Thermal Management of
Solid State Power Switches

Christopher James Frederick Tighe

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy
June 2011

School of Mechanical, Materials and Manufacturing Engineering
University of Nottingham, Nottingham, UK

Abstract

The transient temperature of solid state power switches is investigated using thermal resistance network modelling and experimental testing. The ability of a heat sink mounted to the top of the device to reduce the transient temperature is assessed. Transient temperatures for heat pulses of up to 100ms are of most interest.

The transient temperature distribution inside a typical stack-up of a solid state power switch is characterised. The thermal effects of adding a heat sink to the top of the device are then assessed. A variety of heat sink thicknesses and materials are evaluated. Components of the device stack-up are varied in order to assess their affect on the effectiveness of the heat sink in reducing the device temperature.

Thermal networks are successfully applied to model the transient heat conduction inside the stack-ups. This modelling technique allowed a good understanding of the thermal behaviour inside the stack-up and heat sink during the transient period. The concept of using a heat sink to suppress the transient temperature was validated experimentally on two types of solid state power switch.

Acknowledgements

Contents

Notation

A	$[m^2]$	Area
c_p	$[J/kg\,K]$	Specific Heat Capacity
C	$[J/K]$	Thermal Capacity
E	$[J]$	Energy
Fo		Fourier Number
h	$[W/m^2\,K]$	Heat Transfer Coefficient
I	$[A]$	Current
k	$[W/m\,K]$	Thermal Conductivity
P	$[W]$	Power
q	$[W]$	Heat Transfer Rate
r	$[m]$	Radius
r_s		Residual
R	$[\Omega]$	Resistance
T	$[K\ or\ ^{\circ}C]$	Temperature
V	$[V]$	Voltage
u	$[J]$	Internal Energy
w	$[W]$	Work Rate
x	$[m]$	Length
$Z_\theta(t)$	$[K/W]$	Transient Thermal Impedance

Greek Symbols

α	$[m^2/s]$	Thermal Diffusivity
ρ	$[kg/m^3]$	Density
τ	$[s]$	Time-constant
ω		Relaxation Factor

Glossary

AMB	Active Metal Bonding. A substrate manufacture process where copper is bonded on the ceramic substrate using Silver-Copper-Titanium brazing.
DCB	Direct Bond Copper. A substrate manufacture process where copper and ceramic are bonded directly.
Imide	A thin polymer layer applied to some electronic devices.
Biot Number	A dimensionless number which gives the ratio of the heat transfer rate at the boundary of a model to the heat transfer rate inside the body.

Chapter 1

Introduction

Power electronics involves the processing of electrical power (voltage, current and frequency) by means of solid-state electronics. It is used in a wide range of applications, from mobile phones to space shuttle power supply systems. The aerospace industry also heavily employs power electronics to control electrical supplies within aircraft.

A major design limit for modules of power electronic devices, such as solid-state power switches, is the device temperature; an estimated 55% of electronic component failures are related to temperature[1]. The semiconductor industry is driven to decreasing the size of semiconductor devices. Coupled with the continuing increase in power dissipation, the power density of devices is rising which puts pressure on cooling technologies to develop sufficiently to allow progression in the electronics sector without compromising reliability.

This thesis focuses on the use of solid-state power switches as a replacement for mechanical switches in aircraft. The benefits of using solid-state power switches over mechanical switches include faster switching times, as well as quieter operation and increased lifetime as they have no moving parts. However, solid-state power switches experience increased heat dissipation due to the voltage drop across them. Effective management of this heat generation is required in order to increase the operating

lifetime of the devices. In particular, short duration transient heat losses are a key issue which need to be considered when designing thermal management systems.

Power electronic devices must be designed to handle the largest power dissipation to which they may be subjected. Short surges can occur due to over-current faults, which dissipates heat within devices. In some applications it may be adequate to switch devices off during such transients, but in the aerospace industry it is preferred that aircraft equipment is not switched off unnecessarily. Devices are required to ride-through these transients to allow time for faults to be diagnosed and unnecessary shut downs to be avoided.

In these cases, systems must be designed to handle the transient currents, where heat dissipation may be up to ten times greater than during normal operation. The present method of managing current surges involves over-sizing power modules to provide more thermal mass, which allows more current to be dissipated without exceeding the temperature limit. Power electronic systems designed in this way are over-sized and heavy and are underused for the majority of the time during normal operation.

Some power electronic switches are used for switching at high frequencies. In these applications, a significant amount of heat is generated during the switching process (known as switching losses). This thesis considers the use of power electronics in the 'on' state, which therefore allows switching losses to be ignored.

1.1   Problem Definition

A new device-level thermal design to handle short periods of increased heat generation is required. This would reduce redundancies and give reductions in system size, weight and cost. The new design must therefore be able to reduce the temperature of power electronic devices for transients of up to 100ms in length, the time typically needed to allow for fault diagnosis, or in aircraft, to tolerate lightning strikes.

A typical power electronic stack for a semiconductor device (also known as a die) is shown in Figure 1.1. Typical dimensions for the device stack are listed in Table 1.1. The stack is usually packaged for protection (omitted in Figure 1.1), and the only exposed surface is the bottom of the baseplate, which is normally cooled via convection. Heat is generated within the device and is conducted through the stack to the cooled baseplate surface where it is conducted away from the system.
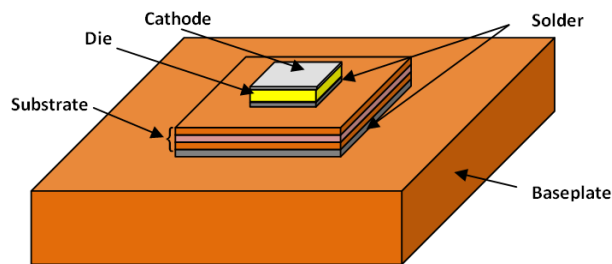
Figure 1.1: A typical device stack.

Layer	Thickness (mm)	Material	Width/Depth (mm)
Cathode	0.02	Aluminium	10
Die	0.4	Silicon	10
Solder	0.1	Solder Sn96.5-Ag3.5	10
Substrate: Copper	0.3	Copper	29
Substrate: AlN	0.6	Aluminium Nitride	29
Substrate: Copper	0.3	Copper	29
Solder	0.1	Solder Sn96.5-Ag3.5	29
Baseplate	5	Copper	60

Table 1.1: Typical thickness and material of the components used in the stack.

Conduction is the main heat transport mechanism for heat generated within the device. All heat generated must be conducted to the baseplate surface in order to be convected out of the system. This setup is not effective for cooling short transients, as the cooling fluid at the bottom of the baseplate is too far away from the heat source to allow it to be effective for the shorter transients. The structure of the stack-up must therefore

be modified to improve thermal device management for short heat surges. Minimising the steady state thermal resistance is not a concern as this is not the limiting factor in the thermal design of the system.

A potential solution to the problem is to use the top surface of the die to provide an additional thermal route for generated heat. A solid heat sink attached to the top of the device would allow the heat to be conducted into the heat sink, where it could be stored during the heat pulse and then conducted back through the device after the heat pulse. A diagram of the modified device stack can be seen in Figure 1.2.

The aim of the project is to investigate the use of the heat sink as a thermal mass close to the die to suppress short thermal transients.
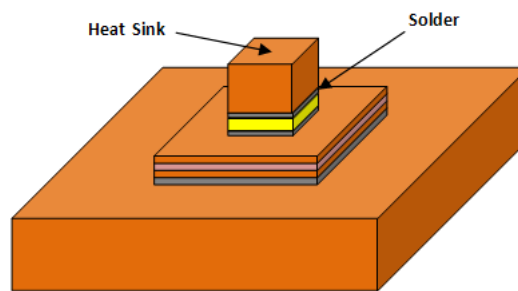
Figure 1.2: The typical device stack with the addition of the heat sink and solder layer on top of the die.

1.2 Objectives

The aim of this thesis is to investigate the heat transfer within a power electronic device stack.

There are two main objectives of this work:

1. To investigate the use of a thermal mass to reduce device temperatures during transients by using thermal resistance networks to model transient heat conduction.

This includes modelling a variety of stack configurations and modifications to assess improvements to the device temperature.

2. To use experimental testing to validate the thermal modelling results. At this stage, the practicality (in terms of cost and ease of manufacture) of the stack designs are considered, as a prototype is manufactured for experimental testing.

1.3 Thesis Structure

The thesis begins with an overview of power electronic devices and heat transfer, presented in Chapter 2. This aims to introduce heat generation mechanisms in power electronics and outline some key heat transfer theory. A literature survey is presented in Chapter 3, which relates the work presented in this thesis to previous work performed and the state-of-the-art. Chapter 4 presents techniques which may be adopted for modelling heat conduction.

The main part of the study consists of two parts:

1. Chapter 5 begins with a detailed description of the method chosen for modelling the transient heat conduction. This leads into the results and analysis from the 1D modelling. A sensitivity analysis of the stack-up on the heat sink performance is performed and presented. The modelling work is then extended into Chapter 6. In this chapter the 2D modelling results are presented, along with discussion into the differences between the 1D and 2D modelling results.

2. The experimental verification of the modelling results is presented in Chapter 7. The experimental facility is detailed, including the apparatus, rig setup and thermal measurement techniques used. The manufacture of diodes with a heat sink on is described followed by the results of the experimental testing. The chapter is concluded with the comparison between the modelling and experimental results, with reasons for any discrepancies discussed. The experimental

work is extended in Chapter 8, in which heat sinks are used on MOSFET devices. Difference in the manufacture and testing procedure compared to the experiments using diodes is explained. The experimental results follow, with a comparison to the numerical results when modelling the experimental setup.

General conclusions and recommendations for further work are finally given in Chapter 9.

Chapter 2

Overview of Power Electronic Devices and Heat Transfer

This chapter aims to explain some aspects of the physics of power electronic devices that are important to heat generation, beginning with the introduction of Breakdown Voltage and Avalanche Breakdown, which are causes of device failure. Following on from this is a section discussing the mechanisms behind the generation of heat in power electronic devices. Different power electronic devices are then discussed which are relevant to the thesis. The theory section on power devices has been written using the following references: [2, 3, 4, 5, 6, 7].

In the second part of the chapter heat transfer theory is discussed. The heat conduction equation is derived, from which important material properties for heat conduction are identified. The chapter then focuses on transient heat conduction. Three transient heat conduction regimes are identified: the early, transitional and late regime. The behaviour of heat conduction through these regimes is analysed, leading to the definition of the Fourier number. Using the Fourier number, a time constant is defined, which is used throughout the thesis when analysing the thermal behaviour inside the stack.

2.1 Power Electronics Overview

In order to model the heat generation mechanism in power electronic devices, it is necessary to first develop an understanding of the devices. Various different types of device exist, each of which is composed of different regions of doped silicon. Although the operation and architecture of these devices differ, they share some fundamental principals of operation. It is assumed that the fundamental behaviour of power electronic devices, such as that of electrons at *p-n* junctions, is known.

2.1.1 Breakdown Voltage and Avalanche Breakdown

Power electronic devices are constructed using regions of different doping. By exploiting the electrical behaviour of these junctions, devices can be created that allow electrical flow to be manipulated and controlled. Knowledge of this background theory including semiconductor doping and junction behaviour under forward and reverse bias electrical flow is assumed, and can be found in numerous text books, such as those listed at the beginning of the chapter.

The *Breakdown Voltage*, BV_{BD}, is a feature of the reverse biased region of an IV curve such as the one shown in Figure 2.1. When sufficient reverse biased voltage is applied to the junction, a sudden rapid increase in current occurs. The current increase leads to a large amount of power dissipation which results in device failure due to the increased heat within the device. This process is termed the *Avalanche Breakdown*.

Avalanche breakdown is caused by a process called *Impact Ionisation*. The kinetic energy of electrons is directly proportional to the electric field applied to them. When electrons with a large enough kinetic energy collide with silicon atoms, the transferred energy can break one of the silicon's covalent bonds, releasing an electron. The released electron may then gain enough kinetic energy from the electric field to break a covalent bond in another atom. The numbers of electrons in the field increase as
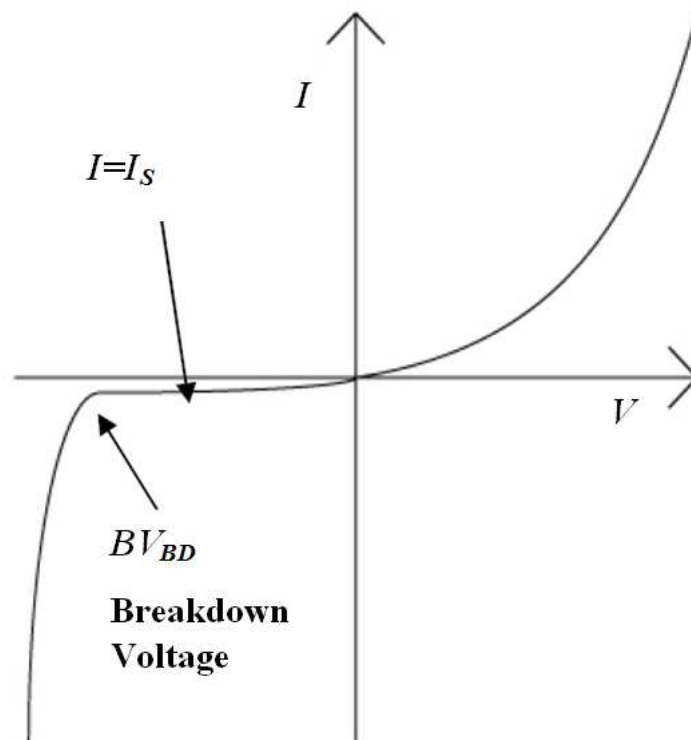
Figure 2.1: The *I-V* characteristics of a *p-n* junction.[2]

covalent bonds are broken, which in turn increases the number of impact ionisations. This cascading effect leads to a dramatic increase in current as the number of electrons increase. Once again, the increased current leads to an increase in power dissipation, which can lead to device destruction through overheating.

2.2 Power Devices

Power devices require a more complicated structure than low-power counterparts. Additional structural features are necessary due to the fact these devices require to handle high power loads. A selection of power electronic devices are discussed which bear relevance to this thesis. The power diode and MOSFET are both used for experimental work, and therefore an understanding of their operation is given.

2.2.1 Power Diodes

Diodes contain a single p-n junction, as shown in Figure 2.2. Two heavily doped regions, the p^+ and n^+ substrate, are separated by a lightly doped n^- region, called the epitaxial. This region of light doping absorbs the depletion layer, allowing large reverse biased voltages to be applied across it without the depletion region growing excessively wide. As such, the width of the epitaxial layer is dependent on the designed breakdown voltage, and is around 10μm/100V. The p^+ and n^+ substrate layers have fixed widths of around 10μm and 250μm, respectively.
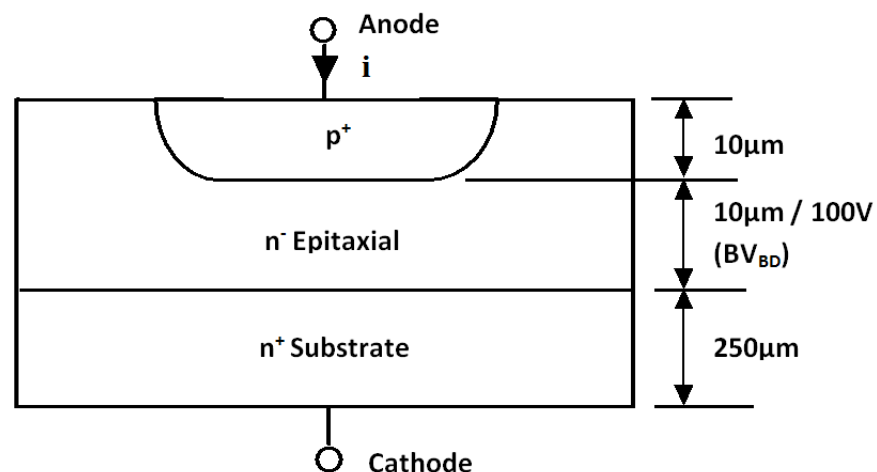
Figure 2.2: Vertical cross-section showing the structure of a power diode.[2]

2.2.2 MOSFETs

Metal-Oxide-Semiconductor Field Effect Transistors (MOSFETs) are used in applications where large off-state blocking voltages and high current-carrying capabilities are needed. Figure 2.3 shows the vertical cross-section of a MOSFET cell. Thousands of cells are used in parallel to create a power MOSFET device with large gain and low on-state resistance.

A MOSFET has three connections: the source, drain and gate. The *n-p-n* junctions between the source and drain makes electrical conduction appear impossible between the two. In order to achieve conduction between the source and drain, a voltage must be applied between the gate and source.
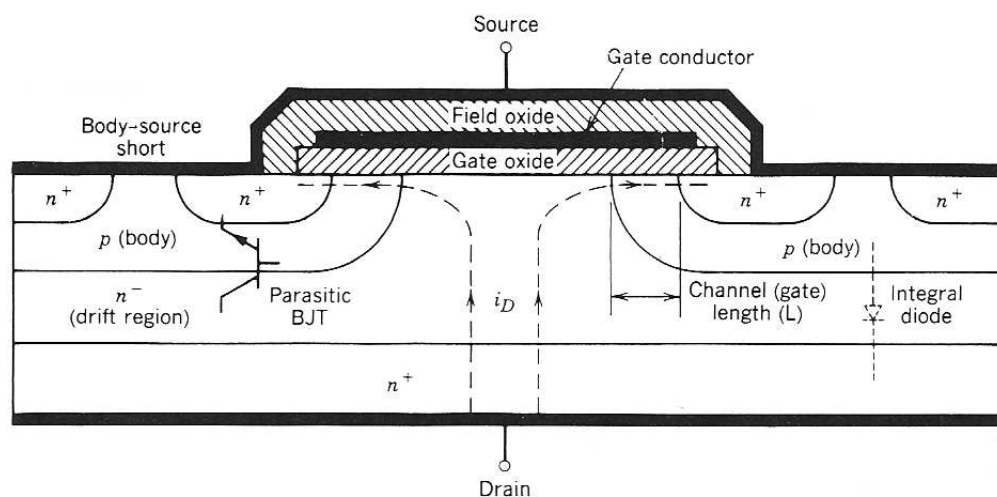
Figure 2.3: A cross-sectional view of a MOSFET cell, showing the doped regions and the three connectors: source, drain and gate.[2]

The gate consists of the *gate conductor* (the metallisation region), a non-conducting silicon dioxide layer (SiO_2, termed the *gate oxide*) and the silicon beneath the gate oxide. Applying a small gate-source voltage (V_{GS}) creates a depletion region at the interface between the silicon and gate oxide. The positive charge on the gate conductor repels holes directly beneath the gate oxide, allowing electrons from the source connection to ionise the acceptor atoms in the *p*-region. These negatively charged ionised acceptors are drawn to the negatively charged gate conductor, creating

a depleted region directly beneath the gate oxide.
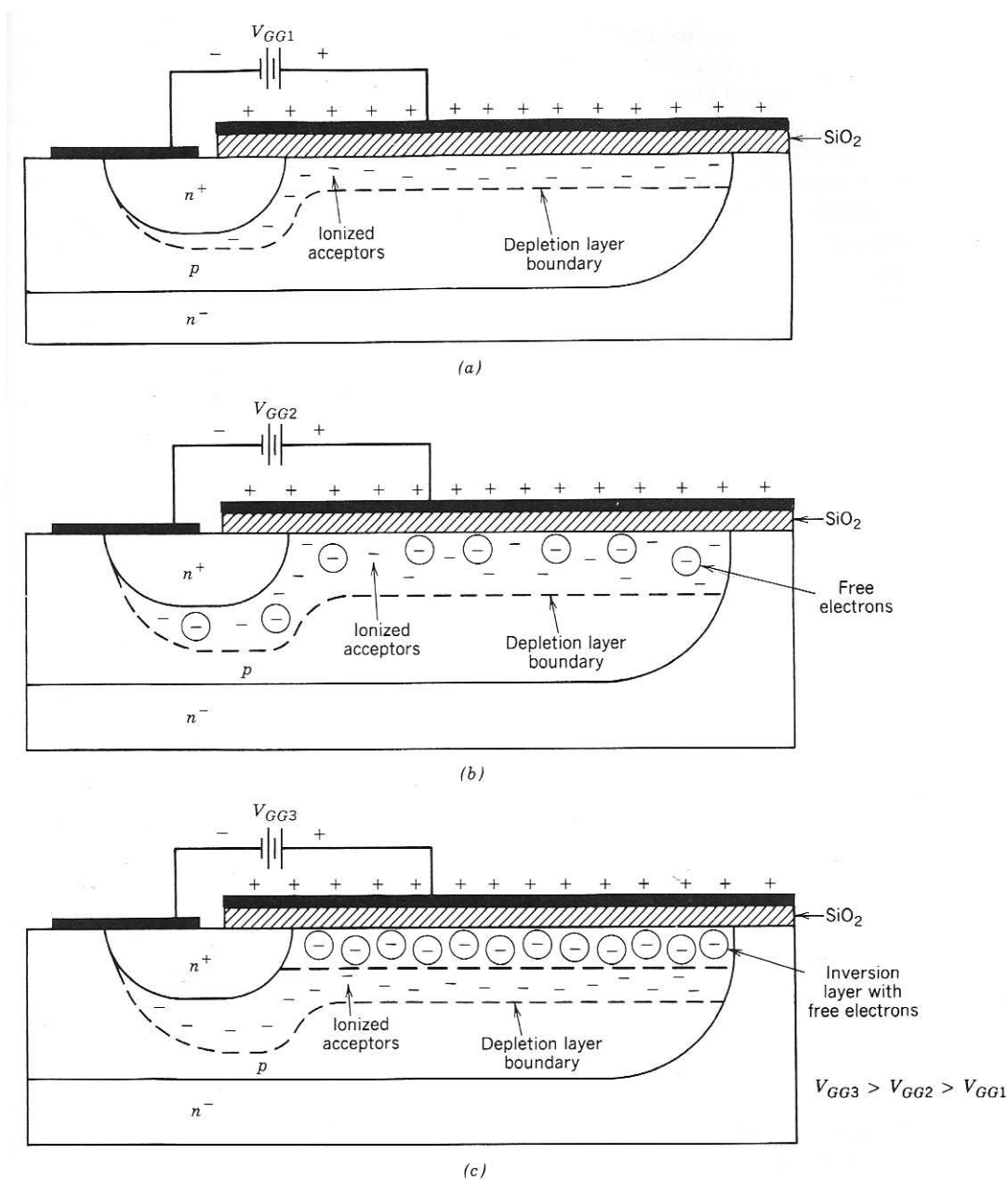
Figure 2.4: The formation of the depletion and inversion layers beneath the SiO_2 gate oxide as the magnitude of V_{GS} increases.[2]

Increasing V_{GS} causes the depletion region to grow in thickness. The electric field at the oxide-silicon interface grows, attracting free electrons to the surface. The layer of electrons is termed the *inversion layer*, providing a conductive path for current between the source and drain. The growth of the inversion layer is shown in Figure

2.4.

The size of the channel created between the source and drain is dependent on the size of V_{GS}; increasing V_{GS} widens the channel width. This reduces the resistance between the source and drain, allowing a greater current flow. For current flow between the source and drain, V_{GS} must be greater than the threshold value of the MOSFET, $V_{GS(th)}$.

2.2.3   IGBTs

The structure of Insulated Gate Bipolar Transistors (IGBTs) is illustrated in Figure 2.5. They are similar in construction to MOSFETs, with the addition of a p^+ injecting layer which forms the drain of the IGBT. This p^+ injecting layer forms a *p-n* junction (J_1) which injects minority carriers into the buffer layer, which is the drain connection in the MOSFET.

The on-state operation of the IGBT works in a similar manner to the MOSFET. A gate-source voltage (V_{GS}, which must be greater than $V_{GS(th)}$) creates an inversion layer that shorts the n^- drift to n^+ source, as demonstrated in Figure 2.4. This allows electron flow though the inversion layer. This results in substantial hole injection from the p^+ drain contact layer into the n^- drift region. The injected holes move across the drift region and reach the p body region, subsequently attracting electrons from the source p body contact with which they recombine. The flow of holes and electrons is depicted in the bottom of Figure 2.5.
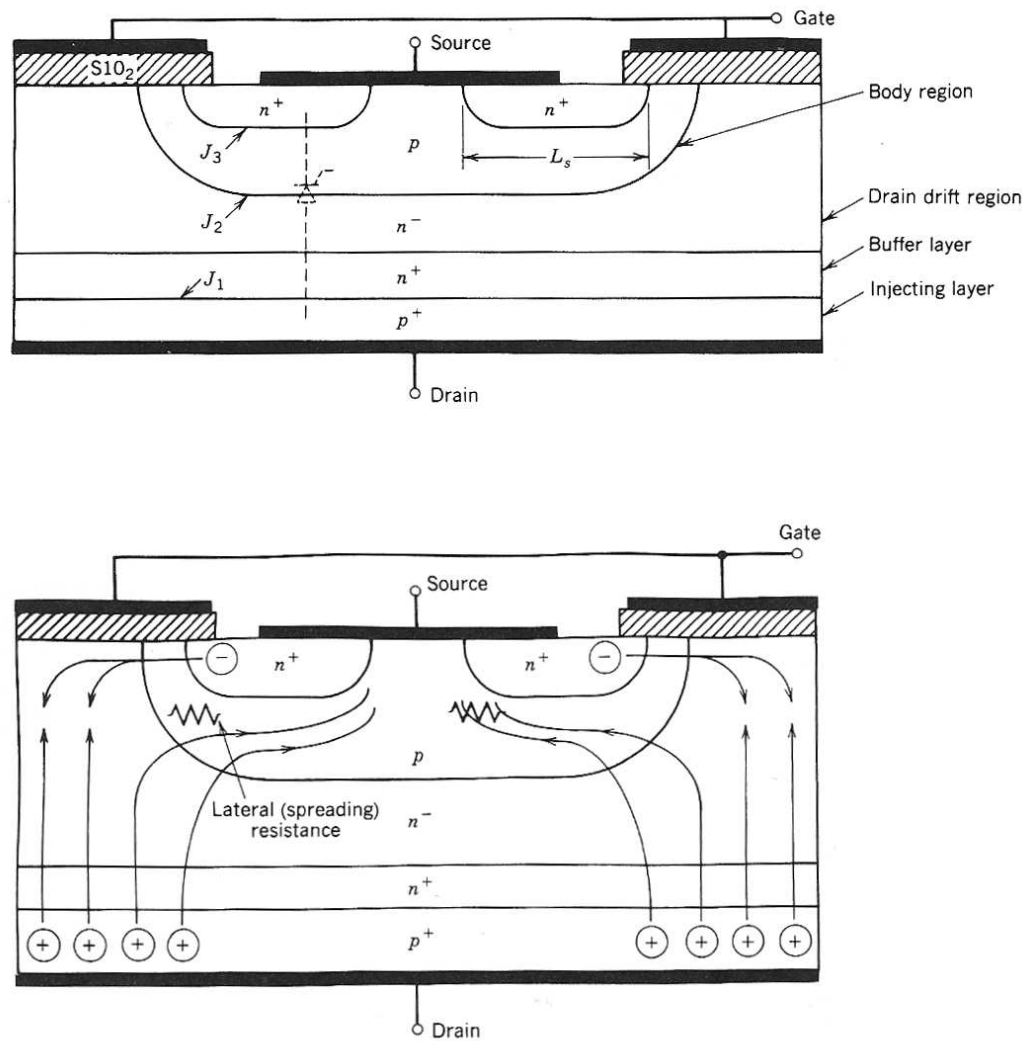
Figure 2.5: *Top*: A cross-sectional view of a IGBT cell, showing the doped regions and the three connectors: source, drain and gate. *Bottom*: The flow of electrons and holes during the IGBT on-state.[2]

2.3 Heat Generation in Power Electronic Devices

Heat generation in semiconductor devices is caused from the power losses within them, and can be calculated by using either of the following equations:

$$P = I^2 R \qquad or \qquad P = \frac{V^2}{R}$$

where I and V are the operating current and voltage, and R is the total resistance through the device. The location of the power loss in the device can be analysed if the local current density and resistances are known.

2.3.1 MOSFETs and IGBTs

The resistance components within the MOSFETs are shown in Figure 2.6. Nearly all power dissipated in a MOSFET occurs during the on-state, and therefore MOSFET design is heavily orientated towards reducing the on-state resistances within the device. The length of the current paths across the large resistance components are minimised, and regions are doped as heavily as the other requirements (such as breakdown voltage) allow. The 'channel' and 'accumulation layer' resistances are affected by the gate-source voltage; larger voltages create larger inversion layers and thus reduce the electrical resistance. The resistive components in Figure 2.6 contribute equally to the total on-state resistance for low breakdown voltages (a few hundred volts or less). At greater breakdown voltages, the drift region resistance is dominant.[2]

The resistances are also temperature dependent. As the device temperature increases, the resistance also increases. This leads to further power dissipation, and as a result, further increases in temperature. This *thermal runaway* causes large increases in device temperature which can lead to device failure.

In transient situations, the current demand through the device increases rapidly. The drain current (I_D) rises until the saturation region is reached, at which point it is said
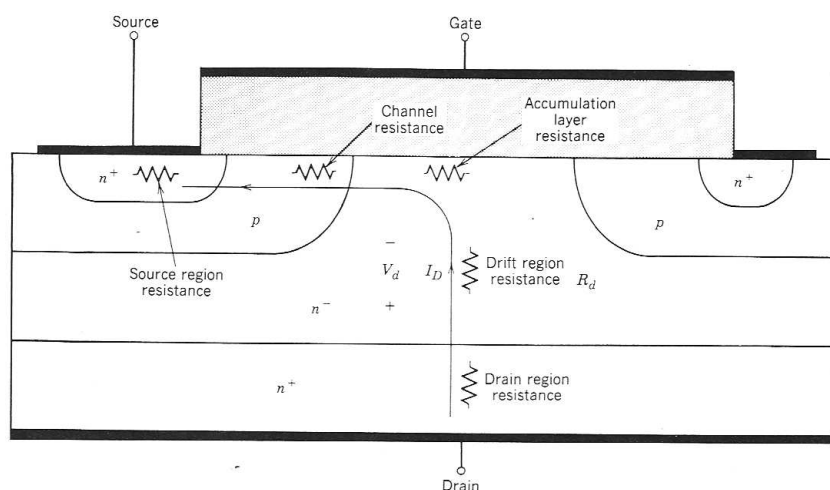
Figure 2.6: The on-state resistances within a MOSFET device.[2]
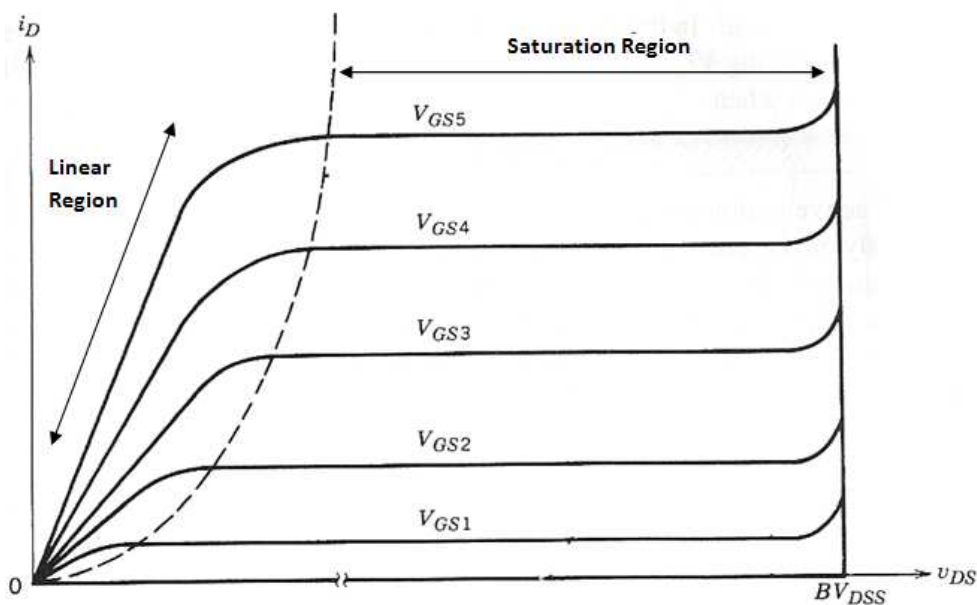
Figure 2.7: The linear and saturation regions of the MOSFET *I-V* curves.[2]

to have 'pinched off'[3]. At pinch off, the channel resistance is the largest resistor in the MOSFET. The power drop across the channel increases with the increase in current demand due to the rise in V_{DS}, resulting in significant heat generation in the silicon region directly beneath the gate contact. Therefore, in transient situations, the majority of the heat generation occurs in the top part of the MOSFET devices.

As IGBTs function in a similar manner to MOSFETs, the heat generation mechanisms

are also very similar. Trivedi and Shenai [8] demonstrated the region of heat genera-
tion in IGBTs during short-circuit (transient) conditions is directly beneath the gate
connection.

Another loss mechanism in power MOSFETs is the flow of charge to and from the gate
capacitance, termed switching losses. In applications where MOSFETs and IGBTs are
switched at high frequencies, these switching losses become significant. However, this
thesis will not be considering these applications, and will therefore focus on device
losses during the on-state only.

2.4 Device Failure Mechanisms

The temperature of any semiconductor device must be kept under control in order to maintain functionality. The theoretical limit to the temperature is the temperature at which the intrinsic carrier density, n_i, becomes equal to the majority carrier doping density in the lightest doped region. At this temperature, termed the intrinsic temperature, T_i, the depletion region at the junction is shorted out by the intrinsic carriers, losing the rectifying characteristics. For power diodes, the lightly doped epitaxial layer has a dopant concentration corresponding to an intrinsic temperature of about 280ºC. The maximum specified temperature from manufacturers is often around 125ºC, which represents the maximum temperature at which the device can reliably perform to certain specifications, for example on-state conduction voltages, switching times and switching losses. There are numerous types of device failure that occur, many of which are related to the temperature and are induced prior to the intrinsic temperature.

Avalanche breakdown, as discussed previously, is a common form of device failure which is a result of impact ionisation. Devices that are heated up too much can also develop thermal runaway, due to the relationship between electrical resistance and temperature. Short-circuit failure mechanisms can be divided into four modes [9]. These failure types are defined by the time at which they occur (in relation to the short-circuit), the most common of which is failure during the short circuit, called 'energy limited failure', and is attributed to thermal runaway.

Devices can also fail due to stack components deteriorating over time. Device stacks are subjected to thermal cycling during operation which causes the stack components to expand and contract. Layers in the stack with different Coefficients of Thermal Expansion (CTEs) expand at different rates, leading to mechanical stresses and cracking. The voids created by the cracks increase the thermal resistance between the device and the coolant. As a result, the steady state temperature of the device increases which can subsequently cause the device to fail.

2.5 Heat Conduction Theory

The theory of heat transfer is well established. Heat transfer occurs due to the migration of energy between locations at different temperatures. Thermodynamic principles based on observations have been generalised into laws, and governing equations have been derived that can be applied to any situation. The thermodynamic theory can be used to deduce the heat conduction theory which characterises the flow of heat through solids. Conduction theory can be used to model the heat conduction through electronic devices, allowing the temperatures inside the devices to be calculated when heat pulses are simulated inside them.

2.5.1 Heat Conduction Equation

The general heat conduction equation is derived from the first laws of thermodynamics and Fourier's equation of heat conduction. This common derivation can be found in sources such as [10]. Fourier's equation is used to calculate the heat flux per unit area, q'', between two points distance dx apart with temperature difference $\mathrm{d}T$:

$$q'' = -k \frac{\mathrm{d}T}{\mathrm{d}x} \qquad (2.1)$$

Thermal conductivity, k, is a material property that describes how easily heat can travel through a solid. The rate of heat conduction between two points at different temperatures is proportional to the thermal conductivity of the conducting material. In other words, the resistance to the heat conduction is proportional to the inverse of the thermal conductivity:

$$Thermal\ Resistance \propto \frac{1}{k} \qquad (2.2)$$

Applying equation 2.1 to a solid and using the first law of thermodynamics produces the 3D conduction equation in partial differential form:

$$\frac{\partial}{\partial x}\left(k\frac{\partial T}{\partial x}\right) + \frac{\partial}{\partial y}\left(k\frac{\partial T}{\partial y}\right) + \frac{\partial}{\partial z}\left(k\frac{\partial T}{\partial z}\right) + \dot{q} = \rho\, c_p\, \frac{\partial T}{\partial t} \qquad (2.3)$$

where $\dot{q}$ is the volumetric internal heat generation with the units W/m³. This can be reduced to 1D, to simplify the discussion:

$$\frac{\partial}{\partial x}\left(k\frac{\partial T}{\partial x}\right) + \dot{q} = \rho\, c_p\, \frac{\partial T}{\partial t} \qquad (2.4)$$

However, the form of the expression used more commonly is:

$$k\,A\,\frac{\mathrm{d}T}{\mathrm{d}x} + q_{gen} = V\rho c_p \triangle T \qquad (2.5)$$

This can be applied to a solid bar through which heat conduction occurs in 1 direction, such as that shown in Figure 2.8. The last term in equation 2.5 contains the expression ρc_p, which is called the thermal mass as it determines how quickly a volume, V, responds to a change in thermal surroundings, such as a heat input.
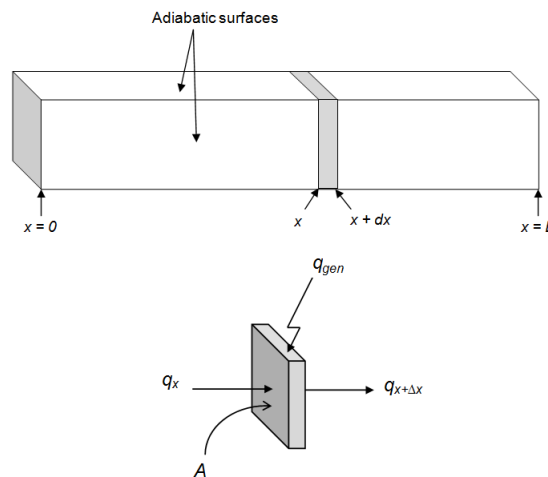
Figure 2.8: 1D heat conduction along a solid bar.

2.5.2   Thermal Diffusivity

The conduction equation can be rearranged to group three important variables toge-
ther: density, specific heat capacity and thermal conductivity.

$$\frac{\partial^2 T}{\partial x^2} + \frac{\dot{q}}{k} = \frac{1}{\alpha}\frac{\partial T}{\partial t} \tag{2.6}$$

In which the *thermal diffusivity* (α) is:

$$\alpha = \frac{k}{\rho\, c_p} \tag{2.7}$$

Thermal diffusivity is an important property for transient heat conduction, as it de-
termines how quickly a material can adjust to its surrounding thermal environment.

2.6 Transient Heat Conduction

When the boundary conditions of a solid system are known (i.e. the heat transfer conditions at the external surfaces and the locations and magnitudes of any internal heat generation) the steady state temperature distribution within the system can be calculated directly. Solution methods can be found in most heat transfer text books, such as Bejan [10], Kreith and Bohn [11]. A general steady state equation for solids, such as the slice shown in Figure 2.8, can be expressed:

$$q_{x+dx} - q_x + q_{gen} = 0 \qquad (2.8)$$

However, a different approach must be taken if the temperature history throughout the transient period is required.

2.6.1 Unidirectional Transient Heat Conduction

When a solid object undergoes a change from thermal equilibrium, the temperature distribution goes through three different regimes, known as the early regime, the transition stage and the late regime. Each one describes the temperature distribution and rate of heat transfer through the system.

To understand these regimes better, an analysis can be performed for unidirectional conduction by considering a solid bar of length *L* at an initial starting temperature of T_i throughout. One end of the bar is suddenly raised to a higher temperature, T_0, and is held constant at the new temperature. The other surfaces are considered adiabatic.

Once the temperature change is applied, heat flow occurs through the bar due to the imposed temperature gradient. The heat flow continues until thermal equilibrium is achieved and all parts of the bar have increased to the new temperature, T_0. During this transient phase, the heat conduction can be categorised into three regimes.

Figure 2.9 shows the temperature profile along the bar at the initial state and during the different regimes. When the temperature has not risen at the far end of the bar, the early regime describes the heat flow. Once the temperature gradients in the bar have decayed and the temperature in the bar can be approximated by a single value, T_0, the heat flow is in the late regime. The transition period is used for the time in between these two regimes.
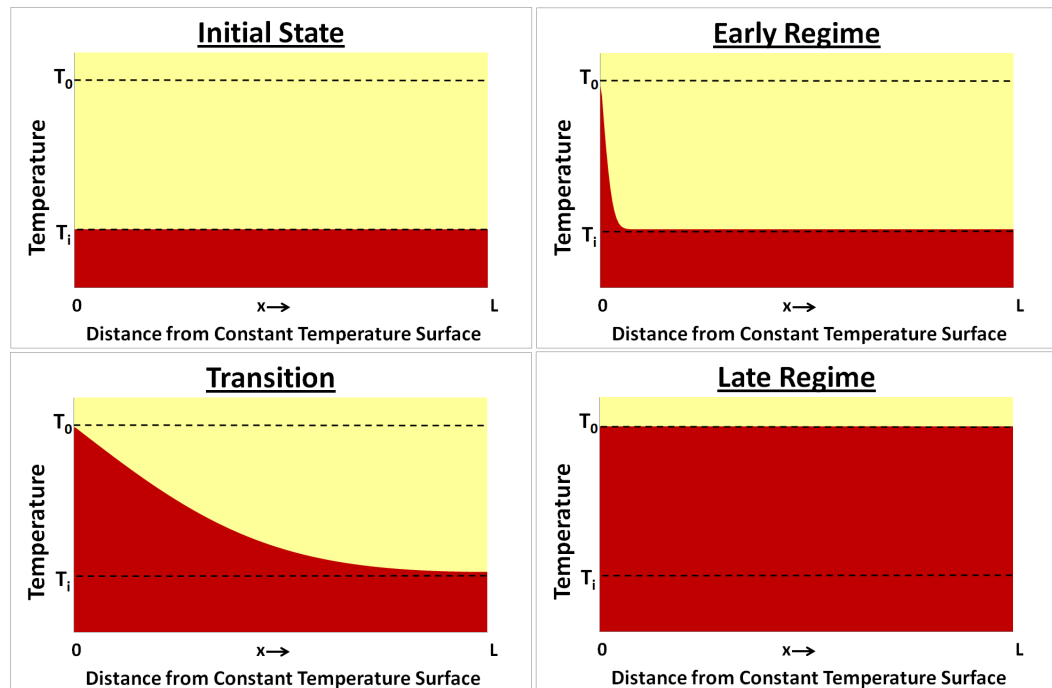
Figure 2.9: Typical temperature profiles along the bar for the three regimes which categorises the conduction heat flow. The bar is at an initial temperature , T_i, before the left surface of the bar is suddenly increased to T_0.

2.6.1.1 The Early Regime

The early regime is of particular importance for the modelling work conducted in this thesis. The durations of the heat pulses are relatively small from a thermodynamic point of view: most of the heat conduction during this time period will behave within the early regime as there is insufficient time to move to the late regime.

As soon as the temperature at the end of the bar is raised, heat flow in the bar occurs due to the created temperature gradient. Whilst the distance that the heat

has penetrated is less than the length of the bar, the temperature curve follows the same general shape. At the constant temperature boundary $(x = 0)$, the temperature is fixed at T_0. The far end of the bar at $x = L$, has a temperature equal to T_i.

The skin layer is a term used to describe the distance that heat has travelled at a certain time during the early regime. This is the distance between $x = 0$ and the point along the bar where the temperature reduces to T_i. The thickness of the skin layer is dependent on the time elapsed and the thermal properties of the bar.
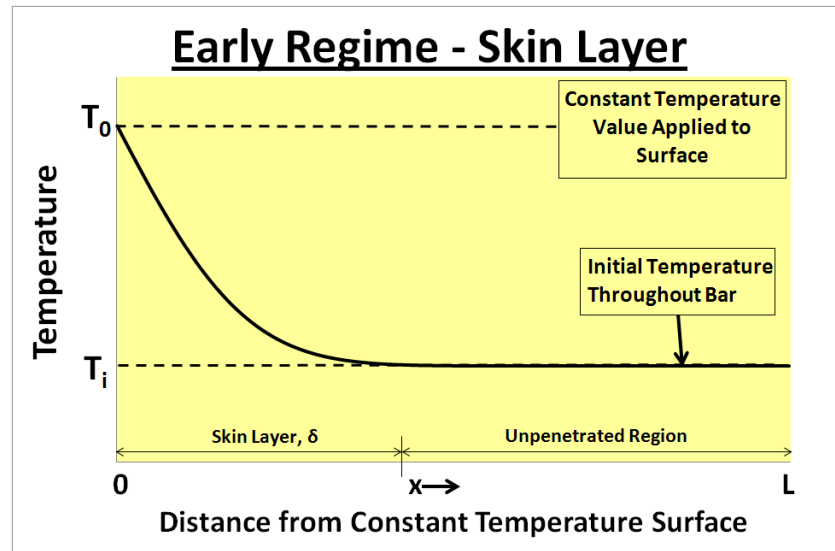
Figure 2.10: The shape of the temperature through the bar with a constant heat surface during the early regime.

An approximation of the depth of the skin layer, δ, can be deduced from the conduction equation, eq.(2.6). As there is no internal heat generation, the equation reduces to:

$$\frac{\partial^2 T}{\partial x^2} = \frac{1}{\alpha} \frac{\partial T}{\partial t} \tag{2.9}$$

Assuming the curvature of the temperature profile over the skin layer, $\frac{\partial^2 T}{\partial x^2}$, is constant, the gradient of the temperature curve can be approximated at two points, $x = 0$ and $x \approx \delta$:

$$\left(\frac{\partial T}{\partial x}\right)_{x \approx \delta} \approx 0 \tag{2.10}$$

$$\left(\frac{\partial T}{\partial x}\right)_{x=0} \approx \frac{T_i - T_0}{\delta} \tag{2.11}$$

At the edge of the skin thickness, $x \approx \delta$, the temperature approaches T_i at which point the temperature gradient becomes zero. Therefore the gradient at the edge of the skin layer is assumed to have a gradient of zero. At $x = 0$, where the constant temperature condition is applied, the gradient is assumed to be equal to the change in temperature over the entire skin layer. Using these approximations, the curvature of the temperature across the skin layer can be written:

$$\frac{\partial^2 T}{\partial x^2} \approx \frac{\left(\frac{\partial T}{\partial x}\right)_{x \approx \delta} - \left(\frac{\partial T}{\partial x}\right)_{x=0}}{\delta - 0} \tag{2.12}$$

$$\frac{\partial^2 T}{\partial x^2} \approx -\frac{T_i - T_0}{\delta^2} \tag{2.13}$$

The right hand side of eq.(2.9) relates the thermal properties of the material to the change in temperature over time. This change, $\frac{\partial T}{\partial t}$, can be approximated during the early regime. The average temperature within the skin layer changes from T_i at $t = 0$ to a value approaching T_0 after a time t.

$$\frac{\delta T}{\delta t} \approx \frac{T_0 - T_i}{t - 0} \tag{2.14}$$

Combining eq.(2.14) with eq.(2.13), a form of the conduction equation is produced using the assumptions made about the temperature profile during the early regime:

$$-\frac{T_i - T_0}{\delta^2} \approx \frac{1}{\alpha}\frac{T_0 - T_i}{t - 0} \tag{2.15}$$

From this equation it is possible to arrive at a formula to approximate the thickness of the skin layer at any time during the early regime.

$$\delta \approx \sqrt{(\alpha t)} \tag{2.16}$$

The early regime is assumed while the heat is still penetrating through the solid, i.e. while the skin thickness is less than the length of the solid thickness. The transition period occurs when the heat has reached the end of the bar.

Once the skin layer extends to the entire length of the bar, the assumptions in eqs.(2.10) and (2.11) can no longer be made. The time at which this occurs is the time at which the transition region begins, t_T. It can be calculated as follows:

$$L \approx \sqrt{(\alpha t_T)}$$

$$t_T \approx \frac{L^2}{\alpha}$$

This transitional time leads to the definition of the Fourier number, Fo, which is an important number used in transient heat conduction. The Fourier number is a dimensionless time that describes the relationship between the distance heat travels in a certain time. A Fourier number of 1 implies that within the time t, the skin layer has fully penetrated a distance L through a solid of thermal diffusivity α:

$$\frac{\alpha\, t}{L^2} \approx 1$$

$$Fo = \frac{\alpha \, t}{L^2} \tag{2.17}$$

The Fourier number can be used to define the times at which the three regimes are dominant during a period of transient conduction. For Fourier numbers less than 1, the heat conduction follows the early regime and Fourier numbers much greater than 1 indicates the late regime is being followed.

Early Regime:

$$Fo \ll 1$$

Late Regime:

$$Fo \gg 1$$

2.6.2 Fourier Number

A Fourier number of 1 implies that heat has 'fully penetrated' a distance L during a time t. This can be better defined by quantifying the expression 'fully penetrated'.

Consider the adiabatic bar at an initial constant temperature T_i, having one end suddenly heated to a constant temperature of T_0. The induced temperature gradient causes heat flow along the bar. This heat flow continues until a new steady state equilibrium is reached, when all parts of the bar are at temperature T_0. The temperature at the far end of the bar can be calculated during this transient phase and expressed as a normalised temperature rise, with respect to the final temperature rise $(T_0 - T_i)$. When plotted against the Fourier number a plot of the normalised temperature rise is produced, as shown in Figure 2.11. It can be seen that a Fourier number of 1 represents that the temperature at the adiabatic end of the bar has increased to 89% of the final temperature rise.
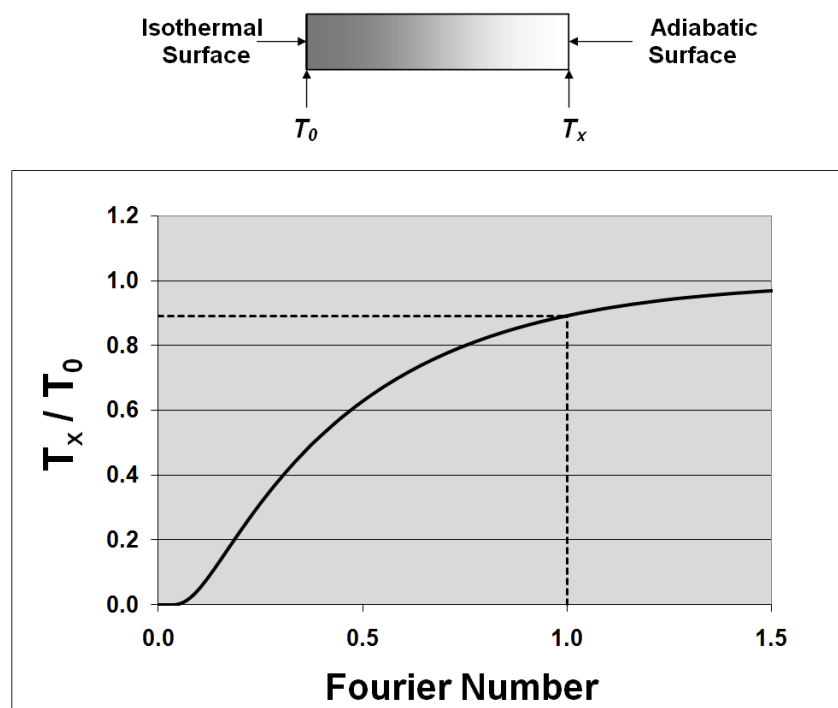
Figure 2.11: Ratio of the temperature at either end of the bar in relation to Fourier number.

2.6.2.1 Significance when Fourier Number Equals 1

The expression 'fully penetrated', which refers to a Fourier number of 1, has been shown to mean that the temperature rise is 89% of the final temperature. This is an important relationship, and will be used later on in the thesis to determine a transient time-constant.

Heisler charts are graphical analysis tools that show the temperature history at the mid-plane of a solid immersed in a fluid. These can also be used to find this correlation. A plot showing the temperature history in the mid-plane of a plate immersed suddenly in a fluid of a different temperature can be found in [10]. The plot lines show the temperature history for different values of $1/Bi$, which is equal to k/hL (Bi refers to the Biot number, see Glossary). The scenario in Figure 2.11 represents a sudden increase in temperature, which is representative of an infinite heat transfer coefficient, therefore $k/hL = 0$. This can be seen in the Heisler Chart in Figure 2.12.

The relative temperature rise for a Fourier number of 1 on the line $1/Bi = 0$ is equal to approximately 0.11, representing a temperature rise at the mid-plane of 89% of the final temperature rise (when the temperature is equal to the fluid temperature).
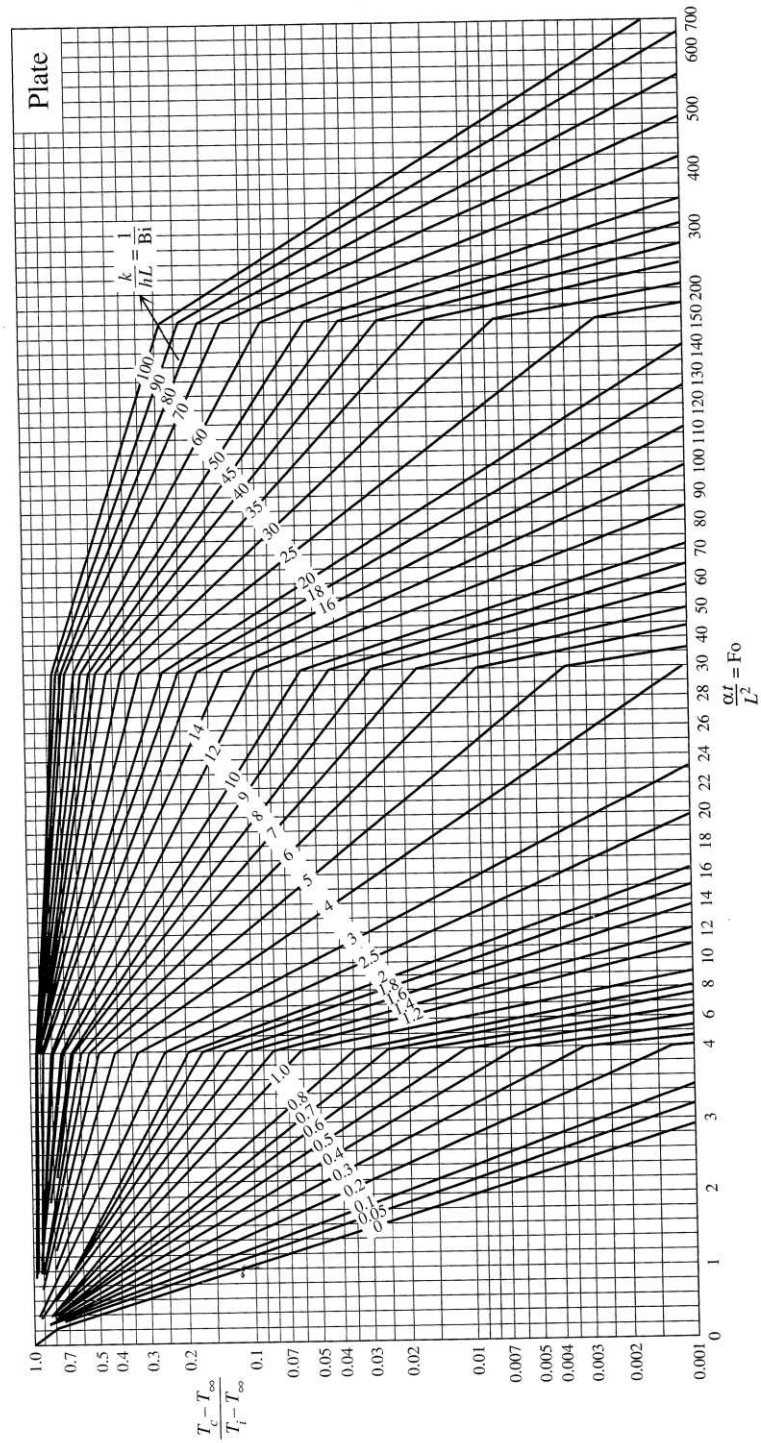
Figure 2.12: A Heisler Chart, representing the temperature rise at the mid-plane of a plate due to immersion of the plate in a fluid.[10]

2.6.2.2   Fourier Number Application to Constant Heat Generation

To describe the scenario of power dissipation in an electronic device more accurately, the isothermal end of the bar is replaced with a constant heat flux surface. In this case, the adiabatic end must be replaced by an isothermal surface (held at T_i) in order to achieve a steady state solution.

Heat is generated at the heat flux surface and is conducted down the bar. The temperature rise at the heat flux surface can be calculated during the transient phase, and normalised with respect to the final steady state temperature rise. This is plotted against Fourier number, and shown in Figure 2.13.

The temperature at the constant heat flux surface reaches 89% of the steady state value when the Fourier number equals $\pi/4$. Thus, when a heat flux surface replaces the raised temperature, the response time of the bar changes.

A further example compares the temperature rise at the constant heat flux surface for two different bars: one with an isothermal end and one that is modelled as semi-infinite. These bars are shown above the plot on right hand side of Figure 2.13.

The temperatures at the heat flux surface will be identical during the early stages of the heat flow. They begin to diverge once the isothermal surface become influential on the temperature at the heat flux surface. A plot of the ratio of the temperatures at the heat flux surfaces against Fourier number at the bottom of Figure 2.13 demonstrates the divergence. The temperature at the heat flux surface with the isothermal end reduces to 89% of that at the semi-infinite heat flux surface after a Fourier number of $\pi/4$.
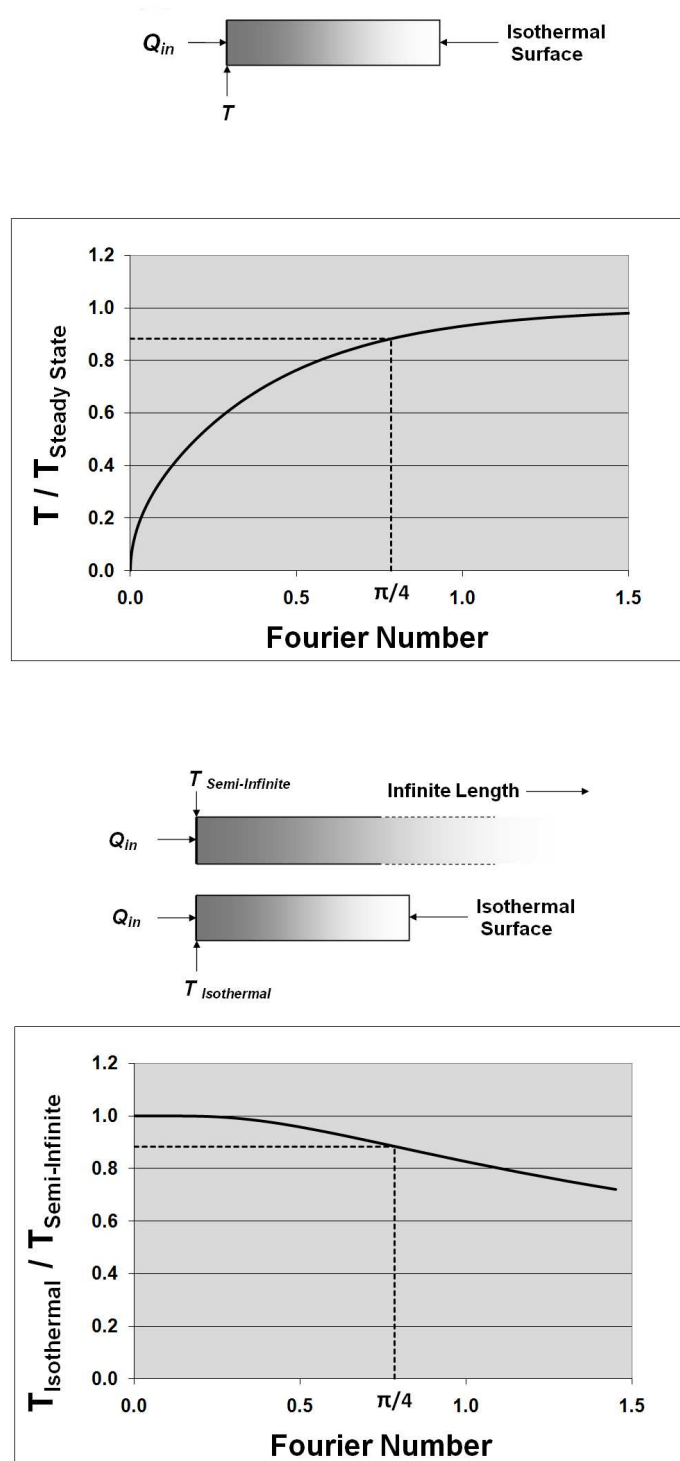
Figure 2.13: *Top:* Temperature rise at the free end of the bar in relation to Fourier number. *Bottom:* Ratio of temperatures at the free end of the bar with different bar length conditions, in relation to Fourier number.

2.6.2.3   Defining the Transient Time-Constant

The Fourier number of $\pi/4$ is very significant to the thesis, and it can be used to formulate an expression for the transient time-constant.

When a heat flux is applied to a surface a Fourier number of $\pi/4$ demonstrates that the heat has 'fully penetrated' a distance L in time t. The Fourier number equation (eq.(2.17) can be rearranged in terms of time. A time-constant (τ) is defined, which is the time taken for the temperature at the constant heat flux surface to be affected by the geometry at a distance L from the surface:

$$\tau = \frac{\pi}{4} \frac{L^2}{\alpha} \tag{2.18}$$

This time constant is used to calculate the time it takes for the heat to fully penetrate a layer of thickness L. This is very useful as it is possible to analytically calculate the time during a heat pulse that each layer in a device stack-up becomes influential on the device temperature.

2.7   Chapter Conclusion

This chapter has provided details of power electronic devices and heat generation mechanisms inside them. The processes of avalanche breakdown and impact ionisation have been described, and their role in device failure explained. In these cases, device failure occurs quickly due to a sudden surge of current through the device.

An overview of the operation of power diodes, MOSFETs and IGBTs has been presented, each of which are relevant to the thesis. It is shown how it is possible to calculate the amount of heat generated in the device at any time. Local electrical resistances and the current flow density within the device dictate where the heat generation occurs during the on-state.

Device failure mechanisms have been identified, including when the temperature exceeds the intrinsic temperature, avalanche breakdown and impact ionisation. These failures are all quick, owing to a sudden increase in current, and temperature. Failures owing to gradual degradation in the stack-up over time have also been discussed, with variations in coefficient of thermal expansion inside the stack being a significant root cause.

The second half of the chapter has introduced some important heat transfer theory. Material properties have been identified using the heat conduction equation, including thermal conductivity and thermal diffusivity.

The three transient heat conduction regimes, early, transitional and late, have been defined. Analysis to the early regime allows the definition of the Fourier number, a highly significant dimensionless time for heat conduction. The Fourier number allows the early and late regime to be defined quantitatively, the transition between the two occurring at a Fourier number of 1.

