THERMAL MODELING OF 3-DIMENSIONALLY STACKED DRAM MEMORY

By

RATNESH RAJ

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN MECHANICAL ENGINEERING

THE UNIVERSITY OF TEXAS AT ARLINGTON

AUGUST 2017

ACKNOWLEDGEMENT

I would like to take this time to thank my master thesis advisor Dr. Ankur Jain for his consistent academic and moral support that he has shown towards me. I am grateful for all the time that he spent with me in helping me understanding the concepts and helping me grow an individual.

I would like to thank Dr. Ratan Kumar and Dr. Miguel Amaya for taking their valuable time to serve on my thesis committee. Also, I would like to thank Ms. Flora Pinegar and Ms. Debi Barton for their invaluable support that they have shown since I started my master program.

I would like to thank all my lab mates in the Microscale Thermophysics Lab for their consistent academic and moral support. I would also like to specifically thank Dr. Daipayan Sarkar for his advice and inputs in helping me with my work

Lastly, this wouldn't have been possible without the consistent support of my friends and family who have always been with me through my highs and lows.

June 22, 2017

ABSTRACT

THERMAL MODELING OF 3-DIMENSIONALLY STACKED DRAM MEMORY

Ratnesh Raj
University of Texas at Arlington, 2017

Advisor: Dr. Ankur Jain

A large fraction of energy consumed in modern microelectronic devices and systems is taken up by memory access operations, which is expected to cause significant temperature rise. Since memory access operations are very short, this is expected to inherently be a transient thermal phenomenon. Despite the critical importance of thermal management in microelectronics, not much work exists on understanding the nature of thermal transport during memory access operations. In this work, a mathematical model to predict the transient temperature rise within a 3D layered memory chip is presented. Most heat-generating memory access processes occur over a short timescale for which the thermal penetration depth is shorter than the die thickness. This enables the modeling of such processes independent of the nature of chip cooling by treating the chip as a combination of semi-infinite and infinite medium layered bodies. A semi-infinite Green's function model is developed for topmost layer of memory. The subsequent layers: 2 to n layers have been modeled on a thermally infinite layer concept. This model is validated against finite element simulation results. The analytical model is used to analyze transient thermal effects of various memory access processes for multiple banks. Finally, a thermal simulator based on 4 layers of

memory is presented which shows the capability of any input sequence of operation along the x, y, z, t directions and resulting temperature rise due to its access. This is used as a limiting case to show that the maximum number of accesses that can be performed on a 4-layered structure where each layer consists of 10x10 grid, i.e. 400 cells. These results will help develop an understanding of optimal layouts and processes for 3D memory chips, eventually leading to co-design tools that simultaneously improve thermal and electrical performance of 3D memory chips.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Chapter 1

Introduction

Thermal management in electronic systems has been an interest for several decades. A key aspect of the analysis lies in the in the increasing computational power of microprocessors. From the data centers to smartphones, around 30% of the power is consumed by the memory chips [1–3]. Specifically, Lefurgy et al. showed that memory in data server had 50% higher consumption of storing power than their processor cores [2]. Large storage of data requirements has also increased the demand of computational capacity in memory leading to increase in number of cores on a single memory chip [4-9], thus requiring efficient performance management of memories. Scaling capabilities due to CMOS technology has opened huge trade-offs for a thermally aware system. Major contributors include the high transistor packing density

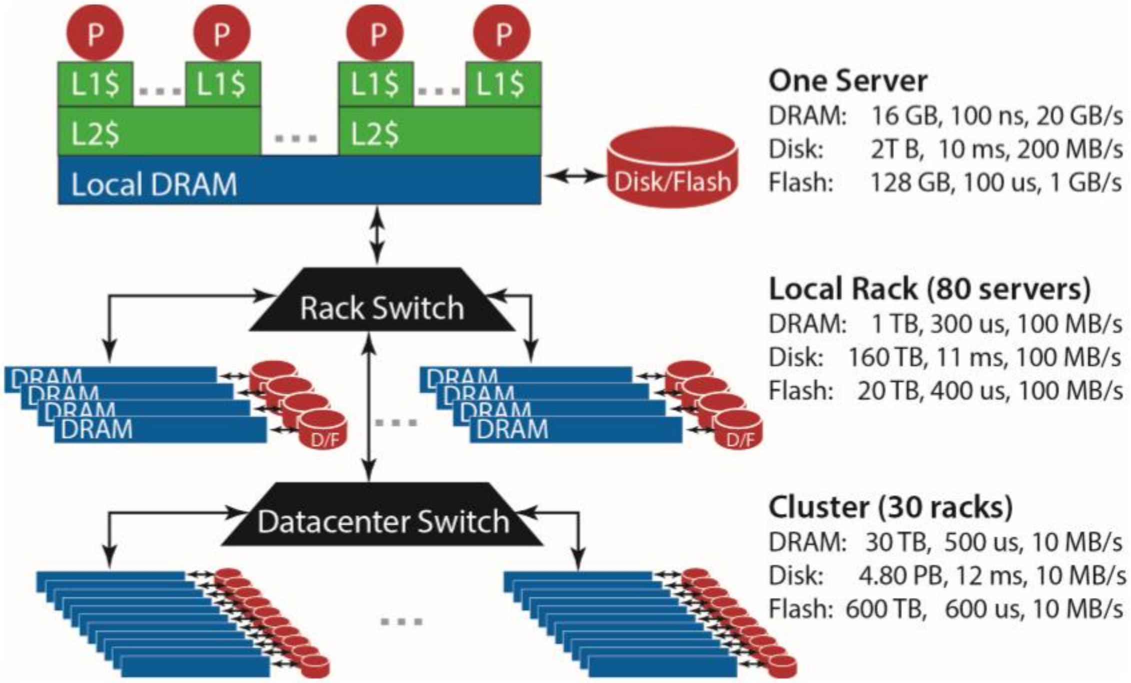

P
P
P
P
L1$
L1$
L1$
L1$
L2$
L2$
Local DRAM
Disk/Flash
One Server
DRAM: 16 GB, 100 ns, 20 GB/s
Disk: 2T B, 10 ms, 200 MB/s
Flash: 128 GB, 100 us, 1 GB/s
Rack Switch
DRAM
DRAM
DRAM
DRAM
D/F
DRAM
DRAM
DRAM
DRAM
D/F
Local Rack (80 servers)
DRAM: 1 TB, 300 us, 100 MB/s
Disk: 160 TB, 11 ms, 100 MB/s
Flash: 20 TB, 400 us, 100 MB/s
Datacenter Switch
Cluster (30 racks)
DRAM: 30 TB, 500 us, 10 MB/s
Disk: 4.80 PB, 12 ms, 10 MB/s
Flash: 600 TB, 600 us, 10 MB/s

Figure 1.1: Schematic shows the storage hierarchy of Google's Warehouse-Scale Computer (WSC) datacenter [1]

as the reduction in size continues as shown in Figure 1.1. On the one hand, this has improved the computational capability, but has significantly limited the design to be thermally friendly by increasing the power consumption, thus allowing heating in the system [7].

This calls for improving the thermal efficiency of the memory microprocessors as they continue to scale further down in size. Figure 1.2 shows the power distribution of the major components in a modern IT setup.

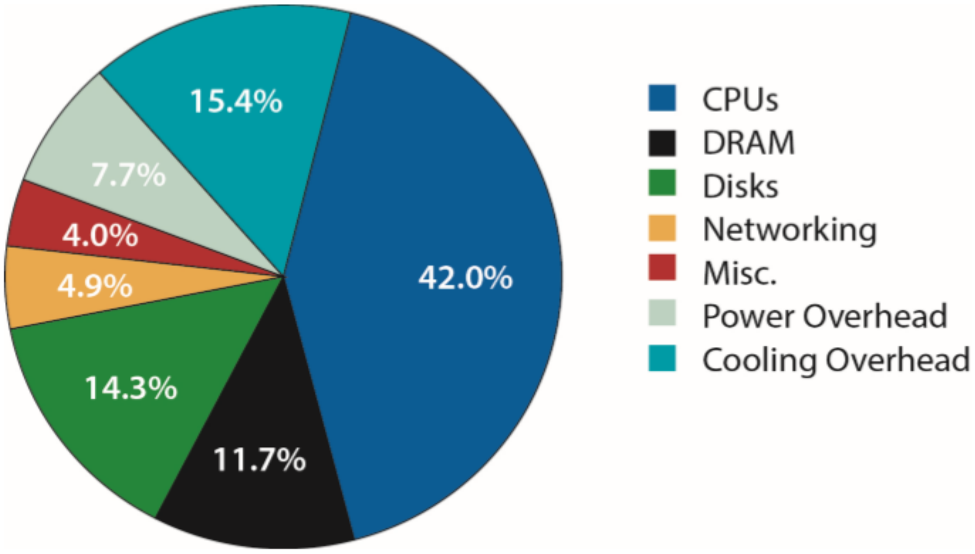


Figure 1.2 Distribution of peak power usage a modern IT/hardware equipment in a data center tower. [1]

Many analytical thermal modeling approaches have been adopted for heat transfer processes on a microelectronic chip [10-15]. These include completely analytical approaches for solving governing energy equation [16-17], as well as finite-difference based numerical computation approaches. Green’s function approach has also been adopted for computing temperature distribution on a microelectronic chip [18]. This paper presents an analytical model for understanding thermal transport and optimizing thermal management of the power distribution within the banks.

A memory chip can be subdivided into two distinct sections: internal computer memory and flash/auxiliary memory. Internal memory is classified as volatile memory “stores” data for a limited amount of time and usually “forgets” after the operation is terminated. Its storage ranges from 512MB to 4GB on conventional memory. Auxiliary Memory is a non-volatile type which stores data even after the operation has been terminated, i.e. hard drives or flash memory [19]. The primary focus of this work deals with internal memory type called DRAM (Dynamic Random-Access Memory). Conventional DRAM architecture contains bit-cells at the smallest addressable region. These are arranged in a grid like manner having columns and rows. An access is directed towards these bit-cells which house transistors and capacitors and are sensed together at the instance of receiving a signal. An array of bit-cells is called bank. These banks are all linked to the same data bus width as the external output bus width. Set of banks are ranks which operate in tandem to service requests from the memory controller. The DRAM modules are placed on the PCB and referred to DIMM (Dual-in-line memory module) which provide an interface to the memory bus [20]. The focus of this analysis is towards a more simplified design which shrinks to two bit-cells that are placed on a single bank of the DRAM as shown in Figure 1.3.

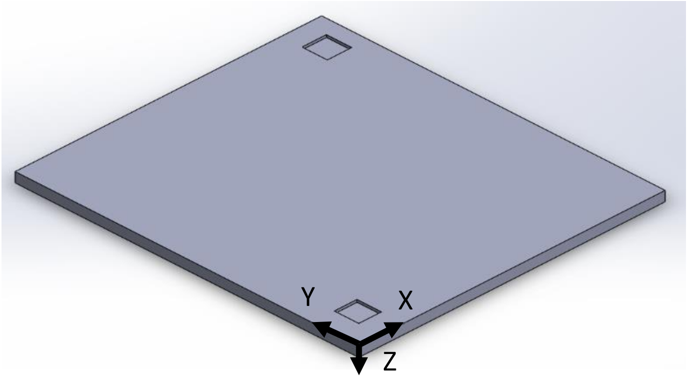


Figure 1.3 3D model of a single bank with bit cells.

When an operation is performed on a module, the power is directed toward a selected rank, bank and row [21]. The power supply on DRAMs can be divided into three different categories: activation power, read/write power, background power [22-24]. The activation power is required when activating a memory array row and in pre-charging the arrays bit lines. The read/write power is when data is transferred. The background power is residual power dissipation mainly consumed by transistor leakage after the access has occurred. The DRAM memory cells store data using capacitors that lose their charge over time and must be periodically recharged. The transistors, embedded in the module, act as amplifiers and read the contents of the row. Consequently, these transistors heat up and dissipate heat which causes the local temperature rise.

Multiple accesses in these transistors leads to an effective temperature rise at the junction thereby, affecting the overall memory chip. An analytical model is presented in this work, which captures the effect of each memory activation, read/write, and background operation. Upon a DRAM read, the sequence is sent to the rank, bank and row. This is then sent to the row buffer which use sense amplifiers to direct the data. Row buffer then directs the data to the memory controller, which directs it to the processor [21]. This accessing operation is solved using the Green’s function solution approach. The methodology is primarily effective due to its unique ability to capture temperature rise occurring in pulse heating fashion [25]. This paper aims at providing the solution to variable 2D hotspots regions, which are a result to accessing of bit cells. In addition, the access rate of the bit-cells allows the temporal aspect to be captured. This allows the model to mimic the behavior of a realistic DRAM operation. Another feature of this papers shows a simplistic sequencing approach which helps emulate the effect of two access one after the other, aiding the architect in designing the DRAM by providing the overall temperature rise in the system.

Chapter 2

Methodology and Validation

This chapter describes the derivation used to derive the 3D thermal model. The continuum Fourier heat conduction partial differential equation is used as the governing energy equation. The model is 3-dimensional in nature due to the nature of the problem and transient due to the time dependency of the model. The governing energy equation is solved using a Green's function approach. The derivation is split into 2 phases:

1) Topmost layer is modeled using the semi-infinite layer theory.

2) Subsequent layers, i.e. 2 to n layers is modeled using the infinite layer concept.

