ABSTRACT

Title: 3D INTEGRATION, TEMPERATURE EFFECTS, AND MODELING

Latise Anitrá Parker, Master of Science, 2005

Directed By: Professor Neil Goldsman, Department of Electrical and Computer Engineering

Practical limits to device scaling are threatening the growth of integrated circuit (IC) technology. A breakthrough architecture is needed in order to realize the increased device density and circuit functionality that future high performance ICs demand. 3D integration is being considered as this breakthrough architecture. In this thesis, the limits to scaling are noted and the feasibility of overcoming these limits using 3D integration is presented. The challenges and considerations, most notably dangerously high chip temperatures, are provided. To address the temperature concern, a mixed-mode simulator that calculates temperature as a function of position on chip is detailed. The simulator captures the important link between individual device and full chip heating. Lastly, circuit simulations and lab experiments are performed to experimentally validate the claims that differences in device activity on chip leads to dangerously high local and overall chip temperatures.

3D INTEGRATION, TEMPERATURE EFFECTS, AND MODELING

By

Latise Anitrá Parker

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Science
2005

Advisory Committee:
Professor Neil Goldsman, Chair
Professor Martin Peckerar
Professor Bruce Jacob

Dedication

This work is dedicated to my mother, Sharon. Mom, you have always been there to support me academically and you have always encouraged me to stay tough and work through the rough times. Your support has brought me to this point.

Acknowledgements

Two years of work can be contributed to one person—God. So first, I want to think Him for making me who I am and for bringing me to this point. Without His ever-present love, strength, and guidance, I would be nothing.

I want to thank my advisor, Dr. Neil Goldsman, for giving me the opportunity to perform my graduate study as a member of his research group. As a senior in undergraduate school, I had no idea what type of research I wanted to conduct. I only knew what I didn't want to do. But he took the time to show me the projects his group was working on and he welcomed me into his group. During stressful times, he was available to talk and offer support. I just want to say thank you for giving me the opportunity to be a part of your team.

To everyone in the lab, I extend my thanks for being the main source of help and support on a day to day basis. You helped with courses and with research. We all worked together to make this thesis possible. Akin, you are who I collaborated with the most. You were always patient and willing to discuss anything. You answered questions, sometimes the same one more then once, and you never seemed frustrated or too busy.

My thanks also go to Professor Tits, my undergraduate advisor, and everyone I worked with in the M.E.R.I.T. program. Each of you contributed to my decision to pursue a graduate degree in the first place. Your faith in my abilities encouraged me to believe that I could handle the combined BS/MS program.

I want to say thanks to my good friend Steve for being my study buddy in undergrad and my first 2 semesters of graduate courses. Theresa, after going from kindergarten all the way through graduate school with you, I can honestly say that I am ever thankful for our friendship (and this crazy twist of fate that has pitted us together for so long). We have been classmates, roommates, and friends during all those years and your intelligence, thoughtfulness and insight has helped me in more ways then you will ever know. To all of my friends in the bowling center, thank you for your support and thank you for listening when I complained about homework, exams, and my 100-page thesis. There is only so much a person can take, and each of you tolerated more. Now we can have peaceful dinners at Bennigan’s.

To the love of my life, Jeffrey, I want to say thank you for teaching me about balance. You showed me the importance of taking time to work and time to play. Without that lesson, I’m sure I would have gone crazy long before I reached this point. More then that, you probably heard the most about my ups and downs in graduate school and life in general. Through it all, you remained by my side. I love you for that—and so much more.

Finally, Mom, Gerryle, Brianna, Gerryle Jr., Angie, Grandma and Grandaddy: thank you for being a supportive family. You have each contributed to this work in your own way, perhaps with words of encouragement or by providing me with an outlet to balance school life with personal life.

To anyone I missed, you know you are important to me and you know there is a page in this thesis that I could not have wrote without you. My journey to this point has been helped along by a huge cast of characters, the names of which I cannot possibly mention without this acknowledgements section being as long as my thesis (it probably already is). God Bless you all.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction and Motivation

Since the invention of the field effect transistor (FET), the electronics industry has seen unparalleled growth. Integrated circuits have evolved from large chips containing only a few transistors to chips with areas less then 100 square millimeters containing over 50 million transistors. These performance gains are possible because designers have found ways to increase the functionality of integrated circuits by realizing faster, smaller devices and by using more devices on a single chip. The performance advances have followed an exponential behavior known as Moore's Law, first observed by George Moore in 1965. The most recent projection of Moore's Law, the 2004 International Technology Roadmap for Semiconductors, predicts processors with device gate lengths down to 20 nm and device densities of over one billion transistors per chip by the year 2016 [1]. This projection is 70 nm smaller and a transistor count of almost twenty times greater then current manufactured technology. While the projections sound realistic given the pace of improvements to date, the reality is that fundamental physical limits may mean the end to Moore's Law. Current methods for realizing faster devices and packing more of them on chip are quickly reaching their limits. The motivation of this work is to present these limiting factors, to introduce three-dimensional integration as a new architecture to overcome these limits, and to present a method for calculating and controlling chip heating in a three dimensional circuit.

Device Scaling Limitations

Device scaling is the shrinking of transistor characteristics to produce smaller, faster devices. Ideally, a transistor's characteristics are scaled up or down by a factor α to maintain an electric field inside the device that is the same as the original device. This "constant field scaling" results in circuit speed increasing in proportion with α and circuit density increasing as α^2 [2]. Unfortunately, ideal scaling theory is reaching its practical limits. As gate lengths become smaller, voltages, gate-oxide thicknesses, and device lengths can not be scaled ideally because of short channel effects, quantum effects, and application tolerance limitations [2].

When device lengths are small, quantum and short channel effects lead to random variations in gate length and other device parameters, leading to large differences in device characteristics. Physical dimensions are limited by leakage currents through various barriers. Voltage scaling is limited by the non-scalable bandgap of silicon and its effect on built-in junction voltages. Supply voltage scaling is limited by the need for sufficient gain for logic functionality. Finally, non-scaling of the subthreshold slope is the limiting factor for threshold voltage [2]. In all, the inability to continue ideal scaling of these parameters will lead (and in some cases, has already led) to a point where process tolerances and proper device behavior cannot be obtained at small feature sizes. The result is an inability produce smaller, faster devices and thus, a potential roadblock in the ability to pack more and more devices on chip.

Interconnect Limitations

The ever-growing demand for functionality and higher performance requires more transistors to be closely packed on a single chip. However, all of these devices must be connected together with interconnects that that do not inhibit the propagation of the signal. While scaling has had a positive effect of allowing designers to incorporate more devices on chip, the impact on interconnects has been just the opposite. Interconnect performance degrades with scaling due to smaller wire pitch and wire cross sections. Also, the increased device density and larger chip area creates the need for an increased number of interconnects that must traverse longer distances. All of these factors lead to an increase in wire resistance and capacitance, thereby increasing signal propagation delay. Also, increasing interconnect loading affects the power consumption in high performance chips, a drawback for low power systems [3].

Solutions to the interconnect problem have been investigated. Designers have tried widening interconnect lines as well as utilizing new materials to combat the increased delay times. At the 250 nm node, copper with low-*k* dielectric was successfully introduced. However, as device features continue to scale, interconnect delays will result in spite of the new materials, and it is believed that no new materials will be available to solve the problem [3]. While widening the metal lines is another solution, it also creates a need to add more metal layers to the fabrication process, which increases the complexity and cost of the process and can also degrade circuit reliability [4]. In all, there is an increasing need for a new architecture that will allow designers to pack more and more devices on chip and connect these devices in an efficient manner.

Thermal Limitations

Ideally, power density remains unchanged when devices are scaled. However, because supply voltages can not be scaled ideally due to the scaling limits previously addressed, power density actually increases for the scaled device. Also, the demand for greater functionality and the presence of smaller devices allows for more devices to be used on a chip. The end result is a higher chip power density because of the use of individual devices drawing more power and more of these devices on chip. In addition, the packaging of many VLSI chips into compact packages, such as multi-chip modules, also creates higher then desired power densities for the packaged component [5]. The energy consumed as power is drawn is converted to heat and results in detrimentally high chip temperatures.