Further Fourier number analysis provided an equation for a time-constant. This time-constant is used to calculate the time it takes for the heat to fully penetrate a layer of specified thickness in constant heat flux applications. This is very useful as it is

possible to analytically calculate the time during a heat pulse that each layer in a device stack-up becomes influential on the device temperature. The time-constant as presented in the chapter, has not been found in any existing literature.

Chapter 3

Literature Overview

This chapter presents a literature review, summarising the state-of-the-art of relevant research and technical areas. The overview begins by discussing cooling technologies used in the semiconductor industry. Different technologies are considered as candidates for reducing the transient temperature in semiconductor dies. Current double-sided device cooling techniques are then examined, with examples given of how it is currently implemented. This is followed by an overview of semiconductor cooling techniques specifically aimed at reducing transient temperatures.

The subsequent section of the literature overview examines different modelling methods for predicting transient semiconductor device temperatures. This includes both analytical and thermal resistance network methods. Different modelling techniques are discussed, together with some published results.

Semiconductor reliability will then be considered. Reliability is shown to be an important issue for semiconductor devices, as the thermal cycling experienced by the stack components induces thermal stresses in the different stack components. This follows on to a discussion of novel composite materials and alternatives to solder for die attachment. The motivation for this examination stems from the desire to optimise the effectiveness of the heat sink, which may involve the employment of novel materials with attractive thermal properties.

3.1   Cooling Technologies for Semiconductors

The importance of keeping electronics cool has been recognised for many years, resulting in a wide range of literature presenting different cooling techniques. The state-of-the-art of electronics cooling has been examined in order to identify any techniques which may be relevant to the problem under investigation.

Various methods have been implemented for cooling electronic devices. Anandan and Ramalingam [1] present an overview of these different methods and technologies. Many things must be considered when matching a cooling technology to a specific scenario, such as cost, the heat transfer rate required and size constraints. Scott [12] classified all of the methods into four categories of different heat transfer effectiveness. These are demonstrated in Figure 3.1 which shows the rate of heat transfer between two surfaces with a temperature difference of 80°C. It can be seen that high heat transfer rates are only achievable through forced liquid convection (e.g. water pumped, jet impinged, sprayed) or liquid evaporation (e.g. heat pipes).
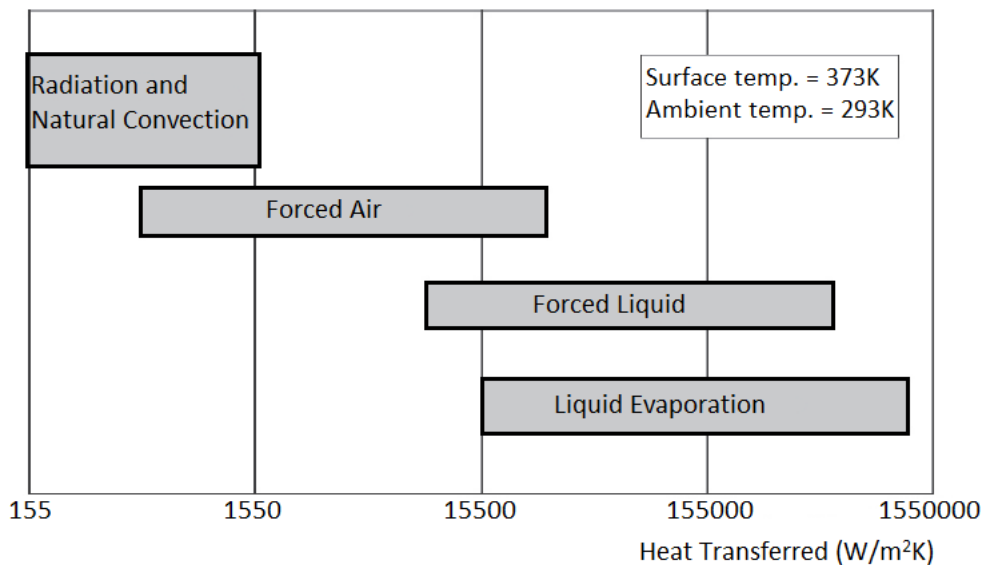
Figure 3.1: Range of conventional heat transfer modes.[1]

3.1.1 Forced Liquid Convection Techniques

Cooling electronic devices by spraying them directly with liquid is discussed by Anandan and Ramalingam [1], who identifies advantages of such a method as:

- the thermal resistance in the bonding layer between heat source and heat spreader is eliminated

- the ratio between power spent and heat removed decreases faster for spray cooling than channel cooling

However, it is also pointed out that the choice of liquid is important; whether it is non-conducting or a di-electric liquid. A thin protective layer is coated on components to protect against electrical shorts.

Cader et al. [13] conduct a test to quantify the ability of spray cooling to handle transient power dissipation at the die. Spray cooling the die is reported to provide a shorter transient time between steady state temperatures. At device switch-on and switch-off, the junction temperature (seen in Figure 3.2) behaves 'like a step function', whereas a lidded device (a completely sealed package with an inert gas injected which surrounds the device) 'follows more of an exponential decay when the power is turned off, and an exponential ramp when the power is turned on'. The unlidded data is not presented in the paper, and is only compared qualitatively.

The resolution of the time axis in Figure 3.2 is relatively poor; only three data points define the transient curve. It should also be noted that data from the lidded device is not presented for comparison. It is therefore difficult to assess the benefits of the spray cooling method.

A further paper by Cader and Tilton [14] gives examples of how spray cooling has been implemented successfully in high heat flux applications.

Garimella [15] states that enhanced heat transfer rates obtained with impinging jets make them suitable for cooling electronic devices with very high heat dissipation rates.
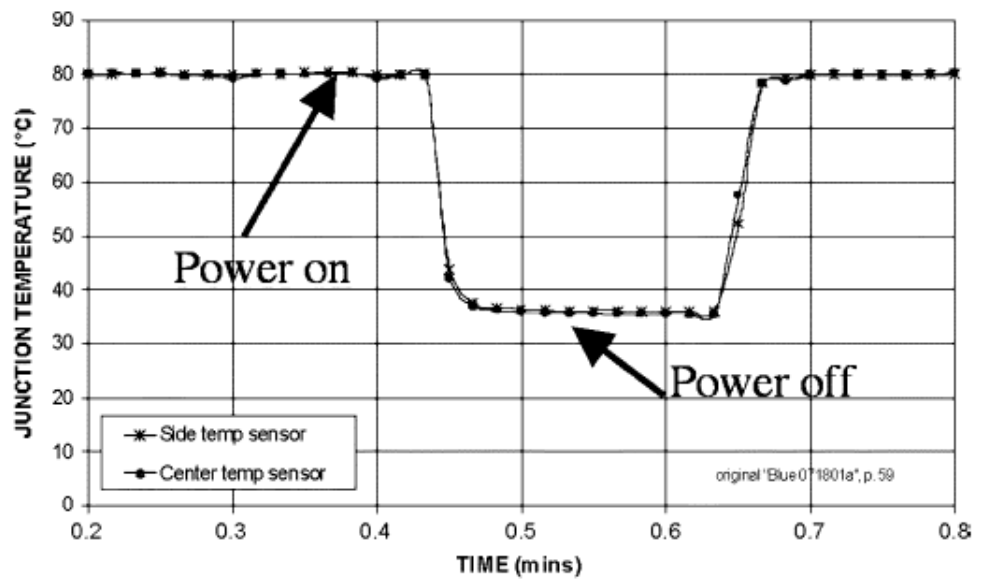
Figure 3.2: Junction temperature transient response to spray cooling of bare die. [13]

The effectiveness of a jet impingement system is dependent on the configuration of the jet setup, demonstrated in Figure 3.3. The heat transfer is increased by submerging the jet [16]. However, the heat transfer rate is reduced when the jet outflow is confined, as it is in many electronics cooling applications [17].

Garimella [18] presents a review of confined jets. Jet geometry configurations have been investigated with the aim of maximising the heat transfer at the impinged surface. Garimella [19] varies the ratio of orifice diameter to height of jet from the surface, and presents the variation of heat transfer in the local area of the jet (Figure 3.4). The results show the significance of the jet configuration on the effectiveness of the jet.

Further enhancements to the heat transfer rate have been achieved through surface enhancements. Finned surfaces have decreased the thermal resistance of liquid jet impingement by approximately 50%. Further roughening of the spreader plate decreased the thermal resistance by as much as 80% [20].

Garimella [15] also discusses the suitability of microchannel cooling for power electronics due to the high transfer coefficients achievable and their compact size. It is
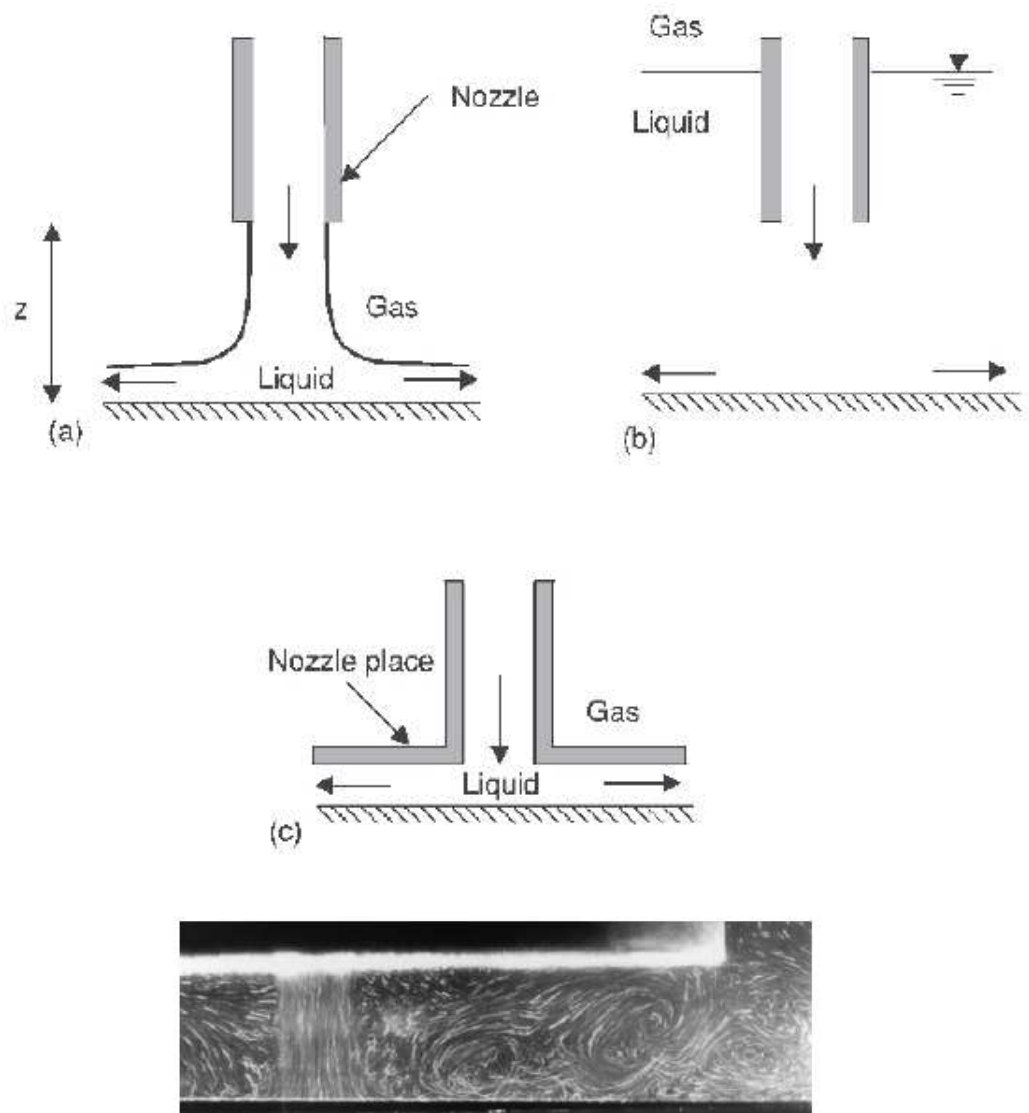
Figure 3.3: *Top*: Jet impingement configurations *a)* free-surface jet, *b)* submerged-jet, *c)* confined submerged-jet. *Bottom*: Flow field visualisation in a confined and submerged liquid jet.[18]

claimed that the large pumping requirements are the main reason they are not widely used on a commercial scale. According to Wilson and Simons [21], in many practical cases the small flow rate within micro-channels produces laminar flow resulting in a heat transfer coefficient inversely proportional to the hydraulic diameter. Thus, smaller channels achieve greater heat transfer coefficients. However, the pressure drop in the channels increases with the inverse of the second power of the channel width.
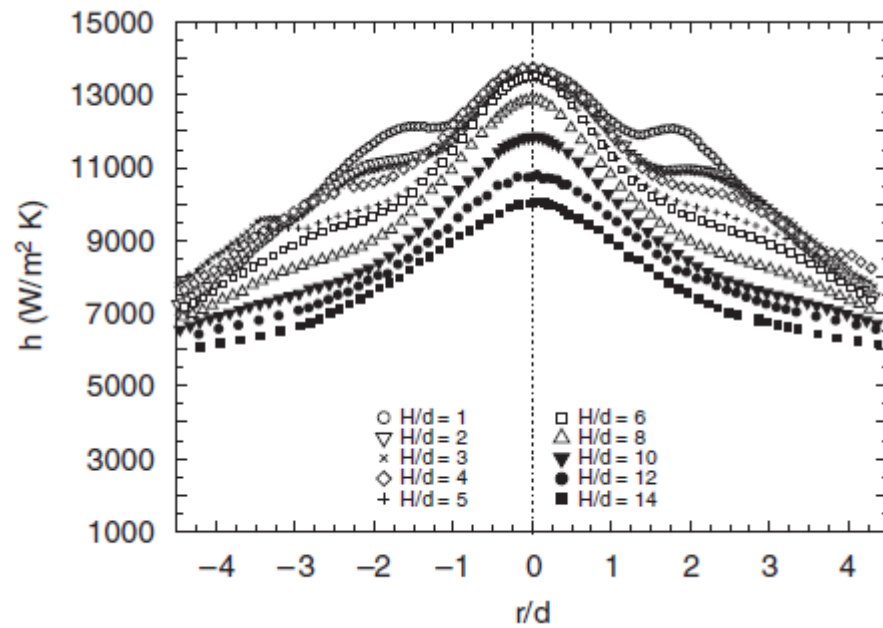
Figure 3.4: Variation in local heat transfer coefficient with jet to surface distance (d) for *Re*=13000 (based on orifice diameter, *d*) where *d*=1.59mm.

Jet impingement and microchannel cooling for high heat flux applications were compared by Lee and Vafai [22] who concluded that microchannel cooling is more effective for areas smaller than 7 x 7 cm [21].
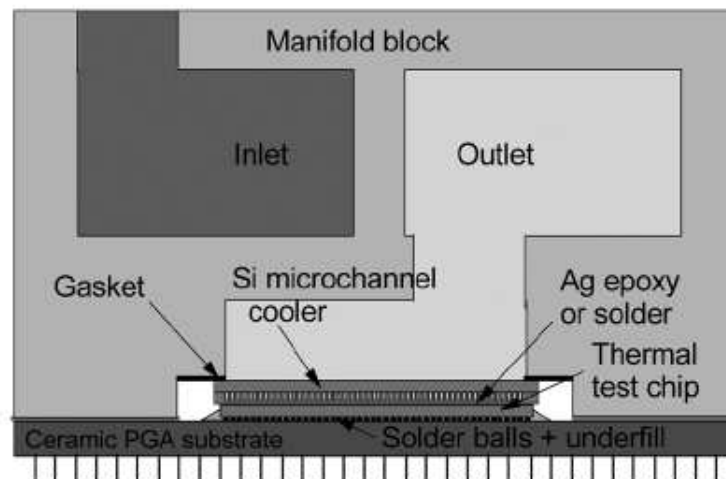
Figure 3.5: Schematic cross-section of the microchannel cooler integrated onto a high-power chip.[23]

Colgan et al. [23] published a practical implementation of a silicon microchannel cooler for high-power chips. Due to practical and cost reasons, a separate microchannel cold plate was bonded to the back of the chip rather than forming the microchannels directly on the chip (Figure 3.5). Power densities in excess of $400W/cm^2$ are reported, for a flow of 1.2l/min at 30kPa.

A further microchannel design is presented by Solovitz et al. [24]. A microchannel was built into the substrate of a high power device as shown in Figure 3.6. Sixty-five microchannels were manufactured, with a width of 100µm and w height of 300µm. The results, also presented in Figure 3.6, demonstrate a thermal performance 'superior to any existing micro-channel heat sink with a comparable electrical assembly'.

Lorenzen et al. [25] demonstrate the use of chemical-vapour-deposited (CVD) diamond as a heat spreader material to enhance a microchannel cooler. The high thermal conductivity of CVD diamond reduces the thermal resistance between the heat source and microchannel, increasing heat diffusivity.

A study by Bergles et al. [26] reports that two-phase microchannels require 20 times less pumping than single-phase liquid microchannels to achieve the same heat sink thermal resistance. However, fluctuations have been witnessed in local heat transfer coefficients when boiling heat transfer in microchannels has been tested.
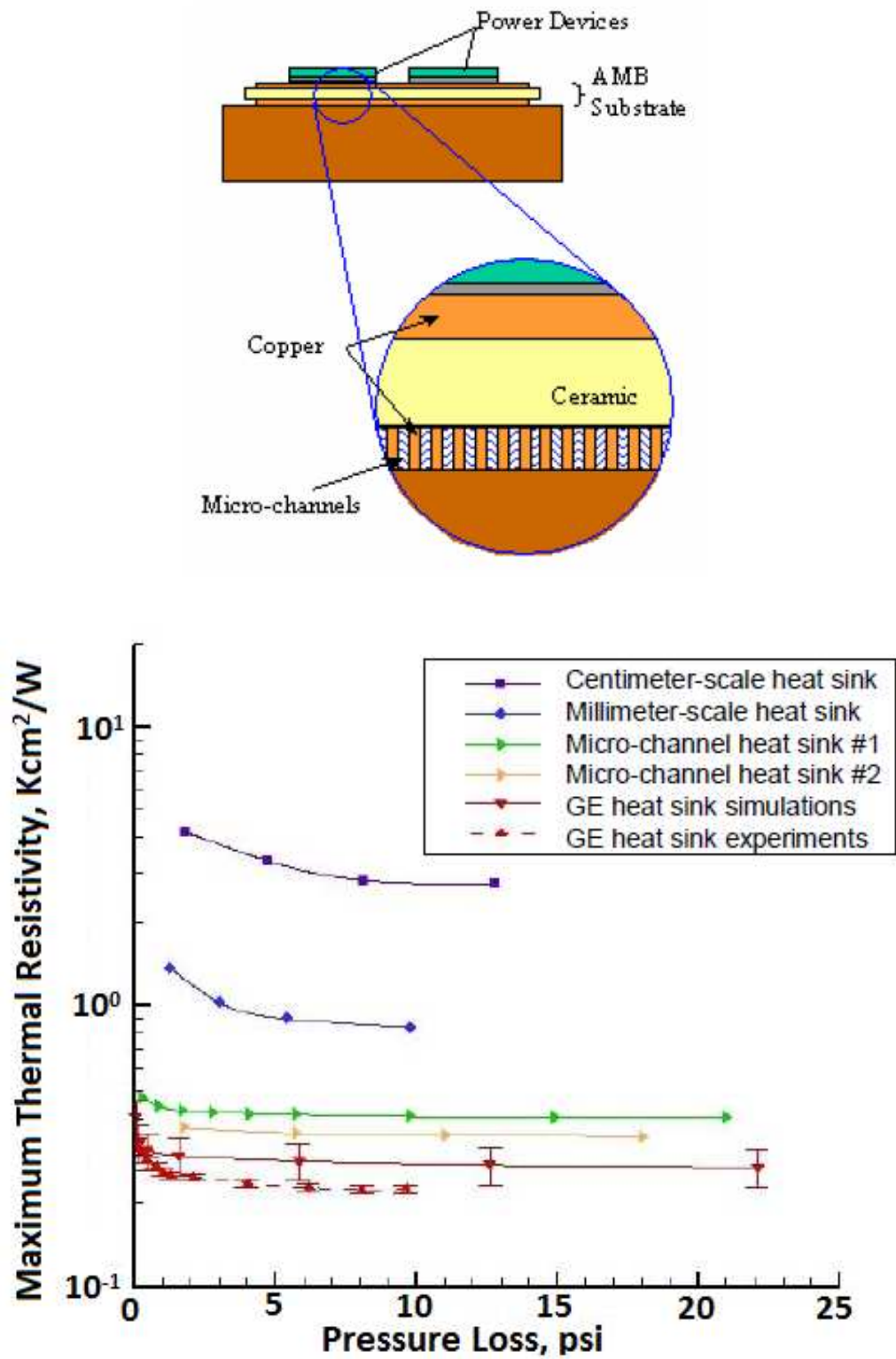
Figure 3.6: *Top:* Conceptual micro-channel heat sink design. *Bottom:*Thermal resistivity for micro-channel heat sink at various diode power dissipation and coolant flow rates. Note units of Kcm^2/W for thermal resistivity (the thermal resistance per square-centimeter of the device) which allows a direct comparison between different sized devices.[24]

3.1.2 Double-Sided Chip Cooling

A range of different double-sided chip cooling techniques have been investigated. Cooling both sides of the chip produces an additional thermal route for generated heat.

Herbsommer et al. [27] and Charboneau et al. [28] present different double sided cooling techniques. Herbsommer et al. [27] point out that most current packaging technologies have poor junction-to-top thermal resistances, making it difficult to have any significant heat flowing to the top of the device. The DUAL COOL^{TM} technology shown in Figure 3.7 creates a high thermal resistance by more than a factor of ten, compared to standard solutions.

The thermal resistance and power dissipated was measured and compared to different constructions: a) a wire bond package b) a clip package with a a 70% die coverage with a 10mm thick clip c) the same clip package but with dual cool technology. The results show a dramatic increase in the heat dissipated through the top of the package due to the decrease in the thermal resistance. As a result, the junction temperature is reduced during operation. The results presented are shown in Figure 3.8.
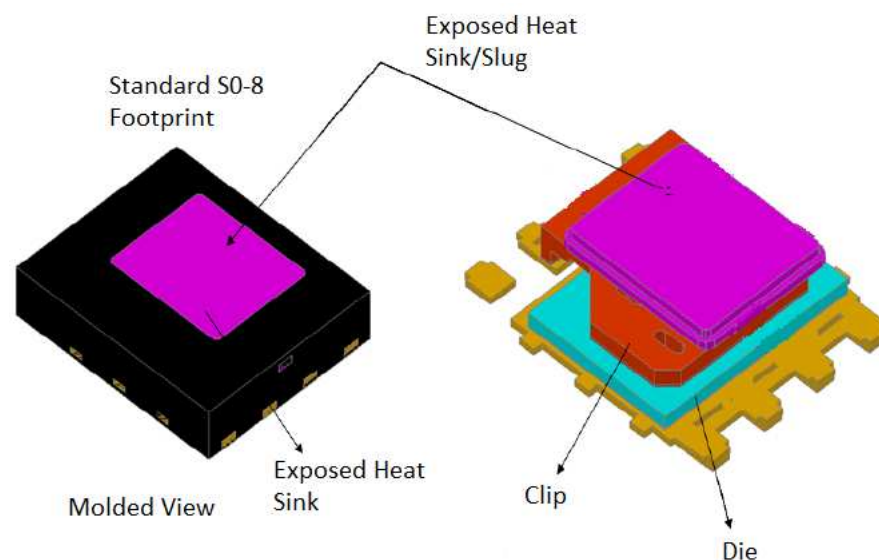
Figure 3.7: Construction of a dual cool device.[27]
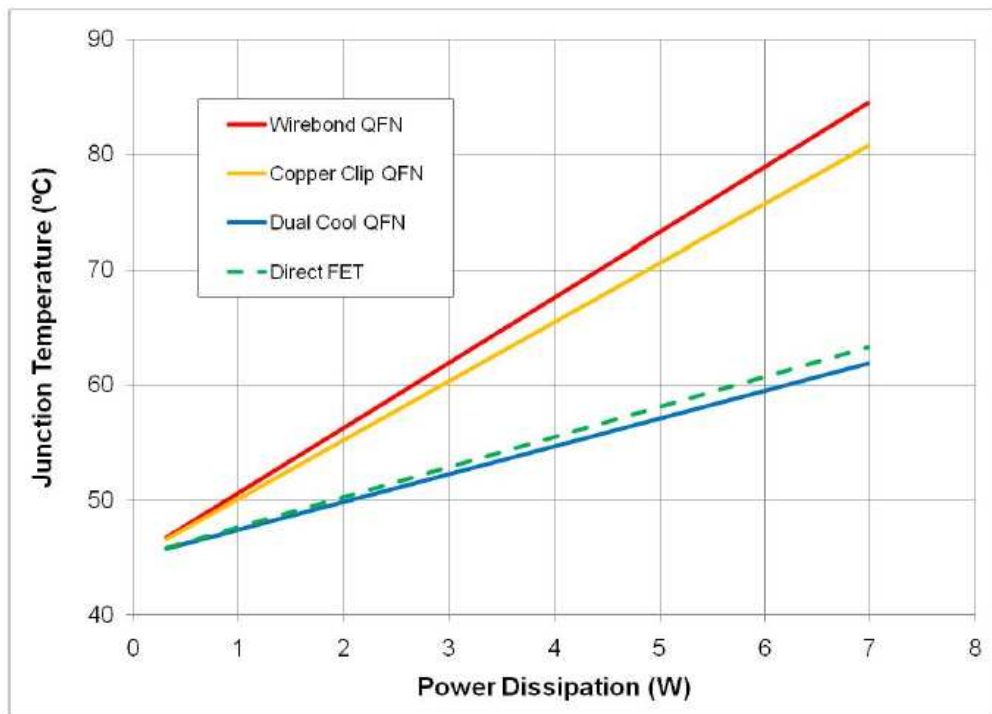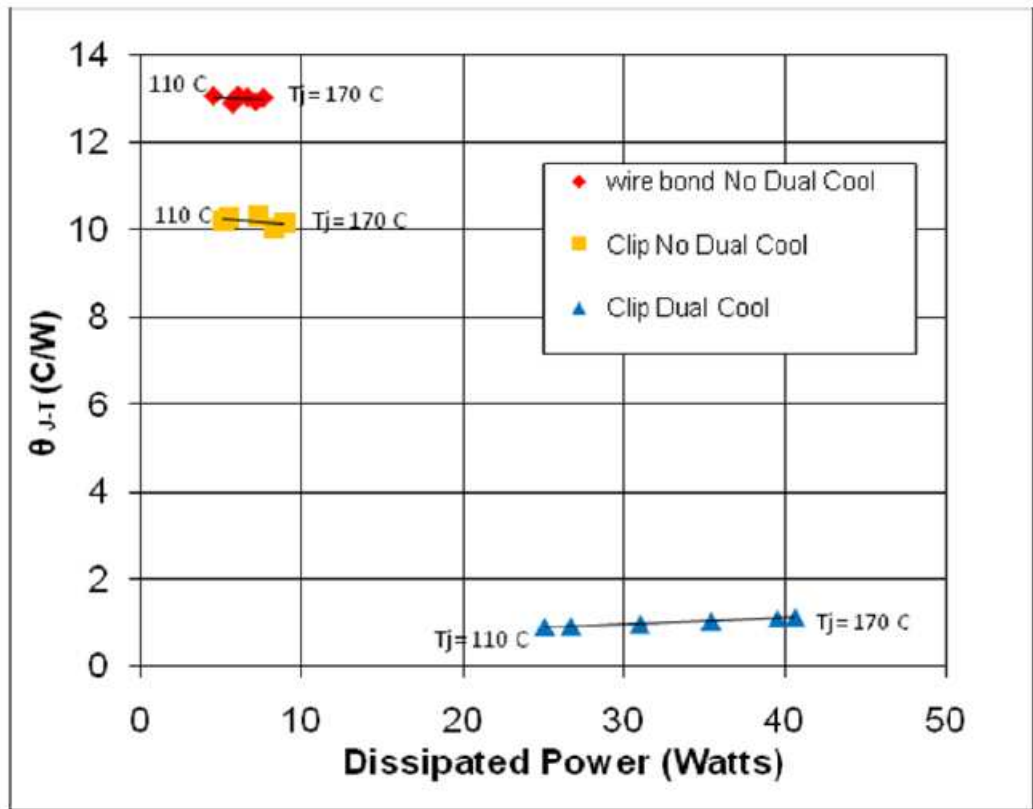
Figure 3.8: *Top:* Thermal resistance junction-top vs power. *Bottom:* Thermal performance comparison for different packaging technologies.[27]

Charboneau et al. [28] demonstrate double-sided liquid cooling on a power semiconductor device. The CoolMOS power module shown in Figure 3.9 is packaged so that both sides of the module can be cooled by parallel forced water convection, as indicated by arrows.

Modelling results were obtained for the packages cooled on a single side and on both sides, as well as experimental results for the package cooled on one side. The results obtained during device switch-on is shown in Figure 3.10. Comparing the modelling curves, it can be seen that the double sided cooling method suppresses the device temperature rise during the transient phase. The steady-state temperature achieved for the double-sided cooled device is less than that of the device cooled on a single-side.
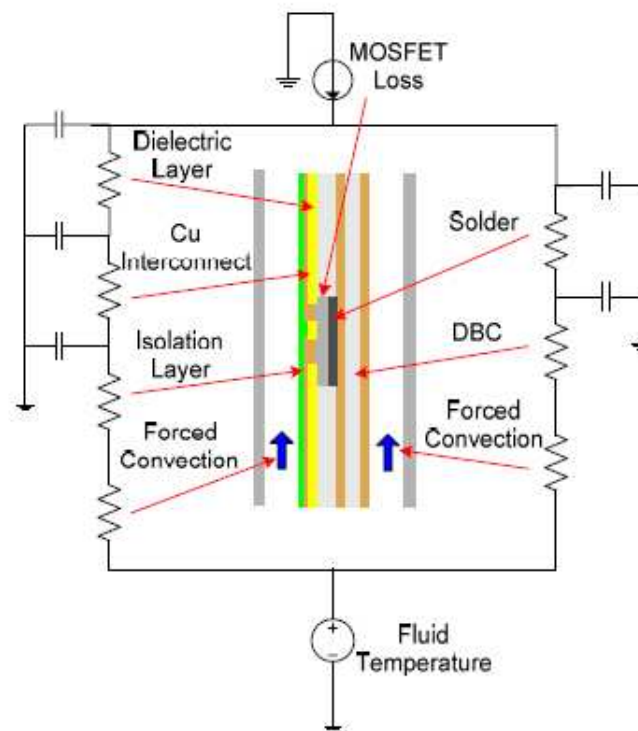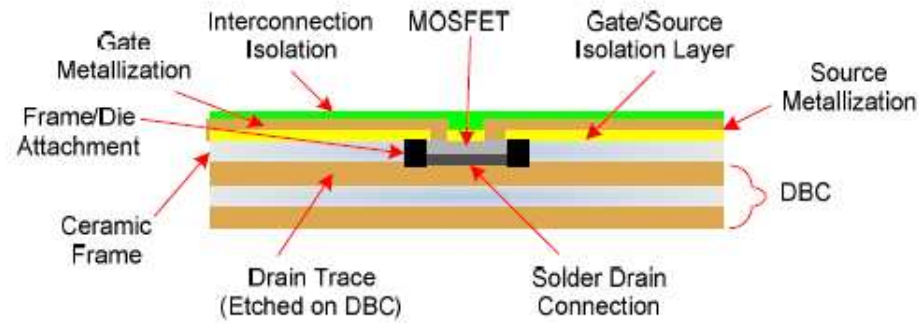
Figure 3.9: *Top:* MOSFET based embedded power structure. *Bottom:* Electro-thermal model for the structure with double-sided liquid cooling. The top MOSFET surface is protected by the Interconnection Isolation, which allows fluid flow over the top surface, as indicated int he bottom diagram. This provides an additional thermal path for heat flow to ambient. [28]
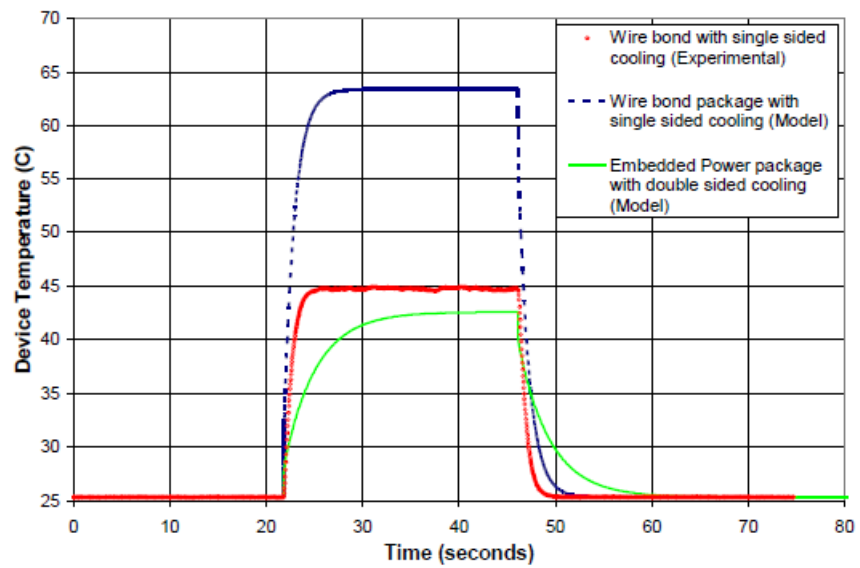
Figure 3.10: The thermal response of the device during switch-on for different cooling techniques.[28]

3.1.3 Thermal Transients

Despite the large volume of literature and research surrounding the electronics cooling industry, relatively little focuses on the thermal management of transients. Work done by Cader et al. [13] and Charboneau et al. [28] consider the transient response of the devices, however on a relatively large time-scale. Thermal designs for shorter transient, in the order of magnitude of milli-seconds rather than seconds, are not as widely considered.

Ngo et al. [29] investigate the geometry of the heat spreader and heat transfer coefficient of the heat exchanger on both steady-state and transient temperature. The geometry examined is shown in Figure 3.11. They state that during a thermal transient, the heat spreader works as a thermal capacitance to store thermal energy, which reduces the peak temperature in the module. However, a large heat spreader does not necessarily result in a better transient thermal performance.
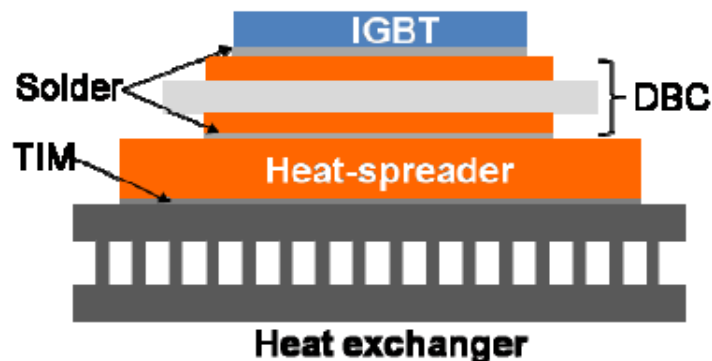
Figure 3.11: Structure for conventional power module with heat exchanger.[29]

Results are obtained from thermal modelling using a finite element analysis package. The thermal impedance is calculated during a transient when different heat spreader thicknesses are used, and also when different heat transfer coefficients are modelled in the heat exchanger.

Figure 3.12 shows that for transients less than 50ms, the heat transfer coefficient does not affect the thermal impedance. Beyond 50ms, a large change to the heat transfer

coefficient has only a limited effect on the thermal impedance. This is similar for the thickness of the heat spreader.
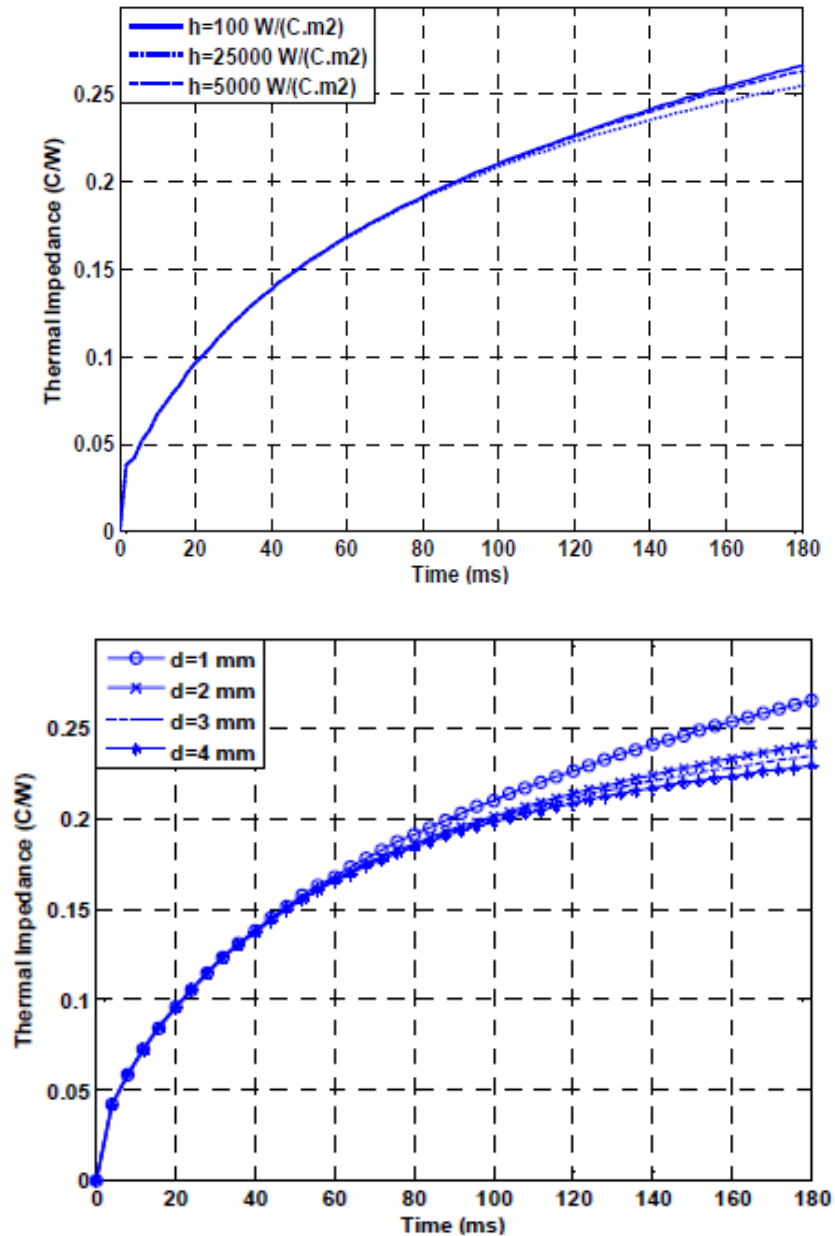
Figure 3.12: Variation in thermal impedance during a pulse: *Top*: for different heat transfer coefficients in the heat exchanger and *Bottom*: for different heat spreader thicknesses (d).[29]

A simpler design procedure is presented to achieve the minimum temperature rise during the pulse, for which the time constant, τ, of the module must be greater than

the pulse length. For any given pulse length, the time constant can be calculated using eqs.(3.1) and (3.2). This allows the thickness of the heat spreader to be calculated based on the table in Figure 3.13. For transients of 120ms or greater, a 3mm heat spreader is preferred, as increasing the thickness beyond 3mm cannot improve the thermal impedance further, as seen in Figure 3.12.

$$\tau_i = R_i C_i = \frac{d_i}{k_i A_i} \rho_i \, c_{pi} \, d_i \, A_i = \frac{d_i^2 \, \rho_i \, c_{pi}}{k_i} \tag{3.1}$$

where d is the layer thickness.

$$\tau = \sum \tau_i \tag{3.2}$$

CALCULATED TIME CONSTANT FOR EACH LAYER IN THE
POWER MODULE

	IGBT	Solder	DBC	Heat-spreader
Time Constant	0.5 ms	0.28 ms	39.4 ms	$d^2 \cdot 8.6$ ms

*d is the thickness of the heat-spreader in mm

Figure 3.13: A table of the calculated time-constant for each layer.[29]

3.2   Modelling the Thermal Response of Devices

Accurate predictions of the temperature response of power semiconductor devices is important for the purposes of design. Being able to model the thermal response of different geometry configurations allows optimum designs to be achieved without the timely and costly procedure of testing each configuration experimentally. Modelling small variations to the design of the device structure allows the sensitivity of design changes to be analysed and an optimum device design to be derived.

3.2.1   Analytical Approaches

Early attempts at modelling the temperature rise in a semiconductor device during a thermal transient took a mathematical analytical approach. The computational power required to perform large series of equations was rare and experimental technology for measuring device temperatures at high frequencies either was not available or was very expensive.

Newell [30] was one of the first to develop an analytical approach for calculating the junction temperature in power devices. The concept of the 'transient thermal impedance' is a quasi-graphical way of calculating instantaneous temperature without resorting to the mathematical complexities of the complementary error functions which appear in analytical solutions of the diffusion equations (which is addressed in Section 4.1)

The paper explains the analogy between heat conduction and electrical conduction, which demonstrates the suitability of the method for modelling heat transfer. Newell notes that the transient thermal impedance, $Z_\theta(t)$, is an "ingenious" and very useful concept for avoiding the solution of the diffusion equation. However, he states that it may more appropriately be called the step-function thermal response, as it is not analogous to the ordinary electrical impedance (it will be referred to as transient thermal impedance in this thesis).

The transient thermal impedance due to a step-function of power P_S is:

$$Z_\theta(t) = \frac{\triangle T(t)}{P_S} = \frac{1}{A}\left(\frac{4}{\pi}\frac{t}{k\,\rho\,c_p}\right)^{1/2}$$

(3.3)

A 'specific' asymptotic transient thermal impedance can be calculated for each layer in the module using eq.(3.4) [30]:

$$AZ_\theta(t) = \left(\frac{4}{\pi}\cdot\frac{t}{k\rho c_p}\right)^{1/2}$$

(3.4)

As $AZ_\theta(t)$ is dependant only on material properties, a universal design chart can be constructed. An example of the chart construction is given for a hypothetical device which consists of a silicon die mounted on a molybdenum layer and a copper package. Material properties and dimensions used are listed in Table 3.1 and the derived plot shown in Figure 3.14.

The solution follows the silicon line on the chart in Figure 3.14 until the specific transient thermal impedance equals the specific thermal resistance of the silicon layer (AR_θ), found using eq.(3.5). The solution then travels horizontally until reaching the molybdenum line, which it then follows until reaching the total thermal resistance of cumulated silicon and molybdenum layer. Again, it travels horizontally until the thermal response meets the copper line which it then follows up until the transient thermal impedance reaches the cumulated thermal response of the silicon, molybdenum and copper layers.

$$AR_\theta = \frac{L}{k}$$

(3.5)

An analytical approach had previously been taken by Carslaw and Jaeger [31], who derived an equation for the device surface temperature for a structure such as that shown in Figure 3.15. The equation for the surface temperature during a pulse is:

	k	ρc_p	L
	(W/mK)	(MW/m^{3}K)	(mm)
Silicon	84	1.72	0.5
Molybdenum	146	2.76	1.5
Copper	388	3.50	5
Aluminium	204	2.46	*not specified*

Table 3.1: Table of the material properties and dimensions of the hypothetical device.
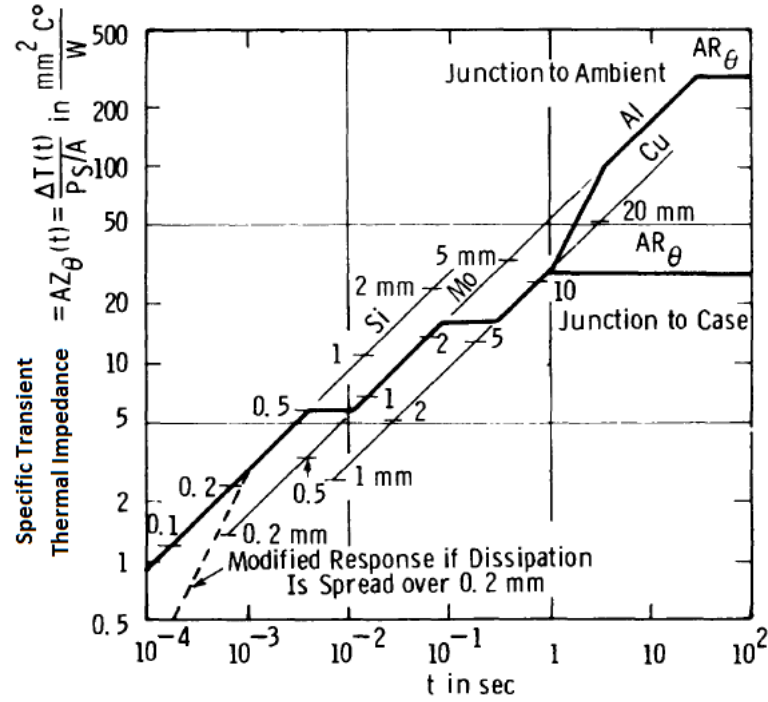
Figure 3.14: Construction of the transient thermal impedance curve for a hypothetical silicon device.[30]

$$\triangle T = \frac{1}{\rho c_p \sqrt{\pi \alpha}} \int_0^t P(t-\tau) \frac{e^{-(x^2/4\alpha\tau)}}{\sqrt{\tau}} d\tau \qquad (3.6)$$

This is derived from the analytical heat diffusion equation, which is discussed in Section 4.1.

Clemente [32] argues that certain assumptions made in the derivation of this analytical solution are unsatisfactory. Carslaw and Jaeger [31] assumed that power is generated in an infinitely thin layer at the surface of the die. Although this assumption may be adequate for pulse durations of around 100μs, shorter pulse widths are misrepresented.

Additionally, Carslaw and Jaeger [31] based the equation on the module being a semi-infinite slab. The temperature at the die attach interface begins to rise after a transient time of a few hundred microseconds, and only within this time period can the thickness of the slab be considered infinite [32].
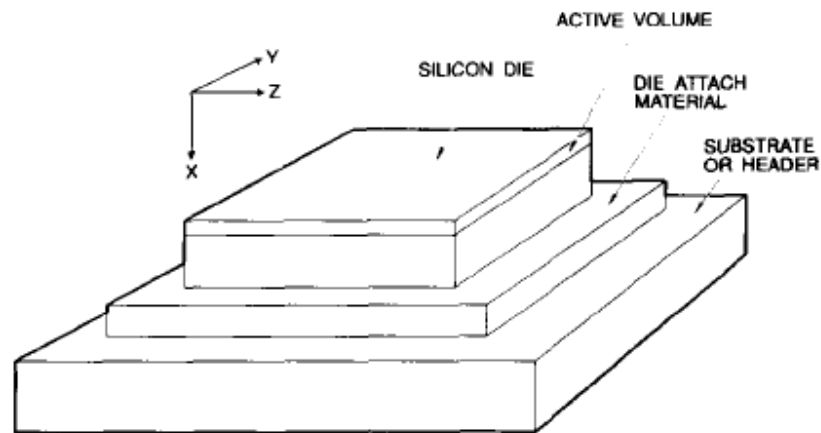
Figure 3.15: The elementary components of a power semiconductor assumed by Clemente. As the analytical model is only valid whilst the wave remains within the die, the structure from the Die Attach Material and below are there for indication only - heat transfer through these layers is not modelled. [32]

Clemente [32] develops a new model, reversing this assumption so that the power dissipation occurs instantaneously in an infinitely short period of time within a finite layer within the die, g. Equation 3.7, is argued to provide a more accurate calculation of peak junction temperature:

$$\triangle T = \frac{E}{\rho c_p g} \qquad for\ 0 < x < g \tag{3.7}$$

where E is the energy density (J/m^2), and can also be expressed as the power density multiplied by the pulse time.

A limit to equation 3.7 is that the heat which dissipates out the finite thickness $'g'$ is not accounted for. It was found that it takes less than 1μs for the temperature

15μm from the active volume 'g' to increase by 10%, and that a volume 30% larger than the size of g has seen a significant temperature rise after 1μs. Although the equation provided by Clemente is improved, it is still not sufficiently accurate. The introduction of a correction factor, ϕ, improves the accuracy of the equation further.

The correction factor is an effective thickness, defined as the value of x that has reached 90% of its final value within the pulse duration:

$$\triangle T = \frac{E}{\rho c_p(g + \phi)} \tag{3.8}$$

Figure 3.16 shows how the correction factor varies with pulse duration for different thicknesses of active volume. A comparison of the temperature rise using the traditional method proposed by Carslaw and Jaeger, and Clemente's improved method using the correction factor (equations 3.8). Two cases are examined using the improved method for two thicknesses of active volume (g=15μm and 100μm) that would be appropriate for 100 and 1000 V devices, respectively. The traditional method always assumes g=0μm, and hence only one curve is shown. The temperature rise from the different scenarios is shown in Figure 3.17.

The temperature prediction at the beginning of the pulse is reduced by using the proposed methods byClemente. The greater the active thickness g, the lower the temperature rise, as the power density is reduced proportionally with g. Clemente's method is more accurate for predicting the die temperature for very short pulse durations, however it is limited to use for times shorter than the transit time across the thickness of the die, calculated to be 300μs.
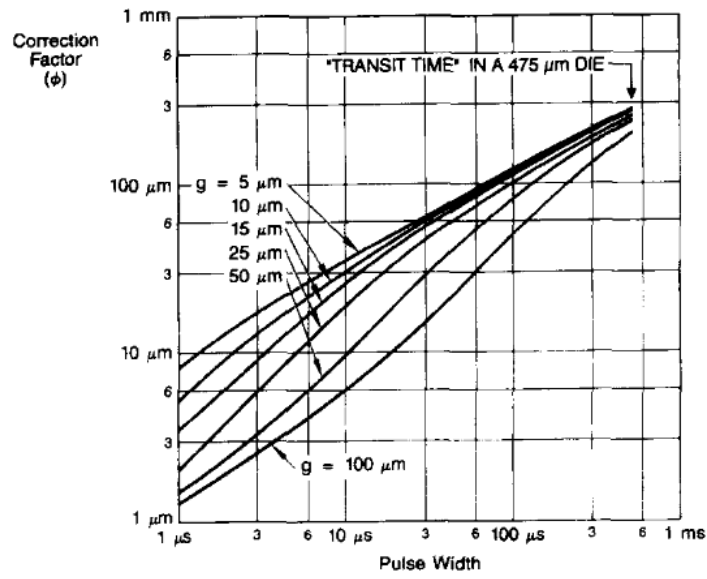
Figure 3.16: Correction factor as a function of pulse width for different thicknesses of active volume. The effective thickness of the active volume should be increased by an amount that depends on the pulse length.[32]
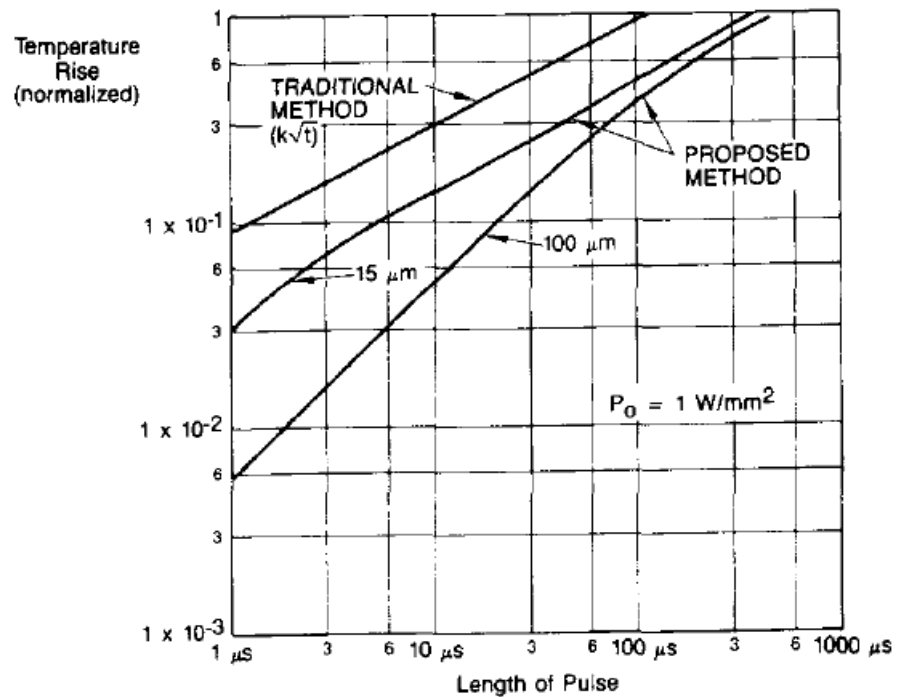
Figure 3.17: Comparison of the temperature rise calculated with the traditional method compared to the proposed new method.[32]

Chambers et al. [33] also use the assumption of a semi-infinite solid to calculate the junction temperature rise at a pulse time, t_p, during a heat pulse of magnitude P:

$$T_j - T_o = \frac{2P}{A\sqrt{\pi \rho c_p k}}\sqrt{t_p} \tag{3.9}$$

where T_j and T_o are the junction temperature at time t_p, and the initial junction temperature, respectively.

An equation for the maximum power dissipation without exceeding the maximum junction temperature (T_{jmax}) is also derived by [33] which is shown in eq.(3.10). The assumption of a semi-infinite slab means the heat conduction remains within the early regime.

$$P_{max} = \frac{A(T_{jmax} - T_o)\sqrt{\pi \rho c_p k}}{2\sqrt{t_p}} \tag{3.10}$$

For times which are beyond the applicability of the semi-infinite model, a conservative assumption that the die (of volume V) has adiabatic sides, which allows maximum power estimates to be made when heat conduction is in the late regime. A new equation for the limit to the power dissipated is developed using energy balances [33]:

$$P_{max} = \frac{\rho c_p V (T_{jmax} - T_o)}{t_p} \tag{3.11}$$

where V is the volume of the die. This new equation can be used until the maximum power equals the steady-state power limit. An estimate of the time limit, $t_{p\,limit}$, can be made

The pulse time limit for which eq.(3.10) is valid ,$t_{p\,limit}$, can be estimated by equating eq.(3.10) equal to eq.(3.11) and solving for t_p. It is found that for a die of thickness d [33]:

$$t_{p\,limit} = \frac{4}{\pi} \frac{d^2}{\alpha} \tag{3.12}$$

The paper continues by analysing the situation where the die is mounted on a heat sink. It explains that the same equations can be used up until the time limit $t_{p\,limit}$. Beyond this time, the heat sink base temperature, $T_{hs\,base}$, and maximum power dissipation, P_{max}, are given by the following equations [33]:

$$T_{hs\,base} = \frac{2P}{A\sqrt{\pi \rho_{hs} c_{phs} k_{hs}}} \sqrt{t_p - t_{p\,limit}} \tag{3.13}$$

$$P_{max} = \frac{A(T_{jmax} - T_o)}{\frac{d}{k} + \frac{2\sqrt{t_p - t_{p\,limit}}}{\sqrt{\pi \rho_{hs} c_{p\,hs} k_{hs}}}} \tag{3.14}$$

$$for: \ t_p > t_{p\,limit}$$

3.2.2   Developing RC Models

RC networks are an important aspect of the thesis. The definition and descriptions of RC networks can be found in Section 4.3.

Obtaining a thermal Resistance Capacitance (RC) network for a semiconductor device structure has received much attention. If an RC circuit can be determined for a given module, it can be used to determine the maximum junction temperature under different heat pulse conditions. This idea has existed for a long time, detailed in publications such as Pritchard [34].

Szekely [35] presents a method for determining thermal RC values of a structure when an experimental transient heating or cooling temperature curve is available. It is stated that the result of the transient thermal measurement is essentially a step-function response.

A transient thermal response (transient temperature curve) can be defined by different time constants, which determine the shape of the temperature curve at different pulse times. Szekely describes how thermal responses of complex thermal structures can be described by the sum of many exponential terms, each with a different time constant (τ) and amplitude (a).

The thermal response function (i.e. the transient thermal response) $a(t)$ can be calculated using eq.(3.15) from the time constants and amplitudes in the system, which can be extracted from the experimental curve.

$$a(t) = \sum_{i=1}^{N} a_i (1 - \exp(-t/\tau_i)) \tag{3.15}$$

The first step of extracting the time constants and amplitudes from the experimental curve involves transforming the response function to the logarithmic time variable. The discrete time constants determined by the curve fitting can be represented as

a continuous spectrum, which would more accurately represent the response. Diffe-
rentiating this continuous response provides an equation which can be deconvoluted
to extract a spectrum that completely describes the behaviour of the system. An
example of the discrete and continuous time-constant spectrum is shown in Figure
3.18.
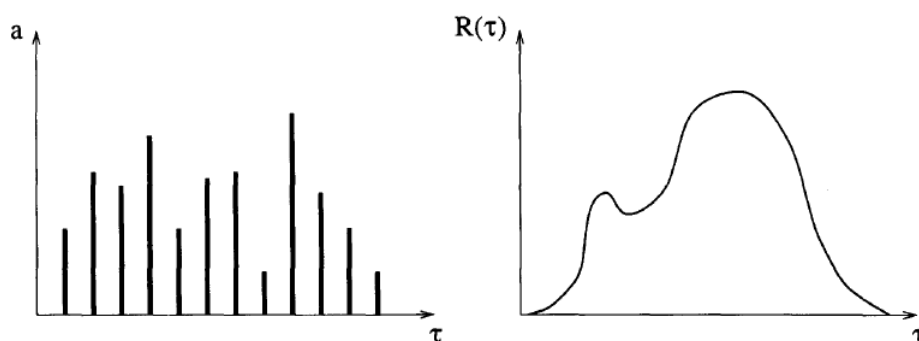
Figure 3.18: Time-constant spectra - *Left:* lumped-element network; *Right:* distribu-
ted thermal network.[35]

Discretising the spectrum allows the calculation of the corresponding RC values for the
Foster network. The Foster network has no physical meaning, and the transformation
of these values to the Cauer network is necessary. Definitions of the Foster and
Cauer networks can be found in Section 4.3. This process is described as a standard
procedure of the linear circuit theory, and refers the reader to Weinberg [36] for the
procedure.

A similar approach is taken by Bagnoli et al. [37] in the development of the TRAIT
analysis, which represents a device heating or cooling curve as an exponential term.
However, Bagnoli et al. question the hypothesis of Szekely, who stated that the time-
constant spectrum obtained can be retained as continuous. Instead, Bagnoli et al.
perform numerical analysis on the discrete time-constant spectrum.

Bagnoli et al. describe the electrical-thermal analogy in detail before the RC network
parameter calculations are laid out. A Laplace Transform of the cooling transient
due to a step in thermal power is developed. The input thermal impedance can be
found using the junction to ambient thermal resistance and the time constant data
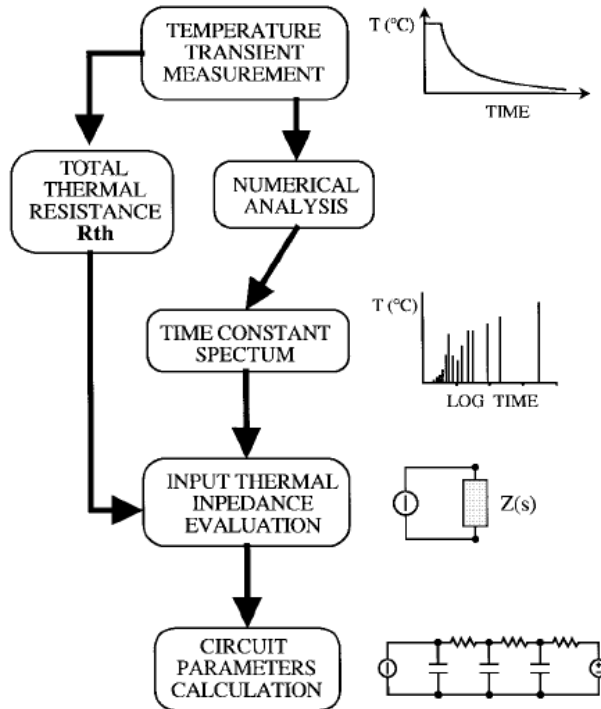
Figure 3.19: Schematic flow chart of the TRAIT method presented by Bagnoli et al. [37].

describing the cooling curve. This is decomposed using the Cauer algorithm, providing the thermal resistances and capacitances for the Cauer network. The process of the TRAIT method is outlined in Figure 3.19.

Whitehead and Johnson [38] explain the process in their paper, which obtains the RC network for an Insulated Gate Bipolar Transistor (IGBT) device on a multi-device module. The transient thermal impedance, $Z_\theta(t)$, (discussed previously in overview of Newell [30]) is defined in eqs.(3.16) and (3.17). A curve fitting procedure is then applied to determine the individual exponential terms - R_{TH}, k_n, α_n and an appropriate number of terms. The normalised representation of the experimental curve is shown in Figure 3.20, which shows the four individual terms and the combination of the terms which fit the experimental response.

$$Z_\theta(t) = \frac{T_j(t) - T_a(t)}{P} = \frac{\triangle T_{ja}(t)}{P} \tag{3.16}$$

$$Z_\theta(t) = R_{TH} \left(1 - \sum_{n=1}^{N} \frac{k_n}{\alpha_n} e^{-\alpha_n t} \right) \qquad (3.17)$$
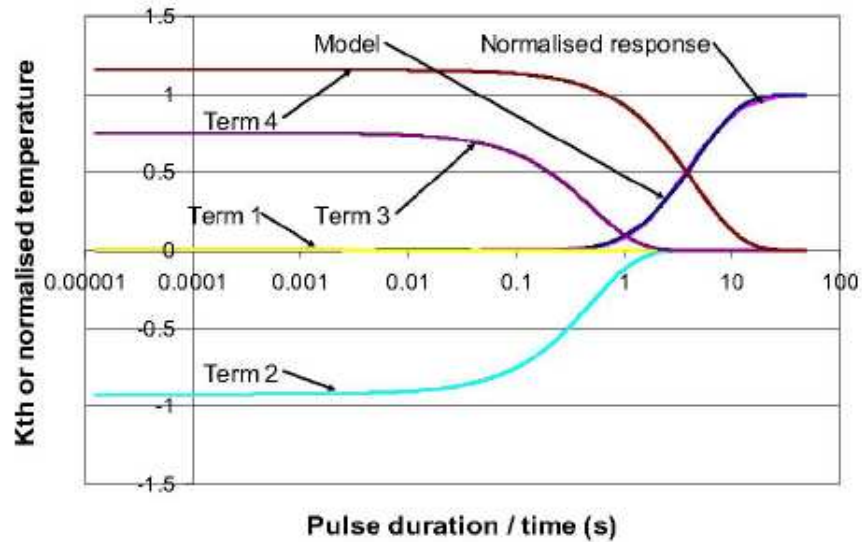
Figure 3.20: Normalised 4th order representation of the transient response.[38]

Using the experimental thermal response curve to extract RC parameters continues to receive much attention. Xu et al. [39] use concepts from linear network theory to develop a 4-rung RC ladder by means of Laplace Transforms. A temperature curve is obtained using the RC model, which is compared to the experimental curve from which it was derived. Igic et al. [40] use a deconvolution approach to obtain a set of RC terms. They then accurately predict the temperature response of a MOSFET with a 3 stage RC circuit. The results from both Xu et al. and Igic et al. can be seen in Figure 3.21.

