Arai, Y., 2015, August. High productivity thermal compression bonding for 3D-IC. *In 3D Systems Integration Conference (3DIC), 2015 International* (pp. TS7-3). IEEE.

[31] R. Daily, G. Capuz, P. Bex and A. Miller, “Understanding the stacked dies interface temperature and its influence during the 3D IC Thermocompression stacking process,” *2012 International Wafer Level Packaging Conference (IWLPC), San Jose, CA, 2012*

[32] Jeong, Y., Choi, J., Choi, Y., Islam, N. and Ouyang, E., 2014, May. Optimization of compression bonding processing temperature for fine pitch Cu-column flip chip devices. *In Electronic Components and Technology Conference (ECTC), 2014 IEEE 64th* (pp. 836-840). IEEE.

[33] Mayer, Michael. *Microelectronic bonding process monitoring by integrated sensors.* 2000.

[34] Laor, A., Athia, D., Rezvani, A., Clauberg, H. and Mayer, M., 2017. Monitoring of thermo-mechanical stress via CMOS sensor array: Effects of warpage and tilt in flip chip thermo-compression bonding. *Microelectronics Reliability*, 73, pp.60-68.

[35] Mayer, M., Schwizer, J., Paul, O., Bolliger, D. and Baltes, H., 1999. In situ ultrasonic stress measurements during ball bonding using integrated piezoresistive

microsensors. *In Proc. Intersociety Electron. Pack. Conf.(InterPACK99)* (pp. 973-978)..

[36] Bex, P., Wang, T., Lofrano, M., Cherman, V., Capuz, G., Sleeckx, E. and Beyne, E., 2017, May. Thermal compression bonding: Understanding heat Transfer by in situ measurements and modeling. *In Electronic Components and Technology Conference (ECTC), 2017 IEEE 67th* (pp. 392-398). IEEE.

[37] Hsieh, M.C., Kang, K., Choi, H. and Kim, Y., 2016, October. Thin profile flip chip package-on-package development. *In Microsystems, Packaging, Assembly and Circuits Technology Conference (IMPACT), 2016 11th International* (pp. 143-147). IEEE.

[38] Sun, P., Xu, C., Liu, J., Geng, F. and Cao, L., 2016, August. Flip chip CSP assembly with Cu pillar bump and molded underfill. *In Electronic Packaging Technology (ICEPT), 2016 17th International Conference on* (pp. 807-811). IEEE.

[39] Lee, M., Yoo, M., Cho, J., Lee, S., Kim, J., Lee, C., Kang, D., Zwenger, C. and Lanzone, R., 2009, May. Study of interconnection process for fine pitch flip chip. *In Electronic Components and Technology Conference, 2009. ECTC 2009. 59th* (pp. 720-723). IEEE.

[40] Murayama, Kei, Mitsuhiro Aizawa, and Takashi Kurihara. “Study of crystal orientation and microstructure in Sn-Bi and Sn-Ag-Cu solder with thermal compression bonding and mass reflow.” E*lectronic Components and Technology Conference (ECTC), 2016 IEEE 66th.* IEEE, 2016.

[41] Li, Yan, and Deepak Goyal, eds. *3D Microelectronic Packaging: From Fundamentals to Applications.* Vol. 57. Springer, 2017.

[42] Lau, John H. “Recent advances and new trends in flip chip technology.” *Journal of Electronic Packaging* 138.3 (2016): 030802.

[43] Matsumura, K., Tomikawa, M., Sakabe, Y. and Shiba, Y., 2015, November. New Non Conductive Film for high productivity process. *In CPMT Symposium Japan (ICSJ), 2015 IEEE* (pp. 19-20). IEEE.

[44] McCracken, M., Mayer, M., Jourard, I., Moon, J.T. and Persic, J., 2010. Symmetric miniaturized heating system for active microelectronic devices. *Review of Scientific Instruments, 81*(7), p.075112.

[45] Mayer, Michael, and Jimy Gomes. “Accelerating reliability assessment with multi-oven racks and sensor chips for wire bonds.” *presentation at NE IMAPS* (2015).

[46] Eslami, M. Reza. *Finite elements methods in mechanics.* Vol. 216. Springer, 2014.

