THERMAL MODELING AND ANALYSIS OF 3-DIMENSINAL MEMORY INTEGRATION

Annmol Cherian[1], Ajay Augustine[1], Vinod Pangracious [2] and Jemy Jose[2]

[1]Department of applied electronics & instrumentation, Rajagiri School of engineering & technology, Kerala, India.
annmolcherian26@gmail.com, ajayaugustine@hotmail.com

[2]Department of electronics & communication, Rajagiri School of engineering & technology, Kerala, India.
vinod.pangracious@rajagiritech.ac.in, jemyjosek@gmail.com

Abstract

Moore's law describes a long-term trend in the history of computing hardware. The conventional methods have reached his limits so new fields has to be exploited. Such a concept is 3-Dimensional integration where the components are arranged in 3D plane. This arrangement can increase the package density of devices. The successful construction of 3D memory can lead to a new revolution in designing and manufacturing high performance microprocessor system on chip. The major problem is the increased temperature effects. It's important to develop an accurate power profile extraction methodology to design 3D memory. The total power dissipation includes static and dynamic component. In this paper the static power dissipation of the memory cell is analysed and is used to accurately model the inter-layer thermal effects for 3D memory stack. Then packaging of the chip is considered and modelled using an architecture level simulator. This modelling is intended to analyse the thermal effects of 3D memory, its reliability and lifetime of the chip with greater accuracy.

Keywords

3D integration, 3Dmemory, Thermal effects

1.1. 3-DIMENSIONAL OR VERTICAL INTEGRATION

Successful fabrication of vertically integrated device dates back to early 1980s. The structures include 3D CMOS inverters where PMOS and NMOS transistors share the same gate, considerably reducing total area of the inverter. The term joint metal oxide semiconductor (JMOS) was used for these structures. In the following years the result on 3D integration remained an area of limited scientific interest. Due to increasing importance of interconnects and the demand for greater functionality on a single substrate, vertical integration has recently become a more prominent research topic. Over the last 5 years, the 3D integration has evolved into a design paradigm manifested at several abstraction levels such as package, die and wafer levels. The quintessence of 3D integration is the drastic decrease in length of the longest interconnects across integrated circuits. The considerable decrease in length is a promising solution for increasing the speed while reducing the power dissipated by the IC.

D.C. Wyld, et al. (Eds): CCSEA 2011, CS & IT 02, pp. 242–256, 2011.
© CS & IT-CSCP 2011 DOI: 10.5121/csit.2011.1223

1.2 CHALLENGES FOR 3D INTEGRATION

Developing a design flow for 3D ICs is a complicated task with many ramifications. Number of challenges at each step of the design process has to be satisfied for 3D ICs to successfully evolve into a mainstream technology. An efficient front-end design methodology and mature manufacturing preprocess at the back end are required to effectively provide large scale 3D systems. They are explained below.

- Technological/manufacturing limitations- new packaging solutions are to be found out, also new utilization methods of vertical interconnects for the propagate signals
- Testing should include die-to-die, wafer-to-wafer or die-to-wafer which is more complicated than the normal ICs.
- Interconnect design and new design methodologies are to be found out to combine different layers.
- Thermal issues-the power dissipation is increased due increased transistor density. Due to increased power density the part of IC away from heat sink has much elevated temperature and hence can accelerate ware out. Therefore this is the fundamental issues to be solved.

2.1 THERMAL MODELING & MANAGEMENT

The primary advantage of 3D integration, significantly the greater packing density, is also the greatest threat to this emerging technology- aggressive thermal gradient among the planes within a 3D IC. Thermal problems, however, are not unique to vertical integration. Due to scaling, elevated temperatures and hot spots within traditional 2D circuits can greatly decrease the maximum achievable speed and affect the reliability of a circuit.

In 3-dimensional integration, low operating temperature is a prominent design objective, as thermal analysis of 3D ICs indicates the escalated temperatures can be highly problematic. Two key elements are required to establish a successful thermal management strategy: the thermal model to characterize the thermal behavior of a circuit and design techniques that alleviate thermal gradients among the physical planes of a 3D stack while maintaining the operating temperature within acceptable levels. The primary requirements of a thermal model are high accuracy and low computational time, while thermal design techniques should produce high quality circuits without incurring long computational design time.

2.2 THERMAL ANALYSIS OF 3D IC'S

A 3D system consists of disparate materials with considerably different thermal properties, including semi conductor, metal, dielectric and possibly polymer layers used for plane bolding. To describe the heat transfer process (only by conduction) within the volume of a system and determine the temperature at each point, T at a steady state requires a solution of

$$\nabla(k\nabla T) = -Q \quad (1)$$

Where k is the thermal conductivity and Q is the generated heat. In integrated circuits, heat originates from the transistor that behave as a heat sources and also from self heating of both the devices and the interconnects (joule heating), which can significantly elevate the circuit temperatures

More specific technique applied at various stages of IC design flow, such as synthesis, floor planning, and placement, routing and maintaining the temperature of a circuit with in specified limits of or elevates thermal gradients among the planes of the 3D circuits.

2.2.1 Closed form Temperature Expressions

A 3D system can be modeled as a cube consisting of multiple layers of silicon, aluminum, silicon dioxide and polyimide. The device on each plane are considered as isotropic heat sources

and are each modeled as a thin layer on the top surface of the silicon layer. In addition, due to the short height of 3D stack, one dimensional heat flow is assumed. Such an assumption vastly decreases the simulation time. Certain boundary conditions apply to this thermal model in order to validate the assumption of one dimensional heat flow. Self-heating of MOSFET devices can also cause the temperature of a circuit to significantly rise. Certain devices can behave as hot spots, causing significant local heating. Although the dielectric and glue layers behave as thermal barriers the silicon substrate of the upper planes spreads the heat, reducing the self heating of the MOSFET's.

