Efficient Thermal Simulation Using Model Reduction Techniques

by

Xiaoming Guo

A thesis submitted to the
Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of
Master of Applied Science

Ottawa-Carleton Institute for Electrical Engineering
Department of Electronics
Faculty of Engineering
Carleton University
Ottawa, Canada

© Xiaoming Guo, 2004

National Library
of Canada

Bibliothèque nationale
du Canada

Acquisitions and
Bibliographic Services

Acquisitions et
services bibliographiques

395 Wellington Street
Ottawa ON K1A 0N4
Canada

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

ISBN: 0-612-93935-9

Our file Notre référence

ISBN: 0-612-93935-9

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this dissertation.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de ce manuscrit.

While these forms may be included in the document page count, their removal does not represent any loss of content from the dissertation.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

A new approach of using direct sub-models is proposed in the thesis to be able to simulate complex thermal systems with very large model sizes. The multi-dimensional model reduction(MDMR) technique is applied to direct sub-models with multiple ports to create reduced sub-models. These reduced sub-models are then used to allow for fast simulation of complex devices and packages with a speed-up of 200 to 500 times. Creation of hierarchical reduced sub-models are also presented. The multi-dimensional macro-models(MDMM) created using MDMR are proved to be boundary condition independent(BCI). MDMMs have shown a number of advantages over compact models derived from resistor network models in terms of creation, accuracy, flexibility and capability of predicting multiple junction temperatures.

A global method of forming models is also proposed and implemented to link reduced sub-models to a base model to explicitly incorporate all boundary conditions(BCs) of both the base model and sub-models. It allows one to perform a hierarchical MDMR of a base model linked with reduced sub-models. Moreover, the global method makes a transient simulation of using reduced sub-models become possible.

Finally, the important issue with respect to how to use a compact or a macro-model is addressed in the thesis. Uniform heat flow ports are proposed to accommodate large temperature gradients in the base models of a substrate or printed circuit board(PCB). It has been shown that uniform heat flow ports usually give much better results than the commonly used uniform temperature ports.

Acknowledgments

I would like to express my deepest gratitude to my academic and research advisor Prof. T. Smy for his guidance and constant support to conduct and complete this work.

Special thanks are extended to Mr. D. Celo, for his help, discussions, suggestions and collaborations. I would also like to thank Prof. P. Gunupudi, Prof. D. Walkey and Prof. M. Nakhla for their invaluable help and guidance.

I would like to specially appreciate the Nortel Scholarship for the financial support during this work. All support from the DOE department staffs, specially from Mr. Jacques Lemieux, is also greatly appreciated. I would also like to acknowledge the Center for Microelectronics Assembly and Packaging (CMAP), Micronet and the Natural Sciences and Engineering Research Council of Canada for funding this work.

Many thanks to all the soccer players and the friends at Carleton for this enjoyable period of time.

Finally, I want to extend my profound appreciation to my family for their love and unconditional support during this study.

Dedicated to my wife and sons
for showering unconditional love
and support

Contents

1	Introduction	1
1.1	Motivation of this work	1
1.2	Overview of thermal modelling	2
1.3	<i>Atar</i> simulator	8
1.4	Introduction to model reduction	9
1.4.1	Concepts of moment matching techniques (MTT)	9
1.4.2	Padé approximation	10
1.4.3	The Arnoldi algorithm	13
1.4.4	Model-reduction through Arnoldi	15
1.5	Electro-thermal analogy	16
1.6	Outline of the thesis	16
2	Multi-Dimensional Model Reduction for Fast Steady-State Thermal Simulations	19
2.1	Introduction	19
2.2	Numerical generation of thermal network equations	20
2.2.1	Explicit temperature BCs	22
2.3	Model reduction technique	24
2.3.1	Computation of the multi-dimensional sub-space	25
2.3.2	Application of MDMR to thermal modeling	28

2.4	Results and discussions	28
2.4.1	GaN MESFET amplifier	30
2.4.2	SOI multi-finger clock driver	31
2.4.3	C4/CBGA package	35
2.5	Summary	39
3	Creation and Application of Direct Sub-Models	41
3.1	Introduction	41
3.2	Creation of direct sub-models	43
3.3	Characterization of direct sub-model	48
3.3.1	Example 1: Power amplifier in GaAs MMIC	48
3.3.2	Example 2: Module of C4/CBGA packages on PCB	59
3.4	Summary	68
4	Creation and Application of Reduced Sub-Models Using MDMR	
	Technique	71
4.1	Introduction	71
4.2	MDMR for a multiple port sub-model	73
4.2.1	Computation of the multi-dimensional subspace	73
4.3	Characterization of reduced sub-models	78
4.3.1	Expansion points	79
4.3.2	Number of ports	82
4.3.3	Number of moments	82
4.4	Reduced sub-model linked to base model	84
4.5	Speed-up of using reduced sub-models	89
4.5.1	Example 1: GaAs MMIC power amplifier	89
4.5.2	Example 2: C4/CBGA packages on PCB	91
4.6	Summary	93

5	Creation and Application of Macro-models Using MDMR Technique	94
5.1	Introduction	94
5.2	BCI characteristics of macro-models	96
5.2.1	Example 1: GaAs power amplifier	96
5.2.2	Example 2: MCMBGA	101
5.3	Applications of macro-models	107
5.4	Summary	114
6	Global Method of Linking Sub-Models	115
6.1	Introduction and Motivations	115
6.2	Global model with a single sub-model	116
6.3	Global model with multiple sub-models	119
6.4	Verification of global model solutions	121
6.5	Hierarchical MDMR using global method	123
6.6	Summary	124
7	Use of Macro-Models with Uniform Heat Flow Ports Rather Than Uniform Temperature Ports	125
7.1	Introduction	125
7.2	Multiple port macro-model characterized by T_p and I_p	127
7.3	Generalized method of linking a macro-model to a base model	129
7.3.1	Number of linking nodes more than number of ports	131
7.3.2	Number of linking nodes equal to number of ports	137
7.3.3	Number of linking nodes less than number of ports	138
7.4	Uniform temperature and uniform heat flow ports	139
7.4.1	Uniform temperature ports	140
7.4.2	Uniform heat flow ports	142

7.5	Results and Analysis	146
7.5.1	Fully detailed solution	147
7.5.2	Uniform temperature ports	150
7.5.3	Uniform heat flow ports	151
7.5.4	Effects of BCs and PCB thermal conductivities	153
7.6	Summary	162
8	Conclusions and Future Work	163
8.1	Summary	163
8.2	Future Work/Directions	166

List of Figures

1.1	a) Perspective chip frequency and b) Power dissipation with time [1] .	3
2.1	a) Characterization of the surface BC's as a fixed surface temperature T_f and mixed heat flow represented by an external resistor (specified by λ) and an ambient temperature T_m ; b) Depiction of both fixed and mixed BC present in package model.	23
2.2	Speed-up factor as function of the number of simulations for all three models.	32
2.3	GaN based multi-finger power device. a) Model view; b) Magnified view of a power cell; c) Temperature profile across four power cells. .	33
2.4	Multi-finger device on SOI. a) Model view; b) Magnified view of multi-finger clock driver device; c) Temperature contours; d) Dependence of maximum T_{rise} on λ	36
2.5	Model of low-profile capped CBGA. a) Schematic and the entire model view; b) Section view showing the sub-layers and other details; c) Heat flow for three different sets of BCs applied in the ROM; d) Dependence of maximum T_{rise} on λ	38
3.1	a) Schematic of a mixed-model including a two-port sub-model attached to a PCB base model; b) Relationship between components in the mixed-model.	45

3.2	Flow chart presenting fully detailed model and direct sub-model to simulate a thermal system(T_b represents the nodal temperature vector of the base model, T_s the nodal temperature vector of the sub-model and T_w the nodal temperature vector of the whole model).	46
3.3	5-section cascode gain PA. a) Microphotograph of the die (2.1×1.0 mm); b) Die <i>Atar</i> model.	50
3.4	Components of die macro-model. a) Detailed model of four finger pHEMT transistor-cell with metal connections; b) Sectional view of GaAs substrate model with metal lines and two out of ten 9-port transistor-cell sub-models.	51
3.5	Sparsity of G_b of the base model of GaAs substrate after attaching 10 transistor sub-models. a) 9-port sub-model; b) 182-port sub-model. .	54
3.6	a) Port configuration on the 5 surfaces($n_{px} = n_{py} = 4$ and $n_{pz} = 3$); b) Error with number of ports.	57
3.7	Components of PA macro-model. a) Fully detailed model of die attached to InP carrier; b) 12-port die sub-model attached to InP detailed model to form PA mixed-model with temperature contours show in the carrier only; c) Temperature contours obtained with the application of PA sub-model.	58
3.8	Evaluation of 9-port transistor-cell direct sub-model. a) Junction temperature at each transistor-cell; b) Nodal temperatures at hottest transistor.	60
3.9	a) C4/CBGA package model; b) 16 packages on a PCB (a cross section of the first four packages is shown).	62
3.10	Sparsity of G_b of the PCB model after attaching 16 package sub-models. a) 64-port sub-model; b) 256-port sub-model.	64

3.11	a) Junction temperature with position of packages; b) Error of junction temperature with position of packages (the number of package locations is indexed from bottom to top in each column and from left to right column. The one in the most left bottom corner is number 1 and the most right top corner number 16).	66
3.12	Temperature contours of PCB from fully detailed model. a) $\lambda = 10^{-12}$; b) $\lambda = 10^{-4}$	67
3.13	a) Error of junction temperature with the number of ports; b) Error of junction temperature with λ	69
4.1	Flow chart presenting reduced sub-model to simulate a thermal system.	72
4.2	Errors of 16-port reduced sub-models with λ at different expansion points. a) $\lambda_{\text{expansion}} = 0$; b) $\lambda_{\text{expansion}} = 10^{-4}$; (c) $\lambda_{\text{expansion}} = 0$ and 10^{-4}	81
4.3	Errors of reduced sub-models as function of λ with different number of ports. a) $n_m = 1$; b) $n_m = 3$	83
4.4	Errors of 16-port reduced sub-models with λ using different number of moments. a) $\lambda_{\text{expansion}} = 0$; b) $\lambda_{\text{expansion}} = 0$ and $\lambda_{\text{expansion}} = 10^{-4}$	85
4.5	Errors of the mixed models with with λ using different number of moments used for the 16-por reduced sub-model. a) $\lambda_{\text{expansion}} = 0$; b) $\lambda_{\text{expansion}} = 10^{-4}$; c) $\lambda_{\text{expansion}} = 0$ and $\lambda_{\text{expansion}} = 10^{-4}$	87
4.6	Errors of the mixed models with with λ using different number of ports and $n_m = 3$. a) $\lambda_{\text{expansion}} = 0$; b) $\lambda_{\text{expansion}} = 10^{-4}$	88
4.7	Temperature contours obtained with application of the reduced sub-models of both transistor cells and PA die	90
5.1	Flow chart presenting macro-model using MDMR to simulate a thermal system.	96

5.2	<i>Atar</i> models showing physical structure and mesh. a) Perspective view of the full amplifier; b) Cross-sectional view of the transistors; c) Perspective view of thermal contours; d) Cross-sectional view of thermal contours.	98
5.3	Junction temperatures of GaAs PA macro-model as function of λ and port temperatures T_p	100
5.4	MCMBGA package model. a) View of structure; b) Bottom view showing solder pads with 1.27 mm pitch.	103
5.5	a) Junction temperatures at each of 40 BC sets; b) Error associated with each set	104
5.6	<i>Atar</i> models and temperature contours	109
5.7	Junction temperatures of GaAs macro-model as function of λ for left-side fixed temperature in the substrate.	110
5.8	Temperature contours of the MCMBGA-PCB model obtained with MDMR technique for fixed temperature at right side. a) Package and PCB from fully detailed model; b) PCB including port area from fully detailed model; c) PCB only for single port MDMM; d) PCB only for 4 port MDMM.	112
5.9	Junction temperatures of MCMBGA model as function of λ . a) Variation of λ_1 for $\lambda_2 = 0$; b) Variation of λ_2 for $\lambda_1 = 0$	113
7.1	Macro-model linking to a base model. a) Two models linked node by node; b) Two models linked by two ports; c) Macro-model representation; d) Macro-model linked to a base model.	128
7.2	Illustration of thermal networks via multiple ports. a) Fully detailed model; b) Uniform temperature ports; c) General ports	130

7.3	a) Number of linking nodes more than number of ports; b) Number of linking nodes equal to number of ports; c) Number of linking nodes less than number of ports	132
7.4	Sparsity of matrix $\tilde{\mathbf{G}}_b$ for back-side fixed BC and linking nodes $n_l=256$. a) Linking a single port macro-model using uniform temperature port; b) Linking a single port macro-model using uniform heat flow port; c) Linking 16-port macro-model using uniform temperature port; d) Linking 16-port macro-model using uniform heat flow port.	145
7.5	a) <i>Atar</i> model of package and PCB; b) Cross-section view of package model; c) 64 positions where the temperatures and heat flows are sampled; d) 4-port configuration	148
7.6	3-D profiles of linking node temperatures and heat flow densities from the fully detailed model at $K=19$ ($W/m.K$) and $\lambda = 0$. a) Linking node temperature profile for <i>back-side</i> fixed BC; b) Heat flow density profile for <i>back-side</i> fixed BC; c) Linking node temperature profile for <i>right-side</i> fixed BC; d) Heat flow density profile for <i>right-side</i> fixed BC	149
7.7	3-D profiles of linking node temperatures and heat flow densities using <i>uniform temperature ports</i> for <i>right-side</i> fixed BC. a) Linking node temperature profile using a single port macro-model; b) Linking node heat flow density profiles using a single port macro-model; c) Linking node temperature profile using 4-port macro-model; d) Linking node heat flow density profile using 4-port macro-model.	152

7.8	3-D profiles of linking node temperatures and heat flow densities using <i>uniform heat flow ports</i> for <i>right-side</i> fixed BC. a) Linking node temperature profile using a single port macro-model; b) Linking node heat flow density profiles using a single port macro-model; c) Linking node temperature profile using 4-port macro-model; d) Linking node heat flow density profile using 4-port macro-model.	154
7.9	Maximum temperature errors as function of PCB thermal conductivities using 4-port macro-model for <i>back-side</i> fixed BC. Isotropic thermal conductivity($W/m.K$) indices as: (1) $K_x = K_y = K_z = 2$; (2) $K_x = K_y = K_z = 10$; (3) $K_x = K_y = K_z = 19$; (4) $K_x = K_y = K_z = 50$; (5) $K_x = K_y = K_z = 100$; (6) $K_x = K_y = K_z = 200$. Anisotropic thermal conductivity indices as: (7) $K_x = K_y = 2, K_z = 19$; (8) $K_x = K_y = 19, K_z = 2$; (9) $K_x = K_y = 19, K_z = 190$. a) Relative errors on PCB; b) Relative errors on package; c) Absolute errors on PCB; d) Absolute errors on Package.	158
7.10	Comparison of 3-D profiles of linking node temperatures and heat flow densities between fully detailed model, macro-model with uniform temperature ports and macro-model with uniform heat flow ports at $K=2$ ($W/m.K$) for <i>back-side</i> fixed BC. a) and b) Fully detailed model, c) and d) 4 uniform temperature ports, e) and f) 4 uniform heat flow ports	159

7.11	Maximum temperature errors as function of PCB thermal conductivities using 4-port macro-model for <i>right-side</i> fixed BC. Isotropic thermal conductivity($W/m.K$) indices as: (1) $K_x = K_y = K_z = 2$; (2) $K_x = K_y = K_z = 10$; (3) $K_x = K_y = K_z = 19$; (4) $K_x = K_y = K_z = 50$; (5) $K_x = K_y = K_z = 100$; (6) $K_x = K_y = K_z = 200$. Anisotropic thermal conductivity indices as: (7) $K_x = K_y = 2, K_z = 19$; (8) $K_x = K_y = 19, K_z = 2$; (9) $K_x = K_y = 19, K_z = 190$. a) Relative errors on PCB; b) Relative errors on package; c) Absolute errors on PCB; d) Absolute errors on Package.	160
7.12	Comparison of 3-D profiles of linking node temperatures and heat flow densities between fully detailed model, macro-model with uniform temperature ports and macro-model with uniform heat flow ports at $K = 2 W/m.K$ and <i>right-side</i> fixed BC. a) and b) Fully detailed model, c) and d) 4 uniform temperature ports, e) and f) 4 uniform heat flow ports.	161

List of Tables

1.1	An analogy between thermal and electrical quantities	16
2.1	Model size and simulation times required for all three models. CPU times: t_f – full solution. t_{red} – model reduction. t_R – ROM solution . . .	30
2.2	Comparison of maximum temperature rise and Nodal Error.	31
3.1	Model size and CPU memory requirement	52
3.2	Simulation time and speed-up as function of the number of ports. The simulation time of fully detailed model is 461.9 seconds.	53
3.3	Errors as function of the number of ports.	55
3.4	Model size and junction temperature obtained with the detailed solution and sub-model application.	59
3.5	Comparison of the simulation times between the hierarchical direct sub-model and the fully detailed model.	59
3.6	Port configurations and resultant number of ports.	61
3.7	Model size and CPU requirement	63
3.8	Simulation time with the number of ports. The simulation time of the fully detailed model is 1,459.8 sec.	63
3.9	Nonzero elements in the matrix $\mathbf{G}_b$ and the base model simulation time with different number of ports. The base model has 104,544 nonzero elements and the fully detailed model has nonzero elements of 2,364,256	65

4.1	The effect of expansion points using the 16-port reduced sub-model. .	80
4.2	Memory size and CPU time for model reduction with the number of ports at $n_m = 1$	82
4.3	Size of reduced sub-models with the number of moments	84
4.4	Time required for each step as function of the number of ports using $n_m = 1$	89
4.5	Model size and reduced sub-model generation time for each component in PA and C4/CBGA package model. ¹ Die model is formed out of a die substrate and 10 transistor reduced sub-models	92
4.6	Solution time and junction temperature obtained with fully detailed solution and reduced sub-model application.	92
5.1	Sets of port temperatures for the six port MDMM	101
5.2	40 sets of boundary conditions for MCMBGA	106
5.3	Computational costs and junction temperatures for the two GaAs PA models attached to a substrate MDMM vs. detailed solution.	110
5.4	Model size, simulation times and junction temperatures for each com- ponent of MCM models	112
6.1	Global model solution times and memory requirements with the num- ber of ports and the reduced mode sizes	122
6.2	Comparison of simulation times between the global method and the attachment method.	123
6.3	Size and accuracy of reduced global model as function of number of moments at the expansion point of $\lambda_{expan} = 0$. The reduced sub-model of package of (76×76) is used.	124
7.1	Maximum temperatures of the PCB and package from fully detailed model at $K=19$ ($W/m.K$) and $\lambda = 0$	147

7.2	Maximum temperatures and errors as function of the number of ports using <i>uniform temperature ports</i> for <i>right-side</i> fixed BC, $K=19$ ($W/m.K$) and $\lambda = 0$	151
7.3	Maximum temperatures and errors as function of the number of ports using <i>uniform heat flow ports</i> for <i>right-side</i> fixed BC, $K=19$ ($W/m.K$) and $\lambda = 0$	154
7.4	Comparison of the relative errors of junction temperatures with λ be- tween using uniform temperature ports and uniform heat flow ports at $K=19$ ($W/m.K$)	156

List of Acronyms

VLSI	Very Large Scale Integration
CAD	Computer-Aided Design
MMT	Moment Matching Techniques
MNA	Modified Nodal Analysis
AWE	Asymptotic Waveform Evaluation
CPU	Central Processing Unit
ROM	Reduced Order Model
MDMR	Multi-Dimensional Model Reduction
MDSM	Multi-Dimensional Sub-Model
MDMM	Multi-Dimensional Macro-Model
MCM	Multiple Chip Module
BGA	Ball Grid Array
C4	Controlled Collapsed Chip Connection
CPGA	Ceramic Pin Grid Array
PCB	Printed Circuit Board
PA	Power Amplifier
SS	Steady State
MMIC	Monolithic Microwave Integrated Circuit

GaAs	Gallium Arsenide
pHEMT	Pseudomorphic High Electron Mobility Transistor
w.r.t	with respect to
FEM	Finite Element Method
FDM	Finite Diference Method
CFD	Computational Fluid Dynamics
BCs	Boundary Condictions
BCI	Boundary Condition Independent
3-D	Three-Dimensional
LU	Decomposition a Matrix into a Product of Lower(L) and Upper(U) Triangular Matrices

List of Symbols

G	discretized thermal conductance matrix
B	source vector
T	vector of nodal temperatures
T_p	port temperatures
I_p	port heat flow
M_p	selector vector maps T_p to solution space
P_j	input power at each heat source volume
S	node selector matrix,
I_{λ_i}	heat flows off each external surface
t_{red}	reduction time, sec
t_{sp}	specify-time, sec
t_{sol}	solution time, sec
T_j	junction temperature, K
A	matrix A
A^T	matrix A transposed
0	matrix of zeros
I	identity matrix
I_q	identity matrix of size $q \times q$

$\mathfrak{R}^N$	the set of all real vectors of size N
$\mathfrak{R}^{N \times N}$	the set of all real matrices of size $N \times N$
$(d_{i,j})$	matrix composed of elements $d_{i,j}$ in the i -th row and j -th column
s	Laplace operator ($= j\omega$)
$\kappa(\mathbf{A}, \mathbf{R}, n)$	Krylov subspace
$\mathbf{Q}$	orthogonal transformation matrix
$\mathbf{m}^{(p,i)}$	i^{th} moment vector for an excitation at port p
λ	heat transfer coefficient ($W/\mu m^2 K$)
n_p	number of ports
n_m	number of moments
λ_{expn}	expansion point with respect to λ

Subscripts

p	associated with a port
b	associated with a base model
s	associated with a sub-model
g	associated with a global model

Chapter 1

Introduction

1.1 Motivation of this work

Faster switching speeds and greater functionality is increasing die power generation and producing significant non-uniformities in on-chip heat flux for silicon-based devices as illustrated in Fig.1.1 (a) and (b) [1]–[3]. Placement of more functions in a smaller package has created higher power densities in the package or module and produced more difficult and complicated thermal design problems. Shrinking the cycle-time from design to production creates enormous pressure on packaging design and fabrication. Therefore, the modelling of electro-thermal processes has become increasingly important during semiconductor device development. It has been shown that package and printed circuit board(PCB) thermal design is a high priority in order to maintain system performance and reliability.

As a matter of fact, reliability of electronic parts is a major concern for manufacturers, since early failures in the field can cost an enormous amount to repair, in many cases far more expensive than the origin cost of the product [4]. A variety of physical failure mechanisms have been identified and more than a half of failures of electronic parts are related to thermally overheating. For instance, thermal stresses

and material degradation are one cause of breakdown in electronics components. Mismatch of the thermal coefficient expansion between two different materials, especially at the interface conditions, could result in the separation of interfaces and bonds between different parts in a module at higher temperature. In addition, fatigue in the solder connections and cracking in substrate are common failures in electronics components when operating at off-limit temperatures [5]. In a hybrid electro-optic system, unacceptable temperature distributions could cause temperature sensitive optical components malfunction or failure. Moreover, thermal and electronic effects couple at all levels, from chip to package and system level. An appropriate thermal management imperative is to maintain the electrical performance of the chips and packages.

Currently, even though the thermal modelling at system level can be carried out for analysis of parametric trends, the predictive capabilities of these models are rather limited in most cases. To predict the operating temperature of various critical components as a function of space and time purely through modelling is still not feasible in most cases [3]. So the efficient thermal modelling including both chip and system level is urgently required. The method should be fast, accurate and able to predict all critical temperatures. Simulation results can then play an important role in the selection and optimization of the thermal management solutions [3].

1.2 Overview of thermal modelling

In thermal modelling, a few terminologies are commonly used. A *detailed* model generally means a model that attempts to represent or reconstruct the physical geometry of a device very closely. Thus a detailed model will usually look physically similar to the actual device geometry [6], [7]. By contrast, a *compact* model is a rather simple thermal representation composed of a limited number (less than 10) of thermal

a)

b)

Figure 1.1: a) Perspective chip frequency and b) Power dissipation with time [1]

resistances connecting the critical parts to the outer surfaces of devices. A compact model is actually a behavioral representation that captures the response of some critical points of a component to various boundary conditions [6]–[8]. In this thesis, the term *sub-model* is introduced. A sub-model can be defined as a mathematical representation of a component, which models the component’s thermal behavior by a linear relationship of surface temperatures and heat flows. The use of a sub-model allows replacement of complex parts of a detailed model with a simple mathematical relationship between port temperatures and thermal heat flows. A full thermal simulation of such a system will involve a detailed numerical model, *base model*, to link sub-models through thermal ports. From the base model perspective, the linking size of a sub-model is $n_p \times n_p$ where n_p is the number of ports. The full size of a sub-model itself could be large or small. A *macro-model* is widely referred to as a model with sufficiently small size and a very short simulation time, such as reduced models or reduced sub-models. In general, the size of a macro-model may be a little larger than a compact model and it is usually not suitable for hand calculation.

A detailed model is commonly created by means of numerical simulators based on computational fluid dynamics (CFD), finite-difference methods (FDM) and finite-element methods (FEM), such as FLOTHERM [9], *Atar* [10]–[12] and TAS [13], [14]. These methods can provide a very detailed description of the heat flow and temperature field in the model and are suitable for detailed design. These methods are, however, computationally intensive and their systematic use is often limited by the complexity of the structure to be modelled and the expected simulation effort. Even with the increasing speed of modern computers it is not possible for the designers to explore the design options for optimum solutions [3].

As a fully detailed thermal model could easily have more than 100,000 block nodes, simplification of the thermal simulation model is inevitable. A compact model is commonly used to simplify and speed up thermal analysis. The traditional and simplest

compact model is the junction-to-ambient thermal resistance model. But it is only limited to the case of comparison from one package to another in a standard environment. When boundary conditions are considered, several techniques have been introduced to assess the junction-to-case resistance as well to form a two-resistor compact model. In order to account for various boundary conditions and temperature variations, Bar-Cohen and Krueger [15], [16] developed the star network, a typical three-resistor topology. Later it was found that the boundary conditions of the packages should be governed not only by a fixed temperature but also for convective heat transfer by a heat transfer coefficient λ and environment temperature T_m . Laisance [17] proposed an optimization approach to determining the resistance values for intuitively chosen network topologies.

In the European community, a great effort has been dedicated to the methodology of thermal compact models. The research project "DELPHI" (Development of Libraries and Physical Models for an Integrated Design Environment) was initiated in 1996 to generate boundary condition independent (BCI) compact models. This project was then followed up by another project SEED (Supplier Evaluation and Exploitation of DELPHI), in which component suppliers evaluated the DELPHI methodology. DELPHI compact models are made up of several thermal resistors that connect a junction node (representing the die) to several surface nodes. Thermal links are also allowed between the surface nodes (shunt resistors) [6]. The basic procedures to form DELPHI compact model include: (1) to select network topology; (2) to apply experimentally validated finite element models (FLOTHERM) of the package to a set of boundary conditions (typical numbers recommended in reference [18] are 38 and 99 BC sets) to get junction temperature and heat fluxes through each of the primary sides. (3) to apply a non-linear global optimization tool to best fit these data by minimizing the *cost function* based on the errors between the compact model and the full detailed model [13]. The accuracy of junction temperatures predicted from such

compact models are typically 5 or 10% over a wide range of boundary conditions.

However, such an approach requires the designer to choose the correct compact model topology without strict guidelines, and to perform a time-consuming optimization including simulations of the fully detailed model. Moreover, sufficient accuracy is usually provided only for a limited parametric domain [19]. DELPHI methodology is limited to use for simple PCB packages, not for other components [20]. Another very important issue, not yet fully studied, is how to use a compact model at board and system level simulation [13], [14], [21].

In addition, a number of issues arise when the fully detailed model presents a complicated module/package. Multiple power sources such as in multi-chip modules and wafer level 3D integrated die produce an obvious need for the prediction of multiple junction temperatures. Higher power and integration levels produce a need for a more accurate representation of the package/board interface and therefore an increase in the number of thermal ports. The presence of these factors complicates both the networks and the optimization procedure and could compromise the accuracy of the model. The accuracy is usually poor as a resistor network compact model is used in the PCB board or system level with complex thermal distributions and large temperature gradients [13].

In the literatures, some other ways are reported to simplify thermal modelling. Use of effective thermal conductivities and effective contact area [22] by ignoring technical and process details as well as detailed geometry of a package can reduce model size [23] by coarser meshing. Genetic algorithms have been proposed to identify the variables such as boundary conditions and thermal conductivities that are most important and then select the optimum compact model from the simplified models [20]. Model reduction(MR) has been introduced to speed up simulation as well [19].

The work described above shows that it is necessary to develop an efficient method for thermal simulations which has the capability of quickly predicting multiple tem-

peratures with a high accuracy. In this thesis, a new method, based on the MDMR technique, is proposed to create thermal macro-model. In general, the model reduction technique captures the behavior of a large set of differential/algebraic equations with a reduced (smaller) system. The application of the MR procedure on a detailed discretized thermal model of the component will produce a much smaller model that matches the characteristics of the original thermal model. The final model produced by MR is a set of small matrices that capture the behavior of the component and can be used to represent the model either by synthesizing a small (non-physical) resistive network or by formulating a set of connection equations. The primary difficulty of using MR to form a thermal compact model is in making the model independent of applied boundary conditions. In general, these boundary conditions (BCs) must be specified before the model reduction is performed, in which case the reduced model becomes boundary condition dependent. However, MDMR can overcome this limitation by capturing the behavior of the detailed model as a function of design parameters, making the model BCI.

Multi-dimensional macro-models (MDMM) created using MDMR technique have been shown to have a number of advantages over resistor network models. The generation of a macro-model is at least an order of magnitude faster than the creation of an optimized network model. The MDMM displays very high accuracy typically better than 0.1%. Finally, the MDMM is very flexible allowing for the prediction of all internal temperatures and presenting no limitations on the external configuration of the compact model. The advantages of MDMM can be summaries as:

1. The creation of the macro-model is far faster than the optimization of a specified resistive network.
2. All internal temperatures of the original detailed model are accessible allowing for the prediction of multiple junction temperatures.

