Circuits and Systems
Mekelweg 4,
2628 CD Delft
The Netherlands
http://ens.ewi.tudelft.nl/

CAS-MS-2011-08

M.Sc. Thesis

Temperature-Constrained Power Management Scheme for 3D MPSoC

Arnica Aggarwal

Abstract

Process technologies are approaching physical limits making further reduction of device size and higher integration challenging. Three-dimensional (3D) integration is emerging as an attractive solution to continue the pace of growth of System-on-Chips. Although vertical interconnection between the stacked dies has substantial benefits in terms of electrical performance, higher integration density aggravates the prevailing challenges of power density and consequently microelectronics cooling. This makes consideration of temperature constraints important while designing power management schemes. Dynamic Voltage and Frequency Scaling (DVFS) schemes in two-dimensional (2D) Multi-Processor System-on-Chip (MPSoC) do not consider thermal relation between various Processing Elements (PE) however this cannot be ignored in 3D stacks. In this thesis a new temperature constraint power management scheme for 3D MPSoC is proposed. Thermal relation between PEs is represented by the effective thermal resistance between them. These values along with PE's operating temperature, utilization and positional information are used to generate weights for each PE and voltage island. These weights are then used for scaling and imposing temporary constraints on operating voltage and frequency (V/F) levels of PEs in the stack. While scaling brings temperatures of all PEs below critical limits, imposing constraints on the V/F levels avoids significant fluctuations in operating temperatures. When compared to 2D DVFS, an improvement of up to 19.55% in overall execution time is achieved, temperatures are maintained at a safe margin from critical limits and stability in operating temperatures was observed.

TU Delft

Faculty of Electrical Engineering, Mathematics and Computer Science

Delft University of Technology

Temperature-Constrained Power Management Scheme for 3D MPSoC

Thesis

submitted in partial fulfillment of the
requirements for the degree of

Master of Science

in

Electrical Engineering

by

Arnica Aggarwal
born in Lucknow, India

This work was performed in:

Circuits and Systems Group
Department of Microelectronics
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

Delft University of Technology

Delft University of Technology
Department of
Microelectronics

The undersigned hereby certify that they have read and recommend to the Faculty of Electrical Engineering, Mathematics and Computer Science for acceptance a thesis entitled **"Temperature-Constrained Power Management Scheme for 3D MP-SoC"** by **Arnica Aggarwal** in partial fulfillment of the requirements for the degree of **Master of Science**.

Dated: 20/12/2011

Chairman: ____________________
prof.dr.ir. A.J. van der Veen

Advisor: ____________________
dr.ir. T.G.R.M. van Leuken

Committee Members: ____________________
dr. Amir Zjajo

____
dr.ir. Said Hamdioui

Abstract

Process technologies are approaching physical limits making further reduction of device size and higher integration challenging. Three-dimensional (3D) integration is emerging as an attractive solution to continue the pace of growth of System-on-Chips. Although vertical interconnection between the stacked dies has substantial benefits in terms of electrical performance, higher integration density aggravates the prevailing challenges of power density and consequently microelectronics cooling. This makes consideration of temperature constraints important while designing power management schemes. Dynamic Voltage and Frequency Scaling (DVFS) schemes in two-dimensional (2D) Multi-Processor System-on-Chip (MPSoC) do not consider thermal relation between various Processing Elements (PE) however this cannot be ignored in 3D stacks. In this thesis a new temperature constraint power management scheme for 3D MPSoC is proposed. Thermal relation between PEs is represented by the effective thermal resistance between them. These values along with PE's operating temperature, utilization and positional information are used to generate weights for each PE and voltage island. These weights are then used for scaling and imposing temporary constraints on operating voltage and frequency (V/F) levels of PEs in the stack. While scaling brings temperatures of all PEs below critical limits, imposing constraints on the V/F levels avoids significant fluctuations in operating temperatures. When compared to 2D DVFS, an improvement of up to 19.55% in overall execution time is achieved, temperatures are maintained at a safe margin from critical limits and stability in operating temperatures was observed.

Acknowledgments

Quoting Napoleon Hill, “Desire is the starting point of all achievement, not a hope, but a keen pulsating desire, which transcends everything.” I had a desire when I set out to commence my Masters’ study. I had a desire to do well. I had a desire to shape my career, to be the best at what I do, to make my parents proud of me. My work here at TU Delft, for my thesis provided me an ideal stepping stone to help me achieve my desire. However, it was not an easy going. There were numerous hurdles, problems, doubts, confusions. Without the help, support, love and nurturing from a number of people, I would not have reached this day where I can confidently say that I see my dreams coming true.

I would like to give a special thanks to my supervisor, Prof. Rene for providing me with the opportunity to work under his supervision. It was an enriching experience to be able to shape my own thesis. The freedom and the invaluable guidance to carry on the research has helped me to really get into the skin of things and confidently do the things my way. I have emerged a more knowledgeable, experienced and confident individual. Thank you Prof. Rene.

I would like to extend my sincere thanks to Amir Zjajo for discussing my work and providing valuable feedback. It gave a better understanding and clearer picture of things.

Sumeet, Thank you for the endless number of brainstorming sessions, for all your patience and constant motivation. Things always looked simpler after discussing them with you. Thank you, for the times you said “I have to” though it was a choice you made.. for being so humble and for being who you are.

Antoon, for instant help and for patiently fixing the things when I messed them up.

Minaksie, for spreading smiles and for quick help with all administrative things.

Aashini, You are the reason that I am here today. When I was scared to look forward, you held my hand and walked along. Thanks for being there.

Radhika, The time we spent together is invaluable. The discussions, chats and gossips, coffees after long working hours, late night work and much more.. Thank you for being around and for not turning my side of the light on..! :)

Sundeep, Thanks for the energy and motivation you instilled in me, no matter what the challenge was, it gave that push.. Also, for all the rice and sambhar u cooked :)

Sakshi and Dushyant, near or away, you have always been my strength.

Momi, Pa, Meghna, Prateek, Ajay, thanks for believing in me. Your love and support has always made survival easier and life a better place.

Arnica Aggarwal
Delft, The Netherlands
20/12/2011

Contents

List of Figures

List of Tables

List of Abbreviations

2D	Two-Dimensional
3D	Three-Dimensional
DFS	Dynamic Frequency Scaling
DVFS	Dynamic Voltage and Frequency Scaling
DVS	Dynamic Voltage Scaling
F2B	Face-to-Back
F2F	Face-to-Face
MPSoC	Muti-Processor System-on-Chip
NoC	Networks-on-Chip
PCB	Printed Circuit Board
PMB	Power Management Block
SoC	System-on-Chip
TIM	Thermal Interface Material
TSV	Through Silicon Via
V/F	Voltage Frequency level or DVFS level.
VFI	Voltage Frequency Island

Introduction 1

1.1 Motivation

Miniaturization of microelectronic devices has led to a tremendous improvement in the performance of electronic products. However, scaling down the feature size of a transistor introduces limitations on the interconnect performance, the process variation and the leakage power consumption [1]. Total power dissipation and power density are at the limits of what packaging and cooling solutions can support [2]. Process technologies are approaching physical limits making reduction in device size and higher integration more challenging. 3-Dimensional (3D) Technology, i.e., vertical stacking of multiple silicon layers is emerging as an attractive solution to continue the pace of growth of SoCs. Through Silicon Via (TSV) provides vertical interconnection between the stacked dies which greatly reduces interconnection length and results in a smaller area footprint. Although 3D technology has some clearly established benefits in terms of electrical performance [1,3,4], it aggravates the prevailing challenges of power density [5] and microelectronics cooling [5–7], limiting the performance and the reliability of a stacked chip [7, 8]. This makes consideration of temperature constraints important while designing power management schemes.

Reducing the power consumption of System-on-Chip (SoC) and Muti-Processor System-on-Chip (MPSoC) has become increasingly important and challenging in electronic system designs, especially when powered by batteries. Low power designing approaches are used at every step of the design process, from software to architecture to implementation. Dynamic Voltage and Frequency Scaling (DVFS) is a commonly used architecture-level power management technique that allows a PE to operate at different voltage and frequency levels according to its changing workload [2, 9–12]. A temperature constraint power management scheme for two-dimensional (2D) ICs addresses the temperature of each PE independently ignoring the thermal relation between PEs. [13] reported that in a 3D IC, thermal conductance in the vertical direction is 16 times of that in the lateral direction. Also, as the depth of a 3D stack increases, the heat transfer in the lateral direction also becomes prominent. Therefore, thermal relation between the PEs can no longer be ignored. Hence, a temperature constrained power management scheme for a 2D IC cannot be directly implemented in a 3D IC.

This thesis aims to propose a new temperature constrained power management scheme for 3D MPSoCs. It uses utilization factor, instantaneous temperature margin, positional details and area of a PE in a 3D stacked IC for calculating new operating DVFS levels. Utilization factor determines how busy a PE is and instantaneous temperature margin denotes the difference between the critical temperature and its actual temperature. Instantaneous temperature margin is monitored to ensure that each PE operates within the allocated temperature limits. Position of a PE refer to its location in the 3D stack, i.e., how far it is from the heat sink. Position and area also have a significant impact on temperature of a PE and power density in the stack. Hence, the effect of the two are also considered.

1.2 Thesis Goals

The differences between 2D power management schemes and an effective scheme for 3D ICs motivates the work presented in this report. The objectives of this thesis are:

- Design an effective temperature constraint power management scheme for 3D MPSoC.
- Include positional and thermal information in the power management scheme in order to address thermal dependencies.
- Keep total power value below the set budget value.
- Effectively maintain temperatures of PEs below critical limits without significant loss in performance.
- Study the effectiveness of 3D islands in a stacked IC.

1.3 Contributions

This work presents a new approach for power management scheme in 3D stacked IC and is compared with a 2D DVFS scheme. The main contributions of this work are as follows:

- Various factors like instantaneous temperature margin of a PE, its positional information, area, and thermal relation between PEs are included in the power management scheme. These are used to calculate weights for PEs in order to effectively select a PE for scaling its DVFS level.
- Reduced total execution time by preventing PEs on the deeper tiers from being turned OFF.
- Effectively maintains temperatures at a safe margin below critical temperature and power below the budget value. Keeping temperature at a safe margin from the critical temperature ensures that the temperature never exceeds the critical limit even under unexpected circumstances like noise in power supply and sudden increase in workload of a PE.
- Shows less fluctuations in temperature due to higher stability in operating DVFS levels. To maintain the performance of devices over time, it is important to avoid fluctuations in temperature.

- Approach is effectively implemented on voltage islands as well as per-core level. 3D islands achieved further reduction in execution time when executing similar workloads. Granularity can be manipulated to achieve benefits of both, voltage islands as well as per-core DVFS.
- The effective resistance matrix for the PEs is derived for the target floorplan. Therefore, the location of PEs in the stack does not affect the algorithm of power management scheme.

1.4 Thesis Organization

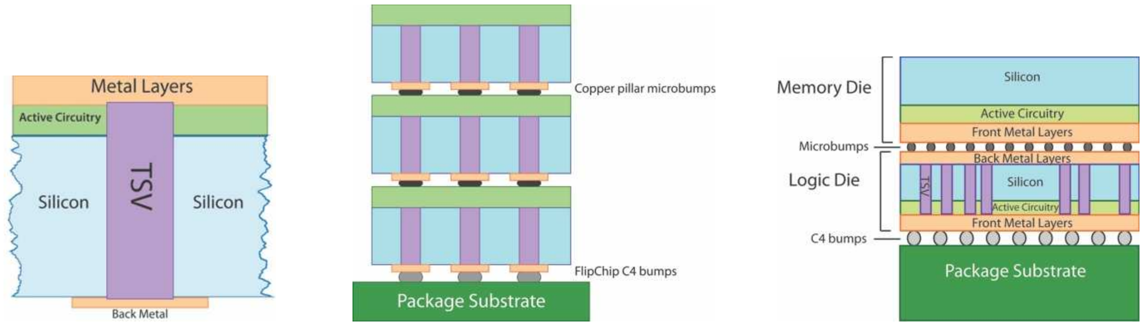
This thesis is organised into the following chapters:
Chaper 2 introduces 3D Integration Technology and basic concepts of power management. Various sources of power dissipation and relation between power and temperature are presented. Power management schemes for 2D MPSoCs and feasibility of extending such schemes to 3D stacks are discussed. Further, thermal modeling of a 3D stacked IC is described and its difference from 2D thermal model is discussed. Lastly, importance of including thermal information in power management of 3D MPSoC is discussed along with a brief discussion on the related work in the field.
Chapter 3 details how the thermal model for the power management control is derived in this work. Heat transfer theory is discussed, a thermal model of a 3D IC is analyzed and the importance of transient analysis is presented, followed by the derived thermal model.
Chapter 4 presents the proposed power management scheme. Implementation of the scheme at voltage island and per-core level is explained. Design details and the complete control algorithm is presented.
Chapter 5 provides the details of simulation environment used for testing the power management scheme. Created testbench to provide inputs to the power management block is described. Details of performance measurements are discussed. Further, the conducted experiments are the obtained results are analyzed. First, a per-core DVFS scheme is demonstrated with lenient temperature constraints. To draw a better comparison, strict temperature constraint is imposed. A comparison is drawn between the new weighted approach and the conventional approach. Followed by the analysis of 3D voltage islands.
Chapter 6 concludes the thesis with a brief discussion on achieved goals and remarks for recommendations for future work.

2 Background

This chapter introduces 3D Integration Technology and basic concepts of power management. The chapter details sources of power dissipation and relation between power and temperature. Various power management schemes for MPSoCs, feasibility of extending such schemes to 3D stacks and need of dc-dc converters are discussed. This chapter also describes the thermal modeling of a 3D stacked IC, its importance in power management of 3D MPSoC and how it is different from 2D thermal models. The chapter is concluded with a brief discussion on the related work.

2.1 3D Integration Technology

Technology scaling has led to a tremendous improvement in the performance of electronic products over decades. The continuation of this trend seems difficult as several process technologies are approaching physical limits making further reduction of device size more challenging by introducing limitations on the interconnect performance, the process variation and the leakage power consumption [1]. Wires consume more than 30% of the power within a microprocessor [4]. Total power dissipation and power density are at the limits of what packaging and cooling solutions can support [2]. 3D integration technology has attracted significant attention in recent past. An example of a 3D integrated IC is shown in Figure 2.1.



(a) The TSV connects the front metal to the back metal layers.

(b) Stacking multiple chips with TSVs together to create a 3D-IC

(c) 3D-IC with TSVs connecting front and back metal layers

Figure 2.1: 3D integration technology using Through Silicon Via(TSV)

Figure 2.1(a) shows how vertical interconnection between the stacked dies is achieved using TSVs which greatly reduced interconnection length and result in a smaller area footprint. It is expected to address interconnect delay related problems and enable integration of heterogeneous technologies [14]. Shorter interconnects would help reduce total power dissipation, but, due to closed packed multi-layer structure, the power density would be significantly high. 3D integration technology thus provides new micro-architecture opportunities to trade-off performance, power and area [4].

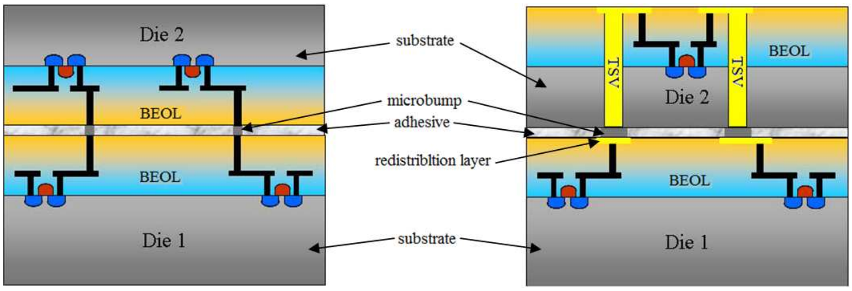


Figure 2.2: Ways of wafer stacking based on the stacking orientation of two device wafers. Left:F2F, Right:F2B [15]

Based on the stacking orientation of two device wafers, there are two different ways of wafer stacking: face-to-face (F2F) and face-to-back (F2B) as shown in Figure 2.2. As the names suggest, in F2F configuration, dies are bonded face-to-face with microbumps, while in F2B configuration, dies are bonded back-to-face and TSVs are used for inter-tier connections. F2F configuration does not require TSVs for bonding if the stack consists of only two tiers (dies), as seen in Figure 2.2. But, bonding more than two dies requires TSVs to provide through silicon bonding of metal layers and the design no longer remains F2F alone. While in F2B configuration, the structure is homogeneous and symmetric with equal lengths of TSVs for equal bulk thickness. For this reason, a F2B bonding for multiple layered 3D stack is considered in this thesis.

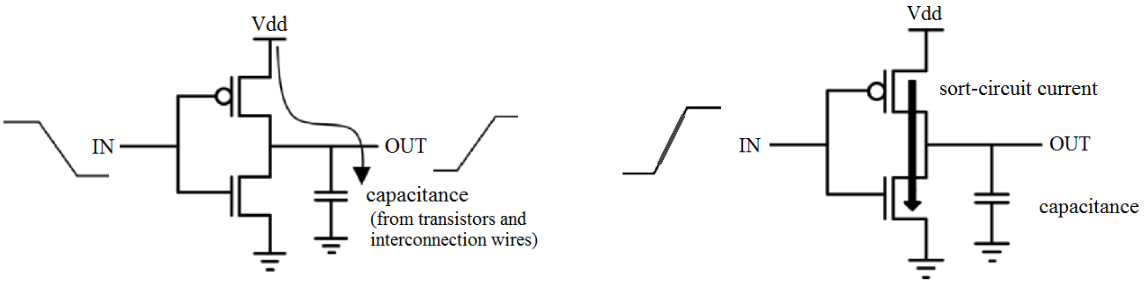
2.2 Power Dissipation

CMOS is a predominant process technology for digital circuits. Power dissipation for these circuits can be accurately modeled using equations, even for complex processors [2]. These models along with the knowledge of system architecture can be used to analyze the system for energy and power consumption. There are two main sources of power dissipation - static power and dynamic power. These are explained in the following subsections.

2.2.1 Dynamic Power Dissipation

Dynamic power is the power dissipated by the device when it is active i.e., when signals are switching. The two main sources of dynamic power are switching power and power due to short circuit current.

Switching power is the power dissipated in the transistors during charging and discharging of the load capacitor, as shown in Figure 2.3. Switching power can be given



(a) Switching currents in an inverter

(b) Short-circuit currents in an inverter

Figure 2.3: Switching and short-circuit currents in an inverter [2].

by the following expression.

$$P_{switching} = C_{eff}.V_{DD}^2.f_{clock} \tag{2.1}$$
$$where, \quad C_{eff} = \alpha.C_L \tag{2.2}$$

Where, α is the switching activity, C_L is the load capacitance, V_{DD} is the supply voltage and f_{clock} is the clock frequency.

An input signal always has a finite slope which causes a direct current path between supply and ground for a short period of time during switching. During this period, the PMOS and the NMOS conduct simultaneously. This short circuit current also results in power dissipation, as shown in Figure 2.3(b). Power dissipation due to short-circuit current can be given by the following expression.

$$P_{sc} = t_{sc}.V_{DD}.I_{peak}.f_{clock} \tag{2.3}$$

Where, t_{sc} is the time duration of the short circuit current and I_{peak} is the internal switching current i.e., the sum of short-circuit current and the current required to change the internal capacitance.

The dynamic power can be given by the following expression.

$$P_{dyn} = P_{switching} + P_{sc} \tag{2.4}$$
$$= (C_{eff}.V_{DD}^2.f_{clock}) + (t_{sc}.V_{DD}.I_{peak}.f_{clock}) \tag{2.5}$$

As long as the short-circuit time(t_{sc}) of the input signal is kept short, switching power dominates in the above equation. Hence dynamic power can be given by the following equation.

$$P_{dyn} \approx C_{eff}.V_{DD}^2.f_{clock} \tag{2.6}$$

The equation shows that dynamic power has direct dependence on f_{clock} and quadratic dependence on V_{DD}. Reducing these parameters help reducing dynamic power.

2.2.2 Static Power Dissipation

The static (or leakage) power dissipation in a digital CMOS circuit is associated with maintaining the logic values of internal circuit nodes between the switching events i.e., when signals hold fixed values [16]. It is expressed by the following relationship [16]:

$$P_{static} = I_{static} V_{DD} \tag{2.7}$$

where I$_{static}$ is the current that flows between the supply rails in the absence of switching activity. Various leakage currents are shown in Figure 2.4. Main sources of leakage current in a CMOS gate are sub-threshold leakage current (I$_{sub}$), gate leakage current and reverse bias leakage current (drain junction leakage) [2].