There are a number of potential problems that result from additional heat being generated. It has been stated that for every 10 ºC increase in temperature, the failure rate of microelectronic devices doubles, leading to a serious long term reliability concern [5]. Typical thermally induced analog failures include input offset voltage and offset voltage drift in differential amplifiers, reference voltage shifts in regulators and data converters, and nonlinearities in multipliers [5]. In digital circuits, logic errors are the main concern. The source of these problems is not overall chip heating, but rather local heating. Since power dissipation is not uniform across the chip, localized heating occurs much faster then heating across the entire chip. The extra heat generated can lead to hot spots—areas where large temperature gradients exist. Thus, there is a need for methods to calculate localized chip temperature in addition to overall chip temperature, and ways to cool hot spots before any timing errors occur or any physical damage is done. Finally, it signals an alert for future VLSI design: any new method employed to allow for the increased packaging of devices or systems should also be temperature considerate.

Thesis Overview

The purpose of this thesis is three-fold:

1. to present 3D integration as a way to group more devices in a single area as well as overcome the limitations of interconnect scaling.
2. to present a methodology for calculating localized chip temperature for a 3D chip.
3. to present experimentation on the effect of device activity on chip temperature and extend these results for verification of the methodology mentioned in 2.

The following is an outline for the remainder of this thesis

- In chapter 2, integration in three dimensions is presented. Design structure, fabrication techniques, vertical interconnect technology, and challenges and considerations are presented.
- In chapter 3, a simulator that calculates chip temperature as a function of position on chip is presented. The specifics of heat theory and thermal networks are explained and the algorithm for the simulator is detailed. Results from simulations are provided.

- In chapter 4, validation of the claims stated in this thesis is presented. Simulations of the effects of temperature on frequency are presented and designs for measuring local chip temperature and selectively heating areas of a chip are provided. Results of experiments conducted after fabrication of the designs are offered.
- In chapter 5, a summary of the findings of this thesis are given.

Chapter 2: Integration in Three Dimensions

Vital to the continued growth of advanced integration is the ability to increase chip functionality. So far, designers have done this by scaling device features to increase the complexity of chip designs, resulting in an increase in the number of devices on chip and an increase in chip area. The fact that scaling not only allows for the use of more devices on chip, but also results in faster devices, is the other important factor in the performance gains to date. However, to continue to meet the demand for highly functional and faster integrated circuits, the limitations discussed previously must be overcome. A paradigm shift from present integrated circuit architecture is needed. The new architecture should meet the ever-growing demand for higher functionality and performance without increasing chip area, exceeding the limitations of interconnect technology, or creating dangerously high local or overall chip temperatures.

At present, integrated circuits are produced in a planar way. Devices are fabricated on a single layer of silicon and are interconnected to form complex circuits. The idea presented in this thesis is to branch out into the third dimension. The suggestion is to create multiple active layers of silicon for fabrication of devices not only laterally, but vertically as well. The devices are then connected using dense, vertical interconnects. The 3D architecture will lead to potentially smaller chip areas, shorter interconnect paths, higher transistor packing density, and flexibility in system design and placement.

The idea of utilizing the third dimension for integration is not a new one. The concept was demonstrated as early as 1979 with the presentation of silicon thin films formed on an insulator using graphoepitaxy, a technique that uses artificial surface relief structures to induce crystallographic orientation in thin films [6]. The idea remained a concept though, with no real strides being made to actually produce working circuits on multiple layers of silicon. This is because of the overwhelming success of device scaling in bulk silicon CMOS technology. However, with the growing limitations to scaling and the menace of interconnect delays, researchers have turned to 3D integration as more of a necessity, rather than a concept, for maintaining chip performance well into the deep sub-nanometer range. Performance models have been developed to show the gains in chip area, performance, and interconnect delay that can be obtained using three-dimensional circuits [3], [7]-[8]. Furthermore, the actual fabrication and testing of devices fabricated on multiple layers of silicon and the implementation of vertical interconnects has been shown [9]-[18]. The idea of three-dimensional circuits has moved from an idea to a reality. The latest research shows the feasibility of the idea and the practicality of implementation.

Design Structure and Requirements

There are 2 approaches to designing a 3D circuit. The first is to divide the entire chip into blocks and place each block on a separate layer, as seen in Figure 1. Since each layer is stacked, the blocks are connected using vertical interconnects. The connections within blocks should be simple, eliminating the need for more then one or two layers of metal for each block layer. The second approach is to fabricate the same kind of device on each layer, for example only n-channel MOSFETs on one layer and only p-channel MOSFETs on another (Figure 2). The individual devices are then interconnected vertically to realize the desired element, such as an inverter. In both cases, there is a significant reduction in chip area and interconnect length over the 2D configuration. However, each approach has other unique benefits.

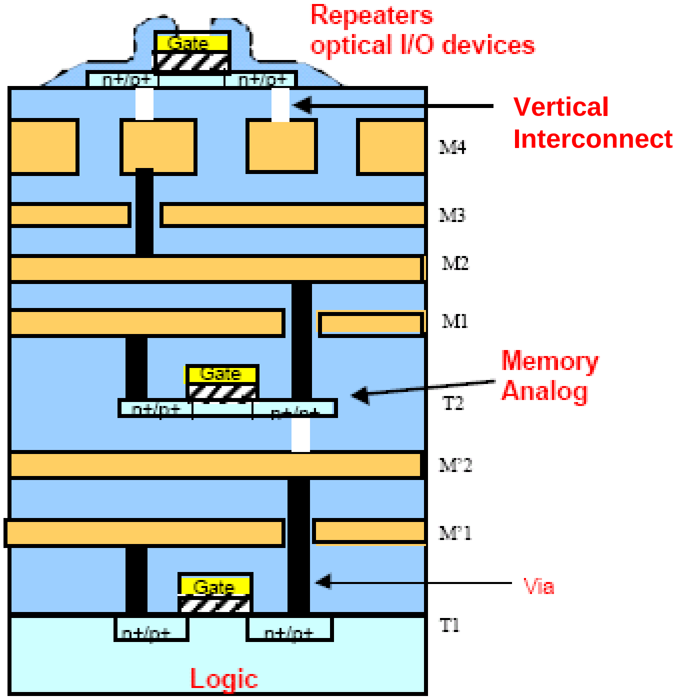


Figure 1: 3D integrated circuit where each layer T is a block of devices. There are separate metal layers for connections on the same layer and for different layers.

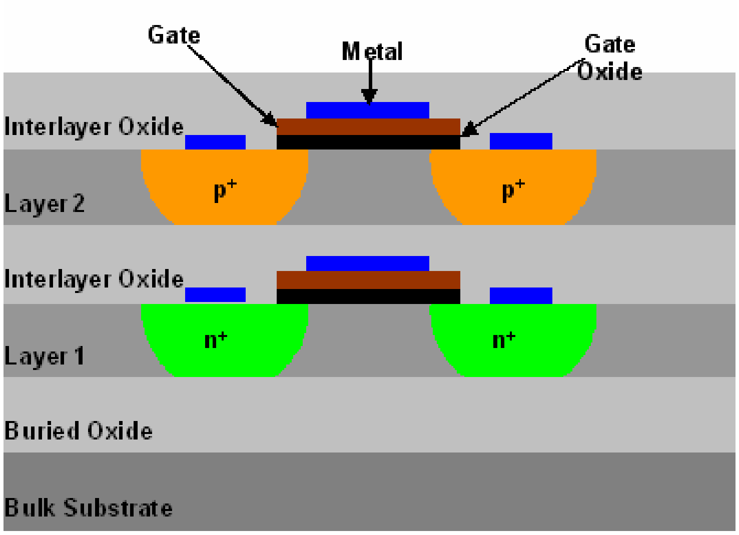


Figure 2: 3D integrated circuit where each layer is a single type of devices.