The model semi-infinite model is validated against the commercial finite element method software ANSYS. The validation is performed for two different cases:

1) Spatially varying heat flux

2) Spatially and temporally varying heat flux

3) Infinite Layer Model Validation

2.1 Mathematical Model

Heat transfer within the 3D memory chip is modeled to obey Fourier law that governs three-dimensional transient heat conduction with no internal heat generation [26]. This equation captures the heat transfer within a single bank which is modeled in three-dimensional manner. The Laplacian below captures the conduction heat transfer that takes places within the bank. The transient term is accounted for since the access of transistors within the bank is transient in nature and releases heat, which leads to increase of bank temperature, thereby increasing the overall temperature of the memory block.

$$\frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}+\frac{\partial^2 T}{\partial z^2}=\frac{1}{\alpha}\frac{\partial T}{\partial t} \tag{1}$$

In equation (1), T refers to as temperature rise above ambient. The bits of data storage cells will henceforth be referred to as patches, and are modeled as three-dimensional body with the following boundary conditions and initial condition. The sides of the geometry in x and y direction are assumed to be adiabatic due to the extremely small thickness of the model as compared to the other two dimensions, as a result heat does not escape from the sides.

$$\frac{\partial T}{\partial x}=0 \quad \text{at x =0, a} \tag{2}$$

$$\frac{\partial T}{\partial y}=0 \quad \text{at y =0, b} \tag{3}$$

The heat flux is assumed to be a thin x-y planar heat source which can be located anywhere in the domain and vary only as a function of space and time. According the reference co-ordinate axis selected in the model schematic, the surface heat flux is applied at the z=0 or the top surface of the body. The reason for selecting this is when power is supplied to access the bit cells, the non-homogeneity can be modeled as a surface heat flux as it only heats the transistors allocated to that bit- cell. The thermal conductivity of the domain is assumed to be isotropic, since the banks are assumed to follow the conventional DRAMs technology that are currently in-use today and not the phase change materials.

$$-k\frac{\partial T}{\partial z} = q(x,y,t) \text{ at z =0} \tag{4}$$

The accessing time for each memory patch is less compared to the time it takes for the thermal wave, after the being accessed, to reach the boundary. Due to this behavior, the model is independent of the physical boundary in z-direction. Hence, a semi-infinite boundary condition is assumed.

$$\lim_{z\to\infty}|T(x,y,z,t)| < \infty \tag{5}$$

It is assumed that the chip is initially at zero temperature,

$$T(x,y,z,0) = 0 \tag{6}$$

Similar methodology is adopted for the subsequent layers after the topmost layer. Thermally infinite layers have been assumed for these layers. The reason for selecting subsequent layers as infinite layers is that the thermal penetration depth of heat after each access in the respective transistor is smaller than the die thickness. In other words, due to the short timescale of the operation/access, the thermal wave generated by these access is never able to see the end boundary. And since these subsequent layers do not see the energy from the above layers, the concept of the infinite layer fits well in describing their thermal energy transport process.

The governing energy equation used for the infinite layer model is:

$$\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} + \frac{g(x,y,z,t)}{k} = \frac{\partial T}{\partial t} \tag{7}$$

This equation is significantly different from the one used for the semi-infinite case. The major difference is in the internal heat generation term. This is used since in the infinite layer, there are no boundary on which the access can be modeled. The volumetric energy source helps to model that part easily since it requires the spatial coordinates and the temporal coordinates.

The spatial and temporal domain for a single infinite layer is modeled as:

$$0 < \mathrm{x} < \mathrm{a} \tag{8}$$

$$0 < \mathrm{y} < \mathrm{b} \tag{9}$$

$$-\infty < \mathrm{z} < \infty \tag{10}$$

$$0 < \tau < \mathrm{t} \tag{11}$$

As expected, the domain is finite in the x and y direction since the edges are bounded. The domain in z-direction is modeled as infinite in both positive and negative directions. The temporal domain starts from 0 to t, where the t is the final time of the operation.

The boundary conditions used for this model are as follows:

X-direction: $\frac{\partial \boldsymbol{T}}{\partial \boldsymbol{x}} = 0 \; at \; x = 0, a$ (12)

Y-direction: $\frac{\partial \boldsymbol{T}}{\partial \boldsymbol{y}} = 0 \; at \; y = 0, b$ (13)

The initial condition is:

$$T(x, y, z, 0) = 0 \tag{14}$$

The planar heating condition in the governing energy equation is modeled as:

$$g(x,y,z,t) = \left\{ \begin{array}{ll} 0, & x < \mathrm{a1},\ \mathrm{x} > \mathrm{a2}, \\ & \mathrm{y} < \mathrm{b1},\ \mathrm{y} > \mathrm{b2}, \\ & t < tl,\ t > tu \\ & z' \neq v \\ 1e7\ \frac{W}{m^2}, & a1 < x < a2 \\ & b1 < y < b2 \\ & tl < \tau < tu \\ & z' = v \end{array} \right\} \tag{15}$$

The Green's function solution approach is selected particularly because the power supplied to access the transistors can be modeled as a pulse which is active only for a finite amount of time and is applied to only specified location having prescribed length and width. This pulse is analogous to power supplied to the bank and treated as surface heat flux.

2.2 Analytical Solution

The general Green's function solution is given by [27]:

$$\begin{aligned} T(x,t) = {} & \int_{x'=0}^{L} G(x,t|x'0)F(x')\,dx' \qquad (initial\ conditions) \\ & + \alpha \sum_{i=1}^{s} \left[\frac{(\rho cb)_i}{k_i}\ G(x,t|x',0)F(x') \right]_{x'=x_i} \qquad (for\ boundary\ conditions\ for\ thin\ film \\ & \qquad\qquad with\ or\ without\ convection\ only) \\ & + \int_{\tau=0}^{t} \int_{x'=0}^{L} \frac{\alpha}{k}\ G(x,t \mid x',\tau)\ g(x',\tau)\ dx'd\tau \quad \begin{pmatrix} for\ boundary\ condition\ with \\ volumetric\ energy\ generation \end{pmatrix} \\ & + \alpha \int_{\tau=0}^{t} d\tau \sum_{i=1}^{2} [\frac{f_i(\tau)}{k_i}\ G(x,t|x_i\,,\tau)] \quad (for\ boundary\ condtions\ of\ prescribed\ heat\ flux, \\ & \qquad\qquad convection\ condition, and\ thin\ flim\ with\ and\ without\ convection) \\ & - \alpha \int_{\tau=0}^{t} d\tau \sum_{i=1}^{2} \left[f_i(\tau) \frac{\partial G}{\partial n'_i}\bigg|_{x'=x_i} \right] \quad (for\ boundary\ condition\ of \\ & \qquad\qquad prescribed\ temperature\ only) \end{aligned} \tag{16}$$

The Green's function temperature solution for the semi-infinite case is given by [27],

$$T(x,y,z,t) = \sum_{i=1}^{P}\left[\frac{\alpha}{k}\int_{\tau=0}^{t}\int_{x'=0}^{a}\int_{y'=0}^{b} \begin{matrix} q_i(x,y)G(x,y,z,t|x',y',0,\tau) \\ dx'dy'd\tau \end{matrix}\right] \tag{17}$$

where,

$$G(x,y,z,t|x',y',0,\tau) = G_{X22}(x,t|x',\tau) * G_{Y22}(y,t|y',\tau) * G_{Z20}(z,t|0,\tau) \tag{18}$$

Large cotime solutions is selected since the Fourier number of the current model is larger than the 0.25. Cotime is a dimensionless representation of the amount of time it takes heat to penetrate to the boundary, as represented by the Fourier number.

$$G_{X22}(x,t|x',\tau) = \frac{1}{a}\left[1 + 2\sum_{m=1}^{\infty} e^{\frac{-m^2\pi^2\alpha(t-\tau)}{a^2}} \cos\left(\frac{m\pi x}{a}\right)\cos\left(\frac{m\pi x'}{a}\right)\right] \tag{19}$$

$$G_{Y22}(y,t|y',\tau) = \frac{1}{b}\left[1 + 2\sum_{n=1}^{\infty} e^{\frac{-n^2\pi^2\alpha(t-\tau)}{b^2}} \cos\left(\frac{n\pi y}{b}\right)\cos\left(\frac{n\pi y'}{b}\right)\right] \tag{20}$$

$$G_{Z20}(z,t|0,\tau) = \frac{1}{\sqrt{\alpha\pi(t-\tau)}} e^{\frac{-z^2}{4\alpha(t-\tau)}} \tag{21}$$

The same methodology is used for the infinite layer case also since they are both derived using the same template as shown in equation 16. The third term in the equation 16 is used to describe the infinite thermal layer, since it accounts for the volumetric heat generation that is happening due to the access of one transistor.

$$T(x,y,z,t) = \frac{\alpha}{k}\sum_{n=1}^{N}\int_{\tau=0}^{t}\int_{x'=0}^{a}\int_{y'=0}^{b}\int_{z'=-\infty}^{\infty} \begin{matrix} g_n(x,y,z,t) * G_{Z11}(z,t|z',\tau) * \\ G_{X22}(x,t|x',\tau) * G_{Y22}(y,t|y',\tau) \\ \delta(z-v) * dz'dy'dx'd\tau \end{matrix} * \tag{22}$$

Where the $\delta(z-v)$ is the used to convert the equation from a volumetric source to planar/surface source as found on memories and microprocessors.

This is used using the Dirac Delta Sifting Property [27]:

$$\int f(z')\,\delta(z-z')dz' = f(z) \tag{23}$$

The sifting property is used for unit impulses, where everywhere the pulse is zero and at the region there is unit impulse which lasts for a finite time domain. Using this property, the volumetric internal heat generation term is converted to planar heat generation since the units of the delta function is [m]$^{-1}$.

Further expanding the terms, the solution for a single infinite layer can be written as:

$$T(x,y,z,t) = \frac{\alpha}{k}\int_{\tau=0}^{t}\int_{x'=0}^{a}\int_{y'=0}^{b} g(x,y,z,t)\; G(x,y,z,t|x',y',v,\tau)\; dx'dy'\,d\tau \tag{24}$$

Where

$$G(x,y,z,t|\,x',\,y',\,0,\,\tau) = G_{X22}(x,t|x',\tau) * \; G_{Y22}(y,t|y',\tau) \; * G_{Z00}(z,t|v,\tau) \tag{25}$$

As seen from the analytical solution for semi-infinite layer derivation, the only difference is the selection of the G$_Z$ term. In the case for the semi-infinite layer G$_{Z20}$ is used since there is still a boundary layer to be considered. Whereas, for the infinite layer, there are no boundaries in the infinite case, hence G$_{Z00}$ is used.

Expanding the unique terms from equation 25, one gets:

$$G_{X22}(x,t|x',\tau) = \frac{1}{a}\left[1+2\sum_{m=1}^{\infty} e^{\frac{-m^2\pi^2\alpha(t-\tau)}{a^2}}\cos\left(\frac{m\pi x}{a}\right)\cos\left(\frac{m\pi x'}{a}\right)\right]$$

$$G_{Y22}(y,t|y',\tau) = \frac{1}{b}\left[1 + 2\sum_{n=1}^{\infty} e^{\frac{-n^2\pi^2\alpha(t-\tau)}{b^2}} \cos\left(\frac{n\pi y}{b}\right)\cos\left(\frac{n\pi y'}{b}\right)\right]$$

$$G_{Z00}(z,t|z',\tau) = \frac{1}{2\sqrt{\pi\alpha(t-\tau)}}\exp\left(-\frac{(z-v)^2}{4\alpha(t-\tau)}\right) \tag{26}$$

Where, the term v signifies the location of subsequent plane.

2.3 Model Validation

The analytical model presented here is validated by comparison with finite-element simulation results for multiple cases.

2.3.1 Spatially Varying Heat Flux

A single patch is used with a surface heat flux as shown in the Figure 2.1. The geometry in ANSYS-CFX is performed by selecting a length scale in z-direction sufficiently long enough such that thermal wave from the surface heat flux does not reach the wall in z-direction, for the semi-infinite model assumptions. The sides in x and y coordinate plane is selected to be adiabatic as in the boundary conditions of analytical model. Material properties of Silicon is used (k = 148 W/m-K, α= 0.000089 m^2/s). Figure 2.2 illustrates good agreement between theoretical model and finite element simulation.

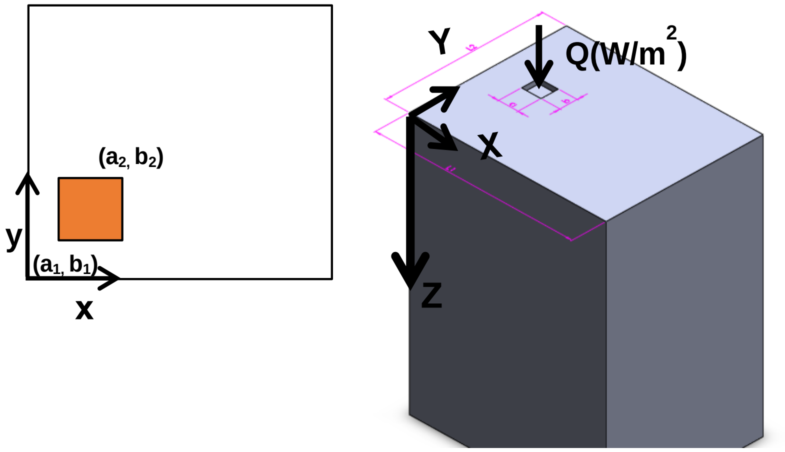


Figure 2.1 Validation model of single patch with heat flux as a function of space having size of 1 mm x 1 mm and the chip size of 10mm x 10mm

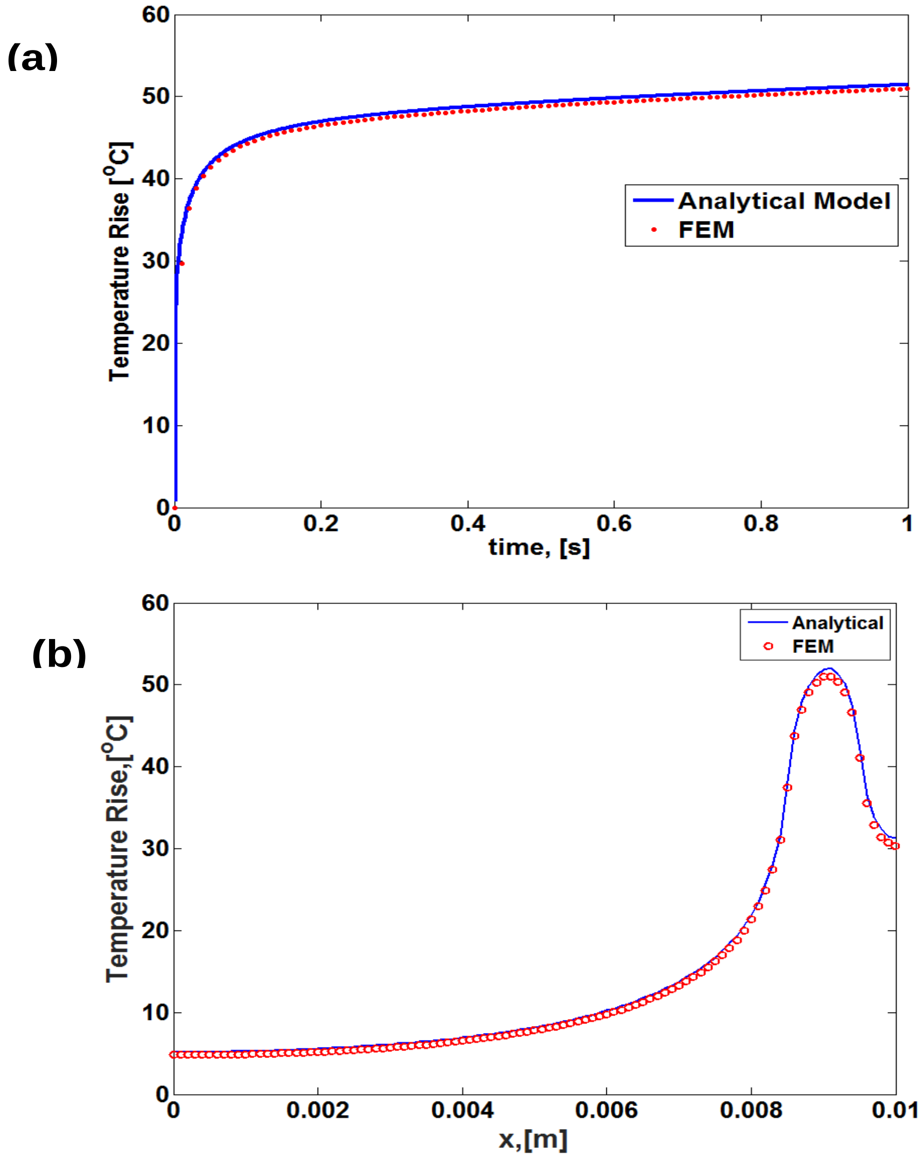


Figure 2.2. Comparison of analytical model (Eq. 7) v/s finite element: **(a)** Temperature rise vs time **(b)** Temperature rise vs x at y = 0.009 m with 10 W constant power applied at the bit-cell.