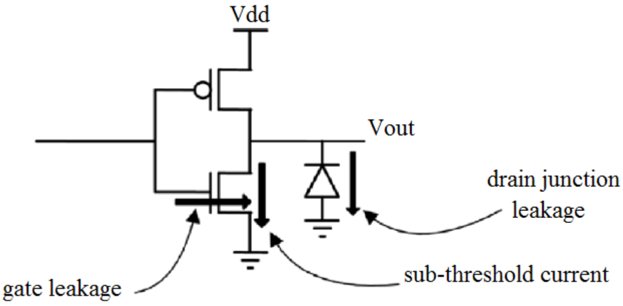


Figure 2.4: Leakage currents through an inverter [2].

Sub-threshold leakage current occurs when a gate is not turned off completely. Its approximate value can be given by the following equation [2]:

$$I_{sub} = \mu C_{ox} V_{th}^2 \frac{W}{L} . e^{\frac{V_{GS}-V_T}{nV_{th}}} \tag{2.8}$$

Where μ is the carrier mobility, C$_{ox}$ is the gate capacitance, V$_{th}$ is the thermal voltage, W and L are the width and the length of the transistor respectively, V$_{GS}$ is the gate-source voltage, V$_T$ is the threshold voltage and parameter n is the function of device fabrication process. Thermal threshold is denoted by kT/e where k is Boltzmann constant, T is Temperature and e is the electron charge. The equation shows leakage current increases quadratically with temperature and exponentially with difference between V$_{GS}$ and V$_T$ and putting a constraint on reduction in V$_T$. This constraint leads to a conflict between dynamic and static power which is discussed in the next subsection. Few of the approaches to minimize leakage current are to use multiple-V$_T$ cells to build circuits and shutting down the power supply to the block when not active.

2.2.3 Conflict Between Dynamic and Static Power Dissipation

Equation 2.6 suggests that reducing V$_{DD}$ can help achieving a lower dynamic power. But, reducing V$_{DD}$(hence V$_{GS}$) has a negative impact on the ON(drive) current of the transistor, thus reducing the speed of the device. A simple approximation of the ON current is presented in the following equation [2].

$$I_{DS} = \mu C_{ox} \frac{W}{L} . \frac{(V_{GS} - V_T)^2}{2} \tag{2.9}$$

Above equation shows the quadratic dependence of I$_{DS}$ on (V$_{GS}$ - V$_T$). Hence, to keep up with performance, V$_T$ should also be reduced when V$_{GS}$ is reduced. But, in Equation 2.8 shows that I$_{sub}$ increases exponentially with (V$_{GS}$ - V$_T$). It is important to keep (V$_{GS}$ - V$_T$) high for good performance but low for less static power dissipation. This leads to a trade-off and hence putting a limit on the supply and threshold voltage.

2.2.4 Total Power Dissipation

Total power dissipation is the sum of static and dynamic power dissipation. As discussed above, static current can be taken care of during physical implementation of the circuits (e.g., use of Multi-V$_T$ cells). Increasing temperature of the device increases the leakage current hence monitoring and controlling temperature is important to control the leakage current and static power. Each PE in a SoC or MPSoC has a specific critical (threshold) temperature exceeding which can lead to temperature-related reliability issues such as time-dependent dielectric breakdown. In this thesis, static power is not addressed directly, but total power is controlled by taking dynamic power and temperature into account. Temperature is monitored and controlled to limit a PE's temperature below its critical temperature value. Dynamic power is controlled by monitoring the activity rate of the core.

2.3 Relation Between Power and Temperature

High operating temperature of a PE has a significant impact on its design [17]. Carrier mobility degrades at higher temperatures making a transistor slower [18]. Resistivity of the interconnect metal is higher at higher temperatures, causing longer interconnect RC delays and degradation in performance. Also, it was seen in Section 2.2, leakage power depends exponentially on operating temperature. Increasing the temperature of a device exponentially decreases its lifetime [17] making a significant impact on its reliability. Hence, it is very important to keep devices below critical temperature making it is an important goal for chip designers.

As discussed in previous sections, technology miniaturization has an unfortunate side effect of increasing power densities which translates into increased heat dissipation. A PE consumes electrical energy and dissipates a part of it during switching of the devices in the form of heat due to the impedance of the electronic circuits. At system-level, temperature of a PE can be controlled by controlling its dynamic power.

Other sources of heat generation in VLSI systems are the leakage energy inside the transistors and electrical current flows through on-chip metal interconnects that connect the transistors. CMOS transistors are not ideal switches. Despite being OFF, they still conduct some amount of current. This leakage current moves charges between power supply and ground, thus drawing energy from the power supply. This energy is wasted without performing useful computation and is dissipated as heat through the resistance in their flow path. In addition to the heat generated inside the transistors, heat is also dissipated when electrical current flows through on-chip metal interconnects that connect the transistors. This is because the interconnects are not ideal electrical conductors and have finite amount of resistance.

In summary, heat is generated from the silicon active surface due to two factors active switching and leakage. The power consumed by the IC is dissipated in the form of heat in the transistors and interconnects, and are eventually removed to the environment by heat transfer.

Energy and power are related by the following equation.

$$Energy = Power * Time \tag{2.10}$$

As P (power), is the rate of energy consumption and Q, is the rate of heat (energy) dissipation, it can be said that

$$Q = P \tag{2.11}$$

Heat transfer equation, as given in [19] for a volumetric system is shown in Equation 2.12 where C_{th} is thermal capacity of the material, R_{th} is the thermal resistance of the material and ΔT is the change in temperature of the control volume. The first term on the left hand side in this equation represents the amount of heat stored in the volume and second term represents the loss of heat from the volume due to heat conduction. The term on the right hand side is a translation of dissipated power as seen in Equation 2.11, hence the equation shows the relation between change in temperature of a volume and the power dissipation. This relation will be explored further in the next chapter where thermal model for the 3D stack will be derived for the power management control.

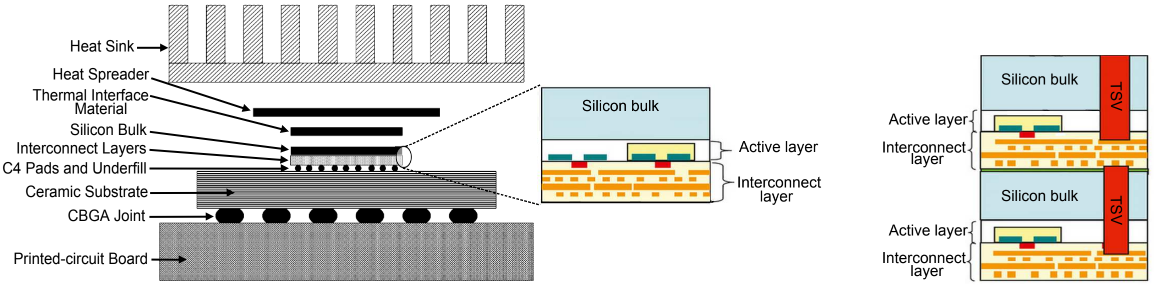
$$C_{th}\frac{dT}{dt} - \frac{\Delta T}{R_{th}} = Q \tag{2.12}$$

2.4 Thermal Modeling

Since power dissipated and resulting temperature are co-related[1], they should be handled simultaneously. To be able to handle temperature effects, an accurate thermal model is necessary. This thesis aims at developing a power management scheme, hence a previously developed thermal simulator is used to develop thermal model. This section discusses previous and related work where thermal models have been developed for use at architecture-level.

A stacked chip package is illustrated in Figure 2.5(a). The heat is generated in the active layer between silicon bulk and interconnects. There are two heat removal paths for such a package model. The primary heat removal path consists of the silicon bulk, thermal interface material (TIM), heat spreader and the heat sink. A significant amount of heat is also dissipated through the secondary heat removal path, i.e., across interconnect layers, pads and to the printed-circuit board (PCB). 3D-IC designs are similarly

[1] Power is dissipated in form of heat, rising the temperature of the device. While rise in temperature increases static power dissipation in the device further increasing the temperature of the device.



(a) Stacked layers in a typical ceramic ball grid array (CBGA) package [17,20].

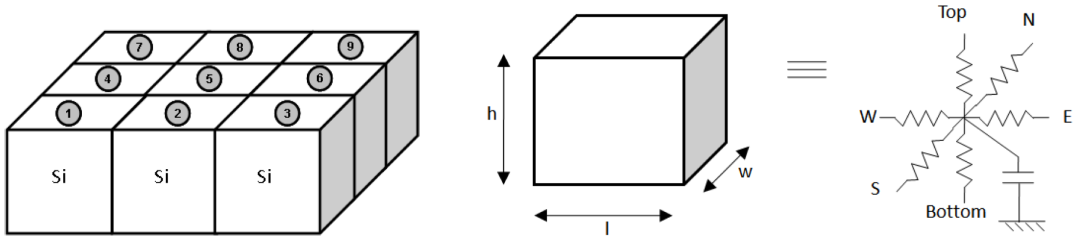
(b) multiple tiers in 3D-IC design [20].

Figure 2.5: 3D stacked layered structure of a chip package.

stacked-layer structures with multiple silicon bulk, active layer and interconnect layers. Figure 2.5(b) shows a 3D-IC design with two tiers, hence having two active layers. Since heat in generated in active and interconnect layer, as the number of tiers increase, generated heat also increases.

There has been a considerable amount of research in developing compact thermal models for 3D ICs. Works [17] and [5, 21] have developed simulation tools HOTSPOT and 3D-ICE respectively to model temperatures on a chip. Both the works use finite-difference based methods.

The models are generated by considering that each layer is divided into "thermal cells" as shown in Figure 2.6(a). Each thermal cell of length l, width w and height h, can be modeled as a node containing six resistances representing heat conduction in all six directions, and a capacitance representing heat storage in the cell as shown in Figure 2.6(b).

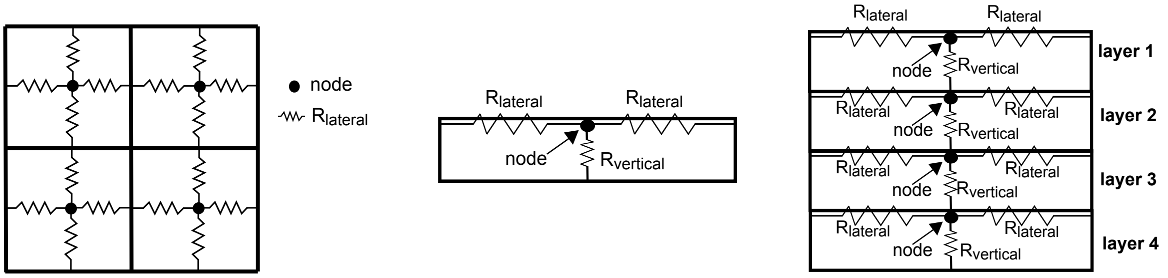


(a) Discretization of a single layer of silicon [5]

(b) equivalent circuit of a single thermal cell [5]

Figure 2.6: Discretization of a single layer of silicon into thermal cells and equivalent circuit of a single thermal cell.

Figure 2.7(a) shows the top view of a large silicon layer divided into 4 thermal cells. Node is considered to be the heat source in a thermal cell. The dimension of these cells determine the accuracy of the resulting thermal model. Smaller the size of the cell,



(a) A large silicon layer divided into 4 thermal cells(top view)

(b) Side view of a thermal cell [17]

(c) Side view of thermal cells on multiple layers

Figure 2.7: Top and side view of partitioned thermal cells on a silicon layer showing lateral and vertical thermal resistances.

more accurate it is. Figure 2.7(b) shows the side of a layer where the lateral resistance (R$_{lateral}$) is the thermal resistance on the same layer between adjacent thermal cells, whereas, vertical resistance (R$_{vertical}$) is the thermal resistance between two thermal cells on adjacent layers. Since 2D-IC has only one pair of active and interconnect layer, therefore only two layers have heat sources. While in 3D-IC, multiple active and interconnect layers provide multiple heat sources in the vertical direction. For example, in Figure 2.7(c), if a heat sink is assumed to be on top of the stack, thermal resistance between a node on layer 4 and the heat sink is more than the thermal resistance between a node on layer 1 and heat sink. This results in a lower transfer of heat from a deeper layer to the heat sink.

The conductance of each thermal resistance and the capacitance of a cell as given in [5, 17] are as follows:

$$\begin{aligned}
g_{top/bottom} &= k_{Si}.\frac{l.w}{(h/2)} \\
g_{north/south} &= k_{Si}.\frac{l.h}{(w/2)} \\
g_{east/west} &= k_{Si}.\frac{w.h}{(l/2)} \\
c_{cell} &= C_{vSi}.(l.w.h)
\end{aligned} \tag{2.13}$$

From these equations, it can be seen that conductance depends on the area of cross-section in the direction of heat flow. The cross-sectional area for heat flow in the vertical direction*(w*l)* for a PE will be much larger than that for the flow in lateral direction*(w*h or l*h)*. Therefore, the heat transfer in the vertical direction is more significant than that in the lateral direction. Since dies in a 3D IC are stacked on top of each other, the heat flow from a PE on one die will strongly affect the temperature of the section of a die just above and/or below it. And, all PEs in a 3D stack are thermally connected, hence the temperature of one PE affects the temperature of all other PEs in the stack.

From the above discussion, two important deductions can me made:

- Conductance in the vertical direction is more than that in the lateral direction.
- Farther the heat source from the heat sink, lesser the heat transferred.

The two deductions based on the difference in the thermal models of 3D-IC and 2D-IC show the importance of considering an appropriate thermal model. Both simulators are derived considering stacking of layers and heat transfer in all directions. But, HOTSPOT thermal simulator does not directly address 3D ICs, whereas, 3D-ICE simulator is developed to support multi-processor 3D ICs. Hence 3D-ICE thermal simulator is used for the purpose of thermal simulations and generating conductance matrix for the power management control in this thesis.

2.5 Power Management Schemes

So far, importance of power management and simultaneous consideration of thermal management in 3D MPSoCs have been discussed. There are various power management schemes for MPSoCs at architecture-level that are currently used in many designs. This section briefly discusses these power management schemes, feasibility of extending such schemes to 3D MPSoCs and work to this thesis.

2.5.1 Voltage Island Partitioning

It was seen in Equation 2.6 that dynamic power is proportional to V_{DD}^2. Reducing V_{DD} in selected blocks can reduce power significantly. It was also seen in **??** that reducing V_{DD} can increase the delay through the gate making the device slower hence putting a constraint on minimum V_{DD}. But, the complete chip can be divided into blocks (islands) where each block operates on different supply voltages. Hence, each block has its independent supply voltage. Depending on the voltage reduction, power saving can be achieved with losses in performance. The chip is first divided into multiple small tiles, and then each tile is allocated to an island. Partitioning of the complete chip in islands in known as Voltage Island Partitioning and is shown in Figure 2.8 where 4 tiles are divided amongst 3 voltage islands. It is a widely practiced in 2D chips. For communication between these islands, level shifters are used which result in some power and area overhead. Island partitioning algorithms decide operating voltage level for each tile considering the performance losses, power energy relationship and overheads due to level shifters. More number of islands can achieve finer control over performance loss and power of each tile, but overheads due to level shifters introduces a trade-off between number of islands and power saving. An algorithm for voltage frequency island partitioning for 2D Networks-on-Chip (NoC) is proposed in [22]. It also shows that optimal number of islands for a 3X3, 4X4 and 5X5 mesh network is either 2 or 3. Increasing the number of islands beyond this optimal value does not result in further improvement of power due to overheads.

Voltage assignments to these islands can be static or dynamic. Static voltage assignment assigns a single, fixed voltage level to each island. Figure 2.8 is an example of

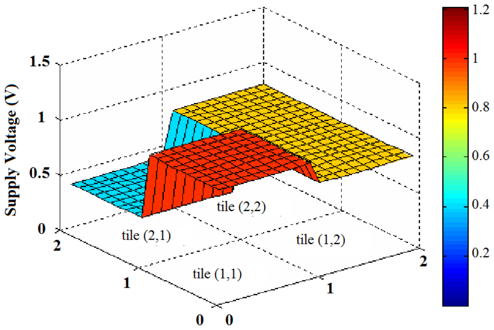


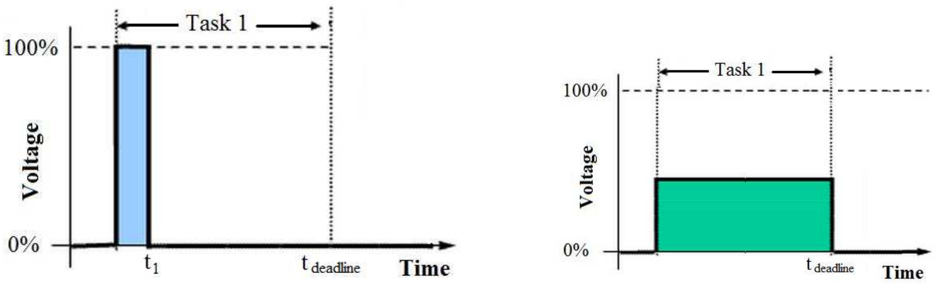
Figure 2.8: Voltage island partitioning for a 2X2 network with static voltage assignment [22].

static voltage assignment. Whereas, in dynamic voltage assignment, the islands are allowed to operate at multiple voltages over time. This approach of allowing voltage to scale dynamically is known as Dynamic Voltage Scaling(DVS) and is discussed later in this section. Works [23] and [24] have applied voltage island partitioning on 3D ICs. [24] compares the 2D Voltage Frequency Islands(VFI) and 3D VFI. Since the work uses VFI, each PE in an island operates at same voltage as well as frequency. Whereas, if the two components are made independent, better flexibility can be provided to each PE. Work [24] proposed a post-placement multiple supply voltage assignment method for partitioning voltage islands. The work has considered an example of 3-tiers, and have divided the islands with static voltage assignment such that each island is a 3D block, each comprising sections of each tier.Voltage islands may be effective in case of 3D MPSoCs when groups of PEs run similar workload. Dynamic voltage islands are considered in this thesis.

2.5.2 Dynamic Voltage Scaling (DVS)

As discussed, lower V$_{DD}$ reduces power dissipation with some degradation in performance. But, if a PE is allowed to adjust its V$_{DD}$ dynamically depending on the performance requirement, the degradation in performance can then be maintained within the desirable limits. Figure 2.9(a) shows an example where deadline of a PE's task is time t$_{deadline}$, whereas task is completed by time t$_1$ when the PE is operating at maximum V$_{DD}$. If the supply voltage is scaled down as shown in Figure 2.9(b), the task not only completes within the allocated time but also reduces the power dissipation. The advantage of using DVS over fixed voltage islands can be seen in Figure 2.10 where voltage is scaled so as to meet the deadlines and PE is not forced to operate at one operating voltage.

Various algorithms have been used over years to intelligently monitor the processor's utilization and activity, to scale the supply voltage accordingly. DVS can be implemented to each PE independently, or on islands of PE, depending on the target application of the MPSoC. Voltage islands introduce an overhead due to level shifters necessary for communication between islands and so does DVS. Also, enabling DVS requires additional circuitry to allow the islands or individual PEs to have multiple



(a) task 1 having a deadline of $t_{deadline}$, operating at 100% V_{DD} and completes the task in time t_1

(b) V_{DD} is scaled such that the task utilizes the completes allocated time i.e., $t_{deadline}$ resulting in power saving

Figure 2.9: Dynamic Voltage Scaling to achieve power reduction.

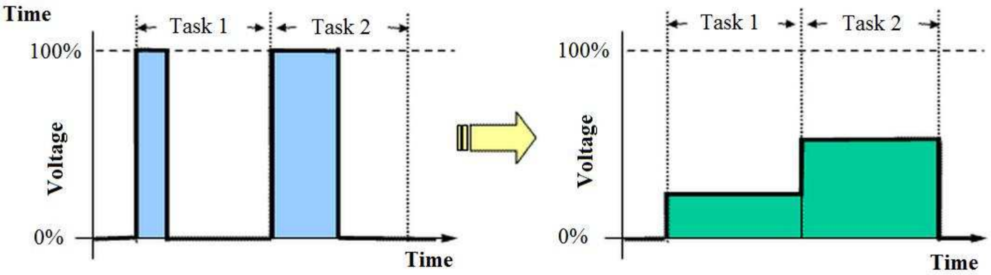


Figure 2.10: Example showing efficient utilization of allocated time by dynamically scaling voltage to achieve power reduction [25].

supply voltages, adding to the overheads. This can be done in two ways.

- By having fixed power grids for the supported voltages and allowing the PEs to select the appropriate supply line using switches; or
- Incorporating voltage converters to generate the required voltages dynamically.