[47] Barkanov, Evgeny. “Introduction to the finite element method.” *Institute of Materials and Structures Faculty of Civil Engineering Riga Technical University* (2001): 1-70.

[48] Turner, M. Jon. “Stiffness and deflection analysis of complex structures.” *journal of the Aeronautical Sciences* (2012).

[49] Szmelter, J. “The energy method of networks of arbitrary shape in problems of the theory of elasticity.” *Proceedings of the IUTAM, symposium on non-homogeneity in elasticity and plasticity, Pergamon Press, New York*. 1959.

[50] Clough, Ray W. “The finite element method in plane stress analysis.” (1960).

[51] Bathe, Klaus-Jürgen. “Finite element procedures: Second Edition.” (2015).

[52] Cardoso, José Roberto. *Electromagnetics through the finite element method: A simplified approach using maxwell's equations.* Crc Press, 2016.

[53] Lockney, Daniel. *NASA Spinoff: Technology Transfer Program.* NASA, n.d. Web. 04 July 2017. <https://spinoff.nasa.gov/spinoff2002/goddard.html>.

[54] Liu, Gui-Rong, and Siu Sin Quek. *The finite element method: a practical course.* Butterworth-Heinemann, 2013.

[55] Bischoff, Jeffrey E., Ellen M. Arruda, and Karl Grosh. “Finite element modeling of human skin using an isotropic, nonlinear elastic constitutive model.” *Journal of biomechanics 33.6* (2000): 645-652.

[56] Maheshwari, B.K., Truman, K.Z., Gould, P.L. and El Naggar, M.H., 2005. Three-dimensional nonlinear seismic analysis of single piles using finite element model: Effects of plasticity of soil. *International Journal of Geomechanics, 5(1),* pp.35-44.

[57] Lin, L., Wang, J., Wang, L. and Zhang, W., 2016. The stress analysis and parametric studies for the low-k layers of a chip in the flip-chip process. *Microelectronics Reliability,* 65, pp.198-204.

[58] Bai, Qiang, Yong Bai, and Weidong Ruan. “Nonmetallic bonded flexible pipe under combined external pressure and bending.” *Advances in Pipes and Pipelines: Flexible Pipes* (2017): 455-474.

[59] Cengel, Yunus A. *Introduction to thermodynamics and heat transfer.* New York: McGraw-Hill, 1997.

[60] Bergman, Theodore L., and Frank P. Incropera. *Fundamentals of heat and mass transfer.* John Wiley & Sons, 2011

[61] Bejan, A. and Kraus, A.D., 2003. *Heat transfer handbook* (Vol. 1). John Wiley & Sons.

[62] Multiphysics, C. O. M. S. O. L. “Heat transfer module user’s guide.” *Documentation within Software Package, Ver* 4 (2014).

[63] Persson, B.N.J., Albohr, O., Tartaglino, U., Volokitin, A.I. and Tosatti, E., 2004. On the nature of surface roughness with application to contact mechanics, sealing, rubber friction and adhesion. *Journal of Physics: Condensed Matter,* 17(1), p.R1.

[64] Yovanovich, M. M. "Thermal contact correlations." *AIAA paper 81* (1982): 83-95.

[65] Aasahi, Noboru, and Masatsugu Nimura. "Heat transfer analysis in the thermal compression bonding for CoW process." *Electronics Packaging (ICEP), 2016 International Conference on.* IEEE, 2016.

[66] Boon Kar, Y., Azrina Talik, N., Sauli, Z., Siow Fei, J. and Retnasamy, V., 2013. Finite element analysis of thermal distributions of solder ball in flip chip ball grid array using ABAQUS. *Microelectronics International, 30*(1), pp.14-18.

[67] Park, A.Y., Ferrone, D., Cain, S., Jung, D.Y., Murray, B.T., Park, S. and Hummler, K., 2013, May. Thermo-mechanical simulations of a copper-to-copper direct bonded 3D TSV chip package interaction test vehicle. *In Electronic Components and Technology Conference (ECTC), 2013 IEEE 63rd (*pp. 2228-2234). IEEE.

[68] Henkel Corporation, Loctite Ablestik 84-1LMISR4 datasheet, April 2014.