For the block layer approach, the immediate benefit is flexibility in block placement and routing. Logic gates on a critical path are placed close together within a block and blocks on a critical path can be placed side by side or vertically for faster signal propagation due to shorter interconnect lines. Furthermore, the ability to build highly integrated systems is evident. Circuits with different voltage and performance requirements can be put on different layers for isolation. Analog and digital components in mixed-signals systems can be placed on different layers to achieve better noise performance due to lower electromagnetic interference [3]. Figure 3, from [3], shows an example of a 3D chip with logic, memory, analog, radio frequency, and optical circuits on separate layers, integrated into a single 3D chip.

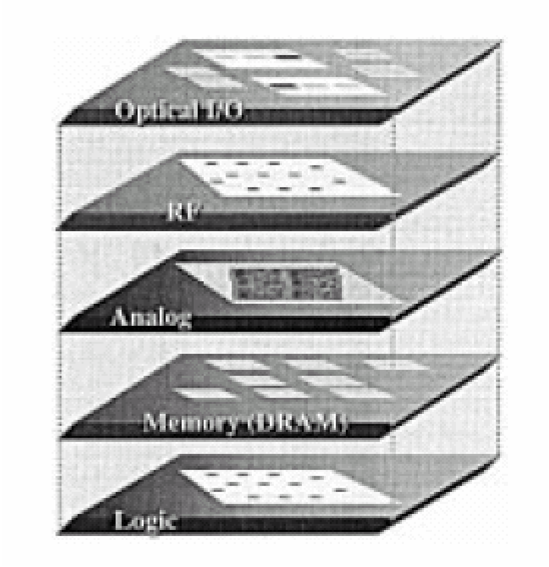


Figure 3: 3D systems-on-chip (SOC) consisting of logic, memory, analog, RF, and optical components, each on a separate layer and integrated onto a single 3D chip.

For the device layer approach, the benefits lie in fabrication and routing. First, separate device layers reduce the process steps for each layer by almost half [8]. This is because there is no need to use wells for fabrication of n- and p-channel devices on the same substrate. Thus, there are no extra ion implants or long thermal anneals at high temperatures. Instead, designers can just use the proper doping depending on the device being fabricated. This approach also allows for better routing capabilities because designers can take advantage of the typically one-third smaller area of n-channel devices over p-channel devices and use this area for routing and interconnects. Finally, single devices are more compact then blocks of devices, so the interconnections between devices are shorter and devices are closer together, allowing for better packing density.

There are basically two design requirements for a 3D circuit: proper device behavior and uniformity in behavior across all devices on all layers. These requirements are satisfied during both design and fabrication of the multiple layers of silicon. For both the device and block layer approaches, obtaining the multiple layers can follow either a sequential or parallel process flow. In a sequential process, each layer is constructed on top of the previous layer. This requires complete fabrication of the devices or blocks on one layer before the next layer is fabricated. This approach requires subjection of devices on previous layers to all of the process steps for fabricating the current layer. Thus, low temperature fabrication processes are required to ensure proper device behavior and uniformity, since subjecting completed layers to high temperatures can degrade the integrity of the previous layers. For a parallel process, each layer is fabricated individually and then all layers are combined together. This approach avoids repeatedly subjecting subsequent layers to process steps and also helps with device uniformity, since each layer is done separately using the same steps. However, this method introduces the problem of alignment. Proper alignment insures good device and circuit behavior after interconnection, so care must be taken to ensure proper device and circuit function.

Fabrication Using Silicon-on-insulator (SOI)

The foundation for 3D integration is the stacking of multiple active layers. One way of achieving multiple stacked active layers is through silicon-on-insulator (SOI). SOI is a layered structure of thin single-crystal silicon either on top of an insulating substrate, such as quartz or sapphire, or separated from a bulk silicon substrate by an insulating layer such as silicon dioxide [19]. The foundation of the SOI structure is a bulk silicon substrate. Above it is an insulating layer, called the buried oxide. On top is a film of silicon. Devices are fabricated on this film, rather then directly on the silicon substrate, as in conventional bulk silicon CMOS.

There are many advantages to utilizing SOI technology, namely increased reliability, faster circuit operation, and low-voltage low-power operation. The advantages are due to differences between SOI and bulk silicon technology. Figure 4 from [20], shows transistors on bulk silicon and silicon-on-insulator.

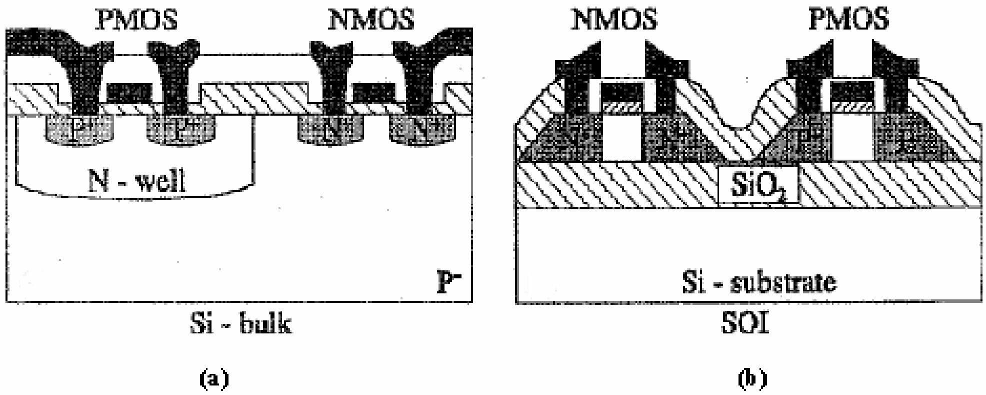


Figure 4: CMOS transistors fabricated on (a) bulk silicon and (b) SOI.

One benefit SOI offers is inherent device isolation, an advantage created by the buried oxide. The thin film of silicon above the buried oxide can be fabricated into "islands" that comprise each device. The buried oxide allows for vertical isolation of the islands so that devices will not share the same substrate, as in conventional CMOS. Thus, vertical isolation is achieved as a benefit of the process, rather then by the sophisticated schemes employed in conventional bulk technology. Separate device substrates eliminate the need for the deep wells that facilitate fabrication of n- and p- channel devices on the same substrate in conventional CMOS. This simplifies process steps by as much as 30 percent and overall circuit area by up to 60 percent [19].

Circuit reliability is also improved in SOI, mostly due to the innate isolation it offers. First, device malfunction due to latch-up is essentially nonexistent in SOI because there is no need for wells or deep trenches for isolation. Thus, there are no parasitic paths between highly doped areas, well, and substrate to cause latch-up. Second, soft errors due to radiation are reduced because the insulating layer in SOI limits the volume of the body of the device to only the thin silicon island, rather then a large bulk substrate. This is a volume reduction of 2 to 3 orders of magnitude [20]. Thus, there is less volume available for a radiated particle to induce ionization as well as less volume available for storing any ionized charge.

Another benefit of the SOI structure is decreased junction capacitance. From Figure 4, it is evident that the source and drain regions extend all the way down to the insulator so that only their lateral sides can serve as junctions. The result is a reduction in the area of overlap of the source and drain regions with the body, thereby significantly reducing junction capacitance as compared to bulk technology. The reduction in the parasitic capacitance significantly improves circuit speed, allowing higher frequency circuits to be constructed.

Lastly, the SOI structure allows for low power and low voltage application. The smaller junctions of SOI films permit better performance over bulk circuits as supply voltages fall. The structure also produces devices with steeper subthreshold voltage slopes then bulk devices. The subthreshold slope determines the gate bias needed to assure an "off" condition in a device. Requirements for a specific on-off current ratio limit the minimum threshold voltage the device can have, which ultimately limits the required power supply voltage for the device [21]. Thus, in SOI the steeper subthreshold slopes allow for lower possible threshold voltages and thus lower supply voltages, all without loss of speed [19]. The effect is lower supply voltages and reduced leakage current, both leading to lower power consumption.