To estimate the maximum rise in temperature on the upper planes of a 3D circuit while considering the heat removal properties of interconnects, a simple closed form expression based on one dimensional heat flow has been developed. The temperature increase ΔT in a 2D circuit can be described by

$$\Delta T = Rth\, P/A, \qquad (2)$$

Where R_{th} is the thermal resistance between the ambient environment and the actual devices, P is the power consumption, and A is the circuit area.

2.2.2 Compact thermal models

Thermal models based on analytic expressions for the temperature of a 3D circuit are discussed. In all of these models, the heat generated within each physical plane is represented by a single value. Consequently, the power density of a 3D circuit is assumed to be a vector in the vertical direction (i.e.: z direction). The temperature and heat within each plane of 3D system however can vary considerably, yielding temperature and power density vectors depend on all three directions. The nodes connected through resistors forming a 3D resistive network.

2.2.3 Mesh based thermal models

This type of model most accurately represents the thermal profile of a 3D system. The primary advantage of mesh-based thermal models is that these models can be applied to any complex geometry and do not depend on the boundary conditions of the problem. A 3D circuit is decomposed into a 3D structure of finite hexahedral elements. (i.e.: parallelepipeds). The mesh can also be non-uniform for regions with complex geometries or non-uniform power densities.

3.1 Chip Power Extraction & Estimation

The MOS technologies have been improving at a dynamic rate. As the continuation of MOORE's law the MOS devices are scaled down aggressively in each generation to achieve higher integration density and performance. Thus the recent trend in semiconductor manufacturing is to build devices in nanoscale. The major problem with nanoscale devices is the increasing power dissipation. The total static power of CMOS technology is comprised of two major components: The device subthreshold current I_{OFF} and the gate tunneling current I_{GATE}. Therefore an accurate analysis of the total static power must include both the channel leakage due to device subthreshold current and the gate leakage due to tunneling current. A brief introduction is given to each of the effects while scaling down the MOS devices is explained in this introduction.

3.1.1 Sub-threshold leakage

From its earlier days MOS devices were well known for the switching applications. By applying suitable gate current the current flow from source to drain is controlled. Moreover by the application of gate voltage less than the threshold voltage makes the device to turn off and the drain current was supposed to be zero.

The problem arises when the whole semiconductor industry shrinking to nanometerscale. When the devices were fabricated in nanometer scale the electric field increased at constant rate because the voltage drops over the gate oxide and the channel stayed the same while their sizes

were reduced, leading to reliability concerns. The supply voltage was reduced to overcome these problems. Consequently the threshold voltage was reduced. As a result the off state current increased and gradually became a limiting factor in scaling of MOS devices. Therefore while modeling a nanometer scale device the “subthreshold” behavior has to be considered.

Basically there are three different regions of MOS operation based on the inversion. They are weak inversion, moderate inversion and strong inversion. The current flow can be due to two reasons drift and diffusion. Under weak inversion the channel surface potential is almost constant across the channel and the current flow is determined by diffusion of minority carriers due to a lateral concentration gradient. Under strong inversion there exists a thin layer of minority carriers at the channel surface and a lateral electric field which causes a drift current. The moderate inversion regime is considered as a transition region between weak and strong inversion where both current flow mechanisms coincidently exist. In the weak-inversion (or subthreshold) region , the drain current depends exponentially on the gate-source voltage The exponential subthreshold behavior is due to the exponential dependence of the minority carrier density on the surface potential which, itself, is proportional to the gate voltage. On a semi-logarithmic scale the transfer (or I_d-V_g) curve in the subthreshold region will, therefore, be a straight line.

3.1.2 Gate leakage

As the device is scaled down the gate oxide thickness is also reduced. Therefore the leakage current can flow from channel to gate attain prominence. And thus the gate cannot be considered as perfectly insulated. This affects the circuit functionality and increases the static power. Also the gate leakage reduces the clock cycle time. The two tunneling mechanisms which causes gate leakage are, Fowler-Nordheim tunneling and direct tunneling. Direct tunneling means tunneling in nano- MOS structures electrons from the conduction band in semiconductor are transferred across the oxide directly (i.e. without changing energy) into the conduction band of metal; probability of direct tunneling is a very strong function of the width of the barrier electron tunnels through (oxide thickness in MOS devices). And Fowler-Nordheim means tunneling of electrons from semiconductor conduction band into oxide conduction band through part of the potential barrier at the semiconductor-oxide interface; most likely to dominate current in MOS structures with oxide 5-10 nm thick. The gate leakage increases exponentially as the oxide thickness is reduced. This limits the down-scaling of the oxide thickness to about 1.5-2nm when looking at the total standby power consumption of a chip.

3.1.3 Hot carrier effects

When a MOS transistor is operated in pinch-off condition, also known as ``saturated case", hot carriers traveling with saturation velocity can cause parasitic effects at the drain side of the channel. This is known as ``Hot Carrier Effects" (HCE). These carriers have sufficient energy to generate electron-hole pairs by Impact Ionization. The generated bulk minority carriers can either be collected by the drain or injected into the gate oxide. The generated majority carriers create a bulk current which can be used as a measurable quantity to determine the level of impact ionization. Carrier injection into the gate oxide can also lead to hot carrier degradation effect. This happens when the threshold voltage changes due to occupied traps in the oxide. Hot carriers can also generate traps at the silicon-oxide interface known as ``fast surface states" leading to subthreshold swing deterioration and stress-induced drain leakage. In general, these degradation effects set a limit to the lifetime of a transistor; thus they have to be controlled.

3.1.4 Drain Induced Barrier Leakage

In long channel devices, the gate is completely responsible for depleting the semiconductor. But in very short channel devices, part of the depletion is accomplished by the drain and source bias. This is shown in figure below. Since less gate voltage is required to deplete the channel,

threshold voltage decreases as length decreases. Similarly, as drain voltage is increased, more substrate is depleted by the drain bias, and hence threshold voltage decreases. These effects are particularly pronounced in lightly doped substrates.