2.3.2) Heat Flux varying as a function of space and time

Second case is modeled to show the behavior of a single patch on a single bank which active only for a finite duration of time. A similar formulation is used as in the above case; however, heat flux is now modeled as function of space and time along with spatial co-ordinate. The solution can now be expressed as:

$$T(x,y,z,t) = \sum_{i=0}^{P}\left[\frac{\alpha}{k}\int_{\tau=0}^{tf_i}\int_{x'=0}^{a}\int_{y'=0}^{b} \begin{matrix} q_i(x,y,t)\ G(x,y,z,t|x',y',0,\tau) \\ dx'dy'd\tau \end{matrix}\right] \quad (27)$$

$$\text{where, } q_i(x,y,t) = \begin{Bmatrix} 0, & 0 < \ t < tl_i \\ q, & tl_i < t < tu_i \\ 0, & tu_i < t < tf_i \end{Bmatrix} \quad (28)$$

The temperature distribution is solved using the error function identity integrals shown in [28]. The finite element model is validated against the theoretical model and results indicate good agreement. Figure 2.4 illustrates the case, when a pulse is active between 0.1-0.3 seconds. As seen at the onset of the pulse, the temperature rise in the domain is around 0.1 second and as the heating is cut off at 0.3 seconds the temperature drops, as expected. Results from the finite element model show a slight rise and this can be due to numerical errors that are enforced in such simulations.

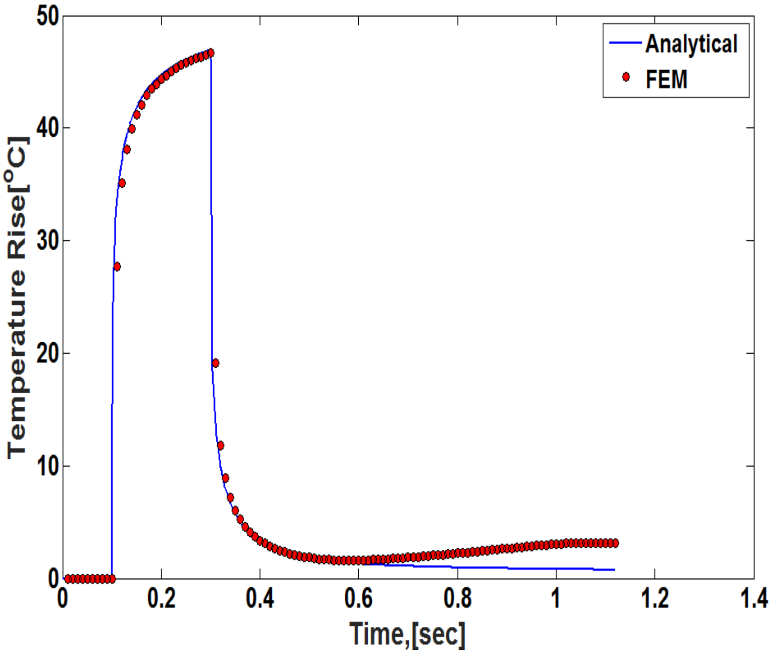


Figure 2.3. Comparison of analytical model (Eq. 12) v/s finite element, temperature v/s time for heat flux varying with time at x = y = 0.001 m

2.3.3) Infinite Layer Model Validation

This third the infinite layer is validated with the ANSYS CFX model. In this case, there is single patch that is placed in the infinite layer as shown in Figure 2.4.

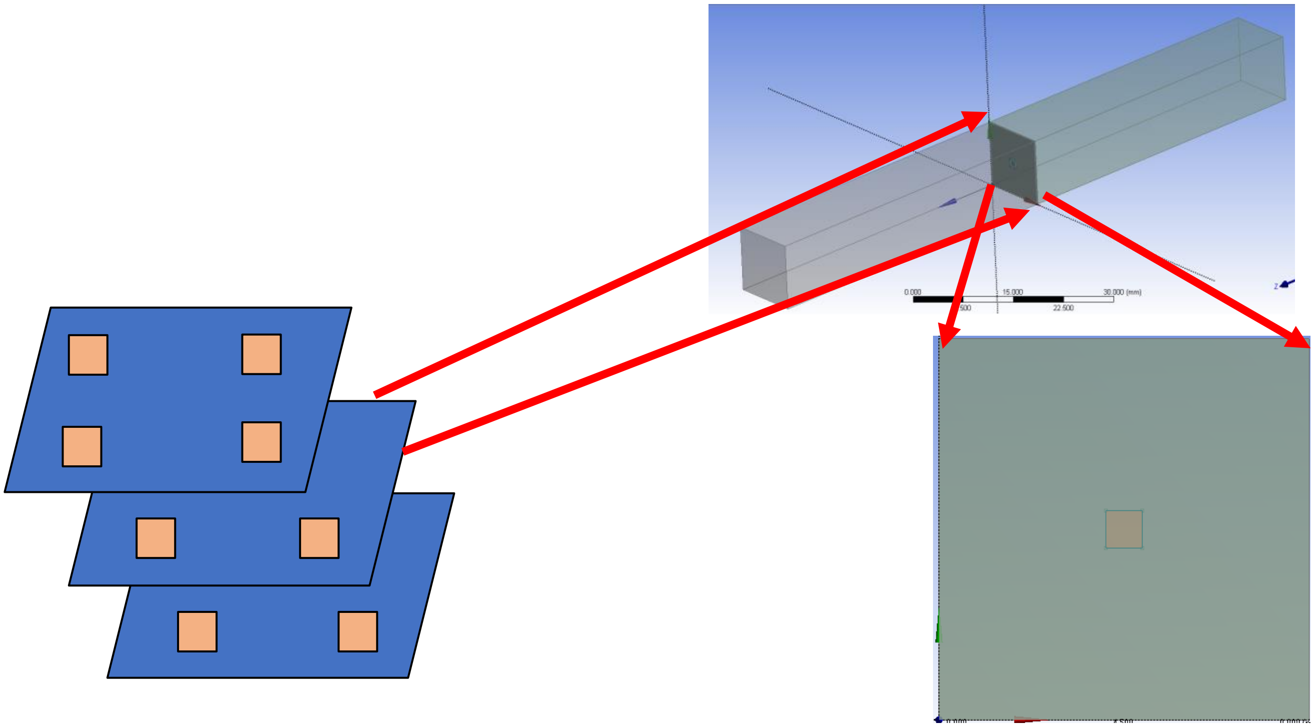


Figure 2.4: Infinite Layer model

This model is validated against FEM model which is shown in Figure 2.5. The single patch is located between 0.0045 ≤ x ≤ 0.0055 and 0.0045 ≤ y ≤ 0.0055 on 0.01 by 0.01 die. The total power supplied to the single patch is 5 W for the time duration 0.01<t<0.05 sec.

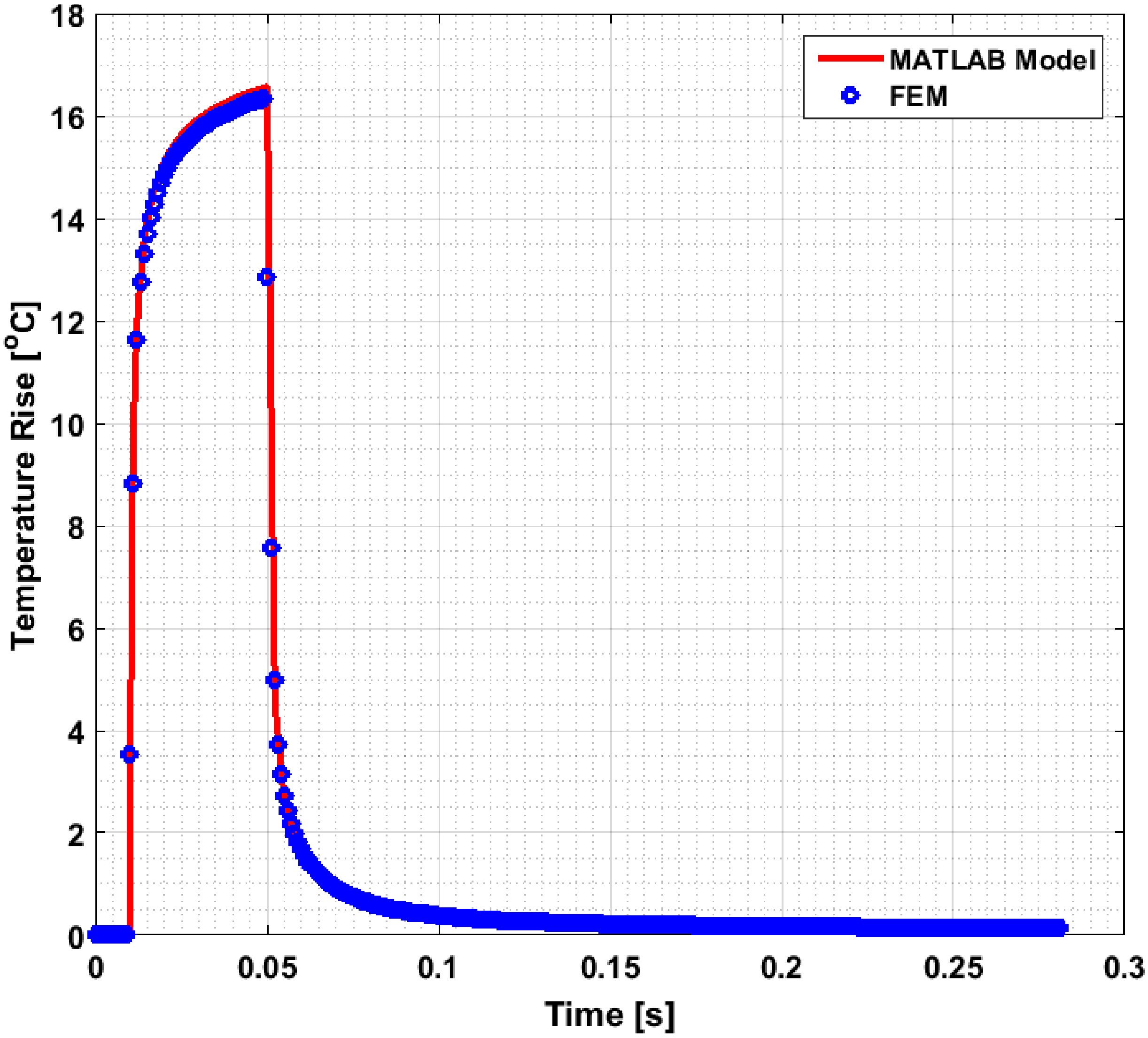
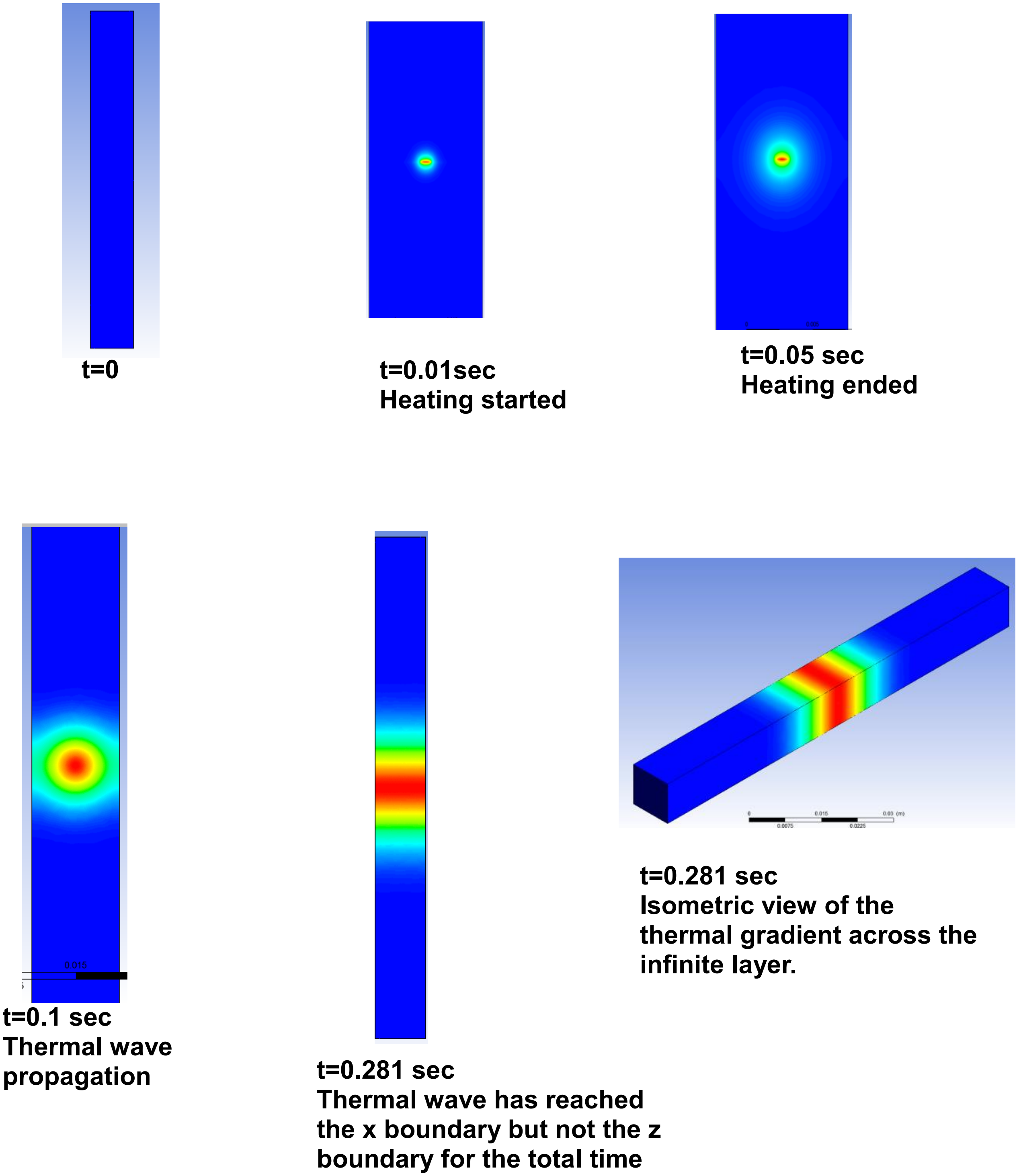


Figure 2.5: Infinite Layer Model with a single patch located 0.0045 ≤ x ≤ 0.0055 and 0.0045 ≤ y ≤ 0.0055 m on 0.01 by 0.01 m. Supplied 5 W power for 0.01<t<0.05 sec. Validation against FEM model is shown.