Ludwig et al. [41] investigate the accuracy of both a Foster RC network model and a CFD model to represent an experimental system. Thermal data obtained from each numerical method is plotted alongside the experimental curve from which the RC terms were obtained. Figure 3.22 demonstrates that the curves show good agreement.
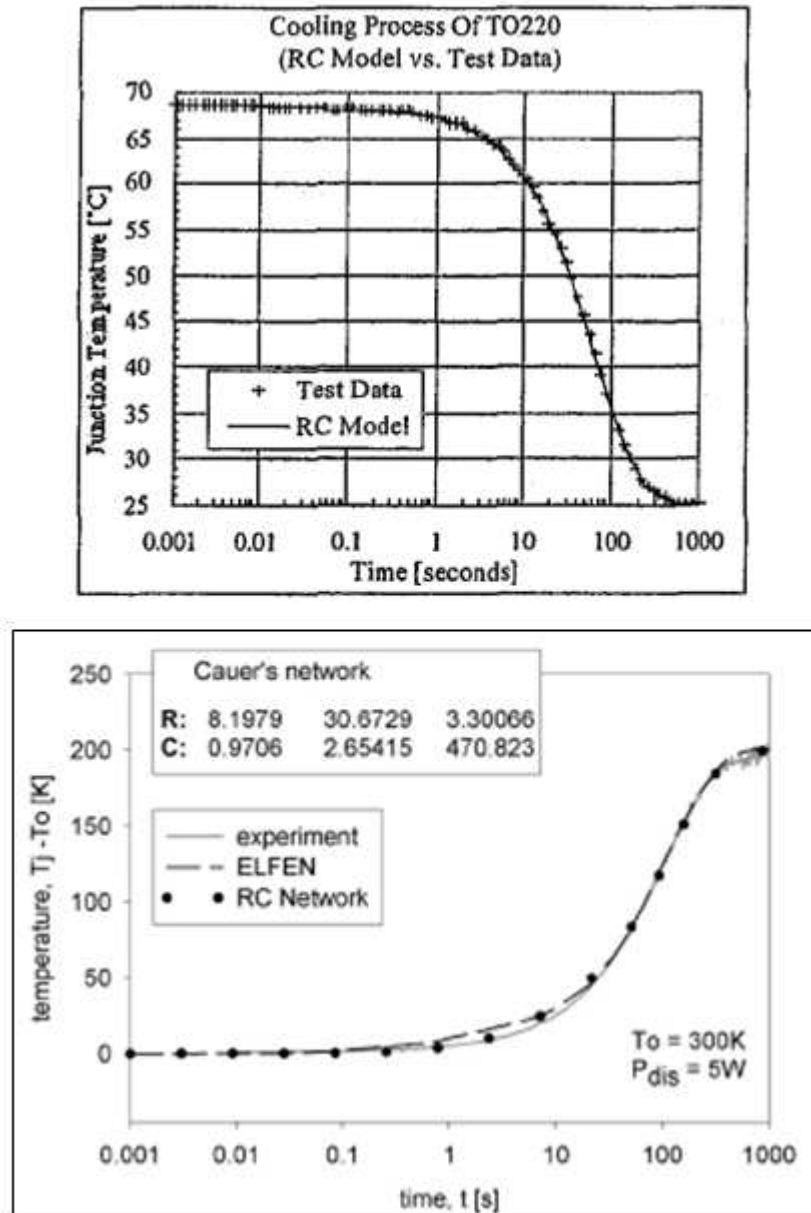
Figure 3.21: Comparison of thermal response curves obtained by experimental and RC Modelling methods. *Top*: Xu et al. [39], *Bottom*: Igic et al. [40]
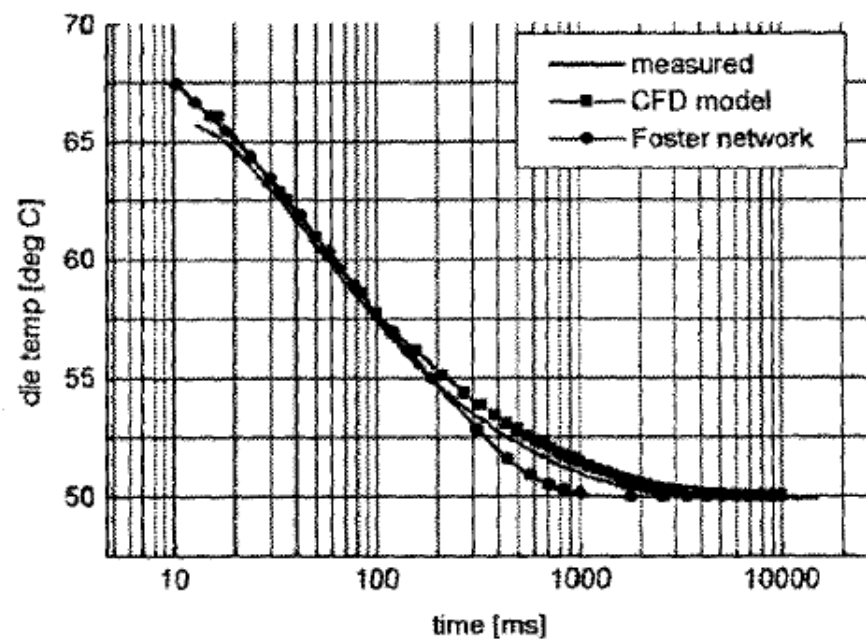
Figure 3.22: Measured and simulated cooling curves.[41]

Cauer networks can be developed based on the physical geometry of the system (dimensions and material properties). In this case, an experimental thermal response curve is not needed, as the RC terms of a Cauer network have a physical meaning and are calculated from physical properties. This approach is used by Jankowski and McCluskey [42] to compare the transient temperature rise of the same power electronic device in different packages. Using equations based on the geometry, the Cauer network RC terms were calculated:

$$R_{th} = \frac{x}{kA} \tag{3.18}$$

$$C_{th} = \rho c_p A x \tag{3.19}$$

$$\tau = R_{th} C_{th} = \frac{x^2 \rho c_p}{k} \tag{3.20}$$

The RC network formulation process is detailed, and analysis is performed to investigate the effect of geometry and heat convection rates in different packages.

An assumption made by Jankowski and McCluskey is that the thermal capacitance and thermal resistance of the entire cross-section is immediately available. When a change of cross-section occurs, heat spreading occurs over a period of time. This is not taken into account byJankowski and McCluskey, which limits the accuracy of their model.

Masana takes this into account when constructing a Cauer RC network of a physical structure. The Variable Angle Method (VAM) is detailed in [43], and summarised in [44] where he applies it to dynamic thermal modelling of semiconductor packages. The demonstration of heat spreading from a heat source into a volume of greater cross-sectional area can be seen in Figure 3.23.
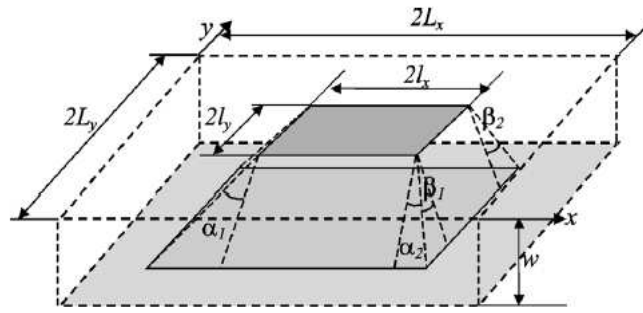
Figure 3.23: Sketch of the volume for VAM.[44]

New equations for the thermal resistances and capacitances are derived mathematically:

$$R_{th} = \frac{1}{4\,k\,l_x} \frac{1}{(\gamma_e \tan\alpha - \tan\beta)} \ln \frac{l_x + w\tan\alpha}{l_x + w\tan\beta/\gamma_e} \tag{3.21}$$

$$C_{th} = V\rho c_p \tag{3.22}$$

where V is the active volume and γ_e and γ_s are geometrical functions:

$$V = 4 \left[l_x\,l_y\,w + (l_x\,\tan\beta + l_y\,\tan\alpha)\frac{w^2}{2} + \tan\beta\,\tan\alpha\,\frac{w^3}{3} \right] \tag{3.23}$$

$$\gamma_e = l_y/l_x \quad \text{and} \quad \gamma_s = L_y/L_x$$

The terms $\tan\alpha$ and $\tan\beta$ are expressions for the spreading angles, and equations for them can be found within the paper. Using these equations Cauer RC terms can be calculated which take into account the dynamics of transient thermal modelling, creating a more accurate representation of semiconductor packages.

Masana [44] demonstrates the method in a worked example for the four layer package shown in Figure 3.24. The VAM expressions defining the VAM were constructed and

solved using a spreadsheet. A finite element method was also used to find the thermal response of the package for comparison. Figure 3.25 shows a comparison of the results obtained which demonstrates a good correlation.
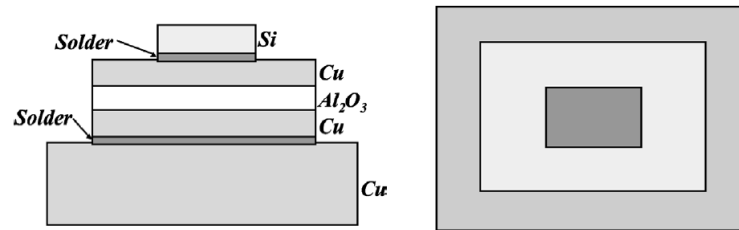
Figure 3.24: The package modelled using the VAM a finite element methods. *Left*: side view, *Right*: Top view.[44]
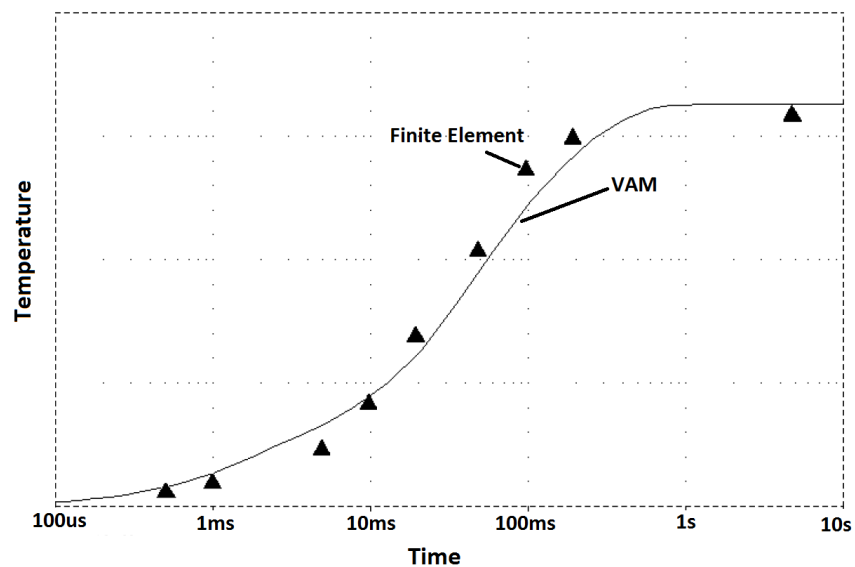
Figure 3.25: Comparison of the results obtained using the VAM (solid line) and finite element method (triangular points).[44]

Due to increasing computer power, Computational Fluid Dynamics (CFD) and Finite Element Analysis (FEA) have become increasingly common for solving transient heat conduction problems, such as the heat flow in semiconductor packages. Details of these methods can be found in the research of Lohner [45] amongst others.

Another numerical method for solving the transient heat flow in electronic packages is the Boundary Element method. Application of the method specifically to transient flow electronic packages is addressed by Guven et al. [46].

3.2.3   Fourier Based Methods

A Fourier based method of solving thermal transients has been presented by Swan et al. [47], in which the heat equation is solved using the Fourier series. The application of the method in the paper is for a co-simulation, modelling thermal transient heat conduction alongside an electrical simulator. It is required to create a thermal model to predict the transient temperatures of devices in real-time, according to a varying power input supplied by the electrical simulator.

Swan et al. state that the Fourier series method was chosen because it can simulate "extremely quickly compared to FDM or FEM simulations". A further benefit to the method is that the Fourier series solution can integrate easily into an electrical simulation. It is, however, pointed out that the Fourier series would not be as appropriate for more complex geometry.

The solution to the 1D form of the heat equation for the *m*-th term in in the Fourier series is given as:

$$\frac{dT_m}{dt} = \frac{2\alpha}{W}\left[\frac{\partial T}{\partial x}W\,(-1)^m - \frac{\partial T}{\partial x}0\right] - \frac{\alpha\pi^2 m^2}{W^2}T_m(t) \tag{3.24}$$

The solution to the 2D heat equation is also provided.

The method is used to solve both a 1D and 2D example, comparing the results in each case to those obtained by modelling the same problem using FLOTHERM, a finite element method simulation package. The 1D example examines heat flow through a series of layers of different materials, due to a heat flux applied at the top surface. The temperature at different points throughout the stack of layers is monitored. A sample of the monitored temperatures are plotted during the transient period, and compared to the temperatures at the equivalent place in the FLOTHERM model (Figure 3.26).

The transient results comparison is shown in a log scale in Figure 3.27. The Fourier method is shown to have very good agreement with the FLOTHERM results throughout the transient period.
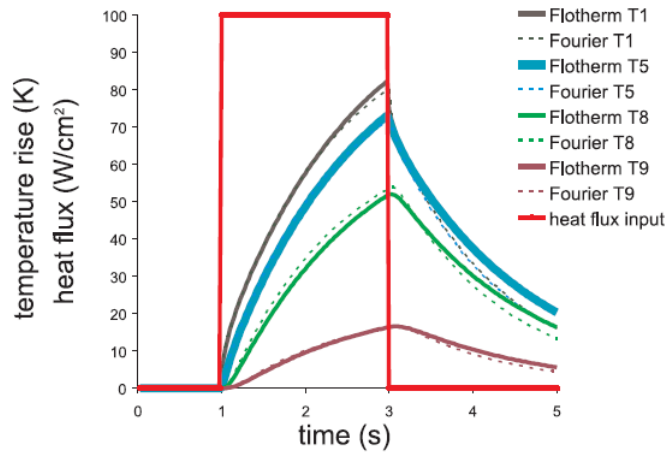
Figure 3.26: Comparison of the 1-D Fourier series method and FLOTHERM. This shows the transient temperature rise due to a step power input. The curves show selected monitoring points at the material interfaces. T1 represents the top of the stack, where the heat is generated.
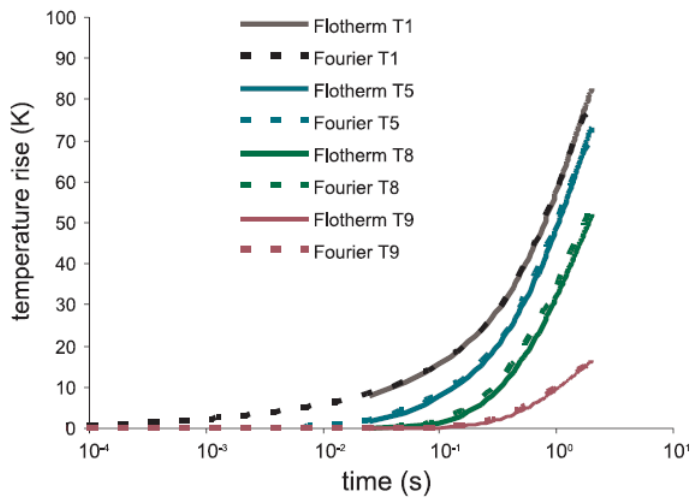
Figure 3.27: Temperature rise versus time, where time=0s is at the moment at which the heat flux is applied to the top of the stack.

The method is validated further by Bryant et al. [48]. In this paper, the heat equation is solved in 3D using the Fourier series, where the mathematics involved in solving the equation in 3D is explained in Swan et al. [49]. An experiment is setup to test the ability of the Fourier series solver to predict device temperatures in real-time, according to a varying power input. Figure 3.28 shows the setup of the experiment,

where four device temperatures (two diodes and two IGBTs) are calculated.

A 3D Fourier series model is created to replicate the problem, alongside a test piece of the setup in Figure 3.28, which is tested experimentally. The varying electrical test loads in the experiment is pre-determined, based on the Artemis urban driving cycle. The temperatures of the experimental test piece are measured using a thermal imaging camera.

Figure 3.29 shows the temperatures of each of the four devices during one of the 60s test periods. Each graph represents the temperature of one device. In each graph, the measured temperature from the thermal imaging camera is compared to the temperatures calculated using the real-time Fourier series method. A very close match is seen in each case. The modelling results are calculated at a higher frequency than the thermal imaging measurements, and therefore show a higher level of detail.

The validation procedure demonstrates the ability of the Fourier series method to calculate device temperatures not only accurately, but in real-time.
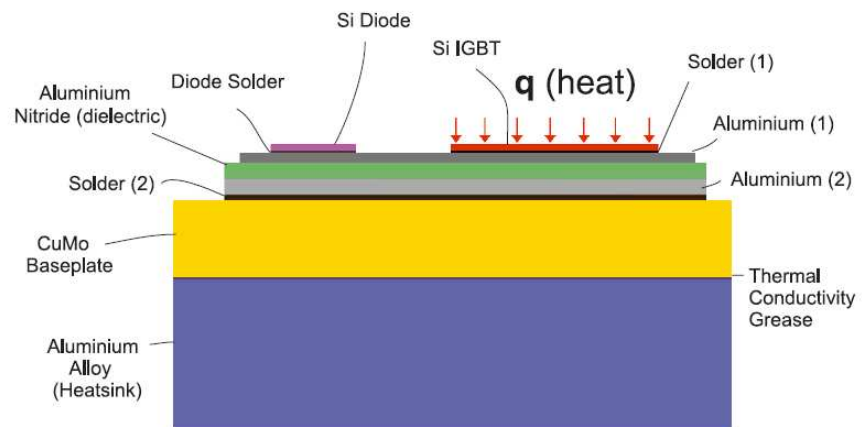
(a) Cross-sectional view of the packaging associated with one IGBT chip and one diode chip.
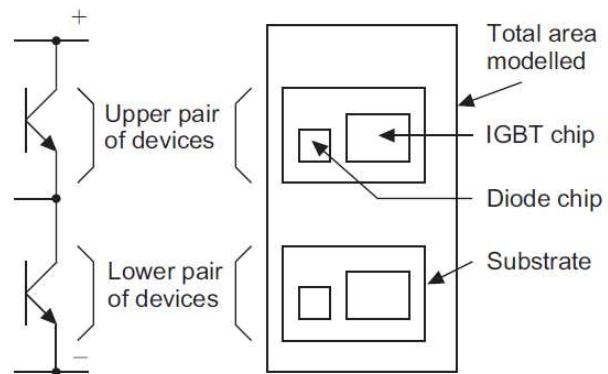
(b) Plan view of chip layout for one inverter phase leg.

Figure 3.28: Structure of the device packaging which is used for the validation.
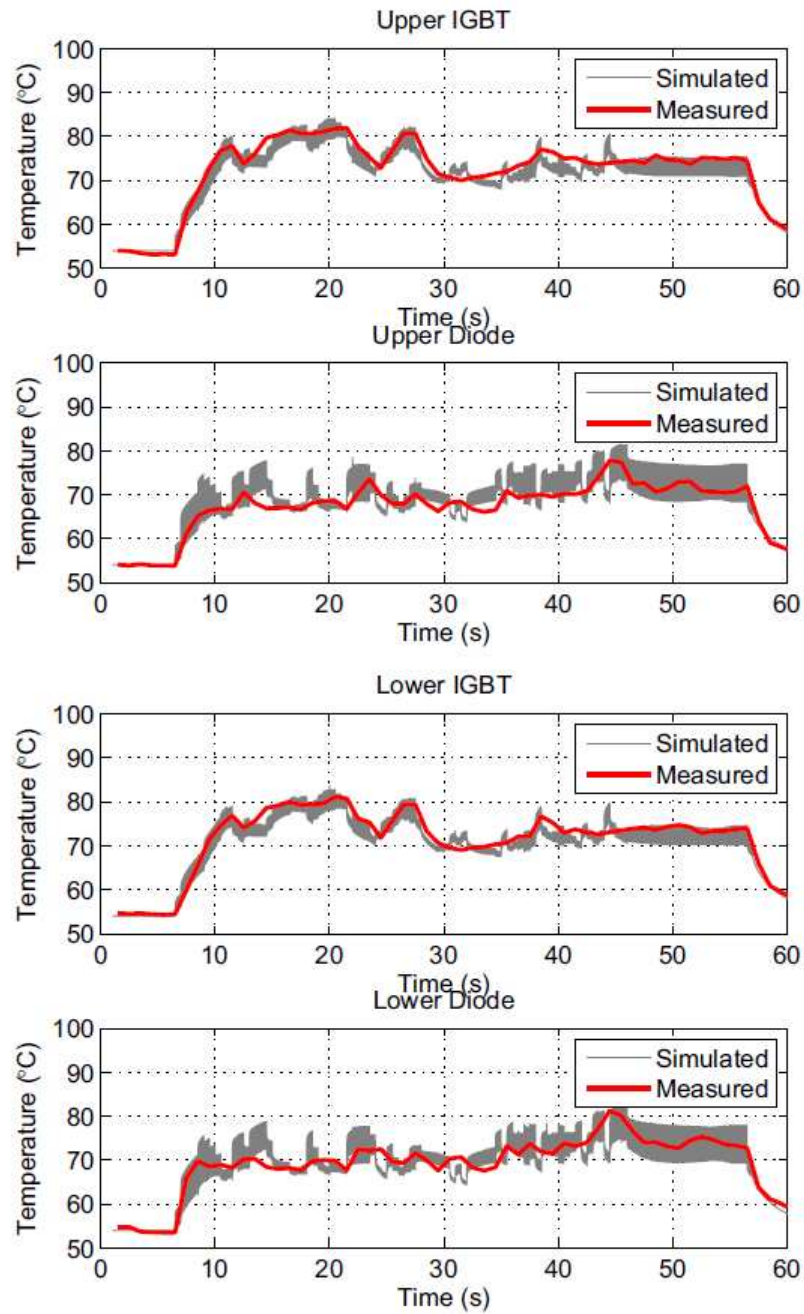
Figure 3.29: Comparison of modelling and experimental device temperatures during one load cycle.

3.2.4   Review of Modelling Methods

The analytical methods discussed have made assumptions, limiting the accuracy and range of applications that each equation is valid for. These limitations include the pulse duration for which the equations are valid, and each analytical method appears only to be valid for very short pulses.

RC networks provide a more accurate modelling method for the entire transient period, up to steady state conditions. Networks are created for each use, and therefore they have a high versatility as they take into account variations in stack structures. However, in order to construct a network for a given stack, it is necessary to know either the structural details of the stack (for a Cauer network), or a transient temperature response curve (for a Foster Network). It should also be noted that the construction of a Cauer network for a 3D structure is not a simple task. Solving the RC network normally requires computational power as a long series of equations must be solved at each time step. Larger and more detailed networks require greater computational power. Formulating and solving RC networks is discussed fully in Section 4.3.

CFD and FEA allow complex scenarios to be modelled with reasonable accuracy when used correctly. The creation of models in CFD software packages is time consuming, and it is easy to misuse the software which can lead to false results.

The Fourier series method is a candidate for use as the modelling method in the thesis. The method is able to predict device temperatures very quickly, although the accuracy of the results cannot be considered as accurate as those achieved using high detail thermal networks, as the infinite Fourier series must be simplified to an expression with a finite number of terms. In addition to this it is not suitable for modelling complex geometries. As solving time is not an issue, the versatility of the thermal resistance network method makes it a better option.

For this thesis, the RC network is the most suitable. Sufficient computational power is available for creating detailed networks and solving the resulting equations. The

networks allow suitably accurate representations of stacks to be created in sufficient detail to provide an accurate thermal response to heat pulses.

3.3 Reliability of Semiconductor Devices

Attaching a heat sink to the top of a power electronic device causes the structure of the stack-up to be modified. This requires consideration of any reliability issues imposed by the new stack design. A stack design which may provide excellent cooling performance may not be able to perform reliably over time. Operating conditions for the devices include thermal cycling due to the device being switched on and off. This can cause thermal stresses in the stack due to joining layers expanding and contracting at different rates.

The reliability of power semiconductor devices in avionics provides additional challenges due to the high reliability required and harsh operating environments. Operating conditions may be between -50ºC and 200ºC, with a temperature change of ±10ºC/min [50]. It is pointed out that material selection in these modules needs to provide the lowest junction to casing thermal resistance, and also to minimise thermal coefficient of expansion (CTE) mismatching which leads to mechanical stresses. Therefore the consideration of reliability when designing power modules and devices is essential.

3.3.1 Solder Joint Integrity

Effects of Solder Voids on Transient Device Temperature

The operation of power semiconductor modules creates stresses in the solder layers which lead to the growth of voids. The voids inhibit the heat flow from the device to the heat sink, causing an increase in temperature. A demonstration of the void growth and effect on the heat flow is shown in Figure 3.30.

Katsis and van Wyk [51] investigate the transient temperature effects in power semiconductor devices due to solder voids. The increase in the voided area of the solder joint increases with thermal cycles, which can be seen in Figure 3.30. The transient temperature rise in a MOSFET was measured for various modules with different
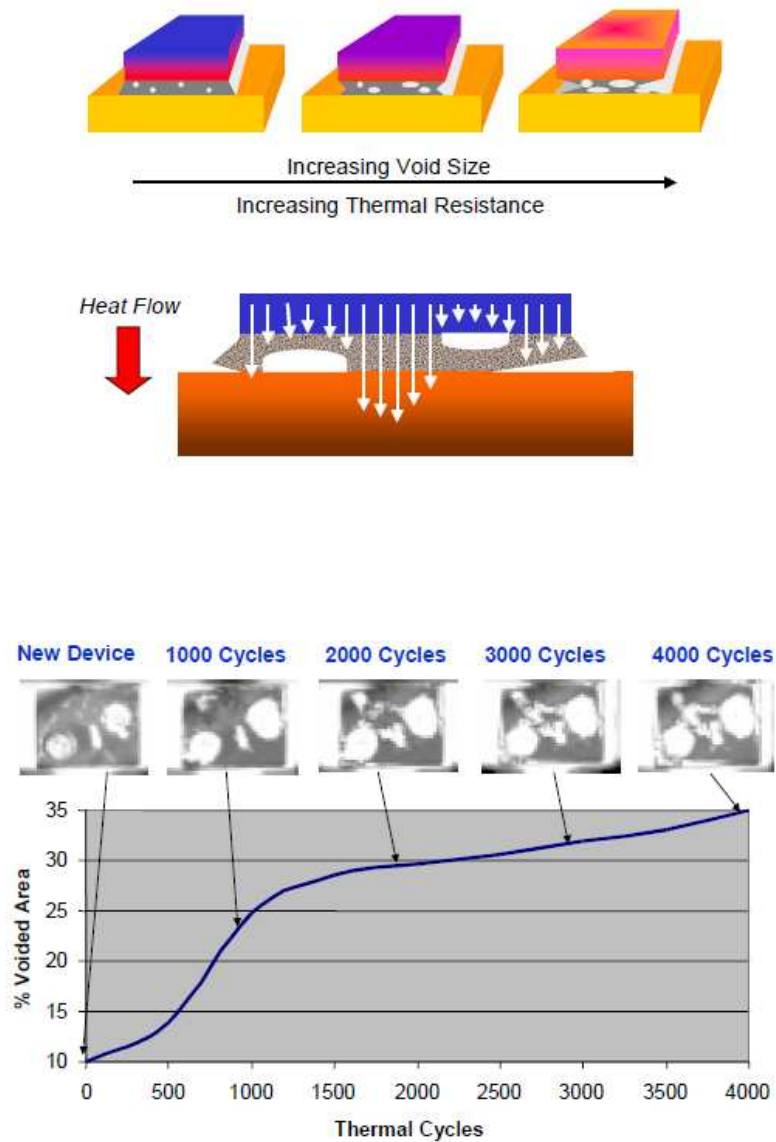
Figure 3.30: *Top:* The affect of solder voids on the heat flow in the power electronic module. *Bottom:* The increase in solder void area with increasing number of thermal cycles.[51]

amounts of solder voids. The predicted results obtained using thermal impedance calculations are shown in Figure 3.31. The experiment results agree with the predicted results, proving that solder voids significantly affect the device temperature for pulses greater than around 5ms.
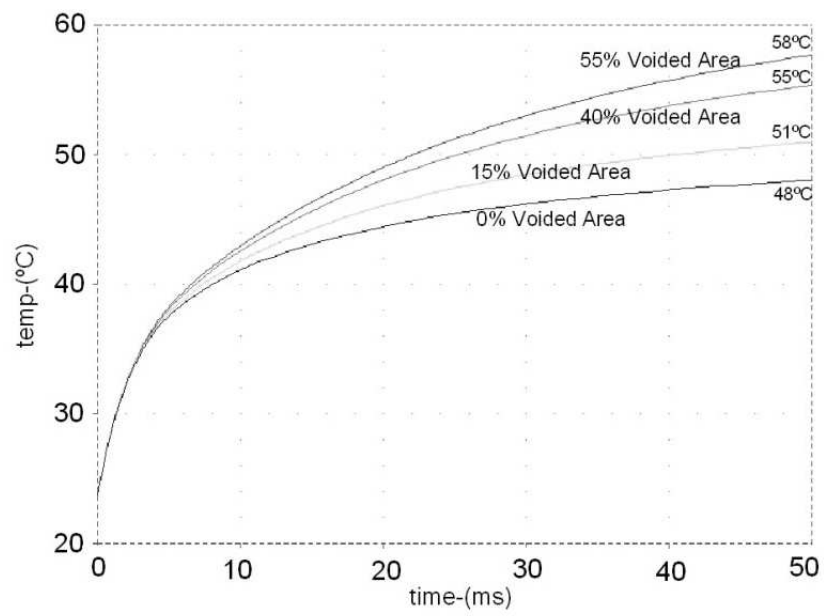
Figure 3.31: Predicted temperature versus time for a MOSFET.[51]

Solder Thickness Effects

The improvement of the fatigue life of solders has been examined by Hayashi et al. [52] and Guth and Mahnke [53]. Both papers investigated the relationship between solder thickness and solder cracks during thermal cycling.

Hayashi et al. [52] used a power chip mounted onto a substrate and a baseplate with various solder thicknesses, cycling the packages between -40°C and 125°C for 1000 cycles. Solder thickness ranged from 50μm to 550μm. The crack length decreased with an increase in solder thickness. Figure 3.32 shows the relationship between the solder thickness and crack length. For solder thicknesses less than 200μm, this rate of change in the crack length was greater than for solder thicknesses over 200μm. Also examined in the paper is the crack growth rate, which showed a similar relationship. The crack in thin solders grows at a faster rate than thicker ones under the same operating conditions (Figure 3.32).
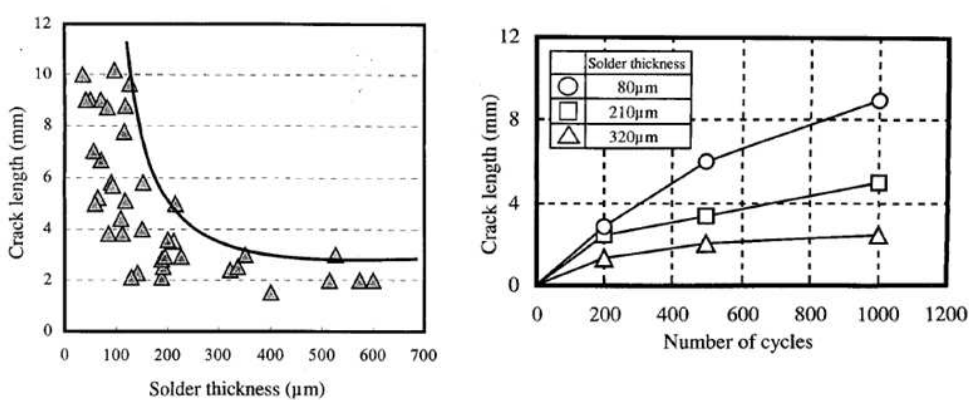
Figure 3.32: *Left:* Relation between solder thickness and crack length after 1000cycles. *Right:* Crack growth rate.[52]

Guth and Mahnke [53] conducted similar tests to investigate the effect of solder thickness on solder cracking. Devices mounted on a substrate and baseplate with different solder thicknesses were put through 8000 thermal cycles. Whereas Hayashi et al. presented a relationship where the stresses increased as solder thickness decreased, Guth and Mahnke found that an optimum thickness exists. Figure 3.33 demonstrates that the minimum crack growth after 8000 cycles was found for solder thicknesses around

150-200μm. The cracks were examined using cross sectional images, and it was found that the crack in the thick solder layer starts at a different location to thinner ones. The formation of the solder fillet in thicker solder layers is less regular, causing a lower reliability.
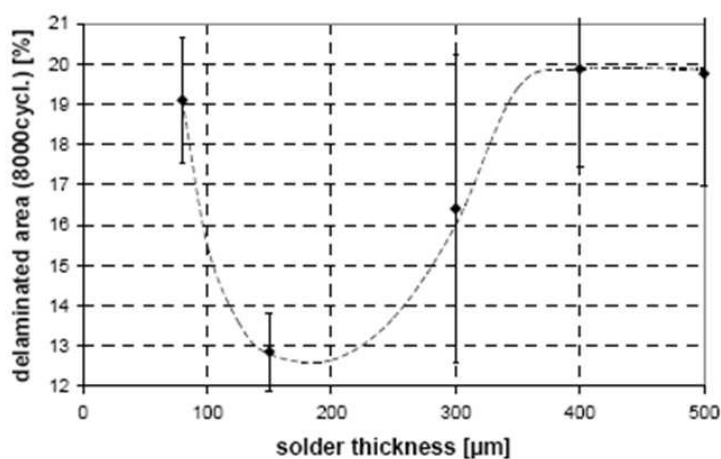
Figure 3.33: Delamination area as a function of solder thickness after 8000 thermal cycles.[53]

Maintaining a constant solder thickness throughout a joint is a difficult and important task, recognised by both Hayashi et al. and Guth and Mahnke. An uneven joint will produce an area of thin solder thickness at one edge and therefore reduce the overall reliability. Both investigated the use of spacers (termed 'wire bumps' by Hayashi et al.) such as those shown in Figure 3.34 to provided a constant and more reliable solder joint.
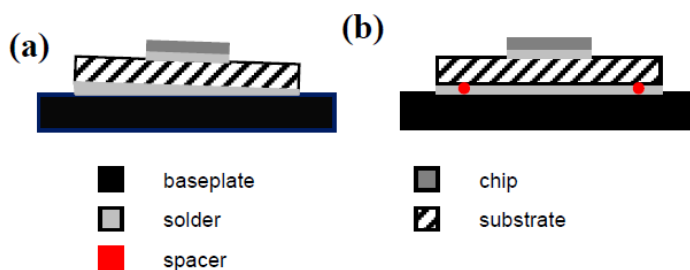
Figure 3.34: *Left*: Conventional design of substrate-baseplate solder connection. *Right*: Homogeneous solder layer thickness due to spacers.[53]

Experiments performed by Hayashi et al. [52] using the spacer technique resulted in

a considerably more consistent solder thickness. After thermal cycling for 150 cycles, a 'considerable crack' is visible in the conventional joints, yet no crack was observed in the wire bump joints even after 400 cycles. The thermal resistance across the joints was measured experimentally. The use of the wire bumps was reported not to adversely affect the thermal resistance before any thermal cycling had been applied. During thermal cycling, the conventional joints endure quick crack growth rates, which increase the thermal resistance over the solder. In comparison, the thermal resistance across the wire bump technique solder barely increases during cycling. The changes in the thermal resistance for the two models are shown in Figure 3.35.
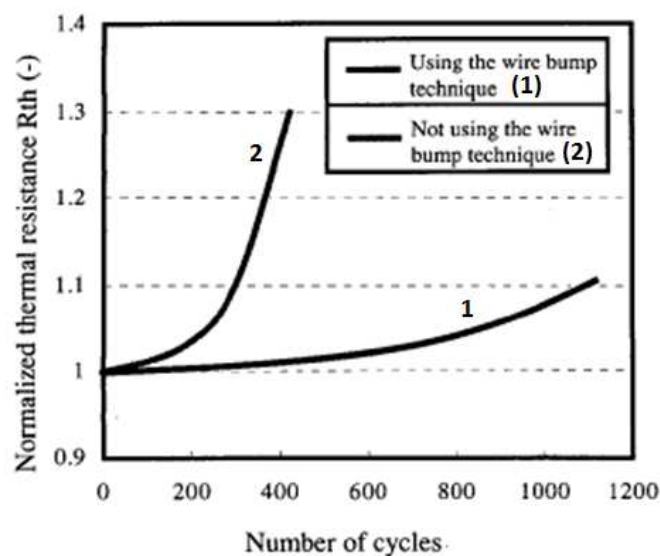
Figure 3.35: Effect of wire bump technology on the thermal resistance under the thermal cycle test. [52]

Guth and Mahnke [53] observed that the delamination in solder joints begins at one edge, and it was found to start in areas where the solder thickness was less than 100μm. A spacer technology, similar in principle to the wire bond technology used by Hayashi et al., was used to achieve a more consistent solder thickness. This resulted in delamination starting in all four corners simultaneously, suggesting a much more even thickness.

The percentage of the area that had become delaminated was measured at intervals up to 10,000 thermal cycles, and the results showed that the use of spacers improved

the reliability by at least a factor of 6. This can be seen in Figure 3.36. It is concluded in the paper that the impact of a homogeneous layer exceeds even that of the solder material.
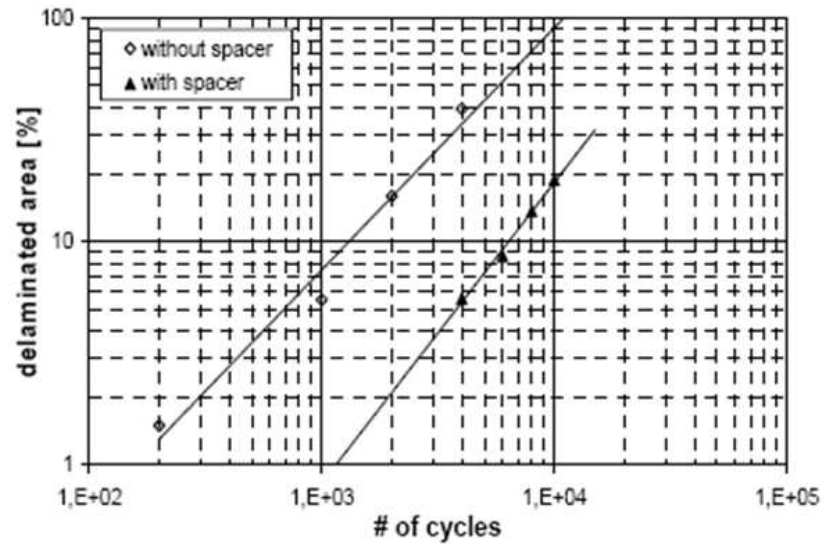
Figure 3.36: The delaminated area vs number of thermal cycles with and without the user of spacer technology. [53]

Solder Materials

The focus on lead free solders has increased since the use of lead in electronics became forbidden in Europe in July 1996 [54]. Exceptions to the ban exist, such as aerospace applications due to the high reliability required under harsh operating conditions.

Guth and Mahnke [53] address the effect of solder material on the reliability of the power modules. The delamination area in the solder joints of modules using lead and lead-free solders was measured during thermal cycling over an 80ºC range. One of the solders tested was an Indium solder (SnAg3.5-In1) due to claims by Nishimura et al. [55] that it achieved high reliability in their lead free IGBT module. However, results achieved by Guth and Mahnke (shown in Figure 3.37) contradicted this, and found it was much worse than the two lead solders tested (Sn-Pb50 and Sn-Pb40-Ag1) as well as the tin silver solder, Sn-Ag3.5. Sn-Ag3.5 was the most reliable solder of the four tested, having the smallest solder area delaminated after 10,000 cycles.
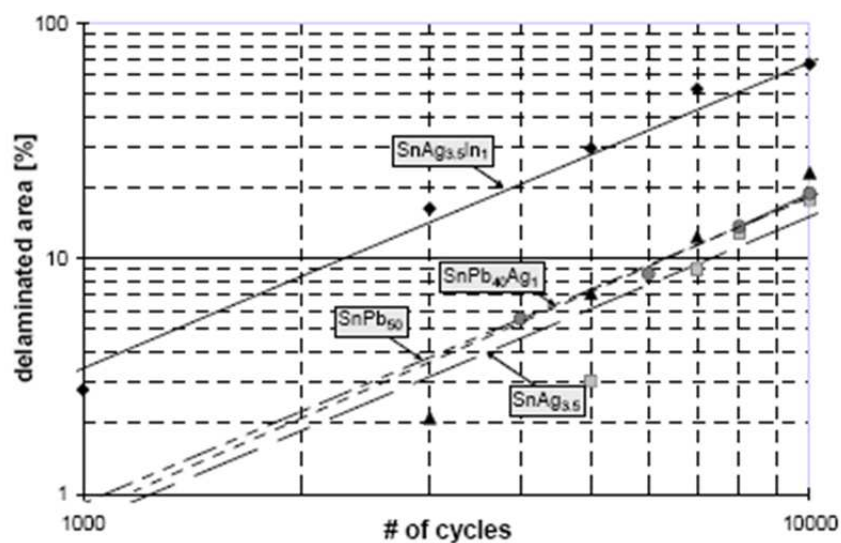
Figure 3.37: Delaminated area vs number of cycles for different solder materials.[53]

Schubert et al. [56] evaluate and compare the fatigue life of lead solder (Sn-Pb-Ag) and a lead-free solder (Sn-Ag-Cu). The experimental work examined the number of cycles to failure for flip-chip assemblies without underfill, and the results are shown graphically in Figure 3.38. Cycling between -50°C and 20°C demonstrated a 1.8 fold

longer fatigue life for the Sn-Ag-Cu bumps than Sn-Pb-Ag bumps. However, when thermal cycling between 50°C and 120°C, a reversal in performance was observed; the Sn-Ag-Cu achieved only 2/3 of the fatigue life of the Sn-Pb-Ag bumps.

Data showing the life cycles under different thermal cycling conditions is shown in Figure 3.39. It can be seen that the Sn-Ag-Cu alloy had a higher lifetime deceleration rate than the Sn-Pb-Ag alloy when thermal cycling increases in severity. The fundamental rule shown in the paper is that when there are higher strains, the Sn-Ag-Cu performs better, but at low strain levels Sn-Pb-Ag is the better performer. In packages with small CTE mismatches and compliant structures, Sn-Ag-Cu it best, whilst Sn-Pb-Ag is best for stiff components with large CTE mismatches.

Ratchev et al. [54] also perform a study on the reliability of different solder connects on Polymer Stud Grid Array (PSGA) packages. Optimised versions of the packages were also tested, which used a different overmould material which provided a CTE closer to the one of the package polymer body. Devices were cycled between -40ºC to 125ºC (1 hour cycle time), until they failed.

The results, shown in Figure 3.40, suggest that the Sn-Ag-Cu solder had a much longer lifetime than the Sn-Pb-Ag solder for both the normal and optimised packages.
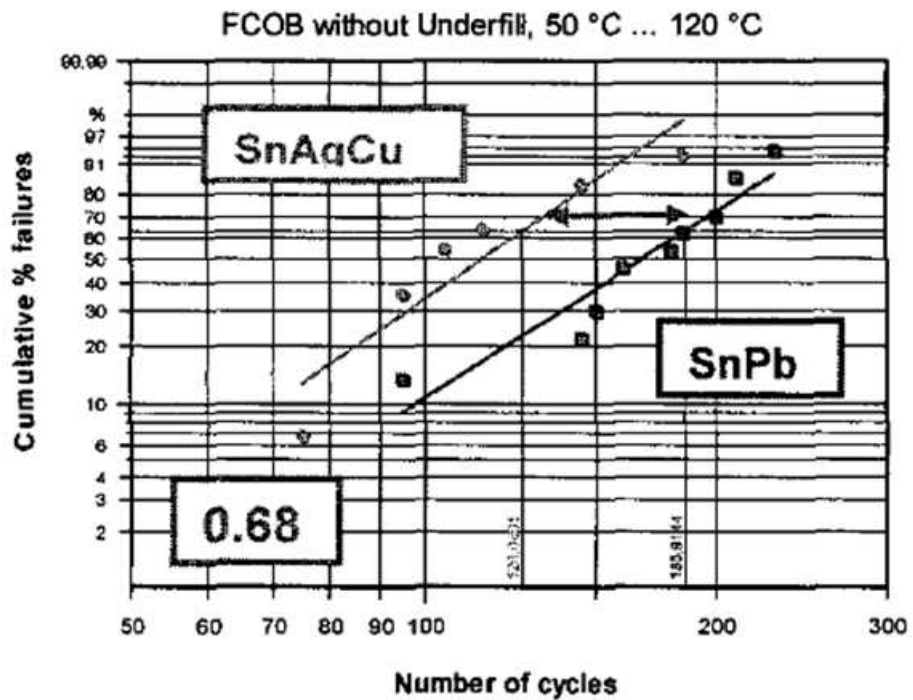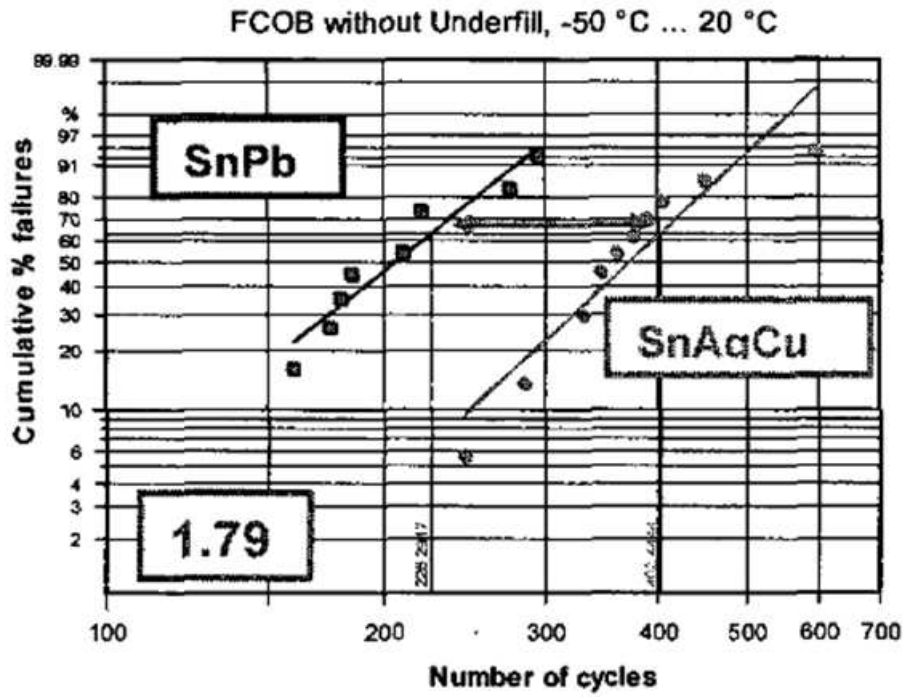
Figure 3.38: The cumulative failure of modules vs number of cycles for two different solder types. The cycling ranges were -50 to 20ºC (top) and 50 to 120ºC (bottom). The values 1.79 and 0.68 are the ratio of fatigue life of SnAgCu to SnPb for each cycling scenario. [56]
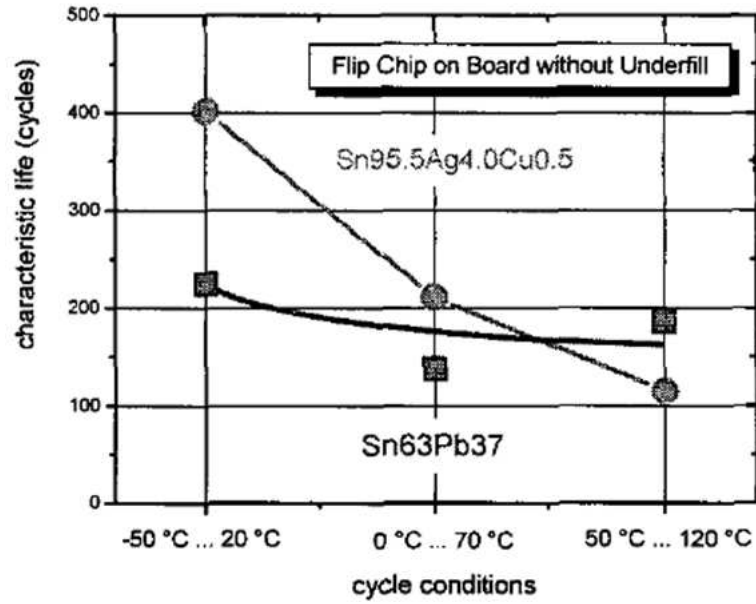
Figure 3.39: Solder joint reliability comparison of Sn-Ag-Cu alloy with Sn-Pb eutectic under different temperature cycling conditions.[56]
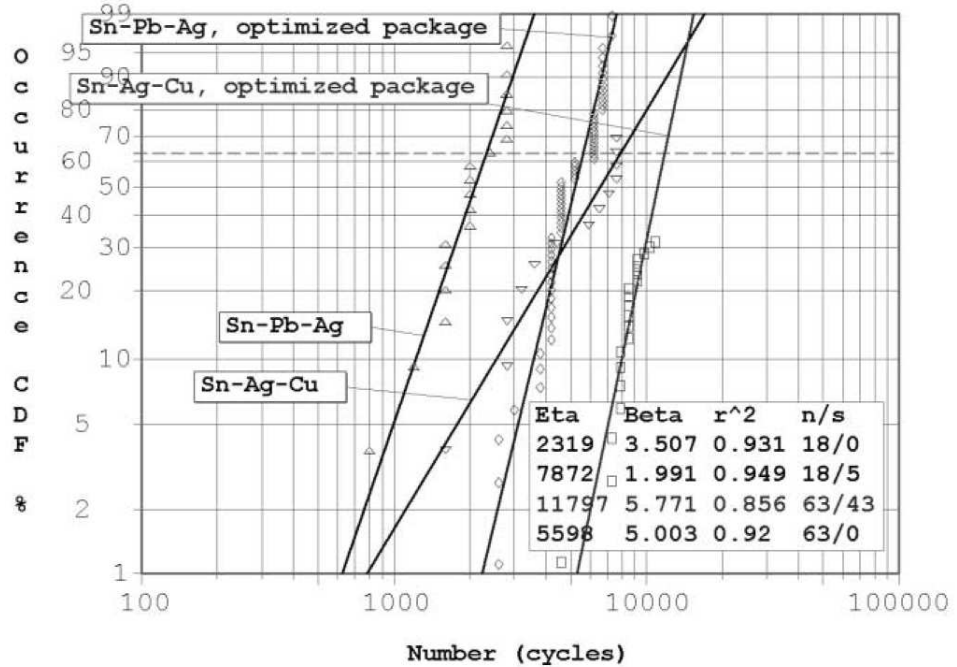
Figure 3.40: Results from thermal cycling experiments of failure percentage vs number of cycles.[54]

3.3.2 Substrate Technologies

Dupont et al. [50] examined the failure modes of power devices in avionics under stressful operating conditions, such as extreme ambient temperatures and high temperature cycles. The failures focused on were ceramic cracks in the substrate, fractures under the copper lead-frame and solder cracking. In experimental work, temperatures were cycled between -30ºC and 180ºC over a full cycle time of 7 minutes. Acoustic analysis showed that after 100 cycles, cracking was found in the ceramic layer beneath the copper metallisation as well as in the solder layer beneath the chip.

A further paper by Dupont et al. [57] attributes the fracturing of the ceramic to high variation in temperature. The paper furthers the experimental work by looking at different substrate technologies in order to evaluate their ability to operate in high temperature environments and under thermal cycling. The different substrates tested were as follows:

- DCB (Al_2O_3) : 5 different samples with and without dimples, with different copper and ceramic thicknesses

- DCB (AlN) : 4 different samples with and without dimples, with different copper thicknesses

- AMB (Si_3N_4) : 1 test vehicle with 400µm copper thickness

- DCB (AlN) : 1 sample, with aluminium metallisation

The test modules were cycled from -30ºC to 180ºC, during which the ageing of the substrate was measured using the capacitance method (see [57]). It was found that reducing the metallisation thickness can increase the number of cycles to failure. Dimples act as local thickness reductions and so provide benefits in reliability. The Si_3N_4 ceramic material used in the AMB substrate is attractive due its high tensile strength, which reduces or completely avoids fracture in the ceramic.

Guth and Mahnke [53] performed an experiment investigating the shape of the substrate in an attempt to increase reliability. The comparison made was between two DCB substrates, one of them with a modified lower metallisation to have a larger surface area than the ceramic area, as shown in Figure 3.41. The metallisation border around the ceramic layer did not have a CTE constricted by the ceramic.
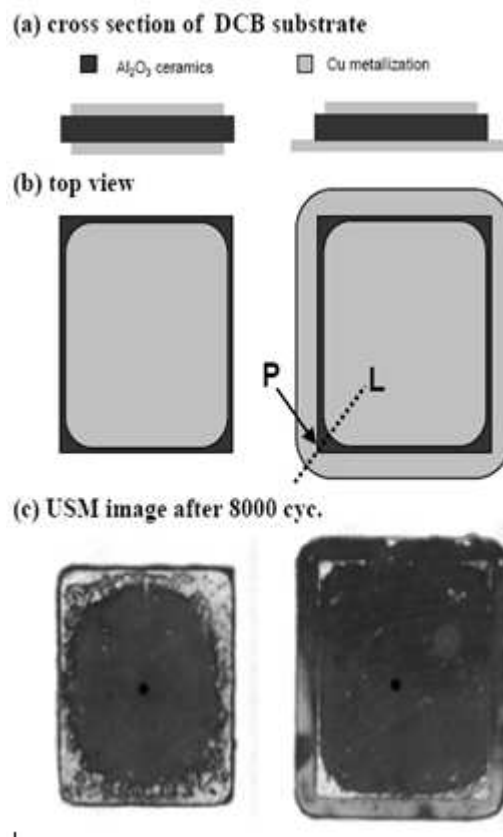
Figure 3.41: The improved reliability by use of an enlarged CU metallisation of the bottom side of the substrate.[53]

The results showed that after 8000 thermal cycles the delamination of the solder layer attaching the substrate to the baseplate for the modified substrate tile was lower than the unmodified one. Investigation into the crack formation showed a difference between the two; the crack in the modified layer began directly beneath the ceramic layer, and only propagated inwards. The unmodified substrate showed a crack that began at the edge of the solder and was therefore potentially able to propagate to occupy the entire layer.

3.3.3 Predictive Modelling of Semiconductor Reliability

Finite element (FE) analysis is being used increasingly in the design of power electronic components. The soldered joints in power electronic modules are highly important from a reliability perspective, yet highly complex, making life prediction modelling difficult. Schubert et al. [56] detail many of the problems involved in modelling thermal stresses in solder joints:

- The micro-structure, such as grain size and dispersion, can vary both in the initial state and also during service. The dependence of the micro-structure on thermo-mechanical properties is not understood very well which makes modelling it difficult.

- The complicated boundary behaviour between the solder and bonded material provides further things to consider. When soldering onto metallisations, soft solders form intermetallics, which grow over time and temperature. The soldering process itself can create anomalies due to different cooling rates, excessive voids, brittle phase formation and concentration gradients of metallurgic composition.

- Properties of the solder, such as Young's Modulus, are nonlinear and temperature dependent.

- The number of possible failure mechanisms is extensive, and can be due to many things such as grain/phase coarsening, grain boundary sliding, matrix creep, and micro-void formation.

Sommer et al. [58] use FE analysis for a high-power IGBT module. The visco-plastic deformation in the solder layers can be evaluated either by the distributions of equivalent plastic strain or by visco-plastic energy consumption. Both these fields can be provided by post-processing tools in commercial FE packages.[58]

The FE results showed that the solder interconnects tend to accumulate stresses at the corners. Delamination due to thermal loading is expected to start there, and the

FE results showed good agreement with experimental results using ultra-sonic imaging which can be seen in Figure 3.42.
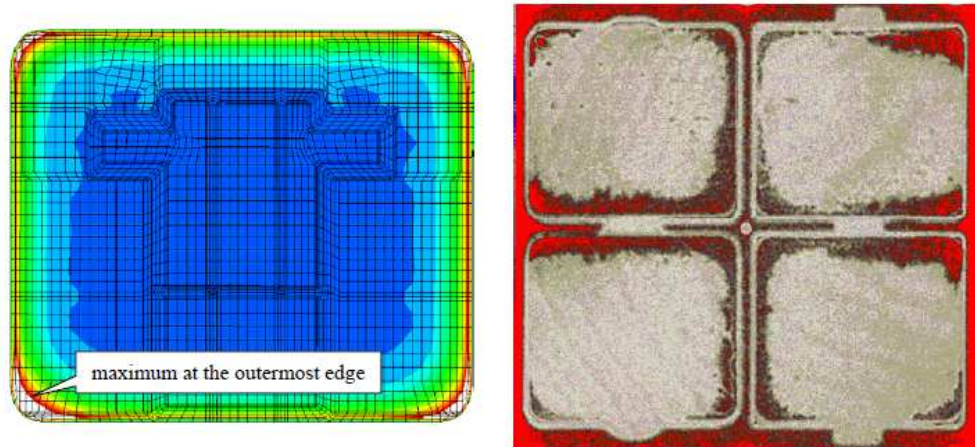
Figure 3.42: *Left:* Equivalent visco-plastic strain distribution in the solder connect from FE analysis. *Right*: Delamination in the solder connect of a 4-device module, witnessed using ultra-sonic imaging.[58]

The identification of failure modes and mechanisms for power electronics modules using FE analysis is addressed by Bailey et al. [59]. The different failure modes are discussed, and the methodology for modelling solder joint fatigue is discussed. Lu et al. [60] use FE analysis to model the propagation of a crack in a solder joint. The different ways of modelling the crack, and ways of determining the lifetime of the module are evaluated. A lifetime prediction model of a solder interconnect is presented in another paper by Lu et al. [61]. FE analysis is combined with statistical techniques to approximate a lifetime calculation based on the linear damage accumulation rule.

3.4   Composite Materials and Solder Alternatives

3.4.1   Thermal Management Materials

A variety of materials exist which are designed to provide a high thermal performance. In addition to maximising factors such as thermal conductivity, the CTE of materials is also considered when designing and choosing such materials in order to decrease thermal stresses and hence increase reliability.

Zweben [62] presents an article summarising 'revolutionary' new thermal management materials.

Reinforcement	Matrix	Inplane Thermal Conductivity (W/m K)	Through-Thickness Thermal Conductivity (W/m K)	Inplane CTE (ppm/K)	Specific Gravity	Specific Inplane Thermal Conductivity (W/m K)
	CVD Diamond	1100-1800	1100-1800	1-2	3.52	310-510
-	HOPG	1300-1700	10-25	-1.0	2.3	565-740
-	Natural Graphite	150-500	6-10	-	-	-
Cont. Carbon Fibers	Copper	400-420	200	0.5-16	5.3-8.2	49-79
Cont. Carbon Fibers	Carbon	400	40	-1.0	1.9	210
Graphite Flake	Aluminum	400-600	80-110	4.5-5.0	2.3	174-260
Diamond Particles	Aluminum	550-600	550-600	7.0-7.5	3.1	177-194
Diamond & SiC Particles	Aluminum	575	575	5.5	-	-
Diamond Particles	Copper	600-1200	600-1200	5.8	5.9	102-203
Diamond Particles	Cobalt	>600	>600	3.0	4.12	>145
Diamond Particles	Silver	400->600	400->600	5.8	5.8	69->103
Diamond Particles	Magnesium	550	550	8	-	-
Diamond Particles	Silicon	525	525	4.5	-	-
Diamond Particles	SiC	600	600	1.8	3.3	182

Figure 3.43: A list of high thermal conductivity materials. Specific thermal conductivity is defined as *'thermal conductivity / specific gravity'* [62]

Due to their natural high thermal conductivity, diamond and graphite feature heavily in the list of materials in Figure 3.43. Diamond has always been a desirable material. Chemical vapour deposition (CVD) diamond is a synthetic material that can be processed in thin films. Thermal enhancement of electronics packages using CVD diamond substrates has already been tested by Fabis [63]. The cost of CVD diamond limits its potential as a serious material, along with the limitation that it can only be produced in thin films.

Highly Oriented Pyrolytic Graphite (HOPG) is a pyrolytic graphite with an angular spread of the crystallites of less than 1 degree, usually produced by stress annea-

ling. The careful orientation of the graphite strands allows an incredibly high thermal conductivity (up to 1700W/mK [62]) along the inplane. The drawback is that it has a very poor thermal conductivity perpendicular to the strand plane, typically 100 times less than the inplane conductivity.

Using diamond in a metal matrix is also a developing idea. Adding diamond particles to a copper matrix can increase the thermal conductivity from that of copper (~385W/mK) up to 1200W/mK, depending on the copper-diamond ratio. The benefits of using diamond in this form to boost thermal performance is the relative versatility of manufacture and machining compared to HOPG and CVD diamond. It also does not suffer from being highly anisotropic, such that orientation need not be considered.

Copper diamond has already been incorporated into electronic packages, and was developed by Davidson et al. [64] as substrates for electronic components in 1995.

Chen et al. [65] used copper diamond as the wick structure in a vapour chamber heat pipe. Figure 3.44 shows the thermal resistance of the heat pipe is reduced dramatically with the use of copper diamond (DiaVC) heat pipe wicks with different copper-diamond ratios.