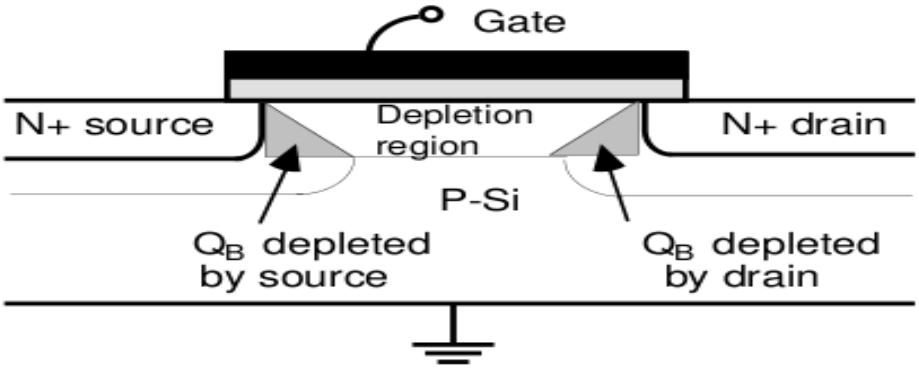

Figure (1)

If the channel length becomes too short, the depletion region from the drain can reach the source side and reduces the barrier for electron injection. This is known as punch through.

In devices with long channel lengths, the gate is completely responsible for depleting the semiconductor. In very short channel devices, part of the depletion is accomplished by the drain and source bias. Since less gate voltage is required to deplete semiconductor, the barrier for electron injection from source to drain decreases. This is known as drain induced barrier lowering (DIBL). DIBL results in an increase in drain current at a given gate voltage. Therefore threshold voltage decreases as length decreases. Similarly, as drain voltage increases, more semiconductors are depleted by the drain bias, and hence drain current increases and threshold voltage decreases.

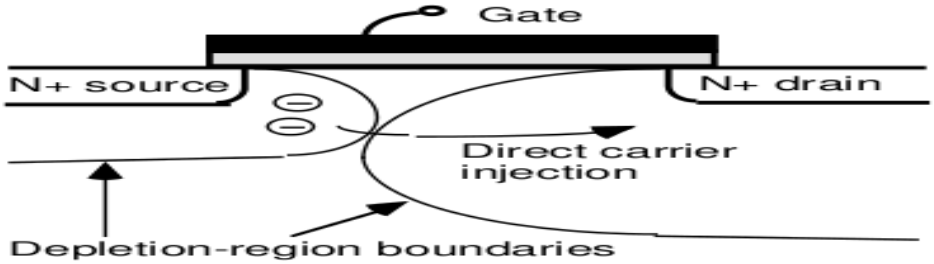

Figure (2)

3.2. BUTT AND SOHI MODEL

The first experimental result regarding the sub-threshold leakage was the Butts and Sohi model. This present a high-level model of sub-threshold leakage that neatly compartmentalizes some different issues affecting static power in a way that makes it easy to reason about leakage effects at the micro-architecture level. But it doesn't include the leakage effects due to temperature. Even then the Butts and Sohi model provides a starting point for the study. The power dissipation according to Butts and Sohi is given by,

$$P_{static} = Vcc \times N \times Kdesign \times Ileakage \qquad (1)$$

VCC is the supply voltage, and N is the number of transistors in the circuit, which could be estimated by comparing it with a circuit of known functionality. Kdesign is a factor determined by the specific circuit topology and accounts for effects like transistor sizing, transistor stacking and the number and relationship of NMOS and PMOS transistors in a circuit. Ileakage is a normalized leakage value for a single transistor- can be named as a unit leakage-and includes all technology-specific effects like threshold voltage (V_T) and also factors in the operating temperature. This model helps in finding various parameters which can be used to control the effect of leakage power savings since Ileakage is proportional to operating voltage and number of transistor used. Unfortunately this will have only little usage in today's manufacturing process. The effect of vdd and temperature are not included in this model. And so it is not suited for actual simulations.

3.3. BSIM MODEL

In order to get an accurate model that includes the different leakage parameters a detailed literature survey was done. The leakage model of a single transistor was found to be as given below:

$$I_{leakage} = \mu 0 \times Cox \times (w \div l) \times e^{b(vdd-vdd0)} \times vt^2 \times (1 - e^{-vdd \div vt}) \times e^{(-|vth| - Voff) \div n \times vt} \quad (2)$$

Low-level parameters are derived using transistor-level simulations: $\mu 0$ is the zero bias mobility, COX is gate oxide capacitance per unit area, W=L is the aspect ratio of the transistor, $e^{b(vdd-vdd0)}$) is the DIBL factor derived from the curve fitting method, Vdd0 is the default supply voltage for each technology (Vdd0 =2.0 for 180nm, Vdd0 =1.5 for 130nm, Vdd0 =1.2 for 100nm and Vdd0 =1.0 for 70nm), vt = kT=q is the thermal voltage, Vth is threshold voltage which is also a function of temperature, n is the subthreshold swing coefficient, Voff is an empirically determined BSIM3 parameter which is also a function of threshold voltage. In these parameters, u0 , COX ,W=L and Vdd0 are statically defined parameters; the DIBL factor b, subthreshold swing coefficient n and Voff are derived from the curve fitting method based on the transistor level simulations; Vdd , Vth and $vt = kT \div q$ are calculated dynamically.