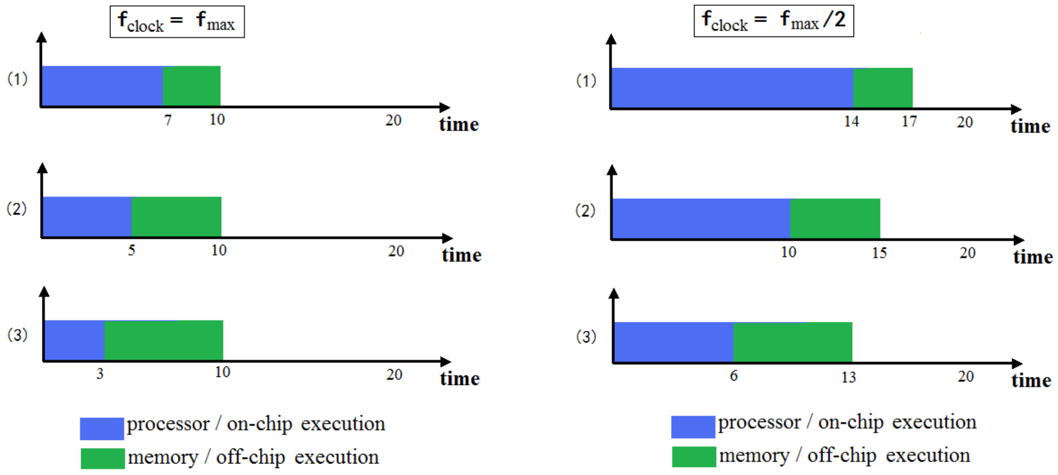
Since former approach requires fixed grid for supported voltage levels, there is a constraint on number of supported voltage levels. Whereas, voltage converters can provide more number of operating voltages. Design issues with voltage converters/regulators will be discussed later in this section. In 90nm and below nodes, there is not sufficient headroom to achieve desired power saving using DVS [2]. Hence, addition power saving by scaling f_{clock} (Equation 2.6) is explored.

2.5.3 Dynamic Frequency Scaling (DFS)

As seen in Equation 2.6, power also depends directly on frequency. But, reducing the frequency leads to increased execution time which relates to performance and energy. The average power value of the PE reduces but the energy saving depends on the type of operation, i.e., memory bound operation or processor bound operation. A memory bound operation spends majority of its execution time in the memory, while a processor bound operation spends majority of its execution time in the processor. This can be explained with the help of Equation 2.14. Energy is the integral of the power dissipation over execution time which gives the following relationship:

$$P \propto V_{DD}^2.f_{clock} \qquad and \qquad E \propto V_{DD}^2.f_{clock}.T_{exe} \tag{2.14}$$

where T_{exe} is the execution time. For example, if f_{clock} is reduced to half, the energy saving depends on the product of f_{clock} and T_{exe}. Reducing f_{clock} to half does not mean that t_{exe} would double because t_{exe} depends on the type of operation i.e., memory bound or processor bound. This can be explained further with the help of Figure 2.11 where three tasks are shown. In Figure 2.11(a), f_{clock} is set to the maximum frequency of the PE and the three tasks execute in 10 time units where task (1) spends 70% (7 units) in processor execution and rest 30% (3 units) in memory execution. Whereas, when the same task in run with f_{clock} set to $\mathrm{f}_{max}/2$ i.e., half of previous case, the execution time does not double as can be seen in Figure 2.11(b). This is because only the processor execution time will get doubled and the memory execution time remains the same. So is the case with task (2) and (3). As the task becomes more memory bound, the exectution time inside the processor reduces, hence achieving more energy saving. Performance penalty can be given by the following equation:



(a) Three tasks operating at f_{max} and with equal execution time.

(b) Same set of tasks running at $f_{max}/2$. Final execution time depending on whether the execution is processor bound or memory bound.

Figure 2.11: Dynamic Frequency Scaling to achieve power reduction.

$$PerformancePenalty(\%) = \frac{increase\ in\ execution\ time}{execution\ time\ with\ maximum\ frequency} * 100\% \tag{2.15}$$

DFS is one of the considered approaches in this thesis.

2.5.4 Dynamic Voltage and Frequency Scaling (DVFS)

Reducing the operating frequency in case of DFS also allows reduction in supply voltage. Reducing the supply voltage in combination with frequency is known as Dynamic Voltage and Frequency Scaling (DVFS). As the speed of a device depends on its operating voltage level, this introduce a constraint on maximum frequency for an operating voltage level. A major requirement for implementing an effective DVFS technique is to accurately predict the time-varying processor workload for a given computational task. As seen in Figure 2.11 and Equation 2.15, more energy is saved achieving underperformance when processor utilization is less. Therefore, monitoring a processor workload and adjusting the operating frequency and voltage based on the its utilization factor can achieve significant power saving. That is, decrease or increase the frequency and voltage when the processor utilization is low or high, respectively. DVFS is a popular power management method and is also used as a thermal management scheme to control on-chip temperatures [26] due to power-temperature relationship. Thus DVFS has been used in this thesis to achieve a temperature constrained power management scheme along with DFS and voltage island partitioning.

2.6 Related Work

Various power management schemes were discussed in the previous section. DVFS, DFS and voltage island partitioning are used in this thesis to build a temperature constraint power management control for 3D-MPSoC. Voltage and frequency scaling can be done on individual PEs or on islands. Work [27] compares per-core[2] DVFS and chip-wide[3] DVFS. The work shows that systems running heterogeneous workloads can benefit from per-core DVFS schemes. As various PEs running different workloads have different performance requirements allowing these PEs to operate at different voltages and/or frequencies achieves higher power saving. This is due to the fact that the PE with lower workload is allowed to operate at a lower operating voltage and/or frequency, independent of the PE with high performance requirement. Also, it was shown that the applications that are highly processor-bound offer fewer frequency-scaling opportunities and hence not much difference in power reduction can be seen in the per-core DVFS when compared with chip-wide DVFS. This is due to the high instruction execution time spent inside a PE.An intermediate case would be to have voltage/frequency islands where each island can have several PEs that operate on same voltage and/or frequency. Similar approach is considered in [12] where chip is divided into Voltage Frequency Islands (VFIs) and cores in a VFI operate at same voltage and frequency. This work also compares the energy and power reduction achieved with different VFI granularities. The work concludes that, increasing the VFI granularity can offer better flexibility in choosing voltage and frequency levels, but does not necessarily translate into better energy-efficiency. Extending the approach of island partitioning to the 3D IC can be done in two ways. First, by considering 2D islands on dies of a stack. Second, by

[2] per-core DVFS refers to the individual setting of the voltage and the frequency levels for the PEs

[3] chip-wide DVFS refers to the single global setting of the voltage and the frequency levels for the complete chip

making islands 3D, i.e., allowing PEs from various dies to form an island. 3D islands can prove to be efficient as the PEs will not only have similar performance requirements but will also be thermally related. Work [24] has proposed a post-placement island partitioning and voltage assignment method for 3D ICs by considering delay caused by power reduction, timing slack, temperature analysis and power density. Islands operating at various supply voltages are used in this thesis to study their effectiveness and to draw a comparison between various approaches.

In [28], a temperature constrained power management scheme for a Chip Multiprocessors (CMPs) using DVFS is proposed but it addresses PEs in a 2D-IC. The temperatures of PEs are considered independently. Since PEs in a 3D stack have strong thermal relation with each other, a 2D temperature constrained power management scheme such as [28] can not be directly extended to 3D chips. Implementation of such a scheme on 3D-IC is studied in this thesis and is compared with the proposed approach. [28] also includes an on-line model estimator for systems with heterogeneous workloads, which is not considered here. [29] analyzes the thermal profile of a 3D stacked MPSoC and proposes an active cool solution using inter-tier liquid cooling along with a DVFS scheme. Sabry et. al. [29] also state that management techniques with passive control elements alone, like DVFS, are incapable of reducing temperature of the 3D stacked MPSoC systems efficiently. They also mention that increased power densities, number of tiers and number of cores increase, raise the temperature of the cores to extreme values in 3D MPSoCs. This results in severe restrictions in high-performance 3D MPSoC design making other cooling methods for a 3D MPSoC important. These may include inter-tier cooling suggested in [29] or thermal TSVs suggested in [30] or thread scheduling along with schemes like DVFS as proposed in [13]. Nevertheless, considering temperature constraints in power management schemes can provide a support to the thermal management unit for such chips and also ensure that temperature of a PE never crosses the critical limits.

System Modeling 3

This chapter describes the system modeling for the power management scheme. First, an overview is presented where importance of power budget, thermal management techniques and DVFS are described. Next, the voltage island and per-core DVFS approaches that are considered in this thesis are explained. Further, the used control strategy is presented with the required system modeling and detailed thermal model.

3.1 Overview

3.1.1 Importance of Power Budget

Power budgets are employed to ensure that actual power consumption of the chip (or constituent logic block) never exceeds the desired fixed value. Operating PEs in an MPSoC at higher voltage or frequency achieves better performance but at the cost of higher power dissipation. These can also lead to unacceptable temperatures on the chip. These thermal and power dissipation problems can be reduced by setting a power budget to the complete chip or on the constituent logic blocks. These budgets restrict the maximum power dissipation of the chip or the logic block at the cost of performance. Excessively low power budgets would lead to higher performance losses.

3.1.2 Thermal Management Techniques

Thermal management techniques can be either reactive or proactive. While the former reacts to the current temperature value of the target PE, latter predicts the future temperature value and acts accordingly. Proactive techniques are usually accompanied by task scheduling where the prior information of temperature values are used to assign tasks accordingly. This helps in keeping PEs with higher predicted temperature less active. DVFS schemes use reactive methods in order to serve performance requirements of the PEs and react to the temperature values when necessary. The power management control does not have the information of the tasks being assigned to the PEs and the execution time of an application. Considering a proactive method to predict temperature would require an additional capability of predicting the future temperatures, information of the tasks being assigned along with the additional memory to record previous temperature values. Thus, this work uses a reactive method to keep the temperatures of PEs below the critical values. Such a power management scheme provides an aid to the actual thermal management scheme which may be necessary in a 3D chip.

3.1.3 DVFS

In 2D ICs, DVFS is achieved by monitoring the workload of the PEs. If the utilization of a PE (or activity) is high, higher voltage and frequency levels are assigned to it. Opposite is done when the utilization is lower. When a power budget is introduced, the algorithm tries to keep the total chip power below the power budget value by adjusting the voltage/frequency (V/F) of the PEs. When the total chip power falls below the budget value, the V/F levels are increased whereas opposite is done when chip power crosses the budget value. Since density of devices on a chip is increasing, temperature has become a major concern in 2D chips as well. For DVFS with temperature constraints in 2D ICs , the temperature of each PE is monitored independently [28]. The effect of temperature on a PE due to another PE is ignored. This is largely accepted in 2D ICs as heat flow in lateral direction is negligible. But, in case of a 3D IC, temperatures of PEs in a stack are highly interdependent, not only in the vertical direction but also in the lateral direction. Hence, monitoring the activity and the individual temperatures of PEs alone is insufficient. Other parameters should be included in the equation. In order to to have less performance losses in a PE, its utilization should be monitored while keeping the total chip power below a set budget value and temperature of PEs under critical temperature values. The temperature of a PE is primarily influenced by its power dissipation, its location within the stack, and in case of heterogeneous system, its area as well.

3.1.4 Approaches

Two power management approaches are studied and considered: per-core DVFS and DVFS on voltage islands.

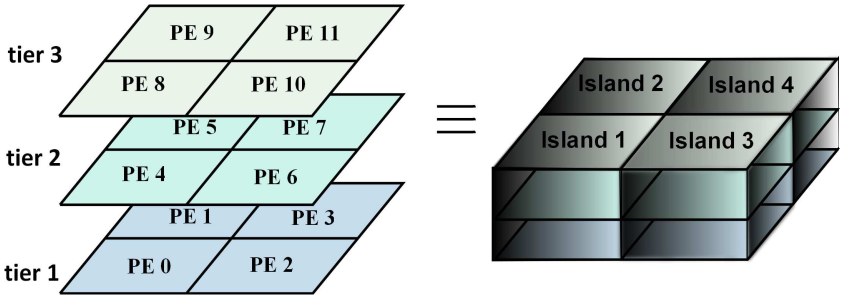


Figure 3.1: 3D voltage islands.

Voltage islands: A stack with PEs is partitioned into 4 islands, as shown in Figure 3.1. A voltage island partitioning and multiple supply voltage assignment technique is presented in [24] considering the delay caused by power reduction, timing slack, temperature analysis and power density. DVFS can be effective on islands if the PEs in an island have similar workloads giving enough opportunities for power saving whereas, if the performance requirements of PEs in an island are different, scaling of V/F levels may highly degrade the performance of active PEs. Voltage island partitioning highly

depends on the type of target application that would run on the PEs and the symmetry in their performance requirements. Special voltage island partitioning algorithms should be considered while assigning islands. In this thesis, the islands are partitioned to study the effectiveness of the 3D islands, therefore, to rule out complexities due to difference in performance requirements, PEs in an island are assumed to have same workloads. DVFS along with DFS is considered in the test case. The voltage-frequency (V/F) combinations can be represented as:
(V1,F1), (V1,F2), (V2,F3), (V2,F4), (V3,F5), (V3,F6).

(V1,F1) and (V3,F6) represent the lowest and the highest operating V/F level, respectively. Six frequencies are used paired with only three voltage levels. This is done in order to utilize the benefit of DFS. This proves to be efficient in thermal management as frequency scaling helps reducing peak power, hence reduces the temperature. Scaling only frequency may mean degradation in the performance without significant energy saving. However, as each voltage level allows two frequencies, in order to reduce the temperature of one PE, its frequency can be scaled down without changing frequency levels of other PEs in the island.

per-core DVFS: This is a special case of voltage islands where each island consists of only one PE. Increasing the granularity of voltage islands can offer better flexibility in choosing operating voltage and frequency levels. This becomes important in cases where PEs have different workloads. However, this comes at an overhead of additional level shifters and voltage converters. Per-core DVFS allows power management block to choose new operating voltage and frequency levels for individual PEs in order to meet temperature and power constraints. As individual PEs are scaled, the change in power density in most cases would be lesser than that in voltage islands. This may prove to be advantageous when PEs are scaled to higher V/F levels because change in temperature of a PE depends on change on power density. Six voltage-frequency combinations can be represented as:
(V1,F1), (V2,F2), (V3,F3), (V4,F4), (V5,F5), (V6,F6).

(V1,F1) and (V6,F6) represent the lowest and the highest operating V/F level, respectively. Each frequency level is coupled with a unique voltage level.

3.2 Control Loop and System Modeling

The control loop for power management scheme is shown in Figure 3.2. The Power Management Block (PMB) takes three inputs from the system to decide new DVFS levels for each PE. These inputs are:

1. *Activity factor (utilization) of each PE.* The activity factor (utilization) of each PE in the previous control period is made available to the PMB. This is assumed to be done by the performance monitor on each PE.

2. *Temperature of each PE.* Temperature sensor on each PE provides the PMB with the current temperature of each PE.

3. *Total chip power.* A power monitor (e.g. power measurement circuit with the power supply circuit) provides the average chip power to the PMB at certain time intervals.

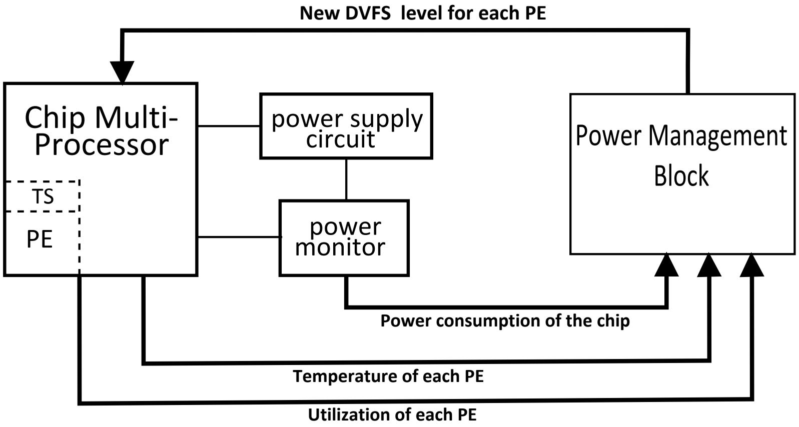


Figure 3.2: Control loop for power management scheme.

In order to design an effective control, it is important to model the dynamics of the controlled system, i.e., the relation between controlled variable and manipulated variable. The manipulated variable is the operating V/F level while the controlled variables are power and temperature. Hence, the relation between power and V/F levels, and the relation between temperature and V/F levels needs to be modeled.

3.2.1 Relation between Power and V/F Levels

DVFS can allow cubic reductions in power density relative to performance loss for each PE in a MPSoC [31]. However, cubic power model may lead to large runtime overhead and high complexity for power management scheme design. However, real MPSoC usually provide a limited DVFS range only, and within this small range, [32, 33] have shown that the relationship between power and DVFS level can be approximated with a linear function. Therefore, the power dissipation of a PE is modeled as

$$P = A * V^2 * F + B \tag{3.1}$$

where A and B are constants, P is power, V and F are voltage and frequency corresponding to a DVFS level. This equation can be looked upon as total power equation where first term denoted the dynamic power while the second term denotes the static power. The value of A varies for different PEs according to the workload as it depends on activity. It can be represented by a generalized value for an intended workload. To remove the constant term and develop a dynamic model equation, the difference equation can be considered.

$$\Delta P = A * \Delta(V^2 * F) \tag{3.2}$$

How this value of A is achieved with be described in Chapter 5. For an intended application, power values are obtained for each set of V/F values. For systems with PEs running heterogeneous workloads, the value of A should be corrected during runtime with feedback from the system. [28] presents method for this on-line correction. For simplicity, this thesis assumes that target application of the PE is known and A is considered to be a static parameter. Since V^2F for each DVFS level is known, ΔP values can be computed.

3.2.2 Relation between Temperature and V/F Levels

Thermal conductance and capacitance equations are given in Equation 2.13 which depend on the dimensional parameters l, w and h. Therefore, thermal conductance between two PEs can be calculated using these equations. However,

- To have a direct relation between temperature and V/F level, these values are not sufficient. Additional information in needed which lead to "effective thermal resistance" that denotes direct relation between the two parameters.
- The thermal resistance or conductance is due to the overlapping area of the PEs. For the PEs that are not adjacent, it is difficult to determine thermal relation between them.
- Also, when the size of PEs is not similar or they are not completely overlapping, it becomes difficult to compute thermal resistance values between them even if they are adjacent.
- Thermal resistance or conductance between PEs in a multiple die stack is cumbersome but is important to be considered.

To deal with these issues, an alternate method to derive thermal relation is required. In this section, the details of effective thermal resistance, and how it is derived using a thermal simulator 3D-ICE is explained. In order to determine relation between temperature and V/F level, it is important to understand the detailed thermal model of 3D ICs. Therefore, details of thermal model are discussed first along with the need of complete thermal model.

Thermal models can be broadly classified as Detailed Thermal Model (DTM) and Compact Thermal Model (CTM). A DTM attempts to represent the physical geometry of a package as accurately as possible. A fully developed DTM provides the best prediction accuracy for all application environments. However, the computational efficiency of DTMs is low and are not suitable for use in system-level design. A CTM on the other hand takes a detailed model and extracts an abstract representation that is still able to preserve accuracy in predicting the temperatures at key points in the package. Most CTM approaches use a thermal resistor network to construct the model, analogous to an electrical network that follows Ohm's law. Transient CTM have also been developed that use thermal resistor as well as capacitor to model the steady-state as well as transient temperature response in the IC, analogous to electrical RC networks. A similar thermal model will be derived in this chapter but at at a higher level for the use in power management scheme.

3.2.2.1 Heat Transfer theory

The heat generated in an integrated circuit must be removed or transferred to the ambient environment to avoid its accumulation. The heat can be transferred in three different ways − conduction, convection and radiation. Conduction is the phenomenon of heat transfer in solids and is the major heat transfer mode in a chip; while, convection refers to heat transfer between a solid surface and a moving fluid, and radiation refers to heat transfer via electromagnetic waves. Heat diffusion equation (Equation 3.3) is a general representation of time dependent heat conduction [19].

$$\rho c_p \frac{\partial T(x,y,z,t)}{\partial t} \ = \ \nabla.[k(x,y,z,T)\nabla T(x,y,z,t)] \ + \ g(x,y,z,t) \tag{3.3}$$

where ρ is the density of the material (kgm^{-3}), and g is the volume power density of the heat source(s) (Wm^{-3}), c_p is the specific heat (J$kg^{-1}C^{-1}$). x, y and z represent directions, T denotes temperature and t represents time. While thermal conductivity k is a function of location and temperature, it can be assumed to be isotropic and temperature independent for the materials and the temperature ranges that are considered here [19].

Thermal models first convert this Partial Differential Equation (PDE) into an Ordinary Difference Equation (ODE) which is then numerically integrated using methods like Euler and Runge-Kutta. The conversion to ODE is an important step which also determines the accuracy of the final model. Converting to ODE using Finite Element Method (FEM) can handle complicated boundaries and geometries with relative ease, whereas Finite Difference Method (FDM) converts the PDE to ODE by applying finite difference approximation. Quality of approximation between grind points in poor as compared to FEM but provides the benefit of good computational efficiency and ease in implementation. CTM like HOTSPOT and 3D-ICE are both based on FDM and allow modeling of temperatures in 3D stacked ICs. 3D-ICE is chosen for the purpose of thermal simulations and thermal modeling.