Time evolution of the thermal gradient across the figure 2.4 is also shown to help visualize the effect of the thermally infinite layer concept in more detail. Figure 2.6 is contour map of the results that were generated in Figure 2.5.



t=0

t=0.01sec
Heating started

t=0.05 sec
Heating ended

t=0.1 sec
Thermal wave propagation

t=0.281 sec
Thermal wave has reached the x boundary but not the z boundary for the total time

t=0.281 sec
Isometric view of the thermal gradient across the infinite layer.

Figure 2.6: Evolution of the thermal gradient across the thermally infinite layer model.

CHAPTER 3

RESULTS AND DISCUSSION

Results for the derived cases in chapter 2 are presented here and discussed. The validation and derivation is used to justify the capability of the model and parametric analysis is described in this section. The results section is presented in three different parts:

3.1) Semi-Infinite Layer Model

3.2) Infinite Layer Model

3.3) Thermal Simulator for 4 layer stacked memory structure

3.1) Semi-Infinite Layer

Subsequent analysis is done by parametrically varying multiple specifications to select the optimum most locations of arranging the patches such that the overall temperature in the bank remained below a critical temperature. Figure 3.1 is obtained by varying the distance of patch1 from patch2, while keeping the second patch at a constant location and the total power supplied to both the patches is kept constant at 10W. This is done to see how the effect that one patch has on another patch, so that the total temperature rise will not exceed a specified temperature rise. 70°C was assumed to be the critical temperature rise [5]. This is a key challenge faced as accurately predicting the optimum length between two bit cells ensures cells operating below the critical temperature value. It is commonly referred to as memory traffic reshaping, which deals with changing the address mapping in DRAM or transferring the data within DRAM. Current work looks at the same problem from a thermal stand point. This optimization process is done to reduce the latency within the memory transfer [29, 30, 31], which thermally can be analogous to physically accessing data between two patches which are much closer to each other. It is seen from Figure 3.1 that as patch1 is shifted closer to patch2, the overall temperature rise reduces. This is due to the symmetrical nature of the architecture. The optimum location for placing the patch1 is x=0.005m.

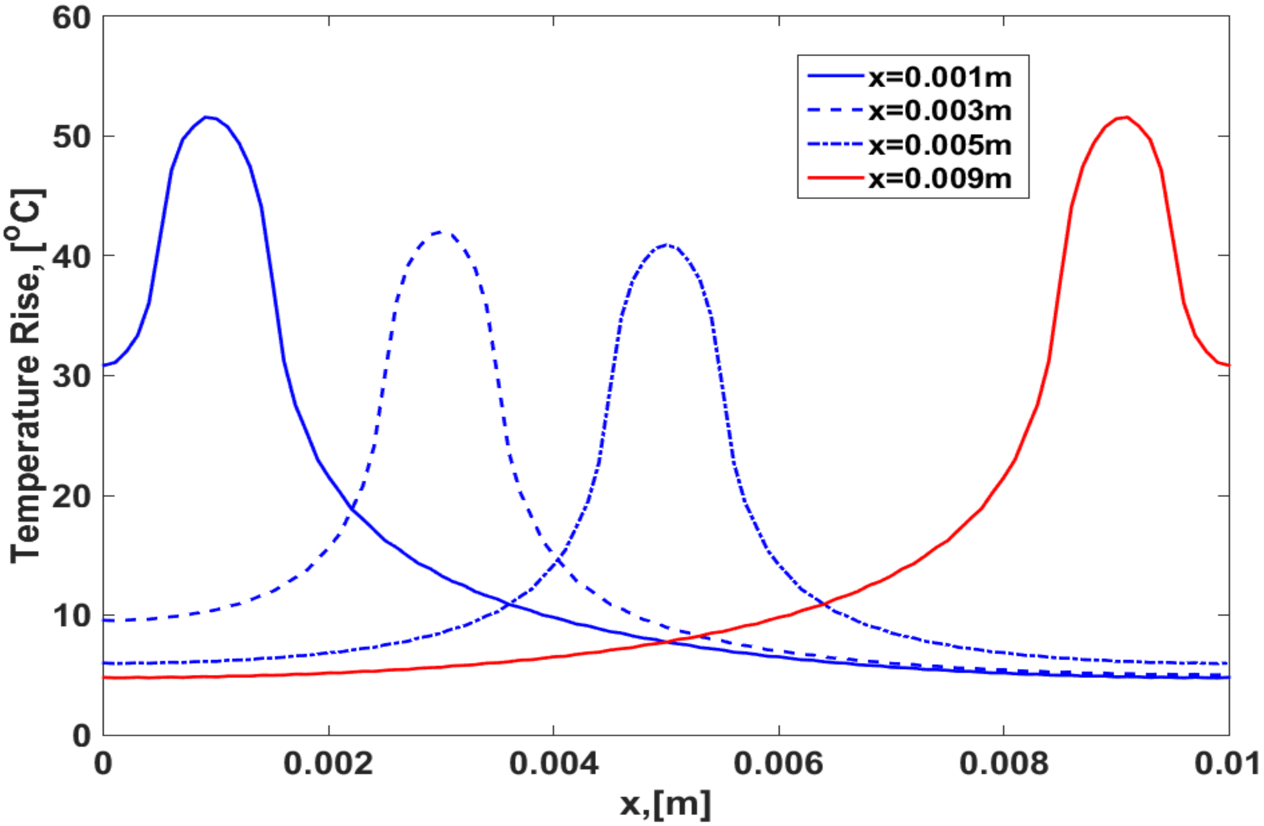


Figure 3.1. The total temperature effect of one patch with varying distance from the other patch. Temperature rise v/s x at different center locations of patch1 (y= 0.001, 0.003, 0.005 m) with fixed center location of patch2 (y= 0.009 m).

Figure 3.2 varies the amount of power supplied to each patch and the effect it has on the total temperature of the bank. For this case, patch1 is kept at 10W while patch2 is varied at 5, 10, and 12W. This test gives an indication of how the temperature rises as different power is supplied for different such read/write operations. This test has practical implications in memory management from an electrical perspective. In past work, a technique for saving memory energy using virtual memory management has been proposed [30]. This technique works by rearranging memory access to reduce the memory trail of each application shifting unused bit-cells to lower power modes. The thermal sequencing performed in this work correlates well with this work as it helps to predict the effect that switching to low power modes will have on the entire memory bank.

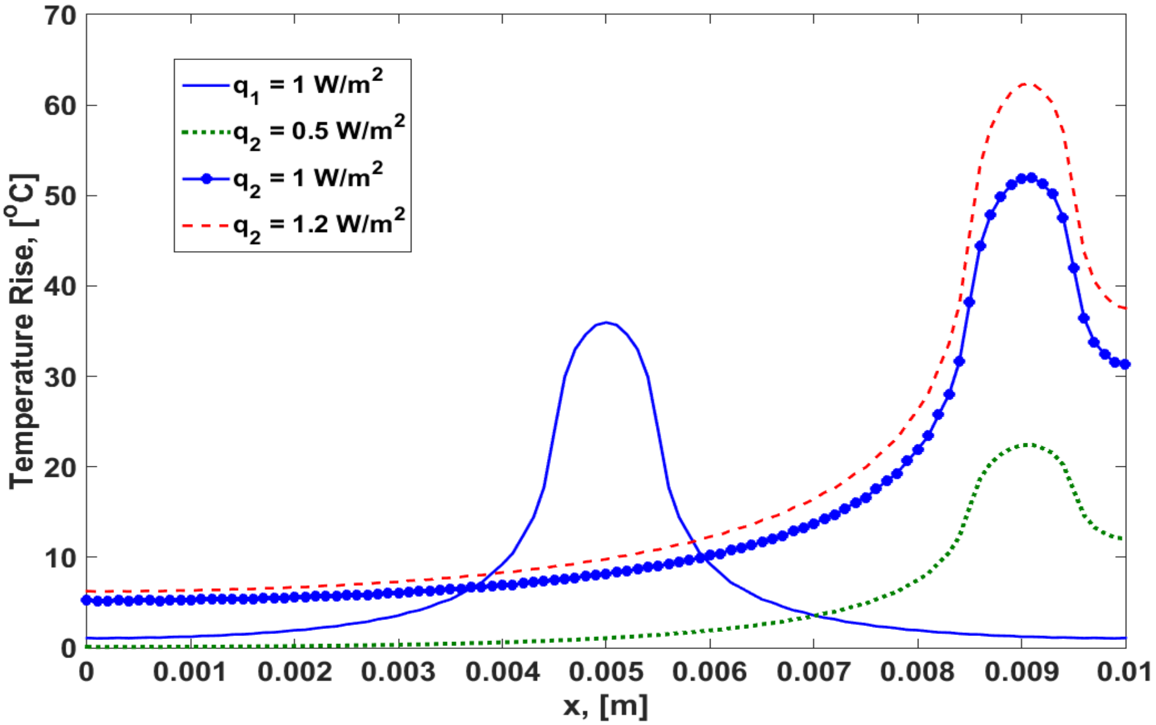


Figure 3.2. Temperature rise v/s x plot for three values for q_2 = [5, 10, 12 W] for a fixed q_1 [10W] at fixed center locations for patch 1 x = y = 0.005m and patch2 x = y = 0.009m.

Figure 3.3 is performed on heat flux which is varying with only time. The results show the variation of total operation time of a single patch at the center of patch1. It is seen that longer the bank is accessed, larger the temperature for the constant power rises. This is expected since the longer patch is activated, the more the energy is supplied, resulting in larger temperature rise. In addition, based on the selection of the adiabatic boundary condition, we see the system converges after the pulse has finished. This behavior is expected since the body reaches steady state condition at final time. The follow up to this analysis is done by checking the temperature rise when two pulses are active for a finite time duration, as that would give an overall and accurate model of how the temperature evolution and decay occurs within each patch. Figure 3.4 shows that

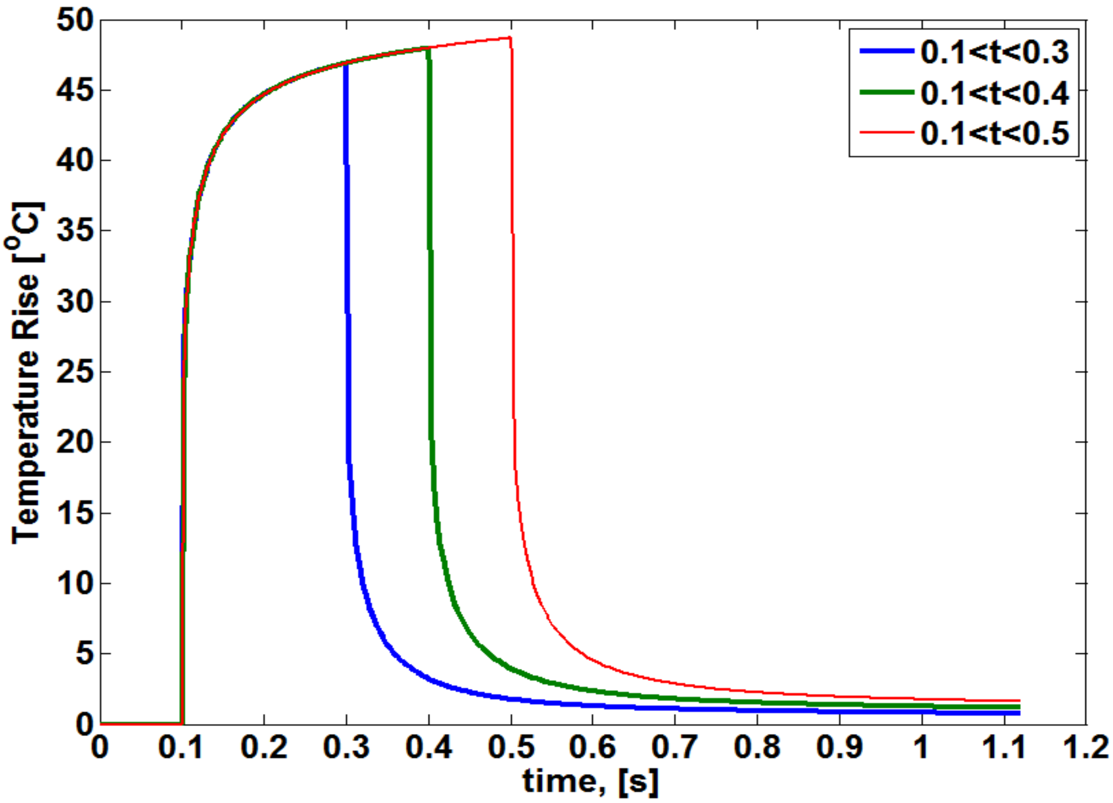


Figure 3.3. Single patch with q (x, y, t) at x = y = 0.005m and z = 0 location, temperature rise v/s time plot for three different pulse widths at the center of the hotspot is shown.

patch1 is active only between 0.1 to 0.3 seconds and then it decays out emitting its residual energy. Figure 3.4 also illustrates temperature rise at two locations, point1 which is center of patch1 (solid red line) and point2 which is the center of patch2. Thermal energy generated due activation of patch1 propagates into medium and its effect is seen at center of patch2 (shown dotted line). At onset of t = 0.4sec patch2 is active, hence we see a temperature jump at t = 0.4-0.6sec and decays out after that duration. The thermal energy generated in medium due to activation of patch1 will reach patch2. This is an accurate modeling of the background power that occurs due to the transistor leakage and its heat dissipated.