[69] "Aluminium nitride AlN - Custom parts." *Precision Ceramics.* N.p., n.d. Web. 27 July 2017. <http://www.precision-ceramics.com/materials/aluminum-nitride/>.

[70] Haynes, William M., ed. *CRC handbook of chemistry and physics.* CRC press, 2014.

[71] Marsden, John, and Iain House. *The chemistry of gold extraction.* SME, 2006.

APPENDICES

Appendix A: Rate Algorithm

```
%file_name = “heatingrate.m”

% The following code is used to calculate the maximum heating rates. It can be applied to other
problems requiring points of the greatest rates of change in a quantity.

maxvals= []; t_maxvals= []; %initialize empty matrices for max rate value storage
[m,i1]=min(abs(x-6)); %time value and index of 6 s
[m,i2]=min(abs(x-9.5)); %time value and its index of 9.5 s
x1=x(i1:i2); %x values are the times between index i1 and i2
figure; %open a new figure to plot the rates

        for j=1:length(y(:,1)); % y-matrix contains the temperatures for each pad for all times
                y1=y(j,i1:i2)'; %transpose of temperatures between pre-defined times
                dx=diff(x1);dy=diff(y1); %difference between adjacent x or y values
                d=smooth(dy./dx,10); %takes the differential and smooths each signal
                xx=(x1(2:end)+x1(1:end-1))/2;%takes the average of the time values to plot
                plot(xx,d); hold on; %plot the resulting rates
                [ma,ima]=max(d); finds the maximum heating rate for each curve and its time index
                maxvals=[maxvals ma];t_maxvals=[t_maxvals x1(ima)]; %stores max, heating rates ad
                times for each rate.
        end
```

Appendix B: Knee-Point Algorithm

```
%file_name = “kpregion.m”
% this code is used to calculate the knee-points of each of the heating curves.

maslo=[ ];%max slope
t_knee=[ ];%time of knee point
T_knee=[ ]; %temp of knee point
T_B_s = [ ]; %max temps

% find times when bond head temps rise and when they drop
    [Tmi,imi]=min(y_BHD);tmi=x_BHD(imi);iri=imi; % index where it starts to rise
    while y_BHD(iri)<Tmi+0.5;iri=iri+1; % adjust based on signal
        if iri==length(y_BHD);asdfasdf;end;
    end;
    [Tma,ima]=max(y_BHD);tma=x_BHD(ima);idr=ima; % index where it starts to drop
    while y_BHD(idr)>Tma-1;idr=idr+1;
        if idr==length(y_BHD);asdfasdf;end;
    end
        figure(1); hold off;

t_B=x_BHD(idr); %time corresponding to imi index value
t_C=x_BHD(iri); %time corresponding to ima index value

figure(2);
plot(x_RTD,y_RTD,'b');hold on
xlabel('Time [s]');ylabel('Temperature [C]')

    for sig=1:length(y_RTD(:,1)); % loop over all signals
        x=x_RTD;y=y_RTD(sig,:);
        % 1) find max slope
        dx=diff(x);dy=diff(y');figure(3);
        hold off;subplot(2,1,1);plot(x,y);hold on
        d_=smooth(dy./dx,10);xx_=(x(2:end)+x(1:end-1))/2;
        [m,i1]=min(abs(xx_-6));[m,i2]=min(abs(xx_-9.5));
        x1=xx_(i1:i2); y1=d_(i1:i2);
        [ma,ima]=max(y1);t_ma=x1(ima);a=axis;axis([a(1:2) -ma/20 ma*1.05])
        plot(t_ma,mean(y(ima+i1-1:ima+i1)),'.');hold on
        xlabel('Time (s)');ylabel('Temperature (C)');
        T_ma=mean(y(ima+i1-1:ima+i1));
        maslo=[maslo ma];T_maslo=[T_maslo T_ma];t_maslo=[t_maslo t_ma];
```

```
        % 2) find pt A
        d=T_ma-ma*t_ma; %y intercept
        t_A=(50-d)/ma; % A time where max slope hits 50 deg C
        t_A_s=[t_A_s t_A]; %times of point A
        T_B_s = [T_B_s max(y(x<t_B))]; %Temps for where RTDs meet max before t_B bond
        head temp drops)
        % 5) do the method of “maximum slope difference” to find the knee-point:
        x_em=[]; % slope difference and ends meet methods
        figure(8); hold off
                for j=2:length(x)-1;
                        x1=x(1:j);y1=y(1:j);x2=x(j:end);y2=y(j:end);
                        [p1,S1]=polyfit(x1,y1',1);yf1=polyval(p1,x1,S1); %linear fit 1
                        [p2,S2]=polyfit(x2,y2',1);yf2=polyval(p2,x2,S2); %linear fit 2
                        title(num2str(sig))
                        plot(x1,y1,x2,y2,x1,yf1,x2,yf2); drawnow;
                        em=[em max(yf1)-min(yf2)];x_em=[x_em x(j)]; %difference
                        between maximum of the first linear fit and the minimum of the
                        second fit. Intersection is when this value is zero
                end
        [masd,i]=max(sd);
        x_em1=interp1(em,x_em,0);
        subplot(2,1,1);hold on
        plot(x_em1,interp1(x,y,x_em1),'g*')
        t_knee=[t_knee x_em1];T_knee=[T_knee interp1(x,y,x_em1)];
        figure(9);
        plot(t_knee(end),T_knee(end),'y.',t_maslo(end),T_maslo(end),'g.');
        title(num2str(sig))
end
```