Steady state analysis : For the steady-state case, the (∂T/∂t) term in Equation 3.3 becomes zero. At steady state, the one-dimensional form of the heat diffusion equation reduces to Equation 3.4:

$$q \ = \ -k\frac{dT}{dx} \tag{3.4}$$

where q is the heat flux (in Wm^{-2}), k is the thermal conductivity of the material (in W$m^{-1}K^{-1}$) and *(dT/dx)* represent temperature gradient within a small distance dx.The equation shows that the heat flux, q (i.e., the flow of heat per unit area and per unit time), at a point in a medium is directly proportional to the temperature gradient at that point. The minus sign indicates that heat flows in the direction of decreasing temperature. If $q = Q/A$, where Q is the heat transfer rate, A is the heat conducting area, and L is the length of a material (integrating dx), then Equation 3.4 becomes:

$$Q = -kA\frac{T_2 - T_1}{L} \tag{3.5}$$

If thermal resistance $R_{th} = (T_1 - T_2)/$Q,

$$R_{th} = \frac{(T_1 - T_2)}{Q} = \frac{1}{k}\frac{L}{A} \tag{3.6}$$

This equation shows resemblance with Ohm's law in electrical circuit theory.

Transient analysis : If both g and k in Equation 3.3 are assumed to be constant, in one-dimensional form of this equation, integrating both sides by the integral variable x from 0 to L (the length of the material),

$$(\rho c_p AL)\frac{dT(t)}{dt} = kA\frac{\Delta T(t)}{L} + Q \tag{3.7}$$

The heat flux q can be denoted by Q/A and further by $g * L$; where g is the volume power density of the heat source(s). The first term of the right-hand side in Equation 3.7 represents heat transferred through the thermal resistance R_{th} (similar to that in Equation 3.6) where $\Delta T = (T_2 - T_1)$. Moving this term to the other side of the equation,

$$C_{th}\frac{dT(t)}{dt} + \frac{T_1 - T_2}{R_{th}} = Q \tag{3.8}$$

where $C_{th} = \rho c_p AL = c_p \rho V$ is defined as thermal capacitance and V is the volume of the material.

From the electrical circuit theories, it is known that C(dV(t)/dt) = $i_c(t)$, i.e., the current flow through an electrical capacitor equals the product of its capacitance (C) and the first derivative of the voltage difference (dV/dt) across it. This resembles the first term on the left-hand side of Equation 3.8. Thermal capacitance describes the heat absorbing capacity of a material. Equation 3.8 shows that the heat flowing through the thermal capacitance (the AC component) summed with the heat flowing through the thermal resistance (the DC component) equals the total heat flowing through the material. The equation also shows energy conservation in the system, as the sum of heat stored and heat conducted is equal to the heat generated. Table 3.1 summarizes the analogy between thermal and electrical parameters.

3.2.2.2 Thermal model of 3D IC

Figure 3.3 illustrates a 3D chip package with multiple PEs. The chip contains multiple vertically stacked silicon layers, each containing PEs and memory modules. One side of the chip is connected to the printed circuit board (PCB) through a substrate while

Thermal quantity	**unit**	**Electrical quantity**	**unit**
Q, Heat transfer rate, power	W	I, current	A
T, Temperature difference	K	V, Voltage difference	V
R_{th}, Thermal resistance	KW^{-1}	R, Electrical resistance	Ω
C_{th}, Thermal capacitance	JK^{-1}	C, Electrical capacitance	F

Table 3.1: Analogy between thermal and electrical parameters.

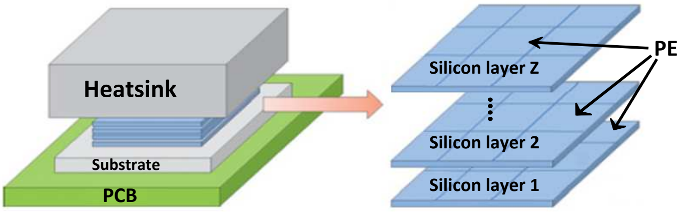
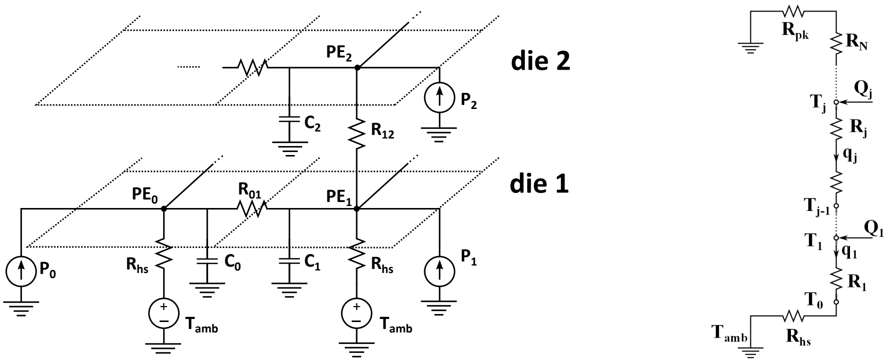


Figure 3.3: A 3D chip package with PEs on vertically stacked silicon layers.

the other side of the chip is attached to a heat sink where most heat is dissipated. Heat flow within a package can be modeled by the analogy between heat transfer and electric circuit phenomena in Resistor-Capacitor (RC) network.



(a) Simplified thermal model of a 3D multi-core system (adapted from [13]).

(b) 1D (vertical) thermal resistive network of a 3D IC [6].

Figure 3.4: Thermal model of a 3D chip.

Figure 3.4(a) illustrates a thermal model of a section of 3D chip in which each PE is represented with thermal model elements, i.e., a resistor that denotes thermal resistance between two PEs, a capacitor that denotes thermal capacitance of a PE, and current sources that denote heat transfer rate or power of a PE. The heat sink is shown at the bottom of the stack which connects to the bottom-most silicon layer (die 1) through a thermal resistance R_{hs}. When ambient is assumed to be at a fixed temperature, the connection between the stack and ambient can be denoted by a fixed voltage source. C_0, C_1, C_2 represent the thermal conductance of PE_0, PE_1, PE_2 respectively. And, R_{ij} represents thermal resistance between PE_i and PE_j. For a thermal model to be accurate, each thermal cell must be small enough so as for the temperature within it is to be assumed uniform. 3D-ICE [21] is used to get a fairly accurate fine-grained thermal model (considering small thermal cells), which is then used to derive a coarse-grain thermal model (thermal resistances between PEs).

Figure 3.4(b) [6] shows a 1-dimensional (vertical) resistive heat transfer model for a 3D stacked chip. Each active layer is represented by a node. T_j and Q_j represent temperature and heat generation at node j, respectively. R_j is the thermal resistance between node *j* and *j-1*. And, q_j represent the heat flow from node j to j-1. R_{pk} and R_{hs} represent the thermal resistance of package and heat sink, respectively. Since temperature at the package end is always greater than the temperature at the heat sink end, the heat flow q_j is shown in the direction from package towards the heat sink. This is true for a resistive model, but the model is incomplete without thermal capacitances.

Resistive model (ignoring thermal conductances) represents steady state condition and heat (analogous to current) will only flow towards the heat sink. Steady state condition would result in following equations for a thermal network shown in Figure 3.4(a)

$$\left.\begin{aligned} T_1 &= T_{amb} + (P_2 + P_1) * R_{hs} \ ; and \\ T_2 &= P_2 * R_{12} + T_1 \ = \ P_2 * R_{12} + T_{amb} + (P_2 + P_1) * R_{hs} \end{aligned}\right\} \tag{3.9}$$

where T_{amb} is the ambient temperature.

But, power dissipation of a PE is not always a steady state function, rather is a function of time. When power of a PE changes, it takes a certain amount of time to reach a steady state value. This results in an AC component that causes the heat to flow into the thermal capacitor. In such a case, the heat will not only flow towards the heat sink, but will also flow in the opposite direction, i.e., towards the package. The ratio of heat flowing in the two directions depends on the ratio of impedances seen in the two direction. In Equation 3.9, a part of P_1 flows towards heat sink via R_{hs} while the other part flows through R_{12}, depending on the R and C values (impedance). The heat flow in lateral direction is being ignored for now. Hence, Equation 3.9 can be rewritten as

$$\left.\begin{aligned} T_1 &= T_{amb} + (x.P_1 + P_2) * R_{hs} \ ; and \\ T_2 &= T_1 + (-y.P_1 + P_2).R_{12} \end{aligned}\right\} \tag{3.10}$$

where x+y = 1.

Minus sign indicates that heat is flowing in opposite direction. $x.P_1$ denote the heat flowing from node 1 towards heat sink, while $y.P_1$ is the current flowing from node 1 to package towards the PCB end.

This is illustrated in Figure 3.5. Since the thermal parameters are assumed to have time-invariant values, these impedances remain fixed. In Figure 3.5, Z_{1hs} is the impedance seen at node 1 in the direction of heat sink and Z_{1pk} is the impedance seen at node 1 towards the PCB end. This results in a fixed ratio of current flowing in different branches from a node. It should be noted that Z_{1pk} is not only a resultant of R_{12} and C_2, but its the total impedance seen in the direction towards PCB. This includes all the branches in vertical as well as lateral direction. This is further explained with the help of following equations.

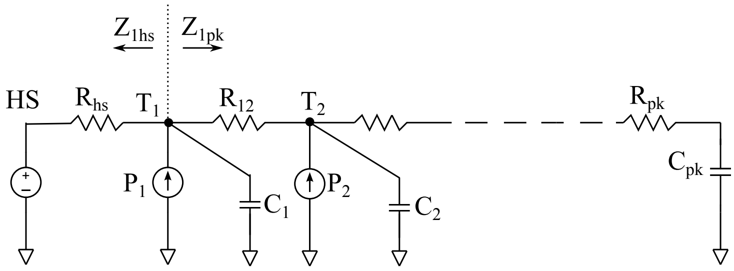


Figure 3.5: 1D thermal network including thermal capacitors.

Considering a case where P_1 changes by ΔP_1 and P_2 remains same. Change in temperature at node 1 and node 2 can be given by

$$\left.\begin{aligned}\Delta T_1 &= x * \Delta P_1 * R_{hs}\\ \Delta T_2 &= \Delta T_1 - y * \Delta P_1 * R_{12}\end{aligned}\right\} \tag{3.11}$$

Since $R_{vertical}$[1], i.e., R_{12} in this case, is very small, $\Delta T_2 \approx \Delta T_1$ (at steady state, i.e., ignoring capacitance). The difference between the two is a strong function of material properties, and the difference in temperature increases as $R_{vetical}$ increases. This becomes more cumbersome if an actual model is considered where a PE node is not only connected to the nodes above or below it, but also on the same plane via $R_{lateral}$[2]. $R_{lateral}$ is often ignored, but, this might be an optimist approach when the stack is deep. Since the conductivity of a die to the ambient decreases with the depth in a stack, the conductance in the lateral direction becomes prominent. The steady-state temperatures depend on the R values between nodes. Hence, using Equation 3.11, temperature change at a node i and node j due to change in power dissipation at node j can be given by Equation 3.12.

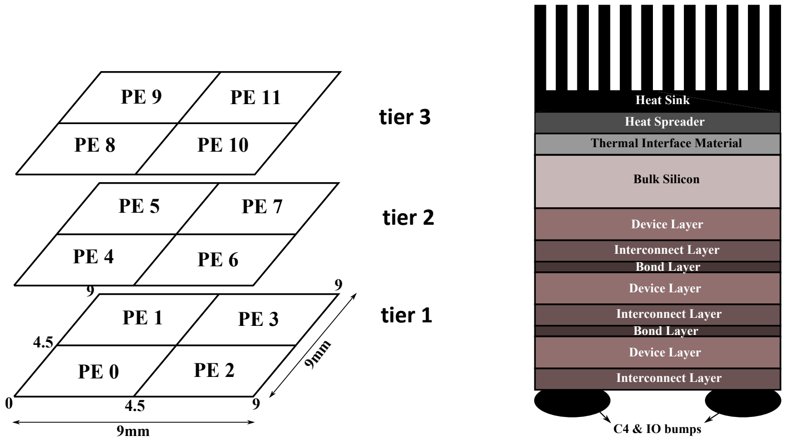
$$\left.\begin{aligned}\Delta T_j &= x * \Delta P_j * R_{jhs}\\ \Delta T_i &= \Delta T_j - y * \Delta P_j * R_{ij} = x * \Delta P_j * R_{jhs} - y * \Delta P_j * R_{ij}\\ \Delta T_i &= \Delta P_j (x * R_{jhs} - y * R_{ij})\end{aligned}\right\} \tag{3.12}$$

where R_{jhs} is the thermal resistance between node j and heat sink, and R_{ij} is the thermal resistance between node i and j.

This is demonstrated with the help of an example. Consider a stack with 3 tiers with a floorplan shown in Figure 3.6(a). Each tier is occupied by 4 PEs. If a case is considered where only one PE dissipates power (10W) at a time while other do not dissipate any power, the simulated steady state temperatures of all PEs (using 3D-ICE simulator)

[1] $R_{vertical}$ is the thermal resistance between vertically adjacent nodes.

[2] $R_{lateral}$ is the thermal resistance between laterally adjacent nodes.



(a) 3D stack with 3 tiers and total of 12 PEs considered in the report.

(b) The physical structure of the considered 3D stack.

Figure 3.6: Floorplan and structure of the considered 3D stack.

are reported in Table 3.2. The geometrical and material parameters used for thermal model and thermal simulations are shown in Table 3.3 and the 3D stack is shown in Figure 3.6(b). In Table 3.2, columns refer to the PE that dissipates power and rows indicate the corresponding increase in temperatures of all PEs. PE 0, 1, 2 and 3 are on the deepest tier of the stack. It can be seen from column "PE0" that PE0 has a significant effect on temperature of PE4 and PE8, i.e., in the vertical directions, and lower yest considerable effect on PE 1, 2 and 3 which lie on the same tier. Column "PE4" shows similar results, but it can been seen that effect on temperature of PEs on the same tier, i.e., PE 5, 6 and 7, is less as compared to that on tier 1. The effect becomes even less on tier 3 which is closest to the heat sink.

		Power Dissipating PE											
		PE0	**PE1**	**PE2**	**PE3**	**PE4**	**PE5**	**PE6**	**PE7**	**PE8**	**PE9**	**PE10**	**PE11**
A	**PE0**	12.44	06.04	06.04	03.03	10.96	05.48	05.48	02.80	09.48	04.84	04.84	02.52
F	**PE1**	06.04	12.44	03.03	06.04	05.48	10.96	02.80	05.48	04.84	09.48	02.51	04.84
F	**PE2**	06.04	03.03	12.44	06.04	05.48	02.80	10.96	05.48	04.84	02.51	09.48	04.84
E	**PE3**	03.03	06.04	06.04	12.44	02.80	05.48	05.48	10.96	02.51	04.84	04.84	09.48
C	**PE4**	10.96	05.48	05.48	02.80	10.97	05.39	05.39	02.74	09.49	04.82	04.82	02.50
T	**PE5**	05.48	10.96	02.80	05.48	05.39	10.97	02.74	05.39	04.82	09.49	02.50	04.82
E	**PE6**	05.48	02.80	10.96	05.48	05.39	02.74	10.97	05.39	04.82	02.50	09.49	04.82
D	**PE7**	02.80	05.48	05.48	10.96	02.74	05.39	05.39	10.97	02.50	04.82	04.82	09.49
	PE8	09.48	04.84	04.84	02.51	09.49	04.82	04.82	02.50	09.50	04.74	04.74	02.44
P	**PE9**	04.84	09.48	02.51	04.84	04.82	09.49	02.50	04.82	04.74	09.50	02.44	04.74
E	**PE10**	04.84	02.51	09.48	04.84	04.82	02.50	09.49	04.82	04.74	02.44	09.50	04.74
	PE11	02.51	04.84	04.84	09.48	02.50	04.82	04.82	09.49	02.44	04.74	04.74	09.50

Table 3.2: Change in temperature (in K) of PEs due to change in power dissipation of a PE in a three tier 3D stack.

Geometrical and material properties	**Explanation**	**Value**
Number of dies	Number of dies in the 3D stack.	3
Dimension	Dimension of the stack (length X width).	9mm X 9mm
Top substrate thickness	Thickness of silicon substrate at the top of stack (below heat sink).	200 μm
Si die thickness	The thickness of silicon dies in the stack.	50 μm
Metal layer height	Height of the layer containing metal (Cu) interconnects.	15 μm
Bond layer thickness	Thickness of the bonding layer between two dies in a stack [4].	10 μm
Effective Silicon thermal conductivity	Effective thermal conductivity of the silicon die; This value accounts for TSV occupancy [13].	$160.11 W m^{-1} K^{-1}$
Effective Silicon heat capacity	Effective heat capacity of the silicon die; This value accounts for TSV occupancy. [13].	$1.66e^{6} J m^{-3} K^{-1}$
Effective metal thermal conductivity	Effective thermal conductivity of the Cu metal layers; This value accounts for the low-k insulation layers and TSV occupancy [4].	$12 W m^{-1} K^{-1}$
Effective metal heat capacity	The thermal capacity of Cu.	$3.4419e^{6} J m^{-3} K^{-1}$
Effective bond layer thermal conductivity	Effective thermal conductivity of the bonding die; This value accounts for TSV occupancy. [13].	$6.83 W m^{-1} K^{-1}$
Effective bond layer heat capacity	Effective heat capacity of the silicon die; This value accounts for TSV occupancy. [13].	$3.99e^{6} J m^{-3} K^{-1}$
Thermal Interface material	Thermal conductivity of the TIM; 3-5 [13].	$5 W m^{-1} K^{-1}$
Thermal Interface material	Heat capacity of the TIM [13].	$4.0e^{6} J m^{-3} K^{-1}$
Heat spreader thermal conductivity	Thermal conductivity of the heat spreader.	$400 W m^{-1} K^{-1}$
Heat spreader heat capacity	Heat capacity of the heat spreader.	$3.55e^{6} J m^{-3} K^{-1}$
Heat sink heat transfer coefficient	Heat transfer coefficient of the heat sink.	$1.0e^{5} W m^{-2} K^{-1}$
Ambient Temperature	Ambient temperature where heat sink is connected to the stack.	$300K$
Thermal cell dimensions	Thermal cell dimensions (length$X$width) used to obtain effective thermal resistance between PEs.	$50\mu m X 50\mu m$

Table 3.3: Dimensional and material parameters used for thermal model.

Since change in temperature depends on resistance values between two nodes, it can be seen that the table is symmetric. Resistance between any two points is the same (seen from either side), hence the effect on temperature also remains the same. In reality, some finite time is required to reach these steady state temperature values. This is due to the thermal capacitance at each node. This delay can be used advantageously. The instantaneous temperature slack, i.e., the difference between critical temperature and current temperature of a PE, can be effectively used for improving the performance of a PE. At a given point of time, a PE may not be at its steady state temperature. In such a scenario, the PE can be allowed to operate at higher voltage/frequency hence allowing better utilization of a PE.

For effective thermal modeling, small thermal cells are required. However, considering these small thermal cells for run time calculations can be very expensive. The thermal relation between individual PEs is required. 3D-ICE thermal simulator is used to derive these relations. Details of how 3D-ICE simulator works can be found in appendix A.

3.2.2.3 Thermal Relation between PEs

The stack floorplan information is used to run a thermal simulation using 3D-ICE. At a time, only one PE is allowed to dissipate a fixed power of 10W for duration of $1ms$[3] (not steady-state) and the temperature of all PEs are noted. Same is done for all PEs to record a table similar to Table 3.2. Following equation can be obtained from the relation derived in Equation 3.12,

$$\left.\begin{aligned} \Delta T_j &= x * \Delta P_j * R_{jhs} \Rightarrow x.R_{jhs} = \frac{\Delta T_j}{\Delta P_j} \; ; \quad and \\ \Delta T_i &= \Delta P_j(x * R_{jhs} - y * R_{ij}) \Rightarrow \underbrace{(x * R_{jhs} - y * R_{ij})}_{\text{Effecctive resistance}} = \frac{\Delta T_i}{\Delta P_j} \end{aligned}\right\} \tag{3.13}$$

The two equation show

- The change in temperature of a PE due to change in its own power dissipation.
- Change in temperature of a PE due to change in power of another PE.

A matrix is created for the values of $(\Delta T_i/\Delta P_j)$ in Equation 3.13, for all PEs. This matrix represents the effective thermal resistance between two PEs, to form a direct relation between ΔP and ΔT. This results is an $N\ X\ N$ matrix where N is the number of PEs in the stack.