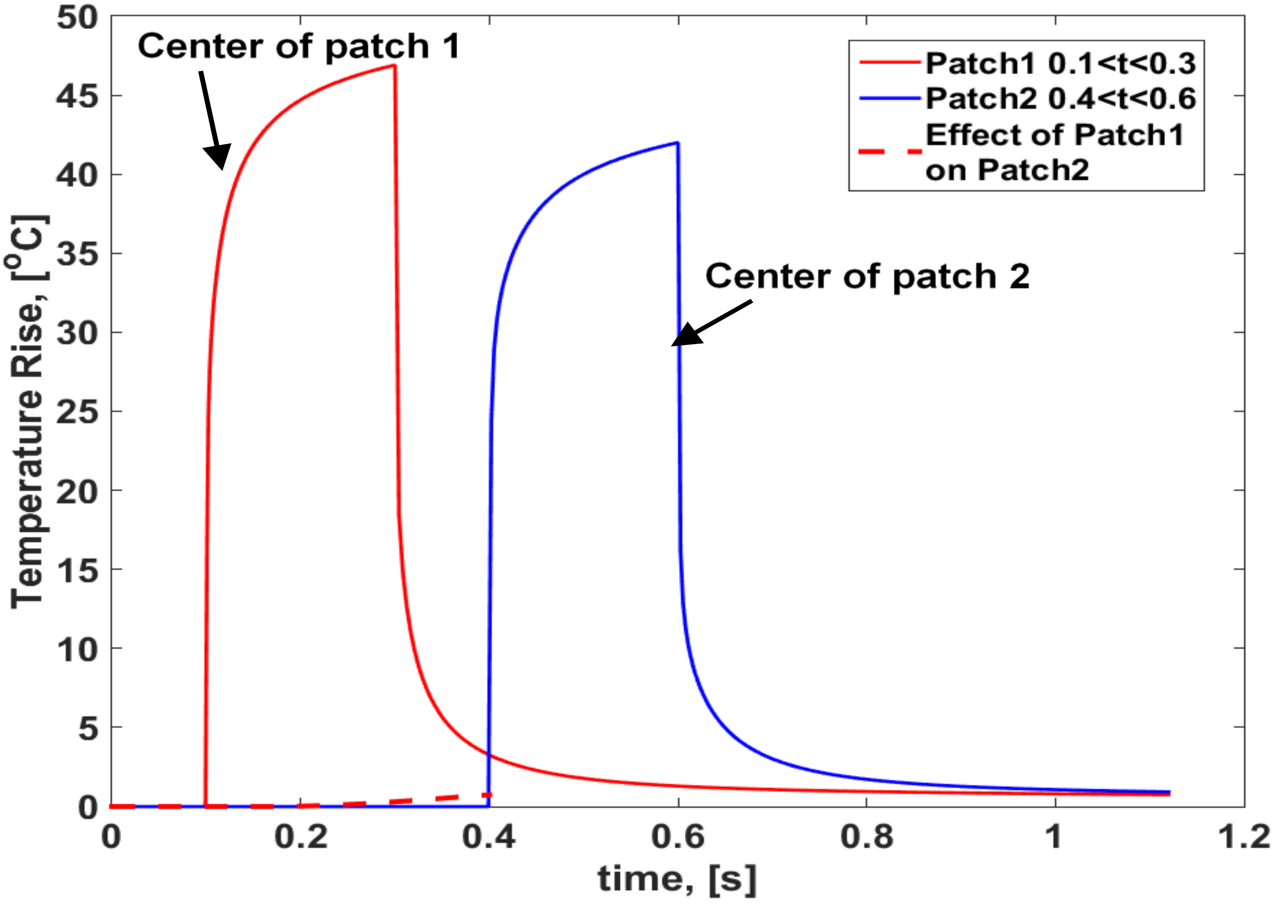
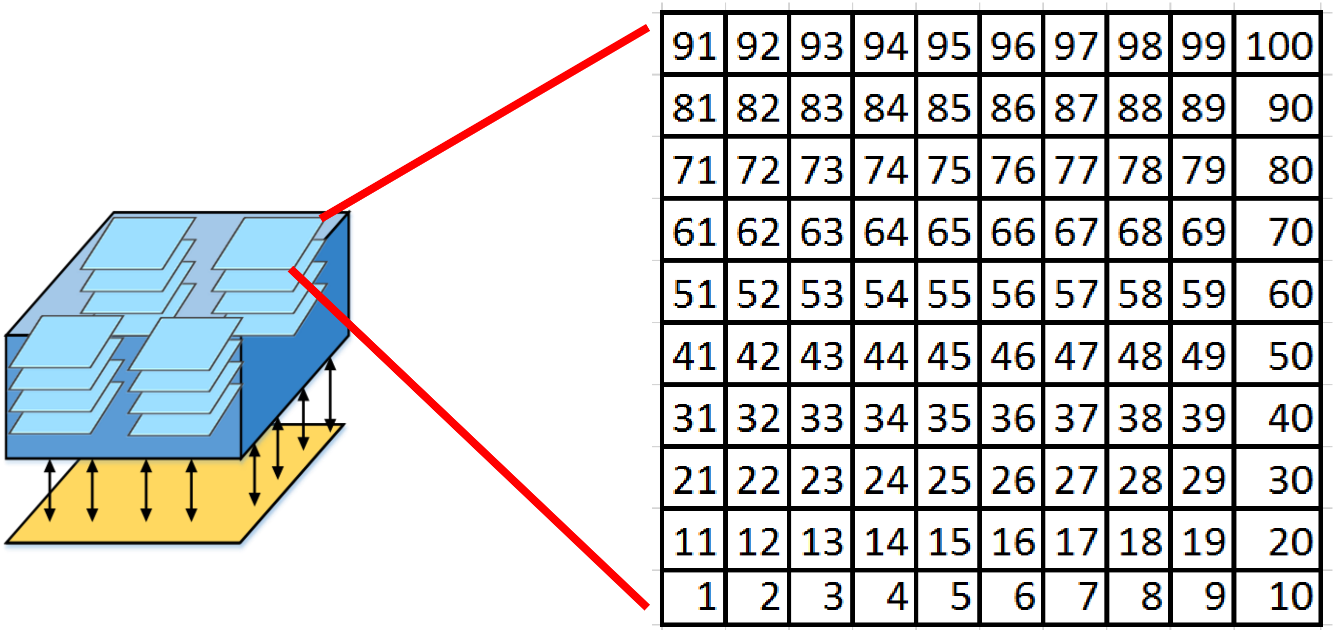


Figure 3.4. Temperature rise v/s time plot of two patches with q_1 (x, y, t) on between $t_1 < t < t_2$ at x = y= 0.005m and q_2 (x, y, t) between $t_3 < t < t_4$ at x = y = 0.009m is shown.

A thermal simulator was created for semi-infinite model where the layer was subdivided into a mesh grid of 10 x 10 as shown in figure 3.5.



91	92	93	94	95	96	97	98	99	100
81	82	83	84	85	86	87	88	89	90
71	72	73	74	75	76	77	78	79	80
61	62	63	64	65	66	67	68	69	70
51	52	53	54	55	56	57	58	59	60
41	42	43	44	45	46	47	48	49	50
31	32	33	34	35	36	37	38	39	40
21	22	23	24	25	26	27	28	29	30
11	12	13	14	15	16	17	18	19	20
1	2	3	4	5	6	7	8	9	10

Figure 3.5: Mesh grid for the top most layer of the 3D memory modeled as thermally semi-infinite body.

The idea of creating the grid like structure is to mimic the actual architecture of a DRAM memory cell as shown in figure 3.6. This is diagram shows how the memory receives it information from the CPU and open the cells using the row and column decoders. The arrangement in a 3D DRAM is shown where there are multiple banks containing individual bit cells, which generate heat.

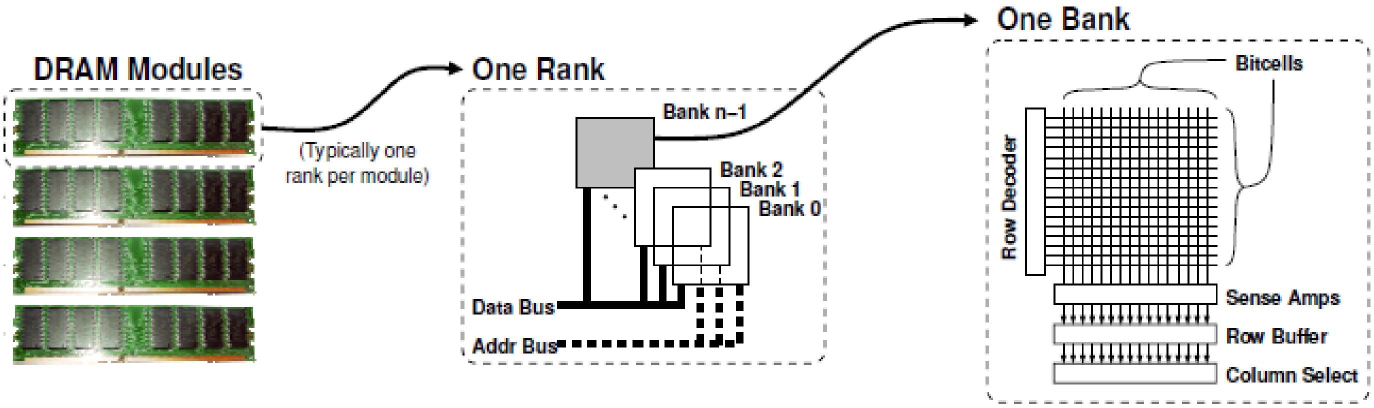
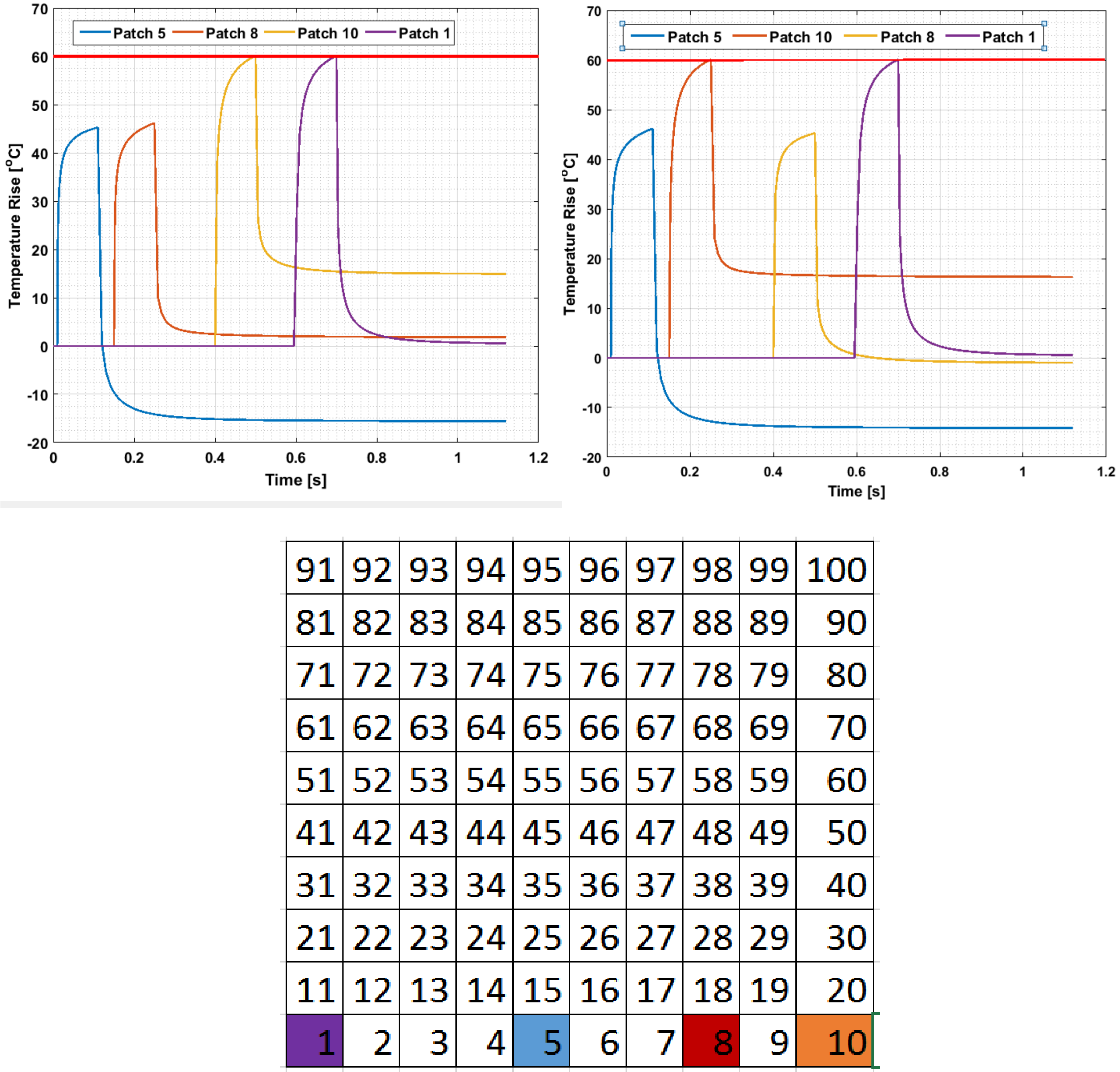