There are commercial options available for fabrication of single layer SOI wafers. One is Separation by Implanting Oxygen (SIMOX), a process in which the buried oxide is formed by internal oxidation of the silicon substrate during a high energy oxygen implantation. A subsequent high temperature anneal is necessary to recover the crystalline quality of the thin silicon film [20]. Wafer bonding, another method, involves fusing together two oxidized wafers (or one oxidized and one bare wafer) at room temperature. After annealing to increase the bonding strength, one of the two wafers is thinned to the proper thickness by grinding, polishing, and etching. The technique provides undamaged crystal quality and a wide range of thickness for both the buried oxide and SOI film [20].

While the single layer methods are sufficient for creating one SOI active layer, the subsequent active layers for the 3D chip must still be fabricated. Each active layer must contain high quality single crystal silicon and the process steps for obtaining the silicon layer must not change the characteristics of the active layer beneath it. SIMOX and wafer bonding processes prove unsuitable for creating multiple layers because of the high temperatures they require. Thus, while designers can use the single layer SOI techniques aforementioned for first layer devices, they must employ other techniques for creating the subsequent layers. Another alternative is to develop a completely new approach to SOI that, while requiring fabrication from bulk rather than utilizing SOI wafers for the first layer, is more suitable for repetition for 3D integration.

There are at present 2 methods suitable for creating multiple layers using SOI: processed wafer bonding and silicon epitaxial growth. The first is a parallel process while the second is a sequential one. The choice of method employed will depend on the requirements of the system, since performance is strongly influenced by the electrical characteristics of the devices fabricated as well as on the manufacturability of the process. The specifics of these two methods are described in the subsections that follow.

Processed Wafer Bonding

Processed wafer bonding is an appealing approach for creating 3D circuits that involves fusing together two or more fully processed wafers. The wafers, usually single layer SOI type, are processed individually to create working 2D circuits complete with planar interconnects. The individual wafers are bonded, face to face or back to back, either directly or indirectly. Direct bonding methods exploit the intermolecular attractive forces that exist between two smooth surfaces brought into intimately close contact. Indirect bonding involves the deposition of an intermediate layer of bonding agent followed by a combination of temperature and force processes to secure the bond. For either method, interlayer vias can be etched before or after the bonding process for vertical interconnections between layers.

One requirement of wafer bonding processes is strong and secure bonds. For direct bonding, in order to achieve sufficient bonding energy at low temperatures, bonding surfaces must be smoothed, flattened, and cleaned. Achieving a smooth surface free of contaminants is critical for maximizing the density of bonding species on the mating surfaces. Defects on the surface reduce reactivity and lower the bond energy [22]. For indirect bonding, bond strength can be optimized by adjusting the bond process parameters, including ambient temperature, pressure, bond temperature, and surface treatments.

Another requirement of wafer bonding processes is preservation of the electrical integrity of the bonded layers. This requirement demands the use of processes that work at low temperature and minimize mechanical stress. High temperature is detrimental to metal layers used for interconnects, as it can lead to undesired metal diffusion through barrier layers. Also, unwanted dopant activation in the active areas of devices can occur at high temperatures. There are reports of direct and indirect bonding processes that are successful at temperatures below 400 ºC [22]. Mechanical stress is reduced by proper preparation of the surfaces for easier bonding and also by embracing low temperature, low pressure techniques.

Perhaps the biggest demand on wafer processes for 3D integration is adequate alignment. Proper alignment is critical for good electrical connections. Each layer must be sufficiently aligned to the layer beneath it if a good vertical interconnect is to be fabricated. One method is transparent alignment, which involves transferring a device layer from its bulk silicon substrate onto a transparent glass "handle" substrate. The new wafer is then inverted and bonded to the layer beneath, using the transparent glass substrate on top for visual alignment. The temporary glass substrate is then removed by laser ablation. This bonding method has been shown to preserve the electrical integrity of long and short channel devices after transfer [12]. Also, alignment can be done using infrared cameras that align through the silicon substrate. Another alternative is flip-chip bonding, a method that achieves alignment in the x- and y- directions using optics. The misalignment is reportedly less then 1 μm and can be done at a temperature of 400 ºC [3].

In all, wafer bonding proves to be a promising method for 3D integration, provided limitations in alignment can be overcome. It is suitable for creating multiple layers in parallel, all with similar electrical properties. Also, it imposes no temperature limitation on device fabrication, since each layer of devices is fabricated separately. Only bond temperature is limited, and this has already proven to be surmountable. The feasibility of integrating 3D ring oscillators and coupling optical signals generated in n^+p diodes on one layer to CMOS circuits on another layer has already been shown [13]. As stated before, wafer bonding using a temporary glass substrate has also been demonstrated. Measurements after layer transfer yielded less then 10 percent degradation in drain current and threshold voltage for 65 nm n-channel MOSFETs and only a slight increase in stage delay of ring oscillator circuits composed of 55 nm CMOS inverters [12]. However, to achieve the state of the art performance enabled by deep nanometer devices, the reported best case alignment of ± 2 μm in [3] must be improved.

Silicon Epitaxial Growth

Silicon epitaxial growth (SEG) is a sequential technique for creating multiple active layers of silicon that uses a hole etched in the insulating layer to grow single crystal silicon seeded from the substrate. The SEG process, shown in Figure 5 from [17], typically takes place in a low pressure chemical vapor deposition reactor. The carrier gas is hydrogen, dichlorosilane supplies the silicon, and hydrochloric acid prevents the formation of polysilicon on the insulator [23]. A thermal oxide is deposited on a bulk silicon substrate and etched to create islands where the silicon will selectively grow. A thin oxide is then deposited in the islands to provide insulation for the island and the substrate beneath. Next, a seed window is opened adjacent to the oxide wells. Silicon is grown first vertically through the seed window and then laterally over the oxide. Device quality material is produced when temperature is between 860 °C and 1000 °C at a pressure of 50 to 150 Torr [23]. The excess silicon grown outside of the islands is removed using chemical mechanical planarization. Thus, silicon islands are obtained selectively within the oxide wells. Subsequent layers are formed the same way.

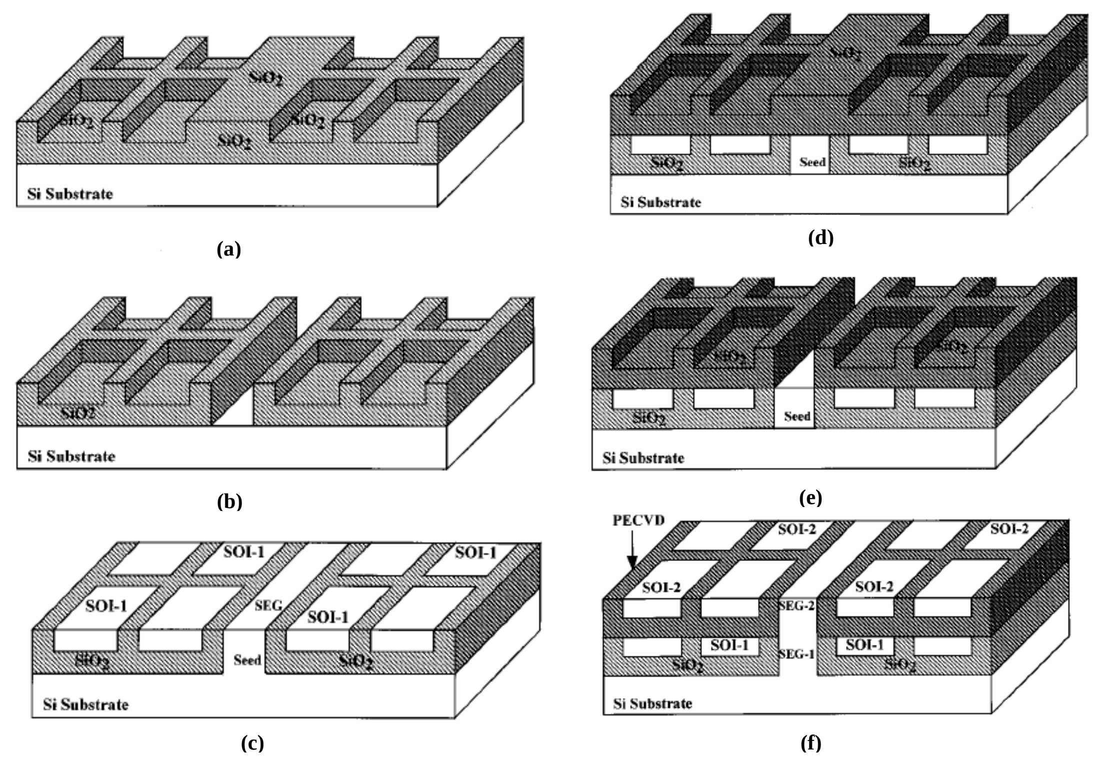


Figure 5: Process steps for fabricating multiple layers of silicon using Silicon Epitaxial Growth: (a) deposition, patterning, and insulation of islands, (b) opening of seed window, (c) silicon growth and chemical mechanical polishing, (d) deposition, patterning, and insulation of 2nd layer islands, (e) opening of 2nd layer seed window, and (f) silicon growth and chemical mechanical polishing of 2nd layer islands.