The above equation is based on two assumptions:

1. Vgs =0 — consider only the leakage current when the transistor is off.
2. V_{ds} =V_{dd} — consider only a single transistor here; the stack effect and the interaction among multiple transistors are taken into account when we model the cell using Equation (3)

The V_{th} in the above equation was extracted from BSIM3 paper and is given by

$$V_{th} = \underbrace{V_{th0ox}}_{A} + \underbrace{K_{1ox} \cdot \sqrt{\Phi_s - V_{bseff}}}_{B} - \underbrace{K_{2ox} V_{bseff}}_{C}$$
$$+ \underbrace{K_{1ox}\left(\sqrt{1 + \frac{Nlx}{L_{eff}}} - 1\right)\sqrt{\Phi_s}}_{B} + \underbrace{\left(K_3 + K_{3b} V_{bseff}\right)\frac{T_{ox}}{W_{eff}' + W_0}\Phi_s}_{C}$$
$$- \underbrace{D_{VT0w}\left(\exp\left(-D_{VT1w}\frac{W_{eff}' L_{eff}}{2 l_{tw}}\right) + 2\exp\left(-D_{VT1w}\frac{W_{eff}' L_{eff}}{l_{tw}}\right)\right)(V_{bi} - \Phi_s)}_{D,E}$$
$$- \underbrace{D_{VT0}\left(\exp\left(-D_{VT1}\frac{L_{eff}}{2 l_t}\right) + 2\exp\left(-D_{VT1}\frac{L_{eff}}{l_t}\right)\right)(V_{bi} - \Phi_s)}_{F}$$
$$- \underbrace{\left(\exp\left(-D_{sub}\frac{L_{eff}}{2 l_{to}}\right) + 2\exp\left(-D_{sub}\frac{L_{eff}}{l_{to}}\right)\right)(E_{tao} + E_{tab} V_{bseff}) V_{ds}}_{G,H}$$

Figure (3)

The parameters mentioned are

V_{tho} : Threshold voltage at V_{sb} =0
K_1 : 1st order body effect coefficient
K_2 : 2nd order body effect coefficient
K_3 : NW coefficient
ϕ_s : Surface potential
Nlx : Lateral non-uniform doping parameter
K_{3b} : Body effect coefficient of K_3
W_0 : NW parameter
D_{vt0w} : 1st coefficient of NW effect on V_{th} for SCL
D_{vt1w} : 2nd coefficient of NW effect on V_{th} for SCL
V_{bi} : Built in voltage of PN junction between S-S
D_{vt0} : 1st coefficient of SCL effect on V_{th}
D_{vt1} : 2nd coefficient of SCL effect on V_{th}
D_{sub} : DIBL coefficient exponent in ST region
E_{ta0} : DIBL coefficient in ST region
E_{tab} : Body bias coefficient for ST DIBL effect

For a specific cell, the leakage current is given by the following equation:

$$I_{cellleakage} = (n_N \times I_N + n_P \times I_P)Kdesign \qquad (3)$$

nN and nP are the number of NMOS and PMOS transistors in the cell, and IN and IP are the calculated leakage current of NMOS and PMOS; when aspect ratio W=L = 1 it is unit leakage; kdesign is the design factor determined by the stack effect and aspect ratio of transistors. Kdesign is derived from transistor-level simulation of an actual design and layout of a cell of interest. Given a cell, the average leakage Icellleakage is derived from the transistor-level simulation with all possible input combinations. Kdesign is the factor which accounts for the transistor aspect ratio (W=L) and the stack effect. (The stack effect refers to the additional reduction in leakage when multiple series-connected transistors are off; for example, sleep transistors take advantage of this.) Unlike the BUTT and SOHI model, here the kdesign does in fact vary with temperature, supply voltage, threshold voltage, and channel length.

3.4. IMPROVED MODEL

The Single kdesign model is suitable for cases where the parameters of N and P transistors are very close. If these two sets parameters of N and P transistors differ significantly, different kdesign should be applied to these two types of transistors. Thus the improved leakage model: double-kdesign model is used. For a specific cell, the leakage current is now given by this different equation:

$$I_{cellleakage} = n_n \times k_n \times I_n + n_p \times k_p \times I_p \qquad (4)$$

kN and kP are the design factors of N and P transistors and they can be derived by the same

method as in the single- kdesign model. For a given cell, divide all possible inputs into two groups: one group inputs will turn off the pull-down network composed of N transistors. The other group will turn off the pull-up network composed of P transistors. Thus the leakage currents are also divided into two groups I1n , I2n ,. . . , Ikn ,. . . and I1p , I2p ,. . . , Ikp ,. . . . Ikn is the leakage current when the pull-down network is turned off, while Ikp is the leakage current when the pull-up network is turned off. kn and kp are given by the following equation:

$$K_n = (I_{1n} + I_{2n} + \cdots + I_{kn} + \cdots) \div (N \times n_n \times I_n) \quad (5)$$
$$k_p = (I_{1p} + I_{2p} + \cdots + I_{kn} + \cdots) \div (N \times n_p \times I_p) \quad (6)$$

N is the number of all possible combinations. The improved leakage model is used to estimate the leakage current of the single transistor.

4.1. 3D MEMORY SIMULATION & MODELING

To model the 3D memory accurately we started with the basic circuitry. The cache memory consists of SRAM cells. The SRAM has a 6 transistor structure. The leakage current of this arrangement is calculated using the double kdesign methodology explained in the initial part of the paper. This leakage current is then used for the 2D and 3D analysis. The leakage current obtained in this method include all the effects which was explained earlier. And so this is the most accurate method for finding leakage current.

The software which used for simulation is HOTSPOT and STiMuL. Hotspot is an embedded tool that takes power dissipation and floor plan as the inputs and gives the temperature profile of the arrangements. The power dissipation given is based on the static power extraction methodology described earlier. Both two dimensional and three dimensional circuits can be modeled using this. The results are tabulated. From this the memory cells are modeled and the temperature profiles were obtained. The temperature profile obtained in this method is more accurate because the process started from circuit level.