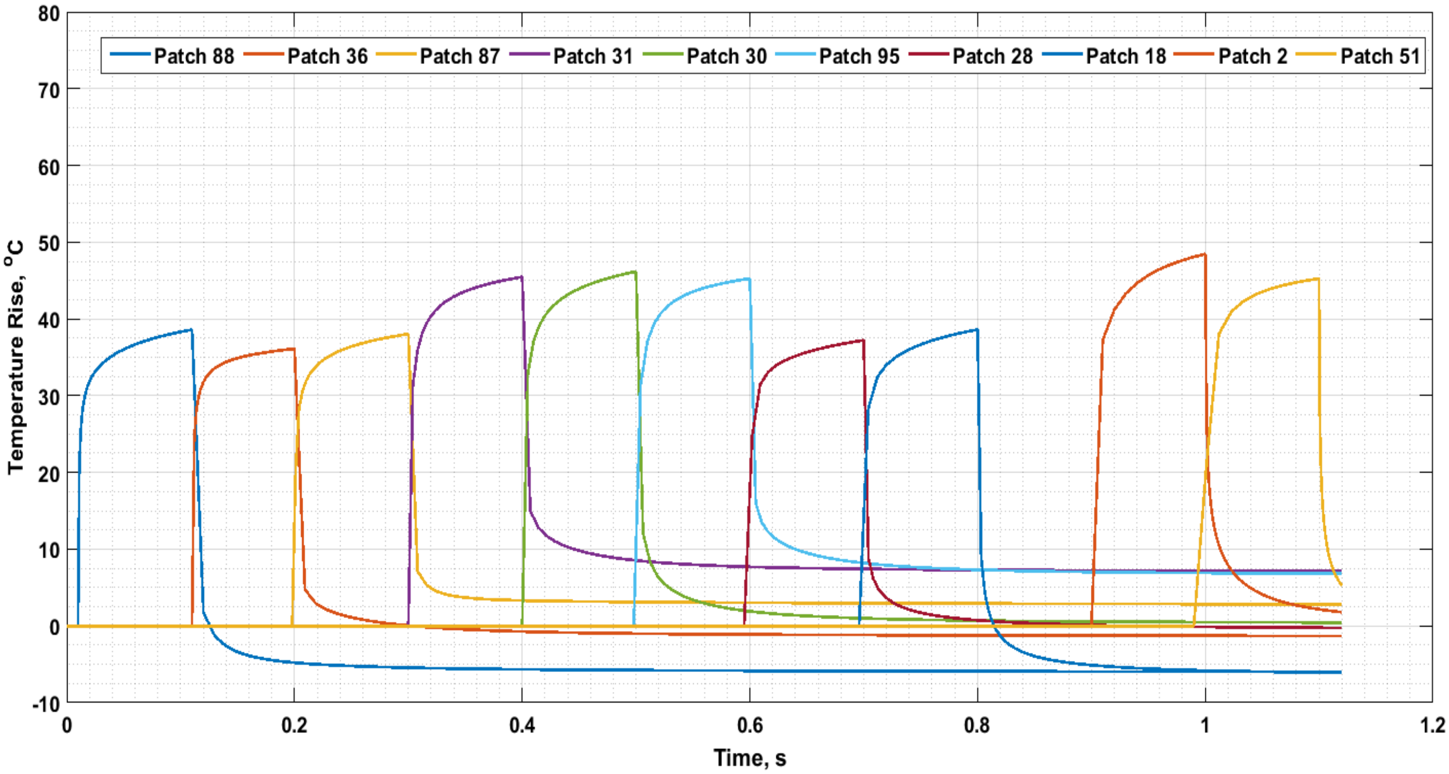
Figure 3.6: Electrical Architecture of access inside a 3D DRAM[21]

Incorporating the derivation from section 2.2 for the semi-infinite body into the mesh grid from figure 3.5, the results generated are used to understand the behavior of 3D memory cell under the influence of a read and write operation. This first case is presented where only the first row of the memory cell is accessed. Only 4 cells in the bank are accessed, which are each supplied 10 W for 0.1 sec. As can be seen from figure 3.7, the cells with the highest temperature rises are patches 1, 10. The reason being that they are located are adiabatic edges of the bank and have the influence of both the x and y direction adiabatic sides. In the figure 3.7, the results shown are for two randomly generated cases, i.e. the patches 1,5,8,10 are accessed randomly in both cases. In both the figures the temperature rise is irrespective of the which cell is accessed first, as the temperature of patches 1 and 10 are always the highest. Since the power supplied to each patch is 10 W, the peak temperature rise from these patches is 60 C. From figure 3.7 it can be seen that there is some latency between the access of the cells, which is generally not the realistic case, since the row and column decoders signal the cells to read and write sequentially without any latency as it will slow the process of operation.



91	92	93	94	95	96	97	98	99	100
81	82	83	84	85	86	87	88	89	90
71	72	73	74	75	76	77	78	79	80
61	62	63	64	65	66	67	68	69	70
51	52	53	54	55	56	57	58	59	60
41	42	43	44	45	46	47	48	49	50
31	32	33	34	35	36	37	38	39	40
21	22	23	24	25	26	27	28	29	30
11	12	13	14	15	16	17	18	19	20
1	2	3	4	5	6	7	8	9	10

Figure 3.7 Thermal simulator case 1 for the semi-infinite model.



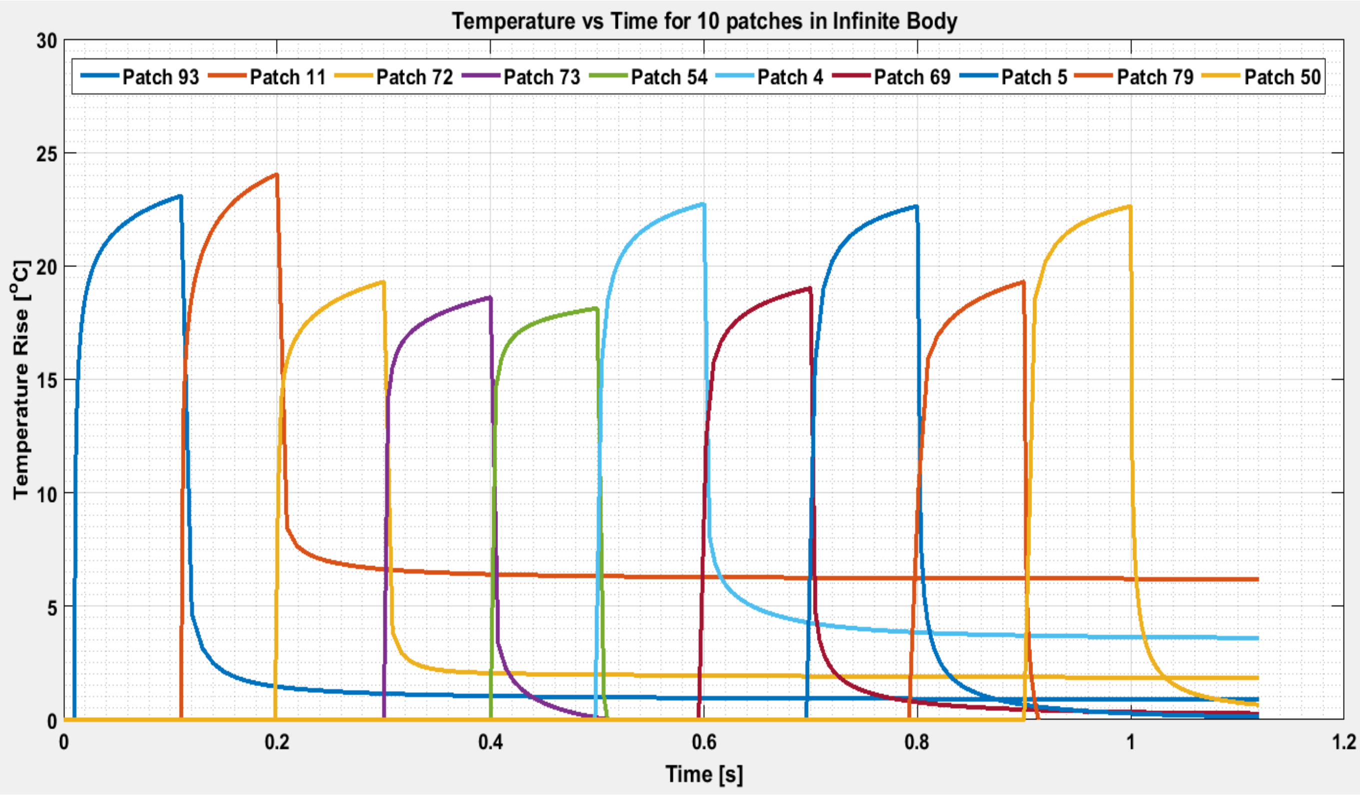
91	92	93	94	95	96	97	98	99	100
81	82	83	84	85	86	87	88	89	90
71	72	73	74	75	76	77	78	79	80
61	62	63	64	65	66	67	68	69	70
51	52	53	54	55	56	57	58	59	60
41	42	43	44	45	46	47	48	49	50
31	32	33	34	35	36	37	38	39	40
21	22	23	24	25	26	27	28	29	30
11	12	13	14	15	16	17	18	19	20
1	2	3	4	5	6	7	8	9	10

Figure 3.8: Thermal simulator case 2 for the semi-infinite topmost layer.

In figure 3.8 second case for the semi-infinite model is shown. This time patches from anywhere in the region are accessed, which is more realistic. In addition, as seen from the temperature vs time graph, majority of the patches are accessed are sequentially, one after the other, which models the realistic scenario that was lacking in the earlier case. A gap between the access was placed in the simulation just to test whether any irregularities in temperature occurs due to a break that might be caused by latency issues in data transfer. However, highest temperature rise occurs right after the latency drop. This is because patch 2 is affected by the effect of the x-direction adiabatic side. The max temperature rise observed by the simulation is 48.6 C in patch 2.

3.2) Infinite Layer

In this section, the analysis was shown based on the single layer thermally infinite model developed from the section 2.2. The model is developed for a single layer one patch on the planar layer. This was further expanded to multiple patches as the work done in the previous section for the semi-infinite concept. The idea was to modulate a thermal layer which has multiple access points or patches as shown in figure 3.6 and can be later coupled in the grand scheme of 3D structure. Due to the confidence gained from the single patch analysis work in the semi-infinite layer work, there was no patch analysis done for the infinite layer. Figure 3.9 shows the behavior of the thermal simulator created for the infinite layer, which consists of 100 cells arranged as shown. The access of these cells is distributed randomly that is generated by the inbuilt command of “randperm” from development software used: MATLAB. The criteria for selecting a random access was to cover a real-world scenario where the access is



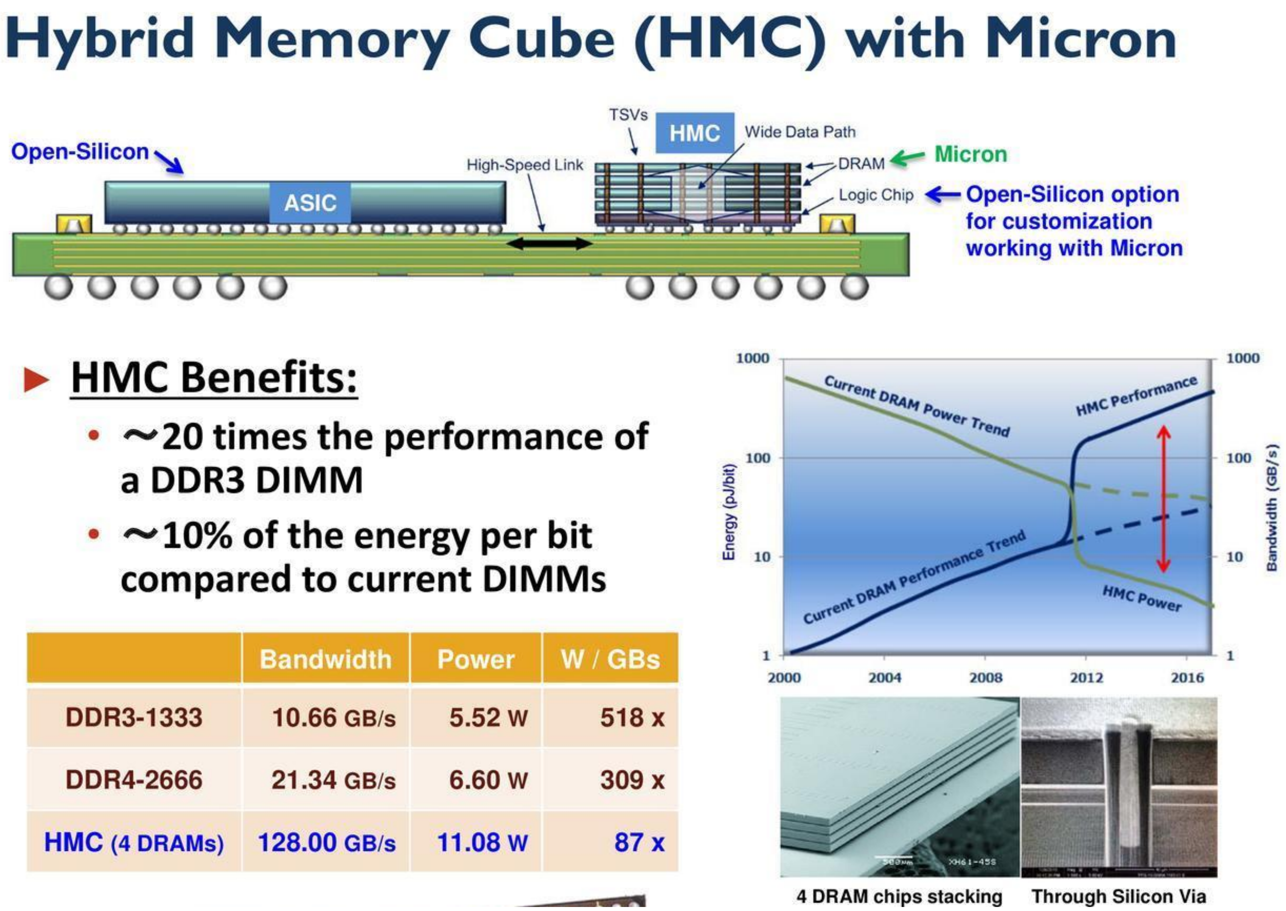
91	92	93	94	95	96	97	98	99	100
81	82	83	84	85	86	87	88	89	90
71	72	73	74	75	76	77	78	79	80
61	62	63	64	65	66	67	68	69	70
51	52	53	54	55	56	57	58	59	60
41	42	43	44	45	46	47	48	49	50
31	32	33	34	35	36	37	38	39	40
21	22	23	24	25	26	27	28	29	30
11	12	13	14	15	16	17	18	19	20
1	2	3	4	5	6	7	8	9	10

Figure 3.9: Thermal simulator case for the infinite middle layer. The grid shows the patches that are accessed and its respective thermal map.

predefined based on the type of operation that is required to be executed. The cells that are selected are accessed sequentially without any latency lag between them. The peak temperatures achieved is 24.23 C, which occurs in patch number 11. This patch experiences the greatest temperature rise, due to its close vicinity to the adiabatic boundary on the x=0 and y=0 edges. In addition, it also seen that the patches which nearest to the adiabatic edges be it x=0, y=0, x=L, or y=L, they all experience the largest temperature rise amongst them. It is primarily since heat does not have any other place to go to and hence is experienced these said patches.