The ability to create single crystal silicon as pure as the underlying substrate is one advantage SEG affords. Provided few faults occur in the growth process, devices should behave the same since they are all fabricated on the same silicon seed. Furthermore, the growth of the silicon is not overly complicated and is compatible with CMOS fabrication processes. Some methods for creating multiple layers of silicon do not offer the ease of process like SEG does, nor do they offer the same crystal quality.

However, SEG is not without its disadvantages, the biggest being thermal budget. While other techniques have much lower process temperatures, SEG relies on temperatures as high as 1000 °C for silicon growth. As previously stated, temperature this high can be very straining on devices fabricated on lower levels. If SEG is to be used to fabricate state of the art devices with small gate lengths, process temperature will need to be lowered significantly to ensure good device performance.

There has been success in forming multiple layers of silicon and even submicrometer devices using SEG. These advances demonstrate the feasibility of 3D integration using SEG. First, there are reports of good crystalline quality SEG at temperatures as low as 750 °C in ultra-high vacuum epitaxy systems [14]. Furthermore, silicon islands as small as 150×150 nm have been developed on 2 layers [17]. At this size, it is possible to accomplish terascale integration (10^{15} devices per die) with devices with gate lengths of 25 nm. Furthermore, fully depleted p-channel MOSFET with gate lengths of less then 100 nm were fabricated on both layers to show the quality of the silicon. The fabricators report normal current-voltage characteristics as well as decent on-off, threshold voltage, and subthreshold slope values. The findings show that good quality silicon can be obtained using SEG and with improvements, the technology could potentially be used for high performance circuits integrated in three dimensions.

Fabrication Using Polysilicon on Insulator

Polysilicon on insulator is a very popular method of creating a second active layer of silicon. As in SOI, an insulating layer is deposited on a substrate. However, polycrystalline or amorphous silicon instead of pure single crystal silicon is deposited on the insulator. In SOI, obtaining single crystal silicon on an insulator at a temperature suitable for 3D applications is difficult. However, obtaining polycrystalline silicon is not as difficult, since there are fabrication steps already available to deposit the poly layer onto the insulator. One commonly used method is low pressure chemical vapor deposition. After the film is deposited, methods to crystallize certain regions are employed in an effort to control grain size and orientation. Thin film transistors fabricated on polysilicon layers have been widely used in active matrix displays for years. The ease of fabrication demands a look at ways of improving device characteristics in order to produce devices that could one day be used in 3D circuits.

The problem with polysilicon thin film transistors is performance. Polysilicon thin film transistors suffer low mobility and high threshold voltages because of grain boundaries. Grain boundaries create trap states due to bonds left dangling because of the change in crystal orientation. These defects can trap electrons at the boundary, inhibiting current flow and degrading device performance. This phenomenon distinguishes thin film transistors from bulk silicon and SOI MOSFETs, which do not suffer from such degradation because the silicon used is pure, single crystal silicon free of defects.

The other problem with thin film transistors on polysilicon is device uniformity, an important requirement for high performance 3D integration. A strategy for modeling variations in device performance due to grain size was developed [24]. It predicts that as grain size and device size converge, variation in device characteristics increases because of variations in grain size, location of grain boundaries, and the number of grains in a single device. This is very important considering high performance devices have nanometer gate lengths. The model also predicts there would be no variation in device performance if grain size and grain boundary can be precisely controlled. Thus, uniformity among devices is a huge concern to be addressed if polysilicon thin film transistors are to be used for 3D circuits. Methods need to be employed to eliminate—or at least control—grain boundaries so that devices can be fabricated away from such locations for better performance and uniformity.

It is believed that the requirement of having single crystal silicon can be relaxed as long as the channel region of the MOSFET lies on a single grain [10]. This belief has prompted many to try and improve thin film transistor performance for use in 3D circuits by obtaining localized regions of single crystal silicon within the amorphous or polycrystalline silicon. If a device could be made to lie on a single grain, then theoretically, it should perform identically to a device formed in highly pure single crystal bulk silicon. Thus, the challenge for polysilicon on insulator for 3D applications is obtaining large single crystal grains of silicon within the polycrystalline layer. Devices fabricated would offer device performance similar to bulk technology. There are at present 2 methods being explored to produce such devices: beam recrystallization and solid phase crystallization. The specifics of these two processes are explained in the following subsections.

Beam Recrystallization

Beam recrystallization is a method commonly used for creating an active layer of silicon atop an existing substrate. Polysilicon or amorphous silicon is deposited on an insulating oxide layer and an intense laser or electron beam is used to induce crystallization. The beam can be focused entirely on one area and used to heat only that area, without much heat being spread through the underlying insulator. Thus, protection of underlying layers is obtained.

Work was done in the early nineties to produce devices and working 3D circuits using beam recrystallization. Mobility and subthreshold slope comparable to bulk silicon technologies were achieved for some devices [25]. However, the leakage current for such devices was unacceptably high. Also, a four-layer 3D device with a primitive function of parallel image signal processing was realized [16], but the techniques used do not appear to be useful for highly scaled circuits due to the high temperature processes employed for interconnections. Lastly, a four layer chip containing a programmable logic array, CMOS gate array, and a CMOS synchronized random access memory was fabricated [18]. The performance was limited by large variations in threshold voltage due to unintentionally doped impurity concentration from vertical isolation layers. It was also limited by fluctuations in gate insulator thickness and carrier mobility due to variation in silicon orientation.

Recently, interest has been renewed in the use of beam recrystallization for thin film transistors. High performance polysilicon thin film transistors using laser beam recrystallization along with low temperature processing have been fabricated [26]. However, it is still difficult to control variations in grain size and grain boundary size. Thus, beam recrystallization will not emerge as a leading technology for 3D integration of high performance devices unless the grain variations can be controlled. This will require obtaining larger single crystal grains that have fewer defects. A beam crystallization method employing sequential lateral solidification to produce single crystal silicon has been preliminarily presented, but further improvements must still be made [27].

Solid Phase Crystallization

Solid phase crystallization (SPC) is another method for fabricating thin film transistors on polysilicon layers. Amorphous silicon is deposited on an oxidized silicon wafer using low-pressure chemical vapor deposition in a silane atmosphere at a temperature of 550 °C. The silicon is then crystallized in the solid phase by thermal annealing in an inert gas such as N_2 or Ar. Low anneal temperatures in the range of 550 °C to 650 °C are selected to hinder the randomness of crystallization. The process, because of its low temperatures, is easily implemented to create multiple layers of polysilicon without thermally damaging devices on lower layers. The drawback: in conventional SPC, there is still no real control over the location of the grain boundaries [28]. Therefore, newer SPC methods have been developed that employ seeding to induce lateral crystallization and allow greater control over grain location [28]-[33].