Drofenik and Kolar [66] investigate the use of composite materials as air cooled heat sinks. A performance parameter Cooling System Performance Index (CSPI) is defined which takes into account the thermal resistance of the heat sink and the volume of the heat sink and fan (units W/K.litre). Figure 3.45 shows a comparison of the CSPI for four different heat sink materials (where AL-MM & Diamond is a aluminium with diamond particles). It is proven that although aluminium is a good, light, low cost option, advanced thermal materials might provide a significant boost in thermal performance.
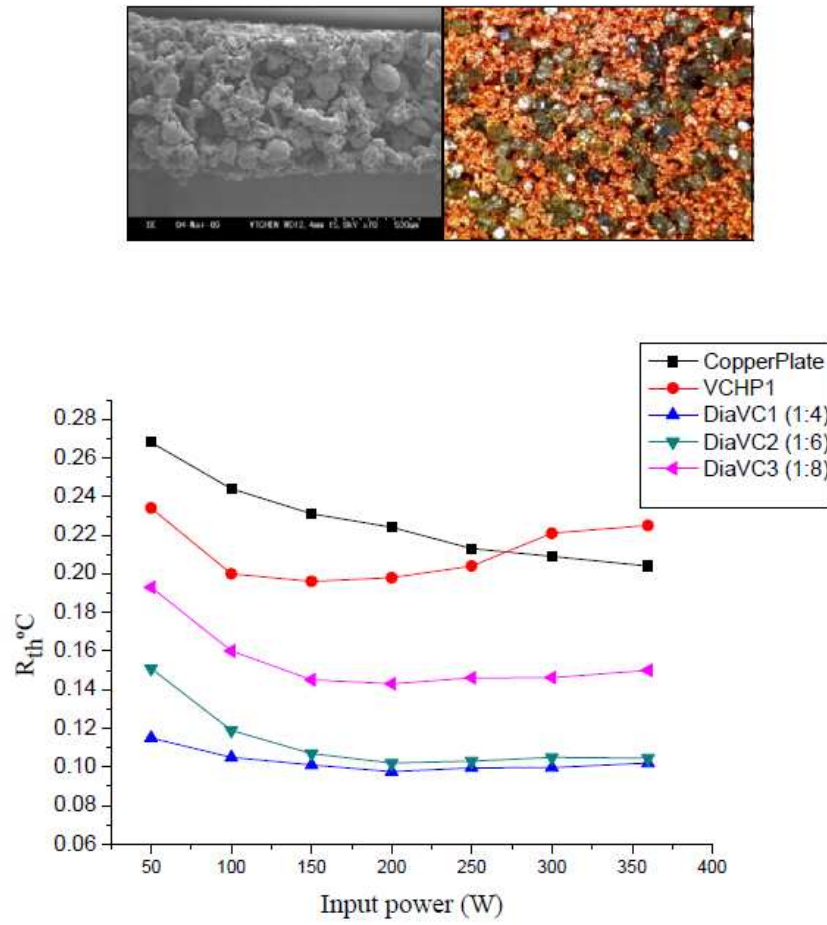
Figure 3.44: *Top:* A magnified cross-sectional photograph of a copper diamond composite. *Bottom:* Effect of different diamond-copper composition volume ratio on the thermal resistance of the vapour chamber heat pipe.[65]cover
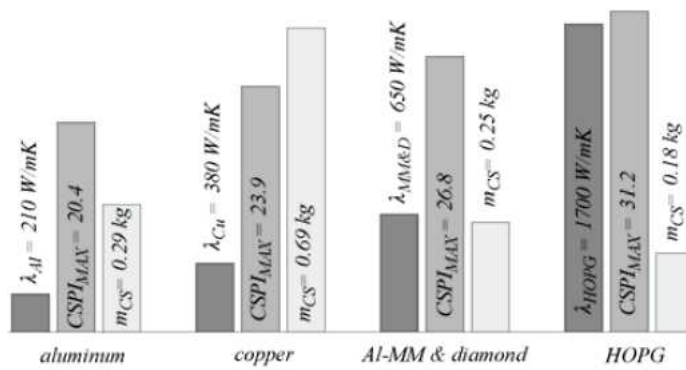
Figure 3.45: The CSPI of four different heat sinks for a 10kW fan. (λ signifies thermal conductivity, m_{CS} signifies mass of cooling system).[66]

3.4.2 Novel Solders and Solder Alternatives

Active solder materials are being developed for use in semiconductor packaging and thermal management systems.[67] The requirement for low thermal resistance is driving the development of solders for direct attachment to silicon and other materials that conventional solder does not adhere well to. The principles behind active solder joining are based in the use of Sn-Ag or Zn-Al-Ag based solders with active rare earth elements such as Ti, Ce and Ga. They have the ability to wet and adhere to most metals and ceramics.[67]

Smith and Redd [67] present one particular active solder. The use of the solder is described, explaining how the rare earth metals allow strong solder bonds to previously un-bondable materials by use of the fluxless process. The active solder is shown to be able to make silicon die attachment improvements over other epoxy and solder joints. It has also bonded a silicon die to a heat spreader, and withstood thermal cycling stresses. Other surfaces to which it has been bonded are aluminium and Al-Si-C.

Narumanchi et al. [68] consider the suitability of different thermal interface materials for power electronic applications. Despite the relatively thin thickness of the interface, the importance of having a good joint is demonstrated. The temperature distribution through the package thickness is simulated for different thermal interface materials of varying thermal resistance. A baseline material with a resistance of $100mm^2K/W$ was simulated, with three other interface resistances 5, 10 and 20 times smaller. The results at the top of Figure 3.46 emphasise the importance of getting a good thermal contact at component interfaces.

The thermal resistance of different commercial thermal greases are also examined. These thermal resistances are also split into contact and bulk resistance for a 75micron bond-line thickness. Thermal resistances as low as $33mm^2K/W$ can be achieved with the thermal grease for a bond-line thickness of $100\mu m$. However, the calculated thermal resistance at which the thermal interface stops being a bottleneck is $3mm^2K/W$, which is not obtainable with the thermal greases.
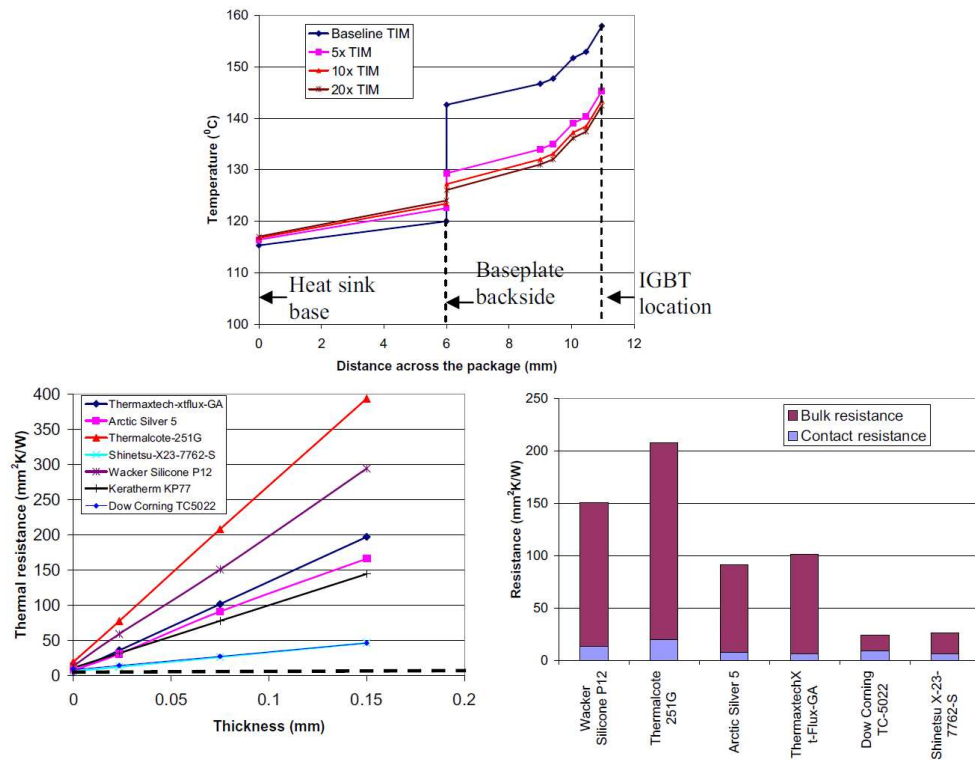
Figure 3.46: *Top:* The temperature through the package when different thermal interface resistances are modelled. *Bottom Left:* Thermal resistance of thermal interface materials with increasing joint thickness. *Bottom Right:* The breakdown of thermal resistance into bulk and contact resistance.[68]

Silver filled pastes are frequently used in electronic packages as they are good conductors of heat and electricity, and require fewer processing steps than a typical solder process [69]. Reliability issues with silver paste are widely reported and attempts have been made to find the optimum silver concentration in terms of reliability and thermal and electrical performance. Opitla and Sinclair [70] investigated the reasons for electrical failure in devices using silver paste. It was found that a major factor in the electrical performance of silver paste is due to an insulating layer forming around the silver flakes.

3.5   Chapter Conclusion

The suitability of various electronic cooling methods for reducing thermal transients has been assessed. Forced liquid convection methods such as spray cooling, jet impingement and micro-channel cooling are often used for cooling semiconductors effectively during steady state operation. However, the suitability of these methods for reducing transient temperatures is limited as they are difficult to apply on or very close to the die. Current double sided die cooling designs also mostly focus on improving the steady state die temperature. A stack design by Ngo et al. [29] was examined which specifically aimed to reduce the transient die temperature. This design adopted a double sided die cooling technique, but results showed improvements for pulses greater than 50ms in length. The design was not shown to be effective at very short transients.

A review of transient thermal modelling techniques presented different methods for predicting the temperature inside a die during heat pulses. Analytical methods, such as those presented by Carslaw and Jaeger [31] and Clemente [32], allow die temperatures to be calculated quickly and simply. However, the assumptions made in each mean they are not suitable for applying to a wide range of stack structures. The short pulse duration for which each is claimed to be valid provides a further limitation to their use. The use of RC networks has shown to provide accurate die temperature predictions over the entire transient period to steady state. Networks can be formed from either experimental data (Foster networks) or from the physical geometry of the stack (Cauer networks), although the construction of Cauer networks for 3D structures is not a simple task. Further drawbacks to the RC network methods is the computational power required to calculate multiple equations at each time step required. It is also not possible to directly calculate the temperature at a given transient time; the temperature history of the die must be calculated in small time steps.

Effects of solder voids on the die temperature and life-time were investigated. The

optimum solder thickness for maximum reliability during thermal cycling was discussed, with slightly contradicting results from Guth and Mahnke [53] and Hayashi et al. [52]. However, both agreed that a constant solder thickness across the joint was crucial. Results which investigated effect of the solder material on the reliability were also discussed, concluding that different results show that different solder compositions perform better under different loading conditions, e.g. the temperature range of the thermal cycling. The impact of substrate technologies on die reliability was also considered, as well as an overview of methods used to predict the reliability of different stack structures under thermal loads.

Composite materials for use in thermal applications have also been discussed. In particular, copper diamond, CVD diamond and HOPG all show very attractive thermal properties which could potentially be used. Davidson et al. [64] have used copper diamond in electronic packages to improve cooling, and Chen et al. [65] adopted it within heat pipes to improve thermal performance. HOPG has been demonstrated by Drofenik and Kolar [66] to improve the thermal performance of an air-cooled heat sink.

Novel solder alternatives have been considered, analysing their thermal and reliability benefits. This focuses on 'active' solder materials and silver filled thermal greases. These materials show promise for use in situations where solder cannot be used,

Chapter 4

Modelling Heat Conduction

This chapter will discuss various transient heat conduction modelling methods. It will present analytical methods which allow the temperature at isothermal and constant heat flux surfaces to be calculated trivially. Manipulation of the constant heat flux surface analytical equation yields an important material property for this thesis, the *coefficient of heat penetration*.

Subsequently RC networks are introduced, which can be used to model more complex thermal situations, following which the differences between Foster and Cauer RC networks are identified. The characteristics of both are considered, with reference to situations where they could be applied. The construction of Cauer networks will then be described in detail, including the methodology for node discretisation, the listing of heat transfer equations and the different boundary conditions relevant to the thesis.

This leads onto methods by which the Cauer RC networks can be solved. The differences between the implicit and explicit methods will be analysed in detail. This includes the formulation of the equations and process by which they are solved. The section finishes with a section discussing which solver method is most suitable in different situations.

The heat transfer theory covered in this chapter is well established, and by no means

novel. It has been compiled from the following references: Bejan [10], Kreith and Bohn [11], Carslaw and Jaeger [31], Grigull and Sandner [71].

4.1 Analytical Equations for Unidirectional Transient Heat Conduction

Heat flow within a semi-infinite bar follows the behaviour of the early regime. The infinite length of the bar provides an unlimited depth at temperature for the heat to penetrate into. Equations can be used to calculate the temperature at any location within the bar, at any time during the heat conduction. The thermal boundary condition applied to the end determines the equation used.

4.1.1 Isothermal Surface

In this scenario, the end of the bar is raised and held at a constant temperature, T_0. The temperature at any distance x from this isothermal surface can be expressed in terms of T_0, and the initial temperature, T_i[10]:

$$\frac{T - T_0}{T_i - T_0} = \mathrm{erf}\left[\frac{x}{2(\alpha t)^{1/2}}\right] \tag{4.1}$$

where $\mathrm{erf}(x)$ is termed the error function, and can be found in tables in textbooks including Bejan [10], Kreith and Bohn [11]. The function can also be approximated by the expression:

$$\mathrm{erf}(x) \approx 1 - 1.5577 \exp\left[-0.7182\left(x + 0.7856\right)^2\right] \tag{4.2}$$

This expression is accurate to within 0.42% of the table values.

The shape of the temperature profile is independent of the initial and surface temperatures and is dependent only on the thermal diffusivity of the material. As the temperature is constrained by two limits, T_0 and T_i, and the isothermal temperature is always known, the equation for temperatures along the bar is relatively straight forward.

4.1.2 Constant Heat Flux Surface

The constant heat flux surface is a more complex situation, as the bar end is subjected to a constant heat flux, q''. As mentioned, this is the situation most relevant to modelling the heat conduction through a semiconductor device during a current overload. Unlike the isothermal surface, the heat flux surface temperature is time-dependent and unrestrained. Consequentially, the equation used to calculate the temperature within the bar is more complex as the heat input is defined, rather than temperature.

At a distance x for the heated surface after a time t, the temperature rise can be calculated[10]:

$$T_{(x,t)} - T_i = 2\frac{q''}{k}\left(\frac{\alpha t}{\pi}\right)^{1/2}\exp\left(-\frac{x^2}{4\alpha t}\right) - \frac{q''}{k}x\operatorname{erfc}\left[\frac{x}{2(\alpha t)^{1/2}}\right] \qquad (4.3)$$

where $\operatorname{erfc}(y)$ is the complementary error function which can be approximated by the expression:

$$\operatorname{erfc}(y) = 1 - \operatorname{erf}(y) \approx 1.5577\exp\left[-0.7182\left(y + 0.7856\right)^2\right] \qquad (4.4)$$

The hottest point at any time is at the surface where the heat flux is applied, $x = 0$. At this location, the equation simplifies to:

$$T_{(0,t)} - T_i = 2\frac{q''}{k}\left(\frac{\alpha t}{\pi}\right)^{1/2} \qquad (4.5)$$

4.2   The Coefficient of Heat Penetration

The temperature rise at the surface for a constant heat flux, eq. (4.5), is not only dependent on the magnitude of the heat flux but also on the material it is travelling through, represented by the thermal conductivity and thermal diffusivity terms in the equation. Rearranging and grouping these two terms, a single expression for the thermal properties of the conducting material can be derived:

$$T_{(0,t)} - T_i = 2q'' \sqrt{\frac{t}{\pi}} \sqrt{\frac{\alpha}{k^2}} \tag{4.6}$$

$$T_{(0,t)} - T_i = 2q'' \sqrt{\frac{t}{\pi}} \frac{1}{\sqrt{k\,\rho\,c_p}} \tag{4.7}$$

$$T_{(0,t)} - T_i \;\propto\; \frac{1}{\sqrt{k\,\rho\,c_p}} \tag{4.8}$$

The density, thermal conductivity and specific heat capacity are all equally important in determining the temperature rise at the surface where the heat flux is applied. The greater the product of these properties, the lower the temperature rise at the surface during the early regime, as the product of the properties determines how much thermal capacitance is available to the heat at any time.

This property is of high relevance for the problem being solved. It has been mentioned in existing literature by Grigull and Sandner [71], and is termed the *coefficient of heat penetration*:

$$Coefficient\,of\,Heat\,Penetration\,(CHP) = \sqrt{k\,\rho\,c_p} \tag{4.9}$$

In the constant heat flux scenario, eq. (4.7) shows that the CHP is a contributing factor to the temperature rise at the heat flux surface. The CHP can also be applied

to the isothermal surface scenario, where the corresponding instantaneous heat flux formula is[10]:

$$q''(t) = -k \frac{(T_i - T_0)}{(\pi \alpha t)^{1/2}} \tag{4.10}$$

This can be rearranged to produce an equation for the instantaneous heat flux in terms of the coefficient of heat penetration:

$$q''(t) = -\frac{(T_i - T_0)}{\sqrt{\pi t}} \sqrt{k \, \rho \, c_p} \tag{4.11}$$

The CHP is an important property during the early regime of transient heat conduction, and is therefore of high importance for this thesis. Materials with a high CHP are desirable in stack-ups in order to reduce the temperature inside the heat generating die. Heat passing through these materials will do so at a much quicker rate than it does through materials with lower CHP values, allowing the heat to be taken away from the die quicker, thus reducing the temperature rise.

4.3 Modelling Transient Heat Conduction Problems

The calculation of transient temperatures becomes less trivial when the geometry extends beyond the semi-infinite case. When heat conduction is multi-directional or through compound layers, temperatures cannot be calculated directly using a simple set of equations. Other numerical techniques must be adopted to calculate temperatures in these cases.

4.3.1 RC Networks

Resistance - Capacitance networks (RC networks) provide a basis for calculating transient temperatures for more complex geometries. RC networks consist of a series of nodes that are used to represent the thermal resistance and capacitance of the structure. Each node corresponds to a section of the structure. The temperature gradient inside each section is considered to be zero. The heat transfer between adjacent nodes can be calculated during a time interval, allowing the resultant temperature change to be calculated at each node due to the net heat gain.

There are two common types of RC networks, the Cauer network and the Foster network. Although both networks are used to represent the resistance and capacitance values of the structure, their topology and construction are different which results in different resistance and capacitance values. A diagrammatic comparison of the two types of network is depicted in Figure 4.1.

For any given system, either circuit can be used to represent the transient thermal response. In the example shown in Figure 4.1, the circuits represent the thermal response of the junction temperature of a semiconductor die due to a power stimulus. Both circuits can be used to calculate T_J, although their construction is very different.

4.3.1.1   The Foster Network

The product of each RC set in the Foster network provides a time constant, τ. These time constants are combined to form a single equation for T_J, in the form:

$$T_J(t) = \sum_{i=0}^{i=N} T_i \left[1 - \exp\left(\frac{-t}{\tau_i}\right) \right] \qquad (4.12)$$

where N is the number of RC terms in the network. In eq.(4.12), the time constants are in an exponential form, meaning that each time constant becomes dominant at different times during the transient phase.

Although a single equation for the temperature of T_J can be formulated from the Foster network, the temperatures at the internal nodes are not calculated. This is because the nodes of the Foster network have no relevance to the physical model it represents - the only node which represents part of the structure is the node for the junction.

Capacitances are used only to connect adjacent nodes; there is no capacitance between the nodes and ambient. This is an accurate representation for an electrical circuit representation as it allows positive and negative current flow on either side of the capacitor. However, no equivalent negative flow exists in the thermal analogy and so capacitors must be connected to the ground to ensure current flow on just one side.

For this reason, the nodes and RC terms in a Foster network have no physical meaning. They are simply a series of time constants that constitute the exponential terms that describe the junction's thermal response to a heat input. As there is no physical correlation, the nodes do not represent specific locations in the system, and therefore cannot represent temperatures at any location.

The Foster network is the preferred network for electro-thermal simulations. These models tend to be only concerned with the junction temperature, which can be calculated reasonably easily once the time constants have been extracted. The time

constants (and RC terms) for a system are typically found by fitting a curve to the measured transient thermal response from and experimental method. Time constants can be extracted from the equation for the fitted curve. Once this is done, calculating the RC terms is trivial.

4.3.1.2   The Cauer Network

The Cauer network is based on the physical structure of the system, with every capacitor connected to ground. Each resistance value represents the thermal resistance between adjacent nodes, and each capacitance value the thermal mass of a node. RC values can be calculated with knowledge of the system geometry and material properties; construction of the Cauer network does not depend on having existing thermal transient data.

Nodal temperatures are calculated from the Cauer network, allowing the temperature distribution inside the system to be obtained. An equation is applied to each node in the bar to calculate the change in temperature over a small time step. During the time-step, net heat gain from adjacent nodes is calculated, as well as heat gained due to internal heat generation.

This method provides a more detailed description of the thermal behaviour within the system during the transient. However, unlike for the Foster network, a single equation for T_J cannot be derived, leading to a longer calculation procedure if only the value of T_J is of concern. The ease of construction and physical relevance makes the Cauer network preferable for many thermal modelling situations.

4.3.2   Cauer Network Construction

4.3.2.1   Domain Discretisation

The construction of Cauer RC networks is easy if the model geometry and material properties are known, or can be assumed.
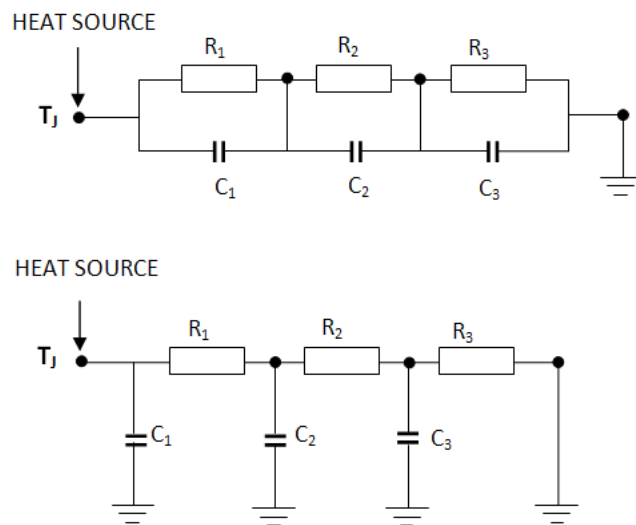
Figure 4.1: The two RC network topologies - Foster (top) and Cauer (bottom).

First, consider a model where heat flow is in just one direction. The domain is discretised with nodes, each node representing a lump of the domain. The temperature within the model is calculated at the nodes, and the number of nodes should take into account the desired resolution of the temperature profile. Solution accuracy is also dependent on the number of nodes used. However, solution calculation time increases with the number of nodes, meaning that calculation time and solution accuracy must be traded-off. This is discussed in more detail in Section 5.3.4.

Figure 4.2 shows an example of a model bar which is discretised with nodes. The section of the bar that each node represents is outlined. The thermal resistance between each adjacent node pair must then be calculated, along with the thermal capacitance of each node. Part of the resultant Cauer network is also shown at the bottom of Figure 4.2.

A Cauer RC network can also be created for 2D heat flow problems. Once the 2D model domain is created, it is discretised in both dimensions. Nodes in the resultant grid have up to four adjacent nodes between which the thermal resistance must be calculated. An example of the discretisation of a 2D model domain can be seen in

Figure 4.3. A typical node is shown in detail, with the thermal resistance between the four adjacent nodes.
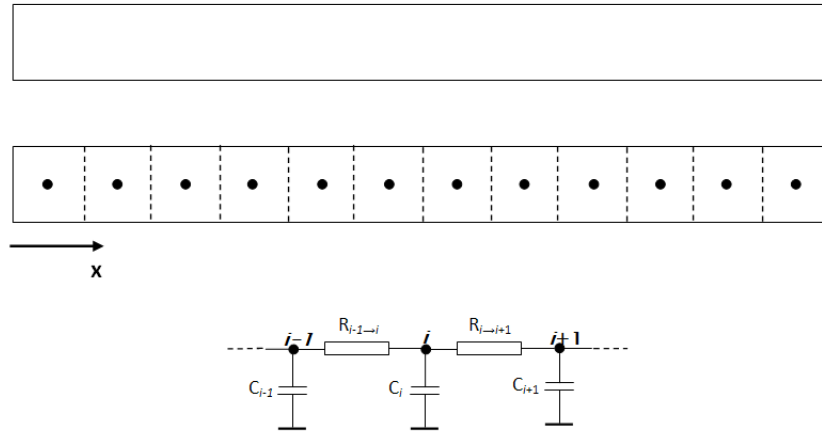
Figure 4.2: The formulation of a Cauer RC Network to represent a model with 1D heat flow.
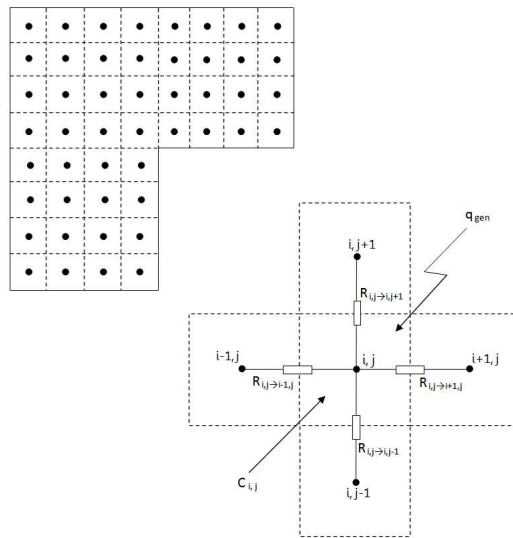
Figure 4.3: The formulation of a Cauer RC network from a 2D model domain.

4.3.2.2 Equations for Thermal Resistances and Capacitances in the Network

The resistance and capacitance values for each node must be calculated once the nodal grid has been constructed. This can be done directly from the geometry and material properties through which the heat is being conducted. A 1D or 2D network

can be used to represent different physical structures by using different equations to formulate the thermal resistances and capacitances.

Cartesian equations are applied to represent a linear structure, where all nodes represent volumes of constant depth. Cylindrical equations can also be applied, which represents the 1D or 2D model as a cylinder, rotated about a central axis.

The shapes that are represented by the equation groups are depicted in Figure 4.4. The relevant equations for each of these scenarios are shown in the Table 4.1.

	1D Cartesian	1D Cylindrical	2D Cartesian	2D Cylindrical
Resistance (x or r)	$R_x =$ $\frac{\triangle x}{k\,A_x}$	$R_r =$ $\frac{ln(r_o/r_i)}{2\,\pi\,k}$	$R_x =$ $\frac{\triangle x}{k\,A_x}$	$R_r =$ $\frac{ln(r_o/r_i)}{2\,\pi\,k\,\triangle y}$
Resistance (y)	—	—	$R_y =$ $\frac{\triangle y}{k\,A_y}$	$R_y =$ $\frac{\triangle y}{k\,A_y}$
Capacitance	$C =$ $\rho\,V\,c_p$	$C =$ $\rho\,V\,c_p$	$C =$ $\rho\,V\,c_p$	$C =$ $\rho\,V\,c_p$
Area (x)	unit area $= 1$	unit area $= 1$	$A_x =$ $\triangle y$	—
Area y	—	$A_y =$ $\pi(r_o^2 - r_i^2)$	$A_y =$ $\triangle x$	$A_y =$ $\pi(r_o^2 - r_i^2)$
Volume	$V =$ $\triangle x$	$V =$ A_y	$V =$ $\triangle y\,\triangle x$	$V =$ $A_y\,\triangle y$

Table 4.1: The Cartesian and Cylindrical thermal resistances and capacitances equations used for the thermal networks.[10]
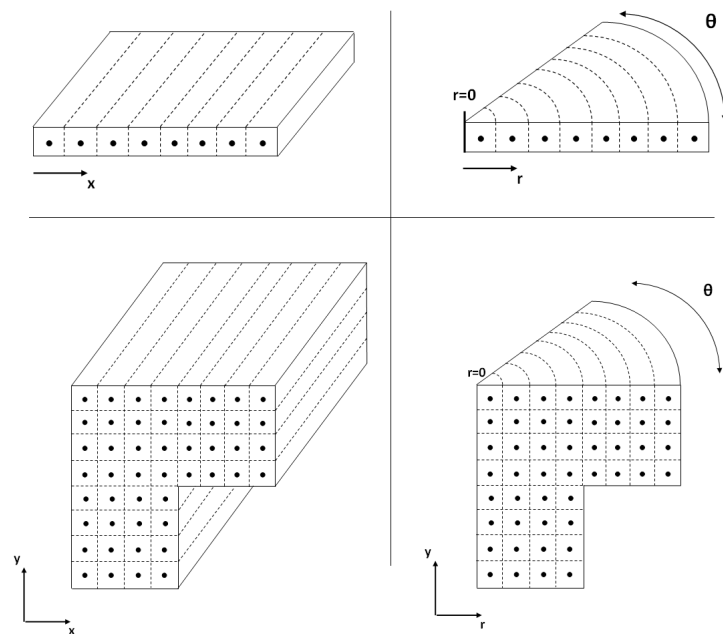
Figure 4.4: The shapes that can be represented by using the different sets of equations: Cartesian (left) and Cylindrical (right).

4.3.3 Boundary Conditions

The surfaces of the thermal resistance networks must have a thermal boundary condition attached to them. The nodes adjacent to a surface with an applied boundary condition need to be treated appropriately to simulate the correct heat transfer across the boundary. 'Dummy' nodes can be used to represent the temperature of an ambient fluid with which the solid exchanges heat.

The different boundary conditions that can be applied, and the alterations required to the nodes adjacent to the boundary in each case, are shown in Figure 4.5. Table 4.2 shows the the appropriate heat conduction equation for each of the boundary conditions, which are described below.

4.3.3.1 Constant temperature surface

The edge of the model has a fixed temperature surface. The thermal resistance between the nodes next to this surface and the constant surface temperature is the

thermal conductive resistance between the node and the surface.

4.3.3.2   Convective surface

Heat is transferred to ambient via convection (or radiation) from the boundary surface. The thermal resistance between the ambient temperature and a node adjacent to the surface is the sum of the conductive thermal resistance between the node and the surface *plus* the convective (or radiative) resistance between the surface and ambient.

4.3.3.3   Adiabatic surface

No heat is transferred across the boundary. The thermal resistance between the adjacent nodes of the boundary and any ambient temperature is considered infinite. No 'dummy' node is needed to represent the boundary condition, and the nodes adjacent to the boundary can be considered to exchange heat with only their neighbouring nodes within the solid.

Constant Temperature Surface
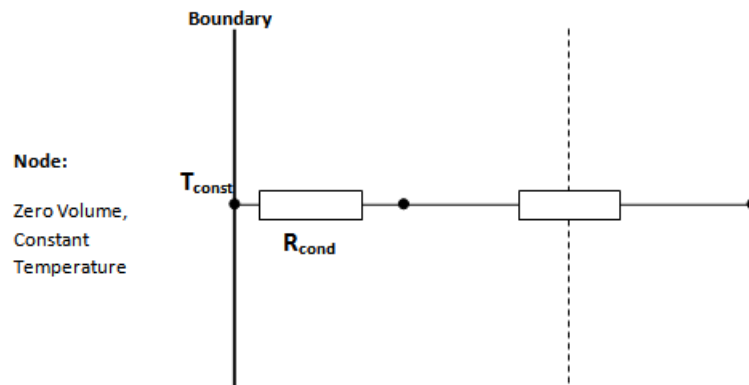
Convective Surface (Thermal Resistance to Ambient Temperature)
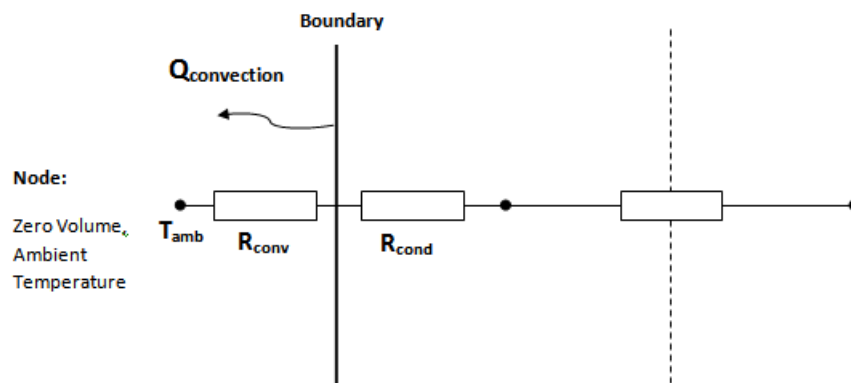
Adiabatic Surface
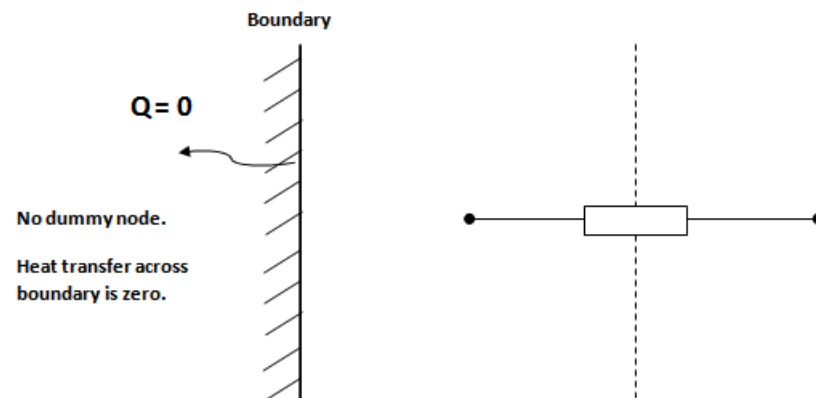
Figure 4.5: The thermal boundary conditions that can be applied and the way each is represented in a thermal resistance network.

Boundary Condition	Heat Transfer
Constant Temperature Surface (T_C)	$Q = -\frac{kA(T_1 - T_C)}{\triangle x}$
Convective Surface	$Q = -\frac{(T_1 - T_A)}{(\triangle x/kA + 1/hA)}$
Adiabatic Surface	$Q = 0$

Table 4.2: The heat transferred from the node to the boundary for each of the boundary conditions.

4.3.4 Solving Thermal Resistance Networks

A completed thermal resistance network for a system arises after the model domain is discretised and the RC values calculated. The network corresponds to the physical structure of the system, and can be used to calculate the heat flow between nodes, thus allowing the temperatures at the nodes to be found at different times during the transient phase.

The temperature profile of the time-dependent system can be calculated in small time steps, $\triangle t$, starting from an initial known temperature distribution. The temperature superscript m denotes the time step being referred to. T^m refers to the temperature at the current time step at any point, whilst T^{m+1} and T^{m-1} refer to the temperature at the next time step and the previous time step, respectively. Similarly, the location of the node being referred to in the x and y direction is denoted by the subscripts i and j, respectively.

The conduction space is divided into discrete volumes (i, j) and time frames (m). A three-dimensional array of temperature values $T_{i,j}^m$ replaces the time-dependent temperature field $T(x, y, t)$. An example of the discretisation in time and space is shown in Figure 4.6, where a node is shown at three consecutive time steps. In the frame showing the current time step, m, the four neighbouring nodes are also shown.

A heat balance equation can then be written for the node i, j, based on the net heat transfer from the adjacent nodes:

$$\rho \triangle x \triangle y\, c_p \frac{\partial T}{\partial t} = q_U + q_D + q_L + q_R \tag{4.13}$$
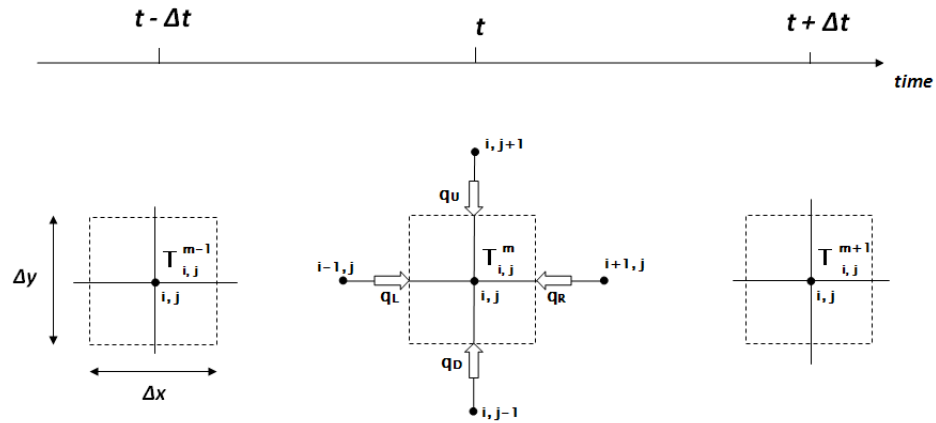
Figure 4.6: Demonstration of the discretisation of the conducting domain in space and time.

This assumes the model has a unit depth in the 3rd dimension. The heat transfer from the neighbouring nodes is labelled in Figure 4.6. The heat transfer along these paths is equal to the temperature difference divided by the thermal resistance:

$$q_U = (T_{i,j+1}^m - T_{i,j}^m)\frac{k\triangle x}{\triangle y} \tag{4.14}$$

$$q_D = (T_{i,j-1}^m - T_{i,j}^m)\frac{k\triangle x}{\triangle y} \tag{4.15}$$

$$q_R = (T_{i+1,j}^m - T_{i,j}^m)\frac{k\triangle y}{\triangle x} \tag{4.16}$$

$$q_L = (T_{i-1,j}^m - T_{i,j}^m)\frac{k\triangle y}{\triangle x} \tag{4.17}$$

Substituting each of the equations (4.14 to 4.17) into eq.(4.13) results in the following expression for the change in temperature at node i, j:

$$\frac{1}{\alpha}\frac{\partial T}{\partial t} = \frac{T_{i+1,j}^m + T_{i-1,j}^m - 2T_{i,j}^m}{(\triangle x)^2} + \frac{T_{i,j+1}^m + T_{i,j-1}^m - 2T_{i,j}^m}{(\triangle y)^2} \tag{4.18}$$

The time derivative $\partial T/\partial t$ must be approximated by means of finite difference. The present-time frame (m) can be compared to either the next time-frame $(m+1)$ or the previous time-frame $(m-1)$. These are termed forward and backward difference methods, respectively:

Forward difference:

$$\frac{\partial T}{\partial t} \cong \frac{T_{i,j}^{m+1} - T_{i,j}^m}{\triangle t} \tag{4.19}$$

Backward difference:

$$\frac{\partial T}{\partial t} \cong \frac{T_{i,j}^m - T_{i,j}^{m-1}}{\triangle t} \tag{4.20}$$

Two algorithms arise from these two options, known as the 'explicit' and 'implicit' method. The differences between the algorithms, and the way each is used to find the nodal temperatures is discussed.

4.3.4.1 The Explicit Method

The explicit method uses the forward difference time derivative. After defining separate expressions for the Fourier number in the x and y direction (as the distances $\triangle x$ and $\triangle y$ may not be equal), eqs.(4.18) and (4.19) can be combined to provide an expression for the temperature at the next time step, based on the temperatures of the nodes at the current time step:

$$Fo_x = \frac{\alpha \Delta t}{(\triangle x)^2} \tag{4.21}$$

$$Fo_y = \frac{\alpha \Delta t}{(\triangle y)^2} \tag{4.22}$$

$$T_{i,j}^{m+1} = T_{i,j}^m + Fo_x \left(T_{i+1,j}^m + T_{i-1,j}^m - 2T_{i,j}^m \right) + Fo_y \left(T_{i,j+1}^m + T_{i,j-1}^m - 2T_{i,j}^m \right) \quad (4.23)$$

As the right hand side of eq.(4.27) only contains information which is known (current node temperatures at and around node i, j and Fourier numbers based on known dimensions and material properties) the temperature at node i, j can be calculated directly. It is for this reason that the explicit method gets its name.

The explicit method requires the equation to only be applied once to each node at each time step to find the temperature distribution for the next time step. Once this has been done, the time-frame $m + 1$ becomes the time-frame m, and the calculation of the temperatures in the next time-frame (now $m + 1$) can begin in the same way.

Time Step Restriction in the Explicit Method   The explicit method provides an attractive option for solving thermal resistance networks. However, a limitation to the method exists which must be considered when constructing the model and choosing the time-step. Small time-step values must be used with the explicit method. Using time-steps that are too large causes oscillations in the solution, creating numerical instability and incorrect results.

The limit to the time-step can be found by grouping the terms for $T_{i,j}^m$ on the right hand side of eq.(4.23):

$$T_{i,j}^m (1 - 2Fo_x - 2Fo_y) \qquad (4.24)$$

In order to avoid instability, the coefficient for $T_{i,j}^m$ must greater than or equal to zero. This provides the limit:

$$Fo_x + Fo_y \leq \frac{1}{2} \qquad (4.25)$$

If $\triangle x = \triangle y$, then $Fo_x = Fo_y$ and the limit becomes:

$$Fo \leq \frac{1}{4} \qquad (4.26)$$

For uni-directional heat flow there is no variation in node position in the j direction. This allows eq.(4.27) to be rewritten excluding the references to nodes in a different j plane.

$$T_{i,j}^{m+1} = T_{i,j}^m + Fo\left(T_{i+1,j}^m + T_{i-1,j}^m - 2T_{i,j}^m\right) \qquad (4.27)$$

In this case, the limitation on the time-step is more relaxed. The extracted term for $T_{i,j}^m$ on the the right hand side of the equation which must remain positive is:

$$T_{i,j}^m(1 - 2Fo) \qquad (4.28)$$

And so the limit on the time-step size must satisfy the condition:

$$Fo \leq \frac{1}{2} \qquad (4.29)$$

The time-step for a 1D resistance network can therefore be twice the size that it can be for a 2D network. This makes the explicit scheme a popular option for solving 1D systems, but the restriction on the time-step makes it less popular for solving 2D systems. The implicit method provides an alternative solver which does not have the same restrictions on the time-step as the explicit method.

4.3.4.2 The Implicit Method

The implicit method uses the backward difference time derivative. Combining eqs.(4.18) and (4.20) produces the implicit method equation:

$$T_{i,j}^{m-1} = T_{i,j}^m + Fo_x \left(T_{i+1,j}^m + T_{i-1,j}^m - 2T_{i,j}^m \right) + Fo_y \left(T_{i,j+1}^m + T_{i,j-1}^m - 2T_{i,j}^m \right) \quad (4.30)$$

In this equation, the only known temperature is $T_{i,j}^{m-1}$, on the left hand side of the equation. Calculating $T_{i,j}^m$ is reliant on knowing the four neighbouring node temperatures at the time-frame m. As none of the values at the time-frame m are known, the equation cannot be solved directly. Instead, an iterative approach must be taken.

An initial guess of the temperatures at time-frame m must be made. The guessed values are inserted into eq.(4.30), which provides a calculated value, $T_{i,j\,calculated}^{m-1}$. There are now two values for $T_{i,j}^{m-1}$: the value calculated and the *known* value, $T_{i,j\,known}^{m-1}$. The difference between these two values is termed the residual (denoted by the symbol r_s):

$$r_s = T_{i,j\,known}^{m-1} - T_{i,j\,calculated}^{m-1} \quad (4.31)$$

If the total of the individual node residuals is below a specified error value (i.e. the known and calculated values of $T_{i,j}^{m-1}$ are all very similar) then the solution for the time-step m is found. The guessed temperature values of $T_{i,j}^m$ are correct, and the process can begin for the next time-frame.

If the total residuals are too high, the temperature values at time-frame m must be modified. This is done by adding the residual, multiplied by a relaxation factor, ω, to the current guessed value of $T_{i,j}^m$:

$$T_{i,j\,modified}^m = T_{i,j}^m + \omega r_s \quad (4.32)$$

The modified values can then be inserted into eq.(4.30) for each node, generating a new calculated $T_{i,j}^{m-1}$ value. The new residual values are calculated and compared to

the benchmark error value. This iterative process continues until the total residuals falls below the error value, allowing the process to begin for the next time step.

The flow diagram for the implicit method iteration procedure is shown in Figure 4.7.
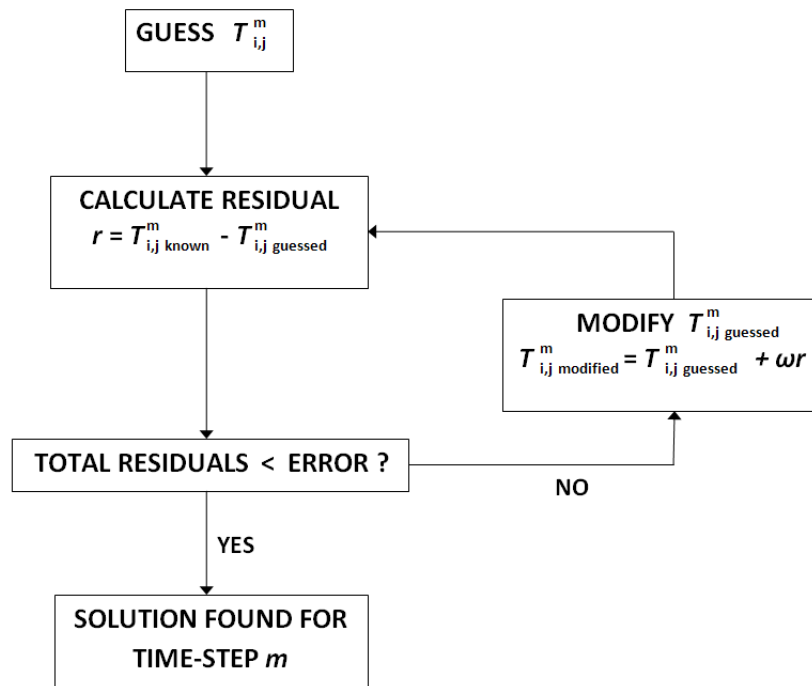
Figure 4.7: Implicit method flow diagram for the calculation T_m.

The Relaxation Factor   Unlike the explicit method, the time-step that can be used for the implicit method is not restricted. Larger time-steps can be used, allowing fewer time-steps to be taken to find the temperature history throughout a certain transient period. However, the iterative nature of the implicit method procedure means that more than one calculation may be required for each node at each time-step.

Choosing a large time-step for the implicit method can lead to oscillations and instability in the residual values. The oscillation magnitude increases, and hence the total residuals never fall below the specified error benchmark, which signifies a solution of sufficient accuracy. These oscillations can be controlled by the relaxation factor. The relaxation factor is the residual coefficient in eq.(4.32), and controls by how much the guessed value of $T_{i,j}^m$ is modified, and takes on the limits:

$$0 < \omega \leq 1$$

For larger time steps, or time steps where a large change in temperature occurs in part of the network (for example due to a large amount of internal heat generation) instability in the solution can occur and the solution will not converge. To avoid instability it is possible to decrease the time step or to decrease the relaxation factor. Decreasing the time step means more time steps are needed to cover the entire solution time, increasing the computational time. However, if the relaxation factor is reduced then the time step may not need to be reduced. Decreasing the relaxation factor increases the number of iterations required at each time step, as the iterative solution is approached in more iterations. This is often the best solution as the increased calculation time imposed is less severe than that of decreasing the time step.

4.3.4.3 Solver Method Selection

When choosing a solver method, the main consideration tends to be be computational time, as it is desirable to arrive at the solution in the shortest time possible.

Generally, the explicit method is preferred for solving 1D problems. The explicit solution requires only one calculation per node at each time-step. Thus, it usually provides a solution quicker than the implicit method, which may require any number of iterations per node at each time-step.

The time-step limitation is a lot more restrictive for 2D problems. For a grid of the same size, the time-step for a 2D problem must be half of that in the equivalent 1D problem. For this reason, implicit method is very often chosen in order to use larger time-steps. Although the iterative procedure involved in the implicit method leads to many more calculations being required at each time-step, the number of time-steps are fewer. More often than not, the total solution time is shorter when using the implicit method, provided the relaxation factor is set to an appropriate value.

The relationship between the time-step and solution accuracy must also be considered. As the explicit method has a restriction on the time-step size, the accuracy of the solution is less dependent on the time-step chosen. In other words, provided the time-step is sufficiently small to avoid numerical instability, the solution can normally be considered suitably accurate. Care must be taken when choosing a time-step for the implicit method. Much greater time-steps can be chosen when using the implicit method which will provide a solution. However, the accuracy of solutions calculated using time-steps towards the upper time-step limit can usually be improved by decreasing the time step.

4.4   Chapter Conclusion

Analytical equations have been presented, which provides a method for the temperatures inside a 1D bar to be calculated trivially. Although these equations are limited to non-compound layers and heat transfer in just one direction, analysis of them provides the definition of the material property coefficient of heat penetration (CHP). The CHP is shown to be highly important for transient heat conduction in constant heat flux scenarios, such as this thesis. Materials with high CHP values are most effective at reducing the maximum temperature whilst the heat conduction through them is behaving according to the early-regime.

RC networks are introduced as a method of modelling more complex heat conduction scenarios, such as multi-dimensional problems or those with compound layers. The Foster network is preferred in electro-thermal simulations, as they calculate only one temperature and tend to be easier to calculate as less terms are needed. However, their physical non-sense means they must be determined from existing transient temperature responses of the structure.

Cauer networks represent physical structures, and can be created based on the structural dimensions and properties of a system. The table of equations which can be applied to solve the networks demonstrates that shapes can be represented either linearly or axisymmetrically. Different boundary conditions can be simulated with ease by modifying the heat equation at the surfaces.

Two different solver methods have been discussed: the explicit and implicit methods. The explicit method allows temperatures to be calculated with one calculation at each time-step, although a restriction on the time-step means small time-steps must be taken. The implicit method uses an iterative procedure which requires more than one calculation process at each time-step. However, larger time-steps may be taken. The explicit solution is usually preferred for 1D problems, and the implicit for 2D and 3D problems.

Chapter 5

Numerical Modelling of Heat Sinks

The numerical modelling chapter describes the modelling approach taken to assess the benefits of using a heat sink to reduce the temperature of devices. It is of primary interest to reduce the temperature for pulses between 100µs and 100ms in length. The setup of the numerical models and the method used to solve them is described. The modelling results are then presented.

Initially, the two model variations are presented - with and without a heat sink. Comparing the die temperatures from each model under the same conditions provides a 'measure of performance' which is used to measure the heat sink effectiveness.

The methodology of creating the 1D thermal resistance models is then described. Dimensions, boundary conditions and details of the heat generation used are listed. Node density is discussed for the descritisation process, and an appropriate number of nodes is selected. Comparisons of sample results from the solver program, a C++ program written for the thesis, are compared to those from commercial software to validate its accuracy.

Modelling results from the 1D networks are then presented, starting off with a heat sink material comparison to ascertain which heat sink materials are most effective

during the transient. This is followed by a sensitivity analysis to assess how the heat sink is affected by changes to the modelling assumptions. A wide range of parameters are investigated including the thickness of the different stack-up components, the heat sink thickness and the location inside the die where the heat is generated. An analysis is performed for each part of the sensitivity analysis, relating the results to heat conduction theory.

5.1 Device Structures

5.1.1 Traditional Device Structure

Power electronic devices are arranged in a stack-up, which provides structural support, electrical isolation and effective cooling to the device. The main components of the stack-up are the device (also known as the die), the substrate and the baseplate, as shown in Figure 1.1. These components are joined by solder layers. A thin aluminium layer (the cathode) is deposited on top of the die which allows electrical connections to be made by bonding aluminium wires to the cathode.

In this configuration, the heat generated in the die is conducted down through the stack to the bottom baseplate surface, which is cooled. The other surfaces have negligible heat transfer across them. A typical device stack-up is shown on Figure 1.1. The thickness and constituent material of each of the layers is listed in Table 1.1.

A time-constant was defined in eq.(2.18), which specifies the time taken for a part of the structure at a distance L from the heat source to noticeably affects the heat source surface:

$$\tau = \frac{\pi}{4} \frac{L^2}{\alpha} \tag{5.1}$$

The equation can be used to calculate the time-constant of each layer in the stack shown in Figure 5.1. The accumulated time-constant of each layer is the pulse dura-

tion at which that layer noticeably affects the temperature at the heat source. The calculated time-constants are shown in Figure 5.1.

The time taken for heat to fully penetrate to the bottom of the baseplate is calculated to be around 177ms, which is outside the pulse width range of interest of 100μs and 100ms. Therefore, designing an improved convection cooling system at the baseplate would not help to reduce the device temperature during short duration current surges.

Layer	Material	Thickness (mm)	Thermal Diffusivity (m^2/s) $\times 10^{-5}$	Time Constant (ms)	Cumulative Time (ms)
Die	Silicon	0.4	8.92	1.41	1.41
Solder	Sn96.5-Ag3.5	0.1	4.22	0.186	1.59
	Copper	0.3	11.6	0.611	2.21
Substrate	AlN	0.6	6.63	4.26	6.47
	Copper	0.3	11.6	0.611	7.08
Solder	Sn96.5-Ag3.5	0.1	4.22	0.186	7.27
Baseplate	Copper	5	11.6	170	177

Table 5.1: The Fourier analysis for each layer beneath the die, showing the thermal diffusivity and time taken for heat to fully penetrate each layer.

5.1.2 Device Structure with heat sink

To absorb short duration thermal transients in the order of a few milliseconds, a high thermal mass could be placed close to the part of the die generating the heat. This can be done by placing a heat sink on top of the die, where the aluminium wire bonds are usually attached. Using solder to attach the heat sink to the top of the die provides a good thermal and electrical contact. This allows the electrical current to be supplied to the device through the heat sink and solder joint.

This modification to the traditional design is shown in the Figure 5.1. The time constants of the layers between the die and the heat sink are calculated using eq.(5.1) and listed in Table 5.2. Heat can reach the proposed heat sink 190μs after the start of the pulse, enabling it to help reduce the die temperature during a short current surge.
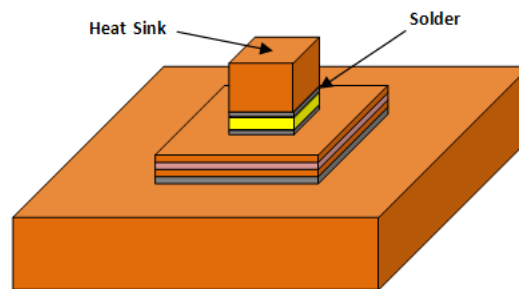
Figure 5.1: The typical device stack with the addition of the heat sink and solder layer on top of the die.

Layer	Material	Thickness (mm)	Thermal Diffusivity (m^2/s) $\times 10^{-5}$	Time Constant (ms)	Cumulative Time (ms)
Cathode	Aluminium	0.02	8.41	0.004	0.004
Solder	Sn96.5-Ag3.5	0.1	4.22	0.186	0.190

Table 5.2: The Fourier analysis for each layer above the die, showing the thermal diffusivity and time taken for heat to fully penetrate each layer.

5.2   Measure of Performance

The benefits of adding a heat sink to a device needed to be quantified. The aim is to minimise the maximum temperature inside the die, which is often the cause of device failure. The maximum temperature in the die is located at the junction between the oppositely doped regions. For this reason, the maximum temperature is often referred to as the 'junction temperature'. The temperature in the die during a heat pulse is dependent on factors such as: the stack-up configuration, the initial starting temperature, the amount of heat generated and the thermal boundary conditions. When evaluating the benefits of a heat sink, these factors must remain constant so they do not influence the temperature within the die.

A dimensionless parameter is defined that is the ratio of the maximum temperature rise with a heat sink to the maximum temperature rise for the same configuration without a heat sink. This parameter, termed the *Measure of Performance* (MoP), is defined as the ratio of the maximum temperature rise in the dies at a given pulse length:

$$MoP = \frac{\Delta T_{max} \; Model \; with \; Heat \; Sink}{\Delta T_{max} \; Model \; without \; Heat \; Sink} \tag{5.2}$$

Lower MoPs correlate to better heat sink performance. The percentage reduction in the maximum temperature rise due to the heat sink is directly calculated from the MoP value:

$$\% \; Reduction \; in \; max \; temp = (1 - MoP) \times 100 \tag{5.3}$$

A MoP value of 1 implies the heat sink does not affect the maximum die temperature. Values less than 1 show that the heat sink is reducing the maximum temperature in the die, and conversely a value greater than 1 implies the heat sink increases the die temperature.
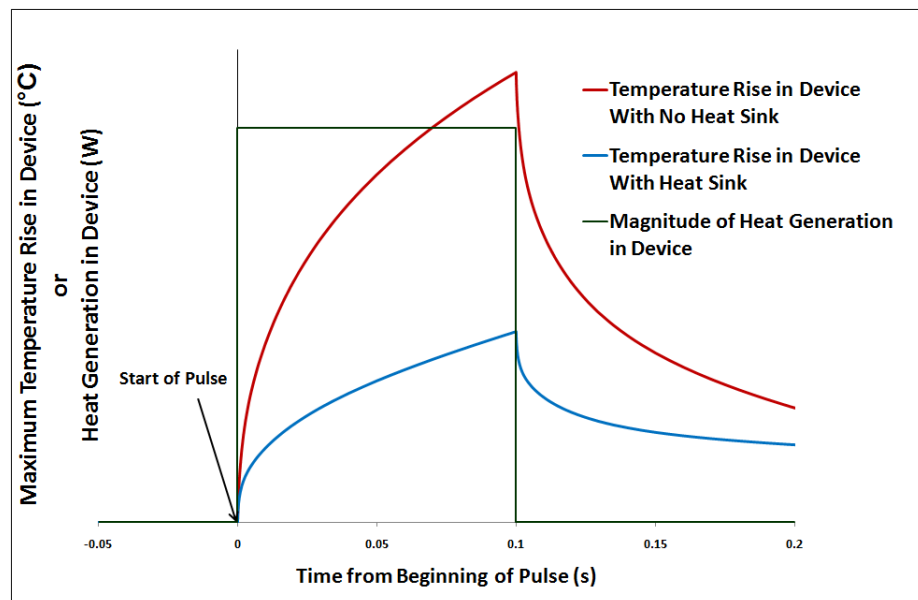
Figure 5.2: An example of the maximum die temperature rise for a stack-up with and without a heat sink during and after a constant heat pulse is applied.

5.3   1D Modelling Methodology

5.3.1   1D Model Structures

Initial modelling work was done assuming one-dimensional heat flow. This involved modelling all layers in the stack with a uniform width and adiabatic sides, such that there is a temperature variation in just one direction.

Two types of stack-ups were created - a traditional device stack (Figure 1.1) and a device stack with the addition of a heat sink (Figure 5.1). Simulating a heat pulse in both the models allowed the MoP of the heat sink to be obtained. The standard thickness of each of the layers in the stack-up corresponds to those listed in Table 5.3, where the layers shown in italics are the layers included only when a heat sink is modelled.

Variations to the heat sink (thickness and material) allowed the heat sink to be optimised. The dimensions of the layers in the stack-up were also varied independently in order to understand how each layer affects the heat sink performance.

Layer	Thickness (mm)	Material
Heat Sink	*varied*	*varied*
Solder	*0.1*	*Solder SnAg 3.5*
Cathode	0.02	Aluminium
Die (Heated Region)	0.04	Silicon
Die (Unheated Region)	0.36	Silicon
Solder	0.1	Solder SnAg 3.5
Substrate: Copper	0.3	Copper
Substrate: AlN	0.6	Aluminium Nitride
Substrate: Copper	0.3	Copper
Solder	0.1	Solder SnAg 3.5
Baseplate	5	Copper

Table 5.3: The standard thickness and material of the stack-up used for the 1D modelling.

5.3.2 Heat Generation in the Model

5.3.2.1 Location of Heat Generation

Table 5.3 shows the component dimensions in the 1D models. The die can be split into two regions - the heated region and the unheated region. The heated section accounts for the top 10% of the die, and is the region of the model where the heat generation is simulated.

Heat is not generated evenly throughout power electronics devices. The location of the heat generation depends on several things, such as the device type and structure, and the operating voltage and the current. Prediction is quite complex, although the heat generation profile can be found from modelling using modelling software such as Silvaco which models the electron flow within devices, allowing current distributions to be predicted.

The thermal modelling work was not done with a particular device in mind, and aimed to provide generic results. Generally, the heat generation tends to occur at the top of devices, near the surface where the *p-n* junction is located (see Section 2.3). It was thus assumed that the heat is generated evenly within the top 10% of the die. This provides a reasonable assumption without basing the thermal models around one particular device.

5.3.2.2 Heat Pulse Shape and Magnitude

The shape and magnitudes of a current pulse curve depends on the type of fault. Lightning pulses feature a sharp increase in magnitude leading to an early peak, which gradually decays back to the normal operating current. A standard lightning pulse used in aerospace applications known as the '10-120' lightning pulse, reaches a peak magnitude after 10μs, which decays exponentially so that after 120μs the pulse is at 50% of the peak magnitude, before reaching approximately zero after 1ms. The shape of the pulse is shown at the top of Figure 5.3.

There are potentially a range of different pulse shapes and for simplicity, a constant magnitude heat pulse was also simulated, as shown at the bottom of Figure 5.3.
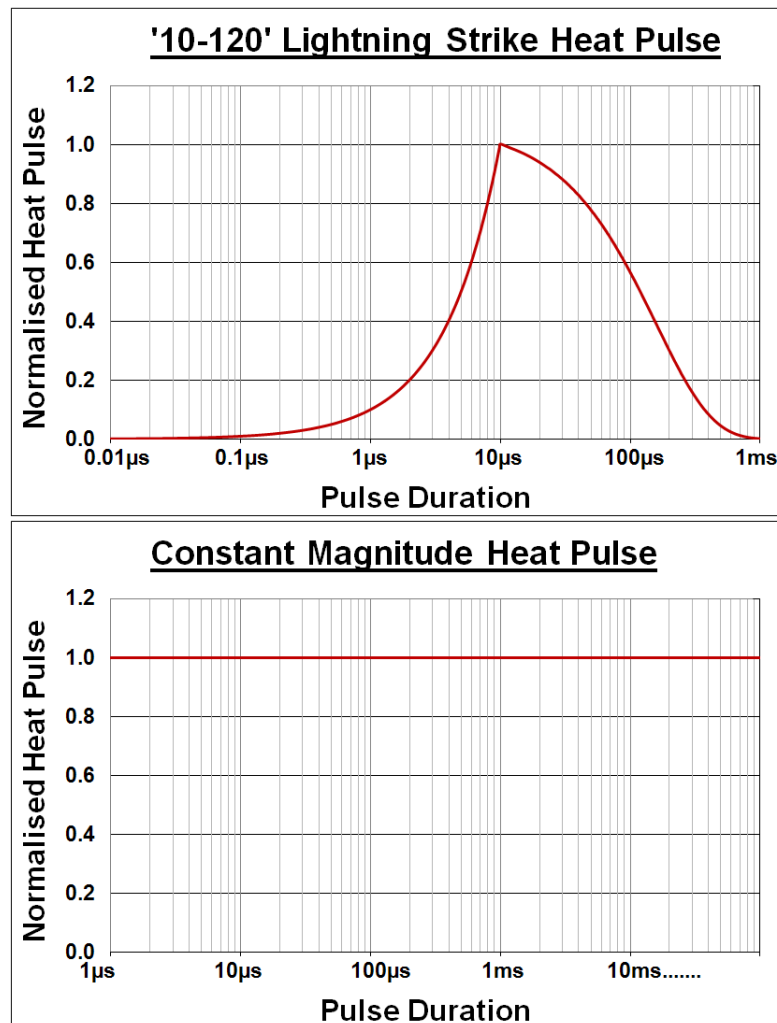
Figure 5.3: The two of heat pulses modelled. *Top*: a '10-120' lightning pulse form. *Bottom*: a constant magnitude pulse.