3.3) 3D simulator

This section primarily deals how the results of the past two sections were used in developing a 3D thermal simulator. The simulator is based on the hybrid memory cube (HMC) concept that was developed by Micron Inc. [32]. The concept is specifically designed for multi-core processing and requirement for high-performance systems. The HMC consists of 4 layers of DRAM which is housed on top a small, high-speed logic layer. These DRAMs are physically connected through silicon via interconnects which couple these vertically stacked dies together. The logic layer is used to handle the DRAM and classify its control of operations. Figure 3.10 shows the architecture of the hybrid memory cube and its location on the motherboard next to processor. It also shows the performance capability as compared to the conventional DRAM structure.



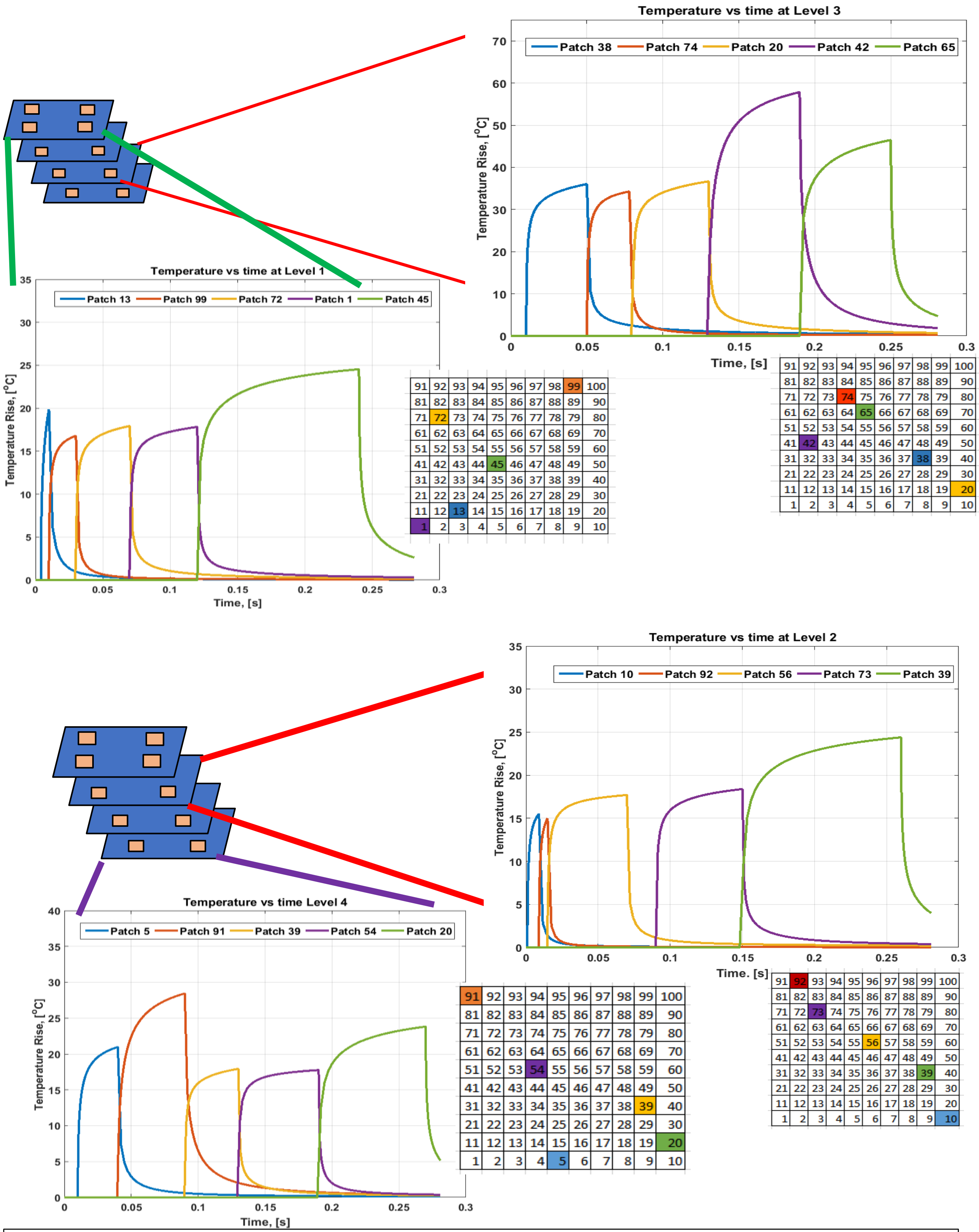
Hybrid Memory Cube (HMC) with Micron

- **HMC Benefits:**
  - ~20 times the performance of a DDR3 DIMM
  - ~10% of the energy per bit compared to current DIMMs

	Bandwidth	Power	W / GBs
DDR3-1333	10.66 GB/s	5.52 W	518 x
DDR4-2666	21.34 GB/s	6.60 W	309 x
HMC (4 DRAMs)	128.00 GB/s	11.08 W	87 x

Figure 3.10: Performance and architecture of the Micron Hybrid Memory cube is shown, which is used a base to model the 3D simulator [32].

Further to elaborate, the 3D thermal simulator is developed using the semi-infinite thermal simulator and the infinite body single thermal simulator. This 3D simulator used the semi-infinite body as the topmost layer and the infinite body as the subsequent 2 to n layers, where n is 4 in this case. The model is developed using MATLAB which takes the help of the inbuilt random sequence generator. To effectively capture the temperature rise, the model accommodates the usage of Temperature in two phases. The temperature captured should be a function of the cell that is being accessed. Infact, majority of the heating will be due to the access of that particular cell. Secondly, the effect of the access of cell 1 onto the neighboring cell. This neighboring cell can be located right next to the accessed cell, on the same row as the accessed cell, or on the same column as the accessed cell, or effectively anywhere else on the same layer. Together the temperature model is effectively captured using the Green's function, which is sole methodology used in solving the 3D Fourier heat equation. In the figure 3.11, the result presented are for a simplistic case, where only 5 cells per layer in a 4-layer structure are accessed. The cells are accessed sequentially without any latency lag. However, the last cell accessed has a longer access time compared to the other cells that are accessed, i.e. it is accessed for 0.04 seconds. The power supplied for each access of 10W of power.



91	92	93	94	95	96	97	98	99	100
81	82	83	84	85	86	87	88	89	90
71	72	73	74	75	76	77	78	79	80
61	62	63	64	65	66	67	68	69	70
51	52	53	54	55	56	57	58	59	60
41	42	43	44	45	46	47	48	49	50
31	32	33	34	35	36	37	38	39	40
21	22	23	24	25	26	27	28	29	30
11	12	13	14	15	16	17	18	19	20
1	2	3	4	5	6	7	8	9	10

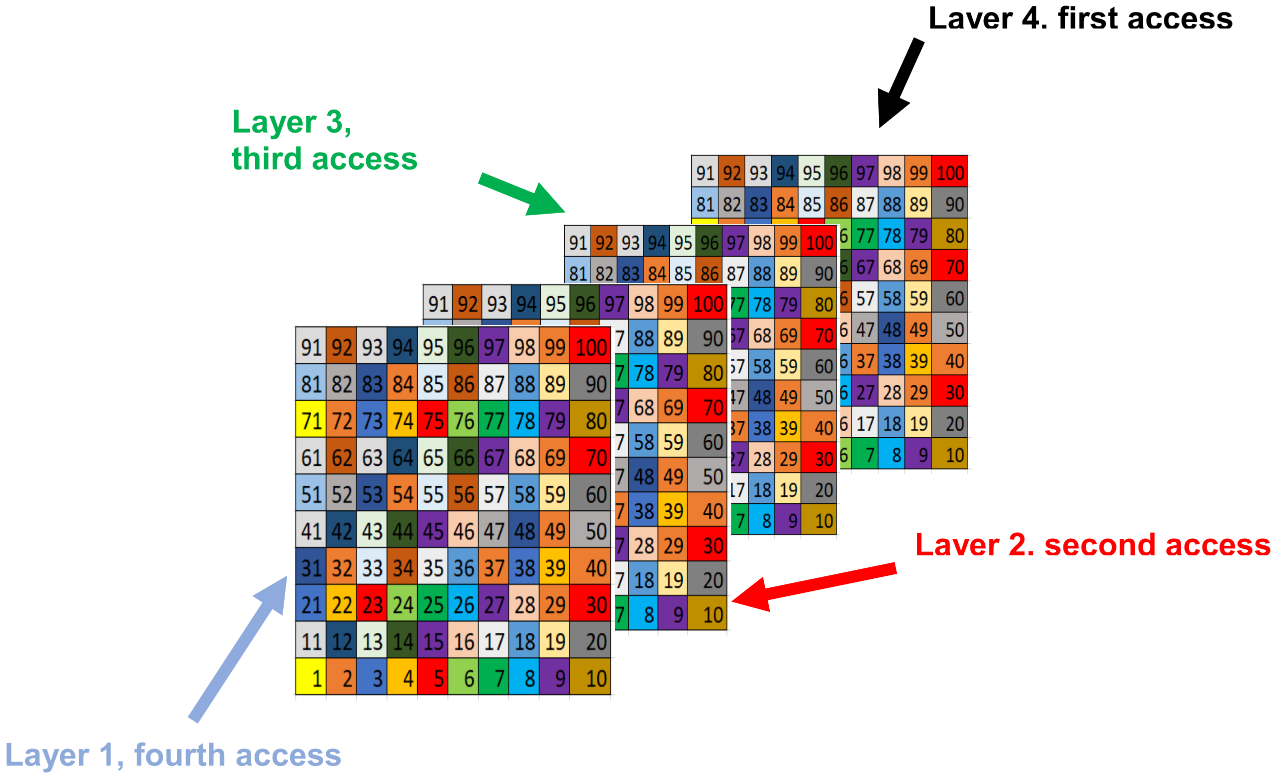
91	92	93	94	95	96	97	98	99	100
81	82	83	84	85	86	87	88	89	90
71	72	73	74	75	76	77	78	79	80
61	62	63	64	65	66	67	68	69	70
51	52	53	54	55	56	57	58	59	60
41	42	43	44	45	46	47	48	49	50
31	32	33	34	35	36	37	38	39	40
21	22	23	24	25	26	27	28	29	30
11	12	13	14	15	16	17	18	19	20
1	2	3	4	5	6	7	8	9	10

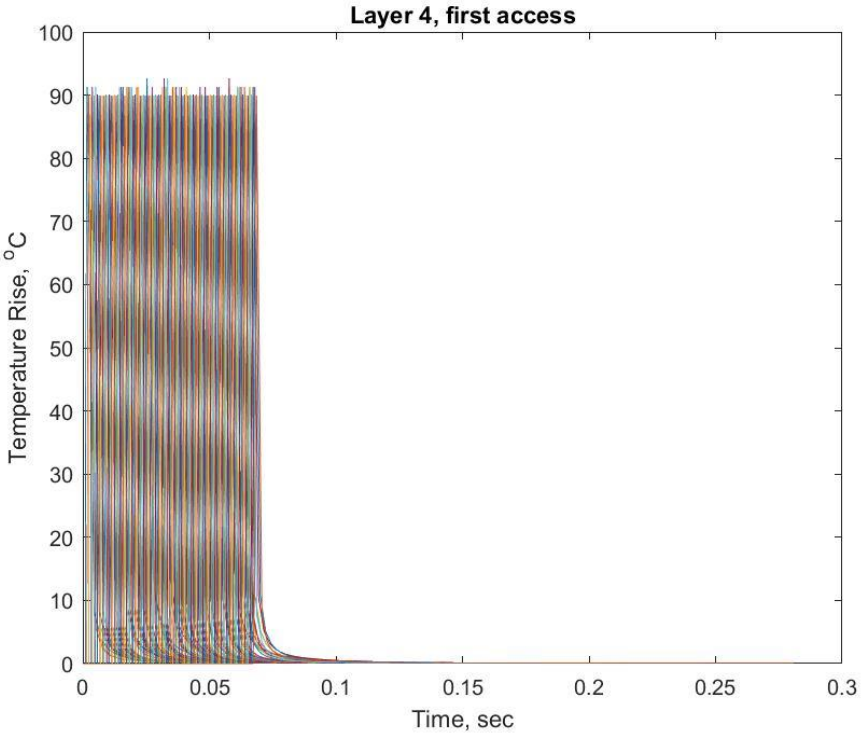
91	92	93	94	95	96	97	98	99	100
81	82	83	84	85	86	87	88	89	90
71	72	73	74	75	76	77	78	79	80
61	62	63	64	65	66	67	68	69	70
51	52	53	54	55	56	57	58	59	60
41	42	43	44	45	46	47	48	49	50
31	32	33	34	35	36	37	38	39	40
21	22	23	24	25	26	27	28	29	30
11	12	13	14	15	16	17	18	19	20
1	2	3	4	5	6	7	8	9	10

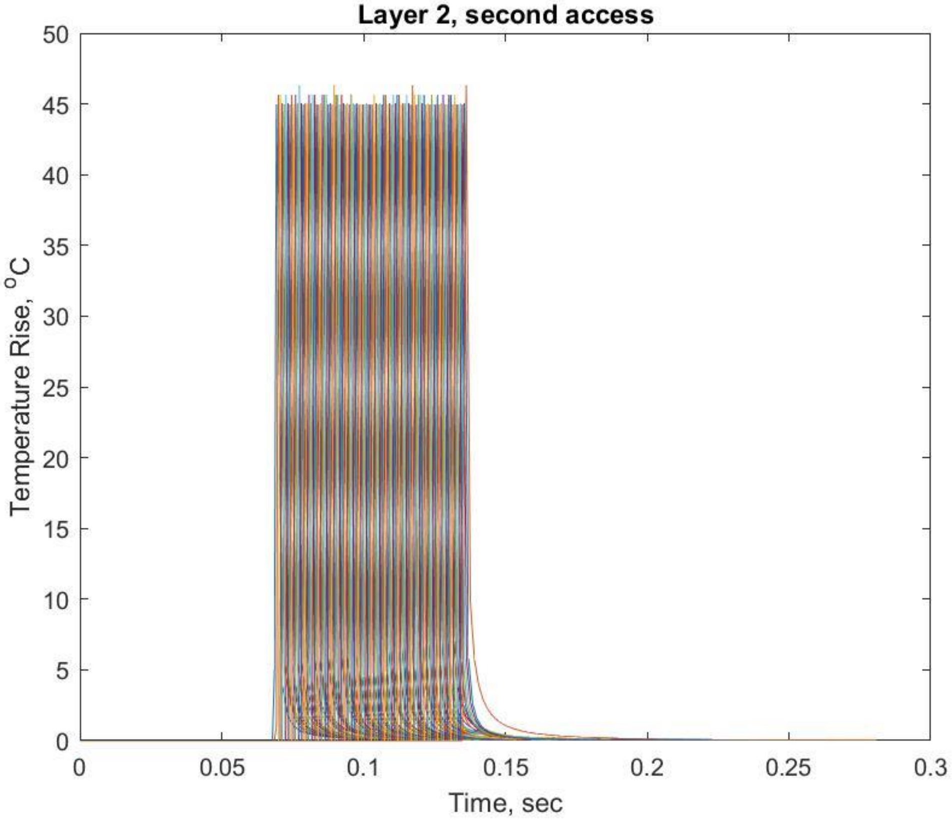
91	92	93	94	95	96	97	98	99	100
81	82	83	84	85	86	87	88	89	90
71	72	73	74	75	76	77	78	79	80
61	62	63	64	65	66	67	68	69	70
51	52	53	54	55	56	57	58	59	60
41	42	43	44	45	46	47	48	49	50
31	32	33	34	35	36	37	38	39	40
21	22	23	24	25	26	27	28	29	30
11	12	13	14	15	16	17	18	19	20
1	2	3	4	5	6	7	8	9	10