The process of crystallization begins with nucleation: the formation of single grains of a specific orientation that form and enlarge as the material is heated. The time it takes for nucleation to begin is the incubation time, while the rate at which the grains enlarge is the grain growth. In solid phase crystallization, designers exploit the fact that the amount of nucleation relative to grain growth decreases with decreasing temperature [29]. This means that at lower temperatures, grains will have more time to enlarge before nucleation of another grain. This allows designers to produce larger grains and thus have some control over grain size. The disadvantage to the process is that crystallization may take several hours because a fraction of the time is spent in the incubation period [29].

While grain size can be controlled somewhat by utilizing lower anneal temperatures, there is still a need for better control over grain size, location, and boundaries. To achieve this control, the use of seeding agents has been investigated. An illustration of the seeding process is shown in Figure 6 from [28]. After depositing amorphous silicon, extra process steps are added to selectively deposit seeding agents to induce lateral crystallization with minimal self nucleation. A sacrificial oxide is deposited and etched to expose seed windows where the seeding agents will be deposited. The crystallization by annealing is then performed. The seeds stimulate lateral crystallization and the grains are grow from the specified locations defined by the windows. The deposited seeding agent and the sacrificial oxide are then removed. Another anneal is employed to enlarge the grains even further. Finally, thin film transistors are fabricated on the grain a safe distance away from the seeds. Theoretically, the transistors are fabricated on a single grain with no boundaries within the channel of the device. Thus, greater control over device variations as well as better device performance is attained.

One seeding agent that has been used in SPC is germanium. Germanium deposited on silicon bonds to form a SiGe alloy. This alloy crystallizes faster then pure silicon, allowing for a reduction in the incubation time and thus the anneal time. Also, germanium does not diffuse fast in silicon, and the amount of germanium needed to induce nucleation is small [30]. These two facts allow for easier removal of the germanium after crystallization because the alloy is confined only to the interface between silicon and germanium. Typically, grain sizes of 1-2 μm can be obtained.

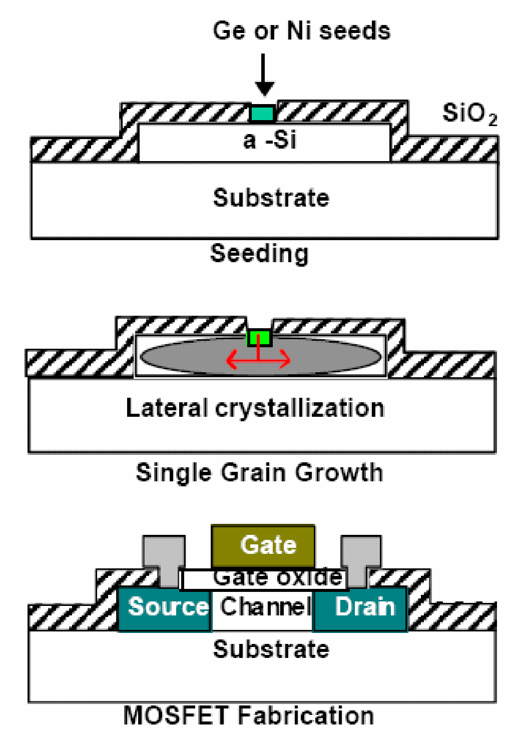


Figure 6: Use of nickel or germanium seed to induce lateral crystallization to create a single crystal region to comprise the channel of the device.

Germanium seeding has been used to fabricate polysilicon thin film transistors with gate lengths of 100 nm [30]. Since the length of the device is smaller then the grain size, the channel lies on single grain silicon. The devices exhibit excellent on-off ratio and off-state performance. Transfer and output characteristics, from [30], are shown in Figure 7 and Figure 8. Seed window sizes have been scaled to dimensions smaller then 1 μm, indicating that seed size is not a limiting factor on scaling the technology for smaller length devices [30]. While there are still no reported 3D circuits constructed using this technology, the method seems promising for achieving vertical integration of high performance circuits with deep submicrometer devices because of its low thermal budget and ability to accommodate small feature sizes.

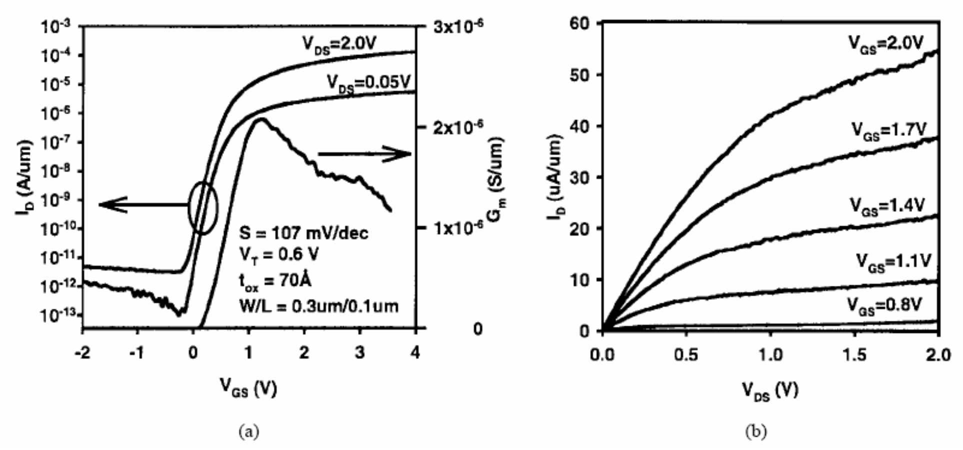


Figure 7: Germanium seeded n-channel MOSFET with gate length of 100 nm: (a) transfer characteristics and (b) current output.

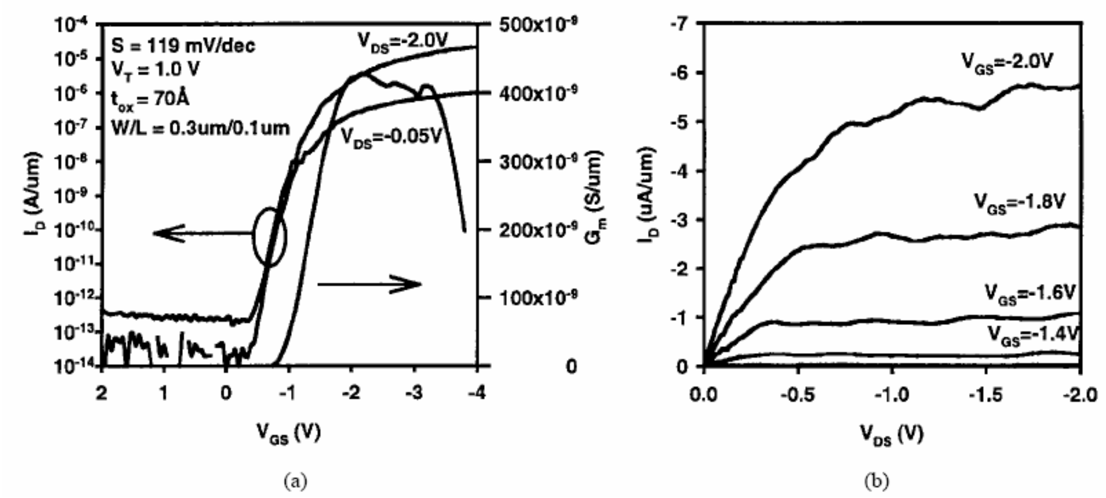


Figure 8: Germanium seeded p-channel MOSFET with gate length of 100 nm: (a) transfer characteristics and (b) current output.

Another seeding agent that has been investigated is nickel. Nickel is used to increase grain size beyond the roughly 1-2 μm obtained using germanium, allowing for fabrication of wider devices. The technique of using a metal such as nickel has been termed Metal Induced Lateral Crystallization (MILC). Nickel deposited on silicon through the etched seed windows bonds to form $NiSi_2$ and acts as a nucleus for crystallization, just as with germanium [10]. The grains grow elongated with major axis aligned perpendicular to the seed strip. After completion of crystallization and removal of the remaining Ni and sacrificial oxide, another anneal is performed to enlarge the grain size in the MILC region. The result is grains with major and minor axis lengths as large as 80 μm and 10 μm, respectively [31]. In addition, the additional anneal can also act as a dopant activation step [28], thereby combining two process steps into one.