3. The accuracy of the temperature prediction is very high because the method does not restrict the model to obey a specific network model.
4. The method allows for design parameters (geometry, heat flow off the surfaces, ambient temperature and multiple power levels) to be incorporated explicitly into the MDMM.
5. The method is very general, allowing for any reasonable number of thermal ports, power sources and external heat flows.

1.3 *Atar* simulator

Atar, developed by the thermal modelling group at Carleton University [10]–[12], is the thermal simulator used in this thesis. A short introduction and background about *Atar* is given as follows. *Atar*, based on FDM, is a novel three-dimensional (3-D) thermal simulation tool for semiconductor integrated devices. The simulator is used to automatically generate an accurate 3-D physical model of the device to be simulated from layout information. The simulator produces an appropriate mesh of the device based on a non-orthogonal rectangular block structure. A 3-D model is built up from the surface of the semiconductor surface, which represents both the substrate of the device and any metallization that is present. A 3-D quad-tree mesh is automatically created such that a fine mesh is produced around heat generation regions, but a moderate number of blocks are used for the entire device. From this model, a network of thermal resistors and capacitors are extracted that mathematically represents the model, which can be solved to obtain the temperature distribution.

Atar can build and link multiple detailed models together. *Atar*'s steady-state (SS) solvers are based on either direct solution of the system or iteration. *Atar* also incorporates a transient solver based on a transmission line matrix (TLM) implementation using a physical extraction of a resistance and capacitance network. The

formulation allows for temperature dependent material parameters and a nonuniform time stepping.

In this thesis, *Atar* is extended to be capable of defining thermal ports and creating sub-models. It also produces all network equations for MDMR. It incorporates sub-models "linking" with a base model to finish a thermal simulation.

1.4 Introduction to model reduction

This section outlines the fundamentals of model reduction techniques following the reference [24]–[27]. Model order reduction techniques have been widely used in the circuit simulations as a way to deal with the large lumped models that result from packaging and interconnect structures. The basic concept is that while these lumped models contain a large number of poles only a fraction of these poles have a significant impact on the circuit response in the frequency range of interest. Model reduction takes advantage of this fact by approximating the original large system with reduced order model that capture the essential characteristics, or the dominant poles of the original system. The reduction methods in the circuit simulation literature can be classified under two main classes: 1) Direct moment matching based on Padé approximation, and 2) Indirect approximations based on finding the leading eigenvalues (those with the largest magnitude).

1.4.1 Concepts of moment matching techniques (MTT)

A brief mathematical description of the underlying concepts of MMTs is given below. Consider a single input/single output system and let $H(s)$ be the transfer function. $H(s)$ can be represented in a rational form as

$$H(s) = \frac{P(s)}{Q(s)} \quad (1.1)$$

where $P(s)$ and $Q(s)$ are polynomials in s . Equivalently, (1.1) can be written as

$$H(s) = c + \sum_{i=1}^{N_p} \frac{k_i}{s - p_i} \quad (1.2)$$

where p_i and k_i are the i^{th} pole residue pair, N_p is the total number of system poles, and c is the direct coupling constant. The time-domain impulse response can be computed in a closed form using inverse Laplace transform as

$$h(t) = c\delta(t) + \sum_{i=1}^{N_p} k_i e^{p_i t} \quad (1.3)$$

In case of large networks, N_p , the total number of poles can be of the order of thousands. Computing all N_p poles will be highly CPU intensive even for a small network and for large networks it is completely impractical. Model-reduction techniques address the above issue by deriving a reduced-order approximation in terms of dominant poles, instead of trying to compute all the poles of a system. Assuming that only dominant poles were extracted, Eq.(1.1) can be rewritten to obtain approximate frequency/time responses, as

$$H(s) \approx \hat{H}(s) = \frac{\hat{P}(s)}{\hat{Q}(s)} = \hat{c} + \sum_{i=1}^L \frac{\hat{k}_i}{s - \hat{p}_i} \quad (1.4)$$

$$h(t) \approx \hat{h}(t) = \hat{c}\delta(t) + \sum_{i=1}^L \hat{k}_i e^{\hat{p}_i t}. \quad (1.5)$$

Note that $L \ll N_p$.

1.4.2 Padé approximation

Consider a system transfer function $H(s)$ that is approximated by a rational function $\bar{H}(s)$ as

$$H(s) \approx \bar{H}(s) = \frac{a_0 + a_1 s + \dots + a_L s^L}{1 + b_1 s + \dots + b_M s^M} = \frac{P_L(s)}{Q_M(s)} \quad (1.6)$$

where $a_0, \dots, a_L, b_1, \dots, b_M$ are the unknowns (total of $L + M + 1$ variables). Considering also the Taylor series expansion of $H(s)$ at $s = 0$ we have,

$$\frac{a_0 + a_1 s + \dots + a_L s^L}{1 + b_1 s + \dots + b_M s^M} = m_0 + m_1 s + m_2 s^2 + \dots + m_{L+M} s^{L+M}. \quad (1.7)$$

Cross multiplying and equating the coefficients of similar powers of s starting from s^{L+1} to s^{L+M} on both sides of Eq.(1.7), the denominator polynomial coefficients can be evaluated as

$$\begin{pmatrix} m_{L-M+1} & m_{L-M+2} & \dots & m_L \\ m_{L-M+2} & \ddots & \vdots & m_{L+1} \\ \vdots & \dots & \ddots & \vdots \\ m_L & m_{L+1} & \dots & m_{L+M+1} \end{pmatrix} \begin{pmatrix} b_M \\ b_{M-1} \\ \vdots \\ b_1 \end{pmatrix} = - \begin{pmatrix} m_{L+1} \\ m_{L+2} \\ \vdots \\ m_{L+M} \end{pmatrix} \quad (1.8)$$

The numerator coefficients can be found by equating the remaining powers of s (from s^0 to s^L) as

$$\begin{aligned} a_0 &= m_0 \\ a_1 &= m_1 + b_1 m_0 \\ &\vdots \\ a_L &= m_L + \sum_{i=1}^{\min(L,M)} b_i m_{L-i} \end{aligned} \quad (1.9)$$

Eqs.(1.8) and (1.9) yield an approximate transfer function in terms of rational polynomials. The computation of the moments $m_0, m_1, m_2, \dots, m_{L+M}$ is discussed in the following section.

Computation of moments

Moment matching techniques such as asymptotic waveform evaluation(AWE) approximate the transfer function of a system using Padé approximation. However, in order to perform Padé approximation, we need to evaluate the moments of the system, which are required by Eq.(1.8) and Eq.(1.9). As an example of moment calculation, consider a linear circuit containing lumped elements only. For such a circuit, the modified nodal analysis (MNA) equations can be written in the frequency domain as

$$\mathbf{G}\mathbf{X}(s) + s\mathbf{C}\mathbf{X}(s) = \mathbf{B} \quad (1.10)$$

where $\mathbf{X} \in \mathbb{R}^n$ is the vector of node voltage appended by independent voltage source currents and linear inductor currents; $\mathbf{C} \in \mathbb{R}^{n \times n}$, $\mathbf{G} \in \mathbb{R}^{n \times n}$ are constant matrices describing the lumped memory and memoryless elements of the network, respectively; and $\mathbf{B} \in \mathbb{R}^n$ is the input vector; and s is Laplace's operator. Note that for impulse input excitation, the laplace transformation $\mathbf{B}$ of the input vector is simply a constant. The moments of this circuit are simply the Taylor coefficients of $\mathbf{X}(s)$ such that

$$\mathbf{X}(s) = \mathbf{M}_0 + \mathbf{M}_1 s + \mathbf{M}_2 s^2 + \dots \quad (1.11)$$

Substitute Eq.(1.11) into Eq.(1.10) as

$$[\mathbf{G} + s\mathbf{C}][\mathbf{M}_0 + \mathbf{M}_1 s + \mathbf{M}_2 s^2 + \dots] = \mathbf{B} \quad (1.12)$$

By equating the coefficients of equal power of s , the following recursive expressions for the moments can be obtained

$$\mathbf{G}\mathbf{M}_0 = \mathbf{B} \quad (1.13a)$$

and

$$\mathbf{G}\mathbf{M}_i = -\mathbf{C}\mathbf{M}_{i-1} \quad i > 0 \quad (1.13b)$$

From Eq.(1.13), one can see that only one LU decomposition is needed to calculation all moments. The k^{th} moment $\mathbf{M}_k$ can therefore be related the zeroth moment $\mathbf{M}_0$ by

$$\mathbf{M}_k = \mathbf{A}^k \mathbf{M}_0 \quad (1.14)$$

where $\mathbf{A} = -\mathbf{G}^{-1}\mathbf{C}$ and $\mathbf{R} = \mathbf{G}^{-1}\mathbf{B}$

Ill-conditioning of Padé

As can be seen from Eq.(1.14), when successive moments are explicitly calculated, they are obtained as powers of $\mathbf{A}$. With the increasing values of k , this process quickly converges to an eigenvector corresponding to an eigenvalue of the largest magnitude [28]. As a result, for relatively large values of k , the explicitly calculated moments $\mathbf{M}_k, \mathbf{M}_{k+1}, \dots$ will not add any extra information to the moments, as all of them contain information only about the largest eigenvalue. In other words, the “moments beyond k ” are almost identical (or parallel to each other), giving rise to the ill-conditioning problem associated with Padé approximation.

1.4.3 The Arnoldi algorithm

The Arnoldi algorithm was suggested to overcome the ill-conditioning problem associated with the Padé approximation. This algorithm [28] first starts by computing a subspace which spans the same space as the moments of the system of Eq.(1.10). This subspace is called the Krylov-subspace and is denoted by $\kappa(\mathbf{A}, \mathbf{R}, n)$. In other words, any vector that is a linear combination of the moments of the system, can be expressed as a linear combination of the columns of $\kappa(\mathbf{A}, \mathbf{R}, n)$. This can be mathematically

expressed as

$$\kappa(A, R, n) = \text{colsp}(\mathbf{K}) \quad (1.15)$$

where,

$$\mathbf{K} = \begin{pmatrix} \mathbf{R} & \mathbf{A}\mathbf{R} & \dots & \mathbf{A}^{n-1}\mathbf{R} \end{pmatrix}. \quad (1.16)$$

$\mathbf{K}$ by itself is likely to be ill-conditioned since the columns of $\mathbf{K}$ are formed based on the sequence $\mathbf{A}^i\mathbf{R}$, which quickly converges to the eigenvector of the largest eigenvalue [28]. To overcome this problem (which is the ill-conditioning problem), the Arnoldi algorithm computes another orthogonal matrix $\mathbf{Q}$ such that

$$\text{colsp}(\mathbf{Q}) = \text{colsp}(\mathbf{K}). \quad (1.17)$$

In contrast to the matrix $\mathbf{K}$, the matrix $\mathbf{Q}$ is well conditioned and as many leading columns of $\mathbf{Q}$ can be constructed to get an accurate solution. The matrix $\mathbf{Q}$ can be computed using techniques such as the *modified Gram-Schmidt algorithm* [28]. A simple implementation of the Arnoldi algorithm which uses modified Gram-Schmidt is given below [24], [26].

Arnoldi Algorithm

```

 $q_1 = \mathbf{R} / \|\mathbf{R}\|_2$  // Get the direction of  $q_1$ 
for  $j = 1$  to  $k$  //  $k$  is the desired number of columns of the  $\mathbf{Q}$  matrix
     $z = \mathbf{A}q_j$ ;
    for  $i = 1$  to  $j$ 
         $h_{ij} = q_i^T z$ 
         $z = z - h_{ij}q_i$ ;
    end for
     $h_{j+1,j} = \|z\|$ ; if  $h_{j+1,j} = 0$ , quit
     $q_j = z / h_{j+1,j}$ ;
end for

```

In a multiple port system, the block moments ($\mathbf{R}$ is a matrix rather than a vector) are introduced [24], [25], [29]. So block Arnoldi process is needed to obtain orthogonal basis $\mathbf{Q}$ of Krylov subspace shown below [25], [29].

Block Arnoldi Algorithm

$[\mathbf{q}_0; \mathbf{R}] = qr(\mathbf{R})$; *qr factorization of $\mathbf{R}$* \\

for $j = 1$ *to* k *do* \\

$\mathbf{z} = \mathbf{A}\mathbf{q}_{j-1}$ \\

for $i = 0$ *to* $j-1$ *do* \\

$h = \mathbf{q}_i^T \mathbf{z}$ \\

$\mathbf{z} = \mathbf{z} - \mathbf{q}_i h$ \\

end for \\

$[\mathbf{q}_j; \mathbf{R}] = qr(\mathbf{z})$; *qr factorization of $\mathbf{z}$* \\

end for \\

$\mathbf{Q} = [\mathbf{q}_0 \ \mathbf{q}_1 \ \dots \ \mathbf{q}_k]$

1.4.4 Model-reduction through Arnoldi

Finding the reduced-order circuit equations can be explained by a change of variables in Eq.(1.10) by mapping the vector $\mathbf{X}$ of dimension n into a smaller vector $\hat{\mathbf{X}}$ of dimension q ($q \ll n$) using an orthogonal matrix $\mathbf{Q}$. This mapping is performed as

$$\mathbf{X} = \mathbf{Q}\hat{\mathbf{X}} \quad (1.18)$$

Using Eq.(1.18), we can rewrite Laplace-domain circuit equations in Eq.(1.10) as

$$\mathbf{G}\mathbf{Q}\hat{\mathbf{X}} + s\mathbf{C}\mathbf{Q}\hat{\mathbf{X}} = \mathbf{B} \quad (1.19)$$

Multiplying both sides of Eq.(1.19) by $\mathbf{Q}^T$ and the reduced system of as

$$\hat{\mathbf{G}}\hat{\mathbf{X}} + s\hat{\mathbf{C}}\hat{\mathbf{X}} = \hat{\mathbf{B}} \quad (1.20)$$

where $\hat{G} = Q^T G Q$; $\hat{C} = Q^T C Q$; $\hat{B} = Q^T B$

1.5 Electro-thermal analogy

A thermal system is found to be analogous to an electrical system if they have similar equations and boundary conditions. The equations describing the behaviors of an electrical system can be transformed into the equations for the analogous thermal system by simply changing the variables. This analogy between a thermal and electrical system allows us to predict the thermal characteristics with widely known electrical basic laws and methods for electrical circuit simulations, such as the model reduction technique described in Section 1.4. The summary of thermal variables, their units and thermal electrical analogies are given in Table 1.1.

Thermal Quantities			Unit	Electrical Quantities			Unit
Temperature rise	ΔT		(K)	Voltage	V		(V)
Heat Flow(Power)	P		(W)	Current	I		(A)
Heat	Q		(J=W.s)	Charge	Q		(C=A.s)
Thermal Resistance	R		(K/W)	Resistance	R		(Ω =V/A)
Thermal Conductance	G		(W/K)	Conductance	G		(S= Ω^{-1})
Thermal Resistivity	ρ		(K.m/W)	Electrical Resistivity	ρ		(Ω .m)
Thermal Conductivity	K		(W/K.m)	Electrical Conductivity	σ		(S/m)
Heat Capacity	C		(J/W)	Capacitance	C		(F=C/V)
Specific heat	c_p		(J/kg.K)	Permittivity	ϵ		(F/m)

Table 1.1: An analogy between thermal and electrical quantities

1.6 Outline of the thesis

In this thesis, efficient thermal simulations based on multi-dimensional model reduction (MDMR) technique are presented. The work can be outlined as follows:

- Chapter 2 presents an introduction work done by the thermal modelling group at Carleton University [30]. This work shows how MDMR allows for

model parameters to be incorporated explicitly into the reduced model. MDMR is used for steady state thermal simulation to produce reduced models which are independent of the boundary conditions applied to the model. This is of significant consequence as from a single MDMR one can now perform many simulations of a detailed model under different boundary conditions.

- In Chapter 3, the work was then extended to creation of direct sub-models. In order to allow for a very large model to be simulated, the method of creating *sub-models* is proposed. Creation of a sub-model is to partition a large model into a couple of smaller models and then link them together through the defined multiple ports to make the large model solvable.
- Chapter 4 extends the MDMR technique to sub-models with multiple ports to create reduced sub-models. A reduced sub-models explicitly incorporates all BCs, so the "connection" parameters of the reduced sub-model with respect to BCs can be specified very quickly in the reduced space. Thus, a typical simulation speed-up of 200 to 500 times will be seen.
- Chapter 5 shows that the macro-models created using MDMR technique are verified to be BCI and have distinct advantages over compact models derived from resistor networks in terms of fast creation, accuracy, flexibility and capability of predicting multiple junction temperatures.
- Chapter 6 presents the *global method* to link sub-models to a base model such that the BCs of both the sub-models and the base are all explicitly incorporated. Therefore, one can perform a hierarchical MDMR after the base model linked with the reduced sub-models by using the global method. Moreover, the global method makes a transient simulation of using reduced sub-models become feasible.

- In Chapter 7, the important issue of using a macro-model at board level and system level simulation is addressed. Uniform heat flow ports and uniform temperature ports are proposed to accommodate different board temperature distribution or thermal properties of substrates or PCBs.
- Finally, conclusions and future work suggestions are given in Chapter 8.

Chapter 2

Multi-Dimensional Model

Reduction for Fast Steady-State

Thermal Simulations

2.1 Introduction

This chapter outlines the work of Multi-dimensional model(MDMR) for steady-state(SS) thermal simulations by the thermal modelling group at Carleton University [30], [31] and is provided as background for the authors work. Model reduction (MR), in general, captures the behavior of a large set of differential/algebraic equations with a reduced (smaller) system. Typically, a set of equations characterized by a sparse matrix is reduced down to much smaller set of strongly coupled equations. The problem can then be solved quickly in the reduced system and the full solution obtained by a mapping back to the large set of original system [30].

In the literature [27], [32], [33], several algorithms were proposed to efficiently simulate large systems of equations by model-order reduction . It should be noted that all these techniques perform model reduction with respect to only a single parameter

(frequency). The MDMR technique is based on Krylov sub-space techniques extended to multiple dimensions [34], [35]. It can form reduced-order models of large linear systems with respect not only to time but also to other design parameters of the system. The use of MDMR for SS thermal simulation can produce reduced models which are independent of the boundary conditions (BCs) applied to the model (including external thermal resistances representing a mixed boundary condition). This is of significant consequence as from a single MR we can now perform many simulations of a detailed model under different boundary conditions.

This chapter proceeds as follows: Section 2.2 first briefly describes the use of *Atar* a finite difference based simulator to generate a thermal model with an explicit incorporation of the BCs in the discretized thermal equations. Section 2.3 outlines the theory behind the MDMR and derives thermal network equations in the reduced space. Section 2.4 presents the results of applying MDMR to realistic examples of IC components and packages. Models are chosen from three different technologies: A Silicon on Insulator (SOI) CMOS process, a Gallium Nitride (GaN) on SiC power amplifier process and a controlled-collapsed-chip-connection/ceramic-ball-grid-array (C4/CBGA) single-chip package. Results are generated and evaluated by comparing with full numerical simulation, the error is found to be very small (in the range of 0.1% of numerical simulation) for all three models considered.

2.2 Numerical generation of thermal network equations

To solve for the conductive heat flow in an IC or package structure requires solving the following partial differential equation [36],

$$\nabla(\kappa(T)\nabla T) + g(x, y, z, T) = \rho C \frac{dT}{dt} \quad (2.1)$$

where $\kappa(T)$ is temperature dependent thermal conductivity (W/mK), $g(x, y, z)$ is the volumetric heat generation (W), ρ is material density (kg/m³), and C is thermal capacity of the material (J/kgK). Equation (2.1) is, in general, non-linear due to the temperature dependent thermal conductivity. This equation is a general form of the heat conduction equation and allows for the calculation of the temperature distribution in a structure containing heat sources for known BCs.

In this thesis, it is assumed that the thermal conductivity is temperature independent, so Eq.(2.1) becomes linear. The first stage in solving this linear equation is to discretize the model to be simulated. For this purpose we have used *Atar* which is a numerical simulator developed at Carleton University [10]–[12]. It is a thermal simulation tool that uses a technology description and layout/geometry information to automatically generate a full 3D model using a non-orthogonal mesh. A thermal model of the structure consisting of resistors, capacitors and power sources is then extracted and the BCs applied [10]. This can be expressed for a SS problem as:

$$\mathbf{G}\mathbf{T} = \mathbf{B} \quad (2.2)$$

where $\mathbf{G}(W/K) \in \mathbb{R}^{N \times N}$ is thermal conductance matrix and N is the node size; $\mathbf{T}(K) \in \mathbb{R}^N$ is the vector of node temperature; $\mathbf{B}(W) \in \mathbb{R}^N$ is the source vector representing internal power generations and applied boundary conditions,

With respect to the boundary conditions for the model, Dirichlet, Neumann and mixed conditions are allowed. *Atar* provides for the application of uniform or non-uniform temperature BCs on each surface of the model. The use of a uniform temperature BC models an average surface temperature which can be assigned to the entire surface or a fraction of the surface (allowing the user to model non-uniformities in the applied temperature distribution). It is also possible to specify an analytical function that describes the temperature distribution at the surface. Mixed boundary condi-

tions are applied in a similar way with the heat flow off the surface $B = \lambda(T_s - T_f) + \beta$ specifying the local heat flow as a function of the surface temperature T_s . The parameters T_f (fixed BC temperature), λ (heat transfer coefficient) and β (outside heat flow) can all be specified for entire surface or for individual surface elements.

Atar in this work is used to build a detailed model, generate the thermal network equations as well as to numerically simulate and solve for the temperature in the detailed model for evaluation purposes. As previously mentioned, FD/FEM programs such as *Atar* provide a very detailed description of the heat flow and temperature field in the model, suitable for detailed design, but are computationally intensive. Computation time increases dramatically with model size because of its dependence on the number of nodes (elements) in the model, limiting the size of a model and forcing model simplifications. To overcome this barrier, the quick model building capability of *Atar* is combined with the MDMR method explained in Section 2.3, creating a more practical design tool useful for simulation of bigger models.

2.2.1 Explicit temperature BCs

For electronic components and packages the simple application of fixed temperatures (T_f) on model surfaces (a Dirichlet BC) does not appropriately capture the presence of external heat drops caused by the convective and conductive cooling. When modelling a package it is often needed to use mixed (Neumann + Dirichlet) BC conditions for some of the model surfaces (Fig. 2.1). A convenient way to represent these mixed boundary conditions is to use an external thermal resistance as shown in Fig. 2.1a. The resistors address the relation between the model and the surrounding by connecting a model surface to a reference temperature, capturing the convective and conductive cooling. According to the Newton's Law of Cooling the thermal transport from a surface to the surrounding can be expressed in the form of $B_s = \lambda A_s(T_s - T_m)$ where B_s is the heat flow transferred from a surface at uniform temperature T_s to

Figure 2.1: a) Characterization of the surface BC's as a fixed surface temperature T_f and mixed heat flow represented by an external resistor (specified by λ) and an ambient temperature T_m ; b) Depiction of both fixed and mixed BC present in package model.

an ambient with a reference temperature T_m , λ ($\text{W}/\text{m}^2\text{K}$) is the heat transfer coefficient and A_s (m^2) is the surface area. As shown in [37] the relation between the external thermal resistance and the coefficient of heat transfer can be expressed as $\theta_{ext} = 1/(\lambda A_s)$ (K/W). Given in this context, the heat transfer coefficient for each surface may represent a generalized heat transfer, including linearized convection or an external conductive resistance. It should be mentioned here that all the external resistances are assumed to be constant (no variation with temperature is taken into account).

We therefore define two types of BC surfaces; those to which a fixed surface temperature T_f is applied and others to for which mixed BC characterized by λ and T_m is used. Having defined boundary conditions in this way, a matrix equation characterizing thermal behavior of a 3D linear model can be written which incorporates all the BCs explicitly:

$$\mathbf{G}_0 \mathbf{T} + \sum_{i=1}^n \lambda_i \mathbf{G}_{\lambda_i} \mathbf{T} = \mathbf{B}_0 + \sum_{i=1}^n \lambda_i T_{m_i} \mathbf{B}_{\lambda_i} + \sum_{k=1}^m T_{f_k} \mathbf{B}_{f_k} + \sum_{j=1}^l P_j \mathbf{B}_{p_j} \quad (2.3)$$

$\mathbf{G}_0 (\text{W}/\text{K}) \in \mathbb{R}^{N \times N}$ is thermal conductance matrix between internal nodes and N is the model size.

$\mathbf{B}_0(W) \in \mathbb{R}^N$ is source vector representing internal power generations.

$\mathbf{T}(K) \in \mathbb{R}^N$ is the vector of node temperature.

$\mathbf{G}_{\lambda_i}(m^2) \in \mathbb{R}^{N \times N}$ where $i \in \{1, 2, \dots, n\}$ is surface node area selection matrix that is used to map the conduction between the mixed BC surface node and the reference to the solution space, n is the number of mixed BCs.

$\mathbf{B}_{\lambda_i}(m^2) \in \mathbb{R}^N$ where $i \in \{1, 2, \dots, n\}$ is surface node area selection vector that is used to map the power flow from mixed BC surface nodes to the solution space.

$\mathbf{B}_{f_k}(W/K) \in \mathbb{R}^N$ where $i \in \{1, 2, \dots, m\}$ is surface conductance selection vector that is used to map power flow from the fixed BC surface nodes to the solution space, m is number of fixed BCs.

$\mathbf{B}_{p_j} \in \mathbb{R}^N$ where $i \in \{1, 2, \dots, l\}$ is selection vector that maps the nodes with input power to the solution space, l is number of power sources.

It should be noted at this point that a direct SS solution of Eq.(2.3) as function of $\lambda_1 \dots \lambda_n$ is made problematic by the presence of the $\sum_{i=1}^n \lambda_i \mathbf{G}_{\lambda_i} \mathbf{T}$ term on the left hand side. The presence of this term requires a completely new solution of the equation for a new set of $\lambda_1 \dots \lambda_n$. In the next section we will address the application of MDMR to Eq.(2.3).

2.3 Model reduction technique

This section briefly outlines the mathematical techniques used to perform the multi-dimensional model reduction of a linear thermal network. In order to find the multi-dimensional solution of Eq.(2.3) the *multi-dimensional sub-space* should be evaluated. This is described in the following sub-section.

2.3.1 Computation of the multi-dimensional sub-space

Evaluation of the multi-dimensional sub-space of Eq.(2.3) is crucial in the computation of multi-dimensional macro-models(MDMs). In order to ensure that the MDMs capture the behavior of Eq.(2.3) w.r.t frequency and other design parameters, sufficient information should be extracted from the original system. This information is generated in the form of multi-dimensional derivatives (moments) of the response of the system in Eq.(2.3). Special techniques based on Krylov sub-space methods are extended in this section to compute the multi-dimensional moments accurately by avoiding ill-conditionality due to a direct moment evaluation [28]. Once the multi-dimensional moments are evaluated they are subjected to QR decomposition to form an orthogonal sub-space [28]. This sub-space is later used in the generation of MDMs. The rest of this section gives a mathematical description of the proposed algorithm.

The moments of $\mathbf{T}$ w.r.t λ_i (where the k^{th} moment of $\mathbf{T}$ w.r.t λ_i is represented as $\mathbf{M}_{\lambda_i}^k$) are computed using the equations:

$$\mathbf{DM}_{\lambda_i}^0 = \mathbf{F} \quad (2.4)$$

$$\mathbf{DM}_{\lambda_i}^1 = T_{m0}\mathbf{B}_{\lambda_i} - \mathbf{G}_{\lambda_i}\mathbf{M}_{\lambda_i}^0 \quad (2.5)$$

$$\mathbf{DM}_{\lambda_i}^2 = -\mathbf{G}_{\lambda_i}\mathbf{M}_{\lambda_i}^1 \quad (2.6)$$

$$\vdots$$

$$\mathbf{DM}_{\lambda_i}^{q_{\lambda_i}} = -\mathbf{G}_{\lambda_i}\mathbf{M}_{\lambda_i}^{q_{\lambda_i}-1} \quad (2.7)$$

where $\mathbf{D} = \mathbf{G}_0 + \sum_{i=1}^n \lambda_{i0}\mathbf{G}_{\lambda_i}$, $\mathbf{F} = \mathbf{B}_0 + \sum_{i=1}^n \lambda_{i0}\mathbf{B}_{\lambda_i} + \sum_{k=1}^m T_{fk0}\mathbf{B}_{f_k} + \sum_{j=1}^p P_{j0}\mathbf{B}_{p_j}$. $\lambda_{i0}, T_{m0}, T_{fk0}$ and P_{j0} are the chosen expansion points of λ_i, T_m, T_{f_k} and P_j respectively and q_{λ_i} is the number of moments to be preserved w.r.t λ_i . It has to be noted that the moments $\mathbf{M}_{\lambda_i}^2 \dots \mathbf{M}_{\lambda_i}^q$ follow a pattern that is amicable to use Krylov sub-space techniques to form an orthogonal sub-space $\bar{\mathbf{Q}}_{\lambda_i}$. The sub-space of

the moments of $\mathbf{T}$ w.r.t λ_i is given as

$$\text{colsp}(\mathbf{Q}_{\lambda_i}) = \text{colsp} \begin{pmatrix} \mathbf{M}_{\lambda_i}^1 & \bar{\mathbf{Q}}_{\lambda_i} \end{pmatrix}. \quad (2.8)$$

Note that the 0^{th} moment w.r.t for all λ_i s is same and is denoted by $\mathbf{M}_{\lambda}^0$. It will be included in the multi-dimensional sub-space later.

Moments of $\mathbf{T}$ w.r.t T_m, T_{f_k} and P_j (for which only one moment exists) are computed using

$$\mathbf{D}\mathbf{M}_{T_{m_i}} = \lambda_{i0}\mathbf{B}_{\lambda_i}, \quad (2.9)$$

$$\mathbf{D}\mathbf{M}_{T_{f_k}} = \mathbf{B}_{f_k} \quad (2.10)$$

and

$$\mathbf{D}\mathbf{M}_{P_j} = \mathbf{B}_{p_j} \quad (2.11)$$

respectively.

Once all the required moments are evaluated, the multi-dimensional sub-space (denoted by $\mathbf{Q}$) is computed such that

$$\text{colsp}(\mathbf{Q}) = \text{colsp}(\mathbf{K}) \quad (2.12)$$

where,

$$\mathbf{K} = \begin{pmatrix} \mathbf{M}_{\lambda}^0 & \mathbf{Q}_{\lambda_1} & \dots & \mathbf{Q}_{\lambda_n} & \mathbf{M}_{T_{m_1}} & \dots & \mathbf{M}_{T_{m_n}} \\ \mathbf{M}_{T_{f_1}} & \dots & \mathbf{M}_{T_{f_m}} & \mathbf{M}_{P_1} & \dots & \mathbf{M}_{P_p} \end{pmatrix} \quad (2.13)$$

This can be achieved by using a standard QR decomposition [28].