If the following matrix is the thermal resistance matrix,

$$R_{th} = \begin{pmatrix} R_{11} & R_{12} & R_{13} & R_{14} \\ R_{21} & R_{22} & R_{23} & R_{24} \\ R_{31} & R_{32} & R_{33} & R_{34} \\ R_{41} & R_{42} & R_{43} & R_{44} \end{pmatrix}$$

then, $R_{12} = R_{21}, R_{32} = R_{23}$ and so on. These values are thermal resistance between two nodes (hence equal). The diagonal elements R_{ii} represent the thermal resistance

[3] The chosen value is equal to the temperature check period. It is explained further in next chapters.

between node i and heat sink. If effective resistance is considered, as shown in Equation 3.13, then the effective thermal resistance matrix will not be symmetric. This is because $R_{eff_{ij}}$ will now represent effective thermal resistance between PE_i and PE_j for the case where PE_j dissipates power and its effect on PE_i is to be calculated. Therefore,

$R_{eff_{ij}} = (x * R_{jj} - y * R_{ij})$; and
$R_{eff_{ji}} = (x * R_{ii} - y * R_{ij})$

The change in temperature of PE_i due to change in power dissipation of PE_j, can now be directly given by

$\Delta T_i = R_{eff_{ij}} * \Delta P_j$.

The diagonal elements $R_{eff_{ii}}$ are used to calculate the effect of power dissipation of PE_i on its own temperature. These are the same as R_{ii}.

$\Delta T_i = R_{eff_{ii}} * \Delta P_i$.

This effective thermal matrix is used in power management model to handle the temperature of PEs based on the actual thermal model. From Equation 3.2, relation between power and V/F level is known, hence relation between temperature and V/F level now be given as:

$$\Delta T_i = R_{eff_{ii}} * A * \Delta(V_i^2 * F_i) \tag{3.14}$$

Since PE takes quite some time to reach its steady state temperature, considering the small time period to create this matrix helps in utilizing the instantaneous temp slack of a PE efficiently.

3.2.2.4 Limitations

Limitations of the derived thermal model are as follows:

- The accuracy of a thermal model depends on the thermal cell size. A fixed cell size may fail to accurately model the temperature change due to high and localized power dissipation, introducing some error. A relatively small thermal size of $50\mu m X 50\mu m$ is considered here, which provides sufficiently accurate figures for the architecture-level knowledge of temperature profile, but, these values do have some errors.

- Some lumped thermal resistances (e.g. the ones in the peripheral parts of heat sink) do not accurately represent the exact thermal resistance according to the analysis in [34].

- In 3D-ICE, heat is assumed to be delivered to the ambiance only via heat sink on top of the stack; hence, boundary conditions are applied only on top of the stack and not on the sides of the layers. PCB is also not considered.

3.3 Summary

- Power management block (PMB) is designed for DVFS on voltage island as well as per-core level, while latter becomes a special case former where each island has only one PE. Six DVFS levels are used in both cases. While per-core DVFS is considered to have six voltage levels corresponding to six operating frequencies, island are considered to have three voltage levels for six frequencies where each voltage level supports two frequencies hence enabling DFS. Since scaling down the frequency reduces power, this can help maintaining temperatures while avoiding complete island scaling when possible.

- PMB takes three inputs from the system to compute new operating V/F levels. These inputs are utilization of each PE in previous control period, current temperature of each PE and total chip power. Temperature inputs are taken at every temperature-check cycle while other two are taken at every control period.

- To design an effective control the dynamics of the controlled system, i.e., the relation between controlled variable and manipulated variable are modeled. The manipulated variable is the operating V/F level of each PE while the controlled variables are power and temperature. Hence, the relation between power and V/F levels, and the relation between temperature and V/F levels are modeled in Equation 3.2 and Equation 3.14 as:

  $$\Delta P = A * \Delta(V^2 * F)$$

  $$\Delta T_i = R_{eff_{ii}} * A * \Delta(V_i^2 * F_i)$$

- To achieve this relation between temperature and V/F level, thermal modeling of 3D stacked IC was studied. It was evident from the conducted experiments that heat flow is prominent in vertical as well as in lateral direction in deep stacks. This is due to reduced thermal conductance between deeper tiers and heat sink. An equation for effective thermal resistance between two PEs in a 3D stack was derived. This equation represents a direct relation between change in temperature of a PE due change in power of another PE in the stack (Equation 3.13).

- Thermal simulator 3D-ICE is used to obtain temperature profile of the target stack by using small grid sizes. By allowing only one PE to dissipate power at a time, the change in temperature of all PEs is recorded. These values are then used to compute the effective resistance matrix of size $N \; X \; N$ ($R_{eff}[N][N]$) where N is the number of PEs in the stack using the relation stated in Equation 3.13. An element $R_{eff_{ij}}$ represents effective thermal resistance between PE_i and PE_j for the case where PE_j dissipates power and its effect on PE_i is to be calculated.

- Heat conduction is a strong function of material properties like thermal conductance and thermal capacitance used in the stack. The geometrical and material properties used for generating effective resistance matrix should be similar to the target stack. The target floorplan is shown in Figure 3.6(a) and the values of geometrical and material parameters considered in this thesis are reported in Table 3.3.

4 Power Management Scheme

This chapter details the algorithm of the proposed power management scheme. Firstly, how the modeled dynamics of the system can be used as static parameters is described. Next, the five stages of the control algorithm are explained.

4.1 Static Parameters

Relation between power and V/F levels, and temperature and power (and V/F levels) were obtained in previous chapter as:

$$\Delta P = A * \Delta(V^2 * F)$$

$$\Delta T_i = R_{eff_{ii}} * \Delta P_i$$

$$\Rightarrow \Delta T_i = R_{eff_{ii}} * A * \Delta(V_i^2 * F_i)$$

For an intended application, ΔP values are obtained for each set of V/F values. R_{eff} values are derived for the target floorplan, hence ΔT value corresponding to each ΔP value can also be computed and these values are recorded as static parameters.

Calculated Parameter	**Explanation**
ΔP (For each DVFS level of all PEs)	ΔP is the change in power of a PE when the immediate next V/F level is chosen. For a PE, $\Delta P_0 = A * V_1^2 F_1$, $\Delta P_1 = A * (V_2^2 F_2 - V_1^2 F_1)$, $\Delta P_2 = A * (V_3^2 F_3 - V_2^2 F_2)$ and so on.
ΔT (For each ΔP value for all PEs)	ΔT is the change in temperature of a PE when the immediate next V/F level is chosen. For a PE, $\Delta T_0 = R_{eff}[0][0] * \Delta P_0$, $\Delta T_1 = R_{eff}[1][1] * \Delta P_1$ and so on.
Normalized R_{eff} matrix	Row i of R_{eff} matrix shows the effective resistance between a power dissipation source and the PE i. Each row of the matrix is normalized to get values ranging from 0 to 1.
$Normalized\ area * weight_e$ $+\ normalized\ height * weight_f$	Area of PEs are normalized to achieve values between 0 and 1. If all PEs are of same area, the normalized value of all PEs will be 1. These normalized areas are multiplied by $weight_e$ (discussed later). If bottommost die in an N-die-stack is die 1 and the topmost die is die N, then the normalized height is given by (die number)/(total number of dies). This normalized height is multiplied by $weight_f$ (discussed later).
$Tmargin$	$Tmargin$ is the decided temperature margin to keep PEs below critical temperature by Tmargin.

Table 4.1: The static information stored beforehand for the use of power management block.

A temperature margin ($Tmargin$) is considered to keep PEs below critical temperature by a certain amount. This helps maintaining the temperature of PEs at a safe distance from critical limit. It is done in order to ensure that the temperatures of PEs always remain below critical temperature even under unexpected circumstances like noise in power supply and sudden increase in workload of a PE.

Some static parameters are computed and stored before run-time for the use in power management block, while some information is stored during run time and is required for the deciding operating V/F levels of a PE. These parameters are listed in Table 4.1 and Table 4.2.

4.2 Control Algorithm

Control period defines the intervals at which the PMB takes inputs and computes the new V/F values. Temperature-check period defines the interval at which temperature input is available to the PMB.

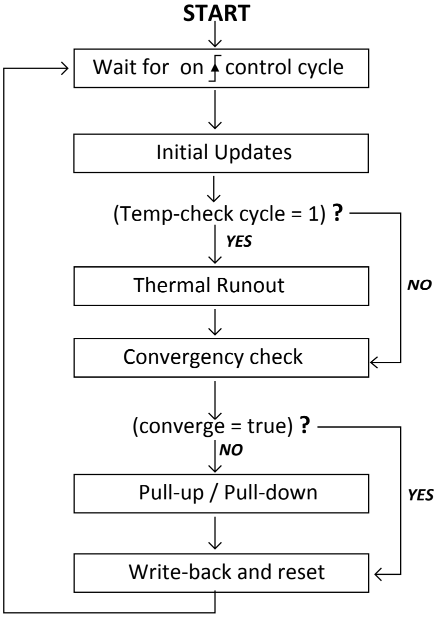


Figure 4.1: Flowchart showing stages in PMB.

When the system starts, no decision is taken in first two control periods. The chip is allowed to start with maximum power dissipation (max DVFS levels for each PE). This is considered to be the warm-up stage for each PE. This is done in order to be able to see the actual activity requirements of a PE. Computations start from the third control period. As shown in Figure 4.1, the complete algorithm is divided into 5 stages:

1. Initial updates
2. Thermal run-out

Parameter	Explanation
delP	$P_{budget} - P_{total}$
uppospe	Set to 0 if a PE cannot be pulled up. (per PE)
upposisl	Set to 0 if an island cannot be pulled up. (per island)
Runoutdenied[i][j]	A 2D matrix of size N X N where N is the number of PEs. $Runoutdenied_{ij}$ denoted that PEs j should not be scaled up due to its effect on temperature of PE i.
LocalTemp	stores a local/estimated value of temperatures of each PE. (per PE)
TempMargin	Temperature Margin is the difference between threshold (T_{th}, i.e., critical) temperature and the actual temperature. (per PE)
OpIndPE	Index of operating V/F level of each island (*0* represents minimum). (each PE)
OpIndIsl	Index of operating voltage level of each island (*0* represents minimum). (per island)
LocalUtil	Local copy of activity factor of each PE. (per PE)
WeightPE	Calculated weight of each PE. (each PE)
WeightIsl	Calculated weight of each island. (per island)

Table 4.2: The dynamic information generated and store by power management block during run-time.

3. Convergence check
4. Pull up or pull down
5. Write-back and reset

When each PE behaves as an independent island, it becomes per-core DVFS. Hence, per-core DVFS is a special case of DVFS in islands. The approach for per-core and islands are explained here in context with the islands. Island approach requires more computations due to additional weight calculations and is more complex due to DFS implementation which requires investigating frequency of each PE before scaling island voltage. However, in case of per-core DVFS, algorithm becomes quite straight forward.

4.2.1 Initial Updates

When a new control period starts, few parameters are updated. These include TempMargin, LocalTemp, delP, LocalUtil and Runoutdenied. TempMargin represents the difference between actual and critical temperature of each PE. It is updated only when a new temperature-check cycle has started. Similarly, LocalTemp is updated when a new temperature-check cycle starts and it hold the actual temperature of each PE. delP represents the difference between total chip power and power budget value. LocalUtil for each PE represents the utilization of each PE and is assumed to be proportional to utilization in previous two control periods. Runoutdenied is a N X N matrix where N denotes the number of PEs in the stack. If the element j of row i of the Runoutdenied array is 1, it means that TempMargin of PE i is critical and PE j should not be pulled up in order to keep PE i below critical temperature. This constraint is set in runout stage, and is lifted in this stage if the temperature of PE i is below the safety margin by 1K, the constraint is removed from one of the elements in row i by making it 0. A restricted PE which is closest to the heat sink is chosen and the corresponding element of row i is set to 0.

Few parameters are updated at the start of every new control period.

- IF value of temperature-check cycle = 1, THEN
    - TempMarigin = T_{th} - Input Temperature.
    - LocalTemp = InputTemp.
- delP = P_{budget} - P_{input}
- LocalUtil = 0.8*(input activity) + 0.2*(LocalUtil of previous control period).
- For each PE (say i), IF TempMargin[PE_i] > (Tmargin+1), THEN last elements of Row i of Runoutdenied which is 1 is set to 0; leaving others unchanged.

4.2.2 Thermal Runout

This step is executed if a new temperature-check cycle has started. This stage is executed to ensure that temperature of each PE is below the safety margin (Tth - Tmargin), and to check if an OFF PE can be turned ON again. If a PE is OFF, possibility of turning it ON again is checked. If new temperature of a PE after turning it ON (LocalTemp + $\Delta T[0]$) is calculated to be below the critical temperature, then the PE is turned ON and other PEs on the same island are pulled down to the minimum operating voltage. Turning a PE ON is considered important therefore island is pulled to the minimum voltage. Temperature margin of each ON PE is checked, if the temperature of a PE (victim) is beyond the safety margin then the PMB reacts to bring the temperature at the safety margin again. This is done by allocating weights to each PE and one with the highest weight is scaled down. If PMB fails to bring the temperature below safety margin, the victim PE is turned OFF. When PMB scales a PE (say PE i) down in order to maintain temperature of PE *victim*, then element i of row *victim* in Runoutdenied is set to 1. Each time the V/F level of a PE is scaled, delP and LocalUtil are updated.

IF Temperature-check cycle = 1, THEN

- For each PE, IF $LocalTemp + \Delta T[0] < T_{th}$,

  - PE is turned ON. All PEs in this island are brought to the minimum operating voltage[1].
  - For each change, LocalTemp and delP is updated.

- For each ON PE, IF $TempMargin[PE] < Tmargin$, PMB reacts to pull temperature below safety margin (Tth - Tmargin). First, the normalized TempMargin

  of each PE is calculated. A local copy of each OpIndPE is created and a weighted equation is considered to decide the target PE.

  For each PE i, assigned a weight;
  $weight_a * (1 - LocalUtil) + weight_b * (normalizedR_{eff}[victimPE][i])$

  In this equation:

  - First term results in a heavier weight for a less active PE; and
  - Second term results in a heavier weight for a PE with stronger thermal relation with the victim.

  PE with heaviest weight is pulled down first.

  If a PE's selected[2] V/F level requires change in island voltage, all PEs in the island are brought down to the corresponding V level. (changes are made to the local copy of OpIndPE). When a new V/F level is chosen for a PE, Localtemp is updated, i.e., local temp[victim] = LocalTemp[victim] - $\Delta P[target] * R_{eff}[victim][target]$.

  If the temperature is still less than above safety margin (Tth - Tmargin), next element in the queue is selected.

  IF temperature cannot be brought within the safe limits, and temperature is above critical temperature, the victim is turned OFF but setting the ON_OFF[victim] signal HIGH, and V/F level *0* is chosen.

  ELSE, the following parameters are updated:

  - delP = delP + ΔP
  - LocalUtil[target] = LocalUtil[target] * newFreq/oldFreq.
  - OpIndPE and OpIndIsl

  The PEs that are pulled down in this stage, should not be allowed to scale up in the coming stages. Hence, Runoutdeny[victim][i] is set to 1 where i denotes PEs that have been scaled down. Also, uppospe[i] is set to 0.

[1] in case of per-core, the core is simply turned ON.

[2] Selected V/F level is inside power management block and is not set on the chip until the last stage. All decisions that are made here are local till write-back stage appears.

4.2.3 Convergence Check

Power value is assumed to be converged if total chip power is between 98% and 100% of power budget value (0 < delP < Pwindow). If not, next step, i.e., pull up or pull down is required, else control skips to step 5.

4.2.4 Pull Up or Pull Down

If total chip is above the budget value, operating V/F levels need to be scaled down in order to bring total chip power below the budget value, whereas if total chip power is less than 98% of the power budget value, the operating voltages should be scaled up in oder to utilize power budget efficiently. All PEs are assigned with a weight and the weight of an island is the average of the weight of PEs on it.

If delP < 0 , pulling down of V/F levels is required. Else if, delP > Pwindow, V/F levels are pulled up.

For both the cases, a weighted equation is considered.

$(weight_c * LocalUtil) + (weight_d * normalized\_temp\_margin) + (weight_e * normalized\_height) + (weight_f * normalized\_area)$

Terms 3 and 4 were calculated and stored beforehand while term and 1 and 2 are needed to be calculated.

Term 1: PE with a larger activity factor implies a busier PE and hence should be the preferred choice to scale up.

Term 2: PE with a larger temperature margin implies a cooler PE, hence should be the preferred choice to scale up.

Term 3: PE that is higher up in the stack, has less effect on the temperatures in the stack, and should be the preferred choice to go up.

Term 4: PE with a larger area (in case of heterogeneous system) results in lesser increase of power density, hence should be the preferred choice while going up.

To be able to account for all these factors, a weighted equation is considered. PE with the largest weight is the preferred choice for scaling the V/F up, while the one with the lowest weight is the first choice for scaling down. Weight of an island is the average weight of PEs on that island.

Pull up: To scale V/F levels of a PE up, its new temperature, i.e., temperature after scaling is check. If the temperature would still be below safety margin (Tth - Tmargin) then only its operating voltage and frequency level is scaled up. Island with the largest weight is chosen, if all PEs are at maximum frequency level for the operating voltage and all can be scaled up, then the island is scaled up. Else, island with the next largest weight is selected. If all PEs are not at the maximum frequency of the operating voltage, the PEs are checked and scaled up while keeping the island at the same voltage. This is done till either no more PE can be pulled up or total power exceeds 98% of budget value. If the value has crossed power budget value, Pulldown stage is called in order to converge. Each time a PE is pulled up, its temperature is updated and parameter delP (Pbudget - Ptotal) is updated.

To pull a PE (targetpe) up, its new temp is checked, i.e., $LocalTemp + \Delta T$. IF it is below the (Tth - Tmargin), AND IF uppospe is 1, THEN only it is pulled up, ELSE uppospe[targetpe] is set to 0.

Island with the largest weight is chosen as the target island.

- If all the PEs in the island are at the maximum frequency for operating voltage, the complete island should be pulled up.
  - If operating voltage is not the max voltage, upposisl[targetisl] =1, uppospe[all PEs on the target isl]=1, new temp of all PEs are within the specified limits, the island is pulled up
  - Else upposisl[targetisl] is set to 0.
- Else, the individual PEs in the target island should be pulled up according to their weights.
  - If uppospe[targetPE]=1, new temp of targetPE is within the specified limit, the PE is pulled up
  - Else uppospe[targetPE] is set to 0.
- If all PEs of the target island are operating on the highest possible V/F levels, island next in the queue is chosen as the target island.
- Each time the operating V/F level of a PE is changed, delP and localTemp of the PE are updated. The process continues till
  - Either delP > Pwindow;
  - OR, pulling up of PEs and island is not possible.
- IF, Pwindow < delP < Pbudget, the power value is converged and control shifts to next stage;
  ELSE, (i.e., delP < 0) PULLDOWN is called.

Pull down: Island with the smallest weight is chosen, if all PEs are at minimum frequency level of the operating voltage, then the island is scaled down. Else, individual PEs are scaled down. This is done till either no more PE can be pulled down or total power is below budget value. If the value is below 98% value, Pullup stage is called in order to converge. Each time a PE is pulled down, its temperature is updated and parameter delP (Pbudget - Ptotal) is updated.

Island with the smallest weight is chosen as the target island.

- If operating voltage of the target island is not the minimum voltage, AND
  - IF all the PEs in the island are at the minimum frequency for the operating voltage, the complete island is pulled down.
  - Else, the individual PEs in the target island is pulled down (according to their weights).
- If all PEs of the target island are operating on the lowest possible DVFS levels, island next in the queue is chosen as the target island.
- Each time the operating V/F level of a PE is changed, delP and localTemp of the PE are updated. The process continues till
  - Either delP > 0;
  - OR, pulling down of PEs and island is not possible.
- IF, Pwindow < delP < Pbudget, the power value is converged and control shifts to next stage;
  ELSE, (i.e., delP > Pwindow) PULLUP is called.

To avoid infinite loop

- A constraint is put on the number pull up - pull down calls; and
- If a PE/island that is pulled up (or pulled down) and is pulled down (or pulled up), it is not allowed to be pulled up (or pulled down) again. Hence, avoiding pulling a PE/island up and down trying to converge.

4.2.5 Write-Back and Reset

The chosen V/F values and the ON-OFF state signals for each PE are implemented on the PEs. Few parameters are reset. These are:

- uppospe. uppospe for PEs that are marked as 1 in Runoutdenied are ignored, rest are set to 1.
- upposisl. upposisl for each island is set to 1.

The value of 1 denotes scaling up is possible.

4.3 Summary

- For an intended application, ΔP values are obtained for each set of V/F values. R_{eff} values are derived for the target floorplan, and ΔT value corresponding to each ΔP value are also computed. These values remain fixed for each V/F level and are recorded as static parameters to avoid on-line computations.

- PMB tries to keep the temperature of each PE at a safe distance from critical limit. It is important in order to ensure that the temperatures of PEs always remain below critical temperature even under unexpected circumstances like increase in workload of a PE or noise in power supply. The value is chosen experimentally and is explained in next chapter.

- In thermal runaway situations, where temperature of a PE crosses the defined safety margin (i.e., critical temperature - decided margin), each PE is assigned a weight which depends on its thermal relation with the victim and its utilization. A PE with less utilization and stronger thermal relation with victim is a preferred choice for scaling down.

- When the total chip power is above the budget value or below 98% of the budget value, then the operating V/F levels are scaled in order to converge to the set budget value. The 2% window is taken for convergence and for avoiding hops around the budget value. The window is chosen only on one side because power should always remain below the specified budget.