Temperature rise in the die depends linearly on the magnitude of the thermal pulse. This is demonstrated in Figure 5.4, which shows the temperature response in the die due to different magnitudes of heat pulse. As the MoP is based upon the ratio of temperature rise, it is independent of the heat pulse magnitude.
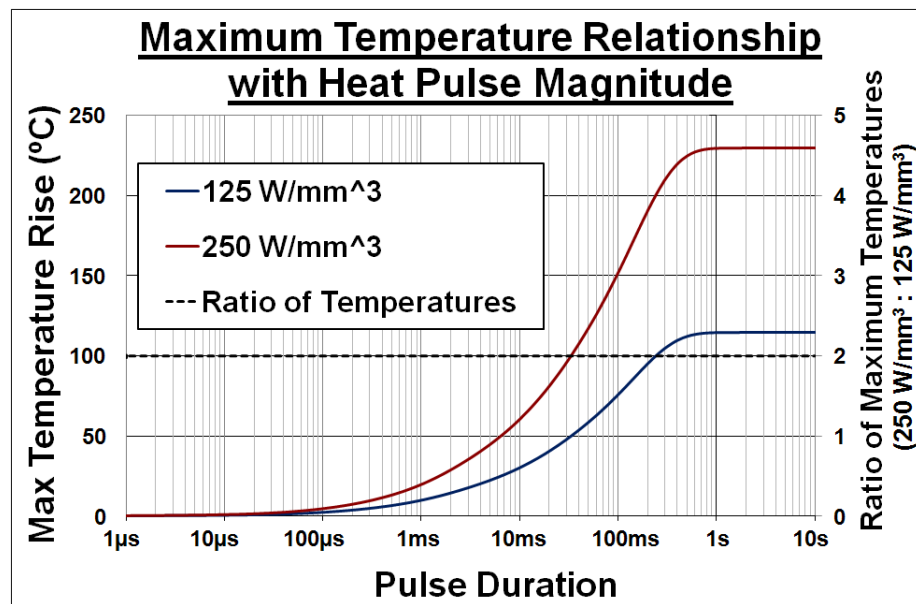
Figure 5.4: The maximum temperature rise observed in the same model when two heat pulses of the same shape but different magnitude, one twice as large as the other, are simulated. The ratio of the temperatures during the pulse is also plotted.

5.3.3 Boundary Conditions

Commercial devices are typically fully enclosed by a package. The packaging reduces heat loss from the upper part of the stack, and so heat transfer from these parts of the structure is minimal. The bottom of the baseplate is generally cooled, by either forced water or air convection, and is the main heat path out of the stack.

The boundary conditions in the numerical modelling reflect these real life boundary conditions. All external surfaces were modelled as adiabatic except for the bottom of the baseplate which was modelled as isothermal (25ºC), the initial starting temperature throughout the model.

The two types of stack-up are displayed in Figure 5.5 which shows the layers, the region of heat generation and the boundary conditions applied.
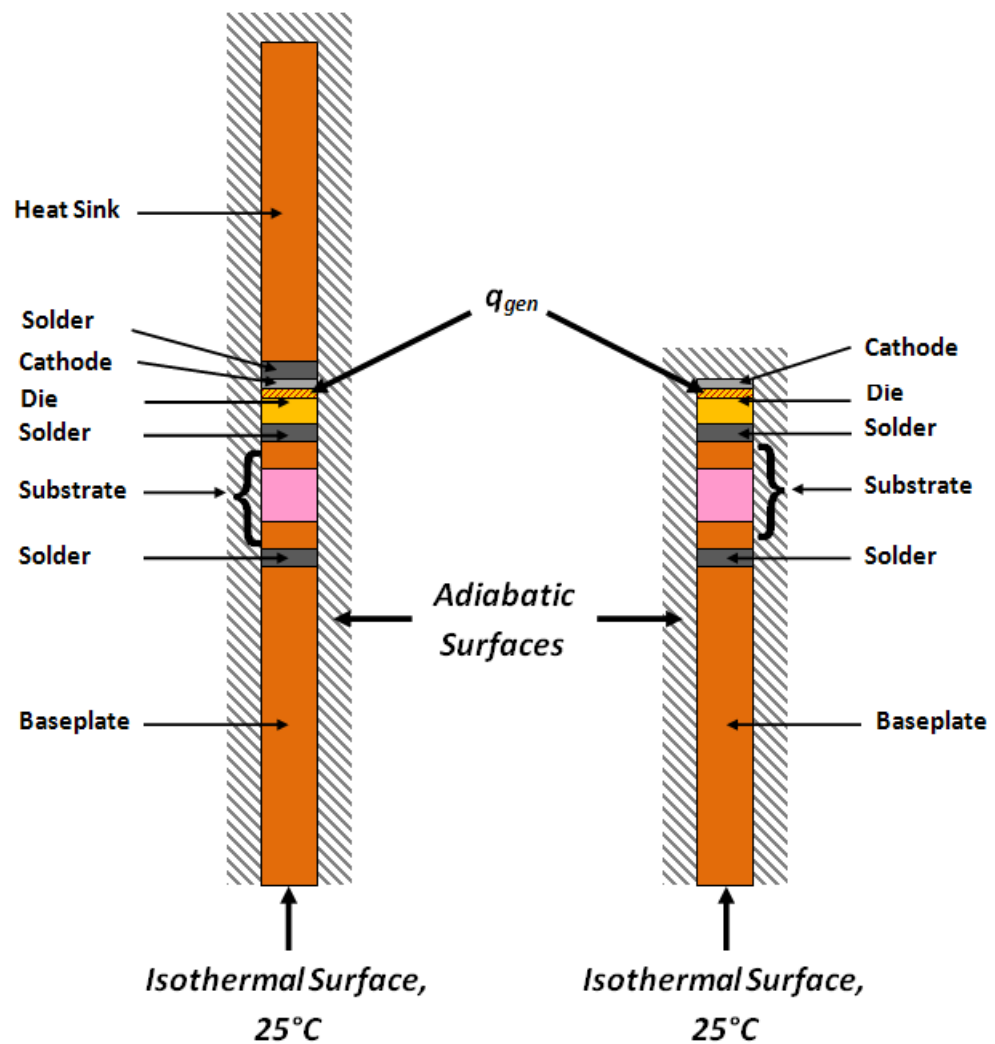
Figure 5.5: The two variations of the 1D model structure: the traditional stack arrangement (*right*) and the arrangement with a heat sink (*left*). All surfaces are modelled as adiabatic except for the bottom of the baseplate which is isothermal.

5.3.4 Domain Discretisation

To create the thermal models the stack-ups were discretised. This procedure is described in Section 4.3.

The node density needed to be greater around the region of heat generation due to the significant temperature gradients there during the early stages of the heat pulse. It was therefore necessary to have a dense node distribution in order to capture the

temperature resolution accurately.

5.3.4.1 Grid Density

The time taken to achieve the solution is related to the number of nodes, as calculations must be performed at each node for each time-step. It is therefore desirable to use as few nodes as possible. However, solution accuracy increases with number of nodes. Discretising the model domain with a large number of nodes will provide an accurate solution, at the expense of an increased solution calculation time. Choosing a suitable number of nodes is therefore important.

Models with different numbers of nodes were created to simulate a 10s constant magnitude heat pulse. The number of nodes used in each layer for the different models is shown in Table 5.4.

The profile of maximum temperature during the pulse was compared to the temperature profile from the model with next fewest nodes. The maximum relative difference between the temperature profiles was plotted against the change in the number of nodes. The plot is shown in Figure 5.6.

The results demonstrate that models with the fewest nodes are the most sensitive to a change in the number of nodes. Increasing the number of nodes from 34 to 67 caused a maximum change of 21% in the temperature rise. A further increase in the number of nodes to 134 causes up to a 17% change in the maximum temperature rise. These errors are large, and therefore a greater number of nodes are needed to provide confidence in the solution accuracy.

Changing the number of nodes from 528 to 654 causes a maximum change of 0.18% to the temperature during the pulse. Adding extra nodes does not change the solution significantly, and it can be assumed the solution is sufficiently accurate. For this reason the decision was made to use the '654 nodes' configuration.

	34 Nodes	67 Nodes	134 Nodes	264 Nodes	397 Nodes	528 Nodes	**654 Nodes**	822 Nodes
Heat Sink	20	40	80	160	240	320	**400**	500
Solder	1	1	2	4	6	8	**10**	12
Cathode	1	1	2	2	4	5	**6**	8
Die Heated	1	1	2	4	6	8	**10**	12
Die Unheated	1	4	8	16	24	32	**35**	43
Solder	1	1	2	4	6	7	**8**	10
Substrate Cu	1	3	6	12	18	24	**30**	40
Substrate AlN	1	3	6	12	18	24	**30**	40
Substrate Cu	1	2	4	8	12	16	**20**	25
Solder	1	1	2	2	3	4	**5**	7
Baseplate	5	10	20	40	60	80	**100**	125

Table 5.4: Table showing the node distributions used to find the point where mesh independence can be assumed. The highlighted column (654 nodes) shows the node distribution decided upon for the 1D modelling.
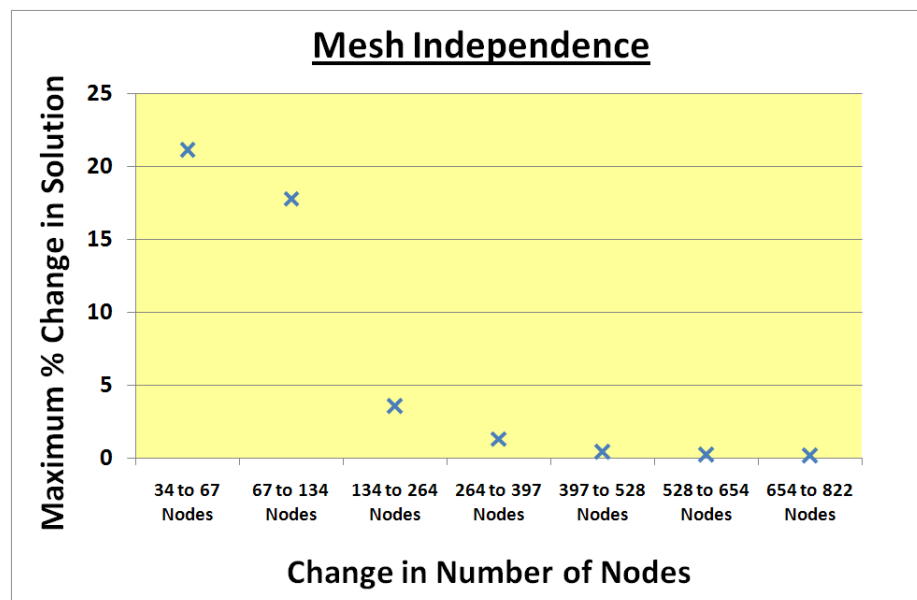
Figure 5.6: Graph showing the maximum change to the solution during the 10s pulse duration when the number of nodes is increased.

5.3.5 Program Justification and Validation

Many commercial software packages are available capable of running the heat transfer simulations required. Two such programs that were available were FLUENT and FLOTHERM. FLOTHERM is a piece of CFD software specialising for the heat transfer in electronic devices and systems, whilst FLUENT is a more general CFD software.

Both software packages had limitations for this problem.

The thermal conductivity of some materials is temperature-dependent, and it was considered necessary to be able model this. A further property which needed to be modelled was variation in thermal conductivity in different planes. Certain materials exhibit a highly anistropic thermal conductivity, in particular Highly Oriented Pyrolytic Graphite which has thermal conductivities of 1500W/mK and 15W/mK in different planes.

FLOTHERM and FLUENT are able to model both of these properties of a material independently, but not concurrently. Although this was not an initial problem when choosing the software to use, it was considered a limitation that may provide future problems should other composite materials be discovered that were needed to be modelled.

Further to this reasons, the format of the output files from the commercial packages were not ideal for the analysis required. Processing the results files outputted from the packages was time consuming, and considering hundreds of cases were to be run this was deemed reasonably significant.

Conduction was the main heat transport mechanism needing to be modelled, and only basic convection at a boundary needed to be simulated. Using the theory discussed in Chapter 2, a C++ program was written to model the heat conduction using thermal resistance networks. This provided control over the format of the output files, and also gave the ability for custom monitors to be added with ease. These monitors allowed the maximum temperature in the die to be displayed during the calculation process, as well as its location. These benefits allowed for quick calculation times (around 15 minutes for a 1D model) which enabled lots of different scenarios to be assessed without excessive computational or post-processing time.

5.3.5.1 Results Comparison with FLUENT

Validation of the C++ program was required to ensure the accuracy of the results. This was done by running identical simulations using the C++ code and in FLUENT, and comparing the results from both programs.

1D validation was performed for both stack-up types (the traditional stack-up and the modified stack-up with the heat sink). A 1D model was created in each program for each model type, using the dimensions listed in Table 5.3 and boundary conditions in Figure 5.5. Consistency in the number of nodes used in the models was ensured by using the number of nodes corresponding to the highlighted column in Figure 5.4.

The models used in the custom program were created as described in Section 4.3.2. The FLUENT models were created in ICEM as 2D surfaces with constant width and adiabatic walls. This created models with temperature variation in only one-dimension. Once meshed (using the blocking technique in ICEM), the model was imported into FLUENT where the initial and boundary conditions could be applied.

The temperature throughout the models was initially constant (300K in FLUENT, 25ºC in the custom program). The isothermal boundary was kept at this temperature throughout the simulation. A constant heat flux equal to 250W/mm^3 was applied to the heated die region of each model. This is the equivalent of around 1000W being dissipated in a device, which typically have cross-sectional areas in the order of 1cm^2. The maximum temperature inside the die was calculated in both programs during a 10s pulse.

The plotted maximum die temperatures throughout the pulse can be seen in Figure 5.7. For both cases (with and without a heat sink) the maximum temperatures calculated by each program show a very strong correlation. The maximum difference between the temperatures at any time was 2.5%. The difference was less than 1% for the majority of the pulse.

These errors can be attributed to the difference in solver methods. The custom program used a central difference method, whereas a second order upwind method

was used within FLUENT. This difference in the solver technique is likely to introduce a difference between the calculated temperatures.

A close match in temperature profiles is clearly visible in Figure 5.7. This provided satisfaction that the C++ program produced accurate results and it was deemed appropriate for the thermal modelling simulations.
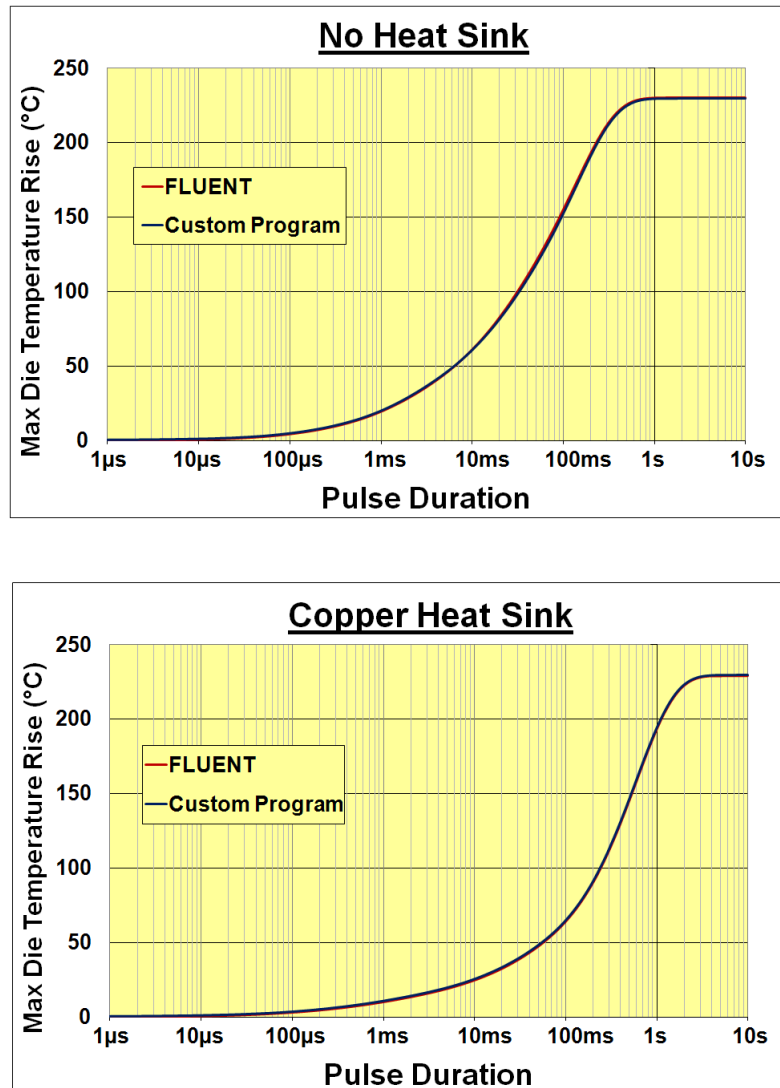
Figure 5.7: The maximum temperature rise in the die during a constant magnitude pulse. Each graph shows results obtained using a commercial program (FLUENT) and also the custom program written for the thesis.

5.4   1D Modelling Results

1D models were run with different stack and heat sink configurations. This allowed a wide range of situations to be explored quickly, allowing the effectiveness of the heat sink in each situation to be evaluated.

5.4.1   Heat Sink Material

An understanding of the way in which the material of the heat sinks affects its performance was required to inform about cost and performance.

Eight materials were chosen for testing including a synthetic diamond, copper, silver, aluminium and iron. Copper is well known to have very good thermal properties compared to cheaper common metals like aluminium and steel, but also costs more.

In addition to these, some composite materials manufactured for thermal applications were also considered: Highly Ordered Pyrolytic Graphite (HOPG), copper with diamond particles and silver with diamond particles (referred to as copper diamond and silver diamond herein). HOPG is a form of high purity carbon with aligned graphite sheets. Silver and copper diamond are composite materials with embedded diamond particles inside the metal matrix, increasing the thermal conductivity of the bulk material.

For each heat sink material, a 1D model with an arbitrary heat sink size of 5mm was created. The boundary condition at the top of the heat sinks were modelled as adiabatic, and the boundary at the bottom of the baseplate modelled as isothermal (at the initial model starting temp of 25ºC). A 10s constant magnitude heat pulse was applied to the models, and the maximum temperature in the die throughout the pulse recorded. A model with no heat sink was also run with the same conditions which allowed the transient MoP curve to be obtained for each heat sink. The maximum temperature rise in the die for each model was plotted, as well as the MoP curve for each heat sink. These are both shown in Figure 5.8.
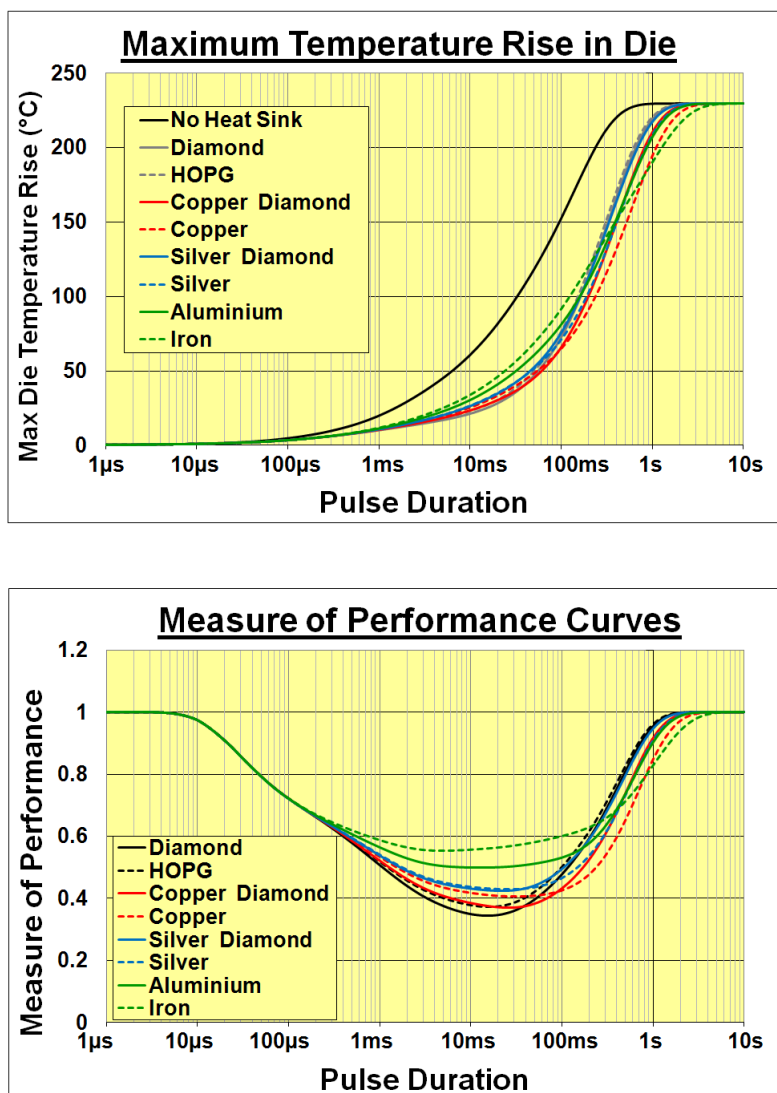
Figure 5.8: The maximum temperature rise in the die (*top*) and the MoP curves (*bottom*) for each heat sink material throughout a 10s constant magnitude pulse.

5.4.1.1 MoP Curve Overview

For pulses less than 4μs in length, the heat is unable to reach either the heat sink or the joining solder layer. Beyond 4μs the heat wave penetrates the solder layer causing the MoP curve to decrease, signifying that the solder layer is reducing the temperature rise in the die. The heat reaches the heat sink after around 190μs, leading to the temperature and MoP curves for the different heats sink materials to deviate from each other. The MoP curves continue to decrease until they reach their minimum value, which happens between 5ms (for iron) and 30ms (for copper) into the heat

pulse.

The MoP curves begin to rise once the heat wavefront has fully penetrated to the end of the heat sink. The time at which this happens is dependent on the thermal diffusivity of the heat sink material, and occurs between from 15ms and 233ms for all the heat sinks, except iron for which this happens after 960ms. Once the thermal wavefront has fully penetrated the heat sink, the temperature gradient between the die and the heat sink decreases until it becomes zero. This reduces the rate of heat transfer into the heat sink, reducing its effectiveness.

The isothermal boundary at the bottom of the baseplate influences the die temperature after about 177ms. As the heat has already reached the top of the heat sink in all cases except for the iron heat sink, the solution approaches steady state (at which point the temperature distribution remains constant). The MoP curves rise steadily, approaching the steady state MoP value of 1. The time at which the curves reach steady state is dependent on the thermal capacity of the heat sink. Those with greater thermal capacity can absorb more heat and thus take longer to reach steady state.

Layer	Material	Thickness (mm)	Thermal Diffusivity (m^2/s) $\times 10^{-5}$	Time Constant (ms)	Cumulative Time Constant (ms)
Cathode	Aluminium	0.02	8.41	0.004	0.004
Solder	Sn96.5-Ag3.5	0.1	4.22	0.186	0.190
Heat Sink	Diamond	5	13.1	15.0	15.2
	HOPG	5	9.20	21.3	21.5
	Copper–Diamond	5	3.72	52.8	53.0
	Copper	5	1.16	169.6	169.8
	Silver–Diamond	5	3.23	61.1	61.1
	Silver	5	1.74	113.1	113.1
	Aluminium	5	0.84	233.4	233.6
	Iron	5	0.21	960.1	960.3

Table 5.5: The thermal diffusivity and time constants of each layer above the die. Each of the heat sink materials examined are listed. The cumulative time constant is also shown which indicates the time taken for heat to fully penetrate each layer.
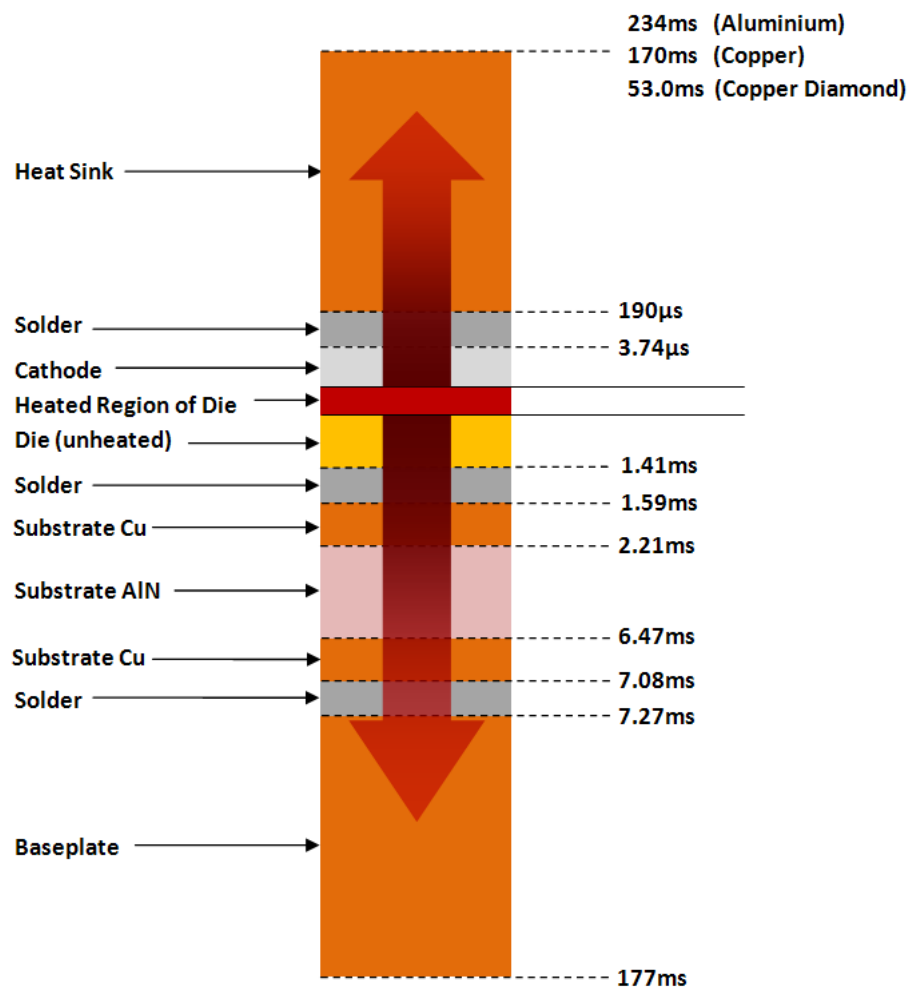
Figure 5.9: 1D stack-up with the heat sink, showing the time constant for heat to penetrate from the die to the specified location.

5.4.1.2   Early Regime of the MoP Curves

At the early stages of heat penetration into the heat sink, the effectiveness of the heat sink is dependent on the coefficient of heat penetration (CHP) of the heat sink material (see Section 4.2). Heat sinks with a higher CHP reduce the die temperature more during this stage, as the materials have a low thermal resistance and high thermal capacitance.

According to Table 5.5, it takes longer than 10ms for heat to penetrate to the end of

any of the heat sinks. Plotting the MoP for each heat sink at this time against the corresponding CHP demonstrates the correlation between them. Table 5.6 lists the CHP for each of the heat sink materials, as well as the MoP after 10ms and Figure 5.10 shows the graph.

	Thermal Capacitance (MJ/m^{3}K)	Coefficient of Heat Penetration (KJ/K.m^2.s$^{1/2}$)	MoP After 10ms	Minimum MoP During Pulse
Diamond	1.782	64.427	0.348	0.343
HOPG	1.631	46.458	0.378	0.372
Copper Diamond	2.420	46.669	0.384	0.369
Copper	3.438	36.993	0.417	0.405
Silver Diamond	1.860	33.407	0.433	0.425
Silver	2.468	32.535	0.437	0.429
Aluminium	2.425	22.244	0.500	0.499
Iron	3.569	16.142	0.557	0.553

Table 5.6: Coefficient of heat penetration of the heat sink materials and the MoP of each heat sink after a 10ms pulse.
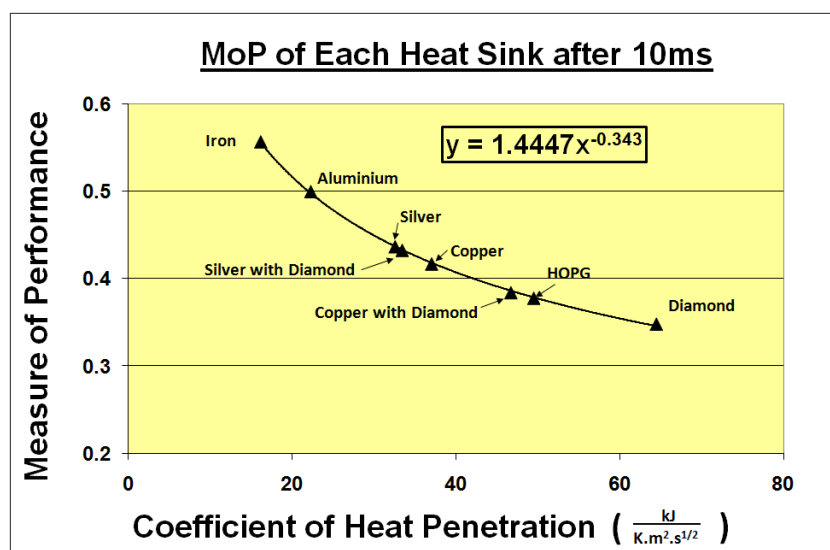
Figure 5.10: A plot of the coefficient of heat penetration and the MoP after 10ms for each heat sink. A power curve can be fitted to the data well, demonstrating a strong relationship.

Whilst the heat is diffusing through the heat sink, but yet to reach the end, the MoP is proportional to the CHP:

$$MoP = a\,CHP^b + c \tag{5.4}$$

where a, b and c are constants.

The CHP describes how much thermal capacity is available within the early regime of the heat flow, and is derived from heat conduction theory in Section 4.1. The constants a, b and c are dependent on the model geometry (size and materials of components), boundary conditions and pulse duration.

The point at which the minimum MoP is achieved is also dependent on the material of the heat sink, and varies between 5ms for iron and 30ms for copper. The minimum MoP values (listed in Table 5.6) are proportional to the CHP of the heat sink, suggesting these values are reached during the early conduction regime. The correlation of the minimum MoP and heat sink CHP is shown in Figure 5.11 which, along with Figure 5.10, suggests that the CHP of the heat sink material is the most important material property to consider in the design of a heat sink that is effective for pulse durations of this order of magnitude. However, as the CHP increases, the gradient of the MoP curve reduces, indicating that further reductions in the MoP require large increases in CHP

5.4.1.3   Late Regime of the MoP Curves

The CHP continues to be the defining property on the heat sink performance until the heat reaches the end of the heat sink, defined by the time constant. This happens at a different pulse duration for each heat sink material and is determined by the thermal diffusivity. Once the heat sink depth becomes exhausted, heat cannot travel any further and so the thermal resistance becomes less relevant. The CHP no longer defines the shape of the MoP curve; the volumetric thermal capacitance ($\rho\,c_p$) of the heat sink becomes the determining material property.
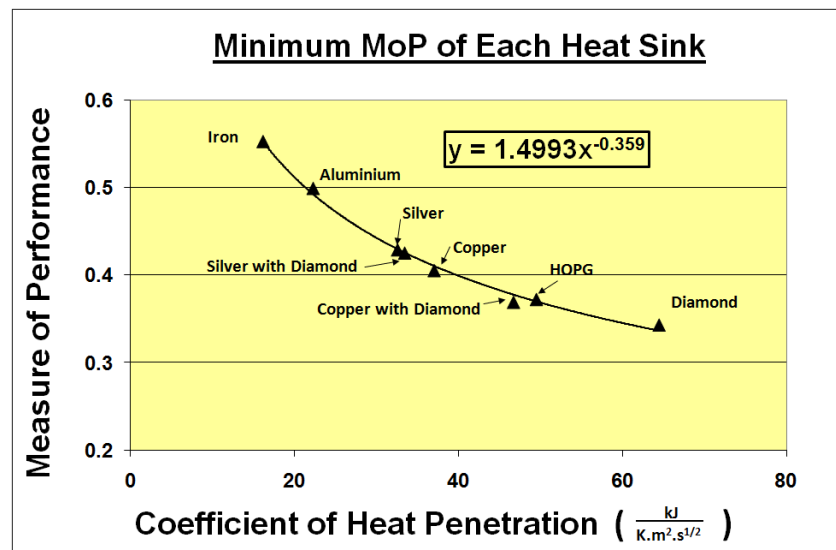
Figure 5.11: A graph showing the coefficient of heat penetration and the minimum MoP achieved by each heat sink during the 10s pulse. A power curve similar to that in Figure 5.10 can be fitted to the data.

Figure 5.8 shows all MoP curves return to a value of 1 at pulse durations between 1s and 5s. It can also be seen that the same steady state die temperature is reached for every model. At steady state, MoP tends to 1 and the heat sink does not affect the die temperature due to the adiabatic boundary condition applied to the top of the heat sink.

A comparison of the temperature profiles through a model with and without a heat sink after 10s is shown in Figure 5.12. It shows a uniform temperature through the heat sink once steady state has been reached. At steady state, all heat sinks show the same temperature profile shown in Figure 5.12, regardless of the material.

The the time taken to increase the heat sink and die temperature to steady state is related to thermal capacities. Heat sinks with lower thermal capacities become saturated quicker, causing steady state to be reached sooner, whereas heat sinks with high thermal capacities have greater thermal inertia, and require more thermal energy to raise the temperature.

Once the heat wavefront has reached the end of the heat sink, the volumetric thermal
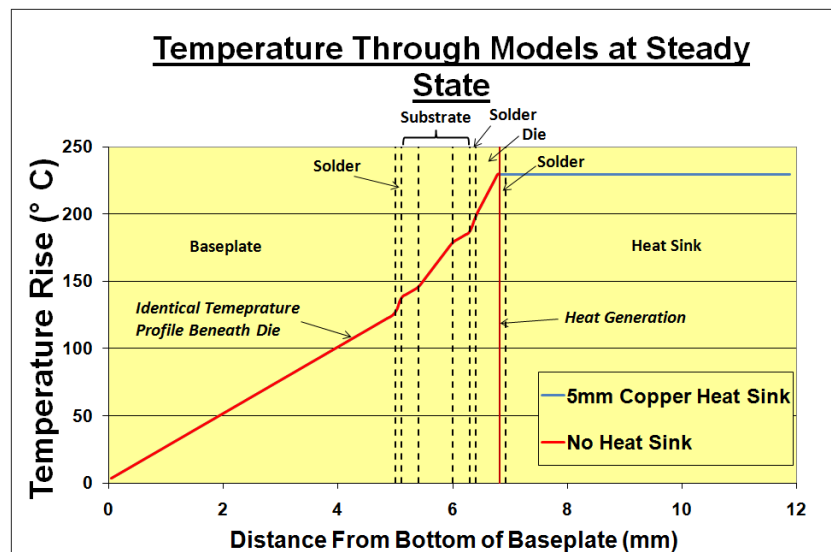
Figure 5.12: Temperature distributions through a 1D model after 10s, comparing a model with and without a heat sink. The blue and red line follow identical paths between the bottom of the baseplate (x=0) and the top of the die - only the red line is visible as it hides the blue line.

capacity dictates how quickly the heat sink saturates, which is reflected in the shape of the MoP curve. Lower MoP curve gradients indicate a slow increase in the die temperature which signifies a heat sink with a high thermal capacity, and vice versa.

5.4.1.4 Heat Sink Material Assessment

The heat sink material that achieved the greatest reduction in die temperature was diamond, which reduced the maximum temperature rise by almost 68% at a pulse duration of 65ms. HOPG and copper diamond performed second best, with almost identical reductions in maximum die temperature of around 64%. Pure copper, with a reduction of 60%, was the next best. Other common materials tested were aluminium (~50% reduction) and iron (~44% reduction).

Diamond and HOPG have a high thermal conductivity, leading to them having the greatest CHP values. At short pulse durations of less than 20ms, these heat sinks performed very well, reducing the maximum die temperature by up to 66% and 63%, respectively. Although diamond was the best performer, the cost and practical diffi-

culty of manufacturing a heat sink out of synthetic diamond makes it an unrealistic prospect.

HOPG performed well due to its high thermal conductivity of around 1500W/mK (compared to 398W/mK for copper). However, it is highly anisotropic and has a thermal conductivity of just 15W/mK in the lateral direction. As the modelling done was in 1D, this anisotropic property was not represented in the results. The performance of HOPG will likely decline once this effect is taken into account, as high thermal resistance in the lateral direction would hamper the heat spreading outwards.

The large reduction in die temperature that the diamond and HOPG heat sinks provide early on in the pulse are short-lived; despite having the lowest MoP values until 20ms into the pulse, after 300ms they have the highest MoP values of all the heat sink materials. Although they have the highest CHP values, they have the lowest thermal capacities, as shown in Table 5.6. During the period where the MoP curve changes from being dependent on the CHP to the thermal capacity, the MoP curves for diamond and HOPG experience sharp increases compared to the other materials. This transition period is when thermal conductivity becomes less important.

Copper and copper diamond have slightly lower CHP values than diamond and HOPG, but still perform well during at short pulse durations. However, as they also have high thermal capacities their MoP curve minima persists longer, providing higher benefits than both diamond and HOPG by the time the pulse reaches 100ms in length. Due to their relatively high thermal capacity, the copper and copper diamond heat sinks take longer to reach steady state, reducing the maximum die temperature for longer than diamond or HOPG.

Iron takes the longest to return to steady state due to its high thermal capacity, and reduces the die temperature the most between 1s and 5s. However, due to its low thermal conductivity of 73W/mK, it is by far the worst performing heat sink for the short pulse durations of most interest.

5.4.1.5   Composite Heat Sink

The possibility of using a composite heat sink was initially thought of at this stage. Combining two materials to form a heat sink in a way that the thermal properties of both were exploited fully could produce a more effective heat sink. HOPG would be ideal for use in a composite heat sink due to its high thermal conductivity to take heat away from the die. Another material with a high thermal capacity could be used alongside HOPG to improve the heat sink performance later in the pulse when pure HOPG is less effective. A high thermal capacity material, such as copper, could store the heat (the density of HOPG is only about a quarter of that of copper) and also conduct heat more effectively in the lateral direction.

Composite heat sinks are investigated as part of the 2D modelling. The results and discussion can be found in Section 6.2.5.

5.5   1D Modelling Sensitivity Analysis

A sensitivity analysis was performed to see how the different stack components affect the MoP of a heat sink. During this analysis, modelling all the heat sink materials would be time consuming and so only three heat sink materials were considered: aluminium, copper and copper diamond.

Aluminium was chosen for its low cost and weight and common availability, copper for its established performance in thermal applications and copper diamond as an available novel material that seems to provide greater benefits than copper.

5.5.1   Heat Sink Thickness

Investigating how the thickness of the heat sink affects the MoP at different pulse durations allows the optimum heat sink size to be chosen. Oversizing the heat sink would increase the cost, size and weight unnecessarily. Undersizing the heat sink would reduce its effectiveness during the heat pulse.

For each of the three heat sink materials different heat sink thicknesses were modelled, ranging from 1mm to 10mm. The rest of the stack remained the same, conforming to the standard dimensions. A 10s constant heat pulse was simulated in the die for each model, and the maximum temperature rise calculated throughout the pulse. The MoP curves obtained were plotted and are shown in Figure 5.13. Each graph shows the MoP curves for the different heat sink thicknesses.

Increasing the thickness of the heat sink extends the pulse duration for which the heat sink is able to reduce the maximum die temperature. The benefits of increasing the heat sink thickness are two-fold:

Firstly, the time taken for the heat wavefront to reach the end of the heat sink is extended. This extends the early regime of the conduction into the heat sink. During this time the CHP of the heat sink determines the heat sink effectiveness.

Secondly, the time taken to saturate the heat sink increases. Once the heat wavefront reaches the end of the heat sink, the thermal capacity of the heat sink becomes important. Increasing the thickness of the heat sink increases the amount of thermal capacity available for the heat. This increases the time it takes to increase the heat sink temperature up to the steady state temperature. Whilst the heat sink temperature is increasing, it is reducing the die temperature, as it is absorbing heat generated in the die.

Heat sinks with greater thermal diffusivity are most sensitive to a change in the heat sink thickness. This is evident in Figure 5.13 which shows that the MoP curves for the copper diamond heat sink have a greater spread compared to the aluminium heat sink.

The greater the thermal diffusivity, the sooner the heat wavefront penetrates the full length of the heat sink. Therefore, in order to maintain the early conduction regime in these heat sinks for a certain pulse duration, a thicker heat sink is required compared to an equivalent heat sink with a lower thermal diffusivity.

The optimum heat sink thickness for each heat sink material can be determined from the MoP curves in Figure 5.13. As the heat sinks are to be designed to handle transients up to 100ms, comparing MoP curves at this pulse duration allows the ideal heat sink thickness to be determined. The heat sink thickness which provides the lowest MoP value at 100ms will have the lowest maximum die temperature at this stage, and therefore would be most suitable.

It follows that if an aluminium heat sink were chosen, a size of 4mm would be adequate. Increasing the size to 5mm decreases the MoP for a 100ms pulse duration by less than 1%, which is insignificant. Similarly, a 5mm copper heat sink would be optimum. A very small benefit could be seen by increasing this to 6mm, but the additional benefits are slim. If a copper diamond heat sink were used, the ideal thickness would be 8mm.
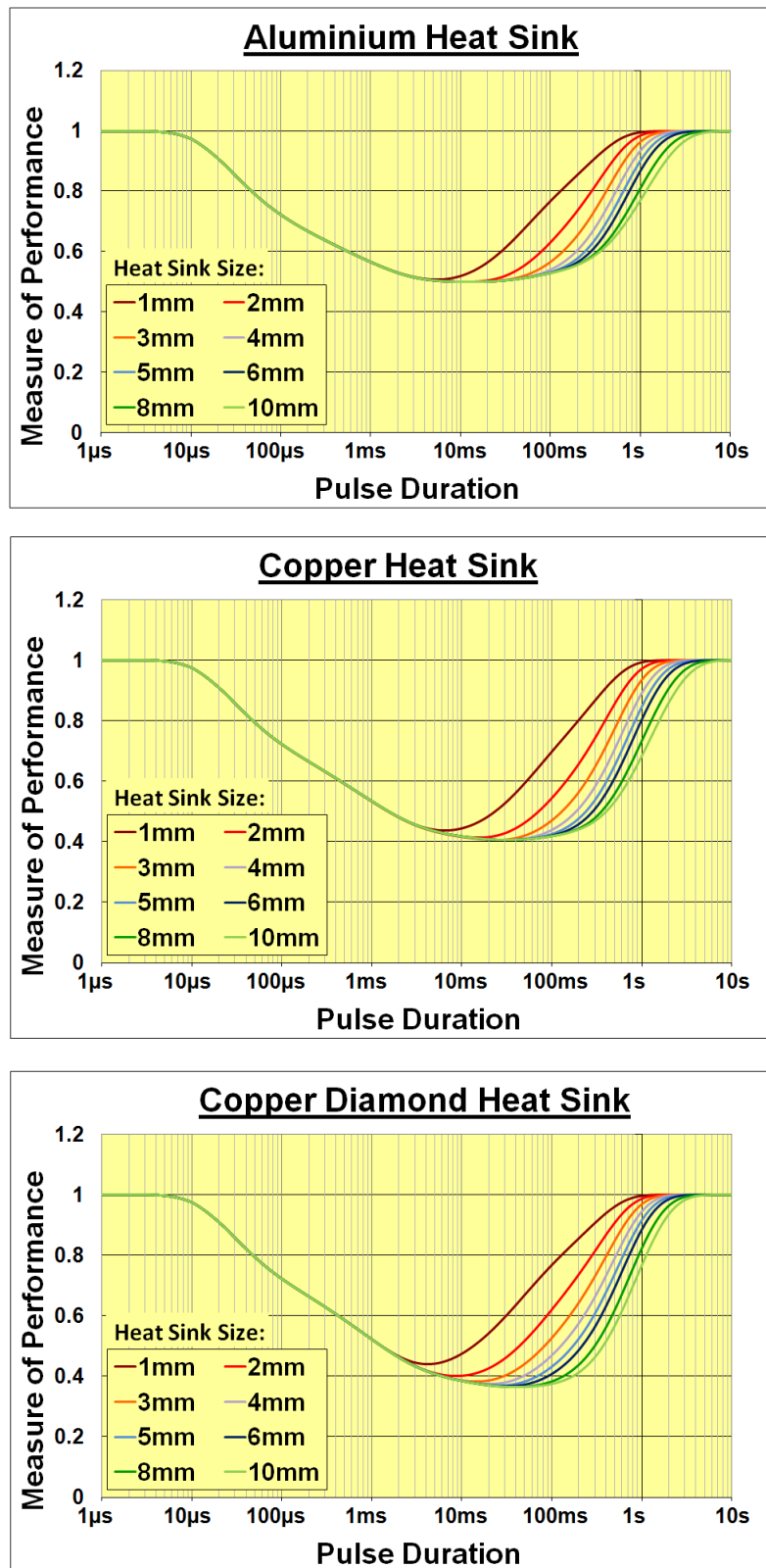
Figure 5.13: Graphs showing the MoP curves for different sized heat sinks during a constant magnitude heat pulse.

5.5.2 Baseplate Thickness

Baseplates provide structural rigidity to the stack-up, acting as an interface between the electrically active part of the structure and the coolant (typically air or water). In addition to its structural role, the baseplate must also be able to conduct heat effectively to the cooled surface.

The effectiveness of the heat sink for stacks with different baseplate thickness was investigated. Both model types (with and without a heat sink) were changed to simulate the heat transfer in the stacks with a variety of baseplate sizes, ranging from no baseplate (in which case the solder layer between the substrate and baseplate was also omitted) to 5mm in thickness. The same boundary conditions were applied, noting that in the 'no baseplate' case the isothermal boundary was applied to the bottom of the substrate.

Aluminium, copper and copper diamond heat sinks with a constant thickness of 5mm were modelled, and the MoP curves obtained for the different baseplate thicknesses. The MoP values throughout the 10s pulse were plotted, shown in Figure 5.14.

The initial separation of the MoP curves representing models with different baseplate sizes occurs at about 8ms into the pulse. At this point during the pulse, the MoP curve for the 'no baseplate' scenario breaks away from the other curves, which remain grouped. At this point the heat wavefront reaches the isothermal boundary at the bottom of the substrate when no baseplate is modelled. According to Table 5.1, the time taken for the heat to fully penetrate to the bottom of the substrate is 7.08ms. This corresponds reasonably well to the pulse duration where the curve deviation becomes noticeable, after around 8ms.

A general trend which is exhibited in all three graphs in Figure 5.14 is that the thinner the baseplate, the less effective the heat sink becomes at longer pulse durations. The reason for this relates to the thermal distance between the die and the isothermal boundary. The MoP curves increase as the solution approaches towards a steady state.

This begins to happen when the heat wavefront reaches the isothermal boundary. The time taken to reach the isothermal boundary is dependent on the baseplate thickness; the thinner the baseplate the shorter the distance the heat must penetrate.

In addition to the time required to reach the isothermal boundary, another determining factor of the time to reach steady state is the thermal mass in the system. There is a significant difference between the thermal mass of the stack with no baseplate and the one with a 5mm baseplate. As a result, steady state is reached sooner for models with thinner baseplates, as there is a smaller thermal mass to absorb the heat. Larger baseplates have a greater thermal mass and hence need more time to raise the baseplate temperature.

Combining the increased time to reach the isothermal boundary and the increased time to raise the baseplate temperature provides an explanation for the trend in the MoP curves.

Although the heat sink provides most benefits when devices are mounted on thick baseplates, the actual die temperatures may not be lower in these cases. The MoP is a measure of the reduction in the maximum temperature rise when a heat sink is used. As the baseplate thickness was varied in both model types (with and without a heat sink) the actual temperatures cannot be deduced from the dimensionless MoP alone. It provides only an indication of how much a heat sink reduces the max temperature in a certain situation.

One further observation is that the effect of varying the baseplate size on the MoP is consistent for different heat sink material graphs. Changing the baseplate thickness will not affect the performance of one heat sink material compared to another.

Figure 5.14: Graphs showing how varying the baseplate thickness affects the MoP curve for a 5mm aluminium, copper and copper diamond heat sink.

5.5.3 Substrate Thickness

Another significant part of the stack-up which can be investigated is the substrate. Substrates are formed from two metal layers (commonly copper) separated by a ceramic layer, in this case aluminium nitride. Substrates are required to provide electrical isolation between the die and the baseplate. As well as providing this electrical isolation, the ceramic layer also reduces thermal stresses in the die. The thermal expansion rate of the ceramic inhibits the expansion of the copper substrate layer, which could otherwise damage the die.

When investigating the thickness of the substrate, the substrate layers were scaled together rather than investigating the effect of scaling each layer independently. The scaling factor varied from 0.25 to 4 times the original substrate thickness in Table 1.1. The hypothetical scenario of no substrate at all was also examined, in which case only one of the substrate solder layers was modelled. The effect of changing the substrate thickness can be seen for the three heat sink materials in Figure 5.15.

The MoP curves show a degree of complexity as a result of changing three layers simultaneously. However, the MoP curves follow a similar pattern for each heat sink material.

The initial separation of the curves occurs when heat wavefront reaches the ceramic layer in the model, which is Aluminium Nitride (AlN). Up to this point, all models are identical beneath the die. AlN has a fairly low CHP, around half of that of copper (19.65 and 36.99 $KJ/K.m^2.s^{1/2}$, respectively). As such, it is ineffective at reducing the die temperature when heat wavefront travels through it. The heat sinks are therefore more effective when the heat is passing through the AlN layer in the stack, as opposed to copper.

The pulse duration at which the heat arrives at the AlN layer is related to the substrate thickness; the thinner the substrate the sooner in the pulse the AlN is encountered. This causes the order in which the curves deviate from the main group to be in

this order. Consequentially, the MoP curves obtained from the models with thinner substrates deviate from the 'no substrate' curve sooner than in the thicker substrate models. At around 10ms into pulse, the thinnest substrate modelled has the smaller MoP value, and the thickest has the greatest. Although models with thicker substrates aren't affected by AlN layer until later in the pulse, the effect the AlN layer has on the die temperature is greater due to its thickness.

A distinct 'w' shape to the MoP curve of the substrate four times larger than the standard size can be seen. The second dip (as it deviates from the 'no substrate' curve) indicates the heat passing through the AlN. As the curves ascend back up to unity, the heat sink provides the most benefit for the models with thicker substrates. This is the same pattern as seen in the baseplate sensitivity analysis, and the trend exists for the same reasons.

The increased thermal distance between the die and isothermal boundary increases the time it take the heat to reach the boundary. Additionally, there is more thermal mass in the stack, which requires more heat (and hence a longer pulse time) to increase to the steady state temperature distribution.
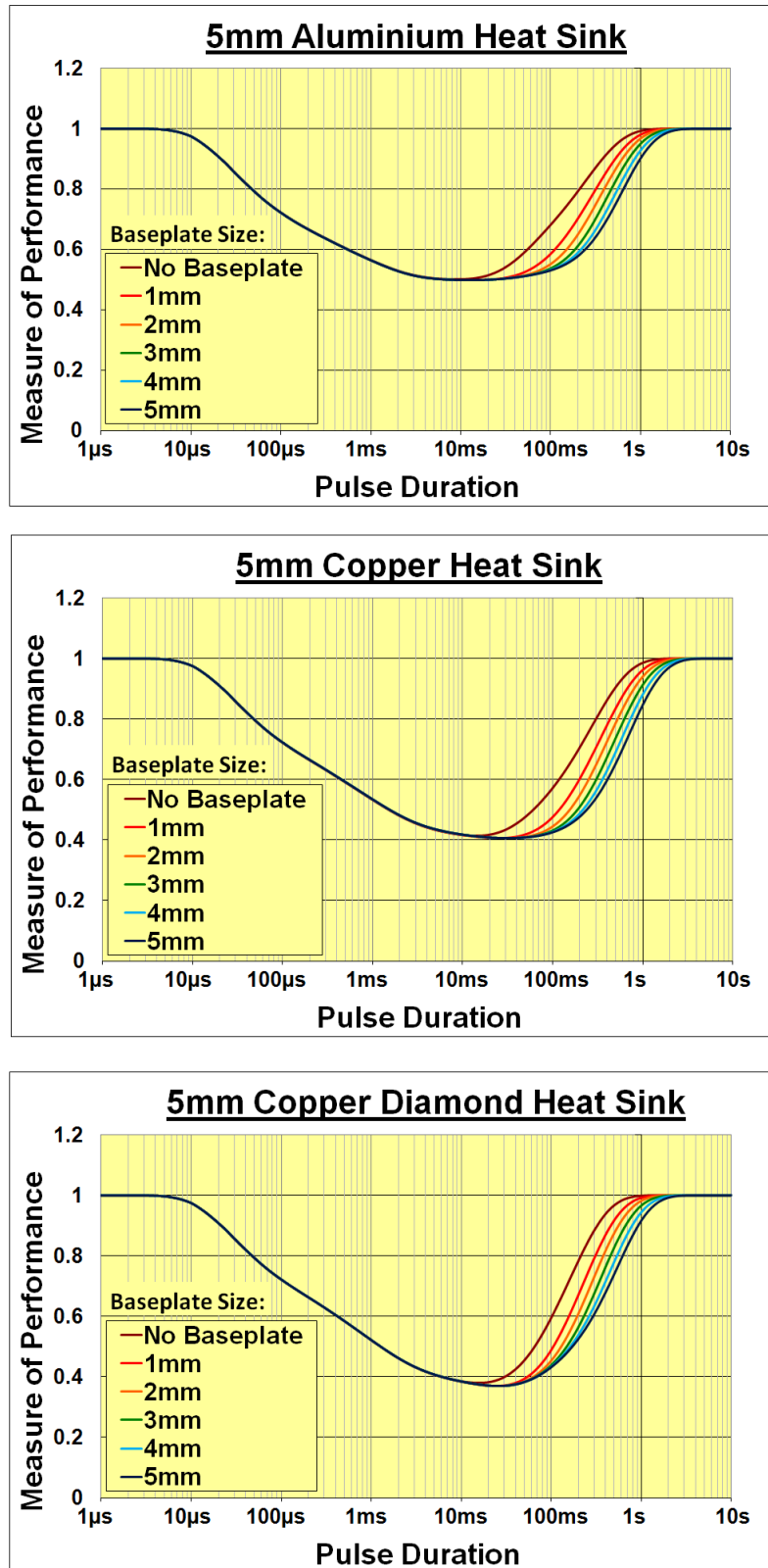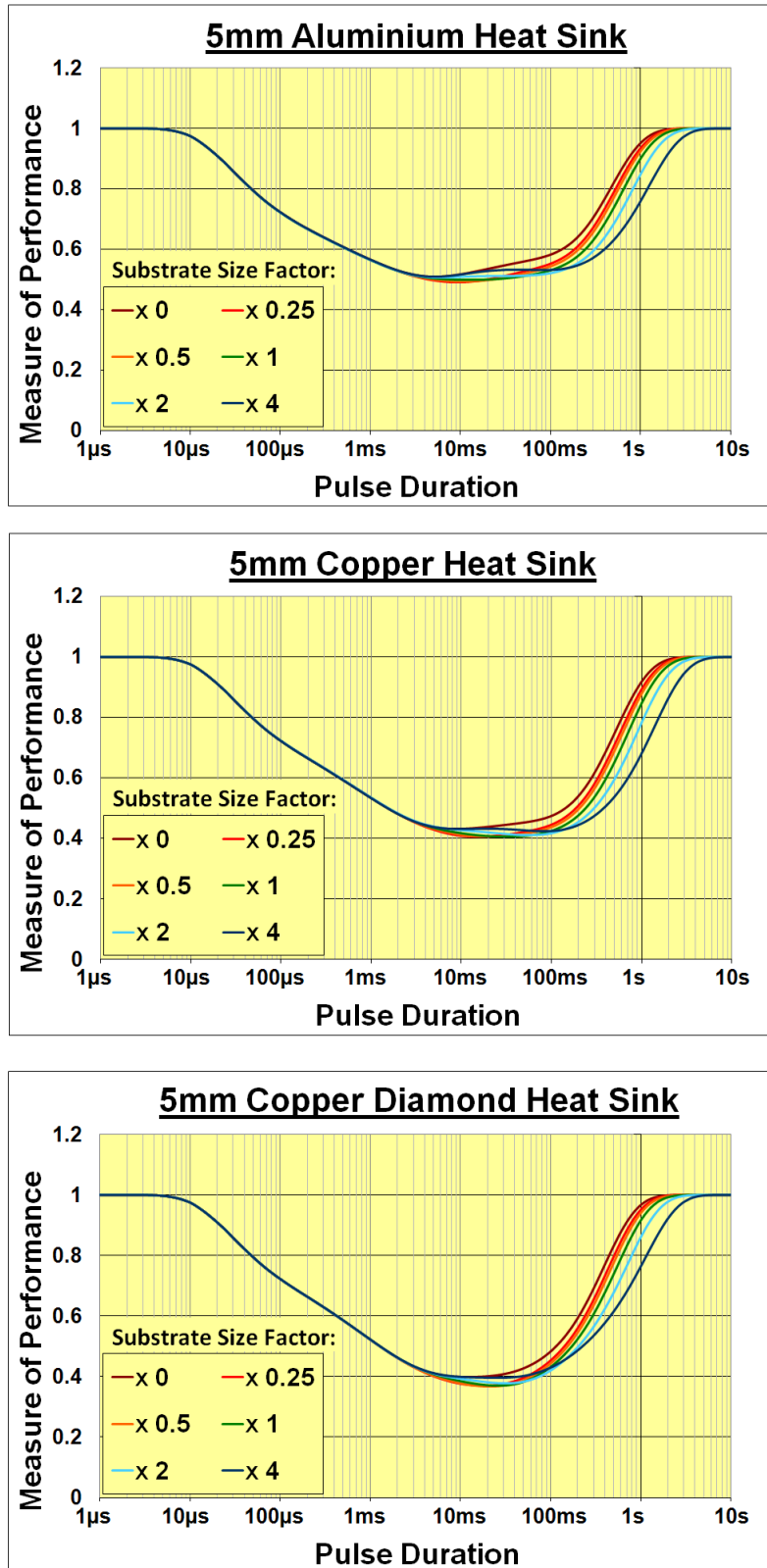
Figure 5.15: Graphs showing how varying the substrate thickness affects the MoP curve for a 5mm aluminium, copper and copper diamond heat sink.

5.5.4 Solder Joint Thickness

Solder joints are an important part of the stack-up. Individual layers must be joined at the interfaces by a material that will provide a strong mechanical bond, a good thermal contact and, for certain joints, a good electrical contact.

Heat dissipation in the die must be conducted away as effectively as possible to minimise the die temperature. To do so, it is necessary to reduce the thermal resistance across the interfaces between the layers to allow the heat to travel between them as easily as possible. This is a common problem in many different thermal management designs. Any air voids that are left within these interfaces when using a thermal glue, grease or solder may causes a significant increase in the thermal resistance across the interface, resulting in greater temperature gradients across it.

In the stack-up, the heat must flow across two solder joints when travelling from the die to the baseplate boundary where it can escape to ambient. Any unnecessary increase in the thermal resistance at these joints results in an increased die temperature, which reduces the device reliability and increases the risk of failure. A solder joint is also used to connect the heat sink to the die. Minimising the thermal resistance between the die and the heat sink will allow the cooling benefits of the heat sink to be maximised, reducing risk of device failure.

To reduce the thermal resistance across solder joints it would be instinctive to minimise the thickness. However, compromise must be made to allow for mechanical integrity in the joints. During thermal cycling, when the structure heats and cools due to fluctuations in device usage, each part of the component expands and contracts due to the change in temperature. Materials expand at different rates when they are subjected to a change in temperature and this is described by the material property called the Coefficient of Thermal Expansion (CTE) which has the units (/K).

The solder must maintain the integrity of a joint between two materials expanding and contracting at different rates during the thermal cycling. Fracturing, which is

common under these conditions, introduces voids in the solder layer, which greatly increases the thermal resistance. This leads to hotter die temperatures and reduced device reliability. The effect of solder thickness on reliability is discussed in Section 3.3 as part of the literature overview.

The solder thickness must be a balance between thermal resistance and reliability. This thickness will vary between packages depending on the materials and the environment under which it is used. It can also be difficult to manufacture precise solder thicknesses, and therefore knowledge of how the solder thickness affects the performance of the heat sinks is desirable.

The solder layers (2 layers in the no heat sink models and 3 layers in the heat sink models) were all varied simultaneously, scaled with respect to the original assumed solder thickness of 0.1mm. The scale factor was varied between 0.25 and 4 times the original thickness, and the 'no solder joint' case was also modelled for comparison. The MoP curves for each of the three heat sinks are shown in Figure 5.16, each graph showing how the solder thickness affects the MoP throughout a 10s pulse.

The heat wavefront reaches each solder layer at different times, affecting the MoP in different ways. The solder joint between the die and heat sink is the first to be influential. Thinner solder joints allow the heat to reach the heat sink sooner. As all the heat sink materials have a greater CHP than solder (which has the lowest of all the materials used in the models), the MoP decreases once it reaches the heat sink due to the heat sink acting as a more effective temperature suppressor than solder. Models with the thinnest solder layers are most effective early on in the pulse.

Heat sink materials with a high CHP are affected most by the solder layer between the die and heat sink. Once the heat reaches the heat sink, the deviation in the MoP curve is dependent upon the ratio of the CHP between solder and the heat sink. At 10ms into the pulse, the difference in MoP between the no solder case and the 400μm case is 0.04 when an aluminium heat sink is used, compared to 0.13 for a copper heat sink and 0.17 for copper diamond one.

The two solder layers in the stack, joining the die to the substrate and the substrate to the baseplate, affect the thermal resistance between the die and baseplate boundary. Like the baseplate and substrate variation, models with thicker solder layers (and hence greater thermal resistance to the isothermal boundary) take longer to reach steady state. This causes a reverse in the MoP trend - thicker solder joints produce lower MoPs than thinner ones. This occurs after around 30ms for the aluminium heat sink, and 100ms for both copper and copper diamond.

Thicker solder layers beneath the die cause an increase in the thermal resistance and the thermal mass in the model. These factors both contribute to an increase in the time required for steady state to be reached, which is a feature of the MoP curves.