In the case when cross-derivatives are required to preserve the network response, the multi-dimensional sub-space is modified to include the cross-derivatives (denoted by $\mathbf{M}_\times$) as

$$\text{colsp}(\mathbf{Q}) = \text{colsp} \begin{pmatrix} \mathbf{K} & \mathbf{M}_\times \end{pmatrix} \quad (2.14)$$

Empirically, it is not essential to include cross-derivatives in-order to preserve the accuracy of typical thermal networks. It is to be noted that once computed, the multi-dimensional sub-space $\mathbf{Q}$ is a constant real matrix independent of frequency and $\lambda_1 \dots \lambda_n$.

We then define:

$$\mathbf{T} = \mathbf{Q}\hat{\mathbf{T}} \quad (2.15)$$

where $\hat{\mathbf{T}} \in \mathbb{R}^q$ and q is the reduced vector of unknowns ($q \ll N$).

Using (2.15) a congruent transformation is performed on (2.3) to give

$$\hat{\mathbf{G}}_0 \hat{\mathbf{T}} + \sum_{i=1}^n \lambda_i \hat{\mathbf{G}}_{\lambda_i} \hat{\mathbf{T}} = \hat{\mathbf{B}}_0 + \sum_{i=1}^n \lambda_i T_m \hat{\mathbf{B}}_{\lambda_i} + \sum_{k=1}^m T_{f_k} \hat{\mathbf{B}}_{f_k} + \sum_{j=1}^p P_j \hat{\mathbf{B}}_{p_j} \quad (2.16)$$

where,

$$\begin{aligned} \hat{\mathbf{G}}_0 &= \mathbf{Q}^T \mathbf{G}_0 \mathbf{Q}; & \hat{\mathbf{G}}_{\lambda_i} &= \mathbf{Q}^T \mathbf{G}_{\lambda_i} \mathbf{Q} \\ \hat{\mathbf{B}}_0 &= \mathbf{Q}^T \mathbf{B}_0; & \hat{\mathbf{B}}_{\lambda_i} &= \mathbf{Q}^T \mathbf{B}_{\lambda_i}; & \hat{\mathbf{B}}_{f_k} &= \mathbf{Q}^T \mathbf{B}_{f_k}; & \hat{\mathbf{B}}_{p_j} &= \mathbf{Q}^T \mathbf{B}_{p_j} \end{aligned} \quad (2.17)$$

The size of the reduced network (2.16) is very small compared to the original network (2.3). The response at any node of the original network can be computed using (2.15) once the solution of the reduced system (2.16) is known.

It can be proved that the time-domain reduced network so formed preserves the dominant eigen-values of the original network (2.3) [27].

2.3.2 Application of MDMR to thermal modeling

The application of MDMR within *Atar* involves incorporating the relationships given in Section 2.3.1 into the simulation procedure. The parameters, matrices and vectors in Eq.(2.3) which describe the detailed model and capture the BC dependence are formulated explicitly by *Atar*. The procedure in section 2.3.1 is then applied to form the reduced model described by Eq.(2.16). This reduced model is then written to file for latter use (it is of course of moderate size as the reduced model is far smaller than original). To obtain a particular solution of a model the values of λ -s are set; *Atar* is then invoked to build the 3D model, load in the reduced model, obtain a solution using the reduced model with the specified BCs and map the reduced solution back to the actual temperatures using $\mathbf{Q}$. At this point the analysis capability of *Atar* can be used to plot temperature distributions and obtain graphical views.

2.4 Results and discussions

Our investigation has revealed that MDMR can successfully be applied to the thermal analysis of a very diverse range of electronic devices and packages, placed under a wide range of boundary conditions. The device and package examples shown in this section are intended to emphasize the broad environment of application of the proposed approach. The importance of obtaining an internal field of the temperature, the large number of elements in the thermal network and the multi-dimensional BCs are common characteristics to all thermal models presented. Three examples will be discussed: 1) A GaN power microwave amplifier die attached to chip carrier. 2) A silicon MOSFET clock driver fabricated in a Silicon-On-Insulator technology and 3) a C4/CBGA large pinout package.

For each example a multi-dimensional model was computed following the method of Section 2.3. All thermal simulations and ROM generations were performed using

a workstation SUNW-Blade100, 2Gb RAM. As expected, cross derivatives were not required to for accurate temperature predictions. To validate the accuracy of the method, a thermal analysis was performed by numerically simulating the detailed model with $Atar$ and the corresponding reduced-order model. In both cases nodal temperatures were obtained, and a node-by-node comparison was performed. As a means of comparing the results from numerical simulation (T_{Atar}) and reduced-order model (T_{ROM}), a prediction error was defined as the absolute value of the difference $T_{Atar} - T_{ROM}$ divided by the correct value T_{Atar} . The multi-dimensional nature of the reduced-order model was evaluated by changing the BCs (sweeping $\lambda_1 \dots \lambda_n$) and either determining the maximum temperature rise in the model or by generating a full 3D temperature field for the entire model.

Of significant interest in the use of the model is the speed-up obtained. This speed-up is, however, difficult to define as it is dependant on the number of simulations the user wishes to undertake on a single model for different BC sets. The CPU cost of generating a reduced-order model (t_{red}) is one LU decomposition of $\mathbf{G}$ and q forward-backward substitutions, in other words essentially equivalent to a direct solution of the detailed model (t_f). As the CPU cost of the MDMR is the same as for a single SS simulation there is no benefit for a single simulation. However, as the CPU cost of a reduced solution (t_R) is very small relative to full solution a speed-up is obtained for multiple simulations at different BC sets. For the purposes of comparison we will define the speed-up factor as a function of the number of simulations for each structure:

$$SF = \frac{n \times t_f}{n \times t_R + t_{red}} \quad (2.18)$$

with n being the number of simulations, and the time of model reduction t_{red} included in the calculation.

Model	Number of nodes	System size		CPU Time (s)		
	Full model	Q	ROM	t_f	t_{red}	t_R
GaN Power Amp	151,221	151,221x18	18x18	321	330	0.64
SOI Clock Driver	58,301	58,301x18	18x18	180	193	0.12
C4/CBGA Package	54,945	54,945x29	29x29	305	323	0.45

Table 2.1: Model size and simulation times required for all three models. CPU times: t_f – full solution. t_{red} – model reduction. t_R – ROM solution

2.4.1 GaN MESFET amplifier

The first example presents a GaN based MESFET microwave power amplifier on a SiC substrate. Such power devices offer a number of advantages over Si and GaAs at high power densities. In the design of such devices effects of device self-heating and thermal coupling between devices are an important consideration [38]. Knowledge of the temperature in the active area of the device is essential for optimizing device/chip design, performance and reliability. The model (Fig. 2.3a) incorporates a full die placed on a chip carrier with a die attach layer present. The component consists of eight power cells with each cell containing a seven finger MESFET fabricated on a GaN mesa (Fig. 2.3b). The simulated die is $2 \times 2 \text{ mm}^2$ on a side representing a $4 \times 4 \text{ mm}^2$ die. Each cell dissipates 2.1 W for a $50 \text{ }\mu\text{m}$ gate finger width, with a total power in the model of 16.8 W. The die is attached to a chip carrier composed of two layers – one a thin die-attach layer of a generally low thermal conductivity and a thicker CuW substrate. This model is an example of where the correct simulation of the junction temperature or 3D temperature field requires meshing from deep sub-micron size (the heat generation region at the drain end of the gate) to the cm scale-lengths of the chip carrier. The BC set for the model includes a uniform temperature at the bottom of the chip carrier, zero heat flow from the top surface of the model and mixed BCs (characterized by λ_1, λ_2) on the right and rear surfaces. It should be emphasized that modelling heat flow off the top surface or through back-end metallization would simply require additional parameters.

Model	BCs ($\lambda_1 \dots \lambda_n$)					Max T_{rise}		Error
	Back (K)	Right	Rear (W/ $\mu\text{m}^2\text{K}$)	Top 1	Top 2	Atar (K)	ROM (K)	%
GaN PA	330	10^{-12}	10^{-12}	–	–	193.5	193.5	10^{-3}
	330	10^{-8}	10^{-6}	–	–	191.8	191.6	10^{-3}
	330	10^{-6}	10^{-5}	–	–	188.9	188.7	10^{-3}
SOI Clock Dr.	300	–	–	10^{-12}	10^{-12}	18.72	18.72	0.02
	300	–	–	10^{-8}	10^{-6}	18.25	18.25	0.06
	300	–	–	10^{-4}	10^{-6}	16.75	16.75	0.04
C4/CBGA	300	10^{-12}	10^{-12}	10^{-12}	–	18.54	18.54	0.04
	300	10^{-6}	10^{-12}	10^{-12}	–	9.09	9.09	0.00
	300	10^{-6}	10^{-6}	10^{-6}	–	9.06	9.05	0.01

Table 2.2: Comparison of maximum temperature rise and Nodal Error.

The original size of the GaN model was 151,221 nodes and application of MDMR produced a model of 18 unknowns (Table 2.1). A comparison of the nodal temperatures from a full simulation and the reduced model showed deviation of less than 0.01% at any node (Table 2.2). The accuracy of the reduced model is clearly seen in Fig. 2.3c which shows the temperature profile across four GaN power mesas at the channel depth. This figure allows for a comparison of the temperatures obtained from the ROM and the full simulation in the region of largest temperature gradient. As can be seen the ROM has very little error. The efficiency of the using the ROM can clearly be seen in the simulation times shown in Table 2.1. The ROM solution is performed in 0.64 seconds, presenting a small number where uncertainty is comparable to the actual solution time. By using the speed-up definition in (2.18) and varying n from 1 to 1000, plot of Fig. 2.2 is obtained. As we can see the ROM is more efficient as the number of the simulations is increased and the speed-up approaches an asymptotic value for very large numbers.

2.4.2 SOI multi-finger clock driver

The second example presents a detailed model of a multi-finger clock driver simulated on SOI technology. As the scaling in bulk silicon technology approaches its limits, SOI

Figure 2.2: Speed-up factor as function of the number of simulations for all three models.

Figure 2.3: GaN based multi-finger power device. a) Model view; b) Magnified view of a power cell; c) Temperature profile across four power cells.

has emerged as a technology for mainstream digital applications to improve device performance [39]. The presence of an oxide layer restricts the initial heat spreading to primarily a 2D plane defined by the top silicon layer, producing a substantially higher thermal resistance and elevated temperatures. Results from [40] show that the junction-to-chip thermal resistance increases by more than a factor of three as more fingers become active, and the conventional use of a single thermal resistance in SPICE type models for SOI devices without differentiating between a single-finger or a multi-finger structure is erroneous. As a result, self-heating, thermal coupling between fingers in a multi-finger device on SOI, and its impact on local interconnects and other devices become an important aspect that needs to be characterized by a very detailed thermal model. Fig. 2.4 shows a detailed model of a multi-finger clock driver device which produces 20 mA and has 10 fingers each of $20 \times 0.2 \mu\text{m}^2$. The model incorporates both device heating effects and ohmic heating in the local metal.

The set of BCs applied to the multi-finger SOI device were: 1) a uniform fixed temperature on the backside of the die, 2) two mixed BCs at the top of finger metalization representing the heat flow out of the structure into the back-end of the chip and 3) a zero heat flow condition on the sides of simulation region. Two different λ -s were used to characterize the mixed BCs as external resistances. The simulation region was made large enough to minimize the coupling effect between the (virtual) adjacent devices and not distort the heat flow in the model.

Fig. 2.4c shows temperature contours for the SOI model under a specific set of BCs ($\lambda_1 = \lambda_2 = 0$ and 300 K in the backside) which forces the heat flow toward the backside. The heat flow in the device is quite complex with the SOI structure causing a dramatic 2D heat flow in the active area before the heat spreads down through the oxide layer to the substrate. The original size of the SOI model was 58,301 nodes and application of MDMR produced a model of 18 unknowns (Table 2.1). A comparison of the nodal temperatures from a full simulation and the reduced

model showed deviation of less than 0.06% at any node (Table 2.2), illustrating the ability of the reduced model to capture a complicated heat flow.

To test the ability of the ROM to capture the response of the model to changes in the BC's λ_1 was swept in the interval 10^{-12} to 10^{-2} for λ_2 constant (this variation in λ_1 represents an extreme change from essentially a fixed boundary condition to an infinite external resistance). Simulations were run for three different values of λ_2 , creating a large number of BC sets which are quickly obtained using the ROM. The small size of the reduced-order model allows the simulation of a very large number of BC sets, which otherwise would have been prohibitively expensive. For each BC set applied to the reduced model a maximum temperature rise (T_{rise}) in the detailed model was obtained. The sweeping of the two parameters represent the effect of different back-end layouts and packaging options for the SOI structure. Fig. 2.4d presents the dependence of T_{rise} as a function of λ_1 and λ_2 . With an increase in either λ -s the associated thermal external resistance θ_{JA} will be decreased resulting in a lower temperature rise in the model. Full *Atar* simulation results for few λ points are also shown in the same plot. It can be clearly seen that the ROM captures the variation in both λ -s very well.

2.4.3 C4/CBGA package

The third example models a C4/CBGA package (schematic shown in the inset of Fig. 2.5a). The C4 interconnect technology offers significant reduction in both the signal delay and the microelectronic packaging volume [41]. In addition, this technology allows electrical interconnect on the ball-array side, while allowing the back of the chip exposed for thermal management options. Thermal simulation of a very detailed model of C4/CBGA package which includes all the package elements is important to control package junction temperature and determine the characteristic thermal resistances θ_{JA} or θ_{JC} of the package.

Figure 2.4: Multi-finger device on SOI. a) Model view; b) Magnified view of multi-finger clock driver device; c) Temperature contours; d) Dependence of maximum T_{rise} on λ .

The entire model incorporates three main parts: a package substrate ($33 \times 33 \times 0.86 \text{ mm}^3$), a silicon digital integrated circuit ($20 \times 20 \times 0.356 \text{ mm}^3$), and an aluminum cap. In the model shown, the ceramic BGA substrate consists of four ceramic layers, two signal layers and two ground layers. The package-to-board interconnection is an array of 1024 quadratic solder balls on 1 mm pitch. In order to take into account the important role of substrate thermal vias 576 thermal vias are included [42]. A total power of 20 W dissipated is assumed to be uniformly spread over a thin layer close to the top side of the chip. The I/O count is 4096 with a C4 grid pitch of $300 \mu\text{m}$ and bump size $200 \mu\text{m}$. A partial view of the model structure of C4/CBGA package is shown in Fig. 2.5b. The set of applied BCs model the package's exposure to the air or other cooling paths to the surroundings. Specifically, the BC set includes a uniform temperature at the bottom of the package (the package/board contact surface) and mixed boundary conditions characterized by λ_1 , λ_2 and λ_3 at each of the external package surfaces.

The application of MDMR on this model required a larger number of moments of $\mathbf{T}$ w.r.t to the λ -s, resulting in larger ROM (Table 2.1). The ability of the reduced model to capture a strongly varying physical situation is illustrated with simulation of three BC sets, chosen to provide an extreme test of the ability of the MDMR to model the effect of changing BCs and also to represent a realistic modelling scenario. A variation in one of the external resistances (top of the package) causes the heat flow to change very significantly. Shown in Fig. 2.5 are temperature contours for three widely differing internal heat flows as determined by the applied BCs. Specifically, the three case are: the entire heat flows off the backside ($\lambda_1, \lambda_2, \lambda_3$ set to zero and $T_{BC} = 300K$), heat flow off both the back and top sides ($\lambda_1, \lambda_2, \lambda_3 = 10^{-5}$ and $T_{BC} = 300K$), and off the top side only ($\lambda_1 = 10^{-3}, \lambda_2, \lambda_3 = 0$ and $T_{BC} = 0K$). It was found that one ROM was able to correctly predict the temperatures for all three cases with a maximum error of 0.04% of any model node temperature.

Figure 2.5: Model of low-profile capped CBGA. a) Schematic and the entire model view; b) Section view showing the sub-layers and other details; c) Heat flow for three different sets of BCs applied in the ROM; d) Dependence of maximum T_{rise} on λ .

As a further test, results shown in Fig. 2.5 illustrate the important effect of varying λ_1 on maximum temperature rise. The relatively small effect of λ_2 and λ_3 (compared to λ_1) is due to the differences in corresponding surface areas where $A_{top} \gg A_{side}$, resulting in $\theta_{ext_{top}} \ll \theta_{ext_{side}}$. As with the SOI model, the ROM captures the variation of the BC's very well. It should be mentioned that the other design parameters such as input power and backside temperature can be varied in the same way, and results can be quickly generated.

2.5 Summary

An approach has been presented to create a reduced-order model based on the model reduction technique. Very large sparse matrices of the original linear system of equations describing a thermal network are replaced by small and dense matrices which form a reduced-order model. This technique reduces the size of the original thermal model by the factor of 3000 to 5000, representing an efficient way of handling multi-scale problems and geometrical complexity.

Multi-dimensional model reduction technique presented here enables the efficient steady-state simulation of a thermal model. By exploiting the multi-dimensional nature of the model reduction, BC parameters can be incorporated explicitly into the reduced model. The final reduced model is therefore independent of the boundary conditions applied to the model (including external thermal resistances). This is considerable use as from a single reduced model we can now perform many simulations of a detailed model under different boundary conditions.

The reduced-order models have been shown to be very accurate, with nodal temperature errors of less than 0.1% over a wide range of physical conditions. Due to the small size of the reduced model, applying multi-dimensional BCs and solving for the full model temperature distribution is easy and efficient with speed-ups of 30 times

typical. Finally the reduced models are simple and small, well suited for use in the design environments employed by the electronic industries.

However, the MDMR technique in some cases are practically limited by the complexity and size of models. To overcome this limitation, the method of sub-models is proposed and will be presented in the next chapters.

Chapter 3

Creation and Application of Direct Sub-Models

3.1 Introduction

In Chapter 2, a great success has been demonstrated by using MDMR producing faster simulations with high accuracy. However, this technique is in some cases practically (not theoretically) limited by the complexity and size of models. For a large model, such as GaAs power amplifier shown later in this chapter with more than 250,000 blocks, a typical SunBlade100 workstation cannot do the LU decomposition of this model. To overcome this limitation, the creation of *sub-models* is proposed. The basic idea of sub-models is to partition a large model into a couple of smaller models and then link them together through multi-ports to make the large model solvable. This method specially possesses advantages in the following situations:

- A large model which includes more than one identical or thermally identical components, such as 10 transistor cells in the GaAs power amplifier shown in Section 3.3.
- An electronic component which is extensively used in different packages or sys-

terms. Thus, the sub-model of the component can be created once and used repeatedly for different packages or systems.

- A full thermal simulation of sub-models will involve a detailed numerical model, referred as *base model*, and “linking” sub-models, creating a mixed system which we refer to as a *mixed-model*. A *hierarchical sub-model* can be performed for a *mixed-model* which has already linked one or more sub-models. This way a large model could be simplified significantly.
- The individual sub-models and the base model could be simulated separately. So parallel simulations could become possible particularly while a lot of design parameters are needed to explore.
- It will be of special benefit if a sub-model can be reduced using MDMR techniques which will be discussed in the following chapter.
- A sub-model can be used as a boundary condition independent (BCI) standard library cell with MDMR technique.
- A reduced sub-model with multi-ports will provide more flexibilities than a resistor compact model to accommodate the cases while a higher temperature gradient is present in the base model.

The methodology of a sub-model generation and application is analogous to an electrical multi-port linear sub-network, a sub-model can be defined as a mathematical representation of a component, which models the component’s thermal behavior by a linear relationship of surface temperatures and heat flows. There is no limitation on the number of thermal ports and the sub-model has sufficient internal information to provide all temperatures of the original detailed model of the component. Sub-models are used to simplify a complex model with multiple instances of the same components. The use of a sub-model allows replacement of complex parts of a detailed model with a

simple mathematical relationship between port temperatures and thermal heat flows and still able to produce all nodal temperatures in the original detailed model at the end of the simulation procedure.

The work presented in this chapter proceeds as follows: first a brief description of the procedure to create a thermal sub-model of an IC component and then a mixed-model of the system is formed by linking the sub-models with the base model. The sub-model predictive capability is evaluated by comparing the results with a fully detailed numerical analysis performed in two example models. Firstly a GaAs power amplifier (PA) attached to an InP carrier and secondly C4/CBGA packages attached to a PCB. Using sub-models, thermal analysis is performed and 3D temperature distribution on the original detailed models are obtained.

3.2 Creation of direct sub-models

A sub-model created directly from a detailed thermal model is referred to as a *direct sub-model*. The interface or contact surfaces between the component and the base model are designated as thermal ports. Including thermal ports as part of the model leads to the creation of the sub-model which can be connected to a base model. Thermal ports allow for links between models, making it possible for the sub-model to exchange heat with the base model. Each thermal port might encompass one or more surface blocks (nodes) of the original model. Defining one surface block per thermal port leads to a fully linked sub-model, mathematically identical to a detailed model which includes both components. Such a sub-model will be large due to the number of ports. Conversely, using a few large ports (many blocks per port) will imply that surface nodes in both the base model and the sub-model are at a common temperature which will produce some errors in the simulation when compared to the fully detailed model.

To demonstrate, a 2-port sub-model with top side heat flow BC is shown schematically in Fig.3.1(a). The geometry of the ports will have an impact on the accuracy and determination of the sub-model internal temperatures due to the distortion caused by uniform surface temperatures imposed on the port area. In addition to the geometry and location, a port is characterized by two parameters: a port temperature T_p and heat flow I_p , analogous to port voltage and port current respectively in an interconnected electrical sub-network.

Thermal network equations characterizing the sub-model in the steady-state condition are generated with *Atar*. By neglecting non-linearities, the thermal network simply contains resistances and power sources, and the system of equations characterizing this thermal network is linear. If both BCs and the thermal ports are explicitly described, the system of equation will have the form [43]–[46]:

$$\mathbf{GT} = \mathbf{B} + \mathbf{M}_p \mathbf{T}_p \quad (3.1)$$

$$\mathbf{I}_p = \mathbf{M}_p^T \mathbf{T} \quad (3.2)$$

where:

$$\mathbf{G}(\lambda) = \mathbf{G}_0 + \sum_{i=1}^n \lambda_i \mathbf{G}_{\lambda_i}$$

$$\mathbf{B}(\lambda, \mathbf{T}_m, \mathbf{T}_f, \mathbf{P}) =$$

$$\mathbf{B}_0 + \sum_{i=1}^n \lambda_i \mathbf{T}_{m_i} \mathbf{B}_{\lambda_i} + \sum_{k=1}^m \mathbf{T}_{f_k} \mathbf{B}_{f_k} + \sum_{j=1}^l \mathbf{P}_j \mathbf{B}_{p_j}$$

where $\mathbf{T} \in \mathbb{R}^N$ is the unknown vector of nodal temperatures in the system, $\mathbf{G} \in \mathbb{R}^{N \times N}$ is discretized thermal conductance matrix and $\mathbf{B} \in \mathbb{R}^N$ is the source vector representing internal power generation and applied boundary conditions. In addition, a number of selection matrices are needed to map BC surfaces and power volumes in the sub-model into the solution space. Surface nodes associated with a mixed BC

Figure 3.1: a) Schematic of a mixed-model including a two-port sub-model attached to a PCB base model; b) Relationship between components in the mixed-model.

Figure 3.2: Flow chart presenting fully detailed model and direct sub-model to simulate a thermal system (T_b represents the nodal temperature vector of the base model, T_s the nodal temperature vector of the sub-model and T_w the nodal temperature vector of the whole model).

are selected using the selector matrices G_{λ_i} and B_{λ_i} and have an associated heat transfer coefficient λ_i . External surfaces at fixed temperature T_{f_k} are selected using the B_{f_k} matrices. B_{p_j} are selector matrices which select nodes associated with the input power P_j at each heat source volume. M_p is a selector vector that maps the port temperatures into the node space of the network (the superscript "T" denotes the transpose), I_p and T_p denote the port heat flows and temperatures respectively.

As we can see the system of Eqs. (3.1) and (3.2) include the port currents I_p in terms of the port temperatures T_p allowing for attachment of the sub-model to the base model. Solving Eq. (3.1) for T and substituting into Eq. (3.2), the expression for the heat flow through ports can be found:

$$I_p = M_p^T G^{-1} B + M_p^T G^{-1} M_p T_p \quad (3.3)$$

By defining an attachment conduction matrix $\mathbf{G}_p$ and source vector $\mathbf{B}_p$ as:

$$\mathbf{G}_p = \mathbf{M}_p^T \mathbf{G}^{-1} \mathbf{M}_p \quad (3.4)$$

$$\mathbf{B}_p = \mathbf{M}_p^T \mathbf{G}^{-1} \mathbf{B} \quad (3.5)$$

Eq. (3.3) can be written in the form:

$$\mathbf{I}_p = \mathbf{B}_p + \mathbf{G}_p \mathbf{T}_p \quad (3.6)$$

This attachment equation is generally quite small as the size of matrix $\mathbf{G}_p$ is only $n_p \times n_p$ where n_p is the number of ports present in the sub-model. We refer to this as the attachment size of the sub-model. The relationship between the sub-model described in Eq. (3.6) and the base model for the specific case of a two-port sub-model is shown in Fig. 3.1(b).

The solution of the two component system shown in Fig. 3.1 where the attached component is described by Eq. (3.1) and the detailed base model described by

$$\mathbf{G}_b \mathbf{T}_b - \mathbf{D}^T \mathbf{I}_p = \mathbf{B}_b \quad (3.7)$$

$$\mathbf{I}_p = \mathbf{B}_p + \mathbf{G}_p \mathbf{T}_p \quad (3.8)$$

$$\mathbf{T}_p = \mathbf{D} \mathbf{T}_b \quad (3.9)$$

where $\mathbf{D}$ is the selector matrix which maps the port temperature into the solution space. $\mathbf{T}_b$ being the nodal temperatures in the base model, can be obtained in a number of ways. As illustrated in Fig. 3.2, the most straightforward method of solving this system is to form a fully detailed model of both components, link them and solve the two sets of equations self-consistently. This procedure follows the top branch of the flow chart and will produce an exact numerical solution to the problem without using the sub-model attachment equation. This method, is however, numerically

intensive and amounts to a full complete detailed solution.

Alternatively, using Eqs. (3.1)-(3.6) a sub-model can be defined and attached to the base model forming a mixed-model. Following the procedure described in the bottom branch of Fig.3.2, the sub-model can be generated from a fully detailed model. The detailed model of the component is initially created in *Atar* with the model equations given in Eq.(3.1) and is valid for any BC set. For a specific set of BCs Eqs.(3.4) and (3.5) represent an instantiation of the sub-model allowing it to be connected to the base model and can be determined by direct methods or using an LU decomposition of $\mathbf{G}$. The connection of the sub-model to the base model creates additional “new” nodal surface temperatures which are the port temperatures. The size of the new system is therefore the original size of the base system increased by the number of ports which is usually insignificant. Solution of the mixed-model determines $\mathbf{T}_b$ and the port temperatures. This then allows the nodal temperatures of the sub-model component to be determined using Eq. (3.1) and the LU decomposition of $\mathbf{G}$ previously done.

3.3 Characterization of direct sub-model

As for a creation and application of sub-models, it is very imperative to know the impact of the number of ports and the port configurations on the memory requirement, CPU simulation time and accuracy. Here given are two examples from real electronic components, a gallium arsenide (GaAs) monolithic microwave integrated circuit (MMIC) amplifier and a module of C4/CBGA packages on PCB.

3.3.1 Example 1: Power amplifier in GaAs MMIC

As an example a sub-model of an MMIC power amplifier(PA) designed and built in Carleton University [47] is created and tested. As part of a fiber optic transmitter system, this PA was realized using GaAs MMIC technology. The simulated model

incorporates a die with a 5-section cascode gain cell structure with layout shown in Fig. 3.3. This die is then attached to an indium phosphide(InP) carrier through top surface pads and contact bumps. Each section of the cascode consists of two power cells each containing a four finger pseudomorphic high electron mobility transistor (pHEMT) with $0.2\ \mu\text{m}$ gate length and $15\ \mu\text{m}$ gate width. A view of the transistor model including a section of GaAs substrate (substrate thickness $100\ \mu\text{m}$) is shown in Fig. 3.4(a). Each cell dissipates $0.03\ \text{W}$ with a total power in the model of $0.3\ \text{W}$. In addition, thermal through vias to the backside and connection pads for DC biasing, radio-frequency (RF) signals and ground are included in the die model shown in Fig. 3.3(b). Finally, the die model is attached to the InP carrier model as shown in Fig. 3.7(a). Attachment to the InP carrier is realized with 12 gold bumps. This model is an example of where the correct simulation of the junction temperature at the gate pinch-off region and the complex heat flow through the GaAs die and InP substrate requires meshing from deep sub-micron to millimeter scale lengths.

In the beginning, only the GaAs Power amplifier die (InP carrier not included) with backside fixed temperature BC is used to study effects of the number of ports and the port configurations on CPU cost, memory requirement and accuracy, etc. The fully detailed *Atar* model of this amplifier consists of 269,547 blocks. A detailed transistor model shown in Fig.3.4(a) was created with 19,900 blocks. The heat flow paths from transistor into the substrate are the metal interconnect lines, the bulk substrate material and thermal ports are defined for these surfaces. As seen from Fig.3.4(b), the transistor sub-model has 5 interfaces connected with the GaAs substrate, so the ports will be located on these 5 surfaces as seen in Fig.3.6(a). Due to the metal lines on the top surface of the substrate, there must be a separate port for each metal line at each side in order to take consideration of the large temperature different between the metal line and bulk substrate. Therefore a minimum of 9 ports is required.

Figure 3.3: 5-section cascode gain PA. a) Microphotograph of the die (2.1×1.0 mm);
b) Die *Atar* model.

Figure 3.4: Components of die macro-model. a) Detailed model of four finger pHEMT transistor-cell with metal connections; b) Sectional view of GaAs substrate model with metal lines and two out of ten 9-port transistor-cell sub-models.