- To scale the the operating V/F levels of PEs, all PEs are assigned a weight which depends on four factors: its utilization, temperature margin, distance from the heat sink, and its area.

  - Utilization is included in the equation for allowing a PE with higher utilization to speed up hence PE is high utilization is a preferred choice for scaling up.
  - Higher temperature margin represents lower operating temperature, chances of thermal runaway on scaling up its V/F are less hence PE with hight temperature margin becomes a preferred choice for scaling up.
  - PE that is closer to heat sink has least effect on the stack temperature, hence is a preferred choice for scaling in V/F up.
  - Increase in power by a specific amount in a larger area results is lower increase in power density which has lesser impact on the stack temperature.

  To account for all these factors, weighted equation is considered, PE with the maximum weight is a preferred choice for scaling up while one with the lowest weight is the a preferred choice for scaling the V/F level down. Weights were chosen experimentally.

Results and Discussion 5

This chapter describes the simulation environment and setup used to perform various experiments on the proposed PMB approach. Further, the performed experiments are described with an analysis of results.

5.1 Simulation Environment

Figure 5.1 show the block diagram of the setup from the top level. As compared to Figure 3.2, the block Testbench behaves as the test system. All inputs for the PMB are generated here, and the output generated by the PMB are tested here.

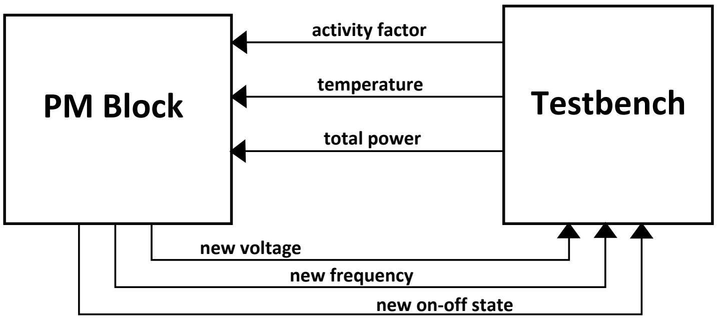


Figure 5.1: Experimental setup.

The complete setup is made using SystemC. The testbench provides 3 inputs to the PMB.

- utilization of each PE
- temperature of each PE
- the total chip power dissipation

Activity factor is generated using the SimpleScalar [35] and a benchmark. SimpleScalar is a cycle accurate computer architecture simulator. Basicmath application from the Mibench benchmark (automotive) is chosen [36]. It keeps utilization of a PE less [37], giving good opportunities for scaling V/F level. The activity trace is generated by executing the application on SimpleScalar for an in-order 32-bit PISA configuration. The trace records and marks each cycle as either a processor-bound cycle or memory-bound cycle. This trace is then used to calculate the activity factor of a PE in each control period.

SimpleScalar configured with Wattch [38,39] is used to achieve power values for different V/F levels. The power values corresponding to each (V^2F) value is then put in a linear regression calculator to achieve the values of A and B in Equation 3.1. When PMB decides the operating voltage and frequency value for a PE, the power value of each PE for the control period is calculated and recorded. The total power dissipated by the stack is then calculated and provided to the PMB as one of the inputs in the next control period.

When temperature check cycle is invoked, the average power value for each temperature cycle is written to input files for 3D-ICE, which is then invoked with the timing information identical to that of temperature-check cycle. The temperature information is generated and the final temperature values for each PE are sent to the PMB. The testbench records certain information about each PE for performance analysis. These include temperature, power, operating V/F level, actual run time, ideal run time and OFF time. It also records total power dissipation for each control period. The complete simulation setup is shown in Figure 5.2.

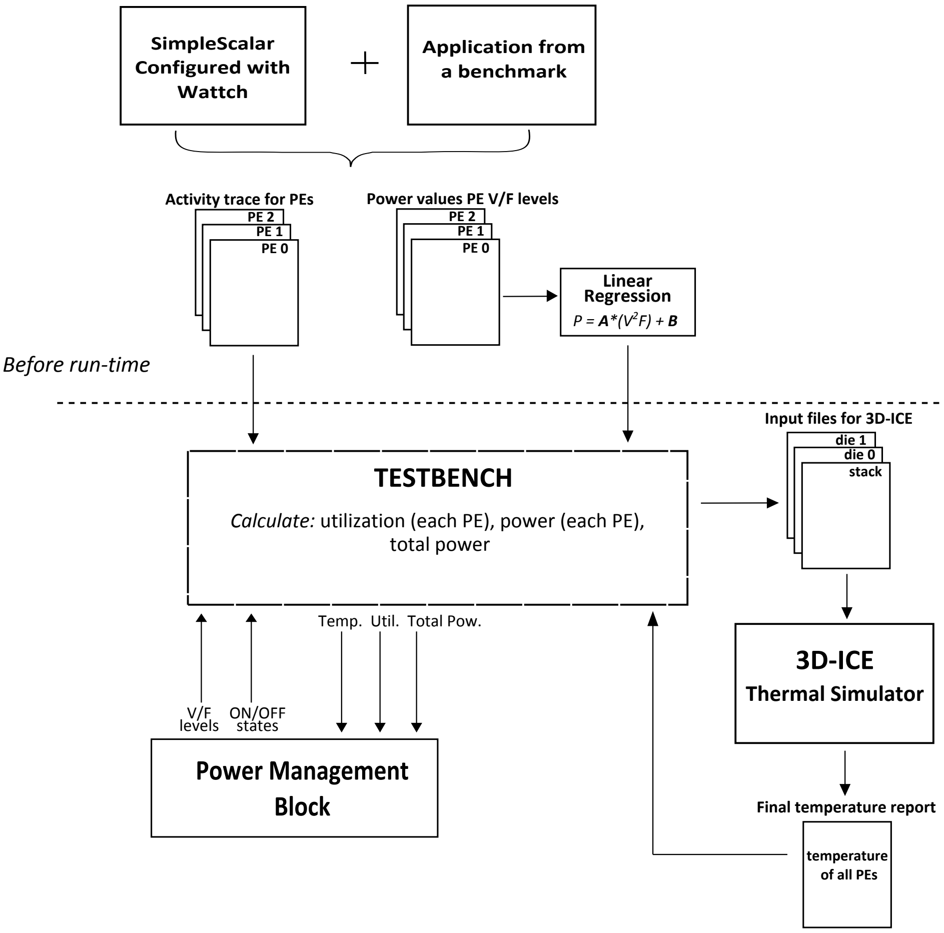


Figure 5.2: Complete Simulation environment.

Performance calculations: The performance loss in percentage is calculated using the following relation:

$$PerformancePenalty(\%) = \frac{increase\ in\ execution\ time}{execution\ time\ with\ maximum\ frequency} * 100\% \quad (5.1)$$

Voltage and frequency switching takes time and also consumes energy. During this switching period, PEs cannot execute any task. These losses are not modeled here but are discussed with results.

5.2 Experimentation

A three tier stack as shown in Figure 3.6(a), reproduced in Figure 5.3 is considered. Each tier is considered to have 4 similar PEs. Each PE can operate on six frequencies: 700MHz, 800MHz, 900MHz, 1000MHz, 1100MHz and 1200MHz. Voltage levels are chosen according to the experiment. A deep sleep mode for each PE is considered where clock to a PE is gated. In this mode, there is no switching power but leakage power still exists.

Two sets on simulation setup were developed. Both were tested on the same target application, using same convergence algorithm, similar conditions for V/F level scaling and similar constraints on power and temperature. However, one setup used the proposed approach for DVFS in 3D MP-SoCs and the other used an approach similar to DVFS in 2D chips where temperature and V/F levels of each PE is considered independently.

A control period of 60,000 cycles at maximum frequency is considered. 60,000 cycles at the frequency of 1200MHz result in a period of 50us, this is chosen to be the control period for the PMB. This value is influenced by [27] and [40]. [27] has analyzed per-core DVFS with on-chip switching regulators. They have shown that voltage transitions can occur on the order of tens of nanoseconds. This would result in a very small amount of overhead during switching of V/F levels. [40] has studied various power delivery networks for 3D ICs and have also discussed the feasibility of including on-chip regulators. This small overhead allows us to use smaller control period than with off-chip regulators. Precise value of 50us seconds is taken for the sake of simplicity. Temperature check cycle of 1ms period is chosen. Frequency temperature inputs require more computations while larger check periods result in large temperature differences. The value of 1ms is chosen on the basis of experiment. The value was kept large enough to monitor the temperature change of about 1K when operating near critical temperature.

Power budget is taken according to the experiment and a 2% window (Pwindow) is used for convergence. For e.g., for a power budget of 100W, the power value of 98W is assumed to be converged to the budget value. This window is important to reduce the number of iterations required to converge and to avoid the fluctuation in the V/F levels. But, Pwindow should be small enough to avoid performance losses due to insufficient budget utilization. Fluctuations in the operating conditions such as power supply noise, change in ambient temperature and higher switching activity than expected can affect the temperatures of PEs. In order to ensure that the temperatures do not cross the critical limit it is important to include a margin that keeps a PE at a safe distance from the critical temperature. The value of this margin was decided on the basis of experiments. The rise in temperature of a PE in one temperature-check cycle was always observed to be less than $2K$. Therefore, a temperature margin (Tmargin) of $2K$ was considered. Since the PMB receives a temperature feedback in every temperature-check period, the sudden rise in temperature can be dealt with. The PMB tries to keep the temperature of each PE $2K$ below critical temperature but does not turn a PE OFF unless the temperature crosses the critical temperature. This margin is considered in both the approaches.

Each PE is assigned with the same task but with an offset of few cycles. If PMB is unable to maintain the temperature of a PE below its critical temperature, its frequency is set to 0 (OFF state). In order to see the actual effect on the performance of a PE, tricks like task migration to cope with the performance losses for OFF PEs are not considered. When the temperature of the PE is below its critical temperature and the PMB decides to turn the PE ON again, it resumes its task.

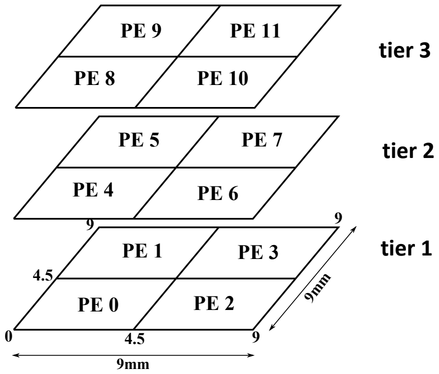


Figure 5.3: 3D stack with 3 tiers and total of 12 PEs that is considered for experiments.

5.2.1 Experiment 1

This experiment was done on the considered 3D stack to compare the two approaches at per-core level when the temperature constraints on the PEs are lenient.

Setup: A per-core DVFS is implemented. Each PE is operated at six V/F levels: (0.8V 700MHz), (0.855V, 800MHz), (0.907V, 900MHz), (0.956V, 1000MHz), (1.003V, 1100MHz) and (1.048V, 1200MHz). The voltage values for each frequency are decided such that, if the Frequency scales by x, then voltage scales by $x^{(1/2)}$. The values of A and B for the relation between power and V^2F are obtained as 0.0083055 and 2.4844, respectively. The co-relation factor of 0.999616 was obtained signifying the strong linear relation between the power and V^2F. Maximum power is dissipated by the stack when all PEs are operating at the maximum V/F level. The stack dissipated a maximum power of 162W therefore the power budget of 164W was chosen to allow PEs to operate at maximum V/F levels. A temperature constraint of 330K was imposed on all PEs. Due to the considered Tmargin, the PMB tries to keep the temperature of each PE under 328K but does not turn a PE OFF unless its temperature exceeds the critical temperature limit of 330K. This margin is considered in both the approaches.

Analysis: PE 0, PE 1, PE 2 and PE 3 are the PEs on the deepest tier of the stack. The temperatures of these PEs is always higher than those on the upper tiers. Figure 5.4 shows the total power dissipation of the stack for every control period and Figure 5.5 shows the temperature profile of PE 0 for each temperature-check period. The stack initially dissipates maximum power by allowing the PEs by operating at maximum V/F levels. The temperature of PE 0 rises to 328K in around 50ms. This is when PMB starts to pull the power of the stack down in order to keep the temperature below 328K.

In case of 2D approach, the V/F level of PE 0 alone is pulled down while with the new approach, the target PE is decided according to the assigned weights in pull down stage. Both approaches were observed to effectively maintain the temperature around 328K till 90ms. Beyond this point, the 2D approach failed to keep the temperature below the assigned margin. The PE was then allowed to operate at minimum frequency and the temperature then increased gradually. While with new approach, the temperature was observed to be maintained below 228K by pulling down the V/F levels of affecting PEs. This can be seen as the reduction in power dissipation at 90ms.

At around 125ms, the temperature of PE 0 in the 2D approach was observed to settle again. This is due the fact that the temperature of PEs on tier 2 reach 328K. This can be seen in Figure 5.6 where at 125ms the temperature of PE 4 reaches the margin. The operating V/F level of PE 4 was then pulled down in order to maintain its temperature below the margin. This results in the reduction of power dissipation. When the temperature value dropped below the margin, the frequency of the PE was increased again, which led to the rise in temperature. As the temperature rises, the V/F level is pulled down again. The constant pull up and pull down can be noticed in the total power dissipation with 2D approach. Figure 5.7 shows the active V/F levels of PE 4, in which constant switching in the V/F levels can be observed. In contrast, the operating V/F levels, temperature profile and power dissipation of the stack with the new approach are observed to be observed to be very smooth and settled. This is due to the fact that when the V/F level of a PE is scaled down due to its effect on temperature of another PE (victim), it is not scaled up again unless the temperature of the victim falls below the margin by 1K. This value was chosen experimentally to avoid constant switching. Also, the value is kept small enough to not lose opportunities of scaling under reduced temperature conditions in order to allow utilization of available instantaneous temperature slack. This constraint is imposed in the runout stage and is removed in the update stage of the control algorithm which helps achieving a better stability.

Conclusion: Although both the approaches took almost the same time to finish the tasks on all PEs, it was evident that the 2D approach failed to maintain the temperatures at a margin from critical temperature, while the new approach could effectively keep the temperature below this margin by effectively scaling the PEs considering the thermal relation. Furthermore, the 2D approach utilized PEs on tier 2 by constantly scaling their V/F levels and allowed the PEs on tier 1 to cross the temperature margin. The new approach on the other hand resulted in a more stable system with all temperature values maintained below the margin at a cost of 2% performance loss in the PEs on tier 2. This value does not account for the losses in switching V/F levels of PEs.

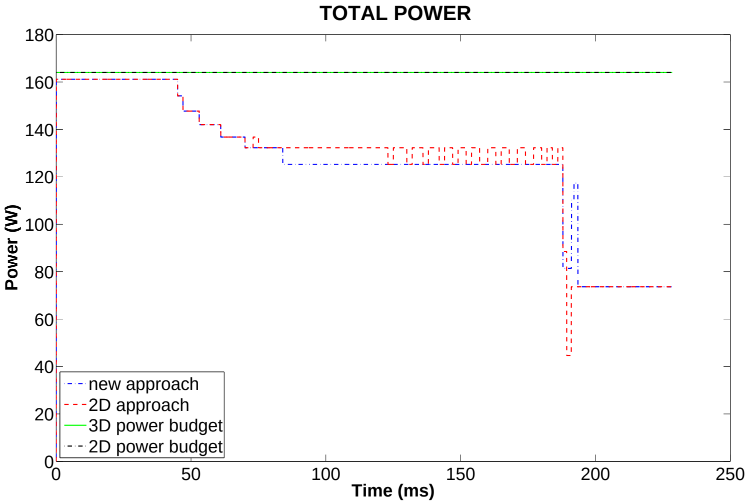


Figure 5.4: Total power dissipated of the considered stack when power budget is 164W and temperature constraint of 330K is imposed on each PE.

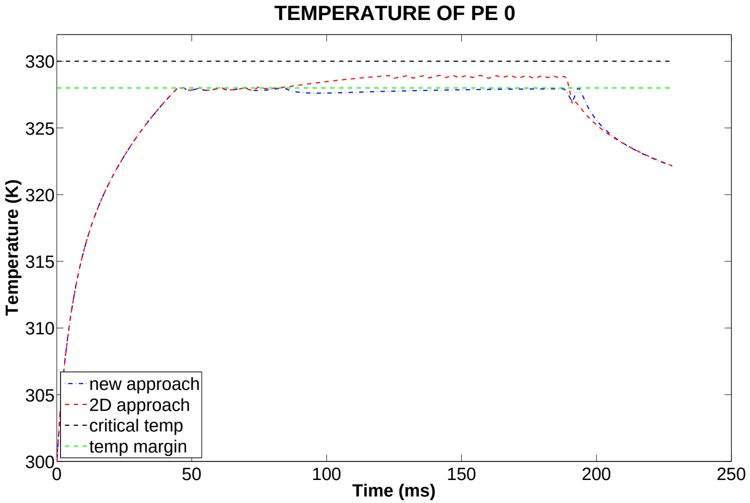


Figure 5.5: Temperature profile of PE 0 when power budget is 164W and temperature constraint of 330K is imposed on each PE.

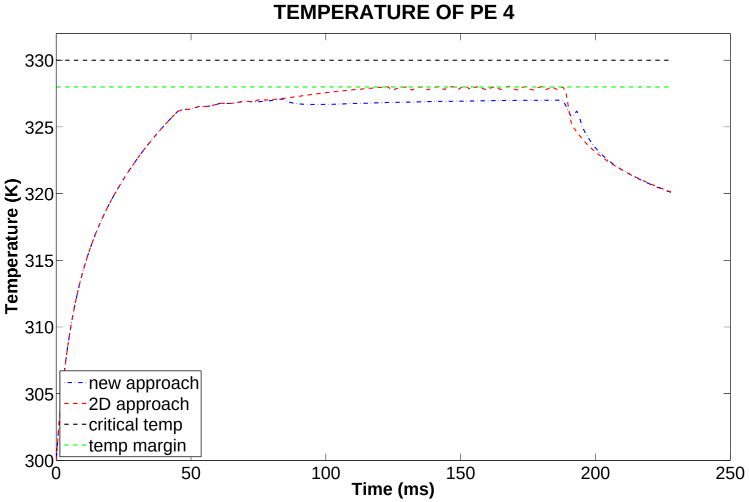


Figure 5.6: Temperature profile of PE 4 when power budget is 164W and temperature constraint of 330K is imposed on each PE.

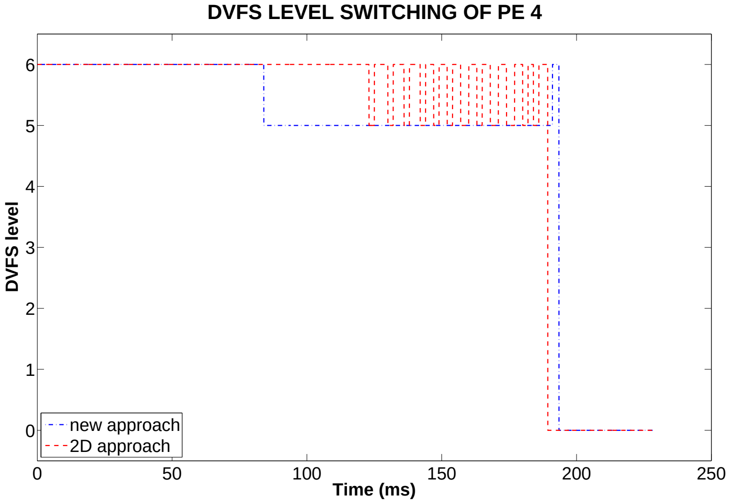


Figure 5.7: Operating V/F levels of PE 4 when power budget is 164W and temperature constraint of 330K is imposed on each PE.

5.2.2 Experiment 2

Experiment 1 showed similar performance results for the two approaches. A major difference was observed in the temperatures of PEs on tier 1 and the V/F level switching on tier 2. Experiment 2 was conducted to test the approaches under more stressed situations.

Setup: A configuration similar to that in Experiment 1 is considered. But, Power budget of 160W was imposed to check the ability to converge to the set budget value. A temperature constraint of 320K was imposed on all PEs. Due to the considered Tmargin, the PMB tries to keep the temperature of each PE below 318K but does not turn a PE OFF unless its temperature crosses 320K. This margin is considered in both the approaches.