Now the thermal model while packaging is to be considered. For this the STiMuL software is used. STiMuL provides the ultimate model for packaging. It includes modeling of 2D structures and also packaging. According to the input parameters the output plot is obtained which gives the temperature distribution along the distance. In this paper first the model without packaging effect was modeled. The results are compared with those obtained in Hotspot. Then the packaging was included and the results are analyzed. Detailed explanation is given in the next section.

5.1. EVALUATION AND RESULTS

The evaluation process has been divided into three.

- The first part consist of modeling of a single transistor considering all the leakage effects such as sub-threshold leakage, gate leakage, hot carrier effects and DIBL. The results include calculation of leakage current of PMOS, NMOS, CMOS and SRAM cell. The plots of variation in leakage current with V_{dd} and variation of leakage current with temperature were obtained. This makes the first stage of the evaluation and is named as thermal modeling.
- In the second step the thermal profile of conventional 2D memory circuits was obtained using HOTSPOT. Here each memory cell having a power dissipation of 10W. For 3D circuits seven cases are considered and the results are explained below in detail. This step is named as circuit modeling.
- As the third step the whole packaging is to be modeled. For this STiMuL is used. Initially 2D circuits were modeled to get familiarized with the software. In STiMuL the heat generating areas are given as the heat sources and the corresponding plots are obtained.

This step is named as package modeling.

5.2. THERMAL MODELING

The leakage current of transistor level circuit was found out based on the conclusions from literature survey. Based on the leakage model a code was written in C language. The aim of the code was to obtain the leakage current. An option was introduced to select among NMOS and PMOS. The parameters such as V_{dd} and temperature can be changed. Thus the plots of leakage current with variation in temperature and V_{dd} are plotted.

For the 3D simulation the leakage current of an SRAM cell was also calculated. The 6T structure of SRAM cell was considered. The results are given below.

Leakage current of NMOS transistor ($I_{leakage}$) $=8.585\text{x}10^{-7}$ A
Leakage current of PMOS transistor ($I_{leakage}$) $=1.692\text{x}10^{-7}$ A
Leakage current of SRAM cell ($I_{cellleakage}$) $= 1.027 \text{ x } 10^{-6}$ A

The variation of leakage current with temperature and supply voltage is found in order to check the accuracy of the thermal modelling process. If the curve follows the ideal curve then the strategies taken are perfect. In this case both the characteristics follow the ideal curve and so the strategies are accurate and perfect. Leakage current with temperature graph attached below.

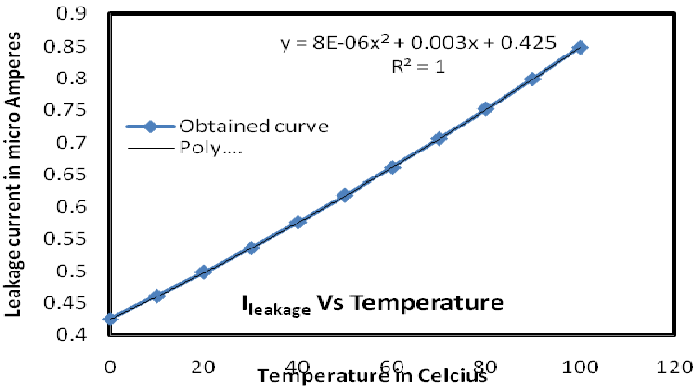

Figure (4)

The above given plot is the variation of leakage current with temperature. The ideal behavior of leakage current with temperature is super linear curve. The result obtained from our simulation follows the ideal curve. Temperature is varied from 0^{o} C to 100^{o}C by keeping V_{dd} constant and $I_{leakage}$ is calculated. The figure 5 shows the variation of leakage current versus supply voltage.

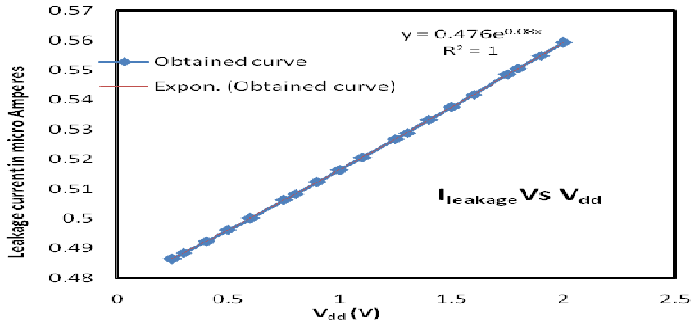

Figure (5)

The above given plot is the variation of leakage current with $V_{dd.\ supply\ voltage\ or\ Vdd}$ is varied from 0v to 2v keeping temperature constant and leakage current was calculated. The ideal characteristics should have a parabolic shape. The curve obtained from the simulation can be approximated to parabolic. Thus the simulation has circuit level accuracy.

5.3. CIRCUIT MODELING

The circuit modeling process consists of finding the temperature of all inner circuit elements. This step was accomplished by the software HOTSPOT. Both 2D and 3D simulation was possible in Hotspot. The 2D simulation results were plots and 3D results were temperature values.

5.3.1. 2D Modeling

Initially 2D circuit models where used for simulation. The circuit has 4 cache memory, eight other circuit components which can be application specific depending on the manufacturer and a crossbar for interconnection. The power dissipation of other circuit is assumed to be uniform low compared to the memory cells. Here the maximum power dissipation is given to the memory cells. Then the operating temperature of memory cell was found to be 331.26k. This range is permissible for proper operation of memory cell. The results and plots are shown below. The 3D simulation in next section is given more importance in this paper.