Memory requirement and simulation time with the number of ports

Shown in Table 3.1 are the sizes of the base model and transistor cell as well as the memory requirements for the executions of the fully-detailed solution and sub-model solution. It shows that use of sub-models can save CPU memory requirement compared to the fully-detailed simulation, but the difference for this case is not so substantial since the substrate model is still relatively large due to the complicated features of metal lines and hollow thermal vias of the GaAs substrate.

Model	Fully detailed model	Substrate model	Single transistor model
Blocks	266,465	67,465	19,900
CPU (Mbits)	872	757	396

Table 3.1: Model size and CPU memory requirement

Table 3.2 shows the simulation time with the number of ports. One can see that the time to generate $\mathbf{B}_p$ and $\mathbf{G}_p$ strongly increases with the number of ports. But the time to obtain solutions of the base model and the sub-models are only slightly affected by the number of ports. It could be mainly attributed to the two facts: (1) linking a sub-model to a base model only adds the same number of nodes as the number of ports, which usually is insignificant compared to the number of nodes of the base model(67,465); (2) the LU decomposition of the sub-model, which is the major CPU cost to obtain the internal temperatures of a sub-model, had been stored and available while $\mathbf{B}_p$ and $\mathbf{G}_p$ were generated. From Table 3.2, it also shows that the use of sub-models can be 3.8 times faster than the fully-detailed model, which takes 461.9 seconds.

It is important to note that even though the number of nodes added to the base model is insignificant, but the added nonzero elements could seriously affect the simulation time of the base model if a large number of ports is used. As it is well known that the original $\mathbf{G}_b$ is very sparse (on average only 6 to 7 nonzero elements associated with each node). But the added $n_p \times n_p$ matrix is completely dense, which means

that $n_p \times n_p$ nonzero elements are added to $\mathbf{G}_b$ as shown in the right-bottom corner of Fig.3.5. The situation would get even serious if multiple sub-models are attached.

For this GaAs PA model, the number of ports did not show any effect on the simulation time of the base model as shown in Table3.2. That is because the number of nonzero elements of the base model is always dominant in all these cases as seen in Fig.3.5. The base model $\mathbf{G}_b$ of GaAs substrate has nonzero elements of 442,863. Only 331,240 ($182 \times 182 \times 10$) nonzero elements are added when 10 sub-models of the transistor with 182 ports are attached. A substantial effect of nonzero elements added to a base model will be shown in the C4/CBGA package model discussed in Section 3.3.2.

(n_p)	Time to generate B_p and G_p (sec)	Time to simulate base model(sec)	Time to simulate 10 sub-models(sec)	Speed-up
9	23.1	90.3	8.8	3.8
20	28.2	91.9	9.1	3.6
56	43.2	108.1	9.1	2.3
88	56.6	135.8	8.8	2.0
182	100.3	150.9	8.9	1.8

Table 3.2: Simulation time and speed-up as function of the number of ports. The simulation time of fully detailed model is 461.9 seconds.

Error of a sub-model with the number of ports

For the transistor cell, if the ports are defined in such a way that every surface block of the transistor is a single port (*port/block*) and then 182 ports are formed, which is mathematically equivalent to the fully detailed model. Ideally, a port spanning more than one block should be chosen on the isothermal plane as it will not then introduce any error. Otherwise, an error would be introduced due to the fact that the surface nodes within a port are all forced to be the port temperature. This implies that the accuracy of a sub-model depends on not only the number of ports but also the port configuration with respect to temperature distributions.

Figure 3.5: Sparsity of G_b of the base model of GaAs substrate after attaching 10 transistor sub-models. a) 9-port sub-model; b) 182-port sub-model.

The accuracy of sub-models are evaluated with respect to the numerical solution of a fully-detailed model for given sets of BC's. The error in nodal temperatures prediction is defined to be the difference in temperatures, usually junction temperatures:

$$\epsilon = \frac{|T_{det} - T_{sub}|}{T_{det} - 300} \quad (3.10)$$

Where T_{det} is the temperature of interest from the detailed model and T_{sub} is the corresponding temperature from the sub-model. The prediction error as shown, can be calculated at each node allowing the computation of the maximum prediction error on the entire model, or at a single point where the maximum temperature in the model is reached.

Listed in Table 3.3 are the errors as function of the number of ports n_p , where n_{px} , n_{py} and n_{pz} is the number of defined ports in X , Y and Z directions, respectively. As expected the error consistently decreases with the number of ports and down to 0 for 182 port sub-model as it is equivalent to the fully-detailed model. Table 3.3 also shows that the error depends on not only the number of ports but also the port configurations. The sub-model with 9 ports or 20 ports gives almost same accuracy. But big improvement can be found as the number of ports increase from 20 to 28 ports by increasing one more port in Z direction. As a fixed BC is applied on the backside of substrate, the temperature gradient in Z direction is expected to be larger than in X and Y directions. Thus the number of ports in Z direction would play a more important role than in X, Y directions.

Number of ports(n_p)	$n_{px} \times n_{py} \times n_{pz}$	Error (%)
9	$1 \times 1 \times 2$	0.43
20	$2 \times 2 \times 2$	0.42
28	$2 \times 2 \times 3$	0.28
56	$4 \times 4 \times 3$	0.11
88	$4 \times 4 \times 5$	0.093
182	port/block	0

Table 3.3: Errors as function of the number of ports.

Hierarchical sub-model

As a final test of the sub-model technique we simulate the whole model which includes both the GaAs power amplifier(PA) die and the InP carrier as Fig.3.7(a). The procedure of creating a sub-model was applied to the mixed-model of the PA die producing a *hierarchical* sub-model which already has 10 transistor sub-models present. The whole simulation is then realized by linking this hierarchical sub-model to the InP carrier. In principle, there is no limitation to create hierarchical sub-models. This way a complicated model will be gradually simplified into a very simple one.

For this specific application a 9-port transistor cell sub-model was found to satisfy the accuracy requirements. The mixed-model of the PA die was therefore formed using ten 9-port transistor sub-models attached into the substrate as shown in Fig.3.4(b). A 12-port hierarchical sub-model is then generated. The set of BCs for the die model includes heat flow off the bottom surface characterized by heat transfer coefficient λ ($W/\mu m^2 K$) and ports on the top surface for interconnection with the InP base model. The number of ports in the die sub-model was imposed by the number of contact pads in the substrate where each pad was naturally assigned as a port. The attachment of the 12-port PA die sub-model into the InP carrier detailed model is shown in Fig.3.7(b) and it creates a higher level mixed-model of the entire system. A fixed temperature BC ($T = 300\text{ K}$) on the outer bumps of the InP carrier was used. The mixed-model simulation was then completed using the 12 port PA die sub-model and a detailed model of the InP substrate comprised of 3,082 blocks. Once the simulation is complete the port temperatures were used to determine first the amplifier substrate temperatures and then the internal transistor temperatures. These internal temperatures can be plotted or visualized. Figs.3.7(b) and (c) present contour plots of the thermal distribution in the InP carrier and the amplifier whereas Fig.3.8(a) plots the surface temperature of a transistor through the gate fingers. To demonstrate the accuracy of the sub-model critical junction temperatures for each transistor cell

a)

b)

Figure 3.6: a) Port configuration on the 5 surfaces($n_{px} = n_{py} = 4$ and $n_{pz} = 3$); b) Error with number of ports.

Figure 3.7: Components of PA macro-model. a) Fully detailed model of die attached to InP carrier; b) 12-port die sub-model attached to InP detailed model to form PA mixed-model with temperature contours show in the carrier only; c) Temperature contours obtained with the application of PA sub-model.

obtained with a fully detailed model and 9-port transistor cell sub-model are shown in Fig.3.8(b).

In Table 3.4 the model sizes and junction temperatures for the entire system are shown. Using the 10 transistor sub-models to build the hierarchical sub-model of PA die, the total model size of only 3,094 nodes was achieved. The accuracy to predict the junction temperature is about 1%.

Model	Model size	T_j (K)
Detailed	269,547	105.4
Transistor	19,900	
Die sub-model	67,465	
InP+Die sub-model	3,094	104.3

Table 3.4: Model size and junction temperature obtained with the detailed solution and sub-model application.

As seen from Table 3.5, the simulation of the direct hierarchical sub-model(174.7 sec) is 2.9 times faster than the fully detailed model(512.4 sec). It also shows that the major CPU cost of the hierarchical sub-model is the creation time of the PA die sub-model due to its large model size of 67,465 as shown in Table 3.4.

Model	To generate B_p and G_p (sec)	To obtain all internal temps (sec)	Total time (sec)
Fully detailed	...	512.4	512.4
Transistor sub-model	23.1	8.8	
Die sub-model	140.6	1.7	
InP+Die sub-model		0.5	174.7

Table 3.5: Comparison of the simulation times between the hierarchical direct sub-model and the fully detailed model.

3.3.2 Example 2: Module of C4/CBGA packages on PCB

We also investigated the use of sub-models for simulation of a package system. Shown in Fig.3.9(a) is the model of a generic C4/CBGA package with 64 lead bumps located on the back-side of the model. The module to be simulated is composed of 16 package

Figure 3.8: Evaluation of 9-port transistor-cell direct sub-model. a) Junction temperature at each transistor-cell; b) Nodal temperatures at hottest transistor.

models attaching to a PCB model as shown in Fig.3.9(b), where the cross-section is cut as well to show the meshing details inside. The left side of the PCB is assumed to be at ambient temperature to simulate the situation where the module is plugged inside a slot of a chassis. A mixed BC is placed on the top of each package model characterized by a heat transfer coefficient λ (note that the dimension of λ is $W/\mu m^2 K$ in this section unless otherwise specified). Due to incorporation of λ in the sub-model, more complicated characteristics would be expected than those of the transistor model. The thermal ports are located on the back-side of the package model, while in the transistor model of the GaAs power amplifier the ports are located on the five different sides as already shown in Fig.3.6. There will be 256 ports needed if each connected block is defined as a single port. The port configurations and resultant number of ports are given in Table 3.6, where n_{px} and n_{py} are the number of ports defined in X and Y directions, respectively.

Number of ports(n_p)	$n_{px} \times n_{py}$
4	2×2
8	4×2
16	4×4
32	8×4
64	8×8
256	16×16

Table 3.6: Port configurations and resultant number of ports.

Memory requirement and simulation time with the number of ports

Since the PCB base model is relatively simple and attached are 16 thermally identical package models, the use of sub-models would be more beneficial in this case than the previous example of the GaAs PA die. Compared to Table 3.1, a substantial CPU memory saving is achieved in this case as shown in Table 3.7. The use of sub-models requires only a memory of about 0.4 Gbits instead of 1.4 Gbits for the original fully-linked model. The overall simulation time of using sub-models is also reduced by

a)

b)

Figure 3.9: a) C4/CBGA package model; b) 16 packages on a PCB (a cross section of the first four packages is shown).

3-16 times depending on the number of ports of the sub-model as listed in Table 3.8. It should be noted that the sub-model with 256 ports is actually equivalent to the fully-detailed model.

Models	Fully detailed model	PCB substrate model	Single package model
Blocks	341,856	13,972	20,540
CPU (Mbits)	1,419	370	391

Table 3.7: Model size and CPU requirement

n_p	Time to generate B_p and G_p (sec)	Time to simulate PCB model(sec)	Time to simulate 16 sub-models(sec)	Speed-up
4	58.1	2.9	31.7	15.7
16	67.0	6.2	33.6	13.9
64	127.4	9.0	32.3	8.7
256	386.1	20.4	35.2	3.3

Table 3.8: Simulation time with the number of ports. The simulation time of the fully detailed model is 1,459.8 sec.

In contrast to the previous example, the number of ports here plays an important role in the simulation time of the base model as seen in Table 3.8. Table 3.9 shows the number of nonzero elements in the G_b of the PCB base model after 16 sub-models of the package are attached and their effects on simulation time. It can be seen that the nonzero elements are mainly from the PCB base model up to 64 ports. However, for 256 ports, the number of added nonzero elements due to attachment is about 10 times more than the nonzero number of the original PCB base model. Therefore, a significantly longer time are required as shown in Table 3.9. The dense blocks of nonzero elements added to the base model due to attachment can be clearly seen from Fig.3.10.

Error of a sub-model with its locations

Shown in Fig.3.11(a) are the junction temperatures of packages versus their locations on the PCB with different heat transfer coefficients(λ). Those simulation results in

Figure 3.10: Sparsity of $\mathbf{G}_b$ of the PCB model after attaching 16 package sub-models.
a) 64-port sub-model; b) 256-port sub-model.

(n_p)	Num. of nonzero elements	Time to simulate base model (sec)
4	104,800	2.9
16	108,640	6.2
64	170,080	9.0
256	1,153,120	20.4

Table 3.9: Nonzero elements in the matrix $\mathbf{G}_b$ and the base model simulation time with different number of ports. The base model has 104,544 nonzero elements and the fully detailed model has nonzero elements of 2,364,256

Fig.3.11 are obtained using a 16-port sub-model with a fixed BC ($T = 300K$) applied to the left-side of the PCB and a mixed BC(λ and $T_m = 300K$) to the top-side of each package. It shows that at a small value of λ , a large temperature gradient is built up from left to right side of the package module as most of heat has to get out through the left side BC. As λ increases gradually the scenario starts to change and the temperature gradient gradually reduces, and as $\lambda > 10^{-7}(W/\mu m^2 K)$, the temperature difference between all packages becomes negligible and all of them stay at almost the same temperature. The reason is that the heat flow due to λ from the top side dominates the case at a large λ .

Fig.3.11(b) shows the temperature errors of the packages with their locations. The number of package locations is indexed from bottom to top in each column and from left to right column. The one in the most left bottom corner is given as number 1 and the most right top corner as number 16. The temperature error of each package is not quite the same because the temperature gradient under each package is different at different locations, while all packages are represented by the same sub-model of 16 ports. The errors decreasing with λ s can also be found from this figure. The reason for this decreasing would be the same as that the temperature gradient decreases as increasing λ s. The temperature contours for $\lambda = 10^{-12}$ and $\lambda = 10^{-4}$ are given in Fig.3.12(a) and (b), respectively.

Figure 3.11: a) Junction temperature with position of packages; b) Error of junction temperature with position of packages (the number of package locations is indexed from bottom to top in each column and from left to right column. The one in the most left bottom corner is number 1 and the most right top corner number 16).

a)

b)

Figure 3.12: Temperature contours of PCB from fully detailed model. a) $\lambda = 10^{-12}$;
b) $\lambda = 10^{-4}$

Error of a sub-model with the number of ports

The results shown in Fig. 3.13(a) illustrate the error as function of the number of ports. Note that the error used here is the prediction error of the maximum temperature. As will be seen that the error decrease significantly as increasing the number of ports in X direction at a small λ due to large temperature gradient. It should be pointed out that in Fig.3.13 the errors for the cases of 64 ports ($n_{px} \times n_{py} = 8 \times 8$) and 32 ports ($n_{px} \times n_{py} = 8 \times 4$) at small λ s are very close. The cases of 16 ports ($n_{px} \times n_{py} = 4 \times 4$) and 8 ports ($n_{px} \times n_{py} = 4 \times 2$) behave similarly as well. However, a substantial difference can be found between 8 ports ($n_{px} \times n_{py} = 4 \times 2$) and 4 ports ($n_{px} \times n_{py} = 2 \times 2$). It is noted that the number of ports in X direction plays a more important role than in Y direction at small λ values due to large temperature gradients in X direction. As λ increases, the error due to the presence of ports decreases as shown in Fig. 3.13(b) due to the decreasing temperature gradient. Gradually, the number of ports both in X and Y directions becomes equally important at a larger λ . As the number of ports is defined in such a way that each surface block of a sub-model is a single port, the temperature error becomes zero for the 256 port sub-model.

3.4 Summary

The creation and application of sub-models with examples of the MMIC PA and C4/CBGA package module on PCB have been successfully performed. By attaching sub-models to a base model, such as a substrate or PCB model, a mixed-model is formed and it can predict all original nodal temperatures, reduce memory requirements and speed-up simulations. Furthermore, the procedure of creating sub-models can be applied even to a mixed-model, which already include sub-models, a hierarchical sub-model can then be generated. Characterization of sub-models has been carried out as function of BCs, number of ports and port configurations, etc. It has

Figure 3.13: a) Error of junction temperature with the number of ports; b) Error of junction temperature with λ

been shown that the accuracy of using a sub-model depends on not only the number of ports but also the port configurations with respect to the temperature distributions. It is suggested that more ports are generally required to cope with a large temperature gradient in the base model in order to achieve good accuracy.

It is very important to note that the temperature error is due to introduction of ports shorting out surface temperatures.

It should be noted that applications of direct sub-models becomes time-consuming whenever BCs of a sub-model are considered as design parameters to be extensively examined since $\mathbf{B}_p$ and $\mathbf{G}_p$ must be updated with any change of BCs, which needs an LU decomposition of a large matrix G of a sub-model as seen in Eq.(3.4) and Eq.(3.5). So, it is strongly motivated to create a reduced sub-model with the multi-dimensional model reduction (MDMR) technique, which will be discussed in the following Chapter 4.

Chapter 4

Creation and Application of Reduced Sub-Models Using MDMR Technique

4.1 Introduction

As mentioned in Chapter 3, a direct sub-model encounters a difficulty as it is applied to a case where BCs are considered as design parameters which we may wish to sweep. As $\mathbf{G}_p$ and $\mathbf{B}_p$ must be formed for any given BC set and this requires performing an LU decomposition on the large matrix $\mathbf{G}$ of the direct sub-model as seen in Eqs.(3.4) and (3.5). The CPU cost of generating a sub-model prior to connection (which we refer to the specify-time as t_{sp}) to a base model for different BCs is very large and makes the use of a direct sub-model of less value when handling different BCs. A secondary disadvantage is the large model size as the LU decomposition needs to be retained to quickly determine the direct sub-model internal temperatures once the port temperatures are obtained from the base model.

These two major problems mentioned above could be resolved if the MDMR tech-

Figure 4.1: Flow chart presenting reduced sub-model to simulate a thermal system.

nique can be implemented to a direct sub-model with multiple ports. As demonstrated in Chapter 2, the application of a MDMR technique not only can reduce the size of the model equations, but also can allow for the parameters such as the external heat flows and the input power levels to be explicitly incorporated into the reduced equations, capturing the thermal behavior of the model with respect to these parameters. The application of this technique can drastically reduce the size of the system and capture the effect of BCs. The reduction then allows G_p and B_p to be quickly calculated in the reduced subspace for a specified set of BCs. The procedure is described in the flow chart of Fig.4.1 where both G_p and B_p are obtained in the reduced subspace. It is important to note that the reduced sub-model is generated only once. We refer the CPU cost of generating a reduced sub-model prior to connection to the model reduction time as t_{red} .

This chapter is organized as follows: First presented is application of the MDMR technique to direct sub-models with multiple ports, and from there to generate a re-

duced sub-model is also explained. The GaAs power amplifier and the generic package models are used as two examples to demonstrate the efficiency of this technique. The reduced sub-models will be evaluated by comparing the reduced sub-models with the corresponding direct sub-model as function of the expansion points (λ_{expn}), the number of moments (n_m) and the number of ports (n_p). Using the MDMR technique with multiple expansion points is also discussed. Finally presented is the use of the reduced sub-models by linking them to the base models. It has been demonstrated that a high accuracy and more than 200 times speeds-up can be typically achieved.

4.2 MDMR for a multiple port sub-model

In Chapter 2, the fundamentals of the theory and procedure about MDMR technique have been presented. Here as an extension of this technique, it will be incorporated to cope with a sub-model with multiple ports following the references [24]–[26], [31], [35], [48]. The situation becomes a little complicated as the moments w.r.t each port must be computed as well. The concept of block moments and block Arnoldi techniques are introduced to overcome the ill-conditioning problem [24], [25], [29]. The first step to do again is to compute the multiple port multi-dimensional subspace.

4.2.1 Computation of the multi-dimensional subspace

As already shown in Chapter 3, the multiple port thermal network has the form as Eqs(4.1) and (4.2) if both BCs and the thermal ports are explicitly described.

$$\mathbf{GT} = \mathbf{B} + \mathbf{M}_p \mathbf{T}_p \quad (4.1)$$

$$\mathbf{I}_p = \mathbf{M}_p^T \mathbf{T} \quad (4.2)$$

where:

$$\mathbf{G}(\boldsymbol{\lambda}) = \mathbf{G}_0 + \sum_{i=1}^n \lambda_i \mathbf{G}_{\lambda_i}$$

$$\mathbf{B}(\boldsymbol{\lambda}, \mathbf{T}_m, \mathbf{T}_f, \mathbf{P}) =$$

$$\mathbf{B}_0 + \sum_{i=1}^n \lambda_i \mathbf{T}_{m_i} \mathbf{B}_{\lambda_i} + \sum_{k=1}^m \mathbf{T}_{f_k} \mathbf{B}_{f_k} + \sum_{j=1}^l \mathbf{P}_j \mathbf{B}_{p_j}$$

and

$$\mathbf{G}_p = \mathbf{M}_p^T \mathbf{G}^{-1} \mathbf{M}_p \quad (4.3)$$

$$\mathbf{B}_p = \mathbf{M}_p^T \mathbf{G}^{-1} \mathbf{B} \quad (4.4)$$

Considering the system represented by Eq.(4.1) when a unit excitation is applied at the p^{th} port, where $p \in 1, 2, \dots, n_p$, Eq. (4.1) becomes [48]:

$$\mathbf{G}\mathbf{T} = \mathbf{B} + \mathbf{m}_p^p \quad (4.5)$$

$$\mathbf{I}_p = \mathbf{M}_p^T \mathbf{T} \quad (4.6)$$

where $\mathbf{m}_p^p$ is a binary vector with non-zero entry at p^{th} port, i.e. the p^{th} column of $\mathbf{M}_p$. For the p^{th} port excitation, the moments of $\mathbf{T}$ w.r.t λ_i (where the k^{th} moment of $\mathbf{T}$ w.r.t λ_i is represented by $\mathbf{m}_{\lambda_i}^{(p,k)}$) are computed using the equations:

$$\mathbf{Dm}_{\lambda_i}^{(p,0)} = \mathbf{F}^p \quad (4.7)$$

$$\mathbf{Dm}_{\lambda_i}^{(p,1)} = \mathbf{T}_{m0} \mathbf{B}_{\lambda_i} - \mathbf{G}_{\lambda_i} \mathbf{m}_{\lambda_i}^{(p,0)} \quad (4.8)$$

$$\mathbf{Dm}_{\lambda_i}^{(p,2)} = -\mathbf{G}_{\lambda_i} \mathbf{m}_{\lambda_i}^{(p,1)} \quad (4.9)$$

$\vdots$

$$\mathbf{Dm}_{\lambda_i}^{(p,q_{\lambda_i})} = -\mathbf{G}_{\lambda_i} \mathbf{m}_{\lambda_i}^{(p,q_{\lambda_i}-1)} \quad (4.10)$$

where $\mathbf{D} = \mathbf{G}_0 + \sum_{i=1}^n \lambda_{i0} \mathbf{G}_{\lambda_i}$; $\mathbf{F}^p = \mathbf{m}_p^p + \mathbf{B}_0 + \sum_{i=1}^n \lambda_{i0} T_{m0} \mathbf{B}_{\lambda_i} + \sum_{k=1}^m T_{fk0} \mathbf{B}_{f_k} + \sum_{j=1}^l P_{j0} \mathbf{B}_{p_j}$; $\lambda_{i0}, T_{m0}, T_{fk0}$ and P_{j0} are the chosen expansion points of λ_i, T_m, T_{f_k} and P_j respectively and q_{λ_i} is the number of moments to be preserved w.r.t λ_i .

As for a multiple port system, block moments (a matrix rather than a vector) are introduced and the block moment $M_{\lambda_i}^j (j \in 1, 2, \dots, q_{\lambda_i})$ of $\mathbf{T}$ w.r.t λ_i are formed as:

$$\mathbf{M}_{\lambda_i}^0 = [\mathbf{m}_{\lambda_i}^{(1,0)} \mathbf{m}_{\lambda_i}^{(2,0)} \dots \mathbf{m}_{\lambda_i}^{(p,0)}] \quad (4.11)$$

$$\mathbf{M}_{\lambda_i}^1 = [\mathbf{m}_{\lambda_i}^{(1,1)} \mathbf{m}_{\lambda_i}^{(2,1)} \dots \mathbf{m}_{\lambda_i}^{(p,1)}] \quad (4.12)$$

$\vdots$

$$\mathbf{M}_{\lambda_i}^{q_{\lambda_i}} = [\mathbf{m}_{\lambda_i}^{(1,q_{\lambda_i})} \mathbf{m}_{\lambda_i}^{(2,q_{\lambda_i})} \dots \mathbf{m}_{\lambda_i}^{(p,q_{\lambda_i})}] \quad (4.13)$$

Combining Eqs.(4.7)-(4.10) and Eqs.(4.11)-(4.13), the block moment $M_{\lambda_i}^j (j \in 1, 2, \dots, q_{\lambda_i})$ of $\mathbf{T}$ w.r.t λ_i can be evaluated from:

$$\mathbf{D} \mathbf{M}_{\lambda_i}^0 = \mathbf{M}_p \quad (4.14)$$

$$\mathbf{D} \mathbf{M}_{\lambda_i}^1 = T_{m0} [\mathbf{B}_{\lambda_i} \dots \mathbf{B}_{\lambda_i}] - \mathbf{G}_{\lambda_i} \mathbf{M}_{\lambda_i}^0 \quad (4.15)$$

$$\mathbf{D} \mathbf{M}_{\lambda_i}^2 = -\mathbf{G}_{\lambda_i} \mathbf{M}_{\lambda_i}^1 \quad (4.16)$$

$\vdots$

$$\mathbf{D} \mathbf{M}_{\lambda_i}^{q_{\lambda_i}} = -\mathbf{G}_{\lambda_i} \mathbf{M}_{\lambda_i}^{q_{\lambda_i}-1} \quad (4.17)$$

Note that the block moments $\mathbf{M}_{\lambda_i}^1 \dots \mathbf{M}_{\lambda_i}^{q_{\lambda_i}}$ follow a pattern that is amicable to use Krylov subspace techniques to form an orthogonal subspace $\mathbf{Q}_{\lambda_i}$. It is important to note that $\mathbf{M}_{\lambda_i}^j (j \in 1, 2, \dots, q_{\lambda_i})$ is the block moment and so the block Arnoldi algorithm are needed to evaluate orthogonal subspace [25], [29]. The subspace of the

moments of $\mathbf{T}$ w.r.t λ_i is given as

$$\text{colsp}(\mathbf{Q}_{\lambda_i}) = \text{colsp} \begin{pmatrix} \mathbf{M}_{\lambda_i}^1 & \mathbf{M}_{\lambda_i}^2 & \dots & \mathbf{M}_{\lambda_i}^{q_{\lambda_i}} \end{pmatrix}. \quad (4.18)$$

Note that the 0^{th} block moment w.r.t for all λ_i s is same and is denoted by $\mathbf{M}_{\lambda}^0$. It will be included in the multi-dimensional subspace later.

The moments of $\mathbf{T}$ w.r.t T_m, T_{f_k} and P_j can also be obtained using

$$\mathbf{Dm}_{T_{m_i}} = \lambda_{i0} \mathbf{B}_{\lambda_i} \quad (4.19)$$

$$\mathbf{Dm}_{T_{f_k}} = \mathbf{B}_{f_k} \quad (4.20)$$

and

$$\mathbf{Dm}_{P_j} = \mathbf{B}_{p_j} \quad (4.21)$$

respectively.

Once all the required moments are evaluated, the multi-dimensional subspace is computed such that [31], [34]

$$\text{colsp}(\mathbf{Q}) = \text{colsp}(\mathbf{K}) \quad (4.22)$$

where

$$\mathbf{K} = \begin{pmatrix} \mathbf{M}_{\lambda}^0 & \mathbf{Q}_{\lambda_1} & \dots & \mathbf{Q}_{\lambda_n} & \mathbf{m}_{T_{m_1}} & \dots & \mathbf{m}_{T_{m_n}} \\ \mathbf{m}_{T_{f_1}} & \dots & \mathbf{m}_{T_{f_m}} & \mathbf{m}_{P_1} & \dots & \mathbf{m}_{P_p} \end{pmatrix} \quad (4.23)$$

This can be achieved by using a standard QR decomposition [28].

The reduced-order equations of Eq.(4.25) can be obtained by transforming Eq.(4.24) and mapping the vector $\mathbf{T} \in \mathbb{R}^n$ of dimension n into a smaller vector $\hat{\mathbf{T}} \in \mathbb{R}^q$ of dimension q ($q \ll n$) using the orthogonal congruent matrix $\mathbf{Q}$ such that

$$\mathbf{T} = \mathbf{Q}\hat{\mathbf{T}} \quad (4.24)$$

Further, using Eq.(4.24) a congruent transformation can then be performed on Eq.(4.1) to give

$$\hat{\mathbf{G}}\hat{\mathbf{T}} = \hat{\mathbf{B}} + \hat{\mathbf{M}}_p \mathbf{T}_p \quad (4.25)$$

where:

$$\hat{\mathbf{G}}(\boldsymbol{\lambda}) = \hat{\mathbf{G}}_0 + \sum_{i=1}^n \lambda_i \hat{\mathbf{G}}_{\lambda_i}$$

$$\hat{\mathbf{B}}(\boldsymbol{\lambda}, \mathbf{T}_m, \mathbf{T}_f, \mathbf{P}) =$$

$$\hat{\mathbf{B}}_0 + \sum_{i=1}^n \lambda_i T_{m_i} \hat{\mathbf{B}}_{\lambda_i} + \sum_{k=1}^m T_{f_k} \hat{\mathbf{B}}_{f_k} + \sum_{j=1}^p P_j \hat{\mathbf{B}}_{p_j}$$

and

$$\hat{\mathbf{G}}_0 = \mathbf{Q}^T \mathbf{G}_0 \mathbf{Q}; \quad \hat{\mathbf{G}} = \mathbf{Q}^T \mathbf{G} \mathbf{Q}; \quad \hat{\mathbf{G}}_{\lambda_i} = \mathbf{Q}^T \mathbf{G}_{\lambda_i} \mathbf{Q}$$

$$\hat{\mathbf{B}} = \mathbf{Q}^T \mathbf{B}; \quad \hat{\mathbf{B}}_{\lambda_i} = \mathbf{Q}^T \mathbf{B}_{\lambda_i}$$

$$\hat{\mathbf{B}}_0 = \mathbf{Q}^T \mathbf{B}_0; \quad \hat{\mathbf{B}}_{f_k} = \mathbf{Q}^T \mathbf{B}_{f_k}; \quad \hat{\mathbf{B}}_{p_j} = \mathbf{Q}^T \mathbf{B}_{p_j} \quad (4.26)$$

Using the reduced space, the reduced sub-model described by the equation $\mathbf{I}_p = \mathbf{G}_p \mathbf{T}_p + \mathbf{B}_p$ can be obtained by calculating the attachment conduction matrix and source vector involving $\hat{\mathbf{G}}$ rather than $\mathbf{G}$;

$$\hat{\mathbf{G}}_p = \hat{\mathbf{M}}_p^T \hat{\mathbf{G}}^{-1} \hat{\mathbf{M}}_p; \quad (4.27)$$

$$\hat{\mathbf{B}}_p = \hat{\mathbf{M}}_p^T \hat{\mathbf{G}}^{-1} \hat{\mathbf{B}} \quad (4.28)$$

The primary advantage of Eq.(4.27) and Eq.(4.28) versus Eq.(4.3) and Eq.(4.4) is the great reduction in the size (many orders of magnitude) of $\hat{\mathbf{G}}$ when compared to $\mathbf{G}$. Therefore the application of model reduction technique on the large set of equations obtained from the detailed model produces a much smaller set of equations (typically a reduction from 100,000 variables to a size of 10-100 is obtained). Using MDMR the attachment size of the reduced sub-model is still, of course, $n_p \times n_p$, however, the internal size of the reduced sub-model used to generate $\hat{\mathbf{G}}_p$ and $\hat{\mathbf{B}}_p$ and obtain the internal temperatures of the reduced sub-model is much smaller.