There has been considerable success in utilizing nickel seeding for 2D integration. Just as with Germanium seeding, 100 nm high performance thin film transistors have been fabricated using MILC with nickel as the seed agent [33]. The maximum process temperature employed was 500 °C, and the dopant was fully activated during the crystallization process. The process is compatible with CMOS technology, and device layers can be fabricated on top of metal lines because of the low thermal budget. With improvements to device performance, this approach seems very likely to succeed with 3D integration of high performance technology.

There has also been success in accomplishing 3D integration using MILC with nickel as the seed [9]-[10]. Inverters integrated in three dimensions, shown in Figure 9, were fabricated using MILC with maximum process temperature of 900 °C [10]. The n- and p-channel MOSFETs have gate lengths of 0.5 μm and 0.4 μm, respectively. The inverter performance approaches that of an SOI inverter, as shown in Figure 10 from [10]. Ring oscillators were also constructed to show the uniformity and repeatability of the device performance. While these 3D circuits have been constructed, they are still far from exhibiting the high performance seen in bulk technology 2D devices. Furthermore, low thermal budget and small feature size, as in [33], were not achieved. The point is that 3D devices have been suggested and fabricated. With further improvements, especially in utilizing the advances in [33] to create 3D structures with low a thermal budget and small feature size, MILC may definitely emerge as a possibility for 3D integration of high performance circuits.

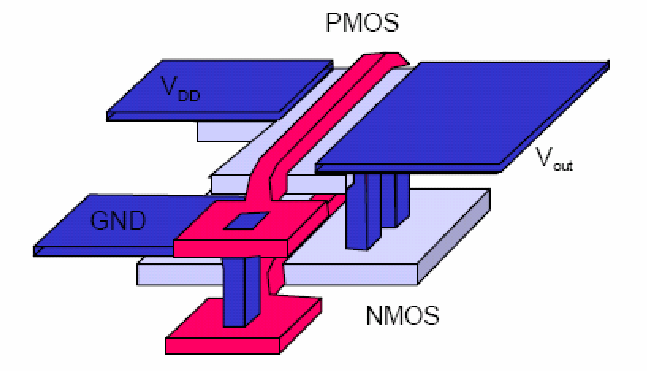


Figure 9: Layout of a 3D inverter fabricated using MILC with nickel as the seeding agent.

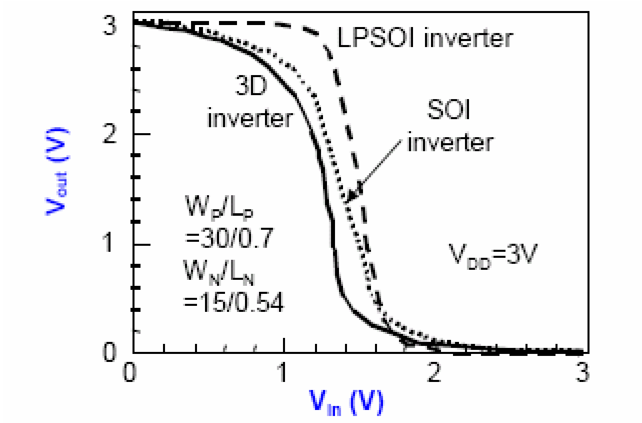


Figure 10: Inverter performance comparison for SOI, Large-grain Polysilicon SOI, and conventional SOI inverters fabricated on the same wafer. The performance of the 3D inverter fabricated using MILC is similar to the conventional SOI inverter.

Vertical Interconnects

In 3D circuits, vertical interconnects will be responsible for connecting active layers. Instead of having long global interconnects, blocks on critical paths are placed close to each on separate levels and connected with vertical interconnects. The proposed structure consists of the various active levels. Each silicon layer will have its own conventional 2D interconnects, but then between each layer there will be a metal layer solely used for vertical interconnects. Dense 3D vias will connect these metal layers.

The nice thing about incorporating vertical interconnects into 3D integration is that current metallization techniques are easily expanded to three dimensions. Copper technology currently being used for 2D interconnects as well current via technology can be used for vertical interconnects without modifications [3]. Thus, no new technologies need to be implemented, only expansion of current techniques.

Furthermore, there are only a few extra process steps for creating the vertical interconnects. In a sequential fabrication process, after the formation of devices or blocks on the level being processed, contacts are made for both vertical and horizontal interconnects. Then, where vertical interconnects are to be defined, dense vias are patterned and filled using conventional techniques. After completion of the next layer, the connections are complete and the vertical interconnect is realized. In a parallel process, the extra steps are even less. All interconnects within a layer are fabricated. The vertical interconnects can either be patterned before or after bonding the layers. If done prior to bonding, vias are etched and can be used as alignment markers for bonding. After bonding, vias are filled in during the metallization process. For after-bonding techniques, etching is used to create the 3D vias for the metallization process. Either way, the additional process steps are few.

Many believe that the huge performance increase that will result from 3D integration is due to the shortening of interconnect lines by using vertical interconnects. In the 3D work that has been done, most of the speed enhancements have been attributed to the vertical interconnects. For example, in [10], the 3D inverter exhibited a delay of 1.4 ns, a delay smaller then an inverter formed completely on a large grain polysilicon area on one level and an SOI 2D inverter. The speed increase is attributed to the reduced length of the interconnect forming the inverter.

The main concern with vertical interconnects continues to be device fabrication process. Subjecting metal layers to high temperature can cause considerable metal diffusion and leakage through barrier layers. The potential for metal contamination or poor interconnections due to diffusion is large and threatening. For true high performance integration, process temperatures will need to be kept below 500 °C for copper technology to protect the reliability of interconnects and to prevent copper diffusion [3]. In short, the 3D interconnect is one of the main reasons for performance gains from 3D integration. As such, its performance will need to be protected throughout the fabrication process.

Challenges and Considerations

While three-dimensional integration appears to be a highly desirable innovation for continuing the trend of Moore’s Law, there are still many roadblocks. As stated previously, low process temperature and good device performance are the main requirements for a 3D technology. Obtaining quality silicon islands for device fabrication at a suitable temperature to fabricate uniform devices is a serious challenge. However, in addition to this fabrication challenge, there are still other challenges to be faced once the challenges of fabrication are overcome.

One consideration is temperature. In 3D integrated circuits, the thermal heat generated is even more important then in two dimensions. As stated previously, high temperature gradients can cause local as well as global chip failure. The same factors that lead to better performance gains in 3D circuits will also lead to higher temperatures. On a 3D integrated circuit, there are even more transistors enclosed in a given area, resulting in an overall increase in heat generated on chip. Furthermore, the SOI structure and the insulating layers, while providing device isolation, have thermal conductivity values generally two orders of magnitude smaller then silicon, thus providing poor thermal insulation. Also, 3D chips will offer systems-on-chip capabilities. This increased functionality of electrical, optical, RF, and other types of signals demand greater surface area for input and output of the signals, thereby reducing the surface area available for heat removal [34].

Thermal packaging will be the biggest factor in keeping 3D chips cool. It has already been stated that thermal packaging technologies with thermal resistance below 0.5 K/W will be necessary to maintain reasonable chip temperature. In [3], it is suggested that providing heat sinks for each active layer can significantly alleviate the thermal problem. No matter what, in order for the 3D chips of the future to operate at maximum performance limits, advancement in cooling and packaging technology will be necessary.

There are also reliability concerns in 3D circuits. As already stated, thermal effects pose a serious reliability risk. In addition, heterogeneous integration and systems-on-chip create the need for a better understanding of the mechanical and thermal behavior of new material interfaces. The use of different fabrication process will create a need for a new and better understanding of bonding interfaces and materials used, as these new processes may lead to new ways for devices to fail.

Furthermore, these new processes may influence existing reliability hazards such as electromigration. There are questions of yield and cost that will have to be dealt with as well. Careful tradeoffs will have to be made between system performance, cost, and 3D manufacturing.