Analysis: Figure 5.8 shows the total power dissipation of the stack for every control period. During initial 20ms, the total power with both the approaches were observed to converge well to power budget value till the temperatures on tier 1 reached 318K. This is when the PMS started to scale down the V/F levels on tier 1. While PMB with 2D approach scaled down the affected PEs, the PMB with the new approach scaled according to the weighted equation. A PE that was pulled down by the PMB with new approach was not allowed to scale up till the temperature of affected PE dropped below 317K (1K below the margin). While both approaches tried to keep the temperature of the PEs below 318K, the PMB with the new approach could successfully converge the temperature while the 2D approach failed to do so. This can be seen in Figure 5.9. PE 0 continued to operate at the lowest V/F level, causing the temperature to cross the critical limit and PMB was forced to turn it OFF. PE 1, PE 2 and PE 3 had similar temperature profiles. When the temperature falls below the critical temperature, the PE is turned ON again, which gradually results in temperature rise and the PE is turned OFF again. The switching of V/F levels of PE 0 can be seen in Figure 5.11. The switching period slowly increases as temperature on the tier 2 increases, resulting in slower cooling of PEs on tier 1. The temperature of PEs in case where the new approach is used are always below the temperature margin. Lower operating V/F levels are chosen to keep temperatures below temperature margin. While in case of 2D approach, the PEs on tier 1 were mostly OFF, ones on tier 2 operated at lowest V/F level due to the rising temperature and the PEs on the top tier operated at the maximum V/F level.

Conclusion: Figure 5.10 shows the higher sum of frequencies achieved with the new approach. As a result, tasks completed faster on the PEs. However, Figure 5.8 shows that the total power dissipation to be lower as compared to the 2D approach. This was achieved by essentially keeping all the PEs in ON state and allowing them to operate on lower V/F levels while with 2D approach, PEs on the topmost tier were operating at highest frequency resulting in higher power. Consequently, better performance was achieved with lower power dissipation. Higher stability could be observed in this experiment as well. PEs took almost 70ms less to complete the allocated tasks. The performance losses for both the approaches are summarized in Table 5.1.

	2D approach (x)	**new approach (y)**	**(x-y)**
Total simulation time	336.05ms	260.35ms	65.7ms (19.55% of x)
Average OFF-time on tier 1	106.5ms	0ms	106.5ms
Average performance loss on tier 1 (including time in OFF state)	78.38%	38.48%	39.9%
Average performance loss on tier 2 (including time in OFF state)	29.28%	37.80%	-8.52%
Average performance loss on tier 3 (including time in OFF state)	0%	29.34%	-29.34%

Table 5.1: Performance losses with the two approaches in Experiment 2.

The reported losses do not account for the overheads due to turning a PE ON/OFF and switching of V/F levels. Since PEs are stalled during the transition time, a performance loss is incurred. However, these losses are negligible when compared to performance improvement obtained from scaling. The difference in simulation time is mainly because OFF PEs on tier 1 had to wait for the temperature to fall below the critical value to resume their operation. This requires PEs on upper tiers to finish their tasks and turn OFF so that PEs on the lower tiers can resume their tasks. The time spent in OFF state after completing the task is not included in the OFF time or performance calculation. It can be seen that that 2D approach allowed PEs on the top tier to operate without any performance loss while the PEs on the deepest tier suffered with constant switching between ON/OFF states. The new approach on the other hand could achieve a balance between the performance losses on the three tiers and maintained the temperatures to avoid switching to OFF state.

A lot of fluctuations in temperature of PE 0 with the 2D approach can be seen in Figure 5.9. These temperature fluctuations alter threshold voltage, carrier mobility, and saturation velocity of MOSFETs. Furthermore, the induced variations in individual device parameters have unique effects on MOSFET drain current [41]. It was seen in Section 2.2.3 that the speed of a device depends on this value of drain current. Degradation in the drain current would mean degradation in performance of a device. To maintain the performance of devices over time, avoiding these fluctuations in temperature is important. The new approach was observed to achieve a stable system with smoother temperature profile and less switching between the V/F levels.

Using techniques like task migration can allow the transfer of the workload of an OFF PE to those operating at higher frequency. However, frequent migration of tasks to cooler PEs, i.e., those closer to heat sink would result in an uneven aging across the stack. Consequently, cooler PEs may fail earlier than the those that are turned OFF more often.

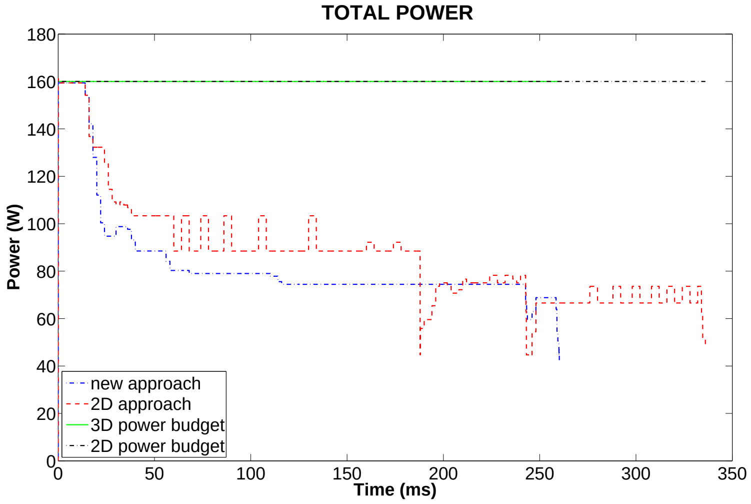


Figure 5.8: Total power dissipated of the considered stack when power budget is 160W and temperature constraint of 320K is imposed on each PE.

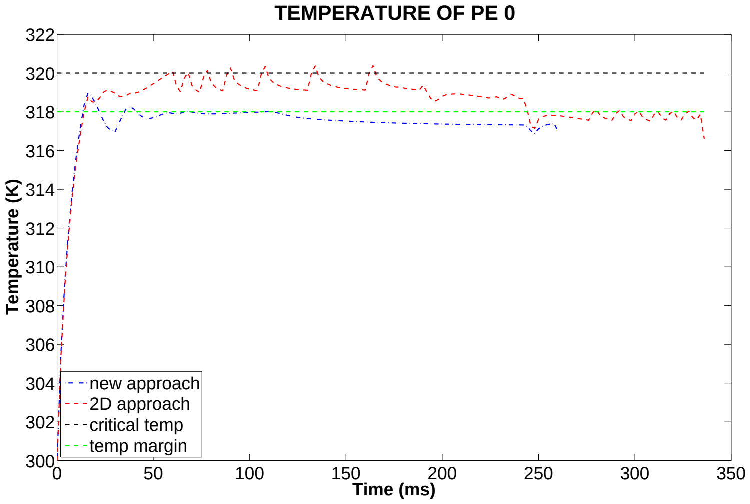


Figure 5.9: Temperature profile of PE 0 when power budget is 160W and temperature constraint of 320K is imposed on each PE.

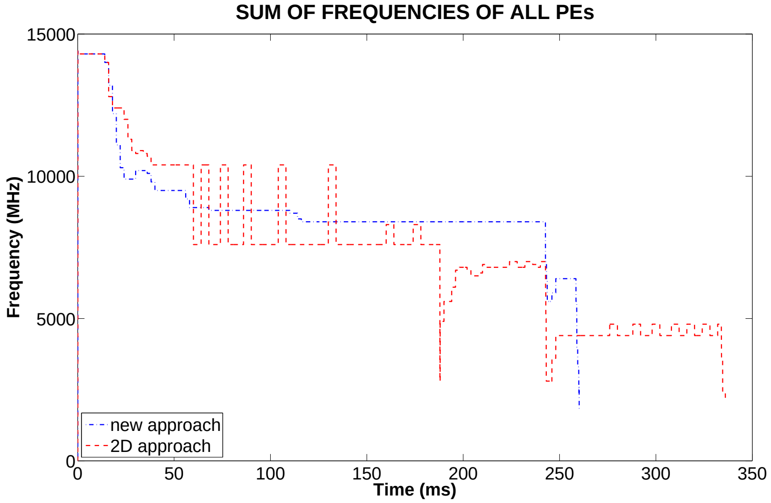


Figure 5.10: Sum of frequencies of all PEs when power budget is 160W and temperature constraint of 320K is imposed on each PE.

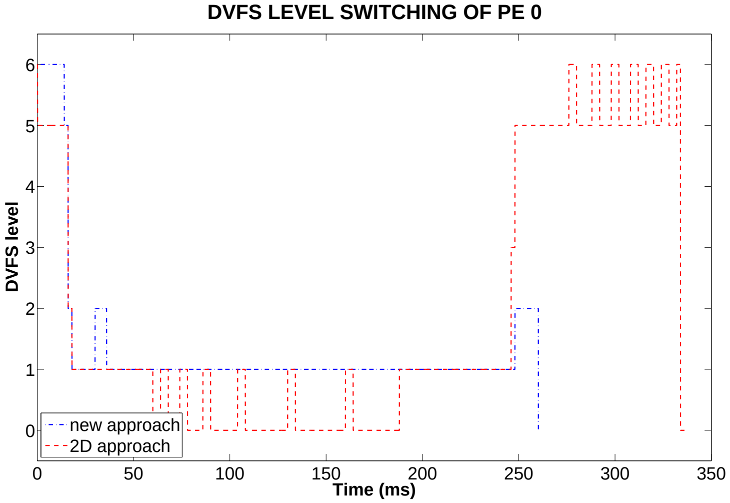


Figure 5.11: Operating V/F levels of PE 2 when power budget is 160W and temperature constraint of 320K is imposed on each PE.

5.2.3 Experiment 3

Experiment 3 is conducted to study the effect of vertical voltage islands in a 3D stack. DVFS in voltage islands is implemented.
Setup: Vertical islands for the considered stack, as shown in Figure 3.1, reproduced in (Figure 5.12). PE 0, PE 4 and PE 8 form Island 1, PE 1, PE 5 and PE 9 form Island 2, PE 2, PE 6 and PE 10 form Island 3, and PE 3, PE 7 and PE 11 form Island 4. Each PE is is allowed to operate at six V/F levels: (0.855V, 700MHz), (0.855V, 800MHz), (0.956V, 900MHz), (0.956V, 1000MHz), (1.048V, 1100MHz) and (1.048V, 1200MHz). Two frequencies are selected for each voltage value in order to include DFS benefits. A power budget of 160W is selected in order to test the ability to converge to set budget value with island. Critical temperature of each PE is taken to be 330K. With the Tmargin of $2K$, the PMB tries to keep the temperature of each PE below 328K but does not turn a PE OFF unless its temperature exceeds 330K.

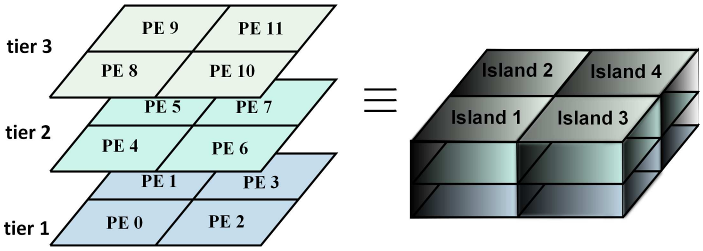


Figure 5.12: 3D stack with 4 voltage islands considered for experiments.

Analysis: Since there is a strong thermal relation between vertically adjacent PEs, assigning them to an island may help in meeting temperatures constraints along with power reduction. Experiment 3 was conducted to study the effect of these vertical voltage islands in 3D MPSoCs. The partitioning of islands and the considered stack was discussed in previous chapters. Power budget was taken to be 160W and a temperature constraint of 330K was imposed on all the PEs. The constraints are similar to those in Experiment 1.

Voltage islands were implemented using both 2D as well as new approach with results exhibiting similarity. Figure 5.13, Figure 5.15 and Figure 5.14 show the total power dissipation of the chip, sum of all frequencies and temperature of PE 0 respectively. New approach was again observed to result in more stable values of power and frequency with an increase of 2.9ms in execution time i.e., 1.36% of the time taken by 2D approach. The similarity in the results is due to the fact that PEs with strong thermal relation are a part of an island and are bound to operate on same voltage level. This constraint on their operating V/F levels ensures a smooth temperature profile as seen in Figure 5.14. The difference lies in relation between these islands. In 2D approach, the islands operate independently and the temeprature of any PE is controlled by the operating levels on the same island, whereas in the new approach, the effect of other islands are also considered and monitored. As a consequence of scaling multiple PEs either up or down at the same time more fluctions in power dissipation and temperature levels were observed compared to the per-core experiments. Since multiple frequencies were used for each operating voltage (enabling DFS), scaling of complete island was avoided when possible. The fluctuations would have been more prominent in the absence of DFS.

Since PEs in an island are bound to operate on same voltage level, their utilization observed to be similar. All PEs in an island suffered similar performance losses leading to earlier completion of tasks compared to the new approach in Experiment 1.

Comparison between Experiment 1 and Experiment 3 In per-core approach, PEs are scaled as and when necessary, while in case of islands, PEs on an island are bound to operate at same voltage levels. The performance losses in Experiment 3 are similar on all the tiers in an island, hence the PEs complete execution at almost the same time. In contrast, per-core scheme in Experiment 1 resulted in faster execution on tier 2 and 3 while slower execution on tier 1 while effectively maintaining the temperature levels. Islands can prove to be effective in cases where workloads of PEs are similar and similar performance results on each of them are expected. When performance of PEs differs, per-core scheme can achieve better performance on tiers higher in the stack while maintaining the temperatures on the deeper tiers. Nevertheless, the results obtained in Experiment 1 show that it can work well in both the cases. The total execution time seen in Figure 5.4 and Figure 5.13 suggest that Experiment 1 took more time to complete the assigned tasks. This is mainly due to the increase in execution time on the deepest tier. The sum of execution time of all PEs for Experiment 1 and Experiment 3 were essentially the same (difference of 0.002% only), which shows the effectiveness of per-core scheme in different scenarios.

DVFS comes with additional overhead of level shifters and voltage converters,which is larger in case of per-core DVFS, while voltage islands reduce these overheads and may become essential in designs with hundreds of PEs. The granularity and depth of islands can essentially be altered in a deep stack to achieve benefits of islands as well as per-core approach. Implementing such a scheme would also need to consider thermal relation between islands in order to control temperatures effectively. Islands higher up in the stack can achieve better performance, while considering the islands thermal relation with each other can effectively scale them down when temperatures on lower dies demand so. The differences are summarized in Table 5.2.

per − core	**voltage islands**
Scaled as and when necessary	PEs in an island are bound to operate on same V level
Higher performance on PEs closer to heat sink	Similar performance throughout the island
Performance losses may differ on different tiers	Similar performance on an island
Larger overhead of level shifters and voltage converters	Depends on the granularity of voltage islands

Table 5.2: Comparison between per-core DVFS and DVFS in voltage islands.

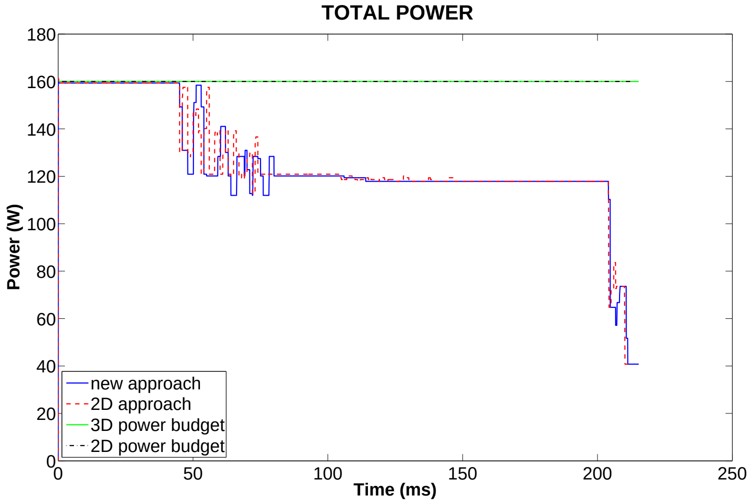


Figure 5.13: Total power dissipated of the island partitioned stack when power budget is 160W and temperature constraint of 330K is imposed on each PE.

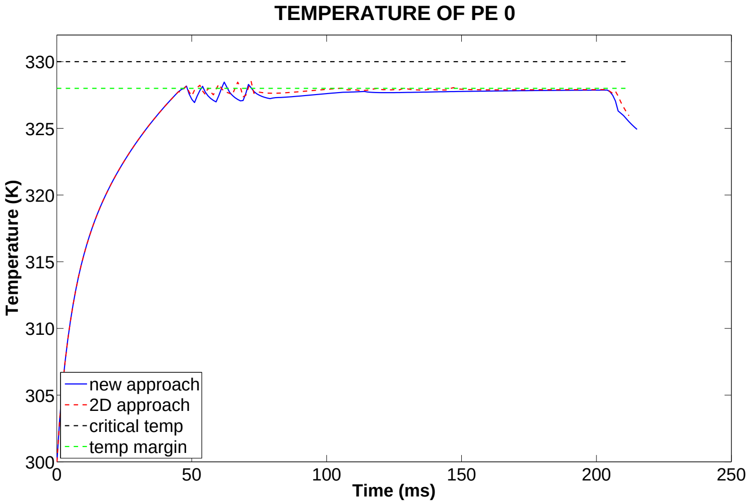


Figure 5.14: Temperature profile of PE 0 when power budget is 160W and temperature constraint of 330K is imposed on each PE.

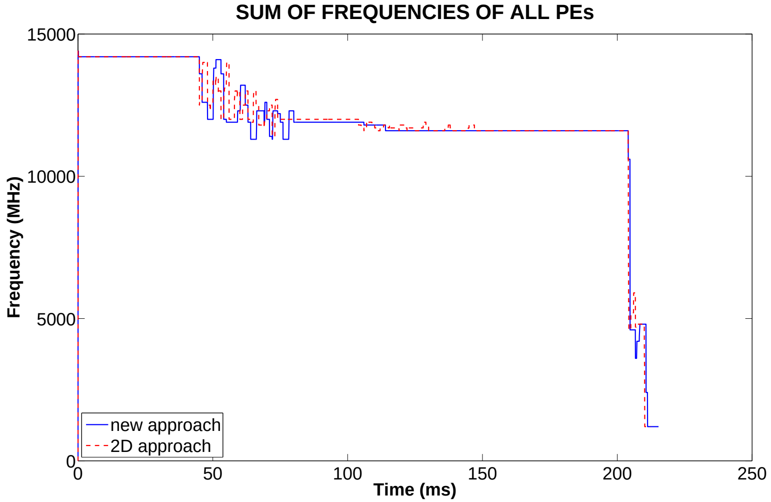


Figure 5.15: Sum of frequencies of all PEs when power budget is 160W and temperature constraint of 330K is imposed on each PE.

Conclusions and Future Work 6

This chapter summarizes the thesis and highlights the goals that were achieved in this thesis. The scope for future work is also presented.

6.1 Summary

Various power management schemes used in 2D ICs were studied and implications of using similar schemes in 3D MPSoCs were analyzed. Higher integration density in 3D stacks aggravates the prevailing challenges of power density and consequently microelectronics cooling. This makes consideration of temperature constraints important while designing power management schemes. DVFS in 2D ICs do not consider thermal relation between various PEs, however this cannot be ignored in 3D stacks. Temperatures of PEs in a 3D stack are highly interdependent.

A new temperature constraint power management schemes is proposed. The PMB accepts three inputs from the system− utilization of each PE, temperature of each PE and total chip power. While utilization and total power are taken at each control period, temperature values are taken at every temperature-check cycle. The Power Management Block (PMB) assigns weight to each PE depending on its instantaneous temperature margin, utilization, location in stack and area. These weights are used to scale V/F levels of each PE to meet power budget constraints while keeping the temperature at a safe margin from critical temperature. Thermal relation between any two PEs is given by effective thermal resistance between them. In cases of thermal runaway, the PMB uses utilization of each PE and the thermal relation between PEs to maintain the temperatures of each PE at a safe margin. Further, it imposes temporary constraints on the operating V/F levels of the PEs to avoid fluctuations in operating temperature due to constant switching of operating level.

In order to show the effectiveness of such a scheme, two sets on simulation setup were developed. Both were tested on the same target application, using same convergence algorithm, similar conditions for V/F level scaling and similar constraints on power and temperature. However, one setup used the proposed approach for DVFS in 3D MPSoCs and the other used an approach similar to DVFS in 2D chips where temperature and V/F levels of each PE is considered independently. Experiments were conducted to test the approach at per-core and island level under lenient and tight temperature constraints. At per-core level, under lenient temperature constraints, the overall performance figures for the proposed and 2D approach were comparable. However, the new approach could effectively maintain the temperatures in the stack and resulted in better stability by restricting repeated switching of operating V/F levels. Furthermore, it effectively maintained the temperatures on each PE at a margin from the critical temperature avoiding any significant fluctuations. In contrast, 2D approach showed significant fluctuations in operating V/F level and consequently fluctuations in operating temperatures were evident. It is important to avoid these temperature fluctuations as they degrade the performance of a device over time. Under lenient temperature constraints, 2D DVFS failed to maintain a margin from critical temperature on the deeper tiers while with tighter constraints, it led to significant switching between ON and OFF states. By effectively scaling the operating V/F levels of various PEs in the stack, PEs were observed to be in ON state under all tested circumstances resulting in an improvement of overall executing time by 19.55%. High performance was obtained on the PEs closer to heat sink when temperature constraints were lenient. Performance losses increased with tightness in temperature constraints in order to maintain temperature of all PEs at a margin from critical value.