The actual model reduction is only done once for each model and requires one LU decomposition and a number of forward-backward substitutions. We refer to this one time CPU cost as the reduction time t_{red} which is roughly equivalent to a steady-state simulation of the fully detailed model. Before attaching a reduced sub-model to a base model the BCs must be specified and a particular instance of $\hat{\mathbf{G}}_p$ and $\hat{\mathbf{B}}_p$ determined. This step involves forming and inverting the small matrix $\hat{\mathbf{G}}$ which has an insignificant CPU cost(usually less than one second). The use of multi-dimensionality is the key to the process, as from a single model reduction we can quickly generate $\hat{\mathbf{G}}_p$ and $\hat{\mathbf{B}}_p$ for a specific set of BCs. After solution of the base model plus reduced sub-model both $\mathbf{T}_b$ and $\mathbf{T}_p$ are known, $\hat{\mathbf{T}}$ can be determined using Eq.(4.25), and the previously found $\hat{\mathbf{G}}^{-1}$, which allows for all the internal nodal temperatures to be found from Eq.(4.24) by the use of the matrix $\mathbf{Q}$.

4.3 Characterization of reduced sub-models

In order to evaluate a reduced sub-model fairly and reasonably, the results of a reduced sub-model will be compared with respect to its corresponding direct sub-model by applying same port temperatures. It will exclude the error due to attachment of a multiple port sub-model to a base model, which has been discussed in Chapter 3. The

error evaluated in such a way is only attributed from model reduction process. The error in nodal temperature prediction is defined to be the difference in temperatures over the maximum temperature rise in the model:

$$\epsilon = \frac{|T_{ds} - T_{rs}|}{T_{ds} - 300} \quad (4.29)$$

Where T_{ds} is the temperature of interest from the direct sub-model and T_{rs} is the temperature of interest from the reduced sub-model.

The model of C4/CBGA package model described in Chapter 3 is employed as an example to characterize the effects of the expansion points, number of ports and number of moments used for creation of a reduced sub-model on the accuracy, memory requirement and CPU cost. It would provide useful insights how to generate a reduced sub-model with a relatively small model size, low CPU cost and memory requirement provided that the design accuracy is satisfied. Note that the dimension of λ s in this chapter is ($W/\mu m^2 K$) unless otherwise specified.

4.3.1 Expansion points

Fig.4.2(a) shows the error ϵ with λ ranged from 10^{-12} to 10^{-4} , which covers most of cooling scenarios that we practically have [18]. The 16-port reduced sub-model is generated using 3 moments at the expansion point $\lambda_{\text{expn}} = 0$. One can see that the error increases for increasing λ away from expansion point. Particularly, as $\lambda > 10^{-6}$, the error dramatically increases with λ , but the maximum is still as low as 0.016%. While the expansion point is chosen at $\lambda_{\text{expn}} = 10^{-4}$ and the results are shown in Fig.4.2(b). Again, it shows that the error increases as the evaluated λ s depart away from the expansion point $\lambda_{\text{expn}} = 10^{-4}$ by its nature. But in this case, a much larger error than the previous case is found as $\lambda < 10^{-6}$ and the maximum error reaches 0.58%. The MDMR technique also allows the use of multiple expansion points to

generate a reduced sub-model. In this case, the maximum error is expected to appear in between expansion points and it should be able to keep the error down over a wider range of λ s. In order to keep the the size(76×76) of reduced sub-model the same as at a single expansion point, 1 moment is intentionally used. It turns out that the maximum error appears in the middle ($\lambda = 10^{-6}$) of two expansion points(0 and $\lambda = 10^{-4}$) and the maximum error is as low as $2.8 \times 10^{-3}\%$ over the whole range as shown in Fig.4.2(c). The details are summarized in Table 4.1. For a wider parameter range, it seems to suggest to use two expansion points to create a reduced sub-model. As for a single expansion point it shows that $\lambda_{\text{expan}} = 0$ is definitely better than $\lambda_{\text{expan}} = 10^{-4}$.

Expansion point (λ_{expan})	0	10^{-4}	0 and 10^{-4}
Number of moments(n_m)	3	3	1
Size of model	76	76	76
Maximum error(%)	1.6×10^{-2}	5.8×10^{-1}	2.8×10^{-3}

Table 4.1: The effect of expansion points using the 16-port reduced sub-model.

Figure 4.2: Errors of 16-port reduced sub-models with λ at different expansion points. a) $\lambda_{\text{expan}} = 0$; b) $\lambda_{\text{expan}} = 10^{-4}$; (c) $\lambda_{\text{expan}} = 0$ and 10^{-4} .

4.3.2 Number of ports

Illustrated in Fig. 4.3 is the error ϵ as function of the number of ports. $n_m = 1$ and $n_m = 3$ are used for this study corresponding to Fig. 4.3(a) and (b), respectively, at the expansion point $\lambda_{\text{expn}} = 0$. It shows that the errors for 4, 8, 16, 32, 64 ports are all very small and no difference can even tell up to $\lambda = 10^{-6}$. Beyond this point, a huge error can be found in Fig. 4.3(a) for the 4-port reduced sub-model perserving only 1 moment. Fig. 4.3(a) and (b) consistently show that a reduced sub-model with more ports gives better performance in all these cases. It is believed that more information of the original model has been captured as more ports are employed producing a large model. Note that the size of the reduced model increases linearly with the number of ports as seen in Table 4.2 and consequently causes the memory requirement and CPU cost to substantially increase accordingly. It has also been shown in Chapter 3 that using a large number ports could slow down the simulation of the base model significantly. So the number of ports should be reasonably small, say 16 ports for this model.

Number of ports(n_p)	Size of reduced model	Memory size (Mbits)	CPU time (sec)
1	10	290	80
4	14	290	90
16	38	294	111
64	134	404	278
256	518	1045	1029

Table 4.2: Memory size and CPU time for model reduction with the number of ports at $n_m = 1$

4.3.3 Number of moments

It might not need to mention that the accuracy of a reduced sub-model is higher if more moments are preserved by a reduced sub-model. However, an extra moment may not necessarily increase the accuracy significantly as shown in Fig.4.4 by comparing

Figure 4.3: Errors of reduced sub-models as function of λ with different number of ports. a) $n_m = 1$; b) $n_m = 3$

the cases of 2 moments and 3 moments. , but the size of the reduced sub-models does increase linearly with number of moments as shown in Table 4.3.

n_p	Single expansion point		Two expansion points	
	$n_m = 1$	$n_m = 3$	$n_m = 1$	$n_m = 3$
4	14	28	28	56
8	22	44	44	88
16	38	76	76	152
32	70	140	140	280
64	134	268	268	536

Table 4.3: Size of reduced sub-models with the number of moments

Generally, a reduced sub-model gives a very high accuracy compared to its direct sub-model provided that the number of ports, expansion points and the number of moments are properly utilized. The accuracy of a reduced sub-model could be improved by increasing the number of ports, moments and use of multiple expansion points. From the above results, it seems that use of multi-expansion points is more effective if the size of reduced model is kept the same. But note that a reduced sub-model is going to be linked to a base model, such as PCB or substrate. The number of ports might be determined in the first place on basis of the BCs and temperature distribution of base model since the accuracy of linking a sub-model greatly depends on the number of ports and port configurations as discussed in Chapter 3.

4.4 Reduced sub-model linked to base model

As an ultimate application of a reduced sub-model, it will be linked to a base model, such as a substrate or PCB, to form a mixed-model. The procedure to link a reduced sub-model to a base model through $\hat{\mathbf{G}}_p$ and $\hat{\mathbf{B}}_p$ would be the same as shown in Chapter 3 for a direct sub-model. The solution of a reduced sub-model must be mapped back using transformation matrix $\mathbf{Q}$ to obtain all internal temperature of the original sub-model. Of course, it would be much quicker to specify $\hat{\mathbf{G}}_p$ and $\hat{\mathbf{B}}_p$

Figure 4.4: Errors of 16-port reduced sub-models with λ using different number of moments. a) $\lambda_{\text{expan}} = 0$; b) $\lambda_{\text{expan}} = 0$ and $\lambda_{\text{expan}} = 10^{-4}$

for a given BC set as it is derived in the reduced model.

This section is intended to briefly analyze the error of a mixed-model using reduced sub-models with respect to a corresponding mixed-model using direct sub-models. An error of $\hat{\mathbf{G}}_p$ and $\hat{\mathbf{B}}_p$ with respect to $\mathbf{G}_p$ and $\mathbf{B}_p$ would affect the internal temperature of the base model and port temperature, the affected port temperature in return causes an error of internal temperature of the original sub-model.

In the Section 4.3, the errors of a reduced sub-model with respect to the corresponding direct sub-model have been studied as function of expansion points, number of moments and number of ports. Shown in the following will be the errors of the mixed-model using the reduced sub-models with these parameters as compared to the mixed model using the direct sub-models. Fig.4.5 illustrates the error of the mixed-model using the reduced sub-models changes with the number of moments. By comparing Fig.4.5 with Fig.4.4, one can see that the basic shapes and relationships of them are very much similar. The errors of the mixed-model using the reduced sub-model are just slightly higher than those of the reduced sub-model itself. The same trend is also found for the errors of the mixed-model changing with the number of ports shown in Fig.4.6. It would be advised from these results that the characteristics of a mixed-model is predominately determined by those of the reduced sub-model. So one may conclude that the dominant error of using a reduced sub-model compared to the fully detailed is actually from the introduction of ports, which has been discussed in Chapter 3, as it is shown that a very little error is associated with a reduced sub-model.

Table 4.4 is the time-frame for use of a reduced sub-model linking to PCB in a simulation. From this table, one can see that the total simulation time increases a lot as a large of number of ports is used. It is very important to note that the main expense contribution is from the process of model reduction, but fortunately enough, this expense costs only once as mentioned before.

Figure 4.5: Errors of the mixed models with λ using different number of moments used for the 16-por reduced sub-model. a) $\lambda_{exp} = 0$; b) $\lambda_{exp} = 10^{-4}$; c) $\lambda_{exp} = 0$ and $\lambda_{exp} = 10^{-4}$.

Figure 4.6: Errors of the mixed models with λ using different number of ports and $n_m = 3$. a) $\lambda_{\text{exp}} = 0$; b) $\lambda_{\text{exp}} = 10^{-4}$.

n_p	Size (n_s)	To generate reduced model $t_{red}(\text{sec})$	To specify $\hat{B}_p$ and $\hat{G}_p$ $t_{sp}(\text{sec})$	To simulate base model $t_{base}(\text{sec})$	To simulate 16 sub-models $t_{sub}(\text{sec})$	Total (sec)
4	14	90	0.0020	2.89	0.38	93.3
8	22	120	0.0024	5.12	0.52	125.6
16	38	132	0.0037	6.21	0.73	138.9
32	70	168	0.0114	7.58	1.32	176.9
64	134	278	0.0734	8.99	2.29	289.3
256	518	1029	1.8348	20.37	15.92	1067.1

Table 4.4: Time required for each step as function of the number of ports using $n_m = 1$.

4.5 Speed-up of using reduced sub-models

As a direct comparison between use of a reduced sub-model and a detailed sub-model presented in Chapter 3, the models of GaAs MMIC Power amplifier and C4/CBGA package module will be used to demonstrate the efficiency and speeds-up of using reduced a sub-model. The details about the power amplifier and the C4/CBGA package module have been given in Chapter 3 where the direct sub-model of the transistor was used. Here will be presented is to create reduced sub-models with MDMR technique and link them to base models to carry out simulations. Comparison between using the reduced sub-model and the fully linked will be given in the end.

4.5.1 Example 1: GaAs MMIC power amplifier

First, a single moment $n_m = 1$ and the expansion point $\lambda_{expn} = 0$ are used as no external heat flow is associated with the transistor cell. The 9-port transistor model was reduced from size of $19,900 \times 19,900$ down to a 11×11 sized model using MDMR. Then a new hierarchical sub-model of the 12-port PA die sized $67,465 \times 67,465$ is formed by linking the GaAs substrate model with 10 reduced sub-models of the transistors. From there the new 12-port hierarchical sub-model of the PA is then reduced down to 60×60 by using MDMR again with $n_m = 3$ and $\lambda_{expn} = 0$. The whole simulation procedure of attachment is the same as using direct sub-models

Figure 4.7: Temperature contours obtained with application of the reduced sub-models of both transistor cells and PA die

shown in Chapter 3. It is completed by linking the reduced sub-model of PA die to the detailed model of the InP carrier through $\hat{\mathbf{G}}_p$ and $\hat{\mathbf{B}}_p$. Once the simulation is complete and the port temperatures and congruent transformation matrix $\mathbf{Q}$ are used to determine the temperatures of PA die first and then the internal transistor temperatures. This is done very quickly as it requires only the solution of very small reduced models and the use of the $\mathbf{Q}$ matrix. These internal temperatures can be plotted or visualized, which is not possible for a compact model of resistor network. Fig. 4.7 present contour plots of the thermal distribution both in the InP carrier and the amplifier. No difference can be identified between Fig.4.7 and Fig.3.7 since the error between the direct sub-model and the corresponding reduced sub-model is so small($< 0.1\%$).

In Table 4.5 a summary of the model size and reduction time for the sub-models used in the PA is presented. It should be noted that all times presented are for CPU time and are quoted for an unloaded SUNBlade 100 workstation with 6 Gig of RAM. The reduction took 46 sec for the transistor sub-model($19,900 \times 19,900$) to reduce down to a 11×11 using MDMR, which is comparable to the time to solve the detailed

transistor model. Generation of the reduced sub-model of PA die of size 60×60 from $67,456 \times 67,456$ took 233 seconds. The size of the base model after attachment are also shown but is simply determined by the number of ports.

Shown in Table 4.6 are the model sizes, solution times(t_{sol} is defined as the CPU times to obtain all nodal temperatures except one-time reduction time) and junction temperatures for the entire system. As we can see, simulation time of the detailed model is large (512.4 sec). Using the 10 transistor reduced sub-models to build a PA die sub-model a reduction to 3,094 nodes was achieved and which was simulated in 0.43 sec. The solution times of the PA die reduced sub-model and the 10 transistor reduced sub-models are 0.27 and 0.32 sec, respectively. So a speed up of 512 times is obtained. Note that the solution time includes all procedure times needed in order to get all node temperatures of the original model.

4.5.2 Example 2: C4/CBGA packages on PCB

It takes 168 sec for the 16-port package model to reduce down to size 76×76 from size $20,504 \times 20,504$ as 3 moments and $\lambda_{expan} = 0$ are used. The simulation is completed by linking 16 reduced sub-models to the PCB base model to form a mixed-model of $14,048 \times 14,048$. Solution of the mixed-model(takes 6.2 sec) using the reduced sub-models gives all node temperatures of the base model and the port temperatures. It takes another 1.3 sec to solve the 16 reduced sub-models by applying the port temperatures and mapping back to the original model. Compared to the fully linked, one can see that size reduction of 300 times and a speed-up of 195 times have been achieved as seen in Table 4.5 and Table 4.6. The junction temperature rise is obtained as a fixed BC is placed on the left side of PCB base model and a mixed BC with $\lambda = 10^{-8}$. Note that the error of predicting the junction temperatures is from the introduction of multiple ports(here 16 ports were used), not from the use of a reduced sub-model.

Model	Detailed Size	Sub-model size		t_{red} (sec)
		Reduced	Attached	
PA				
Transistor	19,900	11×11	9×9	46
Die ¹	67,465	60×60	12×12	233
C4/CBGA				
Package	20,504	76×76	16×16	168

Table 4.5: Model size and reduced sub-model generation time for each component in PA and C4/CBGA package model. ¹ Die model is formed out of a die substrate and 10 transistor reduced sub-models

Model	Size	t_{sol} (sec)	T_j (K)
PA			
Fully detailed	269,547	512.4	105.4
InP+Die direct sub-model	3,094	175	104.3
InP+Die reduced sub-model	3,094	1.0	104.3
	Speed-up of 512		$\Delta T_j = 1.0\%$
C4/CBGA			
Fully detailed	341,856	1,459.8	44.6
PCB+package direct sub-model	14,048	105	43.8
PCB+package reduced sub-model	14,048	7.5	43.8
	Speed-up of 195		$\Delta T_j = 2.0\%$

Table 4.6: Solution time and junction temperature obtained with fully detailed solution and reduced sub-model application.

4.6 Summary

The MDMR technique is successfully extended to a sub-model with multiple ports to form a multi-dimensional reduced sub-model. The characteristics of such a reduced sub-model as function of expansion point, number of ports and number of moments have been carried out. The reduced sub-model represented by a small set of linear equations is able to capture the behavior of a large set of equations of the original detailed thermal model. Size reduction achieved by using MDMR is shown to be three to four orders of magnitude, yet is able to predict nodal temperatures in the original model with high accuracy even for a wide boundary conditions. By attaching the reduced sub-models into a substrate or PCB base model, a high level of thermal analysis in electronic modules was performed. A speed-up of 195 and 512 times has been shown. In addition, the model reduction procedure was applied to a mixed-model which already included reduced sub-models to generate a hierarchical reduced sub-model.

Chapter 5

Creation and Application of Macro-models Using MDMR Technique

5.1 Introduction

Thermal *compact models* are commonly used to simplify and speed up thermal analysis. Existing methods to produce a compact model have been mainly based on the optimization of resistive network models [49], [50]. A variety of increasingly complicated networks have been proposed to improve the accuracy of network based compact models. These models have produced a significant improvement in the predictive capability of the compact model with a typical accuracy 5 to 10% over a wide range of boundary conditions. Typically, these models have one or two thermal ports and a number of external heat flows. Recent work [13] has indicated, however that models with more ports may be desirable when large thermal gradients are present in the base model.

As shown in Chapter 4, a reduced sub-model created using MDMR technique is suf-

ficiently small as expressed in Eqs.(4.24)-(4.28) so that with the addition of two more equations (one defining the temperature at each point of interest and a second defining the heat flows off the external surfaces), a multi-dimensional macro-model(MDMM) can be created and these equations are:

$$\mathbf{T}_{jct} = \mathbf{S}\mathbf{Q}\hat{\mathbf{T}} \quad (5.1)$$

$$I_{\lambda_i} = \lambda_i \mathbf{B}_{\lambda_i} \cdot (\mathbf{Q}\hat{\mathbf{T}} - T_{m_i}) \quad (5.2)$$

where $\mathbf{T}_{jct}$ is a vector whose elements are the nodal temperatures of interest(generally, junction temperatures) and $\mathbf{S}$ is a matrix that selects these nodes from the solution. I_{λ_i} are the individual heat flows off each external surface. As will be shown, MDMR is well suited to produce BC independent thermal macro-models.

As a comparison, the steps involved in an optimized network model are also presented at the very bottom branch in Fig.5.1. Such a procedure includes a costly reduction step (the optimization of the resistor network) which requires many steady-state simulations of the detailed model. However, it should be mentioned, the network model requires no generation of matrix $\mathbf{G}_p$ and vector $\mathbf{B}_p$ for a specific BC set, but does not produces all temperatures of the model.

The chapter proceeds as follows: Firstly presented is the evaluation of BCI of the MDMMs created using MDMR from the models of a power amplifier and a generic multi-chip module ball grid array model(MCMBGA). Secondly the application of the macro-models linked to a substrate and PCB will be performed in Section 5.3. Finally, the advantages of MDMMs will be given in terms of fast creation, high accuracy, flexibility and capability of predicting multiple junction temperatures.

Figure 5.1: Flow chart presenting macro-model using MDMR to simulate a thermal system.

5.2 BCI characteristics of macro-models

As a primary requirement of applications, a compact model or macro-model of an electronic component must be of boundary condition independence (BCI) such that it can be applicable to any user-specific simulations in a package and system level. In the following section, BCI characteristic of macro-models created by MDMR technique will be examined at a wide variety of BCs similar to those presented in [50] and [51] using two examples from diverse electronic components, i.e. integrated GaAs power amplifier and MCMBGA. The MDMR technique presents a new way to create a macro-model or even compact model (really depending on the definition).

5.2.1 Example 1: GaAs power amplifier

In order to correctly determine the internal temperatures of the GaAs power amplifier it needs to be modelled in considerable detail. The detailed model of GaAs amplifier is shown in Fig. 5.2. Figs. 5.2(a) and (b) show views of the detailed model (42,969

blocks) created with *Atar*. The amplifier consists of four 32 μm long transistors formed in a mesa structure on a 1mm square GaAs chip 100 μm thick. Included in the model are six 60 μm gold bumps for attachment to a substrate, and all emitter, base and collector metallization. The metallization includes local heat spreaders and thru-vias to the back-side. Two macro-models with different number of ports were investigated. Each macro-model used the six interconnect bumps to form the interface with the base model. In the first macro-model these were designated to be a single port with a common surface temperature, whereas in the second macro-model six thermal ports (one for each bump) were defined. The design parameters in each model are transistor power levels and the heat flow from the back-side of the chip (which of course would be determined by the packaging and heat sink configuration). Two parameters capture the heat flow off the back-side of the amplifier: the first λ ($\text{W}/\mu\text{m}^2\text{K}$) represents the thermal resistance to the ambient and the second, the ambient temperature T_m (K).

An important feature of model reduction is the ability to predict all nodal temperatures in the model – and from these to extract temperatures for the nodes of interest. The procedure follows three steps: 1) for a given set of port temperatures, solve for the reduced variables $\hat{\mathbf{T}}$; 2) from these use the transformation matrix $\mathbf{Q}$ to obtain the temperatures of the detailed model; 3) finally use the nodal selector vector to extract the temperature value at the junction or any other required node. To illustrate, nodal temperatures were determined for the single port MDMM with an assigned port temperature $T_p = 300$ K. Based on these temperatures, thermal contours for the amplifier model are plotted in Fig. 5.2(c) and (d), where the entire amplifier and a cross-section view is shown. The contours show that heat from the source is flowing in two directions: toward the external surface with mixed BC and to the bumps (port) with fixed temperature. The contribution of high conductivity metal lines and through-via on the temperature distribution can also be seen.

Figure 5.2: *Atar* models showing physical structure and mesh. a) Perspective view of the full amplifier; b) Cross-sectional view of the transistors; c) Perspective view of thermal contours; d) Cross-sectional view of thermal contours.

To evaluate the use of the two macro-models (one and six port versions), simulation results from detailed model and the MDMM, under the same sets of BC, were compared. BC sets were chosen such as to test the models over extreme conditions. Simulation results for the junction temperature and prediction error for different port temperatures temperature as a function of the back-side heat flow are shown in Fig. 5.3. The heat flow off the back-side was varied from essentially zero to a large value that enforces a fixed temperature of 300 K on the back-side of the amplifier. In Figure 5.3(a) the results for the single port model are presented and show that the thermal behavior of the model is captured well by the MDMM. At a small λ , that is no heat flow off the back-side, the junction temperature is linearly related to the port temperature, while at a larger λ , the junction temperature becomes primarily determined by the ambient temperature imposed on the surface where mixed BC is applied, with this surface area being much larger than the port area (bumps). In addition, the prediction error of one port MDMM is shown to be very small, with the maximum value 0.025%. This indicates that the MDMR process performed on the detailed model does not introduce any significant error. The increase in the error value as λ approaches 10^{-4} ($\text{W}/\mu\text{m}^2\text{K}$) is due to the selection of the expansion point (a MDMR parameter) being set to $\lambda = 0$. A general characteristic of the method is that the error becomes zero at the expansion point and increases as we move away, which has been discussed in Chapter 4.

A similar evaluation is performed for the six port MDMM and results are shown in Fig. 5.3(b). Port temperature sets assigned to the model are summarized in Tab. 5.1. Comparing the two macro-models, a smaller prediction error is produced from six-port MDMM. However, the higher accuracy is associated with larger macro-model size, and hence, slightly increased simulation time. The final decision regarding the trade-off of time versus accuracy depends on the user requirements. These results clearly show that both macro-models (one and six-port) are able to capture the thermal behavior

Figure 5.3: Junction temperatures of GaAs PA macro-model as function of λ and port temperatures T_p .

of the detailed model accurately for a wide variety of boundary conditions.

Set #	Port Temperatures (K)					
	T_{p1}	T_{p2}	T_{p3}	T_{p4}	T_{p5}	T_{p6}
1	300	300	300	300	300	300
2	310	300	300	300	300	300
3	310	310	300	300	300	300
4	310	300	310	300	300	300
5	300	300	310	310	300	300
6	310	310	310	310	310	310

Table 5.1: Sets of port temperatures for the six port MDMM

5.2.2 Example 2: MCMBGA

The second example presents a generic MCMBGA model. Multi-chip packages are used frequently to increase the level of integration placing dies with different functions in a single package. They are particularly useful for the integration of incompatible IC technologies, offering advantages for applications when electrical performance and board density is at a premium. The MCMBGA modelled has two die placed on a single package substrate. One die is wire bonded and the other mounted using flip chip attachment. In Fig. 5.4(a). a view of 1/4 of an *Atar* model is shown along with a schematic inset. Die in the upper package (D1) is $8 \times 8 \times 0.266$ mm in size with a nominal 1.25 W power dissipation. D1 is connected to the substrate carrier with flip chip technology where the interconnection between the die and carrier is made through conductive bumps (grid pitch of $300 \mu\text{m}$ and quadratic bumps $200 \mu\text{m}$) placed directly on the die surface. Epoxy underfill fills the area between the die and the carrier, surrounding the solder bumps. Die in the lower package (D2) is $10 \times 10 \times 0.130$ mm in extent, dissipates 2.5 W and is attached to the opposite side of substrate carrier. This die is then wire bonded to the carrier. The substrate carrier then provides the connection from the die to the exterior of the package. The substrate carrier is epoxy/glass laminate $35 \times 35 \times 1.226$ mm in size with a package-to-

board interconnection on the bottom side with 352 solder balls laid out on a square grid (Fig. 5.4(b)) with a constant 1.27 mm pitch [52]. Solder ball composition is 63Sn/37Pb. The package substrate has four conductive layers.

The configuration of the applied BCs is shown in the inset in Fig. 5.4(a). Each BC set includes two mixed BCs representing the heat flow off the top surfaces of D1 and the substrate carrier (characterized by λ_1 and λ_2) modelling the package exposure to air or the effect of a heat sink. Solder balls at the bottom of the package (package-to-board contact surface) are considered as being attached to the PCB model and represented in the macro-models by either a single port or four equally sized ports. A detailed numerical model of MCMBGA consisted of 73,129 blocks. The size of the one-port MDMM was 84 and the size of the four-port model 126.

The reduction procedure for single and four-port produced MDMMs of size 84 and 126, with generation time 290 and 450 sec respectively. After the MDMM is created, the time required to generate $\hat{\mathbf{G}}_p$ and $\hat{\mathbf{B}}_p$ for a specific BC set is respectively 0.24 and 0.37 sec, essentially insignificant from a users perspective.

As to compare with the procedure of a resistor based compact model generation, a detailed solution of this model takes 320 sec and hence, assuming 38 BC sets are to be considered, the time required to generate all the data for the optimization of the resistor network would take a minimum of 3h 23 min. As to conclude the comparison, the reduction procedure described above reduces the compact model generation time by 26 times.

Typically, the evaluation of MDMMs of a complex large input/output (IO) package involves the determination of the error in temperature prediction for a large set of applied BCs. A similar procedure was adapted to evaluate the one and four-port macro-models, varying the BCs (temperatures and heat flows) for surfaces over a large range from 0 to 10^9 (W/m²K). Each set of BCs applied to the macro-model includes: λ_1 , λ_2 representing heat flows off the top surfaces and heat flow through the

Figure 5.4: MCMBGA package model. a) View of structure; b) Bottom view showing solder pads with 1.27 mm pitch.

Figure 5.5: a) Junction temperatures at each of 40 BC sets; b) Error associated with each set

thermal port(s) on the bottom of the package which we will specify as λ_p . The error in temperature prediction is evaluated with respect to the solution of the detailed model. Table 5.2 summarizes 40 (out of 99) BC sets from [18] that were chosen and applied to the detailed model, the one and four-port MDMs. Simulation results for each set of BCs applied into the detailed model and four-port macro-model are shown in the Fig. 5.5. This comparison confirms the very good predictive ability of the macro-model, with all errors being less than $4 \times 10^{-3}\%$, and very little error being introduced by the reduction procedure over a wide range of BCs.

The results in this section clearly indicate the macro-models created from MDMR technique are of boundary condition independence with a very high accuracy over an extreme BC range. Moreover, they have an ability to predict multiple critical temperatures and a general flexibility that allows for multiple thermal ports and external heat flows to be incorporated.