From a thermal perspective, the thinner the solder joints the lower the die temperature. Using thinner solder joints to connect the heat sink to the die allows the heat sink to be active sooner, reducing the die temperature. Heat is also able to reach the cooled baseplate surface sooner when thinner solder layers are used. However, as mentioned reliability must also be considered in choosing the best solder thickness for each application.
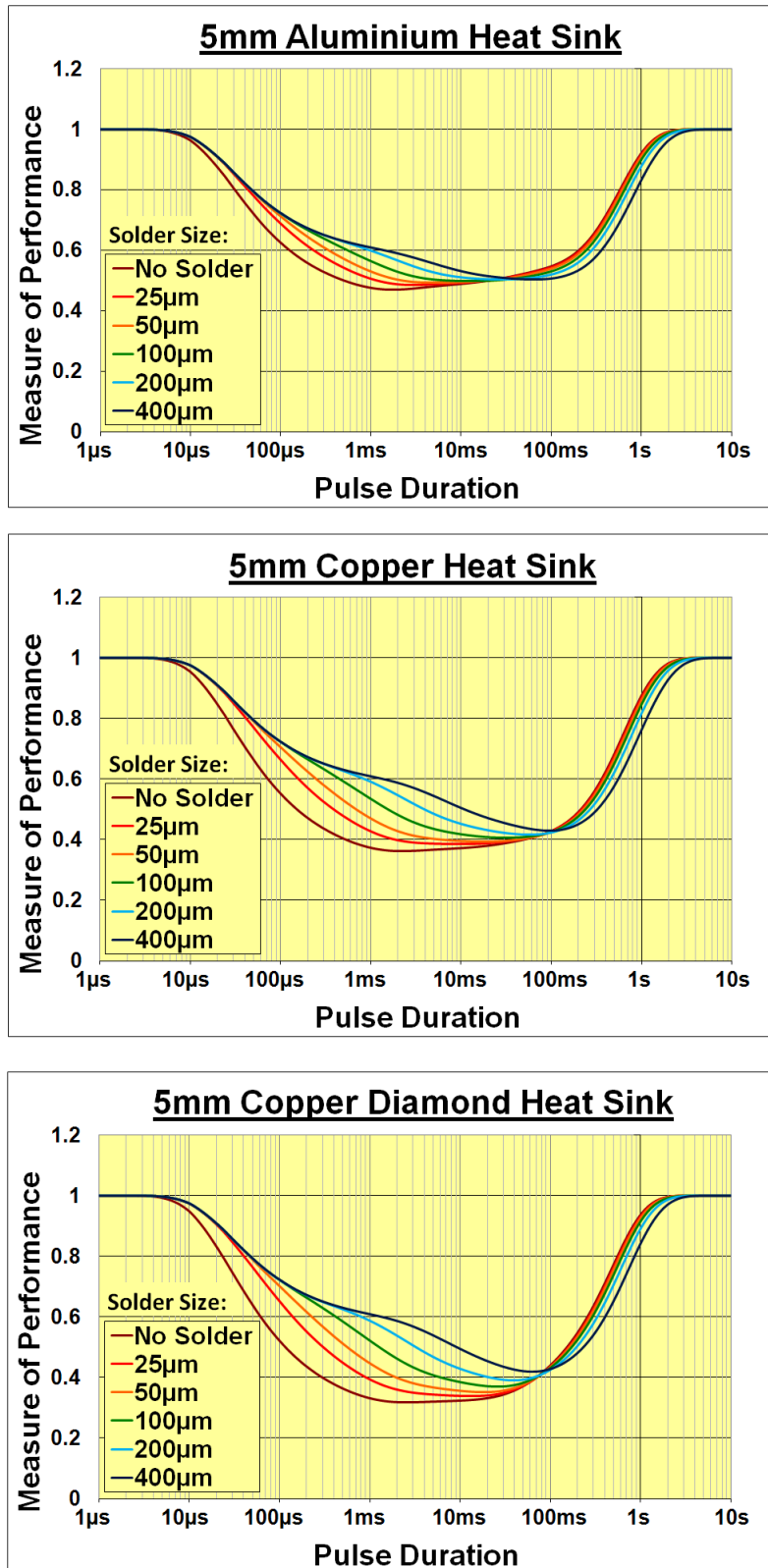
Figure 5.16: Graphs showing how varying the solder thickness affects the MoP curve for a 5mm aluminium, copper and copper diamond heat sink.

5.5.5 Heat Generation Region

The region of the die where heat generation occurs is difficult to predict and dependent on several factors. These factors include the device type (e.g. diode, IGBT, MOSFET), the device construction (e.g. size of *p-n* layer, amount of doping) and the operating voltage and current. Thus far it has been assumed that the heat has been generated in the top 10% of the die. This assumption can be made for MOSFETs and IGBTs at low voltage applications. However, for higher voltage applications more of the device becomes active, leading to the heat generation becoming more widespread throughout the device.

The effectiveness of the heat sink is likely to vary depending on where the heat is generated. When heat is generated further down the die, the distance to the heat sink is increased, thus creating a delay in the heat sink becoming active.

A variety of models were run using the standard stack dimensions (Table 1.1), each simulating the heat generation in a different region of the die. The scenarios considered were: heat generated evenly in middle 10% of die, heat generated evenly in bottom 10% of die, heat generated evenly throughout the die, heat generated evenly in top 50% of die and heat generated evenly in bottom 50% of die. A diagram depicting the examined heat simulation regions is shown in Figure 5.17.

Section 2.3 discusses the heat generation mechanisms in devices. It is explained that for MOSFETs and IGBTs the heat generation during the on-state tends to be in the top of the device. This suggests that the most realistic cases explored where heat is generated at the top, or possibly the middle of the device. The further cases are examined for comparative purposes.

The transient MoP curves were plotted for the different heat generation profiles, and are shown in Figure 5.18. It can be seen that the region of heat generation has a substantial effect on the MoP throughout the pulse. Although this effect is most severe at short pulse durations, it is noticeable until around 400ms into the pulse.
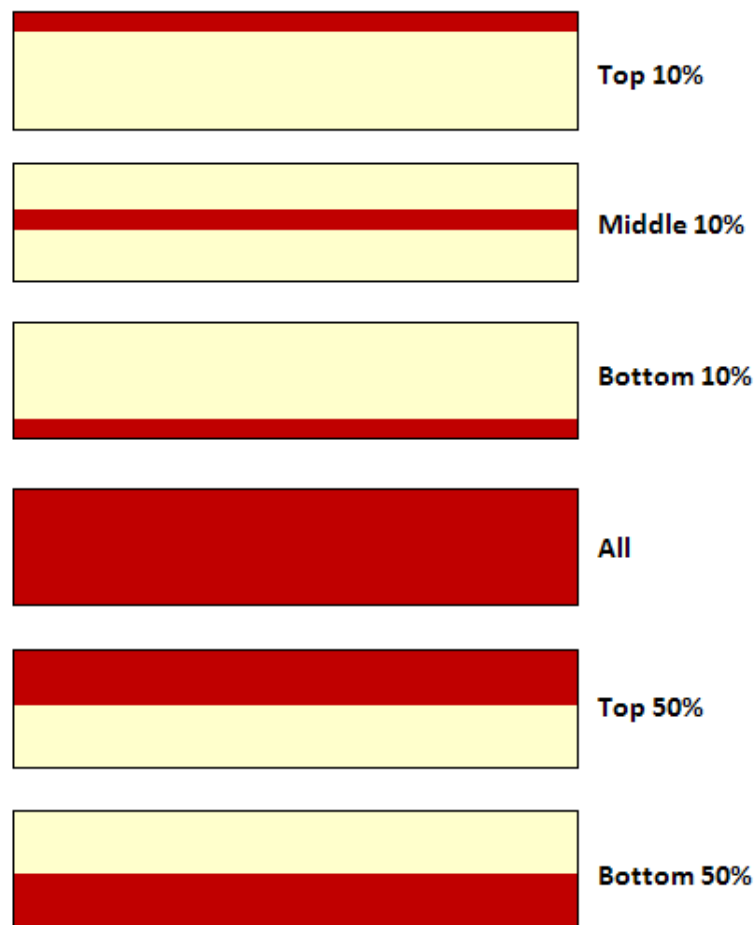
Figure 5.17: Diagram showing the different heat generation profiles modelled within the die, with the heat generation regions (red) and the unheated regions (yellow).

The heat sink is intended to be placed at the top of the die in order to locate it as close to the heat source as possible. When the heat is simulated as being generated lower down in the die, it takes longer for the heat to reach the heat sink. This is seen as a delay in the MoP curve initially decreasing below 1, as the heat sink becomes active later on in the pulse.

After 1ms, when the heat is generated in the bottom 10% of the die the MoP only reaches about 0.97, whereas when the heat is generated in the top 10% the MoP has decreased to between 0.5 and 0.55, depending on the heat sink material. When the pulse has reached 10ms in length the difference in MoP for the two extreme cases is as much as 0.3, and after 100ms it is still as much as 0.1.

It is not until around 400ms that the MoP becomes independent on the heat generation region. The curves are very similar in their approach towards steady state. This is because the variation in thermal resistance between the heat generation and isothermal boundary between the different models is very small. Additionally, the thermal mass of the systems are identical, which indicates the same amount of heat is required to raise the model temperature up to steady state.

The heat sink is most effective when the heat generation occurs as close to the top surface of the die as possible. This reduces the distance the heat must travel to the heat sink. If the heat were generated in the bottom 10% of the die then turning the die upside down would be beneficial, attaching the heat sink close to the region of heat generation.

For maximum heat sink performance, the entirety of the heat should be generated at the top surface. Conversely, when the heat is generated at the bottom of the die surface, the poorest heat sink performance can be expected.
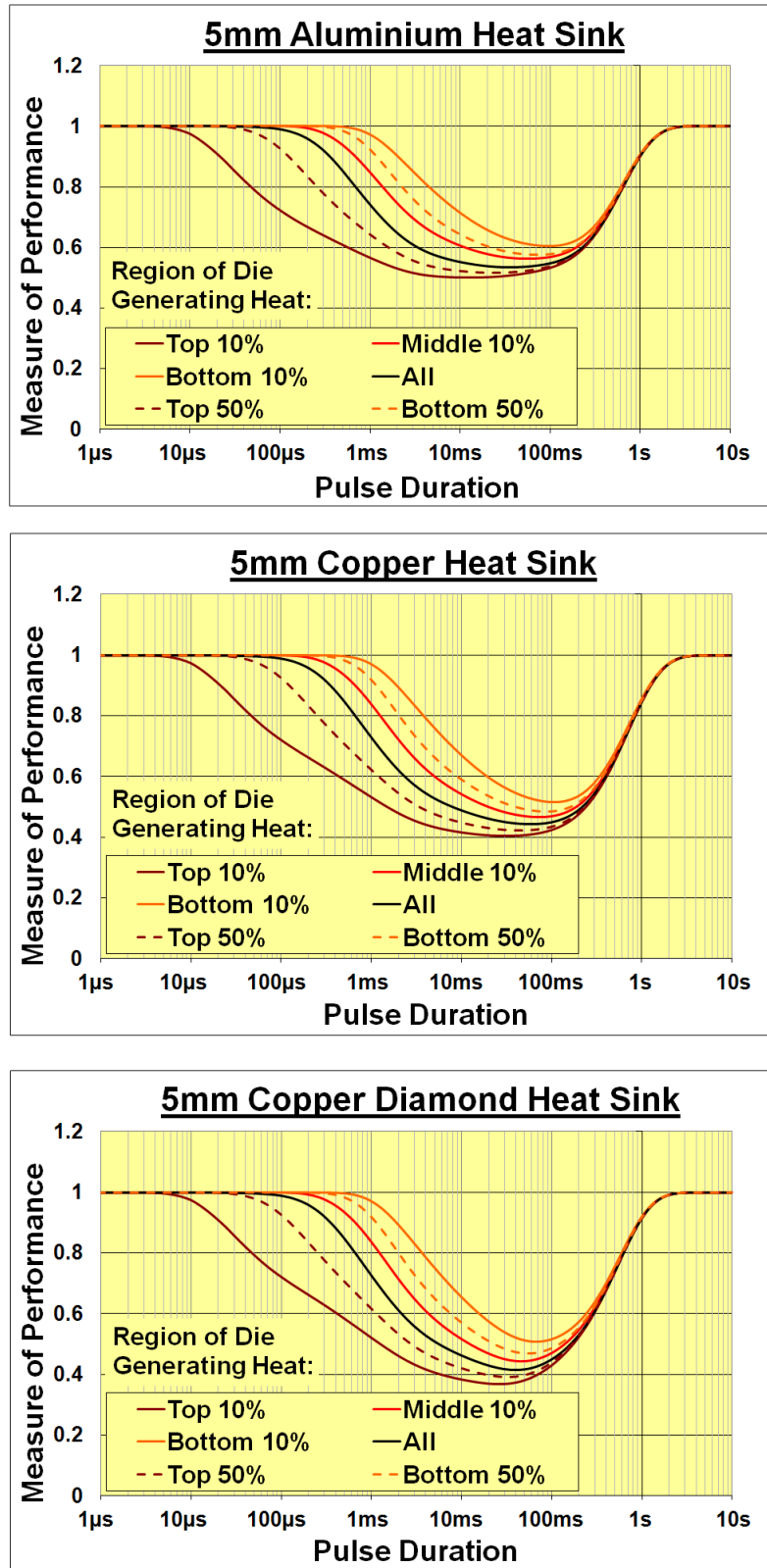
Figure 5.18: Graphs showing how the transient MoP curve changes when the heat generation region within the die is varied.

5.6   Chapter Conclusion

The modelling methodology has been explained in detail. Analysis into the grid density used to model the thermal transients showed that 654 nodes provided a sufficiently accurate solution, which couldn't be improved much much further by adding more nodes. The C++ program written for the thesis is shown to have good agreement with the commercial software FLUENT. This provides confidence in the modelling results obtained. Typically, a 1D model took around 15 minutes to solve for a 10s transient period, which allowed many different scenarios to be explored.

The heat sink performance during a 10s pulse duration has been analysed for a variety of scenarios. In general, the heat sink is most beneficial for pulses between 10ms and 20ms, although reductions to the die temperature are seen for pulses as short as 10µs. It is indicated that heat sinks provide no benefits for long pulses, when a steady state is reached. A MoP value of 0.38 is achievable when using a copper diamond heat sink, occurring at a pulse duration of between 40 to 60ms. For a copper heat sink the minimum MoP value achieved was 0.41, after around 20 to 30ms.

Materials with a high CHP perform better than those with a low CHP early on in the pulse. As the sink sink reaches saturation, the thermal capacity of the heat sink becomes the most important material property. Copper and copper diamond are good candidates for the heat sink material, as both provide significant reductions in device temperature during the pulse range of interest. Aluminium could provide a cheaper, lighter alternative with slightly poorer performance. A 4mm heat sink is shown to be adequate if aluminium were used, whereas a 5mm copper heat sink would be optimum, or an 8mm copper diamond heat sink.

The effects of the stack components on the heat sink performance have been investigated. The baseplate was shown to have significant effect on the heat sink performance for longer pulses. Heat sinks are more effective for thicker baseplates, although thinner baseplates provide lower steady state die temperatures.

For shorter pulses, the solder thickness between the die and heat sink is very influential on die temperature; thinner solder joint are preferable. Analysis on the region in the die where heat is generated has shown to be critical to heat sink performance. Heat sinks perform most effectively when the heat is generated at the top of the die, close to the heat sink.

Chapter 6

2D Modelling

The 1D modelling gave a good indication of how different materials perform as heat sinks, as well as how sensitive the heat sink performance is to changes to different stack-up arrangements. However, the accuracy of the 1D models is limited, as they assume there is no temperature variation across the width of the models. A more accurate representation of the structures can be modelled by using a 2D model domain, which allows the non-uniform width of the stack-up to be modelled, and thus the heat spreading effect of the substrate and baseplate can be accounted for.

The chapter begins by showing the 2D representation of the stack, along with the relevant changes made to the RC network. This is followed by the 2D modelling results, which compares equivalent thermal responses from a 1D and 2D solver. The heat spreading effects are identified from the comparison. A thermal map helps to identify these heat spreading phenomena, which allows the differences in the 1D and 2D measure of performance curves to be explained.

Sensitivity analyses are performed for the heat sink thickness and the heat sink coverage of the die (area of the die covered). The effects of the baseplate's thermal boundary condition on the heat sink performance are also investigated. Composite heat sinks are then investigated, with the aim of utilising HOPG and copper together to improve the heat sink performance.

The chapter concludes with an investigation into the performance of the heat sink during a lightning strike heat pulse profile, as opposed to the constant heat generation scenario used throughout the modelling work.

6.1   2D Model Structures

Constructing the 2D models was done as described in Section 4.3.2 which allowed cylindrical structures to be modelled, as shown in Figure 6.1. The stack-up was modelled as cylindrical, as this could be achieved using a 2D thermal resistance network and applying the cylindrical equations, listed in Table 4.1.
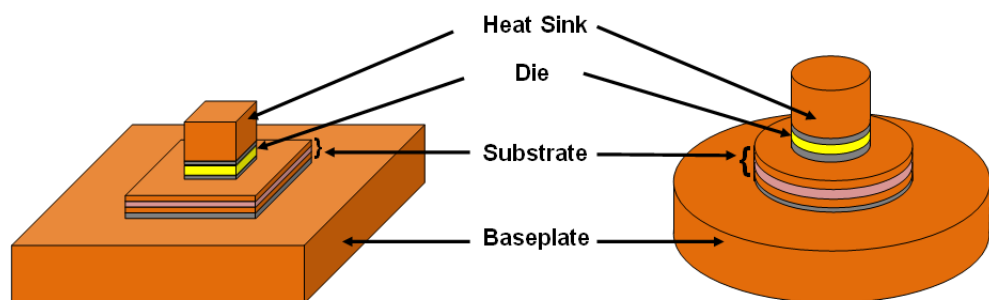
Figure 6.1: Diagram showing the physical structure being modelled (*left*) and the representation of the structure achieved using 2D modelling (*right*).

Two types of model were created - a traditional device stack and a device stack with the addition of a heat sink. Simulating a heat pulse in both the models allowed the MoP of the heat sink to be obtained.

The same standard thickness for each layer was the same as the thicknesses used for the 1D modelling. The width of the models used was not the standard width as listed in Table 1.1. Equivalent widths that provided the same cross sectional area as the

physical stack were calculated using eq.(6.1). A square layer which was width and depth of length w has an area of w^2. A circle with the equivalent surface area would have a radius $r_{equivalent}$:

$$r_{equivalent} = \sqrt{\frac{w^2}{\pi}} \tag{6.1}$$

Table 6.1 lists the width (equivalent radius) and thickness of the layers in the 2D models, along with the materials of each layer. Layers shown in italics are the layers included only when a heat sink is modelled.

Layer	Thickness (mm)	Width (mm)	Equivalent Radius (mm)	Material
Heat Sink	*varied*	*10*	*5.64*	*varied*
Solder	*0.1*	*10*	*5.64*	*Solder SnAg 3.5*
Cathode	0.02	10	5.64	Aluminium
Die (Heated Region)	0.04	10	5.64	Silicon
Die (Unheated Region)	0.36	10	5.64	Silicon
Solder	0.1	10	5.64	Solder SnAg 3.5
Substrate: Copper	0.3	29	16.36	Copper
Substrate: AlN	0.6	29	16.36	Aluminium Nitride
Substrate: Copper	0.3	29	16.36	Copper
Solder	0.1	29	16.36	Solder SnAg 3.5
Baseplate	5	60	33.85	Copper

Table 6.1: The layers modelled showing the 2D dimensions, where the width is used for Cartesian models and the equivalent radius is used for the Cylindrical models.

6.1.1 Heat Generation and Boundary Conditions

A uniform body heat flux was applied to the top 10% of the die, replicating the conditions applied to the standard 1D model. The boundary conditions were also kept the same. All walls were modelled as adiabatic, except for the entire bottom surface of the baseplate which was modelled as isothermal (at 25ºC, the initial starting temperature throughout the model).

The two types of model are displayed in Figure 6.2, which shows the layers in each model, the region of the model generating the heat and the boundary conditions applied. Modelling the side walls as adiabatic is not entirely realistic, but allows the

worst case scenario to be investigated. In reality there would be a small amount of heat transfer across these surfaces.
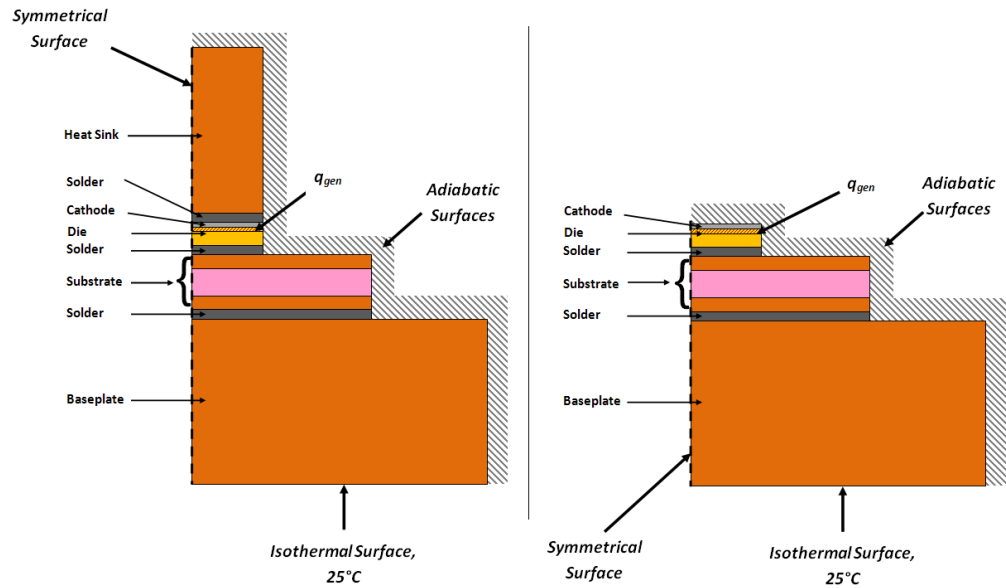
Figure 6.2: The 2D domain setup for the two model types: with heat sink (*left*) and without heat sink (*right*). Standard boundary conditions applied are indicated.

6.2   2D Modelling Results

6.2.1   Heat Spreading Effects

The different layers in the device stack are not all of constant width. As heat reaches the substrate and baseplate, it is able to conduct laterally, spreading into the additional width which becomes available. This creates a non-uniform lateral temperature distribution.

2D modelling allows this heat spreading to be modelled, allowing the effect of the heat spreading on the performance of a heat sink to be investigated. A pair of 2D models were constructed, one with no heat sink and one with a 5mm copper heat sink using the standard layer dimensions in Table 6.1. The models were discretised and a 10s constant magnitude heat pulse simulated in the top 10% of the die, as described in the previous chapter.

The maximum die temperature in each model was calculated during the pulse, which allowed the MoP curve to be obtained. The same situation has already been modelled in 1D. This allowed the 1D and 2D MoP curves of the same structures to be plotted, which is shown in Figure 6.3.

It can be seen that the curves in Figure 6.3 are identical for short pulse durations, up to around 10ms at which point the first divergence occurs. As the width of all layers above the substrate are constant, there should be no temperature variation across the width in the 2D model until the heat has reached the substrate. Thus, the temperatures obtained using the 1D and 2D setup should be identical.

Once the heat reaches the substrate, the additional width in the 2D models provide an extra dimension for the heat to spread. This effect is not likely to be instantaneously obvious as the heat must spread outwards, creating a gradual variation between the 1D and 2D models. The starting point of this divergence is at around 10ms, which
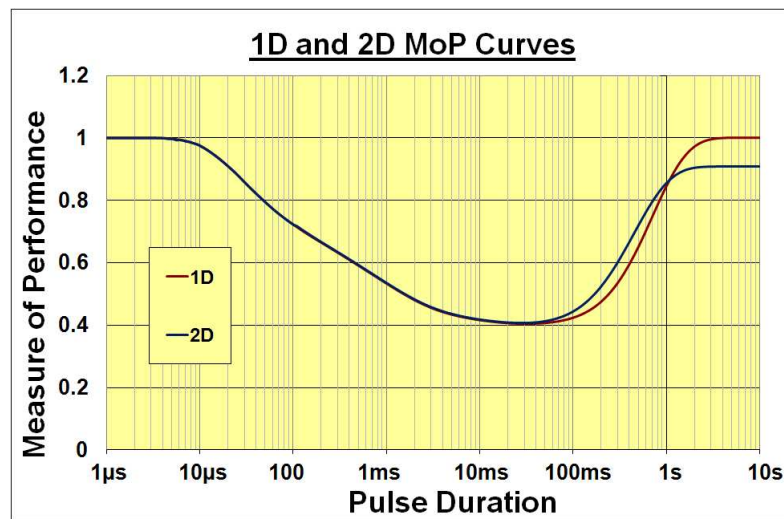
Figure 6.3: A comparison of the MoP curves when a 5mm copper heat sink is modelled in 1D and 2D.

corresponds reasonably well to the time constant for the heat arriving at the baseplate (Table 5.1).

As the heat travels into the substrate and then the baseplate the additional thermal mass in the 2D models becomes gradually more available to the heat, leading to the increasing variation between the 1D and 2D curves. The extra thermal mass within the substrate and baseplate of the 2D models reduces the benefits of the heat sink, causing an increase in MoP compared to the 1D model.

The minimum MoP occurs at a pulse duration of around 30ms to 40ms. This is a very similar value for both the 1D and 2D, models as the divergence of the curves is minimal. The only major difference between the 1D and 2D modelling occurs when the heat has time to exploit the additional thermal mass in the baseplate. Therefore, the 1D modelling provides an accurate MoP curve until the heat spreading effects become significant, at around 40ms into the pulse.

For pulse durations between 40ms and 1s the 2D MoP curves are greater than the 1D curve. They follow a similar gradient, and have a maximum difference of around 0.065. Once the pulse duration reaches 1s the solutions start to approach steady state. For 1D modelling, the steady state value is always 1. However, the steady state MoP

values from the 2D modelling is 0.91. The 1D models showed that the heat sink temperature was uniform throughout and equal to the maximum die temperature. This is not true for 2D models, which is why the steady state MoP value does not return to 1.

Heat spreading occurs within the baseplate which creates additional heat paths between the die and cooled baseplate boundary. Instead of the heat flowing only downwards as it does in the 1D models, heat can spread sideways into the baseplate and the substrate, creating lateral temperature variations. The heat generated at the outer region of the die now has access to more thermal mass. This causes a non-uniform cross sectional temperature within the die, the hottest part of the die being the centre and the coolest the edge.

Figure 6.4 illustrates the steady state temperature distribution through the cross section of a model with a 5mm copper heat sink. Heat which flows upwards into the heat sink from the centre of the die is spread laterally due to these temperature gradients, indicated by arrows in Figure 6.4. From the edge of the heat sink it flows back through the die and follows the heat paths to the baseplate boundary. Heat flow from the centre of the heat sink to the edge is greatest closer to the die. Further inside the heat sink the lateral temperature gradient decreases, which reduces the heat flow, indicated by variation in arrow size.

In the 2D model, the heat sink performs a heat distribution role during steady state operation. Heat passes into it from the centre of the die where it is at its hottest, and then leaves the heat sink at the edges of the die. This evens out the variation in temperature across the die, reducing the maximum temperature in the centre and increasing the temperature at the edges. Models without a heat sink do not receive this heat redistribution as there is no thermal mass on top of the die, which causes the variation in die temperature to be greater. The steady state temperature distribution inside a model with no heat sink is illustrated in Figure 6.5, which demonstrates the hotter die centre and cooler die edge.

A comparison of the temperature from the centre to the edge of the die at its hottest

point is shown in Figure 6.6. Here the levelling out effect caused by the heat sink can be seen. As the MoP is calculated using the maximum temperature in the die, it is expected to be less than 1 at steady state due to the difference in temperature variation when a heat sink is added.
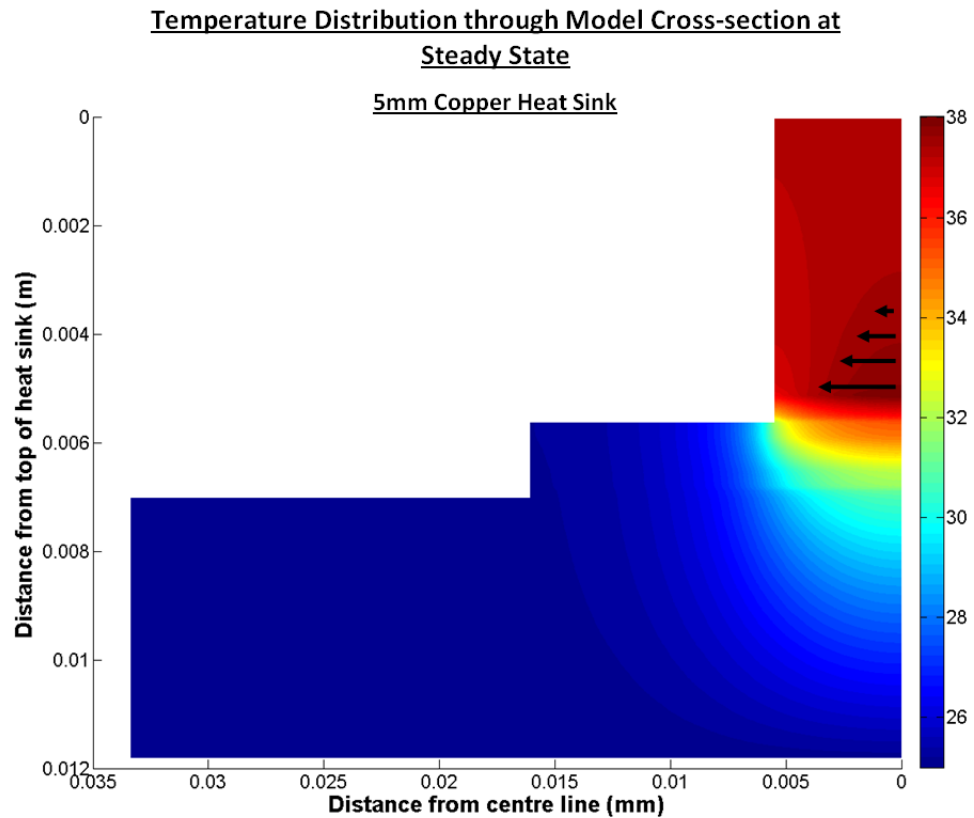
Figure 6.4: The steady state temperature distribution through the cross section of a 2D model which includes a 5mm copper heat sink. The arrows indicate the lateral heat flow in the heat sink which reduces the die temperature in the centre.
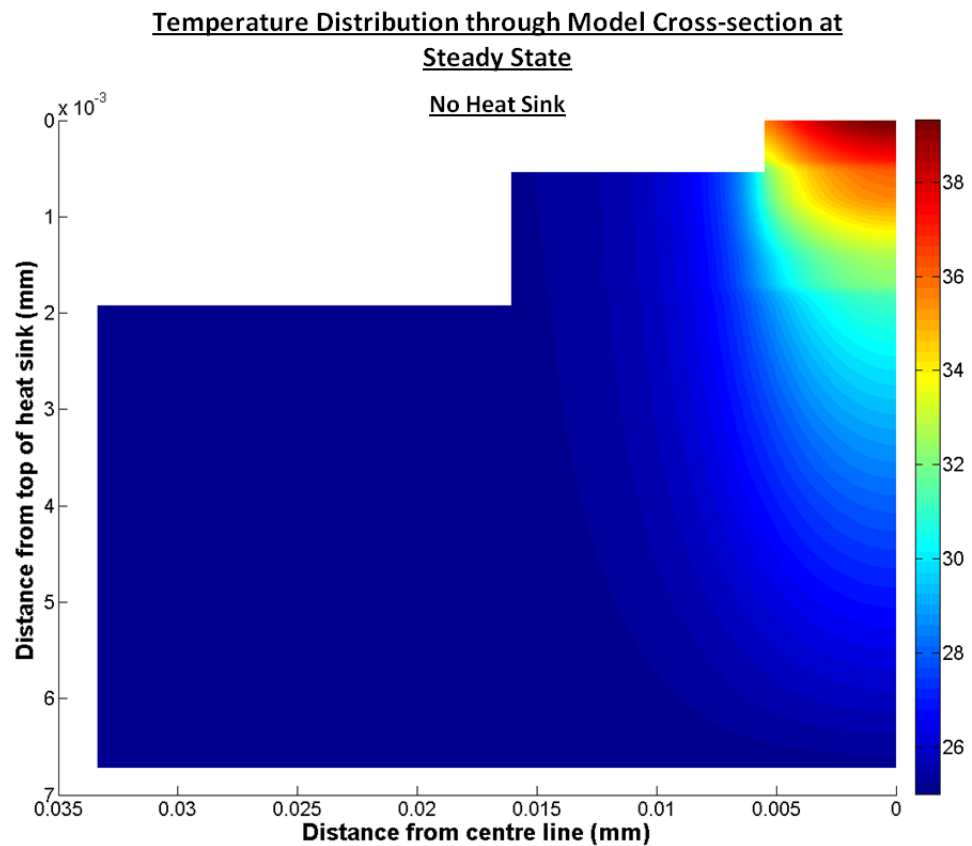
Figure 6.5: The steady state temperature distribution through the cross section of a 2D model which includes a 5mm copper heat sink. With no heat sink present, heat is not distributed across the die which creates a hotter die centre and cooler temperature at the die edge.
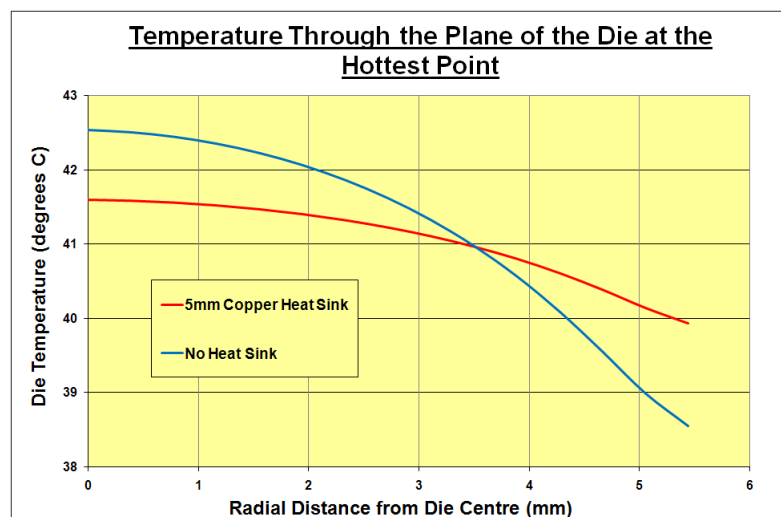
Figure 6.6: A comparison between the temperature through the cross surface of the die at its hottest point. When a heat sink is used the temperature becomes levelled out, reducing it in the middle but increasing it at the edges.

6.2.2 Heat Sink Thickness

The way that the reduction in die temperature during the pulse is affected by the thickness of the heat sink has been investigated using 1D analysis in Section 5.5. It has also been shown that the difference between 1D and 2D modelling is not noticed in the MoP results for pulse durations shorter than around 40ms. However, the 2D modelling has shown that the heat sink reduces the steady state temperature of the die.

A 5mm copper heat sink reduces the maximum steady state temperature rise in the die at steady state by 9%. The magnitude of this die temperature reduction at steady state was investigated by modelling different heat sink thicknesses, from 0.5mm to 5mm. The standard stack setup was used (Table 6.1), and the temperature rise in the die was calculated when the different thickness heat sinks were modelled. The MoP curve for each heat sink was plotted, and can be seen in Figure 6.7.
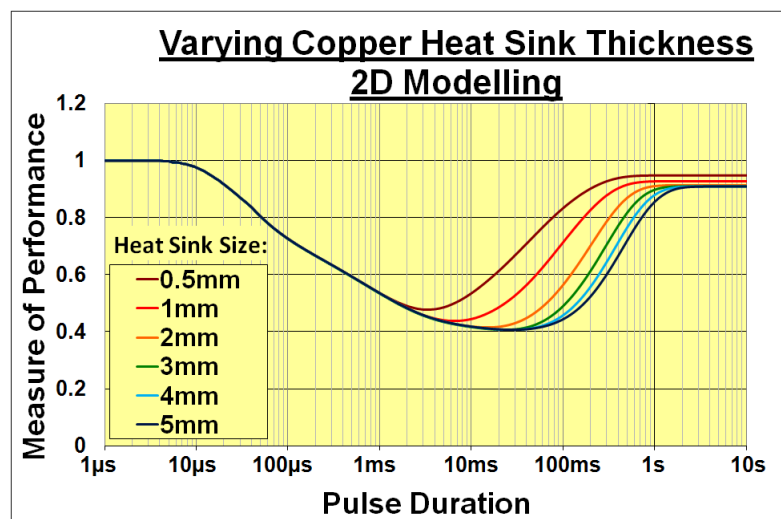
Figure 6.7: The MoP curves throughout a 10s pulse for copper heat sinks of different heat sinks, modelled in 2D.

For the early stages of the pulse, up to 4ms, the MoP curves are the same as seen when modelled in 1D. Beyond this point, the MoP of each heat sink is not consistent with the 1D modelling results, as explained in Section 6.2.1. However, curves behave in the same way and follow the same trends as the 1D MoP curves.

The region of interest, which is where the major variation from the 1D MoP curves lies, is at steady state. Figure 6.7 shows that the steady state MoP value is dependent on the thickness of the heat sink.

The heat sink need not be very large to introduce an effective steady state temperature reduction; a 0.5mm heat sink is able to reduce the steady state die temperature by 5%. This is increased to a 7% reduction by using a 1mm heat sink. The steady state MoP values for the 2mm, 3mm, 4mm and 5mm heat sink are almost indistinguishable from the graph, ranging by just 0.4% up to a maximum of a 9.1% reduction (achieved when using the 5mm heat sink).

As the heat spreading effect occurs in the bottom region of the heat sink, close to the die, the thickness of the heat sink has only a limited effect. Benefits are seen with the heat sinks up to 2mm in thickness, however further increases to the heat sink thickness do not reduce the steady state temperature further.

6.2.3 Baseplate Boundary Condition

All modelling so far has assumed an isothermal boundary condition applied at the bottom of the baseplate. This assumes a worst case scenario for the heat sink, as the cooling at the baseplate is as effective as possible.

The affect of the baseplate boundary condition on the heat sink performance was investigated using 2D modelling. Using the standard stack dimensions in Table 6.1, the MoP curve for the three heat sinks (5mm aluminium, copper and copper diamond) was obtained when an adiabatic boundary condition is applied to the baseplate surface. By comparing the isothermal and adiabatic scenarios, the two extreme conditions are compared: infinite thermal resistance (adiabatic) and zero thermal resistance (isothermal). The two MoP curves from each scenario create an envelope which all other thermal boundary conditions must lie within. This envelope of MoP curves is shown for each heat sink material in Figure 6.8.

Changing the boundary condition does not affect the MoP curve until around 15ms into the pulse. This corresponds well with the time constant for the heat to travel to the bottom of the baseplate, which is calculated as 177ms in Table 5.1. It is not until this point into the pulse that the boundary condition affects the temperature of the die. As this is outside the pulse range at which the heat sink needs to be effective (up to 100ms), the boundary condition is not a very important factor for the performance of the heat sink.

The divergence of the curves happens at the same time for each heat sink, and the behaviour of the curves after this point is near identical. When the baseplate boundary first influences the die temperature, the heat sink is more beneficial when an adiabatic boundary is modelled. When less heat is transferred across the boundary, the heat must be stored within the structure. Thus, the heat sink provides an extra thermal mass for the heat.

This trend continues as the structures heat up to steady state. This occurs after around 3s when the baseplate boundary is isothermal, which is a lot quicker for the

adiabatic models. When no heat is transferred out of the system, all heat stays within the stack-up. The baseplate is a large volume of high thermal capacity material, which requires a large amount of heat to raise up to a steady state temperature distribution. It is not until this happens that steady state can be achieved under these conditions. In the adiabatic scenario, steady state is not achieved as the generated heat cannot leave the model. A steady state MoP value is achieved when the temperature is increasing at the same rate throughout the models.

Steady state values for the isothermal models are based on the thermal conductivity of the heat sink material. The smaller the thermal resistance, the easier heat can travel into the heat sink, towards the edge of the heat sink, and then back through the die to the baseplate as depicted in Figure 6.4. The more heat which follows this thermal path, the smaller the die temperature variation in Figure 6.6. Copper diamond has the greatest thermal conductivity of the three heat sink materials, and therefore gives the lowest steady state MoP of 0.89, followed by copper (0.91) and then aluminium (0.93).

In the adiabatic baseplate scenario, the steady state MoP is not determined by the thermal conductivity, but by the thermal mass of the heat sink. As all generated heat stays within the model, the same amount of heat remains constant between the models at any time period. The temperature at which this is stored depends on the thermal capacity and mass of each material. The summed thermal mass ($specific\,thermal\,capacity \times mass$) of each layer provides the total thermal mass available. The greater this total, the lower the overall temperature in the model at any time during steady state. Moreover, the smaller this total, the sooner the model reaches the steady state temperature distribution.

This trend is reflected in Figure 6.8. Although none of the adiabatic MoP curves reach steady state within the 10s pulse, the adiabatic MoP curve for aluminium is closest to reaching steady state, indicated by the flattest gradient. All three adiabatic MoP curves have increased above their isothermal counterparts by the end of the 10s pulse. The final MoP value which each curve is expected to reach can be calculated

based on the total thermal mass in each model run:

$$Steady\,State\,MoP\,Value = \frac{Total\,Thermal\,Mass\,In\,'No\,Heat\,Sink'\,Model}{Total\,Thermal\,Mass\,In\,'With\,Heat\,Sink'\,Model}$$

(6.2)

The total thermal mass in each of the four models is shown in Table 6.2. Also shown in the table is the calculated expected steady state MoP for each heat sink, based on eq.(6.2).

	Total Thermal Mass in Model (J/K)	Ratio of Total Thermal Mass to 'No Heat Sink' Thermal Mass
No Heat Sink	65.08	-
5mm Aluminium Heat Sink	66.31	0.981
5mm Copper Heat Sink	66.82	0.974
5mm Copper Diamond Heat Sink	66.31	0.981

Table 6.2: The total thermal mass in the 2D models, and the steady state MoP values expected.
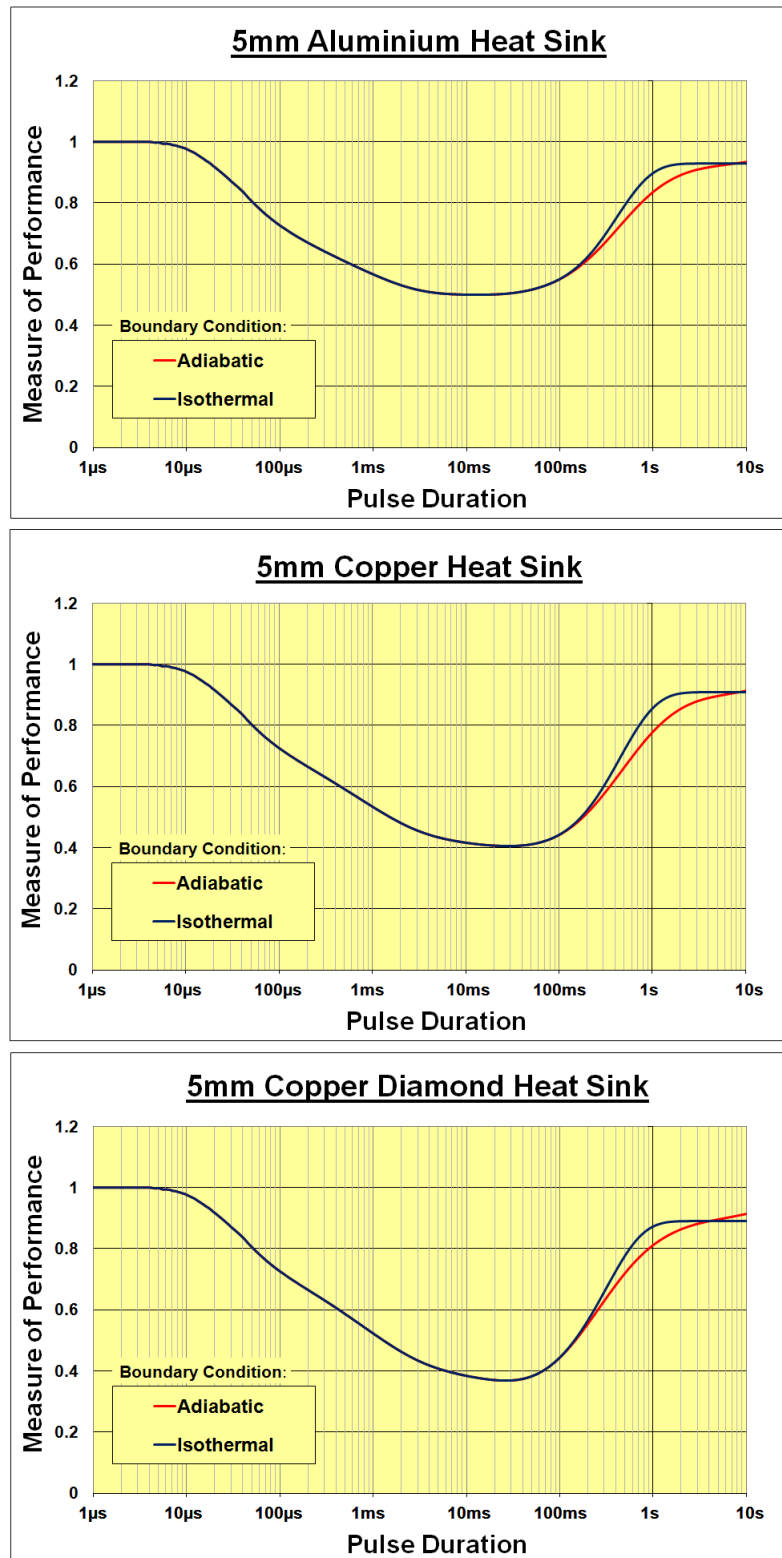
Figure 6.8: Graphs showing the MoP curves from for adiabatic and isothermal base-plate boundary conditions for each heat sink material: aluminium (top left), copper (top right) and copper diamond (bottom).

6.2.4   Heat Sink Coverage of Die

The heat sink has covered the entire surface of the die in all the models. This allows the maximum area for heat transfer between the heat source and the heat sink. When modelling in 1D, this is the only scenario that can be explored. However, 2D modelling allows the contact area between the heat sink and die to be changed and the effect of doing so be evaluated.

The contact area between the die and heat sink was varied, down to a minimum of 80% die area coverage. In each case a 5mm heat sink was modelled. Figure 6.9 shows the MoP curves obtained.
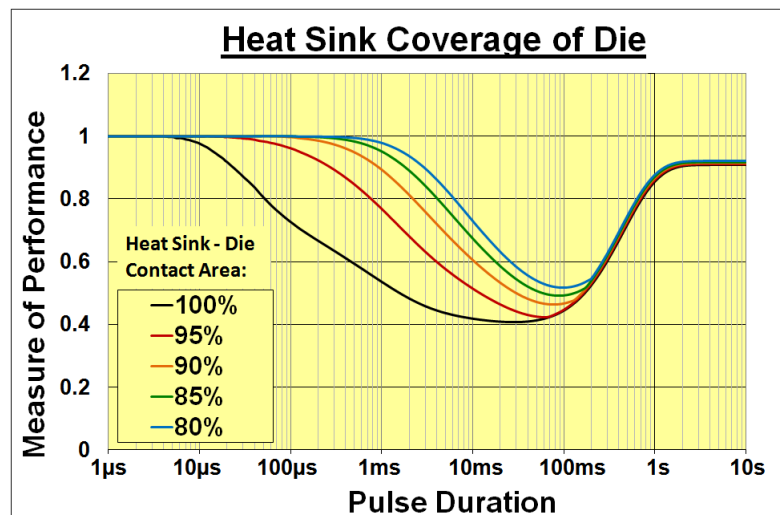
Figure 6.9: The MoP curves when the amount of the die surface in contact with the heat sink varies.

Reducing the heat sink-die contact area affects the heat sink performance significantly. This is particularly true for short pulse durations. When the coverage is reduced from 100% to 95% a maximum difference in MoP during of over 0.25 exists, occurring between 150µs and 550µs into the pulse. The maximum die temperature from the 95% coverage model is greater than the 100% coverage model until the pulse length reaches 60ms.

Further decreases in the contact area diminishes the heat sink effectiveness further. Reducing the coverage to 80% means that the MoP is increased by a minimum of 0.07

between 100μs and 100ms, which is the pulse range of most interest. A maximum difference in MoP of 0.45 is seen at after 1ms into the pulse. This makes the heat sink between 7% and 45% less effective during this region.

The temperature profile across the die surface at its hottest point can be seen in Figure 6.10 at different stages during the pulse. The temperature profiles start with large variations, which smooth out with time. Early on in the pulse a large temperature step occurs towards the edge of the die, where there is no heat sink. As the pulse progresses, heat travels down this temperature gradient which flattens out the temperature step.

Once the heat reaches the baseplate, heat spreading aids in the cooling of the edge of the die. This is evident in the temperature profiles at 10ms and 100ms; the difference in temperature between the edge and the centre of the die is much less pronounced than it is after 100μs.

As the MoP is calculated based on the maximum die temperature and does not take into account variation in temperature across the die, large temperatures at the edge of the die, such as those in Figure 6.10, cause the MoP to be increased. This leads to the shape of the MoP curves seen in Figure 6.9: the MoP decreases as the temperature step at the edge of the die smooths out. The more of the die that is exposed at the edge, the longer it takes for the heat generated at the outer most part of the die to reach the heat sink, resulting in a larger MoP for a longer period of time.
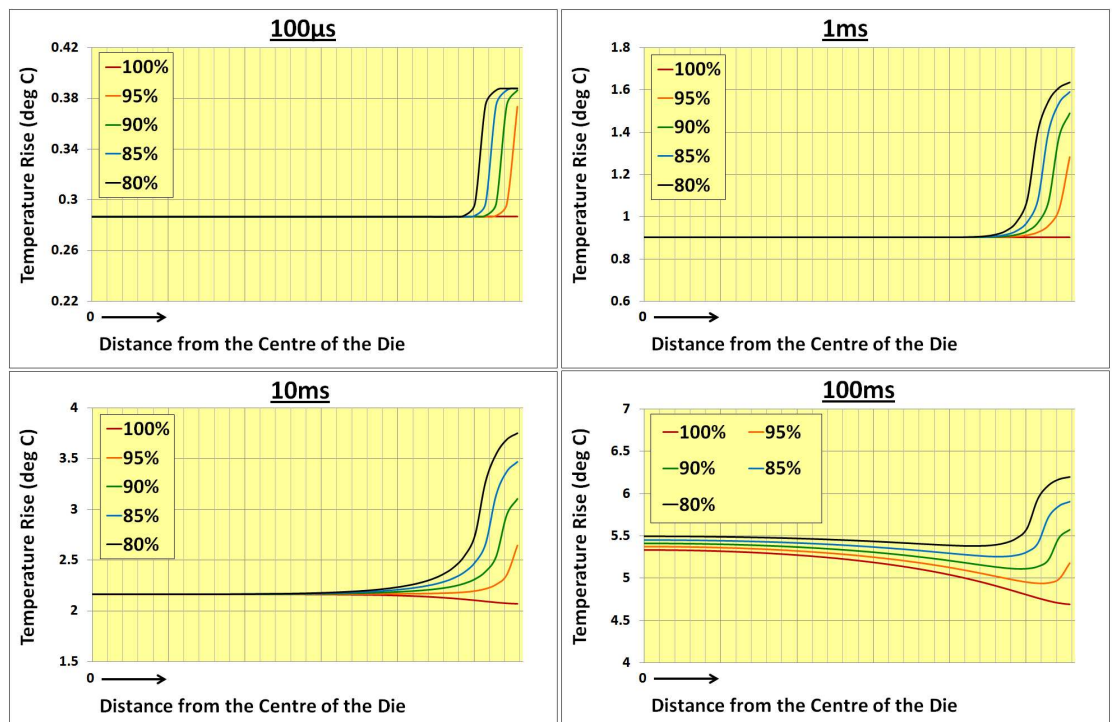
Figure 6.10: The temperature distribution across the surface of the die, at the hottest point, when the heat sink covers a different percentage of the die surface. Each graph shows the distributions at a different stage of the pulse, indicated by the graph headings.

6.2.5   Composite Heat Sinks

In the 1D modelling, the 5mm HOPG heat sink was very good at reducing the die temperature for short pulse durations (see Section 5.4.1). However, for pulses greater than 20ms in length it proved to be less effective than the copper diamond heat sink. The copper heat sink was also more effective than the HOPG once the heat pulse reached 40ms in length. Furthermore, the anisotropic nature of HOPG's thermal conductivity was not modelled at that stage, as the modelling was done in one-dimension. The thermal conductivity of HOPG in the direction modelled is 1500W/mK, but is just 15W/mK in the lateral direction. Once the heat spreading effects have been taken into account, the performance of a HOPG heat is expected to decline further at the later stages in the pulse.

The idea of using a composite heat sink was considered. A heat sink made of HOPG and copper could allow the early cooling benefits of HOPG to be combined with the more sustained cooling benefits that copper provides. HOPG can carry heat away from the heat source quickly due to its high thermal conductivity in the plane of the heat wave. The copper part of the heat sink would allow a higher thermal capacity than a pure HOPG heat sink, and would also provide better heat conduction in the lateral direction.

Different configurations of heat sinks were modelled in 2D using a combination of HOPG and copper regions. In essence, two setups were considered: HOPG layers and HOPG columns. HOPG layers used a layer of HOPG at the bottom of the heat sink, with a layer of copper on top. HOPG columns explored using a copper heat sink with different sized (and numbers of) HOPG columns running vertically from the heat source to the top of the heat sink. Figure 6.11 gives a graphical indication of the structure of the heat sinks.

Figure 6.12 depicts the different HOPG column heat sinks that were modelled. In addition to the composite heat sinks, pure HOPG, copper and copper diamond heat sinks were also modelled for comparison. The maximum die temperature was calcula-
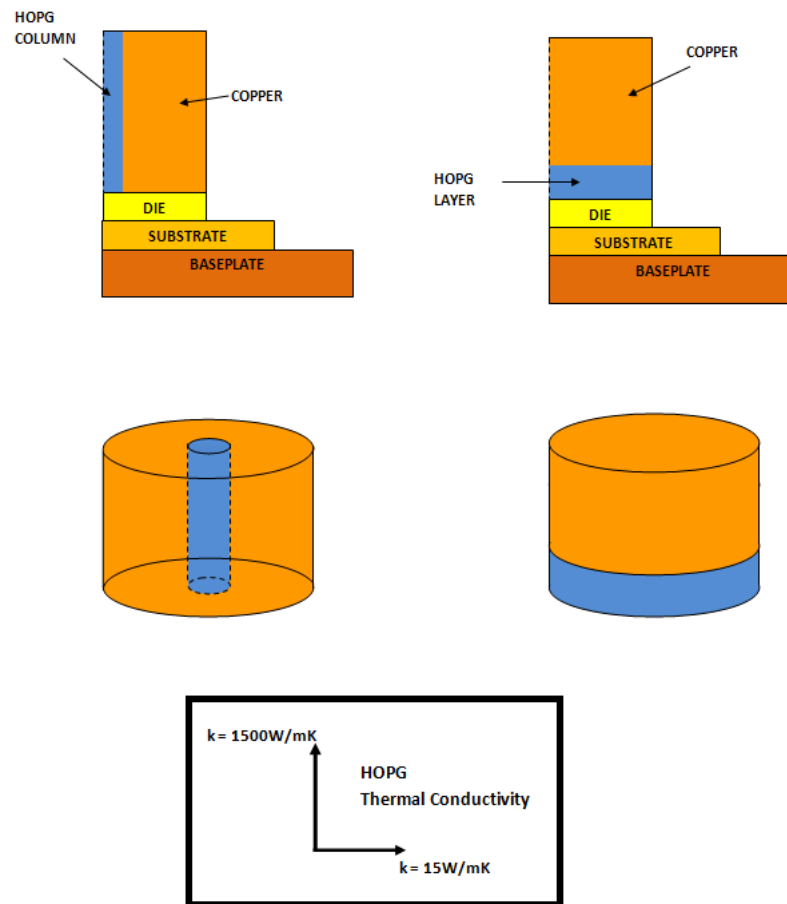
Figure 6.11: A graphical representation of the composite heat sinks - HOPG Columns (left) and HOPG Layers (right). The top diagrams shows the placement of the heat sink on the stack, whist the 3D pictures (*bottom*) better indicate the formation of the heat sink structures.
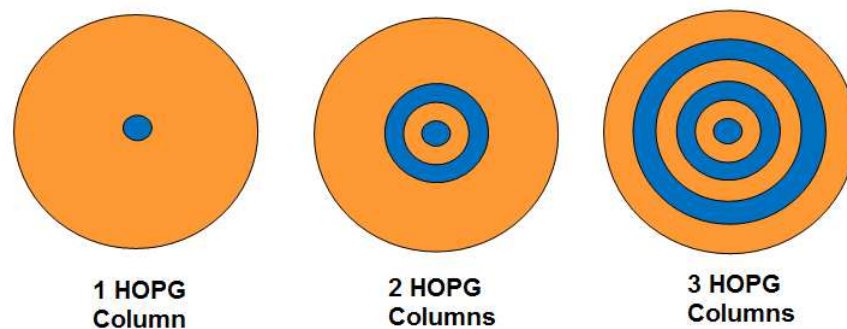
Figure 6.12: Top-view of the three 'HOPG column' heat sinks modelled. The width of each layer is 1mm, except the outer copper layer which is 1.14mm wide.

ted during the 10s constant magnitude heat pulse for each model. Figure 6.13 shows the MoP curves for the different composite heat sinks.
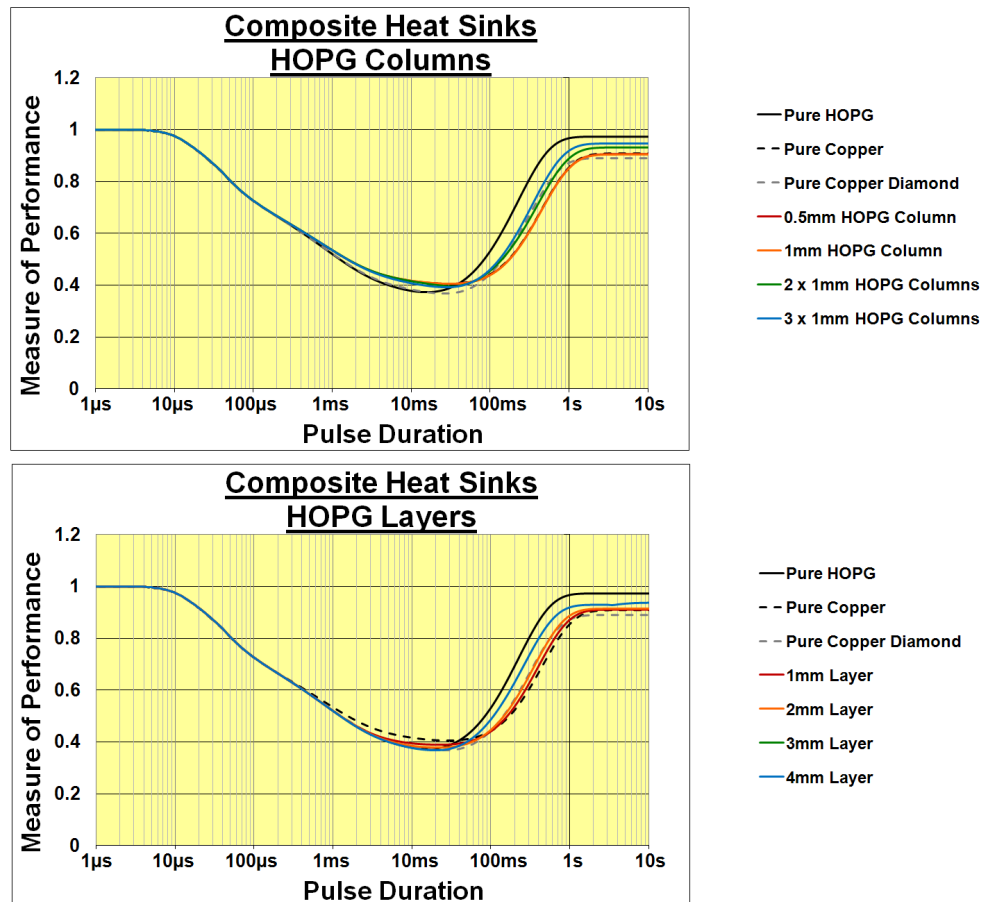
Figure 6.13: The measure of performance of different copper-HOPG composite heat sinks. Top: Composite heat sinks in a column configuration. *Bottom*: Composite heat sinks in a layered configuration.

6.2.5.1   HOPG Columns

When the HOPG was arranged in columns the maximum die temperature was not affected much during the early stages of the pulse. The columns helped to conduct heat away from the die, but only in localised regions of the die. The effect was to lower the average die temperature, but not reduce the maximum temperature.

Figure 6.14 shows the temperature distribution across the surface of the die between the centre and the edge at three different times during the pulse: after 1ms, 10ms

and 100ms.

After 1ms, the heat is yet to reach the baseplate and therefore the heat has not had chance to spread laterally. This means the temperature across the die is constant when a pure copper heat sink is used. It is evident that the addition of HOPG columns reduces the temperature in regions of the die, directly beneath the columns. Although this helps to cool the die as a whole, the maximum temperature of the die is not reduced compared to the pure copper heat sink.

As the pulse length increases to 10ms, heat spreading into the baseplate (and to a lesser extent the substrate) is occurring. The temperature at the edge of the die is noticeably cooler than the centre, which is the location of the maximum temperature. As the HOPG columns are in the centre of the heat sinks, the maximum temperature in the die is lower than that of the pure copper heat sink. The heat sinks with multiple columns do this more effectively, as a greater proportion of the heat sink to die contact area is occupied by HOPG. The heat sink with three columns is the most effective at this point. The die beneath the copper region between the columns now benefits from the reduced temperature around it, and therefore the temperature across the entire die surface is reduced.

By 10ms into the pulse, the thermal capacity of the heat sink is influential on the die temperature. The heat sinks with a greater proportion of HOPG have the greatest maximum die temperature. The heat sinks with two and three HOPG columns have a greater die temperature than the die with a pure copper heat sink. Noteworthy is that by this stage the variation in die temperature that the columns created early on in the pulse are dying out, and the shape of the temperature profile is becoming increasingly similar to that of the pure copper case.