Case 1: 2D 10w uniform PD

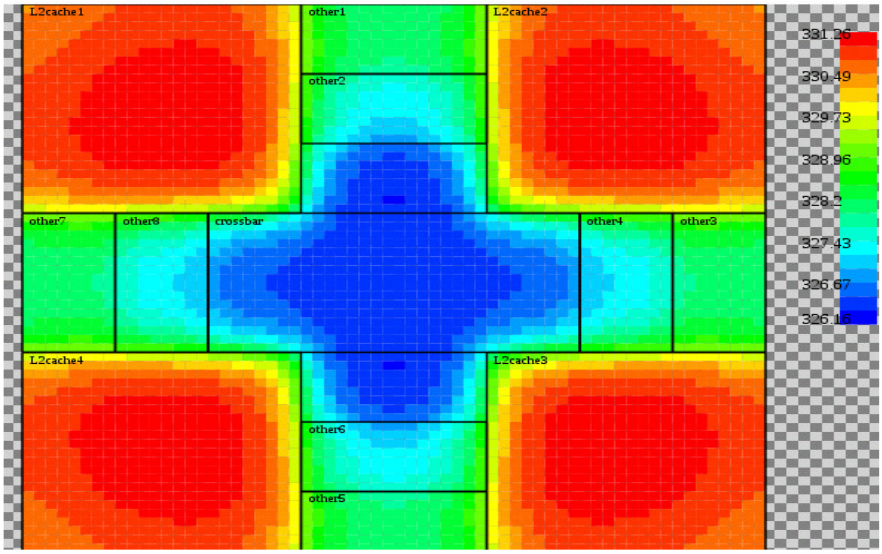

Figure (6)

5.3.2. 3D Modeling

The 3D modeling facility provided in the hotspot software gives the temperature in each layer. The highest temperature in each layer is taken for the comparative study. Five experiments were done for 3D memory model with five layer architecture. The five layer architecture means five memory layer is stacked vertically. The stack consist of five layer of silicon active layer each separated by an epoxy layer. The epoxy layer is used mainly to isolate the silicon layer. Therefore totally there will be 8 layers. The floor plan for the 3D simulation is same as that of the 2D circuit. The same 2D circuits are stacked together to form 3D structure.

The first experiment is with 5 layers of silicon and with uniform power distribution for all the

memory cells. The floor plan and power profile files are the same as that of the 2D case. Here all the layers and all cells in all layers are powered. The resistivity of the epoxy is kept fixed for first three cases.

In the second experiment very thing is same except the power dissipation. The power dissipation is changed to 10 $w/cm2$ for memory cells. The dissipation for other circuits and crossbar are same as in experiment 1.

In the third experiment only layer 1 is powered. Also in layer 1 only the diagonal memory elements are powered. This helps in analyzing the memory when two of the cells are accessed. This strategy is used to emulate the dynamic thermal management. When the 3D structure is working it is not necessary that all the layers work simultaneously. There are some strategies to reduce the operating temperature by selective powering of the memory cells. One such method is diagonal powering of memory cell. Other methods are selective placing of data, selective reading etc.

For the fourth experiment the resistivity of the epoxy layer has changed. The epoxy layer has been inserted to provide a glue interfacing. So the characteristics has to be determined if the epoxy layer is changed. The epoxy resistivity is change in order to study the effect of heat transfer with different materials used as glues. This also helps in transferring the hotspot temperature created during the operation to another spot so that the decay of cell due to hotspot creation is avoided. Three cases are considered.

The fifth experiment is a minor variant of fourth experiment where only the epoxy resistivity is changed. This experiment was done to emulate the effect TSV (through silicon vias) interconnects for 3D stacked chip.

The results are tabulated and given in the table below. The plots are given below with all the five experiments.
Experiment 1: 3D- 5 layer-uniform-3w/cm2- epoxy resistivity=0.25-all powered
Experiment 2: 3D- 5 layer-uniform-10w/cm2-epoxy resistivity=0.25-all powered
Experiment 3: 3D- 5layer-uniform-3w/cm2-epoxy resistivity=0.25-layer1 l1 l3 powered
Experiment 4: 3D- 5layer-uniform-3w/cm2-epoxy resistivity=0.001-all powered
Experiment 5: 3D- 5layer-uniform-3w/cm2-epoxy resistivity=0.005-all powered

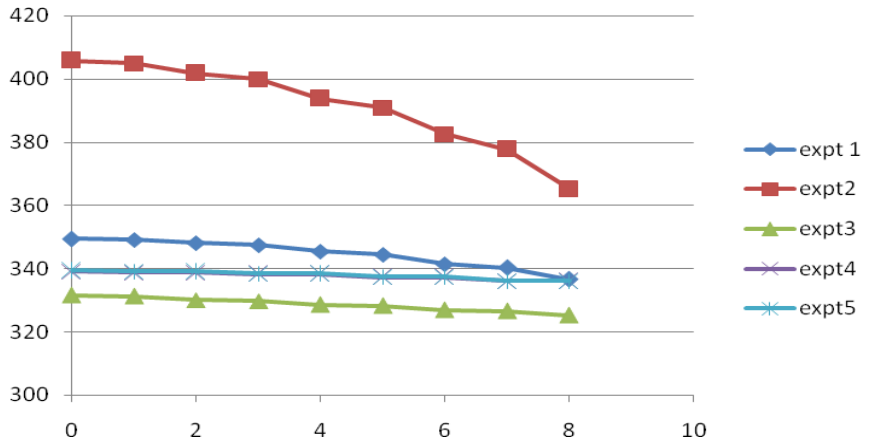

Figure (9)