Index of BCs	λ_1 W/m^2K	λ_2 W/m^2K	λ_p W/m^2K	T_p K
1	5	1	100	330
2	5	10	1000	303
3	5	50	1000	303
4	5	50	3000	301
5	5	100	1500	302
6	15	10	1000	303
7	15	50	3000	301
8	15	100	500	305
9	30	5	500	305
10	30	30	3000	301
11	30	200	10000	300
12	80	5	500	305
13	80	30	2500	301
14	80	500	500	303
15	10	75	1000	302.5
16	50	75	1000	302
17	50	75	3500	300.7
18	100	75	500	304
19	150	5	600	303
20	150	30	3500	300.6
21	150	200	10000	300.15
22	150	200	2000	300.8
23	500	5	500	302
24	500	30	4000	300.3
25	500	200	10000	300.1
26	500	200	2500	300.4
27	10	1000	500	302.5
28	10	10000	500	300.7
29	1000	10	50000	300
30	10000	10	50000	300
31	10000	10000	10000	300
32	5000	5000	5000	300
33	1000	1000	1000	300.3
34	500	500	500	301.2
35	50000	50000	50000	300
36	10000	10000	5	300.14
37	10	10000	5	306
38	10000	10	2	300.62
39	1	1	10000	300.3
40	10^9	10^9	10^9	300

Table 5.2: 40 sets of boundary conditions for MCMBGA

5.3 Applications of macro-models

As a primary requirement, the BCI characteristics of a macro-model created by MDMR has been verified in the previous Section 5.2. The focus in this section will be on the applications of such macro-models over a large BC range. The two issues to be addressed in the evaluation of the macro-models will be the accuracy of the models and the efficiency of the technique to perform thermal analysis of complex systems. The accuracy of the macro-models are evaluated with respect to the numerical solution of a fully detailed model for given sets of BC's. Two macro-models which have been created in Section 5.2 will be used for this study, i.e. GaAs power amplifier attached to a substrate and a generic (MCMBGA) attached to a PCB.

Example 1: Macro-model from GaAs power amplifier

As an application of MDMM, A PA macro-models is used in a system which is comprised of two identical GaAs power amplifiers attached to a ceramic substrate. A fixed temperature boundary condition was placed on the left side of the substrate. A number of different models of the system were created, in order to evaluate the macro-model, address the computational cost of creating and using it, and investigate the impact of the port configuration. First, a full detailed model of all three components was created and simulated with *Atar*. Secondly, two models were built in which the GaAs power amplifiers are represented by either one or six port MDMMs with the substrate model being unchanged. Finally, a model of the substrate only (no amplifiers attached) was simulated. This allows for the determination of the computational cost of using the macro-models.

In Fig. 5.6(a) the model for the substrate and two amplifiers is shown (one amplifier is depicted in full and the other in outline). Two macro-models were attached to a base model and simulation on the system was performed. The resulting temperature contours are shown in Fig. 5.6(b) where the internal temperatures of macro-model

have been determined from the port temperatures and the reduced system of equations.

Table 5.3 summarizes various computational costs of creating and simulating all four models described in the previous paragraph. Compared to detailed model of the system, the use of the either macro-model produces a much smaller model size, and hence shorter simulation time. The one time model reduction cost t_{red} is roughly equivalent to a single solution of the detailed model of the system. The time to specify either MDMM for a particular set of BC's prior to simulation is inconsequential.

In addition, Table 5.3 summarizes the maximum (junction) temperature for each of the two GaAs PAs and the substrate. Results show the use of the one-port MDMM does introduce a small error in the junction temperatures. However, as proved in the previous section, this error is not due to the use of the reduction process, but due to the one-port configuration that is imposing a constant temperature at the contact surface. This effect is also manifest in the large difference of the maximum substrate temperature obtained with one-port MDMM. In contrary, the six port MDMM produces very good results with essentially no error in the amplifier or substrate.

The effect of MDMM port configuration on the temperature distribution and the heat flow in the substrate is illustrated in Fig. 5.6c and d. In Fig. 5.6c it can be seen that the one port MDMM forces all six contact points (interconnect bumps) of the amplifier to be at a single temperature. The single port model is therefore unable to represent the real temperature distribution at substrate-MDMM contact surface. The six port configuration gives a more realistic situation and was found to have very little error with respect to the full model. Finally, Fig. 5.7 shows a comparison of the maximum temperature in the system as function of the back-side heat flow for both the detailed model and the six-port MDMM. For the comparison two cases were simulated whereby a fixed BC was placed on the left side of the substrate of either 300 K and 320 K. Comparison shows that the junction temperatures predicted with

a)

Substrate and two amplifiers

b)

Thermal contours in substrate and two MDMM's

c)

Thermal contours in substrate for single port MDMM

d)

Thermal contours in substrate for six port MDMM

Figure 5.6: *Atar* models and temperature contours

Model	t_{red}	t_{sp} (sec)	t_{sol}	Model size		T_{j1}	T_{j2} (K)	T_{sub}
System (detailed)	N/A	N/A	118	42,969	89,398	92.62	92.73	1.47
System (with MDMM)								
One-port	127	0.04	1.77	40	4,342	91.63	92.0	0.77
Six-port	230	0.13	2.09	90	4,352	92.64	92.80	1.47
Substrate (det.)	N/A	N/A	2.20	N/A	4,340	N/A	N/A	

Table 5.3: Computational costs and junction temperatures for the two GaAs PA models attached to a substrate MDMM vs. detailed solution.

the six-port MDMM are very accurate over the entire range of λ .

Figure 5.7: Junction temperatures of GaAs macro-model as function of λ for left-side fixed temperature in the substrate.

Example 2: Macro-model from MCMBGA model

The evaluation of MCMBGA macro-model as part of a thermal system was completed by attaching it to a PCB and simulating it to predict junction temperatures for each component of the system. The PCB model is of a multi-layer laminate substrate based on FR-4 technology. The model includes four layers of FR-4, two signal planes 30 μm thick with 30 % Cu content and two ground planes 40 μm . Four different models were simulated: a complete detailed model of the system, a one-port MDMM

attached to a PCB detailed model, a four-port MDMM attached to PCB detailed model and finally the PCB detailed model only (with no component attached).

Fig.5.8(a) shows the temperature contours of the package and the PCB obtained from the use of a single port MDMM. In this case the simulation of the PCB and macro-model was undertaken, the port temperature obtained and then used to determine all of the package nodal temperatures. Figures 5.8(b)-(d) show the effect of the port configuration on the PCB temperatures, with Fig. 5.8(b) presenting the contours from a detailed solution of both the package and PCB. As can be seen there is a significant variation in the surface temperature under the package, which can be contrasted to the subsequent figures for the one and four-port MDMMs respectively. In these cases the imposition of a uniform surface temperature on the ports has distorted the contour configuration in the substrate. It should be noted that the degree of distortion depends on the number of ports, which in general is smaller than the number of bumps in the detailed model. Because of this, the contour distortion is produced by the use of any macro-model, without being a consequence of the model reduction procedure.

Solution time, model size and simulated junction temperatures for each model are shown In Table 5.4. It can be seen that the use of either MDMM produced a 100 times speed up in the simulation of the entire system (package MDMM attached to PCB). A junction temperature was defined for each package component. Specifically, at the center of the the package carrier T_{j0} , the center of the bottom die T_{j1} , and the center of top die T_{j2} . The small error of 2% in predicted junction temperature, as mentioned above, is associated with the effect of the port configuration and not with the use of model reduction.

Finally, Fig. 5.9 shows the variation of junction temperature with λ_1 and/or λ_2 as predicted by four-port MDMM attached to the PCB. Compared to detailed model, MDMM captures the heat flow effect very accurately.

Figure 5.8: Temperature contours of the MCMBGA-PCB model obtained with MDMR technique for fixed temperature at right side. a) Package and PCB from fully detailed model; b) PCB including port area from fully detailed model; c) PCB only for single port MDMM; d) PCB only for 4 port MDMM.

Model	t_{sol} (sec)	Model size		T_{j0}	T_{j1} (K)	T_{j2}
System (detailed)	213	73129	77928	43.9	95.5	91.5
System (with MDMM)						
One-port	2.20	84	4801	39.5	92.7	88.7
Four-port	2.20	126	4804	41	93.1	89.2
PCB (detailed)	2.78	N/A	4800	N/A	N/A	N/A

Table 5.4: Model size, simulation times and junction temperatures for each component of MCM models

Figure 5.9: Junction temperatures of MCMBGA model as function of λ . a) Variation of λ_1 for $\lambda_2 = 0$; b) Variation of λ_2 for $\lambda_1 = 0$.

5.4 Summary

Generation of a MDMM using MDMR technique is at least an order of magnitude faster than the creation of an optimized network compact model. All MDMMs presented here were verified to be BCI over an extremely wide BC range with a high accuracy of less than 0.1%. From the applications of such MDMMs by attaching them to a substrate or PCB, high efficiency and flexibility of MDMMs have been indicated that MDMM can predict multiple critical temperatures and allows for multiple thermal ports and external heat flows to be incorporated. Furthermore, use of MDMMs can significantly reduce CPU cost and a speed-up of around 100 times are achieved.

Chapter 6

Global Method of Linking Sub-Models

6.1 Introduction and Motivations

A more general method of forming mixed models, called *global method*, is presented in this chapter to link sub-models with a base model by mathematically forming global model equations. The major motivations of this method can be summarized as follows: (1) The boundary conditions of both sub-models and base models are explicitly expressed, so the global model basically is BCI. Moreover, it allows to apply MDMR to reduce the global model to a much smaller size. In this case the global method can handle BCs variations both from the base model and the sub-models in a reduced space and thus great benefits will be achieved. (2) It becomes feasible to do *hierarchical* model reduction of the global model(i.e. a detailed base model plus previously reduced sub-models) to form a macro-model. The solution from this macro-model can be mapped back to the original global model first by using transformation matrix $\mathbf{Q}_2$ and then mapped to the original problem by using $\mathbf{Q}_1$ to get all internal temperature of sub-models. (3) It is imperative to globally

incorporate a base model and sub-models for a transient simulation (but this is not presented in this thesis).

Note that it will be of less practical value unless the sub-models are reduced because solution of the global equation formed from the detailed base model and direct sub-models is CPU intensive, and more or less like a fully detailed solution.

This chapter is organized as follows: First presented is the formation of global model equations with a single reduced sub-model. Then this method is extended to any number of reduced sub-models linking to a base model. Secondly, the solution of the global model is verified using an example of the C4/CBGA package module. In the end, hierarchical MDMR of the global model is presented as well.

6.2 Global model with a single sub-model

To demonstrate the principle and procedure of forming a global model, a simple case is considered first when only one sub-model is linked to the base model. The theory will be extended to the case of multiple sub-models later. Note that in the following mathematical expressions, the subscript "b" is denoted as base model, "s" as sub model and "g" as global model.

As already presented in Chapter 3, the basic equations for a base model are shown as:

$$\mathbf{G}_b \mathbf{T}_b - \mathbf{D}^T \mathbf{I}_p = \mathbf{B}_b \quad (6.1a)$$

$$\mathbf{T}_p = \mathbf{D} \mathbf{T}_b \quad (6.1b)$$

and for a reduced sub-model they are:

$$\hat{\mathbf{G}}_s \hat{\mathbf{T}}_s = \hat{\mathbf{B}}_s + \hat{\mathbf{M}}_p \mathbf{T}_p \quad (6.2a)$$

$$\mathbf{I}_p = \hat{\mathbf{M}}_p^T \hat{\mathbf{T}}_s \quad (6.2b)$$

$$\mathbf{T}_s = \mathbf{Q}\hat{\mathbf{T}}_s \quad (6.2c)$$

By combining Eq.(6.1) and Eq.(6.2), one can obtain the global equation as

$$\mathbf{G}_g \mathbf{T}_g = \mathbf{B}_g \quad (6.3)$$

where

$$\mathbf{G}_g = \begin{pmatrix} \mathbf{G}_b & \mathbf{0} & -\mathbf{D}^T & \mathbf{0} \\ -\mathbf{D} & \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & -\hat{\mathbf{M}}_p^T \\ \mathbf{0} & -\hat{\mathbf{M}}_p & \mathbf{0} & \hat{\mathbf{G}}_s \end{pmatrix}; \quad \mathbf{T}_g = \begin{pmatrix} \mathbf{T}_b \\ \mathbf{T}_p \\ \mathbf{I}_p \\ \hat{\mathbf{T}}_s \end{pmatrix}; \quad \mathbf{B}_g = \begin{pmatrix} \mathbf{B}_b \\ \mathbf{0} \\ \mathbf{0} \\ \hat{\mathbf{B}}_s \end{pmatrix} \quad (6.4)$$

Solving Eq.(6.3) will give all solutions of $\mathbf{T}_b$, $\mathbf{T}_p$, $\mathbf{I}_p$ and $\hat{\mathbf{T}}_s$ in one step. The procedure seems to be simpler than the attachment method described in Chapter 4 and Chapter 5. From Eqs.(6.4), one may notice that the size of global model is $n = n_b + n_s + 2 \times n_p$, which is even bigger than the fully linked model $n = n_b + n_s$ while the method of attaching a sub-model produces a model of size $n = n_b + n_p$ (usually $\ll n_b + n_s$).

To overcome the size problem of the global model, use of a reduced sub-model(n_s becomes very small) is probably the only practical method.

As we will see, the most important advantage of global method is that all boundary conditions of the base model and the sub model can be explicitly incorporated. As discussed in Chapter 2, the base model matrices with explicit boundary conditions are written:

$$\mathbf{G}_b(\boldsymbol{\lambda}) = \mathbf{G}_{b_0} + \sum_{i=1}^{n_1} \lambda_i \mathbf{G}_{b_{\lambda_i}} \quad (6.5a)$$

$$\mathbf{B}_b(\boldsymbol{\lambda}, \mathbf{T}_m, \mathbf{T}_f, \mathbf{P}) = \mathbf{B}_{b_0} + \sum_{i=1}^{n_1} \lambda_i \mathbf{T}_{m_i} \mathbf{B}_{b_{\lambda_i}} + \sum_{k=1}^{m_1} \mathbf{T}_{f_k} \mathbf{B}_{b_{f_k}} + \sum_{j=1}^{l_1} \mathbf{P}_j \mathbf{B}_{b_{p_j}} \quad (6.5b)$$

and for the reduced sub-model are in reduced space:

$$\hat{\mathbf{G}}_s(\boldsymbol{\lambda}) = \hat{\mathbf{G}}_{s_0} + \sum_{i=1}^{n_2} \lambda_i \hat{\mathbf{G}}_{s_{\lambda_i}} \quad (6.6a)$$

$$\hat{\mathbf{B}}_s(\boldsymbol{\lambda}, \mathbf{T}_m, \mathbf{T}_f, \mathbf{P}) = \hat{\mathbf{B}}_{s_0} + \sum_{i=1}^{n_2} \lambda_i T_{m_i} \hat{\mathbf{B}}_{s_{\lambda_i}} + \sum_{k=1}^{m_2} T_{f_k} \hat{\mathbf{B}}_{s_{f_k}} + \sum_{j=1}^{l_2} P_j \hat{\mathbf{B}}_{s_{p_j}} \quad (6.6b)$$

where n_1 and n_2 are the number of mixed boundary conditions in the base model and the sub model, respectively. m_1 and m_2 the number of fixed boundary conditions in the base model and the sub model, respectively. l_1 and l_2 are the number of power sources in the base model and the sub model, respectively.

By combining Eqs. (6.4), (6.5) and (6.6), the global equations with explicit boundary conditions can be obtained as:

$$\mathbf{G}_g(\boldsymbol{\lambda}) = \mathbf{G}_{g_0} + \sum_{i=1}^{n_1+n_2} \lambda_i \mathbf{G}_{g_{\lambda_i}} \quad (6.7a)$$

$$\mathbf{B}_g(\boldsymbol{\lambda}, \mathbf{T}_m, \mathbf{T}_f, \mathbf{P}) = \mathbf{B}_{g_0} + \sum_{i=1}^{n_1+n_2} \lambda_i T_{m_i} \mathbf{B}_{g_{\lambda_i}} + \sum_{k=1}^{m_1+m_2} T_{f_k} \mathbf{B}_{g_{f_k}} + \sum_{j=1}^{l_1+l_2} P_j \mathbf{B}_{g_{p_j}} \quad (6.7b)$$

where

$$\mathbf{G}_{g_0} = \begin{pmatrix} \mathbf{G}_{b_0} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \hat{\mathbf{G}}_{s_0} \end{pmatrix}; \quad \mathbf{B}_{g_0} = \begin{pmatrix} \mathbf{B}_{b_0} \\ 0 \\ 0 \\ \hat{\mathbf{B}}_{s_0} \end{pmatrix} \quad (6.8)$$

With respect to BCs of the base model, i.e., $i \in \{1, 2, \dots, n_1\}$, $j \in \{1, 2, \dots, l_1\}$

and $k \in \{1, 2, \dots, m_1\}$, the respective matrices and vectors are:

$$\mathbf{G}_{g_{\lambda_i}} = \begin{pmatrix} \mathbf{G}_{b_{\lambda_i}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}; \quad \mathbf{B}_{g_{\lambda_i}} = \begin{pmatrix} \mathbf{B}_{b_{\lambda_i}} \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad (6.9)$$

$$\mathbf{B}_{g_{f_k}} = \begin{pmatrix} \mathbf{B}_{b_{f_k}} \\ 0 \\ 0 \\ 0 \end{pmatrix}; \quad \mathbf{B}_{g_{P_j}} = \begin{pmatrix} \mathbf{B}_{b_{P_j}} \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad (6.10)$$

With respect to BCs of the sub model, i.e., $i \in \{n_1 + 1, n_1 + 2, \dots, n_1 + n_2\}$, $j \in \{l_1 + 1, l_1 + 2, \dots, l_1 + l_2\}$ and $k \in \{m_1 + 1, m_1 + 2, \dots, m_1 + m_2\}$, the respective matrices and vectors are:

$$\mathbf{G}_{g_{\lambda_i}} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \hat{\mathbf{G}}_{s_{\lambda_i}} \end{pmatrix}; \quad \mathbf{B}_{g_{\lambda_i}} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \hat{\mathbf{B}}_{s_{\lambda_i}} \end{pmatrix} \quad (6.11)$$

$$\mathbf{B}_{g_{f_k}} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \hat{\mathbf{B}}_{s_{f_k}} \end{pmatrix}; \quad \mathbf{B}_{g_{P_j}} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \hat{\mathbf{B}}_{s_{P_j}} \end{pmatrix} \quad (6.12)$$

6.3 Global model with multiple sub-models

Eqs. (6.3) and (6.4) can be easily extended to the case when multiple sub-models are involved and linked to the base model. In this case, the basic relationships of the

base model can be extended from Eqs. (6.1) dealing with a single sub model to the following with multiple sub-models.

$$\mathbf{G}_b \mathbf{T}_b - \sum_{i=1}^m \mathbf{D}_i^T \mathbf{I}_{p_i} = \mathbf{B}_b \quad (6.13a)$$

$$\mathbf{T}_{p_i} = \mathbf{D}_i \mathbf{T}_b \quad (6.13b)$$

where $i \in \{1, 2, \dots, m\}$ in the Eq.(6.13b) and m is the number of sub-models. The basic equations for each sub model are similar to Eqs.(6.2) and they are:

$$\hat{\mathbf{G}}_{s_i} \hat{\mathbf{T}}_{s_i} = \hat{\mathbf{B}}_{s_i} + \sum_{i=1}^m \hat{\mathbf{M}}_{p_i} \mathbf{T}_{p_i} \quad (6.14a)$$

$$\mathbf{I}_{p_i} = \hat{\mathbf{M}}_{p_i}^T \mathbf{T}_{p_i} \quad (6.14b)$$

$$\mathbf{T}_{s_i} = \mathbf{Q}_i \hat{\mathbf{T}}_{s_i} \quad (6.14c)$$

where $\hat{\mathbf{G}}_{s_i}, \mathbf{T}_{s_i}, \hat{\mathbf{T}}_{s_i}, \hat{\mathbf{B}}_{s_i}, \mathbf{I}_{p_i}, \mathbf{T}_{p_i}, \mathbf{D}_i, \hat{\mathbf{M}}_{p_i}$ and n_{p_i} denote the respective matrices and vectors corresponding to the i^{th} reduced sub-model ($i \in \{1, 2, \dots, m\}$). By combining Eqs.(6.13) and (6.14), the global equation with multiple sub-models can be obtained the same way as Eq.(6.3), where $\mathbf{G}_g, \mathbf{T}_g$ and $\mathbf{B}_b$ are more complicated than shown in Eq.(6.4), which are:

$$\mathbf{T}_g = (\mathbf{T}_b \ \mathbf{T}_{p_1} \ \mathbf{T}_{p_2} \ \dots \ \mathbf{T}_{p_m} \ \mathbf{I}_{p_1} \ \mathbf{I}_{p_2} \ \dots \ \mathbf{I}_{p_m} \ \hat{\mathbf{T}}_{s_1} \ \hat{\mathbf{T}}_{s_2} \ \dots \ \hat{\mathbf{T}}_{s_m})^T \quad (6.15)$$

$$\mathbf{B}_g = (\mathbf{B}_b \ 0 \ 0 \ \dots \ 0 \ 0 \ 0 \ \dots \ 0 \ \hat{\mathbf{B}}_{s_1} \ \hat{\mathbf{B}}_{s_2} \ \dots \ \hat{\mathbf{B}}_{s_m})^T \quad (6.16)$$

and

$$G_g = \begin{pmatrix} \mathbf{G}_b & 0 & \dots & 0 & -\mathbf{D}_1^T & \dots & -\mathbf{D}_m^T & 0 & \dots & 0 \\ -\mathbf{D}_1 & \mathbf{I} & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -\mathbf{D}_m & 0 & \dots & \mathbf{I} & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & \mathbf{I} & \dots & 0 & -\hat{\mathbf{M}}_{p_1}^T & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 0 & \dots & \mathbf{I} & 0 & \dots & -\hat{\mathbf{M}}_{p_m}^T \\ 0 & -\hat{\mathbf{M}}_{p_1} & \dots & 0 & 0 & \dots & 0 & \hat{\mathbf{G}}_{s_1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -\hat{\mathbf{M}}_{p_m} & 0 & \dots & 0 & 0 & \dots & \hat{\mathbf{G}}_{s_m} \end{pmatrix} \quad (6.17)$$

From Eq.(6.15) or Eq.(6.17), the size of the global model can be obtained by $n = n_b + \sum_{i=1}^m n_{s_i} + 2 \times m \times \sum_{i=1}^m n_{p_i}$. It indicates that the global method dealing with multiple sub-models could make the size of global model extremely large. Again, MDMR is strongly motivated to reduce the size of sub-models n_s and then make the global method to be practically useful.

6.4 Verification of global model solutions

The C4/CBGA package module is used again as an example to demonstrate the application of the global method. The simulation results from the global method and the attachment method described in Chapter 4 have been verified to be exactly *identical*.

Listed in Table 6.1 are the solution time (t_{sol}) and memory requirement as function of the number of ports and the size of the reduced sub-model. The simulation time dramatically increases with number of ports starting at 32 ports($n_s=76$). The added

nonzero elements are believed to be the reason for slowing down the simulation as discussed in Section 3.3.2 because the matrices of reduced models are dense.

Number of ports (n_p)	Size of reduced sub-model (n_s)	Memory requirement (Mbits)	Solution time (t_{sol}) (sec)
4	28	121	11.6
8	44	151	12.4
16	76	154	13.8
32	140	222	25.4
64	268	360	55.1

Table 6.1: Global model solution times and memory requirements with the number of ports and the reduced mode sizes

Shown in Table 6.2 is the comparison of the solution times between the global models and the corresponding attachment models as function of the number of ports and the size of reduced sub-models. It consistently shows that the solution times of the global models are longer than the corresponding attachment models. It can be understood by the fact as seen in Section 6.3 that the sizes of global models are larger than the attachment models and the added nonzero elements of global models are also more than the attachment models.

It should be noted from Table 6.2 that the differences of the solution times are actually not substantial if a small number of ports ($n_p < 32$) and a small size of the reduced sub-model ($n_s < 70$) are used due to the following facts: (1) the size and nonzero elements of the global model and the attachment model are all dominated by the base model; and (2) the difference between the number n_s and n_p are relatively small. So the solution times of the global model and the attachment model are mainly determined by the base model in this case.

However, as the number of ports further increases, the added nonzero elements will become dominant and the size of the global model will increase significantly as well, particularly for the case of multiple sub-models. Moreover, the added nonzero elements of the global model (proportional to n_s^2) will be substantially larger than the attachment model (proportional to n_p^2). So the simulations of the global models are

slowed down significantly compared to the attachment method as seen in Table 6.2.

n_p	(n_s)	Global method	Attachment method		
		$t_{sol}(\text{sec})$	$t_{base}(\text{sec})$	$t_{sub}(\text{sec})$	$t_{sol} = t_{base} + t_{sub}(\text{sec})$
4	14	6.1	2.9	0.38	3.3
8	22	6.3	5.1	0.52	5.6
16	38	7.4	6.2	0.73	6.9
32	70	10.6	7.6	1.32	8.9
64	134	33.9	9.0	2.29	11.3
256	518	2,652.9	20.4	15.92	36.3

Table 6.2: Comparison of simulation times between the global method and the attachment method.

6.5 Hierarchical MDMR using global method

As one may notice that the base models of the examples presented so far are all detailed models. Simulation of a base model could be CPU intensive if the base model is complex enough, such as GaAs PA substrate model, or multiple sub-models with a large number of ports are attached to a base model, such as C4/CBGA package model with 256 ports. Use of the global method can overcome this obstacle by using hierarchical MDMR. As shown previously the global model shown in Eq.(6.7) are formed with explicit BCs both for the sub-models and the base models. The evaluation of BCs can be done in reduced subspace. The MDMR technique presented in Chapter 2 can be directly applied to Eq.(6.7) to form the reduced global model as.

$$\hat{\mathbf{G}}_g \hat{\mathbf{T}}_g = \hat{\mathbf{B}}_g \quad (6.18)$$

where

$$\begin{aligned} \hat{\mathbf{G}}_g &= \mathbf{Q}_2^T \mathbf{G}_g \mathbf{Q}_2; \quad \hat{\mathbf{T}}_g = \mathbf{Q}_2^T \mathbf{T}_g; \quad \hat{\mathbf{B}}_g = \mathbf{Q}_2^T \mathbf{B}_g \\ \hat{\mathbf{G}}_{g_{\lambda_i}} &= \mathbf{Q}_2^T \mathbf{G}_{g_{\lambda_i}} \mathbf{Q}_2; \quad \hat{\mathbf{B}}_{g_{\lambda_i}} = \mathbf{Q}_2^T \mathbf{B}_{g_{\lambda_i}}; \quad \hat{\mathbf{B}}_{g_{P_j}} = \mathbf{Q}_2^T \mathbf{B}_{g_{P_j}} \end{aligned} \quad (6.19)$$

Shown in Table 6.3 are the size and accuracy of the reduced global model with number of moments used for hierarchical MDMR at the expansion point $\lambda_{\text{expansion}} = 0$. The reduced sub-model of package with a size of 76×76 (transformation matrix $\mathbf{Q}_1$, $n_m = 3$ and $n_p = 16$) was used for this demonstration. It can be seen that the junction temperature error of the *reduced global model* is less than 1% compared to the *original global model* if more than 10 moments are employed and the size of reduced global model is still relatively small. The solution time of the reduced global

Number of moments(n_m)	Size of reduced global model	Error(%)
5	13×13	9.5
10	18×18	1.1
15	23×23	0.1

Table 6.3: Size and accuracy of reduced global model as function of number of moments at the expansion point of $\lambda_{\text{expansion}} = 0$. The reduced sub-model of package of (76×76) is used.

model of 23×23 (transformation matrix $\mathbf{Q}_2$ and $n_m = 15$) is only 0.013 (sec). To map back to the nodes of the original global model using $\mathbf{Q}_2$ takes 0.124 (sec) and mapping back to all original nodes of 16 reduced sub-models using $\mathbf{Q}_1$ needs 1.28 (sec). Compared to the same case using the original global method (13.78 sec) as shown in Table 6.1, a speed-up of 10 times has been achieved with an error as low as 0.1% by Using MDMR to the global model.

6.6 Summary

The global method has been successfully proposed and mathematically implemented to link sub-models to a base model. The solutions of a global model were verified to be identical to those of the corresponding mixed model presented in Chapter 4. Hierarchical MDMR of a global model has been demonstrated with a size reduction down to 23×23 , a speed-up of 10 times and an error as low as 0.1%.

Chapter 7

Use of Macro-Models with Uniform Heat Flow Ports Rather Than Uniform Temperature Ports

7.1 Introduction

Use of sub- or macro-models created using MDMR has shown great advantages in terms of speeding-up simulation and reducing memory requirement without yet compromising accuracy as presented in Chapter 4 and Chapter 5. As it is known that in a board or system level simulation, a macro-model or a compact model has to be linked to a *base model*, such as substrate or PCB. A very important issue, but not yet fully investigated, arises with regard to how to use a macro-model or a compact model to predict the junction temperatures with a minimum error [14]. It should be pointed out that the commonly used method of linking a multiple port compact/macro-model with a base model [13], [14] and used in the previous chapters is to assume the temperatures of the linking nodes of the base model within a port to be at the port temperature. However, it has been revealed that this implementation may introduce

a large error when large temperature gradients are present in the substrate or PCB unless a large number of ports are employed. However, use of more ports will produce a longer simulation time and a larger memory requirement as indicated before.

In this chapter, a general method of linking a macro-model to a base model is derived and it turns out that the previously used method, i.e. using *uniform temperature ports*, is a special case of this general method. From this general method, the idea of *uniform heat flow ports* are then introduced to overcome the difficulty of using uniform temperature ports when a large temperature gradient in the substrate or PCB is present. A uniform heat flow port allows the temperatures of the linking nodes of the base model to vary with the temperature gradient while the heat flow density within a defined port area is assumed to be uniform. The results show that the prediction of junction temperature using uniform heat flow ports is more accurate than using uniform temperature ports in most circumstances.