6.2.5.2   HOPG Layers

When the composite heat sink is comprised of layers rather than columns, the temperature distribution across the die surface is similar to that in Figure 6.14. The heat
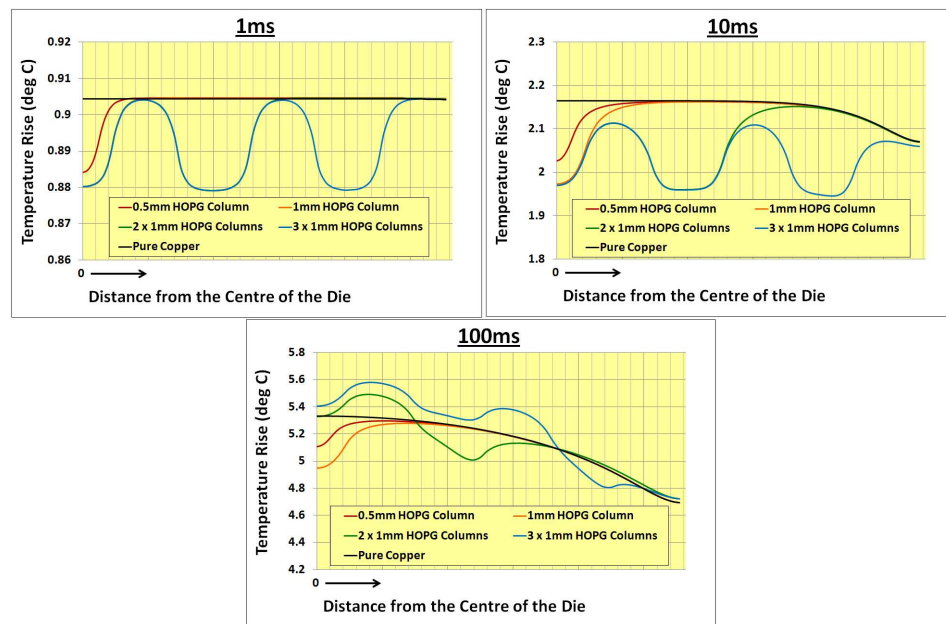
Figure 6.14: The temperature between the centre and the edge of the die for the different composite heat sinks in a column configuration. The graphs show the profiles at different times during the pulse - after 1ms, 10ms and 100ms.

sink composition across the die is uniform, therefore fluctuations in die temperature are not seen.

Figure 6.13 shows the MoP curves from using a heat sink with a layer of HOPG close to the die with a layer of copper on top. Combining the materials in this way allows the maximum die temperature to be reduced early on in the pulse, relative to the pure copper heat sink. In this region, the high thermal conductivity of the HOPG allows heat to be removed from the die quickly and evenly, providing a close match to the pure HOPG MoP curve.

The die temperature when using a pure HOPG increases above the die temperature from a pure copper heat sink after around 40ms into the pulse. This does not happen when a layered heat sink is used. The high thermal capacity of the copper section of the composite heat sink increases the overall thermal capacity of the heat sink. This reduces the die temperature at the later stages of the pulse, from 40ms onwards, compared to the HOPG heat sink. However, the die temperature is greater than it is when a pure copper heat sink is used.

The thickness of the HOPG and copper layers affects the MoP at different pulse durations. A thicker HOPG layer allows the pure HOPG MoP curve to be followed for longer, reducing the die temperature more for pulse durations up to 20ms in length. The copper layer provides benefits at the later pulse durations.

6.2.5.3 Composite Heat Sink Conclusions

Of the different composite heat sinks investigated, those with the layered configuration provided the best performance. The HOPG columns reduced the die temperature well, but only in isolated locations. Consequentially, the maximum die temperature was not reduced much throughout the pulse. The layered heat sinks allowed the thermal properties of HOPG and copper to be combined the most effectively.

During early parts of the pulse, the die temperature was reduced as much as it is when using a pure HOPG heat sink. Later on in the pulse, when the pure HOPG heat sink becomes much less effective, the layered composite heat sink continued to cool the die effectively. At the longer pulse durations, the MoP curve is closer matched to the copper curve than the HOPG curve.

Although combining HOPG and copper to from a hybrid heat sink has shown to decrease the die temperature more than either can on their own, the improvements are limited. The complications in making the hybrid heat sink outweigh the potential benefits. A method of bonding the materials together would need to be determined. This bonding would more than likely introduce an extra thermal resistance at the material interface, decreasing the thermal performance of the heat sink.

6.3   Lightning Strike Simulation

For simplicity and consistency, the heat pulse simulated in the modelling work thus far has been of constant magnitude throughout the duration of the pulse. In reality, the heat pulses endured by power electronics devices are of varying length and forms. The form of the pulse would affect the performance of the heat sink, as it is dependent on factors such as the time at which the pulse peak occurs.

6.3.1   Lightning Strike Form

Lightning strikes tend to have a very early peak, followed by an exponential decay. One known lightning strike form is the '10-120' pulse. The shape of this curve is shown in Figure 6.15. The current increases linearly up to the peak which occurs after 10μs, before decaying exponentially back to the original level in a way such that 120μs after the start of the pulse, the pulse magnitude is half of the peak magnitude.
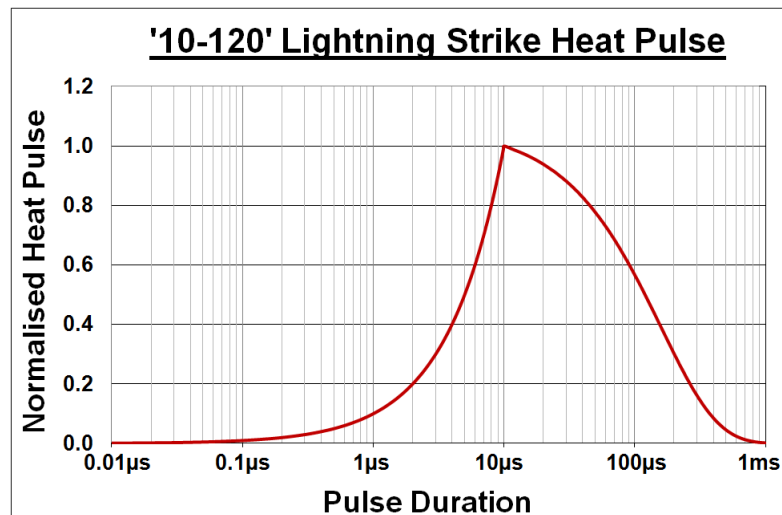
Figure 6.15: The magnitude of the heat generation throughout the '10-120 Lightning Strike' heat pulse.

6.3.2   Heat Sink Performance during Lightning Strike

Modelling work was undertaken to understand how well the heat sinks reduced the die temperature during a lightning strike heat pulse. The standard 2D model setup was used (Table 6.1), along with the boundary conditions shown in 6.2. Heat generation occurred within the top 10% of the die, with varying magnitude such that the heat pulse in Figure 6.15 was simulated, up to a pulse duration of 100ms (from 10ms onwards, no heat was generated).

Aluminium, copper and copper diamond heat sinks were all tested, each 5mm thick. The maximum temperature rise in the die during the pulse is shown alongside the heat pulse in Figure 6.16.
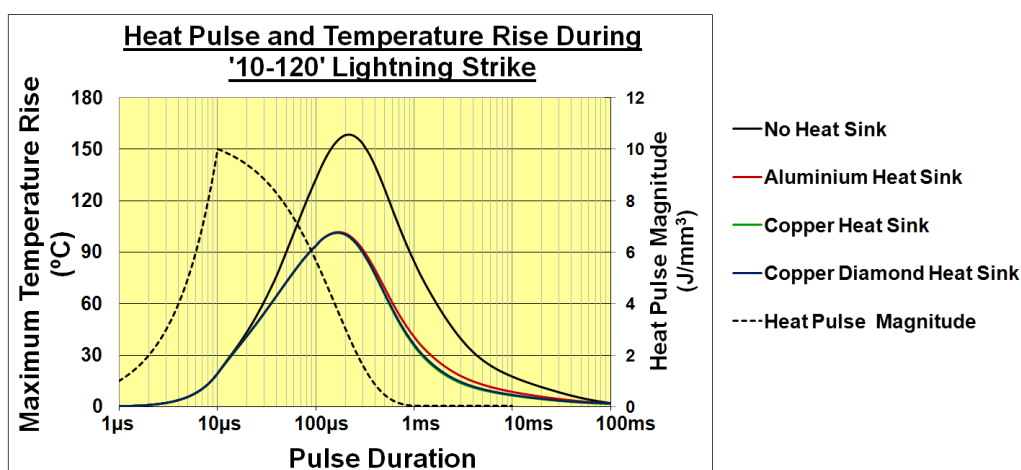
Figure 6.16: The temperature rise in the different models during a lightning heat pulse. The heat generation magnitude is also shown during the pulse.

Interestingly, the peak temperature rise does not correlate with the peak of the heat pulse. The maximum heat pulse magnitude occurs after 10µs, whilst the maximum temperature rise in the die does not occur until around 200µs. Although the heat pulse magnitude decreases after 10µs, heat is still being generated in the die which causes the die temperature to increase until the rate of heat conducting out of the die to the stack (and heat sink if present) exceeds the amount of heat being generated.

When a heat sink is present, an extra heat path is available for the heat leaving the die. Consequentially, the peak temperature in models with a heat sink occurs slightly

earlier than in the model with no heat sink, after around 170μs and 220μs, respectively.

The time taken for the heat to reach the beginning of the heat sink is around 190μs, as indicated by the cumulative time constant for the solder layer above the die in Table 5.2. Therefore, at the pulse duration where the peak temperature occurs, the heat has not fully penetrated the solder layer. Any reduction in the die temperature up to this point is due to the solder layer joining the heat sink.

The corresponding MoP curves for the transient responses of the models to the lightning strike are plotted in Figure 6.17. The temperature curves for each model are also plotted alongside the MoP curves. This allows the MoP to be evaluated at the point of the peak temperature, which signifies the reduction in the maximum temperature experienced during the pulse.
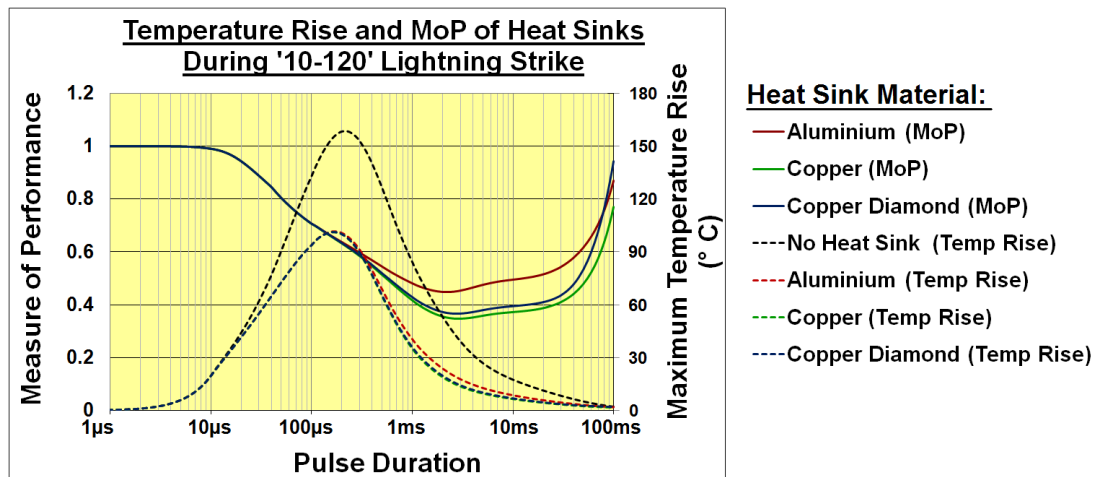
Figure 6.17: The MoP of the three heat sinks during the lightning pulse. The maximum die temperature rise is also plotted to allow the MoP at the maximum temperature rise to be evaluated.

At a pulse duration of 220μs, when the temperature peaks for the model with no heat sink, all three heat sinks provide a very similar MoP of around 0.62. As the heat has only penetrated a small amount of the heat sink, there is little variance in the MoP values (which ranges from 0.621 for copper and 0.627 for aluminium).

The MoP curves decrease from this point during the period when the die is cooling. However, the MoP is less important at these points, as the maximum temperature

has been reached (as seen in Figure 6.17) which is when device failure would occur. Should the device survive the maximum temperature, then it is assumed that the subsequent lower temperatures will not cause any damage.

During the lightning pulse, the presence of a heat sink causes the maximum die temperature experienced to be reduced by almost 38%. This is barely affected by the heat sink material or size, but is heavily dependent on the quality of the solder material and the quality of the solder joint. Using the knowledge learnt in Section 5.5.4, where the affect of solder joint thickness on the MoP was investigated, the MoP at the peak temperature could be reduced by reducing the solder layer thickness. This would allow the heat to penetrate the heat sink sooner. Exposing the heat sink to the heat sooner allows its superior thermal properties to reduce the die temperature from earlier in the pulse.

6.3.2.1 Lightning Strike and Constant Heat Generation Comparison

A comparison of the MoP curves from a lightning strike heat pulse and a constant magnitude heat pulse is made in Figure 6.18, which also shows the lightning strike waveform for reference. The comparison is done for a 5mm copper heat sink. The time at which the MoP begins to increase is very similar in both cases, around 6μs to 8μs into the pulse. The gradients at this point differ; the MoP for the constant magnitude decreases quicker than that of the lightning strike. By 60μs into the pulse, the lightning strike MoP curve decreases below that of the constant heat generation. The minimum MoP value for the lightning strike is lower for the constant heat generation (0.35 and 41, respectively), and also occurs earlier on (3ms and 30ms, respectively). The time of the minimum MoP in the lightning strike case is long after the pulse has reduced to zero.

As the heat generation during the lightning strike varies throughout the pulse, the analysis for the constant heat flux surface cannot be applied. The nature of the thermal response of the die is much more complicated. For the duration of the lightning strike,

around 200µs, the two MoP curves behave similarly. They begin decreasing at the same time, although at different rates. The gradient of the lightning strike curve is more variable, due to the more complex nature of the heat generation.
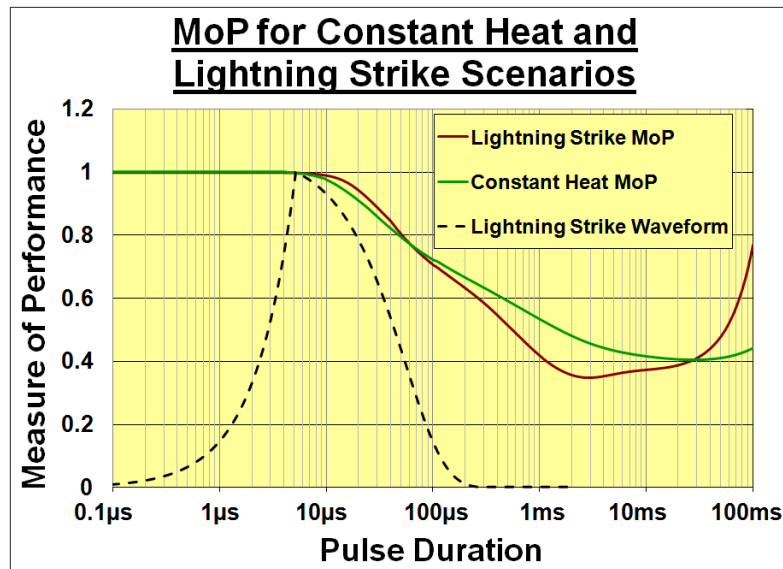
Figure 6.18: A comparison of the MoP curves for a 5mm copper heat sink when a lightning strike and constant heat pulse is simulated. The lightning strike waveform is shown for reference.

6.4 Chapter Conclusion

Modelling in 2D has allowed the heat spreading effects to be analysed. It has been shown that the 2D MoP curve does not differ from the 1D curve for the same situation for pulses less than around 40ms. The 2D modelling showed that the heat sink is less effective for pulse durations between 40ms and 1s when heat spreading effects are taken into account. 2D modelling also shows that the heat sink reduces the steady state temperature of the die. This is due to a flatter temperature gradient being created across the die surface, reducing the maximum temperature in the centre of the die.

A sensitivity analysis into the effects of the heat sink thickness show similar results to the 1D modelling results. The analysis shows that only a small heat sink is required in order to achieve temperature reductions at steady state. It was also seen that the boundary condition applied to the baseplate only effects heat sink effectiveness for pulses longer than 200ms.

Other investigations were into the effects of the heat sink coverage of the die and the use of composite heat sinks. Heat sinks that do not cover the entire die surface are less effective, as hot regions occur around the edge for short pulse durations. This can reduce heat sink performance for pulses as long as 200ms.

HOPG and copper were combined in different ways to try and improve the heat sink performance. It was shown that the heat sink effectiveness can be improved slightly by using a composite heat sink, although the improvements were very small. The most successful composite heat sink consisted of a layer of HOPG close to reduce the die temperature at short pulse durations, with a layer of copper on top to absorb the heat for longer pulse durations. Due to the limited improvements and complications of manufacturing such heat sinks, no more consideration was made to them.

A '10-120' lightning waveform was simulated to investigate heat sink performance for different heat generation profiles. The maximum reduction in the die temperature

was 38% up until the peak of the heat pulse. Heat sink performance was not as material dependent as the constant heat generation models, as the heat was unable to penetrate very far into the heat sink during the pulse.

Chapter 7

Experimental Validation of the Heat Sinks

Computational modelling work conducted indicated that using a copper heat sink on top of a power electronic device can reduce the maximum die temperature rise by 60% during the diagnostic period of current surges. In this chapter the heat sink concept is validated experimentally.

The chapter begins by discussing the different aspects of the experimental facility. The purpose and operation of the apparatus equipment is described, followed by a description of the rig setup, including an electrical diagram. This continues onto the experimental procedure, which states the way that the experiment was conducted.

The process of manufacturing the test pieces is then described, including techniques, equipment used and problems encountered. Before the results section, a discussion of the different possible device temperature measurements techniques is presented.

The results section consists of the MoP curves obtained from test pieces with different heat sink thicknesses. These are then compared to MoP curves obtained from numerical modelling, where the experimental setup is replicated. The chapter concludes with a discussion of the experimental results, with comments on the success of the validation process.

7.1 The Experimental Facility

7.1.1 Experimental Rig

7.1.1.1 Apparatus

Test Piece   Diodes were used as the test device. They were mounted onto a substrate and baseplate, following the numerical model construction closely. Different test pieces were constructed, some following the traditional assembly without a heat sink and some with copper heat sinks of varying sizes. The manufacture of the test pieces is described in Section 7.2.

Test Piece Holder   An aluminium holder provided the housing for the test piece and connections for the coolant pipes. The test piece was clamped into the holder as depicted in Figure 7.1, creating a water tight seal around the baseplate which allows it to be water cooled. Coolant pipes (1 inlet and 2 outlet) were attached perpendicularly to the bottom of the baseplate holder. Water was delivered through the inlet pipe and impinges the centre of the baseplate. Two outlet pipes at the edge of the holder's base provide an exit path for the coolant. This flow is demonstrated by arrows in Figure 7.2.
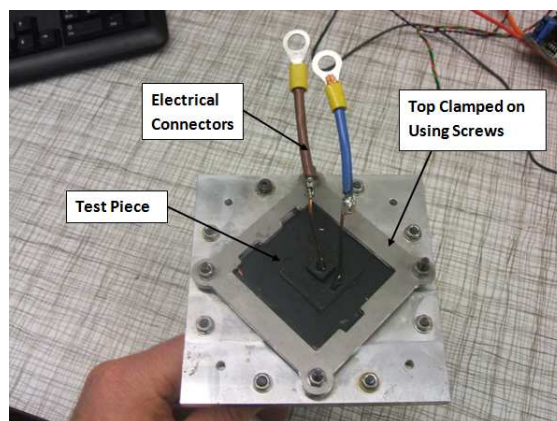
Figure 7.1: A test piece, painted black for thermal imaging, clamped into the test piece holder.
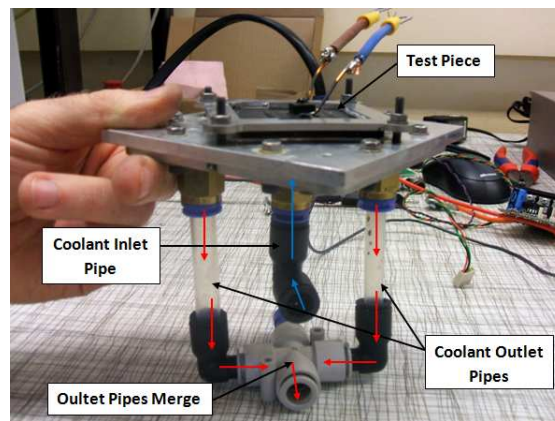
Figure 7.2: The coolant inlet and outlet pipe connections to the test piece holder.

High Current Power Supply   The electrical power pulse which causes the heating within the device was supplied by a high current power supply. The supply used was a Xantrex XDC 10-1200, capable of providing a current of up to 1200A at up to 10V and programmable to provide a constant power output. The power pulse delivered to the device was set at a constant rate of 85W. Limits to the voltage and current were set as 3V and 120A, respectively.

High Current Power Supply Switch   The electrical power pulse was controlled by a switch, which consisted of 3 MOSFETs. A Labview program allowed the pulse logic to be specified, such as the pulse duration (which could be as short as 1ms) and the time between pulses. The gate voltage of the MOSFETs was controlled by the PC to turn the MOSFETs on and off.

Small Current Power Supply   A small constant 50mA current was also supplied to the device when the high current supply was switched off which allowed the voltage drop across the device to be measured . The current had to be sufficiently small so heat generation within the device was negligible. It was also required that this current was very precise and constant. A circuit board was made to provide this current, as shown in Figure 7.3. The circuit required an 8V power supply.
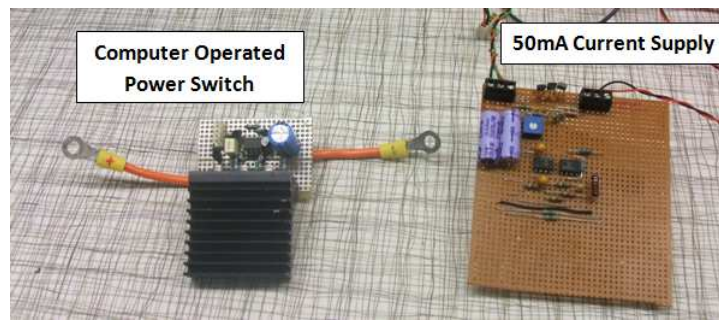
Figure 7.3: The high current power supply switch (left) and the 50mA current supply (right).

Pump   The water coolant was pumped around the piping circuit by a Tool-Temp TT-139 pump. The pump was able to control the flow rate and temperature of the water, thus providing a constant supply of water to the baseplate at a steady pressure and temperature. The achievable water temperature ranges from the temperature of the mains water to around 90ºC. Water was heated using an in-built heating coil and cooled by adding water from the mains. During the experiments, the coolant water was not heated and kept at the coolest possible, typically around 15 to 20ºC.

Computer and Interface Panel   The PC used Labview to control the high current power supply switch and to log voltage and thermocouple measurements. TC 2095 and BNC 2095 terminal blocks were used for the thermocouple and electrical connections. The high frequency forward voltage measurement connections were made to a NI PXI 6280 DAQ card, through a CB-68LP unshielded 68-pin connector block.

7.1.1.2   Rig Setup

The electrical connections were constructed as shown in Figure 7.4. The heating pulse current (I_{heat}) and 50mA measuring current ($I_{measure}$) were connected to the diode in parallel. The switch was used to switch the heating current to the device, and was controlled by the computer. The measuring current remained connected to the diode throughout the experiment. The voltage drop across the device ($V_{measure}$) is measured by the computer when the heating current is switched off.
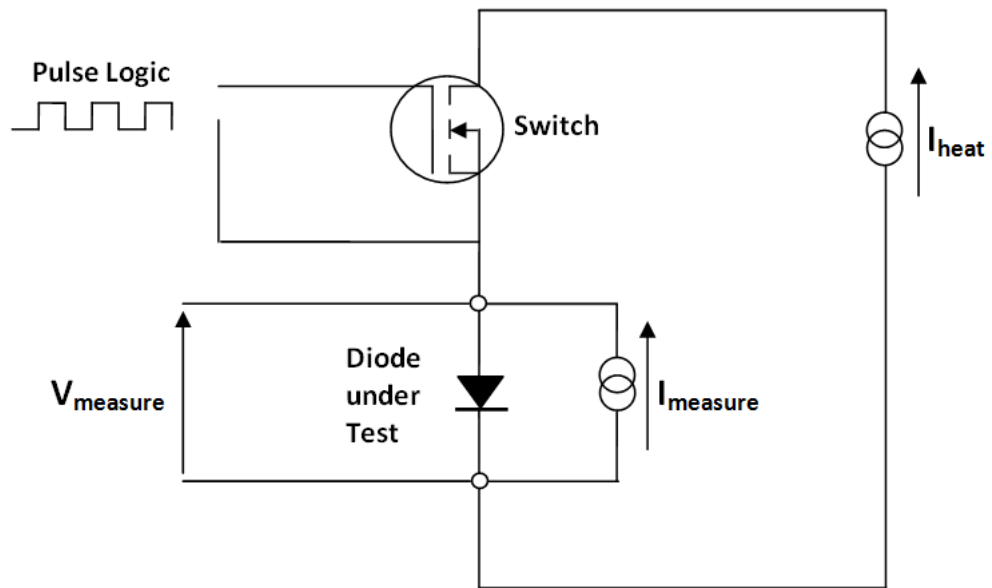
Figure 7.4: Electrical diagram of the experimental setup.

7.1.2 Device Calibration

The linear relationship between the diode temperature and voltage drop across it (V-T) had to be calibrated for each test piece. This relationship was different for each test piece and thus a calibration for each was required.

This was achieved by heating the test piece up to discrete temperatures and allowing it to reach steady state. This meant the device temperature was known more accurately. A heater pad was used as the heat source, varying its temperature using a variable voltage power supply. The test piece was placed on top of the heater pad, using a thermal grease at the interface to provide a better thermal contact. Four thermo-couples were attached to the device at different locations to monitor the temperature. This provided a way of recording when the temperature had reached steady state, and to ensure that the different parts of the test piece were at the same temperature. An insulating cover was placed on the test piece to reduce any variation in temperature.

Measurements were taken at six different temperatures, at which the temperature and voltage drop was measured for a $5s$ period. This allowed enough data points to fit an

equation for the relationship.

7.1.3   Experimental Procedure

7.1.3.1   Procedure Outline

Once the *V-T* calibration had been performed, the test piece was mounted into the holder. Thermocouples were attached to the test piece using a thermal glue, allowing temperatures to be monitored on a laptop using a PicoScope TC-08 USB data logger. Through the monitoring of these temperatures it was known when the test piece was at thermal equilibrium. The pump was switched on, and the temperature set to $0^{\circ}C$. This kept the coolant at the minimum temperature possible. The pump's bypass valve was fully closed, sending the maximum flow rate to the test piece.

High current pulses of different lengths were applied to the device being tested, causing the device to heat up. At the end of the pulse, the voltage drop across the device was measured during the device cool-down. Converting the voltage drop to the corresponding device temperature provided the device cooling curve, which could be extrapolated back to find the device temperature at the end of the pulse. A wide range of pulse durations had to be tested to obtain the thermal response of the device, which proved to be time consuming.

An additional draw back to this method was the time constant involved in the power supply switching on and off. The heating current was not switched instantaneously leading to uncertainty in the pulse length. A current spike was also witnessed in the first 100μs, heating the device more than it should do early in the pulse.

An alternative way to obtain the heating curve for a test piece was through inversion of the cooling curve to provide a heating curve. Applying a 60s pulse to the device allowed it to heat up to steady state. When the pulse ended, the voltage drop was measured for a 100s as it cooled back down to ambient temperature. The resultant cooling curve showed the temperature change during the entire transient period, from

two steady state levels. By normalising the cooling curve, and then inverting it, the equivalent heating curve is produced which provides a more detailed curve than the previous method. The process for inverting the curves and the method of obtaining the heating curve is described in more detail in Appendix D.

7.1.3.2 Labview Data Acquisition Settings

The Labview facility allowed different timing parameters for the pulse and data sampling to be specified.

Pulse Duration (s) The duration the high current power supply switch is open for each pulse. This can take on a value between 1ms and 60s.

Number of Acquisitions The number of times the pulse and data sampling is to be repeated. Repeating the pulses was important to reduce noise in the voltage measurements. The readings from each acquisition were combined to smooth out the noise. Five acquisitions was found to be enough to produce a curve that was unaffected by noise and that could be replicated repeatedly.

Time Between Pulses (s) The time between the pulses when more than one acquisition is performed. The voltage drop across the device is not necessarily measured for the entirety of this period; the sampling period is determined by the following two parameters: Sampling Frequency and Samples per Acquisition. Set to 100s to enable the ambient steady state to be achieved before pulsing again.

Sampling Frequency (Hz) The frequency at which the voltage drop across the device is measured after the pulse is switched off. This can be set up to 100kHz, however readings taken less than 0.1ms after the pulse were heavily affected by noise. A frequency of 10kHz was found to be sufficient.

Samples per Acquisition The total number of voltage readings taken per acquisition. The sampling time per acquisition is thus:

$$Sampling\,time\,per\,acquisition = \frac{Samples\,per\,Acquisition}{Sampling\,Frequency}$$

There was a limitation to the number of samples per run (i.e. total samples for each acquisition) due to the memory of the computer. A maximum of around 5.5million samples could be taken during the run:

$$5.5\,million > Samples\,per\,Acquisition \times Number\,of\,Acquisitions$$

Sampling for the entire 100s at the rate of 10kHz was possible for the 5 acquisitions needed. This corresponded to a total of 1million samples per acquisition which allowed a high resolution transient curve to be obtained. These data points are downsampled after the data acquisition to a number of data points that can be handled easily in a spreadsheet.

Once these parameters had been specified, and the voltage channels selected (as shown in Figure 7.5), the run button allowed the pulsing to commence.

7.1.3.3   Thermal Imaging

A thermal imaging camera was used to view the heating and cooling during the tests. Although the device temperature could not be measured this way for test pieces with a heat sink, it provided an insight into the effects of the heat sink on the test piece temperature as a whole.

A Cedip Titanium thermal imaging camera was used. It was mounted on a tripod to look down on to the top of the test piece. The software used for viewing the images, capturing the data and editing the resultant videos was Altair, specialist software
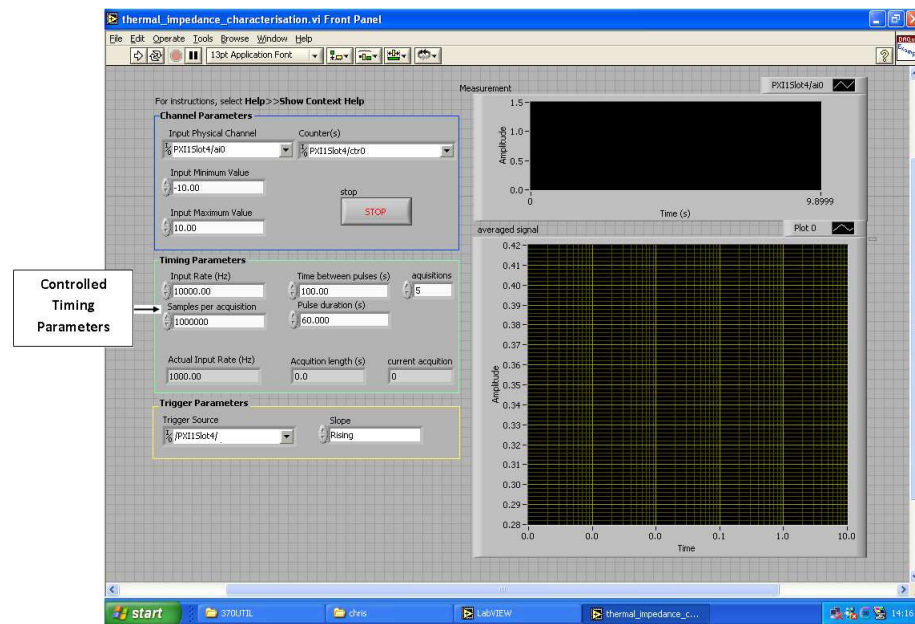
Figure 7.5: The Labview data acquisition computer screen.

provided with the camera. Data was captured at the maximum frequency possible, 383Hz. Images obtained are shown in the experimental results section, 7.4.

At any wavelength, the heat emitted from a surface is some fraction, ε, of that of a black body at the same temperature. Surfaces are assumed to be grey, such that the radiation emitted, $\dot{q}$, is calculated by:

$$\dot{q} = \varepsilon \dot{q}_b$$

where $\dot{q}_b$ is the radiated heat emitted by the equivalent black body. The emissivity coefficient, ε, can be controlled by changing the surface of the body. Polished metal surfaces have lower emissivity values than dulled surfaces, and black is the most effective surface colour. The test pieces were sprayed with a matt black paint in order to maximise the emissivity of the surfaces. Comparisons between thermocouple and thermal imaging measurements showed good agreement. As the thermal imaging was only used for comparative purposes between different test pieces, and not used to measure an exact die temperature, a full calibration was not required.

An example of a test piece sprayed black is depicted in Figure 7.1. Appendix B lists the emissivity coefficient for different materials that are relative to the experimental work. Slightly polished copper has an emissivity coefficient of around 0.15, this can be increased to around 0.91 with the use of a black matte shellac paint.

7.2   Manufacturing the Test Pieces

Diodes were used as the devices for the experimental work. Diodes have a simple architecture and require only two electrical connections: the anode and cathode. Most diodes have aluminium metallisation to allow the aluminium bond wires to be attached with ease. Soldering the heat sink to this surface was difficult, as solder does not bond well to aluminium. A potential solution to this problem is to use an active solder, which claimed to bond to a wider range of materials than normal tin silver (SnAg) solder. This trial proved unsuccessful, and is described in more detail in Section 7.2.1. In order to get around this issue a new set of diodes were obtained from ABB, which had silver presentation on both sides, instead of the traditional aluminium. This allowed the heat sink to be soldered directly to this surface with regular SnAg solder.

Copper baseplates were manufactured to an appropriate size to fit the test piece holder and substrates were obtained from Dynex. The dimensions of the test piece components are given in Table 7.1.

Copper heat sinks were used in the experimental work, due to their performance in the computational work and also their ease of manufacture compared to copper diamond. It was also not possible to obtain any copper diamond to use.

	Thickness (mm)	Width (mm)	Length (mm)
Diode	0.30	12.50	12.50
Diode (Central Metallisation)	0.30	10.50	10.50
Substrate: Copper Layer	0.24	27.30	28.80
Substrate: AlN Layer	0.40	29.00	30.50
Substrate: Copper Layer	0.24	27.30	28.80
Baseplate	3.00±0.05	60.00±0.5	60.00±0.5

Table 7.1: Table of component dimensions for the test pieces. (Unless stated, measured tolerances are ±0.02mm.)

7.2.1 Solders and Soldering Techniques

Soldering the layers presented the main problem during the manufacturing process. To obtain good quality solder joints, with a consistently good bond all over, both proved difficult using the University facilities. Due to the aluminium metallisation on top of the first diodes, a different sort of solder was required as SnAg solder is difficult to bond to aluminium.

7.2.1.1 S-Bond Active Solder

An active solder from a company called S-Bond was found that appeared to be suitable for bonding the copper heat sink to the diode with an aluminium. The description of the solder from S-Bond is that it is a:

> "Modification of conventional lead-free solders with rare earth elements and titanium to allow bonding to metals, ceramics, glasses, and carbon products in air without the use of flux, surface metallisation, or other treatment "

The titanium and rare earth elements are claimed to create an oxygen transport mechanism in the bulk material and allows a reaction at the substrate surface. The oxide layer formed can be broken by vacuum heat treating or mechanical force.[67]

An S-Bond 220 kit in foil form was bought to make the solder joint between diode surface and copper heat sink. Sections of the test piece made with it were sectioned and examined under a microscope, which revealed large voids throughout the joint, especially at the edge. These are shown in Figure 7.6.

There were other difficulties experienced when the solder joint was made due to the need for mechanical force required to break the oxide layer. The process was found to be difficult to produce a consistent solder thickness repeatedly.

In light of these problems, an alternative solution was sought after to produce a reliable joint between the diode and heat sink. This was achieved by using a different diode device with silver presentation on both sides, which allowed direct soldering with a regular tin silver solder.
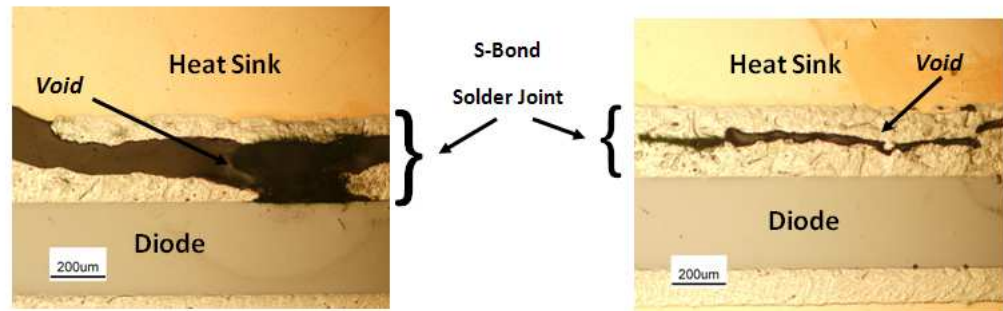
Figure 7.6: The voids present in the S-Bond solder joint between the die and heat sink.

7.2.1.2 Tin Silver Solder Foil

Tin silver (Sn3.5-Ag96.5) solder was used in foil form for the joint between the base-plate and substrate and the substrate and diode. To ensure the best chance of getting good quality solder joints, component preparation and cleanliness was important.

Successful joints depend upon the surfaces having a smooth surface finish. The substrate and diode surfaces had adequate finishes, as they were manufactured in electronic labs with the view to be soldered components. However, the baseplate was manufactured from plate copper and needed polishing to a finer level.

All soldering was done in the University soldering clean room. Components were placed in deionised water in ultrasonic baths for around 4 minutes prior to soldering. Solder foil was cut to have about a 5% larger surface area than the smallest component be joined. The cleaned components were rinsed with acetone to further clean and also to dry them without any water marks. Tweezers were used to handle all parts as any dust or dirt could compromise the entire solder joint.

Test pieces were stacked up with the foil solder in between. The entire stack was then carefully placed into the Electron Mec SRO 702 Solder Reflow Oven. A suitable oven regime was programmed in, with varying temperatures, time at each temperature, pressure and gas mixture.

Thermal results and later sectioning showed that the solder joints made using this method were adequate around 25% of the time. The other times, at least one of the solder joints had a void in it which significantly affected the thermal performance. These voids tended to be isolated and not all along the joints. Examples of the voids seen in the solder foil joints can be seen in Figure 7.7.

7.2.1.3 Tin Silver Solder Paste

Solder paste provided an alternative to the solder foil method outlined above. A tin-silver (SnAg 96.5) solder paste was used which was supplied in a syringe tube.

A benefit of using solder paste is that the surface finish of the soldered surfaces needn't have to be as smooth as they need to be when using solder foil. The flowing nature of the solder paste means it is able to fill in small scratches, unlike the solder foil. This prevents air voids being created at the locations of the scratches.

Masks for the solder foil were made from 0.1mm thick aluminium foil, manufactured to a high tolerance. Each mask was the same size as the small component surface area for each joint. Components were placed in deionised water in ultrasonic baths for around 4 minutes prior to soldering, and washed with acetone for additional cleaning and drying.

The masks were placed on the largest component for each joint, and solder paste was dispensed into the centre of the mask. Using the edge of a spare substrate (which had also been cleaned) the paste was spread within the mask, ensuring all parts were covered. The foil mask allowed the height of the paste to be as close to 0.1mm and as even as possible. Once the solder paste had been applied to all components, they

were stacked up, starting from the baseplate and working up. This was a delicate process as components needed to be placed squarely on top of the solder paste. A diagram of the assembly process can be found in Appendix C.

A different oven was required for components made up with solder paste. Due to the flowing nature of the solder, no vacuum could be used as the paste would be sucked out from the joints. A CIF FT 02 Batch Reflow Oven was heated up to 220°C for 5 minutes before increasing the temperature to 260°C for a further 30 minutes.

The solder joints were of better quality when using the solder paste as opposed to the solder foil, however the solder thickness was harder to control.
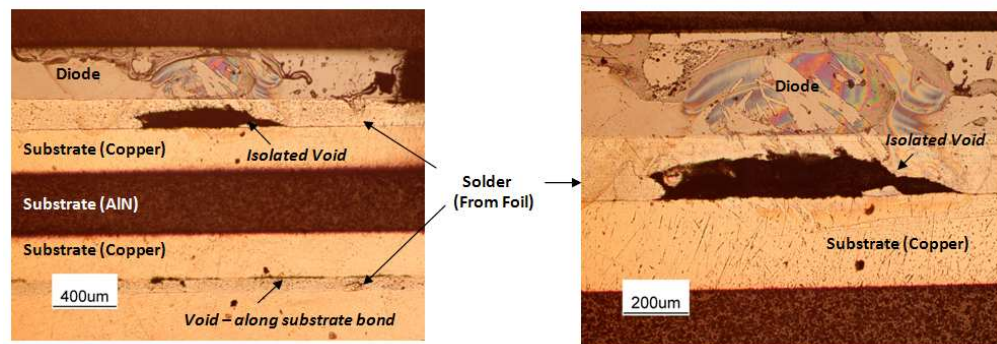
Figure 7.7: Voids seen in the solder joints made using SnAg 96.5 in foil form. The left photo shows a large isolated void, plus small surface voids. The right picture shows the isolated void in greater detail.

7.2.2 Electrical Connections

The substrate architecture is important in terms of how the electrical current is delivered to and from the diodes. When no heat is used, the traditional electrical arrangement can be used. This involves the negative bus bar, which delivers the current, to be soldered to the outer copper strip of the substrate which is electrically isolated from the central copper section. Aluminium wires carry the current from this strip to the top of the diode, bypassing the central part of the substrate in order to avoid short circuiting. The current passes through the diode, and is conducted to the substrate beneath it, to which the positive bus bar is soldered, completing the circuit.

This is shown in Figure 7.8, alongside the electrical arrangement used for the test pieces which used a heat sink.

When the heat sink was soldered to the top surface to the die, the electrical contacts could not be made in the same way and an alternative electrical connection arrangement had to be made. The heat sink and solder were electrically conductive and so the negative connection was made to the heat sink. The bus bar was soldered directly to the top of the heat sink, which simplified the electrical arrangement compared to the aluminium wire bonding.

As the heat sink covered the entire active surface of the diode, the current was supplied evenly across it through the heat sink. Directly soldering to a bare diode wouldn't provide this same coverage. Multiple aluminium wire bonds were used, each connecting to the diode in two places in order to try and distribute the current across the diode as evenly as possible.
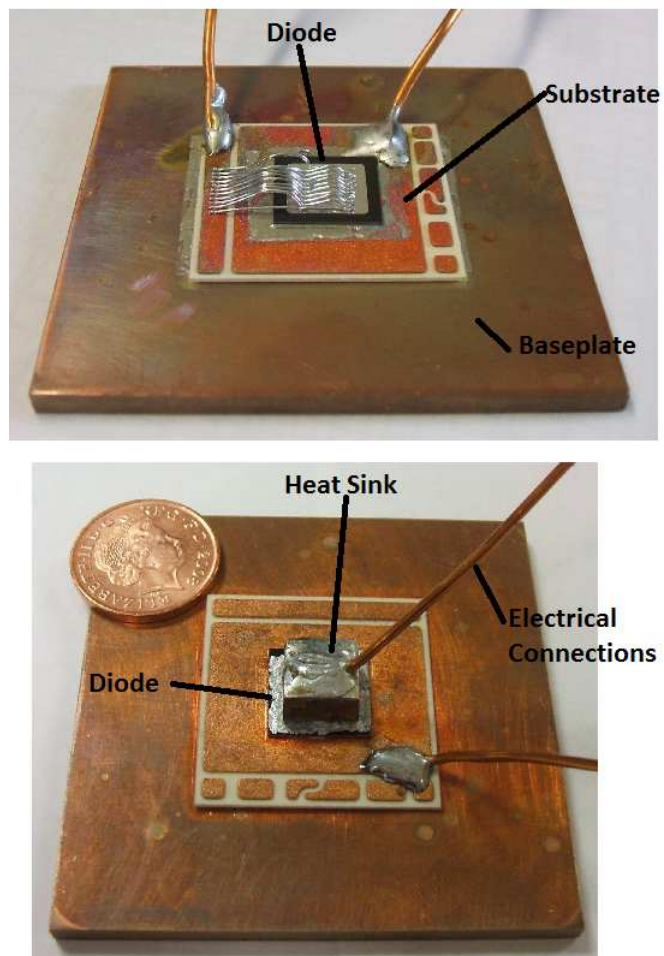
Figure 7.8: The electrical connections made to the test pieces are shown. The test piece with no heat sink (top) requires aluminium wire bonds to supply the current to the top of the diode. When a heat sink is soldered to the diode (bottom) the electrical connection can be soldered directly to the heat sink as it is electrically conductive.

7.3   Device Temperature Measurement Techniques

A fast responding, high resolution temperature measurement method was required to capture the thermal response of the devices to electrical pulses. A range of techniques were considered.

Thermocouples

Using thermocouples was considered impractical as they would only be able to measure the surface temperature, and as the maximum die temperature invariably occurs within the die they would not measure at this location. An additional drawback to the use of thermocouples is the delay in the temperature measurement incurred due to the time required for the heat to travel to the thermocouple from the device through the attaching interface. Fast response thermocouples typically have a response time of around 2ms which, in relation to the pulse durations of interest, is too long to capture the die's thermal response to the pulse.

Thermal Imaging

A high frequency thermal imaging camera was available for use which is able to capture videos at a maximum frequency of 383Hz, allowing the surface temperatures of the test pieces to be measured with a resolution of 2.6ms. However, there are two limitations to this method. The frequency, although very high for thermal imaging, wouldn't be able to capture the thermal transients until 2.6ms into the pulse. Like the thermocouples, a quicker response time is necessary to capture the full thermal response. Further to this, only the temperatures of exposed surfaces visible to the camera can be measured. Therefore, it would be impossible to measure the temperature at the centre of the die when a heat sink is mounted on top.

Voltage Measurement Across Device

The voltage drop across power electronic devices is dependent on the temperature of the device. The voltage-temperature relationship is a linear one and can be found for each device through calibration (measuring the voltage drop across the die for various controlled temperatures). Thus, when a small fixed current is applied to the die, the temperature can be determined by measuring the voltage drop across it. High frequency data logging apparatus allow the temperature of dies to be obtained without any intrusion or destruction.

When the die temperature is measured this way, it is not the maximum temperature which is measured. Typically the maximum temperature is in the centre of the device, and the coolest regions are around the edges. As a result, the local voltage drop in the device vary with the regions of different temperatures. When the overall voltage drop is measured, it provides a temperature which lies in between the average and maximum temperature across the die.

This method was used for measuring the die temperature. Thermocouples and thermal imaging were also used in conjunction with the voltage measurements to monitor conditions such as the coolant temperature and the calibration of the die's voltage-temperature characteristics.

7.4   Experimental Results and Agreement with Numerical Predictions

7.4.1   Experimental Results

Figure 7.9 shows the MoP curves for the four heat sinks obtained from the experimental temperature measurements. Due to noise in the voltage readings at the beginning of the measurements, meaningful data is only provided for time periods of around 1ms onwards. The noise arises from the finite period of time it takes for the high current supply to switch off. The voltage drop across the diodes was therefore larger than expected until the current settled to a constant 50mA.
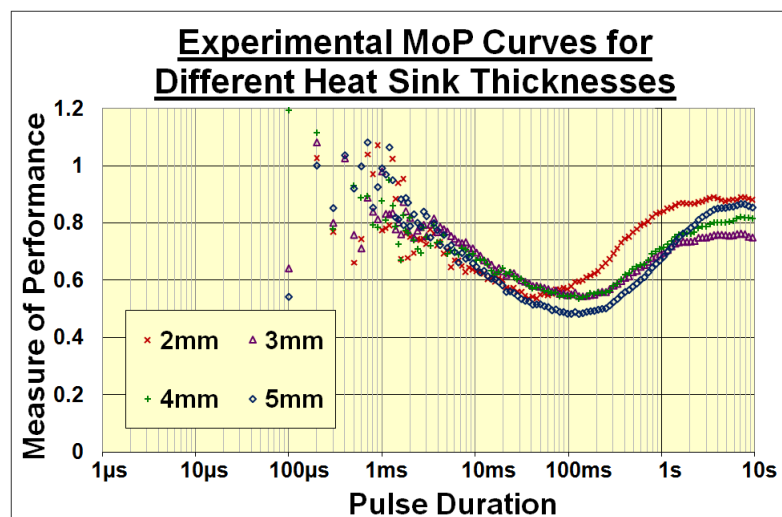
Figure 7.9: The MoP curves for each of the tested heat sink size.

All four curves provide similar MoP values during the time period 1ms to 40ms. This is consistent with the expected behaviour, as seen from the numerical modelling results. After around 40ms the 2mm heat sink reaches its minimum MoP value of 0.54, and then begins to increase towards steady state. The 3mm and 4mm heat sinks behave very similarly to each other, reaching a similar minimum MoP as the 2mm heat sink, although this occurs at a later time of around 100ms. The 5mm heat sink reduces the die temperature for the longest, as expected. A minimum MoP of 0.48 is reached 200ms into the pulse before the MoP increases to steady state.

There is inconsistency in the results for the longer pulse durations. Variation in the MoP curve gradient and the steady state MoP values reached indicate very different steady state thermal resistances between the heat generation and the coolant. This is largely due to the difficulty in getting a consistent solder thickness between the different test pieces. Voids may also be present in some of the solder layers which can unpredictably increase the thermal resistance. This meant the thermal resistance between the diode and baseplate was not the same for all test pieces.

7.4.1.1 Agreement with Original Numerical Results

Numerical models were constructed to replicate the test pieces used in the experiments. A 2D no heat sink and 5mm heat sink model were constructed for comparison with the experimental data. Dimensions for each component corresponded to the measured values listed in Table 7.1.

The models were run using the boundary conditions assumed in the initial modelling work: heat was generated in top 10% of die, and all walls were modelled as adiabatic, except for the bottom of the baseplate which was modelled as isothermal. The maximum die temperature was calculated for a 10s pulse, which allowed the MoP curve to be found and plotted against the experimental MoP curve for the 5mm heat sink.

Figure 7.10 shows that although the experimental and numerical modelling MoP curves have similar shapes, they do not match particularly well. Major discrepancies between the two curves exist, particularly at the early and late pulse durations. There is a big difference in the time at which the heat sink becomes beneficial. This was around 2μs according to the numerical model, compared to around 200μs in the experimental curve. It is clear that in the experiments there is a delay in the heat sink becoming active. The MoP curve gradients are also very different as they approach steady state. The time taken to reach steady state also varies from around 1s in the numerical model to 6s in the experiments.
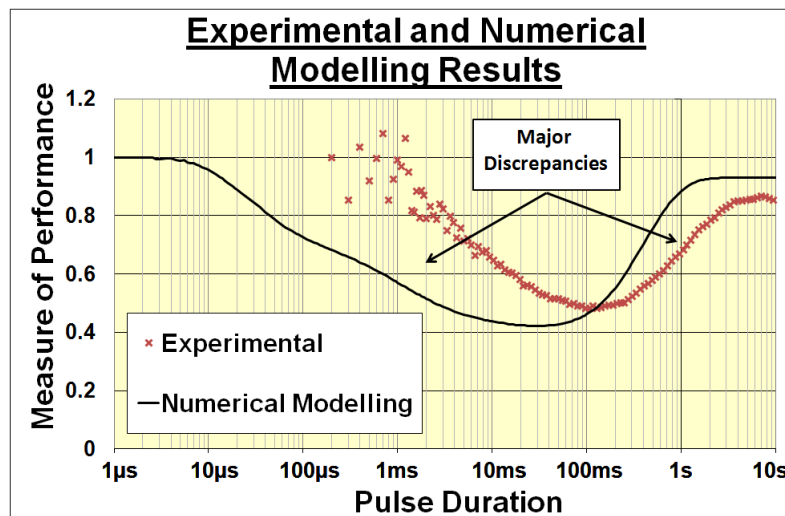
Figure 7.10: Comparison between the experimental and numerical modelling MoP curves for the 5mm copper heat sink. Numerical models based on boundary conditions and heat generation used in Chapter 5.

7.4.1.2 Improvements to the Numerical Model

Alterations were made to the numerical model to replicate experimental conditions more accurately. The modifications focused on improving the accuracy at the beginning at end of the pulse in particular.

The heat generation was changed to the middle 10% of the die. The experimental data was taken during a steady on-state, and therefore this condition is more accurate to the heat generation behaviour in the diodes. During the on-state (as opposed to during high frequency switching) the thermal resistance in the drift region (central region) of the diode is dominant, leading to the majority of the heat being dissipated there. This would increase the time taken for heat to conduct through the top part of the die and into the heat sink, affecting the early region of the MoP curve.

A heat transfer coefficient of 5,000W/m^{2}K was modelled at the baseplate boundary, replacing the isothermal condition previously modelled. This value is realistic for impinged forced water cooling, as used in the experimental setup. Changing the boundary condition in this way reduces the heat transfer to the coolant, therefore increases the time taken to achieve steady state.

Figure 7.11 shows the comparison between the experimental MoP curve and the MoP curve from the improved numerical model. It is evident that the modifications provide a much better match between the curves. The pulse duration at which heat sink first becomes beneficial agrees reasonably well, around 100µs. The minimum value of each curve is also very similar, ~0.48, and it also occurs at a similar pulse duration. The similarity in the gradients of the MoP curves after the minimum MoP is encouraging, as both curves rise up to steady state in a similar fashion.
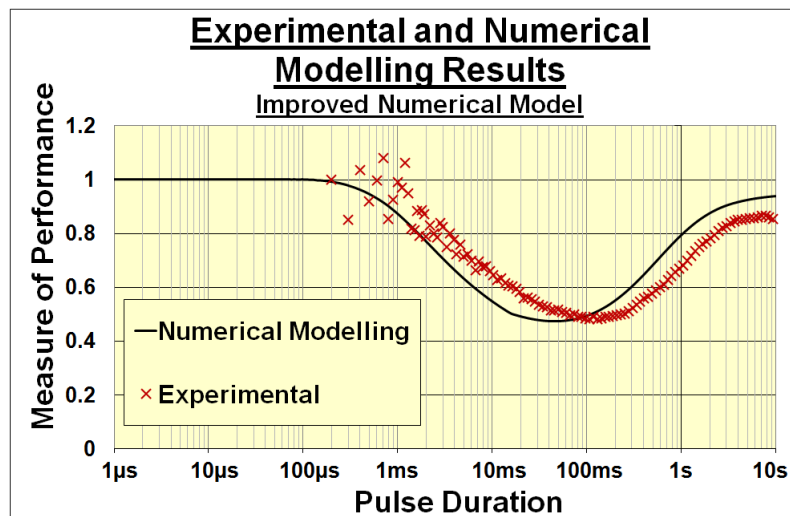
Figure 7.11: Comparison between the experimental and numerical modelling MoP curves once the modifications to the numerical model had been made.

Some discrepancies still exist between the models, demonstrated by a small lag in the curves. This is due to the experimental conditions being very difficult to replicate perfectly, as determination of solder thickness and solder voids is difficult. Small fluctuations in the experimental heat pulse and coolant temperature contribute further to the difficulties.

It is clear from the experimental results that the heat sinks decrease the transient temperature in the die during the transients, and by the magnitude expected. The reasonably close match of the numerical and experimental curves are encouraging, leading to confidence that the numerical model predictions have been validated.

7.4.1.3   Thermal Imaging Snapshots

Thermal imaging videos were taken during some of the pulse testing. Manipulation of the videos enabled the temperature rise in a test piece with and without a heat sink to be viewed concurrently. Snapshots of a heat pulse taken at 40ms intervals are shown for each device in Figure 7.12. Surface temperatures of the test pieces show the suppressing effects of the heat sink in action. During the early region of the pulse, up to around 320ms, the temperature rise in the device with no heat sink is a lot sharper than that seen in the device with a copper heat sink. It is not until approximately 240ms into the pulse that the heat sink becomes noticeably hot. Unlike the test piece with no heat sink, which stops getting noticeably hotter from around 200ms onwards, the temperature of the device with a heat sink is still increasing right up to the end of the 680ms pulse.
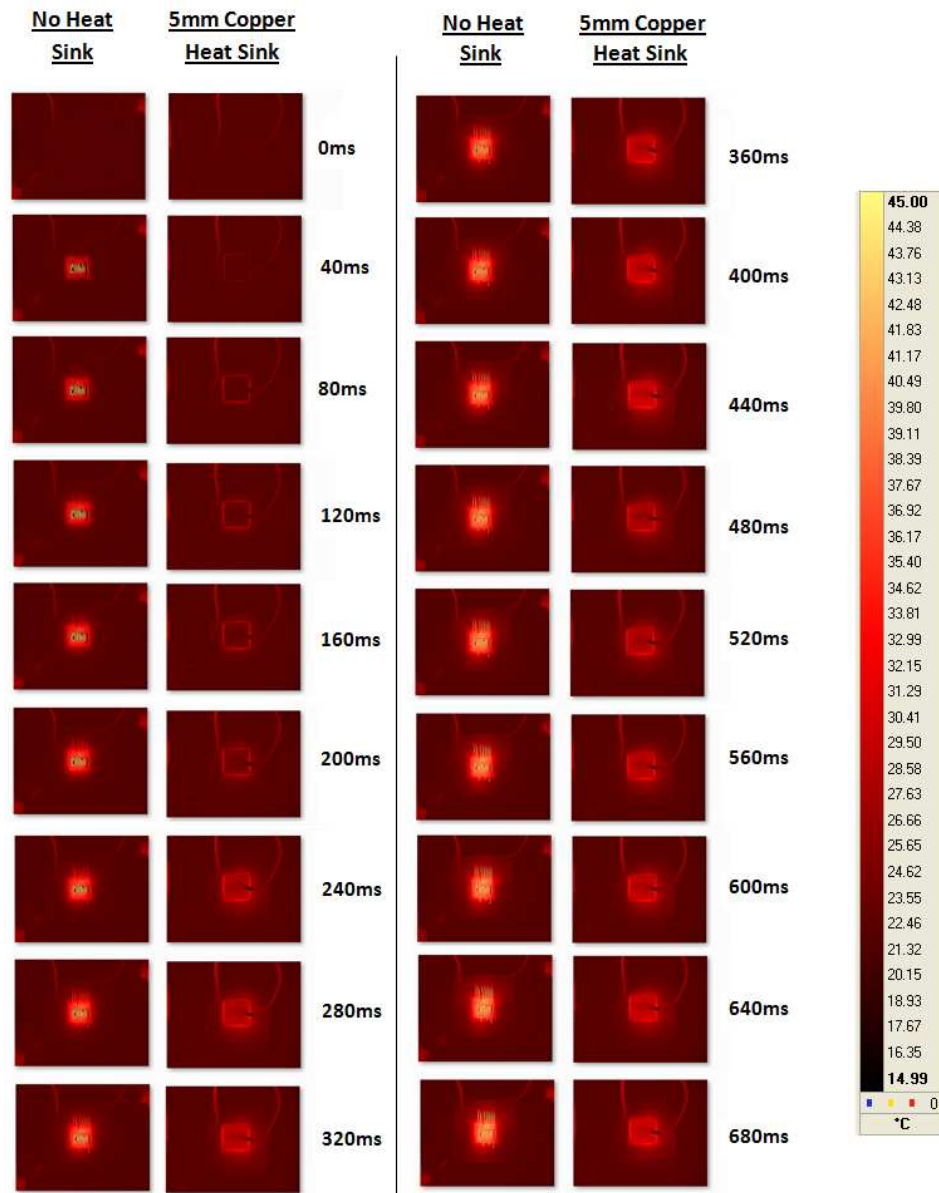
Figure 7.12: A comparison of the thermal imaging shots for the first 680ms of the heating pulse for a device with a heat sink and without a heat sink. The thermal transient is noticeably suppressed when a heat sink is added to the device.

7.5 Chapter Conclusion

The chapter has demonstrated the successful manufacture of test pieces with heat sinks attached to diodes. It is shown that a heat sink can be soldered directly to diode surface, without compromising the function of the diode.

An active solder has been tested to allow direct soldering to a diode with an aluminium cathode, but this proved unsuccessful. Diodes with silver presentation on the top were used instead of the usual aluminium surfaces. This allowed direct soldering using a normal tin-silver solder. Tin-silver solder in paste form provided better quality joints with fewer voids, due to the flowing nature of the paste. However, the joint solder thickness was harder to control than when using solder foil.

A benefit of attaching a heat sink to the die surface was that the electrical connections were simplified. An electrical connection can be soldered to the top of the heat sink, which removed the need for aluminium wire bonds which require specialist equipment.

The experimental testing of the test pieces also proved to be successful. The transient die temperatures were measured at a high frequency. This provided heating curves for the test pieces for pulse durations from 100μs. The temperatures were found by measuring voltage drop across the diode during the cooling period after being heated. This cooling curve was then inverted to provide the equivalent heating curve.

Experimental results showed similar trends to the numerical modelling results. Heat sinks reduced the die temperatures for transient pulses between 1ms and 1s. Thinner heat sinks were generally not as effective for as long as the thicker ones, which agreed with the modelling predictions.

Numerical models were created to replicate experimental conditions, such as stack dimensions, boundary conditions and heat generation. A close match between the experimental and numerical modelling MoP curves was achieved. This gives confidence in the numerical modelling method used.

Chapter 8

Experimental Work on Power Switches

The thesis focuses on the use of a heat sink for suppressing the heat generated in current surges. The first phase of experimental work has successfully demonstrated the heat sinks can be practically applied to diodes. Diodes provided a simple structure on which to attach the heat sink, minimising the assembly problems which needed to be overcome. Having succeeded in proving the ability of the heat sink to reduce the temperature of a diode during heat pulses, the heat sink concept should be validated using complicated power switches, such as MOSFETs and IGBTs. These are the structures which the heat sinks are mostly likely to be used on in practice for power switching.

A MOSFET device was chosen for the testing, and all discussion herein will refer to MOSFETs alone, although most of the principles apply to IGBTs as well.

8.1   Incorporating a Heat Sink into a MOSFET

The operation and structure of different devices is discussed in Chapter 2, which explains that MOSFETs have a more complicated surface structure than diodes. The

additional electrical connection (the gate) required means that three electrical connections must be considered (the gate, source and drain). The heat sink must allow each connection to be electrically active, whilst keeping the connections electrically isolated from each other to avoid short circuiting.

8.1.1 Gate Position

The gate connection is located on the top of a device, occupying a small region of the surface. The position of the gate is typically either in the centre of the device surface, or along one of the edges. These variations are shown in Figure 8.1. The gate voltage is carried from the gate connection to the gate conductor of each of MOSFET channel (of which there are typically thousands in each device [2]) via a gate track embedded under the device surface.

As the heat sink should be made of copper, or other electrically conducting material, it must not connect to both the source connection and the gate connection as this would create an electrical short between the two.