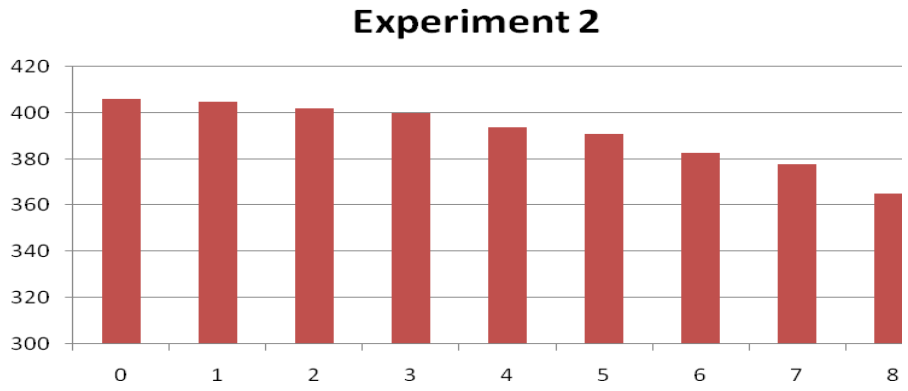

Figure (10)

The temperature value varies largely only with the change in power dissipation. All other case have a similar response when other parameters are kept constant. The temperature value decreased in the case when only one layer is powered. The change in epoxy resistivity from 0.005 to 0.001 dosent cause any change in temperature. But the temperature decreased when the epoxy resistivity was changed from 0.25 to 0.005.

The case 2 was alone considered and plotted in bar chart. The temperature level decreased as the number of layer increased. This is because of the lateral spreading of the heat generated in the cells and no boundary conditions specified.

5.3.3. Comparison with 3 layer silicon & 5 layer silicon

To study the effect of number of layers a comparison is done between 3 layer architecture and 5 layer architecture. The results are shown in the table below. All the parameters except the number of layers are kept constant.

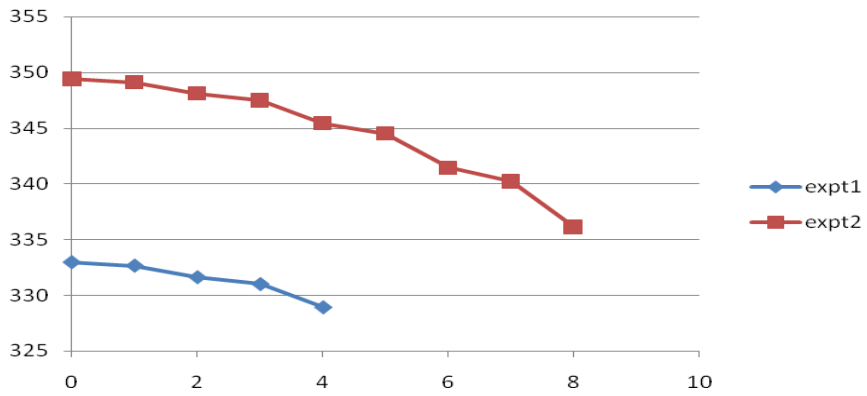

Figure (11)

Experiment 1: 3D- 3 layer-uniform-3w/cm2- epoxy resistivity=0.25-all powered
Experiment 1: 3D- 5 layer-uniform-3w/cm2- epoxy resistivity=0.25-all powered

As the number of layers increases the temperature at the bottom layer of the architecture rises sharply to higher degrees. But the decrease of temperature due to heat dissipation is more or less same with any number of layers.

5.4 PACKAGE MODELING

The circuit modeling with the hot spot software was compared with the results of StiMuL. Exact modeling of 3 D structures is not available with StiMuL and so only approximate modeling is done. The result obtained while simulating in StiMul is given below.

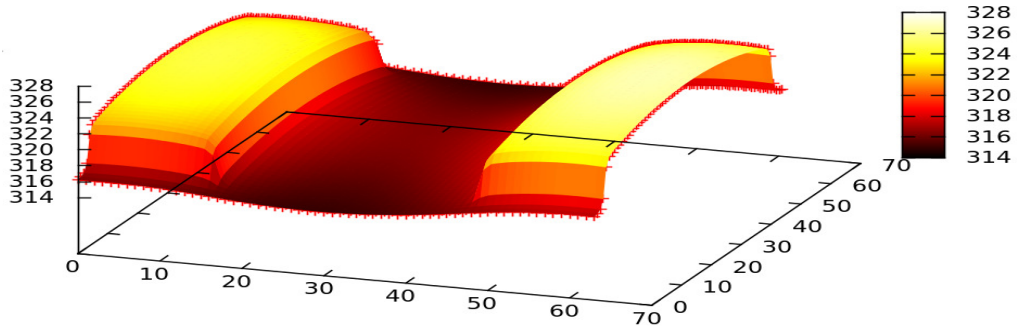

Figure (12)

The next step is to include packaging and heat sink. The inclusion of heat sink should reduce the temperature dissipation. Also the packaging has to be considered to ensure the proper working of the chip. The simulation result obtained while including packaging is given below.

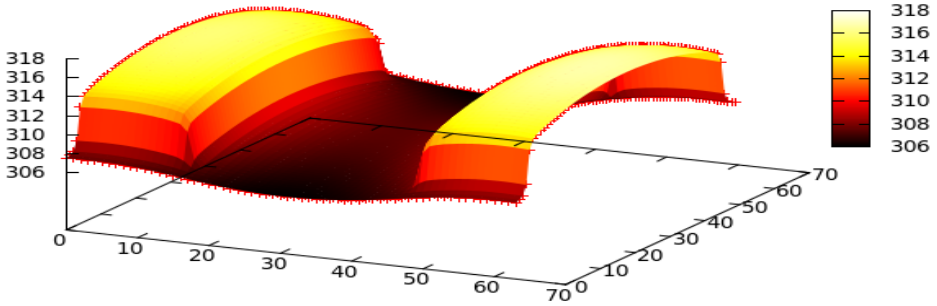

Figure (13)

The maximum temperature obtained in hotspot simulation was 331 k. When the same condition was applied in StiMul the result was 328k. This is clear from the figure (13).
When the heat sink and packaging was included the temperature reduced to 318 k. The obtained results are desirable since they lie in the required range.