This chapter is organized as follows. First an overview of a macro-model characterized by T_p and I_p is given in Section 7.2. Presented in Section 7.3 is the general mathematical implementation of linking a macro-model to a base model. From there the uniform temperature ports and uniform heat flow ports are then presented and discussed respectively in Section 7.4. The results and analysis based on the generic C4/CBGA package mounted on PCB are given in Section 7.5 under a variety of BCs and PCB effective thermal conductivities. The comparison between the results from uniform temperature ports and uniform heat flow ports is given as well. Finally presented in Section 7.6 are the conclusions.

7.2 Multiple port macro-model characterized by $\mathbf{T}_p$ and $\mathbf{I}_p$

As presented in Chapter 3 and Chapter 4 the basic relationships for a macro-model with multiple ports are given by:

$$\mathbf{GT} = \mathbf{B} + \mathbf{M}_p \mathbf{T}_p \quad (7.1a)$$

$$\mathbf{I}_p = \mathbf{M}_p^T \mathbf{T} \quad (7.1b)$$

$$\mathbf{I}_p = \mathbf{B}_p + \mathbf{G}_p \mathbf{T}_p \quad (7.1c)$$

$$\mathbf{G}_p = \mathbf{M}_p^T \mathbf{G}^{-1} \mathbf{M}_p \quad (7.1d)$$

$$\mathbf{B}_p = \mathbf{M}_p^T \mathbf{G}^{-1} \mathbf{B} \quad (7.1e)$$

where $\mathbf{G}$ and $\mathbf{B}$ are the thermal conductance matrix and source vector of the macro-model; $\mathbf{M}_p$ is a selector matrix that maps the port temperatures into the nodal space of the macro-model; $\mathbf{I}_p$ and $\mathbf{T}_p$ denote the port heat flow and port temperature respectively. These equations may be either in a detailed space or a reduced space using MDMR, but the analysis and discussion given below will be general.

Fig.7.1 schematically illustrates how the sub-model use all the links shown in Fig.7.1(a) are replaced by ports as shown in Fig.7.1(b). A macro-model with multiple ports can generally be characterized by two parameters, i.e. port temperature $\mathbf{T}_p$ and heat flow off the port $\mathbf{I}_p$, analogous to the port voltages and port currents respectively in an interconnected electrical network as shown in Fig.7.1(c). Illustrated in Fig.7.1(d) is a macro-model linking to a base model. Depending on local meshing, a port may include one or more linking nodes of a base model as shown in Fig.7.1(b).

Figure 7.1: Macro-model linking to a base model. a) Two models linked node by node; b) Two models linked by two ports; c) Macro-model representation; d) Macro-model linked to a base model.

7.3 Generalized method of linking a macro-model to a base model

As mentioned that the previous method of linking a compact/macro-model to a base model was to assume all the linking nodes of the base model within a port to be uniform at the port temperature. Here a method will be presented which allows a general format to deal with the relationship between all the linking nodes of the base model and the defined ports of the macro-model.

The mathematical implementation starts with the basic relationship between port temperature and port heat flow of a macro-model given in Eq.(7.1c) as

$$I_p = B_p + G_p T_p \quad (7.2)$$

The original base model equation without attaching any sub-models can be generally expressed as follows.

$$G_b T_b = B_b \quad (7.3)$$

where G_b is thermal conductance matrix between all blocks in the base model. T_b is all the block temperature vector and B_b the source vector from the heat generations and the related BCs. The basic thermal resistor networks are schematically illustrated in Fig. 7.2. As can be seen that the fully detailed model does not introduce any linking nodes since they are connected all through block-to-block. However, as a based model is linked to a macro-model through ports, linking nodes of the base model are introduced to use as an interface in order to link the *abstract* macro-model characterized by G_p and B_p . The temperature of these *linking nodes* are designated as T_l . The blocks which provide these linking nodes are specifically designated as *attachment blocks* and their temperatures as T_a . For convenience, the index numbers of the attachment blocks in the base model are assumed to be at the end of the

Figure 7.2: Illustration of thermal networks via multiple ports. a) Fully detailed model; b) Uniform temperature ports; c) General ports

T vector such that Eq.(7.3) can re-written in a form that the temperatures of the attachment blocks are explicitly shown as

$$\begin{pmatrix} \mathbf{G}_{b_{11}} & \mathbf{G}_{b_{12}} \\ \mathbf{G}_{b_{21}} & \mathbf{G}_a \end{pmatrix} \begin{pmatrix} \mathbf{T}_{b_1} \\ \mathbf{T}_a \end{pmatrix} = \begin{pmatrix} \mathbf{B}_{b_1} \\ \mathbf{B}_a \end{pmatrix} \quad (7.4)$$

The corresponding sub-matrices and sub-vectors related to the original $\mathbf{G}_b$, $\mathbf{T}_b$ and $\mathbf{B}_b$ are given as

$$\mathbf{G}_b = \begin{pmatrix} \mathbf{G}_{b_{11}} & \mathbf{G}_{b_{12}} \\ \mathbf{G}_{b_{21}} & \mathbf{G}_a \end{pmatrix}; \quad \mathbf{T}_b = \begin{pmatrix} \mathbf{T}_{b_1} \\ \mathbf{T}_a \end{pmatrix}; \quad \mathbf{B}_b = \begin{pmatrix} \mathbf{B}_{b_1} \\ \mathbf{B}_a \end{pmatrix} \quad (7.5)$$

Depending on the number of the defined ports of the macro-model relative to the number of the linking nodes of the base model, the implementation of linking the macro-model to the base model will be slightly different and thus three cases as shown in Fig.7.3 are separately discussed as follows.

7.3.1 Number of linking nodes more than number of ports

For the case that the base model of the substrate or PCB is needed to be meshed at high resolution, the number of linking nodes in the base model would be much more than the number of ports as shown in Fig.7.3(a). As a result, more than one linking node will be included within a port. To deal the temperatures and heat flows of these linking nodes with the their corresponding port quantities ($\mathbf{T}_p$ and $\mathbf{I}_p$) becomes the key issue to be addressed, i.e. how to determine the port temperature from the temperatures of the underlying linking nodes and how to distribute the port heat flow into the linking nodes.

Assuming that n_{l_i} are the linking nodes located underneath the i^{th} port, the total linking nodes n_l will be $n_l = \sum_{i=1}^{n_p} n_{l_i}$, where n_p is the number of ports. Once the

Figure 7.3: a) Number of linking nodes more than number of ports; b) Number of linking nodes equal to number of ports; c) Number of linking nodes less than number of ports

temperatures of the linking nodes are known, the port temperature can be determined from the defined relationship between the temperatures of the linking nodes of the base model and the port temperatures such as [53]:

$$\begin{pmatrix} \mathbf{T}_{p_1} \\ \mathbf{T}_{p_2} \\ \vdots \\ \mathbf{T}_{p_{n_p}} \end{pmatrix} = \mathbf{W}_T \begin{pmatrix} \mathbf{T}_{l_{11}} \\ \vdots \\ \mathbf{T}_{l_{1n_{l_1}}} \\ \mathbf{T}_{l_{21}} \\ \vdots \\ \mathbf{T}_{l_{2n_{l_2}}} \\ \vdots \\ \mathbf{T}_{l_{n_p1}} \\ \vdots \\ \mathbf{T}_{l_{n_p n_{l_{n_p}}}} \end{pmatrix} \quad (7.6)$$

$\mathbf{W}_T \in \mathbb{R}^{n_p \times n_l}$ is a weighted-factor matrix that defines the relationship between the linking node temperatures and the port temperatures. $\mathbf{W}_T$ has a form as

$$\mathbf{W}_T = \begin{pmatrix} ft_{11} & \dots & ft_{1n_{l_1}} & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & ft_{21} & \dots & ft_{2n_{l_2}} & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & ft_{n_p1} & \dots & ft_{n_p n_{l_{n_p}}} \end{pmatrix} \quad (7.7)$$

Note that there actually is no constraint to the coefficients of ft_{ij} in the matrix $\mathbf{W}_T$ as far as they are physically reasonable or they can give better prediction of junction temperatures.

The distributions of the port heat flows into the linking nodes of the base model

can be similarly defined by the relationship [53]

$$\begin{pmatrix} \mathbf{I}_{l_{11}} \\ \vdots \\ \mathbf{I}_{l_{1n_{l_1}}} \\ \mathbf{I}_{l_{21}} \\ \vdots \\ \mathbf{I}_{l_{2n_{l_2}}} \\ \vdots \\ \mathbf{I}_{l_{np1}} \\ \vdots \\ \mathbf{I}_{l_{npn_{l_{np}}}} \end{pmatrix} = \mathbf{W}_I \begin{pmatrix} \mathbf{I}_{p_1} \\ \mathbf{I}_{p_2} \\ \vdots \\ \mathbf{I}_{p_{n_p}} \end{pmatrix} \quad (7.8)$$

where $\mathbf{W}_I \in \mathbb{R}^{n_l \times n_p}$ is a weighted-factor matrix that shows how to distribute the heat flows into the base model. The matrix $\mathbf{W}_I$ has a form as follows

$$\mathbf{W}_I = \begin{pmatrix} fi_{11} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ fi_{1n_{l_1}} & 0 & \dots & 0 \\ 0 & fi_{21} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & fi_{2n_{l_2}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & fi_{np1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & fi_{npn_{l_{np}}} \end{pmatrix} \quad (7.9)$$

Note that the constraint for the coefficients in matrix $\mathbf{W}_I$ is that the sum of each column must be equal to 1 to conserve the heat flow of each port. i.e. $\sum_{j=1}^{n_{l_i}} fi_{ij} = 1$,

$i = 1, 2, \dots, n_p$. If the coefficients are the area ratio of the linking nodes they will satisfy this constrain by construction.

From Fig.7.2(c) the following relationship regarding the heat flows to the linking nodes can be found

$$I_l = \mathbf{G}_l(\mathbf{T}_l - \mathbf{T}_a) \quad (7.10)$$

where $\mathbf{G}_l$ is the thermal conductance matrix between the attachment blocks to their corresponding linking nodes. As there is no cross resistance between the linking nodes, we have $\mathbf{G}_l \in \mathbb{R}^{n_l \times n_l}$ which is diagonal as below.

$$\mathbf{G}_l = \begin{pmatrix} \mathbf{G}_{l_{11}} & \dots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \mathbf{G}_{l_{1n_{l_1}}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathbf{0} & \mathbf{0} & \mathbf{G}_{l_{21}} & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \mathbf{G}_{l_{2n_{l_1}}} & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{G}_{l_{npn_{l_1}}} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{G}_{l_{npn_{l_{np}}}} \end{pmatrix} \quad (7.11)$$

Substituting Eq.(7.6) into Eq.(7.2) and then into Eq.(7.8), Eq.(7.10) can be re-written as follows

$$\mathbf{W}_I(\mathbf{G}_p \mathbf{W}_T \mathbf{T}_l + \mathbf{B}_p) = \mathbf{G}_l(\mathbf{T}_l - \mathbf{T}_a) \quad (7.12)$$

Re-arrange Eq.(7.12) and one can obtain

$$(\mathbf{W}_I \mathbf{G}_p \mathbf{W}_T - \mathbf{G}_l) \mathbf{T}_l + \mathbf{G}_l \mathbf{T}_a = -\mathbf{W}_I \mathbf{B}_p \quad (7.13)$$

From Fig.7.2(c) the following relationship about the heat flows to the attachment blocks can be found

$$I_a = G_l(T_l - T_a) \quad (7.14)$$

With the addition of the heat flow from the sub-model, the original base model equation (7.4) becomes

$$\begin{pmatrix} G_{b_{11}} & G_{b_{12}} \\ G_{b_{21}} & G_a \end{pmatrix} \begin{pmatrix} T_{b_1} \\ T_a \end{pmatrix} = \begin{pmatrix} B_{b_1} \\ B_a \end{pmatrix} + \begin{pmatrix} 0 \\ -I_a \end{pmatrix} \quad (7.15)$$

Substitute Eq.(7.14) into Eq.(7.15) and then re-arrange Eq.(7.15) in another form as

$$\begin{pmatrix} G_{b_{11}} & G_{b_{12}} & 0 \\ G_{b_{21}} & G_a - G_l & G_l \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} T_{b_1} \\ T_a \\ T_l \end{pmatrix} = \begin{pmatrix} B_{b_1} \\ B_a \\ 0 \end{pmatrix} \quad (7.16)$$

The base model *including the linking nodes* after attaching a macro-model can be obtain by combining Eq.(7.13) and Eq.(7.16) as given below

$$\begin{pmatrix} G_{b_{11}} & G_{b_{12}} & 0 \\ G_{b_{21}} & G_a - G_l & G_l \\ 0 & G_l & W_I G_p W_T - G_l \end{pmatrix} \begin{pmatrix} T_{b_1} \\ T_a \\ T_l \end{pmatrix} = \begin{pmatrix} B_{b_1} \\ B_a \\ -W_I B_p \end{pmatrix} \quad (7.17)$$

The base model equation including the linking nodes after attaching a macro-model can be written in a more general way as

$$\tilde{G}_b \tilde{T}_b = \tilde{B}_b \quad (7.18)$$

where

$$\tilde{\mathbf{G}}_b = \begin{pmatrix} \mathbf{G}_{b_{11}} & \mathbf{G}_{b_{12}} & 0 \\ \mathbf{G}_{b_{21}} & \mathbf{G}_a - \mathbf{G}_l & \mathbf{G}_l \\ 0 & \mathbf{G}_l & \mathbf{W}_I \mathbf{G}_p \mathbf{W}_T - \mathbf{G}_l \end{pmatrix}; \quad \tilde{\mathbf{T}}_b = \begin{pmatrix} \mathbf{T}_{b_1} \\ \mathbf{T}_a \\ \mathbf{T}_l \end{pmatrix}; \quad \tilde{\mathbf{B}}_b = \begin{pmatrix} \mathbf{B}_{b_1} \\ \mathbf{B}_a \\ -\mathbf{W}_I \mathbf{B}_p \end{pmatrix} \quad (7.19)$$

The simulation is completed by solving Eq.(7.18) first to obtain all block temperatures and the surface nodal temperatures of the base model. Eq.(7.6) will give the port temperatures after having the linking node temperatures. Applying the port temperatures back to Eq.(7.1a), all internal temperatures of the macro-model can be determined. If Eq.(7.1a) is in the reduced subspace, then they must be mapped back to the original space by using transformation matrix $\mathbf{Q}$.

7.3.2 Number of linking nodes equal to number of ports

If every linking node of the base model matches exactly the corresponding port as shown in Fig.7.3(b), there exists only one option of linking the linking nodes to the ports as the following equations.

$$\mathbf{T}_p = \mathbf{T}_l \quad (7.20)$$

and

$$\mathbf{I}_l = \mathbf{I}_p \quad (7.21)$$

which implies $\mathbf{W}_T$ and $\mathbf{W}_I$ shown in Eq.(7.7) and Eq.(7.9) all become identity matrices. Then Eq.(7.17) in this case is simplified as

$$\begin{pmatrix} \mathbf{G}_{b_{11}} & \mathbf{G}_{b_{12}} & 0 \\ \mathbf{G}_{b_{21}} & \mathbf{G}_a - \mathbf{G}_l & \mathbf{G}_l \\ 0 & \mathbf{G}_l & \mathbf{G}_p - \mathbf{G}_l \end{pmatrix} \begin{pmatrix} \mathbf{T}_{b_1} \\ \mathbf{T}_a \\ \mathbf{T}_p \end{pmatrix} = \begin{pmatrix} \mathbf{B}_{b_1} \\ \mathbf{B}_a \\ -\mathbf{B}_p \end{pmatrix} \quad (7.22)$$

7.3.3 Number of linking nodes less than number of ports

In a case that the number of linking nodes are less than the number of ports which may occur as the base model is coarsely meshed and a large number of ports are used in the macro-model as shown in Fig.7.3(c). To deal with this situation, we will create new linking nodes in order to match up the number of ports. Therefore, it is very similar to the case discussed in Section 7.3.2. However, the difference should be noted that the new created linking nodes have to share the attachment blocks with the existing linking nodes, i.e. $n_a < n_l$ where n_a and n_l are the numbers of attachment blocks and linking nodes, respectively. The heat flows to the linking nodes of the base model are given by

$$\mathbf{I}_l = \mathbf{G}_l(\mathbf{T}_p - \mathbf{W}_a \mathbf{T}_a) \quad (7.23)$$

where $\mathbf{W}_a \in \Re^{n_p \times n_a}$ is a matrix which indicates the connection relationship between the attachment node and the linking nodes and it can be given as:

$$\mathbf{W}_a = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \quad (7.24)$$

The heat flow into each attachment node is the sum of all heat flows of the linking

nodes which connect the same attachment nodes and it is

$$\mathbf{I}_a = \mathbf{W}_a^T \mathbf{I}_l = \mathbf{W}_a^T [\mathbf{G}_l (\mathbf{T}_p - \mathbf{W}_a \mathbf{T}_a)] \quad (7.25)$$

The original base model after adding the heat flows to the attachment nodes as in Eq.(7.16) becomes

$$\begin{pmatrix} \mathbf{G}_{b_{11}} & \mathbf{G}_{b_{12}} & 0 \\ \mathbf{G}_{b_{21}} & \mathbf{G}_a - \mathbf{W}_a^T \mathbf{G}_l \mathbf{W}_a & \mathbf{W}_a^T \mathbf{G}_l \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{T}_{b_1} \\ \mathbf{T}_a \\ \mathbf{T}_p \end{pmatrix} = \begin{pmatrix} \mathbf{B}_{b_1} \\ \mathbf{B}_a \\ 0 \end{pmatrix} \quad (7.26)$$

Re-arranging Eq.(7.23), the following relationship can be obtained

$$\mathbf{G}_l \mathbf{W}_a \mathbf{T}_a + (\mathbf{G}_p - \mathbf{G}_l) \mathbf{T}_p = -\mathbf{B}_p \quad (7.27)$$

Therefore, the base model including the linking nodes after attaching a macro-model using uniform temperature ports can be obtain by combining Eq.(7.26) and Eq.(7.27) as given

$$\begin{pmatrix} \mathbf{G}_{b_{11}} & \mathbf{G}_{b_{12}} & 0 \\ \mathbf{G}_{b_{21}} & \mathbf{G}_a - \mathbf{W}_a^T \mathbf{G}_l \mathbf{W}_a & \mathbf{W}_a^T \mathbf{G}_l \\ 0 & \mathbf{G}_l \mathbf{W}_a & \mathbf{G}_p + \mathbf{G}_l \end{pmatrix} \begin{pmatrix} \mathbf{T}_{b_1} \\ \mathbf{T}_a \\ \mathbf{T}_p \end{pmatrix} = \begin{pmatrix} \mathbf{B}_{b_1} \\ \mathbf{B}_a \\ -\mathbf{B}_p \end{pmatrix} \quad (7.28)$$

7.4 Uniform temperature and uniform heat flow ports

From the above analysis, if the number of linking nodes is equal or less than the number of ports, there is nothing to do with the relationship between linking nodes and ports since it is basically an one-to-one relationship. Practically, it is not a commonly seen scenario unless a large number of ports is used. The following will

deal with the more common situation where the number of linking nodes is much more than the number of ports.

7.4.1 Uniform temperature ports

As shown in Fig. 7.2 (b) for uniform temperature ports, the linking nodes within a port in the base model are forced to be the same temperature, i.e. port temperature. In this case, all the linking nodes within a port can be represented by one linking node and as there are only n_p independent linking temperatures, we can set

$$\mathbf{T}_l = \mathbf{T}_p \quad (7.29)$$

So the columns of $\mathbf{W}_T$ presented in Eq.(7.7) corresponding to the linking nodes within the same port will emerge to one column and hence $\mathbf{W}_T$ becomes an identity matrix($n_p \times n_p$) if uniform temperature ports are employed.

As a macro-model is linked to the base model, the heat flows to the attachment blocks of the base model can be determined by Eq.(7.30)

$$\mathbf{I}_a = \mathbf{G}_l(\mathbf{W}_p \mathbf{T}_p - \mathbf{T}_a) \quad (7.30)$$

where $\mathbf{W}_p \in \mathbb{R}^{n_a \times n_p}$ is a matrix which indicates the connection relationship between

the ports and the linking nodes and it can be expressed as:

$$\mathbf{W}_p = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} \quad (7.31)$$

The original base model after adding the heat flows to the attachment nodes as in Eq.(7.30) becomes

$$\begin{pmatrix} \mathbf{G}_{b_{11}} & \mathbf{G}_{b_{12}} & 0 \\ \mathbf{G}_{b_{21}} & \mathbf{G}_a - \mathbf{G}_l & \mathbf{G}_l \mathbf{W}_p \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \mathbf{T}_{b_1} \\ \mathbf{T}_a \\ \mathbf{T}_p \end{pmatrix} = \begin{pmatrix} \mathbf{B}_{b_1} \\ \mathbf{B}_a \\ 0 \end{pmatrix} \quad (7.32)$$

The total heat flow to the attachment blocks within a port should be equal to the port heat flow. Mathematically it is

$$\mathbf{I}_p = \mathbf{W}_p^T \mathbf{I}_a \quad (7.33)$$

Substituting Eq.(7.2) and Eq.(7.30) into Eq.(7.33), the following equation can be obtained

$$\mathbf{G}_p \mathbf{T}_p + \mathbf{B}_p = \mathbf{W}_p^T \mathbf{G}_l (\mathbf{W}_p \mathbf{T}_p - \mathbf{T}_a) \quad (7.34)$$

and thus

$$(\mathbf{G}_p - \mathbf{W}_p^T \mathbf{G}_l \mathbf{W}_p) \mathbf{T}_p + \mathbf{W}_p^T \mathbf{G}_l \mathbf{T}_a = -\mathbf{B}_p \quad (7.35)$$

Therefore, the base model *including the linking nodes* after attaching a macro-model using uniform temperature ports can be obtain by combining Eq.(7.32) and Eq.(7.35) as given

$$\begin{pmatrix} \mathbf{G}_{b11} & \mathbf{G}_{b12} & \mathbf{0} \\ \mathbf{G}_{b21} & \mathbf{G}_a - \mathbf{G}_l & \mathbf{G}_l \mathbf{W}_p \\ \mathbf{0} & \mathbf{W}_p^T \mathbf{G}_l & \mathbf{G}_p - \mathbf{W}_p^T \mathbf{G}_l \mathbf{W}_p \end{pmatrix} \begin{pmatrix} \mathbf{T}_{b1} \\ \mathbf{T}_a \\ \mathbf{T}_p \end{pmatrix} = \begin{pmatrix} \mathbf{B}_{b1} \\ \mathbf{B}_a \\ -\mathbf{B}_p \end{pmatrix} \quad (7.36)$$

As one will see that there is a dense matrix ($n_p \times n_p$) in the right bottom corner. As the number of ports increases, this dense matrix will greatly slow down the solution of Eq.(7.36). Figs. 7.4(a) and (c) show the sparsity of matrix $\tilde{\mathbf{G}}_b$ as a single and 16 uniform temperature ports are used, respectively. In the right bottom corner, a dense matrix with 1×1 or 16×16 can be seen, respectively.

7.4.2 Uniform heat flow ports

From Section 7.3, uniform heat flow ports can be derived. The schematic of attaching a macro-model to a substrate or PCB through uniform heat flow ports is shown in Fig.7.2(c). Each linking node is connected to the port through a current source, the magnitude of the current source is reasonably assumed to be proportional to the linking node area. Use of uniform heat flow ports allows the temperatures of the linking nodes to vary to cope with a situation with a large temperature gradient. The port temperature is assumed to be a weighted average of the temperatures of the linking nodes within the port depending on the area ratio of the linking nodes. So

the $\mathbf{W}_T$ in this case is

$$\mathbf{W}_T = \begin{pmatrix} w_{11} & \dots & w_{1n_{l_1}} & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & w_{21} & \dots & w_{2n_{l_2}} & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 & \dots & w_{n_p 1} & \dots & w_{n_p n_{l_{n_p}}} \end{pmatrix} \quad (7.37)$$

where

$$w_{ij} = A_{ij}/A_i \quad j = 1, 2, \dots, n_{l_i}; \quad i = 1, 2, \dots, n_p \quad (7.38)$$

where w_{ij} is the weighted factor for the linking node j within the port i , and A_i is the port area of the port i . $\sum_{j=1}^{n_{l_i}} w_{ij} = 1$, $i = 1, 2, \dots, n_p$.

The distribution of heat flows to the attachment blocks of the base model is assumed also to be proportional to the area ratio of the surface nodes within a port. Thus $\mathbf{W}_I$ can be obtained as

$$\mathbf{W}_I = \begin{pmatrix} w_{11} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ w_{1n_{l_1}} & 0 & \dots & 0 \\ 0 & w_{21} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & w_{2n_{l_2}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_{n_p 1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_{n_p n_{l_{n_p}}} \end{pmatrix} = \mathbf{W}_T^T \quad (7.39)$$

The base model equation Eq.(7.17) after attaching a macro-model can be expressed

as

$$\begin{pmatrix} \mathbf{G}_{b_{11}} & \mathbf{G}_{b_{12}} & 0 \\ \mathbf{G}_{b_{21}} & \mathbf{G}_a - \mathbf{G}_l & \mathbf{G}_l \\ 0 & \mathbf{G}_l & \mathbf{W}_T^T \mathbf{G}_p \mathbf{W}_T - \mathbf{G}_l \end{pmatrix} \begin{pmatrix} \mathbf{T}_{b_1} \\ \mathbf{T}_a \\ \mathbf{T}_l \end{pmatrix} = \begin{pmatrix} \mathbf{B}_{b_1} \\ \mathbf{B}_a \\ -\mathbf{W}_T^T \mathbf{B}_p \end{pmatrix} \quad (7.40)$$

From Eqs.(7.40), one can see that a dense matrix is formed ($n_l \times n_l$ where n_l is total number of linking nodes and $n_l = \sum_{i=1}^{n_p} n_{l_i}$). Since the number of linking nodes is larger than the number of ports, the simulation time may be slightly larger than use of uniform temperature ports mainly due to the added non-zero elements as discussed in Chapter 3. It has been shown that as the number of linking nodes increases up to certain values, say 300, this dense matrix may greatly slow down the solution of Eqs.(7.40) in some cases.

Figs.7.4(b) and (d) show the sparsity of matrix $\tilde{\mathbf{G}}_b$ when uniform heat flow ports are used. In the right bottom corner, a large dense matrix with 256×256 (256 linking nodes) can be seen no matter whether a single port or 16 ports are used.

Figure 7.4: Sparsity of matrix $\tilde{\mathbf{G}}_b$ for back-side fixed BC and linking nodes $n_l=256$. a) Linking a single port macro-model using uniform temperature port; b) Linking a single port macro-model using uniform heat flow port; c) Linking 16-port macro-model using uniform temperature port; d) Linking 16-port macro-model using uniform heat flow port.

7.5 Results and Analysis

Basically, there are three factors which affect the accuracy of using a macro-model as compared to the fully detailed model. They include: (1) an error from a macro-model itself due to introduction of multiple ports unless the ports are defined in such a way that every single linking node of the macro-model is used as a single port; (2) an error associated with the procedure of creating the macro-model, such as, optimization of resistor values and MDMR technique. As has been shown in Chapter 4 that the error from MDMR is actually negligible ($< 0.1\%$); (3) the assumptions of dealing the linking node temperatures and heat flows of the base model with the port temperatures and heat flows of the macro-model. Mathematically, they are about the coefficients in the matrices $\mathbf{W}_T$ and $\mathbf{W}_I$, which will be especially addressed in the following.

The model under investigation shown in Fig.7.5 (a) and (b) is a generic model of C4/CBGA package on PCB. The details about this model has been presented and explained in Chapter 3. The PCB is simplified to a single material with an effective thermal conductivity. The package model has 20,312 blocks and the PCB model only 1,052 blocks.

A variety of temperature distributions on the PCB model are explored by changing the BCs of the PCB and package model and the PCB effective thermal conductivities. A fixed BC is imposed on either the right-side or the back-side of PCB, and a mixed BC is placed on the top of package characterized by the heat transfer coefficient λ . The effective thermal conductivities of the PCB are varied in a wide range from 2 up to 190 ($W/m.K$) isotropically and anisotropically.

To evaluate the use of a macro-model with either uniform temperature ports or uniform heat flow ports, the relative error is defined to be the corresponding temperature difference over the maximum temperature rise in both the base model and the

macro-model as:

$$e_j = \frac{|T_j^f - T_j^m|}{T_j^f - 300} \quad (7.41)$$

where $j = 1$ denotes the PCB and $j = 2$ for the package; T_j^f is the maximum temperature rise obtained from the fully detailed model and T_j^m from the macro-model linked to the base model.

As known from Section 7.3, the key difference between using uniform temperature ports and uniform heat flow ports to link a macro-model is how to deal with the linking nodes. Therefore, 3-D profiles of the temperatures and heat flows of the PCB linking nodes will be presented to help evaluate uniform temperature ports and uniform heat flow ports. The temperatures and heat flows on the linking nodes of PCB are sampled at the defined locations as shown in Fig.7.5(c). As seen in Fig.7.5(d) is a typical 4-port configuration.

7.5.1 Fully detailed solution

As a reference base, the maximum temperature of both the PCB and package models are obtained and listed in the Table 7.1. 3-D temperature and heat flow density profiles on the PCB surface are plotted from the fully detailed simulation as shown in Fig. 7.6. As a matter of fact, one can see that neither the temperature nor heat flow density profile of the linking nodes is uniform, which means that either use of uniform temperature ports or use of uniform heat flow ports will introduce an error to the simulation compared to the fully detailed results if insufficient number of ports is used.

Fixed BC	$T_1(K)$	$T_2(K)$
Back-side	311.20	326.38
Right-side	334.20	349.28

Table 7.1: Maximum temperatures of the PCB and package from fully detailed model at $K=19$ ($W/m.K$) and $\lambda = 0$

Figure 7.5: a) *Atar* model of package and PCB; b) Cross-section view of package model; c) 64 positions where the temperatures and heat flows are sampled; d) 4-port configuration

Figure 7.6: 3-D profiles of linking node temperatures and heat flow densities from the fully detailed model at $K=19$ ($W/m.K$) and $\lambda = 0$. a) Linking node temperature profile for *back-side* fixed BC; b) Heat flow density profile for *back-side* fixed BC; c) Linking node temperature profile for *right-side* fixed BC; d) Heat flow density profile for *right-side* fixed BC

7.5.2 Uniform temperature ports

Uniform temperature ports have been used in the previous chapters. The results presented here are used mainly for the comparison purpose in order to demonstrate the advantages of using uniform heat flow ports. For a single port, Fig.7.7(a) shows the profiles of the surface temperature and heat flow density when the PCB is imposed to the right-side fixed boundary condition. In this case, a large temperature gradient is expected to occur in order to simulate the worse case [13]. One can see from Fig. 7.7(a) that the temperature profile is just as expected to be a uniform, but clearly different from Fig.7.6(c) from the fully detailed simulation. The heat flow in Fig.7.7(b) is also significantly different from in Fig.7.6(d). From Table 7.4, one can see that there is a relative error of 11.89% on the PCB and 2.64% on the package. The results from the same conditions as above except 4 ports employed are shown in Figs.7.7(c) and (d).