Appendix C: Thermal Gradient Code

```
%file_name = “thermgradientmag.m”

temp2_rtd = y_RTD0; %temperature values measured
time_vector_target = x_RTD0; % time values measured
max_rtd = max(max(temp2_rtd(:,:))); %maximum temp. value experienced for all pads of interest
min_rtd = min(min(temp2_rtd(:,:)));  %minimum temp. value
grid_x =
[0,1,2,3,4,5,6,7,0,1,2,3,4,5,6,7,0,1,2,3,4,5,6,7,0,1,2,3,4,5,6,7,0,1,2,3,4,5,6,7,0,1,2,3,4,5,6,7,0,1,2,3,4,5,6,7,0,
1,2,3,4,5,6,7]; %grid for the x values of contour plot
grid_y =
[7,7,7,7,7,7,7,7,6,6,6,6,6,6,6,6,5,5,5,5,5,5,5,5,4,4,4,4,4,4,4,4,3,3,3,3,3,3,3,3,2,2,2,2,2,2,2,2,1,1,1,1,1,1,1,1,0,
0,0,0,0,0,0,0]; %grid for the y-values of the contour plot
grid_x2 = [ -1400, -1000, -600, -200, 200, 600, 1000, 1400, -1400, -1000, -600, -200, 200, 600, 1000, 1400,
-1400, -1000, -600, -200, 200, 600, 1000, 1400, -1400, -1000, -600, -200, 200, 600, 1000, 1400, -1400, -
1000, -600, -200, 200, 600, 1000, 1400, -1400, -1000, -600, -200, 200, 600, 1000, 1400, -1400, -1000, -600,
-200, 200, 600, 1000, 1400, -1400, -1000, -600, -200, 200, 600, 1000, 1400]; %x-axis values of the pad
locations
grid_y2 =
[1400,1400,1400,1400,1400,1400,1400,1400,1000,1000,1000,1000,1000,1000,1000,1000,600,600,600,600
,600,600,600,600,200,200,200,200,200,200,200,200,-200,-200,-200,-200,-200,-200,-200,-200,-600,-600,-
600,-600,-600,-600,-600,-600,-1000,-1000,-1000,-1000,-1000,-1000,-1000,-1000,-1400,-1400,-1400,-
1400,-1400,-1400,-1400,-1400]; %y axis values of the pad locations
num_lines = 25; %25 was previous default value; red_line = 1.0;
max_T =[];min_T =[]; maxGradient = []; TempRange = [];%initialization of variables

        for i = 1:length(time_vector_target) %sweep through all of the time points
                figure(3);
                pitch = 400; %400 um pitch
                grid_x3 = [-1200,-800,-400,0,400,800,1200,-1200,-800,-400,0,400,800,1200,-1200,-800,-
                400,0,400,800,1200,-1200,-800,-400,0,400,800,1200,-1200,-800,-400,0,400,800,1200,-
                1200,-800,-400,0,400,800,1200,-1200,-800,-400,0,400,800,1200]; %x-axis values for
                interpolated central locations
                grid_y3 =
                [1200,1200,1200,1200,1200,1200,1200,800,800,800,800,800,800,800,400,400,400,400,4
                00,400,400,0,0,0,0,0,0,0,-400,-400,-400,-400,-400,-400,-400,-800,-800,-800,-800,-800,-
                800,-800,-1200,-1200,-1200,-1200,-1200,-1200,-1200]; %y-axis value for the interpolated