CONCLUSION

This paper proposes the use of device and circuit level temperature dependent static leakage extraction methodology as an efficient and accurate methodology for modeling temperature induced thermal effects in 3D stacked memory chips. A through study of different type of leakage mechanism in MOS transistor was conducted to formulate the leakage extraction methodology and the module was very much helpful to extract a fast and accurate power profile of a circuit. This paper discusses the detailed mechanism of how the temperature induced leakage extraction module can be interfaced with a standard 3D thermal modeling tool like HOTSPOT.

By applying this methodology we have obtained the most accurate thermal model for 2D and 3D integrated chips. We tried to address the complete product development issues of 3D integrated chip through this paper by introducing the architectural level modeling and also package level modeling. We are able to establish this is an effective methodology for 3D integration, especially in circuits like memory. This methodology can also be extended to other areas like multi-processor system on chip (MPSoC).

ACKNOWLEDGEMENTS

This paper is the beginning of our ride to intellectual sorority and blossom the innate thoughts. We have not traveled in a vacuum in this journey. There are some people who made this journey easier with words of encouragement and more intellectually satisfying by offering different places to look to expand my theories and ideas. First and foremost we would like to express our sincere gratitude to God almighty whose power helped us all the way. Then we express our heartfelt thanks to Ms. Liza Annie Joseph, Head of Department, Applied Electronics & Instrumentation, and Prof. Asha Paniker, Head of Department, Electronics & Communication Engineering.

REFERENCE

[1] Vasilis.F.Pavlidis, Eby.G. Friedman,"3d integrated circuit design", Morgan Kaufmann series in system on silicon.

[2] John F. McDonald,"Thermal and Stress Analysis Modeling for 3D Memory over Processor Stacks" ,Rensselaer Polytechnic Institute, Troy, NY 12180, "SEMATECH Workshop on Manufacturing and Reliability Challenges for 3D IC's using TSV's"

[3] Stephen Tarzia, "A Survey of 3D Circuit Integration", March 14, 2008.

[4] Igor Loi, Luca Benini," An efficient distributed memory interface for many core platform with 3D stacked DRAM", University of bologna,Italy.

[5] H.Mangalam and K.Gunavathi "Gate and subthreshold leakage reduced SRAM cells"DSP Journal, Volume 6, Issue 1, September, 2006

[6] Howard Chen, Scott Neely, Jinjun Xiong, Vladimir Zolotov, and Chandu Visweswariah "Statistical Modeling and Analysis of Static Leakage and Dynamic Switching Power" IBM Research Division, Thomas J. Watson Research Center.

[7] Yan Zhang, Dharmesh Parikh, Karthik Sankaranarayanan, Kevin Skadron, Mircea Stan "HotLeakage: A Temperature-AwareModel of Subthreshold and Gate Leakage for Architects" University of Virginia Charlottesville, March 2003

[8] Weidong Liu, Xiaodong Jin, James Chen, Min-Chie Jeng, Zhihong Liu, Yuhua Cheng, Kai Chen, ansun Chan, Kelvin Hui, Jianhui Huang, Robert Tu, Ping K. Ko and Chenming Hu" BSIM3v3.2.2 MOSFET Model", University of California, Berkeley, 1999

[9] Howard Chen, Scott Neely, Jinjun Xiong, Vladimir Zolotov, and Chandu Visweswariah" Statistical Modeling and Analysis of Static Leakage and Dynamic Switching Power",IBM Research Division

[10] Pierre Michaud," STiMuL: a Software for Modeling Steady-State Temperature in Multilayers - Description and user manual", INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE.

[11] S. Das, A. Chandrakasan, and R. Reif, "Design tools for 3-D integrated circuits," in Proc. of ASP-DAC, 2003.

[12] HOTSPOT, University of virginia, http://lava.cs.virginia.edu/HotSpot/HotSpot-HOWTO.htm

Authors

Annmol Cherian Vazhappilly student for B.Tech degree in Applied Electronics & Instrumentation at Rajagiri School of Engineering & technology affiliated to Mahatma Gandhi University. The project currently working on is thermal modeling & 3D integration of memory. She is an author of the paper named "Thermal Aware Static Power Extraction Methodology for Nanoscale Integrated Circuits". Participated in national level conferences and presented the paper.

Ajay Augustine student for B.Tech degree in Applied Electronics & Instrumentation at Rajagiri School of Engineering & technology affiliated to Mahatma Gandhi University. The project work under taken was thermal modeling & 3D integration of memory. He is an author of the paper named "Thermal Aware Static Power Extraction Methodology for Nanoscale Integrated Circuits". Participated in national level conferences and presented the paper.

Vinod Pangracious received the B.Tech degree in Electronics Engineering from Cochin University of Kerala in 1995 and M.Tech degree from IIT Bombay India in 2000. Currently pursuing PhD at University of Pierre and Marie Curie Paris France. He is currently an Associate Professor at Rajagiri School of Engineering & Technology India. He has authored and co-authored 10 publications in these areas. He is an internationally experienced electronics engineering professional with extensive expertise in memory design, high speed digital circuit design, logic library development, verification, test and characterization. His research interest focus on design methodologies for integrated systems, including thermal management technique for multiprocessor system on chip, novel nanoscale architectures for logic and memories, 3D integration and manufacturing technologies.

Jemy Jose kakkassery Completed B.Tech in Electronics and Communication Engineering, from Amal Jyothi College of Engineering, affiliated to Mahatma Gandhi University, Kerala. Worked as lecturer at Jyothi Engineering College for one and a half years. Presently pursuing M.Tech in VLSI and Embedded System, at Rajagiri School of Engineering and Technology, India. Project under taken was Automation of Partial Reflection Radar System using LabVIEW- from **VSSC-ISRO.** Currently doing research on Predictive Leakage Analysis on Nano-scale CMOS Technology.