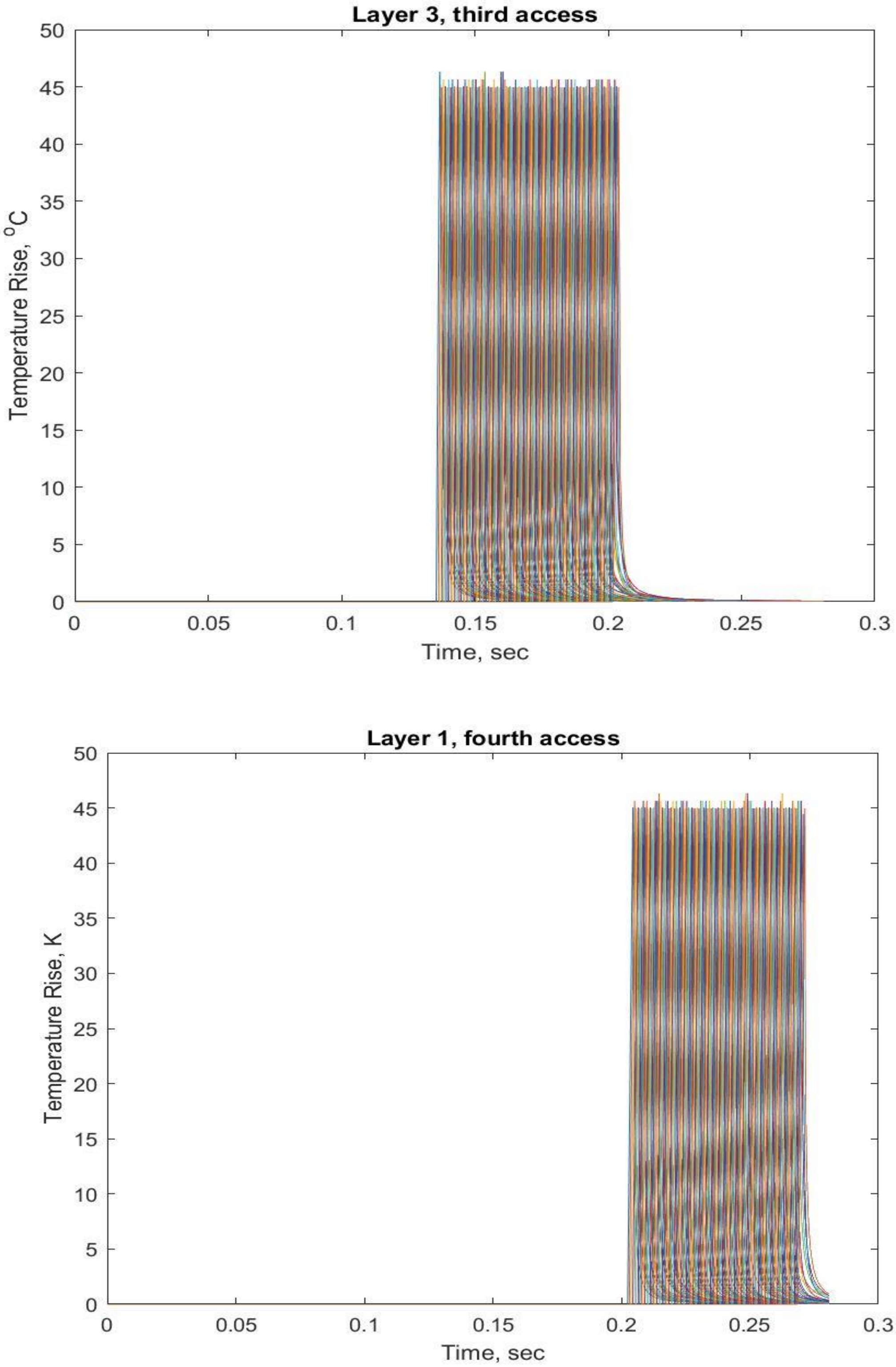
Figure 3.11: Thermal simulator for a 3D structure where the topmost layer is semi-infinite and subsequent layers are infinite.

As seen from the figure 3.11, the peak temperature occurs at Layer 3, which is accessed first. And the subsequent layers are of lower temperatures. To further expand this analysis, all the cells in the layers were accessed, i.e. 400 cells throughout the 4 layers. Figure 3.12 shows this limiting case, where everything is accessed sequentially. However, unlike the previous case, the all the cells in that particular layer are accessed first and then cells from a different layer is accessed. This is presented to show case a limiting case that the model has as in the case any memory access, all the cells might not be accessed. In the event that they are accessed, this model is capable of handling that scenario and outputting the desired results.




Layer 4, first access
Temperature Rise, °C
Time, sec
100
90
80
70
60
50
40
30
20
10
0
0
0.05
0.1
0.15
0.2
0.25
0.3


Layer 2, second access
Temperature Rise, °C
Time, sec
50
45
40
35
30
25
20
15
10
5
0
0
0.05
0.1
0.15
0.2
0.25
0.3



Layer 3, third access

Layer 1, fourth access

Figure 3.12: Thermal simulator for a 3D structure in which all the cell of each layer is accessed. This shows a limiting case of this model.

CHAPTER 4

CONCLUSION AND FUTURE SCOPE

An insight is offered into effectively modeling a single bank in 3D DRAM memory chips using the Green’s Function solution. The results presented are stepping stone towards thermal management of the multi bank and eventually the 3D stacked memory block which would aid in designing a thermally favorable memory algorithm and can be used to capture the thermal cost of a sequence of read operations on a stacked 3D DRAM module. This work is helpful in analyzing the effect that any memory patch access has on its self-including the neighboring patches. It is helpful for a memory architect to understand the thermal cost of such a memory operation as it helpful devise a more efficient access architecture which is the considers the effects of latency and temperature. Even though this work is focused solely on tacking the temperature rise aspect of the sequence, the toolbox is helpful for architects while they are designing the architecture.

Future work can be done in accurate modeling of the through silicon via (TSV) interconnect between the interlayers. The effect of having a through-silicon via might aid the thermal heating process even more and add to effectively capture the temperature rise. In addition, development of an energy efficient algorithm will aid in mitigating the current challenges present in the industry. A physical 3D model can also be tested and validated with this temperature toolbox. This will give more strength to this toolbox and give a better understanding of how the temperature rise can be mitigated from a thermal perspective. This will aid the electrical architect when he tests his algorithm on this model, as it will give him accurate results as compared to other commercially available software, were the computation cost of the running the simulation is also more than what it takes to run in this temperature toolbox. The reason being, this toolbox is using exact integrals as part of its analytical model where the other software might use the finite element techniques which use approximations and refinement to reduce their errors.

APPENDIX A

NOMENCLATURE

a	length of the bank (m)
b	width of the bank (m)
q	heat flux (W/m^2)
P	number of patches
t	time (s)
tf	final time (s)
tu	upper limit of pulse with respect to time (s)
tl	lower limit of pulse with respect to time (s)
τ	dummy index for time
α	thermal diffusivity (m^2/s)
k	thermal conductivity (W/m-K)

REFERENCES

[1] Barroso L. and Holzle U., 2009, “The datacenter as a computer: An introduction to the design of warehouse scale machines,” Morgan & Claypool Publishers, Vermont, USA, Chap. 1-3

[2] Lefurgy C, Rajamani K., Rawson F., Felter W., Kistler M., and Keller T., 2003, “Energy management for commercial servers,” Computer, 36(12), pp. 39–48.

[3] Ware M., Rajamani K., Floyd M., Brock B., Rubio J., Rawson F., and Carter J., 2010, “Architecting for power management: The IBM POWER7 approach,” Proc. 16th High Performance Computer Architecture, Bangalore, pp. 1–11.

[4] Borkar S., 2007, “Thousand core chips: a technology perspective,” Proc. 44 Design Automation Conference, New York, pp. 746–749.

[5] Intel, 2016, http://ark.intel.com/products/53575/

[6] Agrawal A., Khaitan S. K., 2008, “A new heuristic for multiple sequence alignment,” Proc. International Conference on Electro/Information Technology, Ames, pp. 215–217.

[7] Bergman K. et al., 2008, “Exascale computing study: Technology challenges in achieving exascale systems,” Technical Report, DARPA.

[8] Khaitan S., McCalley J., Raju M., 2010, “Numerical methods for on-line power system load flow analysis,” Energy Systems, 1(3), pp. 273–289.

[9] Raju M., Khaitan S., 2009, “Domain decomposition based high performance parallel computing,” International Journal of Computer Science Issues, 5, pp. 27-32

[10] Kandlikar S.G., 2014, “Review and Projections of Integrated Cooling Systems for Three-Dimensional Integrated Circuits,” ASME J. Electron. Package 136(2), pp. 1-11.

[11] Venkatadri V., Sammakia B., Srihari K., and Santos D., 2011, "A Review of Recent Advances in Thermal Management in Three Dimensional Chip Stacks in Electronic Systems," ASME. J. Electron. Package, 133(4), pp. 1-15.

[12] Meng. J, Kawakmai K., and Coskun A., 2012, "Optimizing Energy Efficiency of 3-D Multicore Systems with Stacked DRAM under Power and Thermal Constraints," Proc. 49th Design Automation Conference, San Francisco, pp. 648-655.

[13] Weis C., When N., Igor L., Benini L., 2011, "Design Space Exploration for 3D-stacked DRAMs," Proc. Design, Automation and Test, Grenoble, pp. 1-6

[14]Rahman R., and Reif. R., 2001, "Thermal Analysis of Three-Dimensional Integrated Circuits," Proc. International. Interconnect Technology Conference, Burlingame, pp. 157–159.

[15] Dai F., Yu D., Zhou J., Wu X, Jing X., Song C., and He H.,2013, "Study of Equivalent Thermal Modeling and Simulation of 2.5D/3D Stacked Dies module", Proc. 14th International Conference on Electronic Packaging Technology, Dalian, pp.498-502.

[16] Choobineh, L., and Jain, A., 2015, "An explicit analytical model for rapid computation of temperature field in a three-dimensional integrated circuit (3D IC)", International Journal Thermal Science, 87, pp. 103-109

[17] Choobineh, L., and Jain, A., 2012, "Analytical solution for steady-state and transient temperature field in vertically integrated three-dimensional integrated circuits (3D ICs)", IEEE Transactions Components & Packaging Technologies, 2(12), pp. 2031-2039

[18] Janicki M, Mey G.D, Napieralski A., 2007, “Thermal analysis of layered electronic circuits with Green’s functions,” Microelectronics Journal, 38(2), pp.177–184

[19] Stevens L., Goddard W., and Lynott J., 1964, “Data storage machine” US Patent 3134097.

[20] Mittal S., 2012, “A Survey of Architectural Techniques for DRAM Power Management,” International Journal of High Performance System Architecture, 4(2), pp. 110-119.

[21] Loh G.H., 2008, “3D-Stacked Memory Architectures for Multi-Core Processors,” Proc. 35th International Symposium on Computer Architecture, Beijing, pp. 453-464.

[22] Cooper-Balis E., and Jacob B., 2010, “Fine-grained activation for power reduction in DRAM,” Micro, 30(3), pp. 34–47

[23] Micron, 2016, “Calculating memory system power for DDR3.” http://download.micron.com.

[24] T. Vogelsang, 2010, “Understanding the energy consumption of dynamic random access memories,” Proc. 43rd International Symposium on Microarchitecture, Atlanta, pp. 363–374,

[25]Haji-Shiekh A., 1990, “Peak temperature in High-Power Chips”, IEEE Transactions on Electron Devices, 37(4), pp. 902-907

[26] Ozisik, N., 1980, "Heat Conduction," 2nd ed., John Wiley & Sons, Inc.

[27] Beck J. V., Cole K.D., Haji-Sheikh A., and Litkouhi B. 2011, “Heat Conduction Using Green's Function,-Series in Computational and Physical Processes in Mechanics and Thermal Sciences”, 2nd ed., Taylor and Francis Group, Abingdon, Oxford, UK.

[28] Ng. E.W, and Geller M., 1969, “A table of Integrals of the Error Functions”, Journal of Research of the National Bureau of Standards-B. Mathematical Sciences, 73B (1).

[29] Amin A., and Chishti Z., 2010, “Rank-aware cache replacement and write buffering to improve DRAM energy efficiency,” Proc. 16th International Symposium on Low power electronics and design, Austin, pp. 383–388.

[30] Ayoub R., Indukuri K., and Rosing T., 2010, “Energy efficient proactive thermal management in memory subsystem,” International Symposium on Low-Power Electronics and Design, Ausitn, pp. 195–200.

[31] De La Luz V., Kandemir M., and Kolcu I., 2002, “Automatic data migration for reducing energy consumption in multi-bank memory systems,” Proc. 39th Design Automation Conference, New Orleans, pp. 213–218.

[32] Micron, 2017, https://www.micron.com/products/hybrid-memory-cube/all-about-hmc.

BIOGRAPHICAL INFORMATION

Ratnesh Raj received his Bachelor of Technology degree in Mechanical Engineering from Vellore Institute of Technology, Tamil Nadu, India in 2014. After completion, he was working at Indian Institute of Science, Bangalore, India. He started his Master of Science in Mechanical Engineering at University of Texas at Arlington from August 2015. During his master’s degree, he also did his internship at Continental Carbon Company as a reliability intern from January 207 to June 2017. He received his Master of Science degree in August 2017.