                central locations
                tempgrid = reshape(temp2_rtd(:,i),[8 8])'; %reshape into 8 x 8 matrix - transpose for pad
                numbering (1,1) = 0, (8,8) = 63
```

```
tempx = grid_x3; tempx = tempx'; %transpose both grids
tempy = grid_y3; tempy = tempy';
subplot(1,3,1);
plot(time_vector_target(:),temp2_rtd(:,:));
xlabel([num2str(time_vector_target(i)*1000),' ms'],'fontSize', font_size);
line([time_vector_target(i),time_vector_target(i)],[min_rtd,max_rtd],'LineWidth',red_line
'color','r');hold off;
ylabel('Degrees Celsius');
title ('RTD Temp.');
axis square;
subplot(1,3,2);

set(gca,'FontSize',axis_font); %contour plot produced here
g = gridfit(grid_x,grid_y,temp2_rtd(:,i),0:1:7,0:1:7,'regularizer','diffusion');
contourf([1:8],[1:8],g,num_lines); xlim([1,8]);ylim([1,8]);hold on;%     colorbar;
colorbar('Location','southoutside')
max_T = [max_T max(temp2_rtd(:,i))];
min_T = [min_T min(temp2_rtd(:,i))];
title([num2str(max_T(end)-min_T(end),'%0.4f [°C]'),' range', ],'fontSize', font_size);
set(gca,'XTickLabel',");set(gca,'YTickLabel',");set(gca,'XTick',[1:8]);grid;
axis square;
hold off;
subplot(1,3,3); %gradient quiver plot produced here
hold on;
F = scatteredInterpolant(grid_x2',grid_y2',temp2_rtd(:,i),'natural');
FF = F(tempx,tempy);
FFF = (reshape(FF,7,7))';
[DX,DY] = gradient(FFF,400,-400);
DXY = sqrt(DX.^2 + DY.^2);
DX = DX'; DY = DY';
DDX = reshape(DX,49,1);
DDY = reshape(DY,49,1);
quiver(tempx,tempy,DDX,DDY)
scatter3(grid_x3,grid_y3,F(grid_x3,grid_y3),[],F(grid_x3,grid_y3),'filled')
scatter3(grid_x2,grid_y2,temp2_rtd(:,i),[],'k','filled');
set(gca, 'CLim', [min(min((F(grid_x3,grid_y3)))), max(max(F(grid_x3,grid_y3)))+1])
colorbar('Location','southoutside')
view(0,90)
set(gca,'YTick',[-1400,-1000,-600,-200,0,200,600,1000,1400] ); % y axis labels
set(gca,'XTick',[-1400,-1000,-600,-200,0,200,600,1000,1400] ); % x axis labels
padnum = num2str((1:1:64)');
padlabel = cellstr(padnum);
dx = 50; dy = 0.1; % displacement so the text does not overlay the data points
```

```
    text(grid_x2+dx,grid_y2+dy,padlabel);
    xlabel({['Distance from Origin {\mu}m]'],['Time =
    ',num2str(time_vector_target(i)*1000),' ms']}) % x-axis
    ylabel('Distance from Origin {\mu}m'); title({['Max Temperature Gradient =
    ',num2str(max(max(DXY))), '{\circ}C/{\mu}m'],['Min Temperature Gradient =
    ',num2str(min(min(DXY))), '{\circ}C/{\mu}m']})
    axis square;
    hold off;
    maxGradient = [maxGradient max(max(DXY))];
end;
```