In vertical voltage islands, the PEs with strong thermal relation are grouped together. This was observed to effectively maintain operating temperatures of all PEs at a margin from the critical temperature. Multiple frequencies were used for each operating voltage (enabling DFS) to avoid scaling of complete island when possible in order to keep fluctuations in total power and temperature insignificant. These PEs are bound to operate on same voltage levels, hence similar performance for similar workloads was observed across an island under all tested circumstances. Proposed approach at island level for similar workloads showed 6% reduction in overall execution time as compared to per-core level. However, different workloads on PEs in an island disturbs the balance between their performances. The proposed per-core scheme achieved an effective blend of the advantages from the vertical islands and a per-core DVFS scheme. Under lenient temeprature constraints, it allowed PEs on the upper tier to achieve higher performance while maintaining the temperatures on the lower tiers at decided safety margin, which would allow high perforamce application to run without significant losses. As the temeprature constraints get tighter, the performance losses on top tier increase in order to meet temperature contraints.

In a 3D stack with hundreds of PEs, voltage islands may become essential and a more practical approach due to the overhead of level shifters and voltage converters required to implement DVFS schemes. The depth of islands can essentially be altered in a deep stack to allow higher utilization on PEs near heat sink (benefit of per-core). However, the information of thermal relation between islands becomes important in such a scenario to effectively monitor the temperature. The proposed approach accounts these interdependencies in order to be able to scale closely related islands for maintaining temperature at a margin from the critical temperature. The approach showed an improvement of up to 19.55% in total execution time by considering these interdependencies for scaling V/F levels and preventing the PEs deeper in the stack from being turned OFF.

The thermal model and effective resistance matrix for the PEs are derived for the target floorplan. Therefore, the floorplan of the stack does not affect the power management scheme.

6.2 Future Work

The proposed power management block was designed keeping in mind the differences between 2D ICs and 3D ICs and the differences were successfully demonstrated. In order to do so, several simplifications were introduced which may need consideration in different scenarios.

1. The work assumes homogeneous workload on each PE and parameters relating power and operating V/F levels were computed off-line. For chips where PEs run heterogeneous workload, these parameters need to be updated using on-line feedback. [28] have demonstrated such on-line estimators with Recursive Least Square method with direction forgetting proposed in [42].

2. Although utilization of each PE was considered in form of weighted equations, direct priority for workloads heavier than a decided value may help monitoring performance losses in the PEs while still taking care of temperatures.

3. Effect of V/F level scaling on dependencies in a multiprocessor architecture and accurate performance figures would need integration of PMB with modified cycle accurate full system simulators like SESC [43].

4. Physical implementation of DVFS schemes in 2D ICs is essentially achieved by using multiple Vdd grid lines, dc converters and level shifters. However, in case of 3D stacked ICs, tiers are powered using TSVs. Further research is required for understanding how multiple supply voltages can be delivered on different tiers, its feasibility and consequent overheads.

3D-ICE Simulator A

3D-ICE is a Linux based generic simulation platform written in C to simulate the transient thermal analyses of vertically stacked 3D ICs with/without micro-channel liquid cooling [21].

A.1 Introduction

3D-ICE accesses all the information needed to emulate a 3D IC from two different types of input file:

- Stack Description File
- Floorplan File

A.2 Stack Description File

The stack description file (*.stk) is a netlist that specifies all the physical and geometrical properties of the 3D-IC for the simulation. The extension of the file is not relevant-it will beparsed independent of its presence or content. The stack description file contains six main sections (mandatory and optional) and they must be declared following this order:

1. Materials
2. Conventional Heat Sink
3. Channel (not used)
4. Die
5. Stack
6. Dimensions

A.2.1 Materials

The first section of the file contains the list of materials and their properties to be used in the simulation. At least one material must be declared. Materials are declared with the syntax,

```
material MATERIAL_ID : thermal conductivity DVALUE ;
volumetric heat capacity DVALUE ;
```

where

- MATERIAL_ID is a unique identifier to refer to this material,
- thermal conductivity is expressed in $W\mu m^(-1)K^(-1)$,
- volumetric heat capacity is expressed in $J\mu m^3K$.

Example

```
        material SILICON :
               thermal conductivity      1.30e-4 ;
               volumetric heat capacity  1.628e-12 ;
```

A.2.2 Dies

A die is a group of layersstacked together to form a single entity that is used when declaring the sequence of stacked elements of the 3D IC later in the file. This can represent an actual IC die in the stack. You can declare multiple dies, and use a single die multiple times during the stack description. The Dies section is a mandatory section and must contain at least one die element. A die must contain one source layer (the term source layer is used to denote those layers of the stack which contain active electronic components, and hence, provide the heat source for the simulation) and zero or more passive layers. The source layer can be placed at any location in the stack of layers in a die.

```
        die DIE_ID :
                 [layer IVALUE MATERIAL_ID ; ]
                 source IVALUE MATERIAL_ID ;
                 [layer IVALUE MATERIAL_ID ; ]
```

where

- DIE_ID is the unique identifier used to refer to the declared die;
- IVALUE is the height of the layer (in μm);
- MATERIAL_ID is the (previously declared)identifier of the material composing the layer.

The order of the layers within the die reflects their vertical disposition in the 3D IC, i.e., the first layer declared is the top most layer in the die (closer to the ambient) while the last one is the one at the bottom (closer to the PCB).

A.2.3 Conventional Air-Cooled Heat Sink

This is an optional section which includes a conventional air-cooled heat sink in the 3D IC. All the faces of the 3D IC stack are modeled as adiabatic walls by default. When the Conventional Heat Sink is specified, the top surface of the stack is connected to the ambient via a thermal resistance.

```
conventional heat sink :
        heat transfer coefficient DVALUE ;
         ambient temperature DVALUE ;
```

where

- heat transfer coefficient of the heat sink is expressed in $W\mu m^2 K$;
- ambient temperature is the ambient temperature expressed in K.

A.2.4 Stack

This section builds the vertical structure of the stack. The stack is composed of Dies (as previously declared) and layers.

```
stack :
        [layer LL_ID DVALUE MATERIAL_ID ; ]
        [die DD_ID DIE_ID floorplan "PATH" ; ]
```

where

- LL_ID, CC_ID and DD_ID are identifiers used to name the stack elements and they can be used in the simulator code to refer to the corresponding element. They must be unique for each element.
- MATERIAL_ID is the identifier of the material (as previously declared) composing the declared layer.
- DVALUE is its height of the layer (in μm).
- DIE_ID is the identifier of a die (as previously declared) and PATH is the path to the floorplan file. This floorplan will be placed on the declared source layer in the definition of the die. The floorplan files contain information of the location and power dissipation activity of various floorplan components for the given die. The same DIE_ID can be used multiple times (with different identifiers DD_ID) in a stack with the same or different floorplans, if identical/similar dies exist in a single IC.

A.2.5 Dimensions

The last section of the Stack Description File declares the xy dimensions of the entire chip and the discretization sizes for the thermal cells (all in μ).

```
        dimensions :
                 chiplength DVALUE , width DVALUE ;
                 cell length DVALUE , width DVALUE ;
```

Height of the thermal cell is taken to be equal to the layer height.

A.3 Floorplan File

Every die in the stack must be related to a F̈loorplan File(̈*.flp), which essentially provides the power dissipation profile (or heat sources) for the simulation. Each Floorplan file must contain the list of functional blocks (cores, caches, memories, etc), their positions, and the power dissipation as a function of time. Every functional block, here called floorplan element, is a rectangular area inside the die, laid out in the source layer. Each floorplan element has a unique identifier− the name it is assigned. In addition, the position and the dimensions of each floorplan element are given (in μ) based on the same Cartesian coordinates that was used for building the stack, with the origin at the SOUTH-WEST corner of the source layer.
A floorplan element in the Floorplan File is declared using the following syntax.

```
        IDENTIFIER :
                 position DVALUE , DVALUE ;
                 dimension DVALUE , DVALUE ;
                 power values DVALUE [ , DVALUE ] ;
```

where

- IDENTIFIER is the unique identifier used to name the floorplan element. This string must be unique within the floorplan file it belongs to but it can be used on a different file.
- position, expressed in (in μm), is the (x,y) coordinate of the SOUTH-WEST corner of the floorplan element.
- dimension is the (length, width) dimensions of the floorplan element (in μm).
- The DVALUE(s) against the keyword power values are the list of power dissipation values (expressed in W) of the floorplan element for each time slot (scroll down for the explanation of time slots in 3D-ICE) separated by commas.

A.4 Used stack file

The stack file used for thermal simulations and for generating effective resistance matrix is shown below.

```
material silicon :
   thermal conductivity       1.60e-04 ;
   volumetric heat capacity   1.66e-12 ;
material metal :
   thermal conductivity   12.0e-06 ;
   volumetric heat capacity   3.4419e-12 ;
material bond :
   thermal conductivity   6.83e-06 ;
   volumetric heat capacity   3.99e-12 ;
material TIM :
   thermal conductivity   5.0e-06 ;
   volumetric heat capacity   4.0e-12 ;
material spreader :
   thermal conductivity   4.00e-04 ;
   volumetric heat capacity   3.55e-12 ;
conventional heat sink :
   heat transfer coefficient  1.20e-07;
   ambient temperature        300 ;
die     die_1 :
         source      50     silicon;
          layer      15     metal;
die     die_2 :
         source      50     silicon;
          layer      15     metal;
          layer      10     bond;
die     die_3 :
          layer    1000     spreader;
          layer      50     TIM;
          layer     200     silicon;
         source      50     silicon;
          layer      15     metal;
          layer      10     bond;
stack:
   die       die3      die_3    floorplan   "fp_die3.flp";
   die       die2      die_2    floorplan   "fp_die2.flp";
   die       die1      die_1    floorplan   "fp_die1.flp";
dimensions :
chip length       9000  ,   width      9000;
cell length        100  ,   width      100;
```

A.5 Sample floorplan File

Floorplan for tier 1 of the 3D stack considered in this thesis is shown below. The power values were generated by the testbench for the purpose of thermal simulation. For generating effective thermal resistance matrix, repeated simulations were performed where only one PE was allowed to dissipate power at a time.

```
PE11:
      position       0 ,         0 ;
     dimension    4500 ,      4500 ;
  power values   10,  15;
PE12:
      position       0 ,      4500 ;
     dimension    4500 ,      4500 ;
  power values   10,   10;
PE13:
      position    4500 ,         0 ;
     dimension    4500 ,      4500 ;
  power values   10,   10;
PE14:
      position    4500 ,      4500 ;
     dimension    4500 ,      4500 ;
  power values   10,   10;
```

A.6 Running 3D-ICE

The simulator is run using the following command
./Emulate3DICe stackfile.stk time_slot_DVALUE delta_DVALUE
where

- slackfile.stk is the path to the Stack Description File containing the description of the 3D-IC.
- time_slot_DVALUE is the duration of each time slot (in seconds) for which power values specified in the Floorplan File(s) are held constant. This value, multiplied by number of power values gives the total time is seconds.
- delta_DVALUE is the time step value (in seconds) for the numerical integration of the system equations.

e.g. ./Emulate3DICe stackfile.stk .001 .0001
This command will generate the thermal profile for the stack described in stackfile.stk. The power values given provided in floorplan file as assumes to be constant for the interval of 1ms. And thermal simulation is done with time steps of 0.1ms each. For floorplan file with 2 power values, the output thermal profile would give temperature values from 0.1ms to 2ms at time steps of 0.1ms where power dissipation of PE11 for first 1ms is 10W and for next 1ms is 15W.

Bibliography

[1] Joohee Kim, Jun So Pak, Jonghyun Cho, Eakhwan Song, Jeonghyeon Cho, Heegon Kim, Taigon Song, Junho Lee, Hyungdong Lee, Kunwoo Park, Seungtaek Yang, Min-Suk Suh, Kwang-Yoo Byun, and Joungho Kim. High-frequency scalable electrical model and analysis of a through silicon via (tsv). *IEEE Transactions on Components, Packaging and Manufacturing Technology*, 1(2):181–195, 2011.

[2] Michael Keating, David Flynn, Rob Aitken, Alan Gibbons, and Kaijian Shi. *Low Power Methodology Manual: For System-on-Chip Design*. Springer Publishing Company, Incorporated, 2007.

[3] International technology roadmap for semicondctors, report 2009.

[4] Bryan Black, Murali Annavaram, Ned Brekelbaum, John DeVale, Lei Jiang, Gabriel H. Loh, Don McCaule, Pat Morrow, Donald W. Nelson, Daniel Pantuso, Paul Reed, Jeff Rupley, Sadasivan Shankar, John Shen, and Clair Webb. Die stacking (3d) microarchitecture. In *Proc. MICRO-39 Microarchitecture 39th Annual IEEE/ACM Int. Symp*, pages 469–479, 2006.

[5] A. Sridhar, A. Vincenzi, M. Ruggiero, T. Brunschwiler, and D. Atienza. 3d-ice: Fast compact transient thermal modeling for 3d ics with inter-tier liquid cooling. In *Proc. IEEE/ACM Int Computer-Aided Design (ICCAD) Conf*, pages 463–470, 2010.

[6] A. Jain, R. E. Jones, R. Chatterjee, S. Pozder, and Zhihong Huang. Thermal modeling and design of 3d integrated circuits. In *Proc. 11th Intersociety Conf. Thermal and Thermomechanical Phenomena in Electronic Systems ITHERM 2008*, pages 1139–1145, 2008.

[7] Chong Sun, Li Shang, and R. P. Dick. Three-dimensional multiprocessor system-on-chip thermal optimization. In *Proc. 5th IEEE/ACM/IFIP Int Hardware/Software Codesign and System Synthesis (CODES+ISSS) Conf*, pages 117–122, 2007.

[8] Feihui Li, C. Nicopoulos, T. Richardson, Yuan Xie, V. Narayanan, and M. Kandemir. Design and management of 3d chip multiprocessors using network-in-memory. In *Proc. 33rd Int. Symp. Computer Architecture ISCA '06*, pages 130–141, 2006.

[9] Canturk Isci, Alper Buyuktosunoglu, Chen-Yong Cher, Pradip Bose, and Margaret Martonosi. An analysis of efficient multi-core global power management policies: Maximizing performance for a given power budget. In *Proc. MICRO-39 Microarchitecture 39th Annual IEEE/ACM Int. Symp*, pages 347–358, 2006.

[10] G. Semeraro, G. Magklis, R. Balasubramonian, D. H. Albonesi, S. Dwarkadas, and M. L. Scott. Energy-efficient processor design using multiple clock domains with dynamic voltage and frequency scaling. In *Proc. Eighth Int High-Performance Computer Architecture Symp*, pages 29–40, 2002.

[11] Qiang Wu, P. Juang, M. Martonosi, and D. W. Clark. Voltage and frequency control with adaptive reaction time in multiple-clock-domain processors. In *Proc. HPCA-11 High-Performance Computer Architecture 11th Int. Symp*, pages 178–189, 2005.

[12] S. Herbert and D. Marculescu. Analysis of dynamic voltage/frequency scaling in chip-multiprocessors. In *Proc. ACM/IEEE Int Low Power Electronics and Design (ISLPED) Symp*, pages 38–43, 2007.

[13] Changyun Zhu, Zhenyu Gu, Li Shang, R. P. Dick, and R. Joseph. Three-dimensional chip-multiprocessor run-time thermal management. 27(8):1479–1492, 2008.

[14] K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat. 3-d ics: a novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration. 89(5):602–633, 2001.

[15] Chang-Tzu Lin, Ding-Ming Kwai, Yung-Fa Chou, Ting-Sheng Chen, and Wen-Ching Wu. Cad reference flow for 3d via-last integrated circuits. In *Proc. 15th Asia and South Pacific Design Automation Conf. (ASP-DAC)*, pages 187–192, 2010.

[16] Jan M. Rabaey, Anantha Chandrakasan, and Borivoje Nikolic. *Digital Integrated Circuits.* ISBN 0-13-090996-3. Prentice Hall, 2nd edition edition.

[17] Wei Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. R. Stan. Hotspot: a compact thermal modeling methodology for early-stage vlsi design. 14(5):501–513, 2006.

[18] Donald A. Neamen. *Semiconductor Physics and Devices.* ISBN 0070617120. Mc-Graw Hill, 3rd edition edition.

[19] Wei Huang. *HotSpotA Chip and Package Compact Thermal Modeling Methodology for VLSI Design.* PhD thesis, University of Virginia, Charlottesville, Virginia, 2007.

[20] Yuan Xie. Processor architecture design using 3d integration technology. In *Proc. 23rd Int. Conf. VLSI Design VLSID '10*, pages 446–451, 2010.

[21] url = http://prex.sourceforge.net/doc/power.html.

[22] U. Y. Ogras, R. Marculescu, P. Choudhary, and D. Marculescu. Voltage-frequency island partitioning for gals-based networks-on-chip. In *Proc. 44th ACM/IEEE Design Automation Conf. DAC '07*, pages 110–115, 2007.

[23] C. Seiculescu, S. Murali, L. Benini, and G. De Micheli. Comparative analysis of nocs for two-dimensional versus three-dimensional socs supporting multiple voltage and frequency islands. 57(5):364–368, 2010.

[24] Shih-An Yu, Pei-Yu Huang, and Yu-Min Lee. A multiple supply voltage based power reduction method in 3-d ics considering process variations and thermal effects. In *Proc. Asia and South Pacific Design Automation Conf. ASP-DAC 2009*, pages 55–60, 2009.

[25] url = http://prex.sourceforge.net/doc/power.html.

[26] A. K. Coskun, T. S. Rosing, and K. C. Gross. Proactive temperature balancing for low cost thermal management in mpsocs. In *Proc. IEEE/ACM Int. Conf. Computer-Aided Design ICCAD 2008*, pages 250–257, 2008.

[27] Wonyoung Kim, M. S. Gupta, Gu-Yeon Wei, and D. Brooks. System level analysis of fast, per-core dvfs using on-chip switching regulators. In *Proc. IEEE 14th Int. Symp. High Performance Computer Architecture HPCA 2008*, pages 123–134, 2008.

[28] Xiaorui Wang, Kai Ma, and Yefu Wang. Adaptive power control with online model estimation for chip multiprocessors. 22(10):1681–1696, 2011.

[29] Mohamed M. Sabryz, David Atienzaz, and Ayse K. Coskuny. Thermal analysis and active cooling management for 3d mpsocs. In *Proc. IEEE Int Circuits and Systems (ISCAS) Symp*, pages 2237–2240, 2011.

[30] José L. Ayala, Arvind Sridhar, and David Cuesta. Invited paper: Thermal modeling and analysis of 3d multi-processor chips. *Integr. VLSI J.*, 43:327–341, September 2010.

[31] C. Bostak J. Ignowski M. Millican W. H. Parks R. McGowen, C. A. Poirier and S. Naffziger. Power and temperature control on a 90-nm itanium family processor. *IEEE Journal of Solid-State Circuits*, 41:229–237, 2006.

[32] Ramya Raghavendra, Parthasarathy Ranganathan, Vanish Talwar, Zhikui Wang, and Xiaoyun Zhu. No power struggles: Coordinated multi-level power management for the data center. In *ASPLOS*, 2008.

[33] Xiaorui Wang; Ming Chen. Cluster-level feedback power control for performance optimization. pages 101–110, 2008.

[34] C. J. M. Lasance, H. I. Rosten, and J. D. Parry. The world of thermal characterization according to delphi-part ii: Experimental and numerical methods. *IEEE Transactions on Components, Packaging, and Manufacturing Technology, Part A*, 20(4):392–398, 1997.

[35] http://www.simplescalar.com/.

[36] http://www.eecs.umich.edu/mibench/.

[37] M.R. Guthaus, J.S. Ringenberg, D. Ernst, T.M. Austin, T. Mudge, and R.B. Brown. Mibench: A free, commercially representative embedded benchmark suite. In *Workload Characterization, 2001. WWC-4. 2001 IEEE International Workshop on*, pages 3 – 14, dec. 2001.

[38] http://www.eecs.harvard.edu/ dbrooks/wattch-form.html.

[39] David Brooks, Vivek Tiwari, and Margaret Martonosi. Wattch: A framework for architectural-level power analysis and optimizations. In *In Proceedings of the 27th Annual International Symposium on Computer Architecture*, pages 83–94, 2000.

[40] Sachin S. Sapatnekar. Addressing thermal and power delivery bottlenecks in 3d circuits. In *Proceedings of the 2009 Asia and South Pacific Design Automation Conference*, ASP-DAC '09, pages 423–428, Piscataway, NJ, USA, 2009. IEEE Press.

[41] R. Kumar and V. Kursun. Impact of temperature fluctuations on circuit characteristics in 180nm and 65nm cmos technologies. In *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, page 4 pp., may 2006.

[42] Xue Liu, Xiaoyun Zhu, P. Padala, Zhikui Wang, and S. Singhal. Optimal multivariate control for differentiated services on a shared hosting platform. In *Proc. 46th IEEE Conf. Decision and Control*, pages 3792–3799, 2007.

[43] http://sourceforge.net/projects/sesc.