Consideration must therefore be made for the shape and design of the heat sink. Any heat sink attached to the MOSFET must cover only the source connection region of the surface. Further to this requirement, it must also allow access for the gate connection wire to be attached. Of the two possible arrangements shown in Figure 8.1, the edge gate connection provides the better architecture for attaching a heat sink; as the gate wire could be attached as normal.
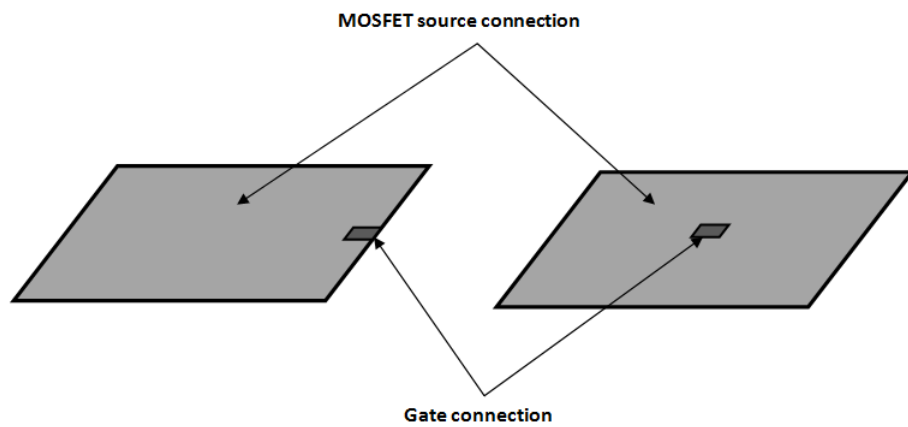
Figure 8.1: The two common gate connection locations on a MOSFET device. *Left*: An edge gate connection. *Right*: A centre gate connection.

8.2   Manufacturing the MOSFETs with Heat Sinks

The objective of the experimental work was to prove that a heat sink could reduce the device temperature of a MOSFET for short thermal transients. This involved identifying a suitable device with an edge gate connection. The chosen device would be prepared for experimental testing by attaching a heat sink and ensuring all necessary electrical connections were made.

8.2.1   The MOSFET device

The MOSFET device chosen was the Infineon CoolMOS IPC60R045CP. Figure 8.2 shows the dimensions and surface architecture of the device. The overall device dimensions are 6.58mm by 10.51mm. The gate connection is located at the bottom of the device in the centre of the short edge. The dimensions of the gate are approximately 0.6mm by 0.6mm.

Gate tracks are visible around the edge of the track and also as 'fingers' from the edge of the device towards the centre. These tracks distribute the gate voltage more evenly across the device surface. A thin film imide layer is present on top of the track, which provides electrical isolation between the gate track and the remainder of the device surface, which is an active source region.

8.2.2   Packaging the Devices

It was decided that the devices should be packaged by a commercial company to improve the quality and consistency of the solder joint between the device and substrate. Semelab packaged the devices inside an open topped TO258 package. The drain, source and gate connections were all made between the device and the appropriate pins on the package.

In total 26 devices were packaged by Infineon for the testing. 6 packages were hermetically sealed, which is customary for commercial packages in order to protect the
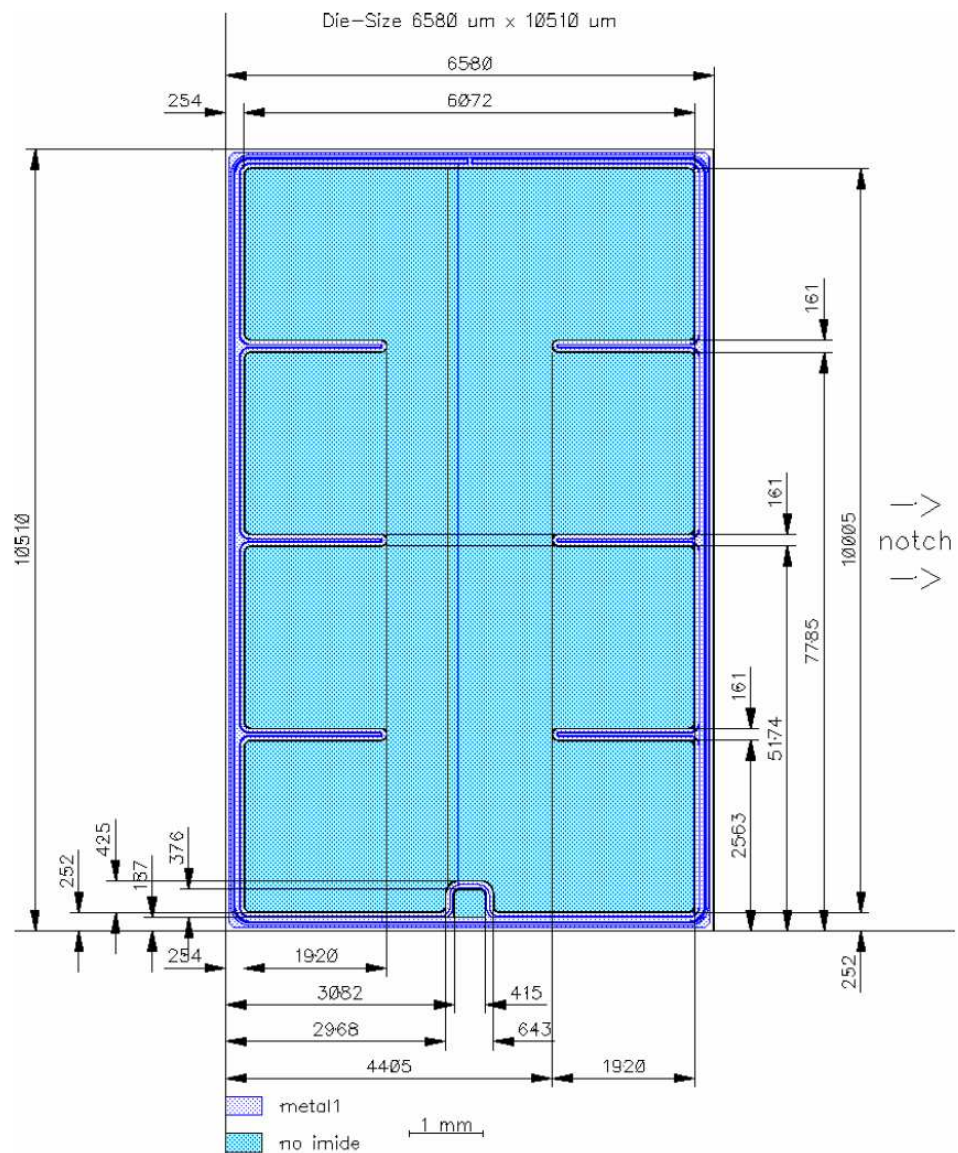
Figure 8.2: The Infineon CoolMOS IPC60R045CP device dimensions. The gate connection is visible as a small square at the bottom of the device.

device. These devices were used to provide temperature data for the no heat sink scenario. The remaining 20 devices were provided without the sealing process (unlidded), which allowed the heat sinks to be attached to the surface before testing. A photo of an unlidded packaged device as received can be seen in Figure 8.3.
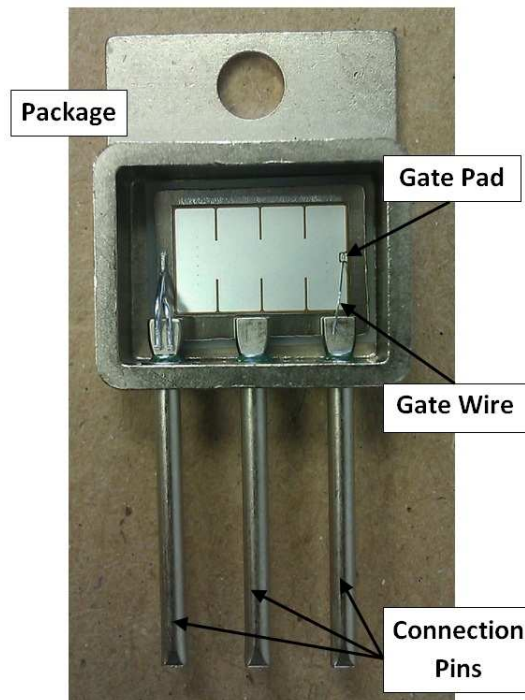
Figure 8.3: An unlidded MOSFET in a TO268 package, as received from Semelab.

8.2.3   Heat Sink Design

The shape of the heat sink was designed to cover as much of the die as possible without shorting the source and gate connections. The gate wire which connects the gate pin to the gate connection pad can be seen in Figure 8.3. Due to the relative location of the gate pin to the gate connection pad, the wire cuts across part of the die. Gate wires are thin and delicate, as they only have to carry small voltages. For this reason, moving the gate wires was considered undesirable as breaking it would make the device unusable.

The heat sink design therefore needed to consider the path of the gate wire, ensuring the two do not touch. Figure 8.4 shows the CoolMOS drawing with the gate wire and heat sink shape indicated. The green dotted line shows the path of the gate wire, and the ideal heat sink footprint (shaded in red).
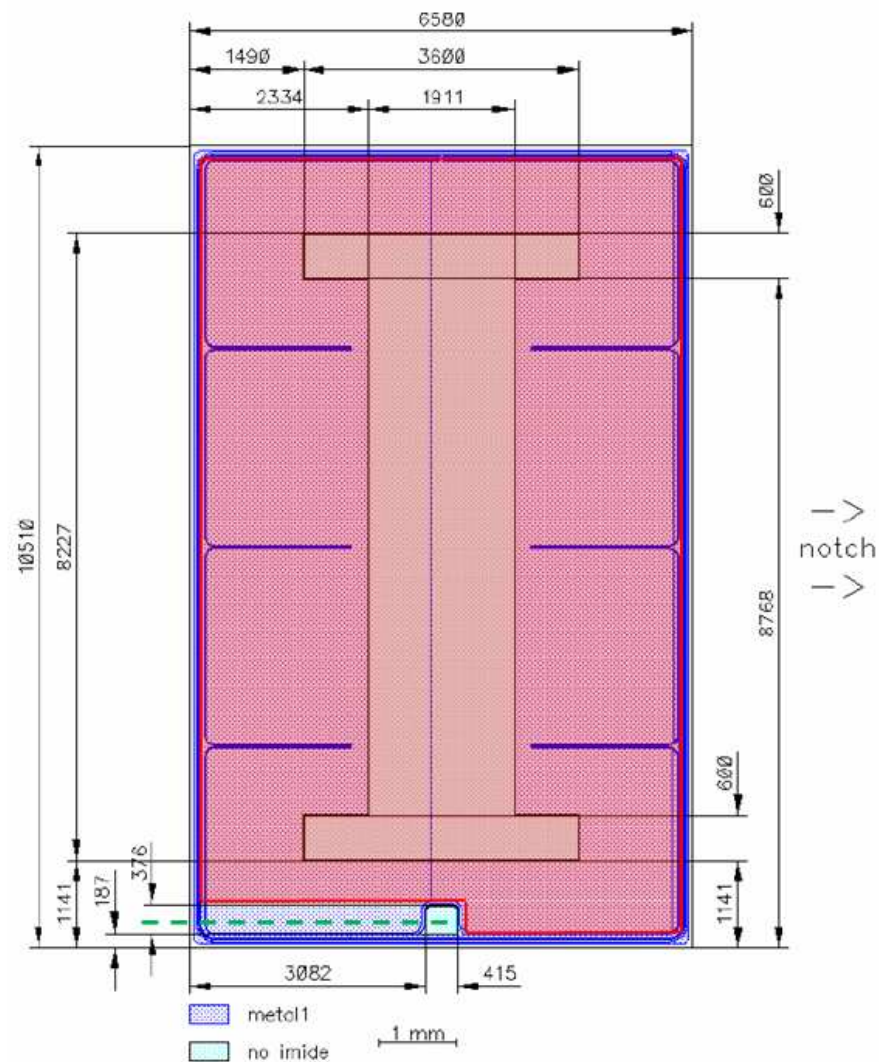
Figure 8.4: The Infineon CoolMOS IPC60R045CP device showing the path of the gate wire (dotted green line) and the heat sink footprint (red shaded area).

8.2.4 Silver Epoxy

It was not possible to use solder to attach the heat sinks to the devices due to the imide layer on the device surface which protects the gate track. Solder would not bond to the imide region of the device, and the heating process would damage the imide layer, shorting the gate and source. An alternative was required which provided a good bond, as well as a good electrical and thermal conductivity between the device and heat sink.

A silver epoxy was identified as a good candidate for a solder replacement. Silver particles added to a high temperature epoxy makes it electrically conductive. The thermal properties of the epoxy are also enhanced by the presence of the silver. The selected silver epoxy was '40-3900' Silver Filled Epoxy Resin, supplied by Epoxies Etc. Unlike many silver epoxies, which have short shelf lives and must be stored in a cold environment (in the region of around -15ºC), this silver epoxy was able to be stored at room temperature.

The electrical and thermal properties of the epoxy are compared to those of solder in Table 8.1. Both the electrical and thermal performance of the epoxy are worse than that of solder. The electrical resistance is around ten times greater than that of solder, and the thermal conductivity around 5 times worse.

	Thermal Conductivity (W/mK)	Electrical Resistivity ($\mu\Omega$-cm)
Silver Epoxy	14.4	100
Solder (Sn96.5-Ag3.5)	78.0	10 - 15

Table 8.1: A comparison of thermal and electrical properties of a tin-silver solder and the silver epoxy used as a substitute.

8.2.5 Heat Sink and Device Preparation

Copper heat sinks were made to an appropriate shape which allows space for the gate connection and wire, as shown in Figure 8.4. The heat sink thickness used was 5mm.

In order to keep the MOSFET surface bare to allow a heat sink to be attached, the source connection from the package pin to the source on top of the MOSFETs were omitted during the packaging procedure. Therefore, an electrical connection had to be made from the heat sink to the package source pin. This was done by soldering one end of an enamelled copper wire to the top of the heat sink, and the other end to the source pin. The wire was soldered to the top of the heat sink before it was attached to the device, which minimised the chance of damaging the epoxy joint once it had cured. Once the heat sink was attached using the silver epoxy and fully cured, the other end of the wire could be soldered to the source pin.

8.2.6 Attaching Heat Sinks to the MOSFET Devices

It was necessary to ensure the gate and source remained electrically isolated. To do so, an epoxy that was not electrically conductive was used to mask the gate connection pad and the gate wire. The epoxy used was a regular high temperature epoxy. It was painted onto the gate wire and gated pad and left to cure prior to attaching the heat sink.

The silver epoxy, supplied in separate tubs of resin and hardener, was mixed well at a ratio of 1:1. A thin layer was applied to the base of the heat sinks, which were then carefully placed on top of the devices. A small amount of pressure was applied to the top to ensure any air gaps were pushed out, whilst taking care not to damage the device. The silver epoxy was cured in a solder oven at 90ºC for 15 minutes. Although the epoxy data sheet advised that the curing procedure could take place at room temperature (for 24hours), it was found a better quality and more reliable joint was achieved by curing at an elevated temperature of 90ºC.

A photo of a MOSFET after the heat sink assembly process can be seen in Figure 8.5.
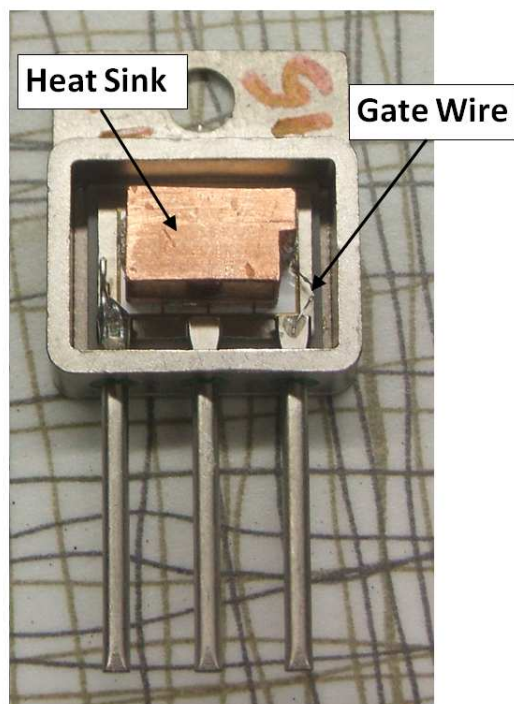
Figure 8.5: An example of a heat sink attached to a MOSFET. It can be seen how the heat sink is shaped to accommodate the gate wire and gate connection pad.

8.3   Experimental Procedure

The experimental testing was performed using the same rig that was used for testing the diodes. Some changes to the rig were required to accommodate differences. These changes include the way the packages were mounted onto the copper baseplates, the electrical setup, and the water coolant pump.

8.3.1   Mounting Devices in Rig

When the diodes were tested in the earlier experimental work, each substrate was soldered onto its own baseplate. It was decided not to do this for the MOSFET devices. It was not known what the bottom of the package was made of, and therefore a good solder bond could not be guaranteed. In addition, it was not desirable to subject the devices to any unnecessary high temperature procedure.

A single copper baseplate was used for all the devices. An M4 tapered hole was made part way through the baseplate which allowed the devices to be attached to the baseplate using a grub screw, passing through the hole in the TO258 packages. To improve the thermal resistance at this joint, a thermal paste was used as a filler in the baseplate - package gap.

8.3.2   Water Coolant Pump

Since the experimental work was performed with the diodes, a new water pump had become available for use. The pump allowed far more accurate water temperature control, as it incorporated a refrigeration unit as well as a heating element. This allowed a more consistent water temperature to be used.

A drawback to the new pump was that the pumping pressure was much lower, meaning lower water velocities and heat transfer coefficients could be achieved at the baseplate surface. This pressure was found to be sufficient for the experimental work and therefore the new pump was used.

8.3.3 Electrical Setup

The experimental procedure varied slightly from that used for the diodes. Three electrical connections had to be made - the gate, source and drain. As explained in Chapter 2, electrical current flows between the source and drain, when a sufficient voltage ($>$ 5V for the MOSFET used) is applied between the gate and source.

Changes had to be made to the electrical setup to accommodate the change in electrical device being tested. The pulse logic from the computer had to control the gate voltage to the switch MOSFETs and also the MOSFET under test. When an 'on' signal is produced by the computer, both the switch MOSFETs and the MOSFET under test had to be switched on in order to allow the heating current through. During 'off' signals, both the MOSFETs had to be off - the switch MOSFET to block the heating current and the MOSFET under test to allow the reverse diode voltage reading to be taken.

The same high current power supply was used to supply the (heating) electrical current between the source and drain, I_{heat}. Pulse logic from the computer provided the power for a fibre optic signal, such that the LED signal was on when the computers pulse logic was on, and off when the logic was off. This signal triggered a 15V gate voltage power supply which was connected to the gate of the two MOSFETs.

After switching the heating current off, the voltage drop across the MOSFET had to be measured to convert into a temperature. This was done by measuring the voltage across the diode region of the MOSFET. However, in a MOSFET this diode is in reverse to the normal electrical flow. This meant the 50mA power source used to provide the small, constant current for voltage measurements needed to be connected in reverse.

The diagram showing the electrical setup of the experiment is shown in Figure 8.6.
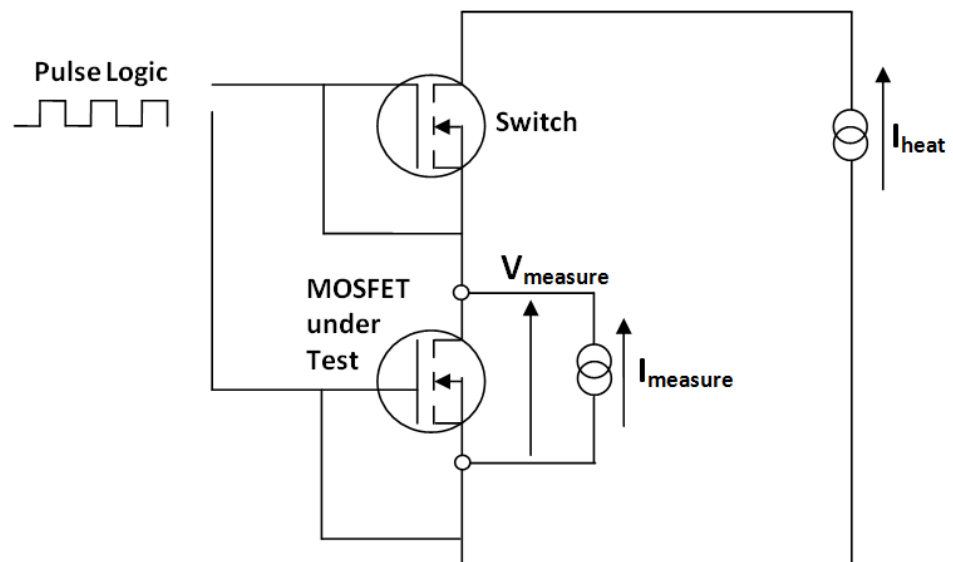
Figure 8.6: Electrical diagram of the experimental setup.

8.4 Experimental Results for the MOSFET Devices

Three MOSFET devices with heat sinks were tested successfully. Each of these devices, as well as a lidded device with no heat sink, were heated up to steady state for 60s. The reverse diode voltage was then measured for 100s during cooling. The transient cooling curve was then inverted to provide the equivalent transient heating curve due to an electrical pulse. The heating curve from each device with a heat sink was compared to the heating curve for the device with no heat sink. The comparison allowed the MoP curve for each to be calculated. This was the same process as described in Section 7.1.3.

8.4.1 Experimental MoP Curves

Figure 8.7 shows the three MoP curves obtained during testing. It can be seen that the three curves have a reasonable level of consistency, and have similar features.

The shape of the MoP is a familiar one, and comparable to those seen in both the numerical modelling and the first experimental results. The MoP value starts at 1 before decreasing with an initially shallow gradient to a minimum value, before rising sharply back towards unity. However, the times at which these features are seen are not as expected.

The heat sinks become effective much later than seen in the previous experimental work using the diodes; the MOSFET temperature is not affected at all until around 100ms into the pulse. Once the heat sinks become effective, their effects are quite slow; after 1s into the pulse, none of the curves have reached a MoP value of less than 0.85. It is not until around 10 to 25s into the pulse that the minimum MoP values are reached.

Although the curves are not as expected in these respects, the three curves show consistency with each other which gives confidence in their repeatability. Devices two

and three in particular are very similar, differing only in the minimum MoP value reached (0.47 for device two and 0.58 for device three). Device one reaches a minimum MoP value of just 0.75, which occurs slightly early than the other two devices. Nevertheless, the consistency between the curves is great enough to suggest the results represent the true behaviour of the heat sinks.
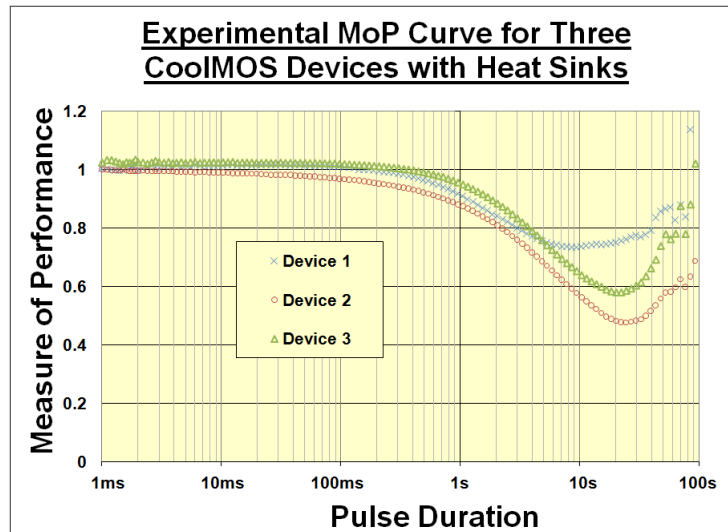
Figure 8.7: The experimental MoP curves obtained for three MOSFET devices with 5mm copper heat sinks.

8.5 Replicating Experimental Results through Numerical Modelling

The experimental MoP curves obtained demonstrated different characteristics to those anticipated. However, as the solder was replaced with silver epoxy, and the package arrangement was different to those modelled previously, it was thought that modelling the new conditions may provide numerical results that correspond to the experimental results.

8.5.1 Model Composition

The MOSFET arrangement was modelled, initially in 1D. The package structure was analysed by sectioning a packaged MOSFET with a heat sink attached. This allowed the layers to be viewed and measured, allowing a more accurate model to be created. The annotated section photos can be seen in Figure 8.8.

The dimensions and material of each layer was determined, which provided the list of model layers in Table 8.2:

Layer	Material	Thickness (mm)
Heat Sink	*Copper*	*5*
Epoxy	*Silver Epoxy*	*0.1*
MOSFET	Silicon	0.2
	Aluminium	0.32
Substrate	Alumina	0.52
	Aluminium	0.16
Package Base	Copper	1.14
Thermal Paste	Thermal Paste	0.1
Baseplate	Copper	6

Table 8.2: The layer material and thicknesses used in the numerical model for replicating the MOSFET experimental results. The layers in italics represent those included only in the models with a heat sink.

Figure 8.8: A photo of the sectioned MOSFET package with an attached heat sink. The labels show the visible components of the package.

8.5.2 Modelling Results

1D modelling was performed to allow a series of model variations to be tested. The main aim of the modelling was to try and emulate the delay in the heat sink effect. 1D modelling was chosen as many situations could be modelled quickly. As shown in Chapter 6, the MoP curve produced from 1D and 2D modelling is identical for pulse durations less than around 10ms. It was also necessary to model the transient temperature for up to 100s, the experimental results demonstrate that little has happened within the 10s pulse duration that is normally modelled. Therefore 1D modelling was initially used to attempt to replicate the MoP curve during the shorter pulses, and then use 2D modelling to attempt to replicate the longer pulses.

Models were set up according to the dimensions and materials in Table 8.2. A model with no heat sink was modelled, as well as a model with a 5mm heat sink, which included the layers in italics in Table 8.2. The heat was generated in the top 10% of

the die, and a constant temperature boundary was applied to the baseplate surface.

In order to replicate the shape of the experimental MoP curve, changes to the model were required. The very late reduction in die temperature suggests a large thermal resistance between the die and heat sink. For the heat sink to be able to reduce the die temperature effectively at pulse durations between 10s and 100s, the heat flow must also be restricted to the cooled baseplate surface. This can be done by increasing the thermal resistance of the thermal paste which used used at the interface between the substrate and baseplate.

To represent these changes, the thermal properties of the silver epoxy and thermal paste were each reduced, dividing them by 100, leading to increases in the thermal resistance across these layers. The modified material properties are shown in Table 8.3. The results obtained from the described modelling setup are shown in Figure 8.9.

It can be seen that modifying the thermal properties of the silver epoxy and thermal paste allow the experimental results to be replicated well. For pulses less than 1s, the heat sink has almost no effect on the die temperature. Both the modelling and experimental MoP curves begin decreasing at around 100ms, and have a value of around 0.95 after 1s. From this point, all curves continue to decreases uniformly.

After 5s, the first deviation is seen. This deviation represents inconsistency in the silver epoxy and thermal paste layers. The modelling curve closely matches the shape of 'Test Piece 3' up until a pulse duration of 10s. The shape of all the curves are very similar in the later stages of the pulse.

As the modelling results have replicated the experimental results well, it is possible to be confident in the experimental procedure. The poor performance of the heat sink can be attributed to poor thermal interfaces where silver epoxy and thermal paste was used. This is likely due to the presence of voids and air gaps in the joints.

	Thermal Conductivity k (W/mK)	Specific Heat Capacity c_p (J/kg K)	Density ρ (kg/m^3)
Silver Epoxy	0.14	12	11.3
Thermal Paste	0.09	14	11.3

Table 8.3: The reduced thermal properties of the silver epoxy and thermal paste.
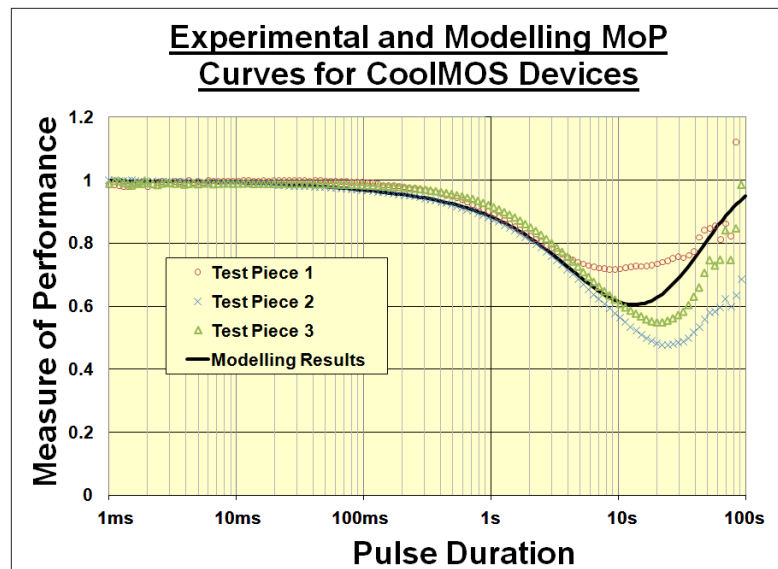
Figure 8.9: A comparison of the CoolMOS experimental and modelling MoP curves.

8.6   Chapter Conclusion

Heat sinks have been successfully attached to a MOSFET device. The heat sink design accommodated the gate connection and imide layer on top of the device. It was not possible to use solder due to the presence of an imide layer on the device and so silver epoxy was used as an alternative, which was electrically and thermally conductive. The epoxy was cured at 90°C as attempts at curing the epoxy at room temperature were unsuccessful as the joint was unable to conduct large currents.

It has been demonstrated that the transient MOSFET temperature can be measured directly after a heating pulse is applied. This was done for devices with and without a heat sink. The experimental results showed that the heat sink did not reduce the die temperature for pulses shorter than 100ms. Minimum MoP values, which ranged from 0.48 and 0.76, were seen between 10s and 25s into the pulse.

Numerical modelling of the experimental setup was able to replicate the experimental results. The thermal resistance across the silver epoxy joint was found to be substantially higher than expected. This is likely due to the presence of air, either as voids or as an air layer across the interface. The thermal properties of the epoxy may also not be as high as claimed by the manufacturer, and should be verified. A thermal paste used to attach the packaged device to the copper baseplate also had a larger thermal resistance than anticipated. This is also likely to be due to regions of air at the interface. The close match between the experimental and modelling MoP curves further demonstrated the ability of the thermal modelling tool to simulate thermal transients for a range of situations.

A better quality thermal interface between die and heat sink should be created. The process of attaching the heat sink to the MOSFET should be examined and improved. Replacing the solder with a silver epoxy introduced unknowns into the quality of the heat sink joint. Solder has been commonly used as an electrical joining material for many decades, and its behaviour and performance is well understood. The same cannot be said about the silver epoxy, and all properties were assumed based on

the manufacturers data sheet. Investigation into the best curing method could be undertaken, with a view to obtaining a high quality electrical, thermal and mechanical joint, similar to those obtained using solder.

Alternatives to the silver epoxy could also be considered. Attractive electrical and thermal properties are the most important consideration when selecting the interface material. The thermal paste was used in the experiments for convenience, and its use in the future should be avoided. Thermal epoxies or solders could achieve a better quality joint than the thermal paste.

Chapter 9

Conclusions and

Recommendations

The concept of a heat sink for suppressing transient temperature rise in solid state power switches has been developed. The heat sink consists of an electrically conductive block being thermally and electrically attached to the entirety of the upper device surface.

Heat dissipation is a significant factor which must be considered when designing solid state power switch systems. Systems must be designed so that the switches are able to withstand the largest power dissipation which they may be subjected to, which are usually transient current surges. The present method of managing current surges involves oversizing power modules to provide more thermal mass, which allows more current to be dissipated without exceeding the temperature limit. Systems designed in this way are over-sized and heavy and are underused for the majority of the time during normal operation. New methods of suppressing transient temperature rises are required in order to increase the efficiency of power electronic systems.

The research presented in this thesis has importance in developing thermal management systems for solid state switches. The investigation focuses on suppressing thermal transients between 100µs and 100ms in length. The concept of using a heat

sink on top of the device surface was considered in order to reduce device temperatures during transients.

A variety of designs were modelled using thermal resistance networks to calculate the transient temperatures inside the device when a heat pulse was simulated. Device temperatures for different designs were compared using a dimensionless Measure of Performance. A sensitivity analysis of the stack layers was conducted by varying the thickness of each independently. Thermal modelling results obtained from each model variation allowed the influence of each stack layer on the transient device temperature to be judged.

Experimental validation of the modelling results was performed to prove the benefits of the heat sink suggested from the modelling work. Test pieces were constructed based on the different designs modelled, using diodes as the power electronic device. Current was pulsed through the diodes which caused heating from the I^2R losses. Transient heating curves were obtained which allowed comparison to the equivalent modelling results under the same conditions.

The scope of the research is similar to that of some other researchers. Cooling semiconductor devices has received much attention in recent years due to the increase in power density capabilities. The majority of this work however focuses on steady state solutions, and relatively little thermal designs are specifically aimed to reducing transient temperatures. The use of double-sided cooling has received some attention as a means of providing an additional thermal path to heat. Again, these designs have been made with steady-state temperature reductions in mind as opposed to short term suppressing effects. The novelty of this research is the focus on very short transients less than 100ms in length.

9.1 Conclusions

9.1.1 Thermal Modelling of Transients in Power Electronic Stack-Ups

This research has successfully demonstrated that thermal RC networks can be used to accurately model transient heat pulses in power electronic devices. Validation of the technique has been achieved using a commercial CFD package and experimental testing. It was possible to examine a variety of situations using the method, allowing a sensitivity analysis on the stack-up to be performed.

9.1.2 Heat Sink Performance during Thermal Transients

Placing the heat sink on top of the die provides an additional thermal path for the generated heat. This reduces the die temperature during transient heat pulses.

The thermal modelling results in Chapter 5 show that the heat sink was able to reduce the die temperature for pulses longer than around 5μs. The heat sink is generally most beneficial for heat pulses between 10ms and 20ms. A MoP value of 0.38 is achievable when using a copper diamond heat sink. This occurred at a pulse duration of around 40 to 60ms. For a copper heat sink the minimum MoP value achieved was 0.41, after around 20 to 30ms.

Heat sink thickness was shown to affect the effectiveness of the heat sink. It was found that for a 100ms pulse the optimum thickness for an aluminium, copper and copper diamond heat sink was 4mm, 5mm and 8mm, respectively. Materials with higher coefficients of heat penetration benefit more from increases in heat sink thickness during the 100ms pulse.

The baseplate has a significant effect on the heat sink performance pulses longer than 20ms. At shorter pulse durations, the solder thickness between the die and heat sink is very influential on die temperature. A thinner solder thickness is preferred in order

to reduce the thermal resistance between the die and heat sink. However, device reliability must also be considered when choosing solder thickness.

Analysis on the region in the die where heat is generated has shown to be critical to heat sink performance. Heat sinks perform most effectively when the heat is generated at the top of the die, close to the heat sink. When heat is generated lower in the die it has further to travel to the heat sink, which causes a delay in the time when the heat sink is able to begin reducing the die temperature.

9.1.3   Effect of Different Heat Sink Materials

An analysis of different heat sink materials was performed in Chapter 5. It was found that the pulse duration at which the heat sink is most effective at reducing the die temperature is dependant on the heat sink material.

Early on in the pulse, whilst heat is conducting through the heat sink in 'early regime', the Coefficient of Heat Penetration (CHP) of the heat sink material dictates its performance: materials with a high CHP reduce the die temperature more than those with a low CHP. As the heat sink becomes saturated and heat conduction into the heat sink enters the 'late regime', the thermal capacity of the material has most effect on the heat sink performance.

Copper and copper diamond are good candidates for the heat sink material, as both provide significant reductions in device temperature during the pulse range of interest. Aluminium could provide a cheaper, lighter alternative with slightly poorer performance.

9.1.4   Heat Spreading Effects

Thermal modelling of the stacks was performed in 2D in Chapter 6. This allowed the effects of heat spreading inside the substrate and baseplate to be identified.

The 1D and 2D MoP results did not differ for pulse durations of less than 40ms. For pulse durations greater than 40ms, the heat sinks become less effective when modelled in 2D. This is due to the increased thermal capacity modelled in the substrate and baseplate.

As the models approach thermal steady state, the MoP in the 1D modelling returned to 1. This was not seen in the 2D modelling. At steady state the heat sink reduces the thermal gradient across the die surface, reducing the maximum temperature. This is seen in the MoP curve: a steady state value of around 0.95 is achieved for heat sinks greater than 2mm in thickness.

9.1.5   Experimental Validation of Heat Sink Performance

Chapter 7 demonstrated the practical application of a heat sink, using diodes as the experimental device. Heat sinks were successfully soldered onto the top of diodes, which had a silver presentation on the top surface. An electrical connection was made to the heat sink, replacing the traditional aluminium wire bonds.

Experimental results showed that the minimum MoP was around 0.48, at a pulse duration of 100ms. Modelling results were obtained by replicating the experimental conditions. A reasonable match between experimental and modelling MoP curves was achieved.

In the experiments the die temperature was found by measuring the voltage drop across the diode, which is proportional to temperature. This allowed temperatures to be measured at a frequency of 10kHz. Experimental noise in the voltage measurements meant temperatures could only be determined accurately for pulse durations longer than around 1ms.

The method by which experimental temperatures are measured could be improved. Aside from the noise issue mentioned, the process of calibrating the voltage and temperature characteristics of each device was time consuming and difficult to control.

The voltage drop across the diode had to be measured at different temperatures. At each temperature, the test piece had to be at a constant temperature to ensure accuracy in the readings. Insulating the test piece to achieve this was difficult and time consuming. Therefore, an improved temperature measurement method could allow more accurate results to be obtained at a higher frequency, that are not as vulnerable to experimental noise.

9.1.6 Application of Heat Sink to Different Power Electronic Devices

Heat sinks were used on MOSFETs, demonstrating that they can be used on a variety of solid state switches that have a more complicated construction than diodes. Heat sinks were successfully glued to the top of MOSFETs using a silver filled epoxy. The epoxy was a suitable replacement for solder, as it provided an electrical connection from the heat sink to the MOSFET.

The experimental results were not as expected. A reduction in die temperature was not seen until around 100ms into the pulse. However, a minimum MoP value of 0.48 was achieved, although it was not until around 20s into the pulse. Similar results were obtained from three different test pieces, suggesting high repeatability.

It was possible to replicate the experimental results through numerical modelling. This was done by modelling very poor thermal properties for the silver epoxy and thermal paste (which joined the substrate and baseplate). This increased the time of the minimum MoP to around 20s, as experienced in the experimental results. It is expected that improving the quality of these thermal joints would change the transient thermal response of the devices so that the heat sinks reduce the die temperature within the 100ms pulse duration which is of most interest.

9.2   Contribution of the Thesis

- A heat sink design has been developed which specifically aims to reduce the temperature of solid state switches for pulse-widths between 100µs and 100ms. The concept of the heat sink is novel and has not been researched or implemented before.

- The normally bare top-side of a solid state switch has been exploited to provide an additional thermal route for heat generated in the device. This has been proven to work experimentally without compromising the electrical functions of the device.

- Transient device junction temperatures have been modelled using thermal resistance networks for heat pulses from 1µs to 10s. This is an unusual situation to model, and the methods previously used to model transient junction temperatures has been developed further. Experimental validation has shown the modelling method is able to accurately predict the transient die temperature for a variety of situations.

- The effects of the different stack layers on the device temperature during a short heat pulse have been investigated. The most significant factors for the heat sink performance have been identified, which allows a solid state switch power module design to be optimised for minimising transient temperatures.

- Inverting the experimental transient cooling curve to provide the equivalent transient heating curve is a novel method for producing a detailed transient temperature profile.

9.3   Recommendations for Future Work

The first recommendation for expanding on the work performed in this thesis would be to extend the experimental side of the project. Initial headway has been made on proving the novel heat sink concept on devices other than diodes. Chapter 8 details the experimental procedure used for testing MOSFETs with heat sinks on, and the problems encountered. Although the results were not as expected, they demonstrate that the heat sink provided some transient benefits to the MOSFET temperature. The chapter concludes by suggesting ways of improving the experimental procedure that should be investigated. The main improvement would be the method of attaching the heat sink to the MOSFET in order to achieve a better thermal contact.

Little is known about the electrical and thermal properties of the silver epoxy used. Investigation into its behaviour would increase the chance of being able to use it successfully. For example, research into the properties of the epoxy after different curing temperatures, and at different operational temperatures would be valuable information. Extending the work in this way would show that the heat sink can be used with versatility, proving that it could be implemented on a wide range of power electronic devices.

Thermal modelling work could be extended by modelling heat pulses in a power module with multiple devices. Thermal modelling work conducted thus far has considered a single device on a substrate and baseplate, therefore only one heat source can affect the device temperature. When many devices share a substrate, heat generated in one device may cause additional heating in an adjacent device as heat is conducted through the substrate. The effect of this heat sharing on the transient device temperature could be modelled, and any significant effects examined.

A further strand of work which should be considered is device reliability. Thermal cycling causes solder fracture and delamination due to mismatches in CTEs of the layers it is bonded to. As copper has a CTE over three times greater than that of silicon, a significant stress can be expected to be exerted on the solder layer which joins

them during heating and cooling. The magnitude of this stress could be investigated through finite element analysis using a package such as ABAQUS. Repetitive thermal cycling must also be examined which simulates the device switching on and off, as this is a common cause of solder failure. This could also be done using finite element analysis, although experimental validation would also be ideal. Acoustic measurements of solder joint delaminations are possible which allows non-destructive monitoring of the solder joint integrity. Assessing the solder joint between the heat sink and device during repetitive cycles would give a good indication as to the long term reliability prospects of using the heat sink in commercial applications.

Bibliography

[1] S.S. Anandan and V. Ramalingam. Thermal Management of Eletronics: A Review of Literature. *Thermal Science*, Volume: 12, Pages: 5–26, 2008.

[2] N. Mohan, T.M. Undeland, and W.P Robbins. *Power Electronics*. John Wiley and Sons, 2003.

[3] S.M. Sze and K.Ng. Kwok. *Physics of Semiconductor Devices*. John Wiley and Sons, 2007.

[4] D.A. Neamen. *An Introduction to Semiconductor Devices*. McGraw-Hill, 2006.

[5] S. Dimitrijev. *Understanding Semiconductor Devices*. Oxford University Press, 2000.

[6] V.A. Vashchenko and V.F. Sinkevitch. *Physical Limitations of Semiconductor Devices*. Springer, 2008.

[7] C. A. Harper. *Electronic Materials and Processes Handbook*. McGraw-Hill, 2003.

[8] M. Trivedi and K. Shenai. Internal Dynamics of IGBTs During Short Circuit Switching. In *Proceedings of the Bipolar/Bicmos Circuits and Technology Meeting*, Pages: 77–80, 1996.

[9] S. Lefebvre, Z. Khatir, and F. Saint-Eve. Experimental Behavior of Single-Chip IGBT and COOLMOS Devices Under Repetitive Short-Circuit Conditions. *IEEE Transactions on Electron Devices*, Volume: 52, Pages: 276–283, 2005.

[10] A. Bejan. *Heat Transfer*. John Wiley and Sons, 1993.

[11] F. Kreith and M. Bohn. *Principles Of Heat Transfer*. CL-Engineering, 2000.

[12] W.A. Scott. *Cooling of Electronic Equipment*. John Wiley and Sons, 1974.

[13] T. Cader, L.J. Westra, and R.C. Eden. Spray Cooling Thermal Management for Increased Device Reliability. *IEEE Transactions on Device and Materials Reliability*, Volume: 4, Pages: 605–613, 2004.

[14] T. Cader and D. Tilton. Implementing Spray Cooling Thermal Management in High Heat flux Applications. In *Proceedings of the 9th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, Pages: 699–701, 2004.

[15] S. V. Garimella. Advances in Mesoscale Thermal Management Technologies for Microelectronics. *Microelectronics Journal*, Volume: 37, Pages: 1165–1185. Elsevier Science Publishers B. V., 2006.

[16] D.J. Womac, S. Ramadhyani, and F.P. Incropera. Correlating Equations for Impingement Cooling of Small Heat Sources with Single Circular Liquid Jets. *Journal of Heat Transfer*, Volume: 115, Pages: 106–115, 1993.

[17] N.T. Obot, W.J.M. Douglas, and A.S. Mujumdar. Effect of Semi-confinement on Impingment Heat Transfer. In *Proceedings of the 7th International Heat Transfer Conference*, Pages: 395–400, 1982.

[18] S. V. Garimella. Heat Transfer and Flow Fields in Confined Jet Impingement. *Annual Review of Heat Transfer*, Volume: 11, Pages: 413–494, 2000.

[19] S. V. Garimella. Confined and Submerged Liquid Jet Impingement Heat Transfer. *ASME Journal of Heat Transfer*, Volume: 117, Pages: 871–877, 1995.

[20] H.A. El-Sheikh and S.V. Gurimella. Enhancement of Air Jet Impingement Heat Transfer Using Pin-Fin Heat Sinks. *IEEE Transactions on Components and Packaging Technologies*, Volume: 23, Pages: 300–308, 2000.

[21] J. Wilson and R. Simons. Advances In High-Performance Cooling For Electronics. *Electronics Cooling*, Volume: November, 2005.

[22] D.Y. Lee and K. Vafai. Comparative Analysis of Jet Impingement and Microchannel Cooling for High Heat Flux Applications. *International Journal of Heat and Mass Transfer*, Volume: 42, Pages: 1555–1568, 1999.

[23] E.G. Colgan, B. Furman, M. Gaynes, W.S. Graham, N.C. LaBianca, J.H. Magerlein, R.J. Polastre, M.B. Rothwell, R.J. Bezama, R. Choudhary, K.C. Marston, H. Toy, J. Wakil, J.A. Zitz, and R.R. Schmidt. A Practical Implementation of Silicon Microchannel Coolers for High Power Chips. *IEEE Transactions on Components and Packaging Technologies*, Volume: 30, Pages: 218–225, 2007.

[24] S.A. Solovitz, L.D. Stevanovic, and R.A. Beaupre. Micro-channel Thermal Management of High Power Devices. In *Proceedings of the 21st Applied Power Electronics Conference and Exposition*, 2006.

[25] D. Lorenzen, J. Bonhaus, W.R. Fahrner, E. Kaulfersch, E. Worner, P. Koidl, K. Unger, D. Muller, S. Rolke, H. Schmidt, and M. Grellmann. Micro Thermal Management of High-Power Diode Laser Bars. *IEEE Transactions on Industrial Electronics*, Volume: 48, Pages: 286–297, 2001.

[26] A.E. Bergles, V J.H. Lienhard, G.E. Kendall, and P. Griffith. Boiling and Evaporation in Small Diameter Channels. *Heat Transfer Engineering*, Volume: 24, Pages: 18–40, 2003.

[27] J.A. Herbsommer, J. Noquil, C. Bull, and O. Lopez. Novel Thermally Enhanced Power Package. In *Proceedings of the 25th Annual IEEE Applied Power Electronics Conference and Exposition*, Pages: 398–400, 2010.

[28] B.C. Charboneau, F. Wang, J.D. van Wyk, D. Boroyevich, Z. Liang, E.P. Scott, and C.W. Tipton. Double-Sided Liquid Cooling for Power Semiconductor Devices using Embedded Power Packaging. In *Proceedings of the 40th IAS Annual Meeting Industry Applications Conference*, Pages: 1138–1143, 2005.

[29] X.C. Ngo, K.D.T. Ngo, and G-Q. Lu. Thermal Design of Power Module to Minimize Peak Transient Temperature. In *Proceedings of the International Confe-*

rence on Electronic Packaging Technology and High Density Packaging, Pages: 248–254, 2009.

[30] W.E. Newell. Transient Thermal Analysis of Solid-State Power Devices - Making A Dreaded Process Easy. *IEEE Transactions on Industry Applications*, Volume: 1A-12, Pages: 405–420, 1976.

[31] H.S. Carslaw and J.C. Jaeger. *Conduction of Heat in Solids*. Oxford University Press, 1959.

[32] S. Clemente. Transient Thermal Response of Power Semiconductor to Short Power Pulses. *IEEE Transactions on Power Electronics*, Volume: 8, Pages: 337–341, 1993.

[33] B. Chambers, T.Y. Tom Lee, and W. Blood. Steady State and Transient Thermal Analysis of Chip Scale Packages. In *Proceedings of the 6th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, Pages: 68–75, 1998.

[34] R.L. Pritchard. *Electrical Characteristics of Transistors*. McGraw-Hill, 1967.

[35] V. Szekely. A New Evaluation Method of Thermal Transient Measurement Results. *Mircroelectronics Journal*, Volume: 28, Pages: 277–292, 1997.

[36] L. Weinberg. *Network Analysis and Synthesis*. R.E. Krieger, 1962.

[37] P. Bagnoli, C. Casarosa, M. Ciampi, and E. Dallago. Thermal Resistance Analysis by Induced Transient (TRAIT) Method for Power Electronic Devices Thermal Characterisation - Part One. *IEEE Transactions on Power Electronics*, Volume: 13, Pages: 1208–1219, 1998.

[38] M.J. Whitehead and C.M. Johnson. Junction Temperature Elevation as a Result of Thermal Cross Coupling in a Multi-Device Power Electronic Module. In *Proceedings of the 1st Electronics Systemintegration Technology Conference*, Pages: 1218–1223, 2006.

[39] Y. L. Xu, R. Stout, and D. Billings. Electronic Package Thermal Response Prediction to Power Surge. *The Seventh Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, Volume: 1, Pages: 366–371, 2000.

[40] P. M. Igic, P.A. Mawby, M.S. Towers, and S. Batcup. Thermal Model of Power Semiconductor Devices for Electro-Thermal Circuit Simulations. Pages: 1–6, 2002.

[41] M. Ludwig, A. Gaedke, O. Slattery, J. Flannery, and S.C. O'Mathuna. Characterisation of Cooling Curves for Power Device Die Attach using a Transient Cooling Curve Measurement. In *Proceedings of the 31st IEEE Annual Power Electronics Specialists Conference*, Pages: 1612–1617, 2000.

[42] N.R. Jankowski and F.P. McCluskey. Modeling Transient Thermal Response of Pulsed Power Electronic Packages. In *Proceedings of the IEEE Pulsed Power Conference*, Pages: 820–825, 2009.

[43] F.N. Masana. A Closed Form Solution of Junction to Substrate Thermal Resistance in Semiconductor Chips. *IEEE Transactions on Components, Packaging, and Manufacturing Technology, Part A*, Volume: 19, Pages: 539–545, 1996.

[44] F.N. Masana. A New Approach to the Dynamic Thermal Modelling of Semiconductor Packages. *Microelectronics Reliability*, Volume: 41, Pages: 901–912, 2001.

[45] R. Lohner. *Applied CFD Techniques: An Introduction based on Finite Element Methods*. Wiley, 2008.

[46] I. Guven, E. Madenci, and C.L. Chan. Transient Two-Dimensional Heat Conduction Analysis of Electronic Packages by Coupled Boundary and Finite Element Methods. *IEEE Transactions on Components and Packaging Technologies*, Volume: 25, Pages: 684–694, 2002.

[47] I.R. Swan, A.T. Bryant, and P.A. Mawby. Fast Thermal Models for Power Device

Packaging. In *Proceedings of the IEEE Industry Applications Society Annual Meeting*, Pages: 1–8, 2008.

[48] A.T. Bryant, N.-A. Parker-Allotey, I.R. Swan, D.P. Hamilton, P.A. Mawby, T. Ueta, T. Nisijima, and K. Hamada. Validation of a Fast Loss and Temperature Simulation Method for Power Converters. In *Proceedings of the 6th International Integrated Power Electronics Systems (CIPS) Conference*, Pages: 1–6, 2010.

[49] I.R. Swan, A.T. Bryant, N.-A. Parker-Allotey, and P.A. Mawby. 3-D Thermal Simulation of Power Module Packaging. In *Proceedings of the IEEE Energy Conversion Congress and Exposition*, Pages: 1247–1254, 2009.

[50] L. Dupont, Z. Khatir, S. Lefebvre, R. Meuret, B. Parmentier, and S. Bontemps. Electrical Characterizations and Evaluation of Thermo-Mechanical Stresses of a Power Module Dedicated to High Temperature Applications. In *Proceedings of the Conference on Power Electronics and Applications*, 2005.

[51] D.C. Katsis and J.D. van Wyk. Void Induced Thermal Impedance in Power Semiconductor Modules: Some Transient Temperature Effects. In *Proceedings of the 36th IAS Annual Meeting Industry Applications Conference*, Pages: 1905–1911, 2001.

[52] K. Hayashi, G. Izuta, K. Murakami, Y. Uegai, and H. Takao. Improvement of Fatigue Life of Solder Joints by Thickness Control of Solder with Wire Bump Technique. In *Proceedings of the 52nd Electronic Components and Technology Conference*, Pages: 1469–1474, 2002.

[53] K. Guth and P. Mahnke. Improving the Thermal Reliability of Large Area Solder Joints in IGBT Power Modules. In *Proceedings of the 4th International Conference on Integrated Power Systems*, Pages: C2–4, 2006.

[54] P. Ratchev, B. Vandevelde, and I. De Wolf. Reliability and Failure Analysis of Sn-Ag-Cu Solder Interconnections for PSGA Packages on Ni/Au Surface Finish.

IEEE Transactions on Device and Materials Reliability, Volume: 4, Pages: 5–10, 2004.

[55] Y. Nishimura, K. Oonishi, A. Morozumi, E. Mochizuki, and Y. Takahashi. All Lead Free IGBT Module with Excellent Reliability. In *Proceedings of the 17th Interntional Symposium on Power Semiconductor Devices and Integrated Circuits*, Pages: 79–82, 2005.

[56] A. Schubert, R. Dudek, E. Auerswald, A. Goldhart, B. Michel, and H. Reichl. Fatigue Life Models for SnAgCu and SnPb Solder Joints Evaluated by Experiments and Simulation. In *Proceedings of the 53rd Electronic Components and Technology Conference*, Pages: 603–610, 2003.

[57] L. Dupont, S. Lefebvre, Z. Zhatir, and S. Bontemps. Evaluation of Substrate Technologies under High Temperature. In *Proceedings of the 4th International Conference on Integrated Power Systems*, Pages: C2–1, 2006.

[58] J-P. Sommer, T. Licht, H. Berg, K. Appelhoff, and M. Bernd. Solder Fatigue at High-Power IGBT Modules. In *Proceedings of the 4th International Conference on Integrated Power Systems*, Pages: C2–5, 2006.

[59] C. Bailey, H. Lu, and T. Tilford. Predicting the Reliability of Power Electronic Modules. In *Proceedings of the 8th International Conference on Electronic Packaging Technology*, Pages: 1–5, 2007.

[60] H. Lu, S. Ridout, C. Bailey, L. Wei Sun, A. Pearl, and M. Johnson. Computer Simulation of Crack Propagation in Power Electronics Module Solder Joints. In *Proceedings of the International Conference on Electronic Packaging Technology and High Density Packaging*, Pages: 1–6, 2008.

[61] H. Lu, T. Tilford, C. Bailey, and D. Newcombe. Lifetime Prediction for Power Electronics Module Substrate Mount-down Solder Interconnect. In *Proceedings of the International Symposium on High Density packaging and Microsystem Integration*, Pages: 1–10, 2007.

[62] C. Zweben. Revolutionary New Thermal Management Materials. *Electronics Cooling Magazine*, Volume: May, 2005.

[63] P.M. Fabis. Thermal Engineering of Electronics Packages using CVD Diamond. In *Proceedings of the 15th Annual IEEE Semiconductor Thermal Measurement and Management Symposium*, Pages: 98–104, 1999.

[64] H.L. Davidson, N.J. Colella, J.A. Kerns, and D. Makowiecki. Copper-Diamond Composite Substrates for Electronic Components. In *Proceedings of the 45th Electronic Components and Technology Conference*, Pages: 538–541, 1995.

[65] Y-T. Chen, J-M. Miao, D-Y Ning, T-F Chu, and W-E Chen. Thermal Performance of a Vapor Chamber Heat Pipe with Diamond-Copper Composition Wick Structures. In *Proceedings of the 4th Microsystems, Packaging, Assembly and Circuits Technology Conference*, Pages: 340–343, 2009.

[66] U. Drofenik and J. W. Kolar. Analyzing the Theoretical Limits of Forced Air-Cooling by Employing Advanced Composite Materials with Thermal Conductivities > 400W/mK. In *Proceedings of the 4th International Conference on Integrated Power Systems*, Pages: C13–1, 2006.

[67] R. W. Smith and R. R. Redd. Active Solder Joining of Thermal Management and Electronic Packaging. In *Proceedings of the 3rd International Brazing and Soldering Conference*, Pages: 79–82, 2006.

[68] S. Narumanchi, M. Mihalic, K. Kelly, and G. Eesley. Thermal Interface Materials for Power Electronics Applications. In *Proceedings of the 11th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, Pages: 395–404, 2008.

[69] S.M. Wentworth, B.L. Dillaman, J.R. Chadwick, C.D. Ellis, and R.W. Johnson. Attenuation in Silver-Filled Conductive Epoxy Interconnects. *IEEE Transactions on Components, Packaging, and Manufacturing Technology, Part A*, Volume: 20, Pages: 52–59, 1997.

[70] R. L. Opitla and J. D. Sinclair. Electrical Reliability of Silver Filled Epoxies for Die Attach. In *Proceedings of the 23rd Annual Reliability Physics Symposium*, Pages: 164–172, 1985.

[71] U. Grigull and H. Sandner. *Heat Conduction*. Hemisphere Publishing Corporation, 1984.

[72] J. Wilson. Thermal Conductivity of Solders. *Electronics Cooling Magazine*, Volume: August, 2006.

[73] R. Siegel and J. Howell. *Thermal Radiation Heat Transfer*. Taylor & Francis, 2002.

[74] M. F. Modest. *Radiative Heat Transfer*. Elsevier Science, 2003.

Appendix A

Thermal Properties of Materials

	Thermal Conductivity k (W/mK)	Specific Heat Capacity c_p (J/kg K)	Density ρ(kg/m^3)
Alumina	18	880	3690
Aluminium	204	896	2707
AlN	160	740	3260
Copper	398	384	8954
Copper Diamond	900	440	5500
Diamond	2330	509	3500
HOPG	1500	709	2300
Iron	73	452	7897
Silver Diamond	600	310	6000
Silver	429	235	10500
Silver Epoxy	14.1	1200	1130
Silicon	148	712	2330
Solder (Sn96.5-Ag3.5)	78	250	7400
Thermal Paste	9.24	1400	1130

Table A.1: The thermal properties of the materials used in the modelling work.[10, 7, 72, 62]

Appendix B

Emissivity Coefficients of Selected Surfaces

Material	Finish	Temperature (deg C)	Emissivity Coefficient, ε
Aluminium	Highly polished	225-575	0.039-0.06
	Commercial sheet	100	0.09
	Heavily oxidised	97-537	0.2-0.33
	Aluminium oxide	275-500	0.63-0.42
Copper	Highly polished	37	0.02
	Polished	115	0.04-0.05
	Commercial, scraped shiny	22	0.072
	Slightly polished	37	0.15
	Black oxidised	37	0.78
Paint	Black shiny lacquer, sprayed on iron	25	0.875
	Black matte shellac	75-145	0.91
	Flat black lacquer	37-97	0.96-0.98
Silicon Carbide		150-650	0.83-0.96
Silver	Polished, pure	225-625	0.02-0.032
	Polished	40-370	0.022-0.031

Table B.1: Radiation emissivity coefficients for selected materials.[73, 74]

Appendix C
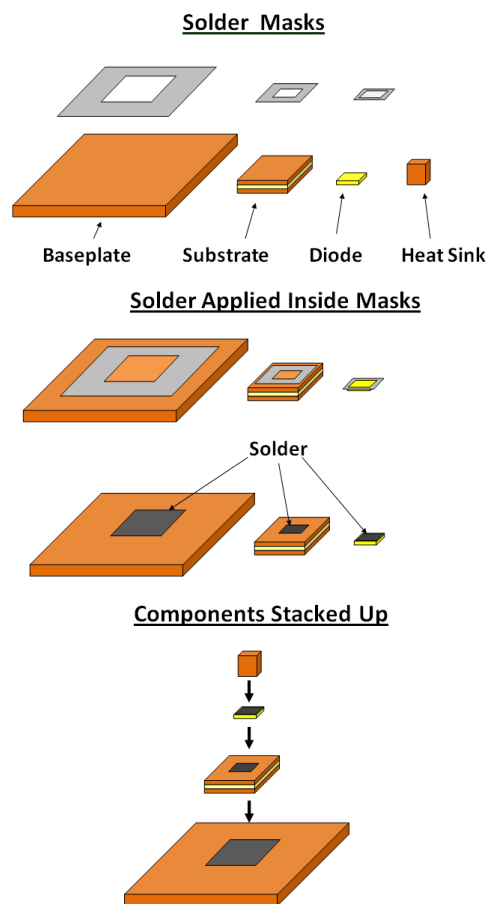
Solder Paste Methodology

Figure C.1: Solder paste application to the components and construction of the test pieces.

Appendix D

Cooling Curve Inversion

The relationship between the transient heating and cooling curve between two steady state temperatures is demonstrated. When the cooling curve is known, the heating curve can be found. At any time from the beginning of the pulse, the temperature of the heating curve ($T_{H,t}$) is found from the temperature of the cooling curve ($T_{C,t}$) at that pulse time:

$$T_{H,t} = T_{C,max} - T_{C,t}$$

where $T_{C,max}$ is the maximum temperature of the cooling curve (at $t = 0$).

Similarly, the cooling curve can be found from the heating curve:

$$T_{C,t} = T_{H,max} - T_{H,t}$$

where $T_{H,max}$ is the maximum temperature during the heating pulse (at the end of the pulse).

Figure D.1 demonstrates the inversion of a typical cooling and heating curve inside the die. The inverted cooling curve perfectly matches the measured heating curve.
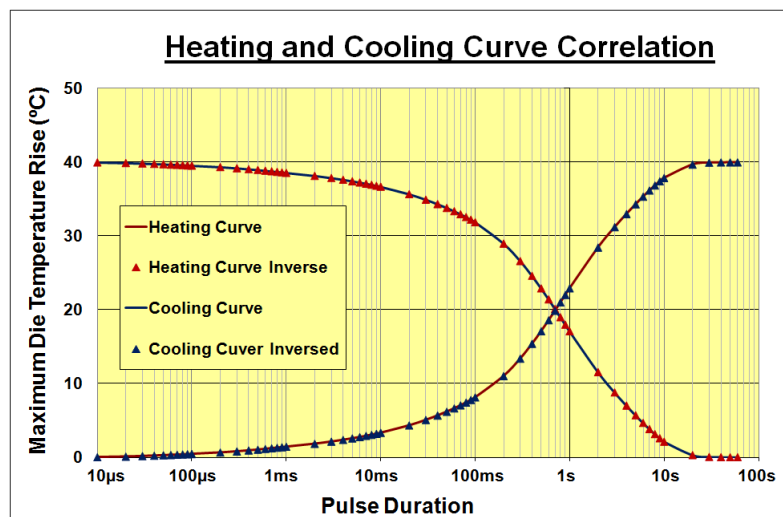
Figure D.1: The measured cooling and heating curves (solid lines) are inverted (triangular points).