The errors with the number of ports are given in Table 7.2. It is shown that the errors are monotonically decreased with increasing the number of ports. As the number of ports is increased to 64, which is equal to the number of the surface blocks of the macro-model, the situation is equivalent to the fully detailed model and thus the relative error is going down to zero for 64 ports. All these results are consistent with the previous ones in Chapter 3. It may be noted that all temperatures both on the PCB and the package model as shown in Table 7.2 are lower than those from the fully detailed model.

It may be noticed that the relative errors of the maximum temperature on PCB are generally much larger than those on the package model. The reasons could be: (1) the maximum temperature on the PCB is usually much lower than the maximum temperature on the package model, the same magnitude of temperature error will give a larger error on the PCB than on the package according to Eq.(7.41); (2) the maximum temperature of the PCB is actually from one of the linking node tempera-

tures since there is no power source in the PCB, but the maximum temperature of the package is indirectly related to these linking node temperatures in a very complicated way through the port temperatures and Eq.(7.1). Intuitively, it may be understood from the fact that the position of the maximum temperature on the package model is physically away from these linking nodes. In a word, the relative errors of the package model would be less sensitive to the errors of the linking node temperatures caused by introduction of uniform temperature ports.

Num of ports	$T_1(K)$	$T_2(K)$	$e_1(\%)$	$e_2(\%)$
1	330.13	347.97	11.89	2.64
4	330.96	348.10	9.46	2.38
16	332.88	348.97	3.86	0.61
64	334.20	349.28	0	0

Table 7.2: Maximum temperatures and errors as function of the number of ports using *uniform temperature ports* for *right-side* fixed BC, $K=19$ ($W/m.K$) and $\lambda = 0$.

7.5.3 Uniform heat flow ports

As for the newly proposed uniform heat flow ports, it is assumed that the heat flow density within a port area to be uniform and the integration over the ports are equal to the port heat flow. The port temperature for the macro-model is obtained by the area-weighted average of all the linking node temperatures within the port. All the linking node temperatures within a port are allowed to change.

As a direct comparison, the results from the same conditions employed in Section 7.5.2 except using uniform heat flow ports are given in Table 7.3 and Fig.7.8. By comparing Table 7.2 with Table 7.3, one can see that the relative errors have been significantly reduced after using uniform heat flow ports. Note that the errors go down to zero as well as the number of ports reaches 64. As seen from Fig.7.8(b) for a single port macro-model and Fig.7.8(d) for a 4 port macro-model, the heat flow densities are uniform as required within a port, and the temperature profiles shown

Figure 7.7: 3-D profiles of linking node temperatures and heat flow densities using *uniform temperature ports* for *right-side* fixed BC. a) Linking node temperature profile using a single port macro-model; b) Linking node heat flow density profiles using a single port macro-model; c) Linking node temperature profile using 4-port macro-model; d) Linking node heat flow density profile using 4-port macro-model.

in Fig.7.8(a) and (c) are much close to those shown in Fig.7.6 from the fully detailed model. Note that the relative error in Table 7.3 does not decrease monotonically with increasing the number of ports anymore, but increases first and then decreases with the number of ports. The reason for this non-monotonic behavior could be due to the two interacting facts: (1) use of more ports conceivably gives better solution as discussed in Chapter 3 and zero error will be found for 64 ports; (2) at the same time, however, if more ports are defined and then the number of linking nodes underneath each port will be reduced, so a lower number of linking node temperatures will be averaged for the port temperatures and thus lead to a less accurate prediction of the junction temperature of the macro-model. As an indirect proof of this explanation, it can be seen from Table 7.4 that this non-monotonic behavior does not appear when the temperature gradient on the PCB is small as a back-side fixed BC is applied. In this case, only a small difference would be expected either averaging more or less number of linking node temperatures. Therefore, this relationship seems to happen only if a large temperature gradient on PCB presents, such as in a case when the right-side fixed boundary condition is applied.

It is worthy mentioning that using uniform heat flow ports the maximum temperature on the PCB model for a single port macro-model as shown in Table 7.3 is higher than that from the fully detailed model, which was never observed when uniform temperature ports are used. From the comparison between Table 7.2 and Table 7.3, it shows that use of uniform heat flow ports gives better performance than use of uniform temperature ports, particularly when a large temperature gradient exists on the PCB.

7.5.4 Effects of BCs and PCB thermal conductivities

To further validate the use of uniform heat flow ports, more parameters are examined to study the effects of boundary conditions and PCB effective thermal conductivities

Figure 7.8: 3-D profiles of linking node temperatures and heat flow densities using *uniform heat flow ports* for *right-side* fixed BC. a) Linking node temperature profile using a single port macro-model; b) Linking node heat flow density profiles using a single port macro-model; c) Linking node temperature profile using 4-port macro-model; d) Linking node heat flow density profile using 4-port macro-model.

Num of ports	$T_1(K)$	$T_2(K)$	$e_1(\%)$	$e_2(\%)$
1	334.54	349.10	1.00	0.36
4	333.72	348.66	1.41	1.25
16	333.47	349.15	2.13	0.26
64	334.20	349.28	0	0

Table 7.3: Maximum temperatures and errors as function of the number of ports using *uniform heat flow ports* for *right-side* fixed BC, $K=19$ ($W/m.K$) and $\lambda = 0$.

on the use of a multi-port macro-model. Presented in the following are the comparison between the results from uniform heat flow ports and from uniform temperature ports all compared to the corresponding fully detailed model.

Boundary Conditions

The boundary conditions associated with the PCB model can be varied by imposing a fixed $T = 300K$ BC to either the back-side or the right-side of the PCB. The boundary condition with the package model can be changed by the heat exchange coefficient λ . The left-side of Table 7.4 shows the relative errors of the junction temperatures on the package models with the boundary conditions by using uniform temperature ports. As observed and discussed in Chapter 3, the errors are decreased as increasing λ because of decrease of the temperature gradients on the PCB as λ increases. The errors of the maximum temperatures on the package model are slightly smaller for the right-side fixed BC than the back-side fixed BC. It is probably attributed to the large difference of the junction temperature rises of 26.38 K for the back-side fixed BC and 49.28 K for the right-side fixed BC shown in Table 7.1. The reason for the large difference of junction temperatures with BCs are that (1) the back-side area of PCB model is much larger than the right-side and (2) the position of the junction is closer to the back-side than to the right-side as seen in Fig.7.5.

Shown in the right-side of Table 7.4 are the corresponding results by using uniform heat flow ports. One can see that the relative errors of using uniform heat flow ports are consistently smaller than using uniform temperature ports under exactly the same conditions.

PCB thermal conductivities

Both isotropic and anisotropic thermal conductivities of the PCB from a wide range from 2 to 200 ($W/m.K$) were investigated. The bar graphs shown in Fig.7.9 are

n_p	Uniform temperature ports				Uniform heat flow ports			
	$\lambda = 0$		$\lambda = 10^{-8}$		$\lambda = 0$		$\lambda = 10^{-8}$	
	Back	Right	Back	Right	Back	Right	Back	Right
	$e_2(\%)$	$e_2(\%)$	$e_2(\%)$	$e_2(\%)$	$e_2(\%)$	$e_2(\%)$	$e_2(\%)$	$e_2(\%)$
1	3.17	2.64	2.96	2.28	2.72	0.36	2.69	1.43
4	2.95	2.38	2.78	2.17	2.50	1.25	2.51	1.75
16	0.77	0.61	0.75	0.59	0.56	0.26	0.62	0.47
64	0	0	0	0	0	0	0	0

Table 7.4: Comparison of the relative errors of junction temperatures with λ between using uniform temperature ports and uniform heat flow ports at $K=19$ ($W/m.K$)

the errors with isotropic/anisotropic PCB thermal conductivities as a fixed boundary condition is applied to the *back-side* of the PCB model and 4-port package macro-model is used. The relative error shown in Fig.7.9 (b) consistently demonstrates that the use of uniform heat flow ports gives better results than use of uniform temperature ports with respect to predicting the junction temperature. From Fig.7.9 (b), one can also see that the difference of the relative errors are reduced between from uniform heat flow ports and from uniform temperature ports as the PCB thermal conductivity increases, particularly in Z direction. It would be suggested that the improvements of use of uniform heat flow ports is more pronounced at a relatively lower PCB thermal conductivity.

As use of uniform temperature ports, it is interesting to see that the temperature errors as shown in Fig.7.9 (c) and (d) on both the PCB and package models decrease monotonically with increasing the isotropic thermal conductivities, however, the relative errors of the PCB shown in Fig.7.9 (a) conversely increase with increasing thermal conductivities because the maximum temperatures on the PCB and the package drop even faster with the PCB thermal conductivities.

With a very low PCB thermal conductivity, such as $K=2$ ($W/m.K$), using uniform heat flow ports could produces a large temperature error on the PCB as can be seen in Fig.7.9(c), but the relative error on the PCB is not too bad as shown in Fig.7.9(a) mainly because of a high temperature rise at such a low thermal conductivity even

though it is not the scenario people want to see in practice. Irrespective of large errors on the PCB, surprisingly small errors on the package model are obtained and it may be attributed to the weighted average of the linking node temperatures for the port temperatures.

Due to a uniform heat flow density, a peaked linking node temperature distribution would always be expected to occur in the central area of ports. It is believed to be the reason why the maximum temperature could be higher than that from fully detailed model at a single port. The peaked shape will become significant as the PCB thermal conductivity are decreased and the number of ports is reduced. But, in reality, the linking node temperature distribution becomes less peaked as decreasing the thermal conductivities by comparing Fig.7.6 with Fig.7.10. It implies that use of uniform heat flow ports would not be desirable for a very low PCB thermal conductivity unless a sufficient number of ports are used. Fig.7.10 presents the results from the fully detailed model, the macro-model with uniform temperature ports and uniform heat flow ports at $K=2$ ($W/m.K$). It is obvious to see a large temperature error on the PCB model using uniform heat flow ports.

Similar study where a fixed boundary condition applied to the *right-side* of PCB has also been carried out and the results are illustrated in Fig.7.11. Except for the very low thermal conductivities, such as $K=2$ ($W/m.K$) particularly in Y direction, use of uniform heat flow ports gives smaller errors than use of uniform temperature ports both on the PCB and the package models. Fig.7.12 presents the results from the fully detailed model, the macro-model with uniform temperature ports and uniform heat flow ports at $K=2$ ($W/m.K$). It is not difficult to see that the uniform temperature ports are a better representation of the linking node temperatures than uniform heat flow ports for this special case.

Figure 7.9: Maximum temperature errors as function of PCB thermal conductivities using 4-port macro-model for *back-side* fixed BC. Isotropic thermal conductivity($W/m.K$) indices as: (1) $K_x = K_y = K_z = 2$; (2) $K_x = K_y = K_z = 10$; (3) $K_x = K_y = K_z = 19$; (4) $K_x = K_y = K_z = 50$; (5) $K_x = K_y = K_z = 100$; (6) $K_x = K_y = K_z = 200$. Anisotropic thermal conductivity indices as: (7) $K_x = K_y = 2$, $K_z = 19$; (8) $K_x = K_y = 19$, $K_z = 2$; (9) $K_x = K_y = 19$, $K_z = 190$. a) Relative errors on PCB; b) Relative errors on package; c) Absolute errors on PCB; d) Absolute errors on Package.

Figure 7.10: Comparison of 3-D profiles of linking node temperatures and heat flow densities between fully detailed model, macro-model with uniform temperature ports and macro-model with uniform heat flow ports at $K=2$ ($W/m.K$) for *back-side* fixed BC. a) and b) Fully detailed model, c) and d) 4 uniform temperature ports, e) and f) 4 uniform heat flow ports

Figure 7.11: Maximum temperature errors as function of PCB thermal conductivities using 4-port macro-model for *right-side* fixed BC. Isotropic thermal conductivity($W/m.K$) indices as: (1) $K_x = K_y = K_z = 2$; (2) $K_x = K_y = K_z = 10$; (3) $K_x = K_y = K_z = 19$; (4) $K_x = K_y = K_z = 50$; (5) $K_x = K_y = K_z = 100$; (6) $K_x = K_y = K_z = 200$. Anisotropic thermal conductivity indices as: (7) $K_x = K_y = 2$, $K_z = 19$; (8) $K_x = K_y = 19$, $K_z = 2$; (9) $K_x = K_y = 19$, $K_z = 190$. a) Relative errors on PCB; b) Relative errors on package; c) Absolute errors on PCB; d) Absolute errors on Package.

Figure 7.12: Comparison of 3-D profiles of linking node temperatures and heat flow densities between fully detailed model, macro-model with uniform temperature ports and macro-model with uniform heat flow ports at $K = 2 W/m.K$ and *right-side* fixed BC. a) and b) Fully detailed model, c) and d) 4 uniform temperature ports, e) and f) 4 uniform heat flow ports.

7.6 Summary

On basis of the above results and analysis, the following conclusions may be summarized:

- A general method of linking a macro-model to the base model is proposed and mathematically implemented. On basis of it, the idea of uniform heat flow ports are presented and studied.
- Both for back-side and right-side fixed BCs, the results in term of predicting junction temperatures from use of uniform heat flow ports are generally better than from use of uniform temperature ports except for an extremely low thermal conductivity, such as 2 ($W/m.K$), which usually is not a realistic case since the junction temperature would be way above the acceptable limits in a real life.
- If each surface blocks of the macro-model is defined as a single port, then use of uniform temperature ports and uniform heat flow ports become identical and equivalent to the fully detailed simulation.
- The use of uniform heat flow ports can better handle the case where a large temperature gradient on the base model presents
- It has been observed that the maximum temperature on the base model predicted by use of uniform heat flow ports could be higher than that from the fully detailed model, but use of uniform heat temperature ports always gives lower the maximum temperature than that from the fully detailed model.
- As the thermal conductivity K increases, the difference between the results from uniform temperature ports and from uniform heat flow ports becomes smaller. The errors of the maximum temperatures on PCB are generally much larger than those on the package model.

Chapter 8

Conclusions and Future Work

This chapter contains a summary of the work presented in this thesis. In addition, the direction of future work is also discussed.

8.1 Summary

1. The MDMR technique has been successfully used in thermal simulations by the *thermal modelling group* at Carleton university. MDMR can explicitly incorporate the design parameters such as BC parameters into the reduced model. This is considerable benefit as from a single reduced model we can now perform many simulations of a detailed model under BCs.

Moreover, the reduced-order model is capable of generating a full 3D temperature distribution in the original model. The reduced-order models have been shown to be very accurate, with nodal temperature errors of less than 0.1% over a wide range of physical conditions. Due to the small size of the reduced model, applying multi-dimensional BCs and solving for the full model temperature distribution is easy and efficient with speed-ups of 30 times typically. Therefore, the reduced models are well suited for use in the design environments employed

by the electronic industries.

2. The method of *direct sub-models* has been proposed to cope with very complex systems with large model sizes. A large model can be partitioned into a couple of smaller models which are represented by a simple mathematical relationship between the port temperature and port heat flow.

The creation and application of direct sub-models have been performed with examples of real electronic components. Use of direct sub-models can predict all original nodal temperatures and reduce memory requirements and CPU cost. The procedure of creating a direct sub-model can be applied to a mixed-model which already includes a couple of sub-models, then a hierarchical direct sub-model can be generated and the original complex model has been gradually simplified. There is no limitation to creating higher level hierarchical direct sub-models as required. It has been shown that a high accuracy can also be achieved provided that an appropriate number of ports and port configurations are defined properly.

3. The MDMR technique has been extended to direct sub-models with multiple ports to form multi-dimensional *reduced sub-models* and multi-dimensional *macro-model*(MDMM). Reduced sub-models have been characterized as function of expansion point, number of ports and number of moments. Size reduction achieved by using MDMR is shown to be three to four orders of magnitude, yet is able to predict nodal temperatures in the original model with high accuracy even for wide boundary conditions. A speed-up of about 200 to 500 times has been achieved. In addition, MDMR can be applied to a mixed-model which already included reduced sub-models to generate a hierarchical reduced sub-model.

A MDMM created using MDMR technique has been testified to be boundary

condition independent(BCI). A MDMM has also shown to have a number of advantages over compact models of resistor networks. The generation of the model is at least an order of magnitude faster than the creation of an optimized network model. The MDMM displays very high accuracy typically better than 0.1%. In addition, the MDMM is very flexible allowing for the prediction of all internal temperatures and presenting no limitations on the external configuration of the compact model.

4. The *global method* has been presented to provide a more general way to link reduced sub-models to a base model by mathematically forming global model equations. The global method allows BCs of both the sub-models and the base model to be explicitly incorporated, and so allows the MDMR technique to be easily applied to reduce the size of the global model. Therefore, the reduced global model can handle BC variations of both the base model and sub-models in a reduced space. More importantly, the global method makes the thermal transient simulations of using reduced sub-models possible. The great benefit of thermal transient simulations using reduced global models will be presented in the future.
5. As for board and system level simulations, a very important issue regarding how to treat the relationship between the linking nodes of the base model and the defined ports of the macro-model has been addressed in the thesis. A generalized method of dealing with this issue has been developed and mathematically implemented. Uniform heat flow ports rather than uniform temperature ports are proposed to accommodate different thermal distributions and thermal properties of a substrate or PCB.

In term of predicting junction temperatures, it has been shown that the results from use of uniform heat flow ports are generally better than from use of uniform

temperature ports except for an extremely low thermal conductivity, such as 2 ($W/m.K$), which is usually not a realistic case since the junction temperature would be way above the acceptable limits. In addition, it has also been proven that the use of uniform heat flow ports can better handle the case where a large temperature gradient on the base model is presented

As either the number of ports or the thermal conductivity K of the base model increases, the difference between the results from uniform temperature ports and from uniform heat flow ports becomes smaller. As the number of ports increases such that each surface block of the macro-model is defined as a single port(port/block), then use of uniform temperature ports and uniform heat flow ports turn out to be identical and equivalent to the fully detailed simulation.

8.2 Future Work/Directions

The work presented in this thesis may be further extended and could be applied in various fields other than the thermal simulations.

1. Further study of linking a macro-model to a base model
2. Efficient thermal transient analysis on basis of the global method and MDMR techniques.
3. The MDMR technique for optimization of large systems with respect to design parameters.
4. The MDMR technique used in the other fields, such as Microelectromechanical systems (MEMS).

Bibliography

- [1] K. Azar, "Power consumption and generation in the electronics industry. a perspective," in *Sixteenth IEEE SEMI-THERM Symp.*, Mar 2000, pp. 21–23.
- [2] S. V. Garimella, Y. K. Yoshi, and et al, "Thermal challenges in next generation electronics systems-summary of panel presentations and discussions," *IEEE Transactions on Components and Packaging Technologies*, vol. 25, no. 4, pp. 569–575, Dec. 2002.
- [3] Y. Joshi, M. Baelmans, D. Copeland, C. J. M. Lasance, J. Parry, and J. Rantal, "Challenges in thermal modeling of electronics at the system level: summary of panel held at the thermic 2000," *IEEE Transactions on Components and Packaging Technologies*, vol. 24, no. 4, pp. 611–613, Dec. 2001.
- [4] H. J. Eggink, G. M. Janssen, and J. H. J Janssen, "Using compact models in the early thermal design of electronics," in *Nineteenth IEEE SEMI-THERM Symp.*, Mar. 2003, pp. 324–331.
- [5] W. Luiten, "Sense and nonsense thermal requirements," in *nineteenth IEEE SEMI-THERM Symp.*, Mar. 2003, pp. 332–335.
- [6] S. Shidore and A. Sahrapour, "Delphi compact models revolutionize thermal design," in *www.coolingzone.com*, JULY. 2001, pp. 1–6.

- [7] J. Parry and H. Rosten, "Development and application of compact models of packages based on delphi methodology," in *Eurotherm 58 Conference*, Sep. 1997, pp. 161–168.
- [8] H. Vinke and C. J. M. Lasance, "Recent achievements in the thermal characterization of electronic devices by means of boundary condition independent compact models," in *Thirteenth IEEE SEMI-THERM Symp.*, Jan. 1997, pp. 32–39.
- [9] J. D. Moerlose and W. Temmerman, "Thermal analysis of a chip on board(cob)," in *Thirteenth IEEE SEMI-THERM Symp.*, Mar. 1997, pp. 112–118.
- [10] T. Smy, D. Walkey, and S. K Dew, "A 3-d thermal simulation tool for integrated devices - atar," *IEEE Trans on CAD of IC and Syst*, vol. 20, no. 1, pp. 105–115, Jan. 2001.
- [11] T. Smy, D. Walkey, and S. K Dew, "Transient 3d heat flow analysis for integrated circuit devices using transmission line matrix method on a quad tree mesh," *Solid-State Electronics*, vol. 45, no. 7, pp. 1137–1148, Jul. 2001.
- [12] T. Smy, D. Walkey, M. Shams, and R. V. Joshi, "Simulation of local heating in vlsi backend structures using atar," in *Advanced Metallization Conference 2001*, Oct 2001, pp. 355–.
- [13] J. DeVoe and A. Ortega, "An investigation of board level effects on compact thermal models of electronic chip packages," in *Eighteenth IEEE SEMI-THERM Symp.*, Mar. 2002, pp. 8–14.
- [14] J. DeVoe and A. Ortega, "An assessment of the behavior of compact thermal models of electronic packages in a printed circuit board level environment," in *Eighteenth IEEE SEMI-THERM Symp.*, May 2002, pp. 69–76.

- [15] E. Sansoucy, G. Refai-Ahmed, and K. V. Karimanal, "Thermal characterization of tbga package for an integration in board level analysis," in *The Eighth Inter-society Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, May 2002, pp. 91–98.
- [16] A. B. Cohen and W. B. Krueger, "Thermal characterization of chip packages-evolutionary developments of compact models," in *Thirteenth IEEE SEMI-THERM Symp.*, Mar. 1997, pp. 180–197.
- [17] C. J. M. Lasance, "Thermal characterization of electronic parts with compact models: Interpretation, application, and the need for a paradigm shift," in *Thirteenth IEEE SEMI-THERM Symp.*, Mar. 1997, pp. 40–48.
- [18] C. J. M. Lasance, "Creation and evaluation of compact models for thermal characterization using dedicated optimization software," in *Fifteenth IEEE SEMI-THERM Symposium*, Mar. 1999, pp. 189–200.
- [19] T. Bechtold, E. B. Rudnyi, and J. G. Korvink, "Automatic order reduction of thermo-electric models for mems: Arnoldi versus gyan," in *The Fourth Int. Conf. On Advanced Semiconductor Devices and Microsystems*, Oct. 2002, pp. 333–336.
- [20] T. Nishio, Y. Yamada, and K. Koyamada, "A compact modeling approach using a genetic algorithm for accurate thermal simulation," in *IEMT/IMC Symposium*, Apr. 1998, pp. 167–172.
- [21] J. DeVoe and A. Ortega, "On the performance of compact thermal models of electronic chip packages in conjugate board level simulation," in *Nineteenth IEEE SEMI-THERM Symp.*, Mar. 2003, pp. 313–318.

- [22] C. V. Loh, K. C. Toh, D. Pinjala, and M. K. Iyer, "Development of effective compact models for depopulated ball grid array packages," in *IEEE Electronic Packaging Technology Conference 2000*, Dec. 2000, pp. 131–137.
- [23] J. Culham and M. Yovanovich, "Factors affecting the calculation of effective conductivity in printed circuit boards [thermal analysis]," in *The Sixth Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, May 1998, pp. 460–467.
- [24] R. Achar and M. Nakhla, "Simulation of high-speed interconnects," *Proceedings of the IEEE*, vol. 98, no. 5, pp. 693–728, May 2001.
- [25] R. Khazaka, "Projection based techniques for the simulation of rf circuit and high speed interconnects," in *Ph.D Thesis*, 2002.
- [26] P. Gunupudi, "Circuit-reduction techniques and application to high-speed systems and interconnects," in *Ph.D Thesis*, 2002.
- [27] A. Odabasioglu, M. Celik, and L. T. Pileggi, "Prima: passive reduced-order interconnect macromodelling algorithm," *IEEE Transactions on Computer-Aided design*, vol. 17, no. 8, pp. 645–653, Aug. 1998.
- [28] J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM Publishers, Philadelphia, PA, 1997.
- [29] I. M. Elfadel and D. D. Ling, "A block rational arnoldi algorithm for multi-point passive model-order reduction of multiport rlc networks," in *IEEE/ACM international conference on Computer-Aided Design*, Nov. 1997, pp. 66–71.
- [30] D. Celo, P. Gunupudi, R. Khazaka, D. Walkey, T. Smy, and M. Nakhla, "Fast simulation of steady-state temperature distribution in electronic components us-

- ing multi-dimensional model reduction,” *IEEE Transactions on Components and Packaging Technologies*, accepted for publication, Apr. 2004.
- [31] P. Gunupudi, R. Khazaka, M Nakhla, T. Smy, and D. Celo, “Passive parameterized time-domain macromodels for high-speed transmission-line networks,” *IEEE Transactions on Microwave Theory and Techniques*, vol. 51, no. 12, pp. 2347–2354, Dec. 2003.
 - [32] L. T. Pileggi and R. A. Rohrer, “Asymptotic waveform evaluation for timing analysis,” *IEEE Transactions on Computer-Aided design*, vol. 9, pp. 352–366, Apr. 1990.
 - [33] E. Chiprout and M. S. Nakhla, “Analysis of interconnect networks using complex frequency hopping (cfh),” *IEEE Transactions on Computer-Aided design*, vol. 14, no. 2, pp. 186–200, Feb. 1995.
 - [34] P. Gunupudi and M. Nakhla, “Multi-dimensional model reduction of vlsi interconnects,” in *Proc. IEEE CICC*, May 2000, pp. 499–502.
 - [35] P. Gunupudi and M. Nakhla, “Analysis of transmission line circuits using multi-dimensional model reduction techniques,” in *Proc. IEEE Meeting on Electrical Performance and Electronic Packaging*, Oct. 2000, pp. 43–46.
 - [36] H. Carslaw and J. L Jeager, *Conduction of Heat in Solids, 2nd ed.*, Oxford University Press, London, 1959.
 - [37] L. C Thomas, *Heat Transfer-Professional Version, 2nd ed.*, Capstone Publishing Corporation, Tulsa, Oklahoma, 1999.
 - [38] C. K. Tam, K. L. Koon, Z. R. Hu, and A. Rezazadeh, “Self-heating effects on gan fets for high power microwave applications,” in *Proc. EDMO 2001*, 2001.

- [39] D. Allen and et al., "A 0.2 μm 1.8v soi 550mhz 64 b power pc microprocessor with cu interconnects," in *Proc. ISSCC Conf.*, Oct. 1999, pp. 438–439.
- [40] R. V. Joshi and et al., "High performance soi/cu srams and memories in microprocessors," in *Advanced Metallization Conference 2001*, Oct. 2001, pp. 347–.
- [41] G. Kromann, R. D. Greke, and W. Huang, "Hi-density c4/cbga interconnect technology for a cmos microprocessor," *IEEE Transactions on Components, Packaging, and Manufacturing Technology-Part B*, vol. 19, no. 1, pp. 166–173, Feb. 1996.
- [42] Z Johnson, K. Ramakrishna, B. Joiner, and M. Eyman, "Thermal sub-modeling of the wirebonded plastic ball grid array package," in *Thirteenth IEEE SEMI-THERM Symp.*, Jan 1997, pp. 1–9.
- [43] X Guo, D. Celo, P. Gunupudi, R. Khazaka, D. Walkey, T. Smy, and M. Nakhla, "The creation of compact thermal models of electronic components using model reduction," *IEEE Transactions on Components and Packaging Technologies*, submitted for publication, Nov. 2003.
- [44] X Guo, D. Celo, P. Gunupudi, R. Khazaka, D. Walkey, T. Smy, and M. Nakhla, "A general method for the connections of thermal macro-models," *IEEE Transactions on Components and Packaging Technologies*, submitted for publication, Apr. 2004.
- [45] X Guo, D. Celo, P. Gunupudi, R. Khazaka, D. Walkey, T. Smy, and M. Nakhla, "Multi-dimensional compact thermal models of ic components using model reduction," in *Twentieth IEEE SEMI-THERM Symp.*, Mar. 2004.
- [46] D. Celo, X. Guo, P. Gunupudi, R. Khazaka, D. Walkey, T. Smy, and M. Nakhla, "The creation of compact thermal models of electronic components using model reduction," in *ITHERM2004*, June. 2004.

- [47] Q. Xia, "Investigation of gaas pHEMT distributed amplification for broadband modulator driver applications," in *M. A. Sc. Thesis*, 2003.
- [48] P. Gunupudi, M. Nakhla, and R. Achar, "Multi-point multi-port reduction of high-speed distributed interconnects using krylov-space techniques," in *Proc. IEEE International Symposium on Circuits and Systems*, Jun. 1999, vol. 6, pp. 242–245.
- [49] A Aranyosi, A. Ortega, R. Griffin, S. West, and D Edwards, "Compact thermal models of packages used in conduction cooled applications," *IEEE Transactions on Components and Packaging Technologies*, vol. 23, no. 3, pp. 470–480, Sep. 2000.
- [50] C. J. M. Lasance, H. Vinke, and H. Rosten, "Thermal characterization of electronic devices with boundary condition independent compact models," *IEEE Transactions on Components, Packaging, and Manufacturing Technology-Part A*, vol. 18, no. 4, pp. 723–731, Dec. 1995.
- [51] C. J. M. Lasance, "Two benchmarks to facilitate the study of compact thermal modeling phenomena," *IEEE Transactions on Components and Packaging Technologies*, vol. 24, no. 4, pp. 559–565, Dec. 2001.
- [52] "Jedec registration for the square ebga (mo-151)," .
- [53] D. Walkey, "Thermal ports," in *Private communication*, Dec. 2003.