Another area of concern is routing and placement along with computer assisted design. In two dimensions, there are numerous design tools available to figure out the best placements of different blocks on a chip. There are also computer programs available for layout; there are design rules in place for fabrication. The same thing will have to happen in three dimensions. There will be a need for computer assisted programs for 3D chip layout and simulators for 3D circuit behavior. Device models will need to be obtained. The road to 3D integration is a long one, and the research to date has not even begun to scratch the surface.

Summary

Integration in three dimensions is a highly practical and highly logical step for the integration industry. It follows naturally to utilize the third dimension to continue the success of 2D integration brought on by scaling. The ability to pack more transistors, reduce interconnect length, and incorporate different technologies onto a single chip will transform the semiconductor field.

Work has been done to achieve multiple layers of silicon and fabricate devices on these layers. The techniques of wafer bonding, silicon epitaxial growth, beam recrystallization, and solid phase recrystallization are 4 methods currently being researched for potential 3D integration. Each of the processes offer advantages and disadvantages and each still needs a lot of improvement before the world will see 3D integration of today’s high performance, ultra small devices using any of these processes. Based on the work done, high performance 3D integration is still a long way off, considering most of the work presented thus far has only touched the surface by either achieving rudimental 3D structures or simply achieving working devices on multiple levels. 3D integration is still an idea and a research concept, but most believe with serious attention to the possibility, 3D integration will be the saving grace for the future of integration.

Chapter 3: 3D Full-Chip Temperature Simulation

As stated previously, chip temperature is a big concern in current VLSI technology. In addition, it will be a major concern in the future for 3D integrated circuits. As more devices are packed onto chips and as devices are scaled, the overall power density on chip is increasing, resulting in overall increase in chip temperature. The increases, however, are not just global increases for overall temperature, but local increases as well. A potentially big problem with temperature increase is the dangerously high temperature gradients that exist locally. These local gradients result in “hot spots” on chip—areas of detrimentally high temperature. These hot spots can cause local logic errors and even physical damage. In 3D circuits, heat is generated on each layer and a coupling exists between layers, resulting in even higher local temperatures. Thus, in 3D circuits, hot spots may be an even bigger concern. What is needed is a methodology for locating hot spots, calculating the temperature at these locations, and figuring out ways to eliminate such spots.

This chapter presents a model for heat flow in a 3D circuit and an algorithm that uses this model, in conjunction with a device simulator, to calculate chip temperature as a function of position. The model builds on work presented in [35]. While previous work has been done in estimating chip temperature [3], [36], there is a need for a simulator that connects individual device heating with full chip temperature. The methodology presented in this work takes into account the individual contributions of each device on each layer to the local and full chip temperature. It employs spatial dependent device operation as well.

3D Heat Flow Model

To calculate chip temperature as a function of position, it is necessary to understand how heat is generated in an integrated circuit and the methods that allow the spread of the heat. Of course, individual devices on a chip generate heat as they are operated. In a 3D circuit, there are multiple layers of devices generating heat. This heat spreads laterally as well as vertically to other layers. The chip package is used to safely dissipate heat to the ambient. This takes place as heat is conducted through the substrate, through a heat spreader, and finally through a heat sink. To calculate temperature, it is necessary to use equations derived from basic heat theory along with the architectural knowledge of a chip and its packaging.

To calculate the localized temperatures for a chip, start with the general heat flow equation given in equation (3.1). The first term is recognized as the heat stored by a device while the second term can be noted as the net flux. The final term is the heat sourced in the device, $(J_n + J_p)\nabla\varphi$, and consists predominately of Joule heating. The goal is to transform the differential equation into a lumped heat flow equation to reduce the number of mesh points required because of the order of magnitude difference in the scales of a single device and the entire chip.

$$C\frac{\partial T}{\partial t} = \nabla \cdot (\kappa \nabla T) + H \tag{3.1}$$

The first step in transforming the differential equation is to eliminate the temperature dependency of thermal conductivity κ in order to make the equation linear using Kirchoff's transformation presented in [5] and shown in equation (3.2).

$$\overline{T} = T_o + \frac{1}{\kappa(T_o)}\int_{T_o}^{T} \kappa(\tau)\, d\tau \tag{3.2}$$

The linearization begins by taking the gradient of transformed temperature $\overline{T}$. Note that T_0 is the ambient temperature and is constant so that $\nabla T_0 = 0$. The final gradient of transformed temperature is given in equation (3.3).

$$\begin{aligned}
\nabla\overline{T} &= \nabla T_0 + \frac{1}{\kappa\left(T_0\right)}\nabla\int_{T_0}^{T}\kappa\left(\tau\right)d\tau \\
\nabla\overline{T} &= \frac{1}{\kappa\left(T_0\right)}\left[\kappa\left(T\right)\nabla T - \kappa\left(T_0\right)\nabla T_0\right] \\
\nabla\overline{T} &= \frac{1}{\kappa\left(T_o\right)}\kappa(T)\nabla T \qquad (3.3)
\end{aligned}$$

The gradient equation allows substitution of $\kappa(T_o)\nabla\overline{T}$ for $\kappa\nabla T$ in the general heat flow equation (3.1) to eliminate the temperature dependency of thermal conductivity κ. The resulting differential heat flow equation is a linear equation in terms of *T* and $\overline{T}$, as seen in equation (3.4).

$$\begin{aligned}
C\frac{\partial T}{\partial t} &= \nabla\cdot\kappa\left(T_o\right)\nabla\overline{T} + H \\
C\frac{\partial T}{\partial t} &= \kappa\left(T_o\right)\nabla\cdot\nabla\overline{T} + H \\
C\frac{\partial T}{\partial t} &= \kappa\left(T_o\right)\nabla^2\overline{T} + H \qquad (3.4)
\end{aligned}$$

The next step is to find the relation between T and $\overline{T}$. Knowing this relationship will enable the calculation of transformed temperature back to absolute temperature. It will also enable the writing of the linear heat flow equation completely in terms of $\overline{T}$. Begin with the known temperature dependence relation for thermal conductivity κ given in equation (3.5). Note that *n* is a constant determined experimentally for the material under consideration.

$$\kappa(T) = \kappa(T_o)\left(\frac{T}{T_o}\right)^n \tag{3.5}$$

Substitute the relation for conductivity into Kirchoff's Transformation and perform the integration. The final expression for $\overline{T}$ in terms of T is given in equation (3.6).

$$\overline{T} = T_0 + \frac{1}{\kappa(T_0)}\int_{T_0}^{T} \kappa(T_0)\left(\frac{\tau}{T_0}\right)^n d\tau$$

$$\overline{T} = T_0 + \int_{T_0}^{T}\left(\frac{\tau}{T_0}\right)^n d\tau$$

$$\overline{T} = T_0 + \left[\frac{1}{T_0^n}\frac{\tau^{n+1}}{n+1}\right]_{T_0}^{T}$$

$$\overline{T} = T_0 + \left[\frac{1}{T_0^n}\frac{\tau^{n+1}}{n+1}\cdot\frac{T_0^{n+1}}{T_0^{n+1}}\right]_{T_0}^{T}$$

$$\overline{T} = T_0 + \left[\frac{T_0}{n+1}\left(\frac{\tau}{T_0}\right)^{n+1}\right]_{T_0}^{T}$$

$$\overline{T} = T_0 + \frac{T_0}{n+1}\left[\left(\frac{T}{T_0}\right)^{n+1} - 1\right] \tag{3.6}$$

Since it is desired to have absolute temperature T in terms of $\overline{T}$, rearrange equation (3.6) to solve for T. Once transformed temperature is known, the absolute temperature can easily be found using this transformation. The result after rearranging and simplifying is given in equation (3.7).

$$T = T_0\left(\frac{(n+1)(\overline{T} - T_0)}{T_0} + 1\right)^{\frac{1}{n+1}} \tag{3.7}$$

Because it is necessary to write the linear heat flow equation completely in terms of $\overline{T}$, the partial derivative $\frac{\partial T}{\partial t}$ in (3.4) must be transformed and substituted. This is easily done by using the chain rule to find the partial time derivative of absolute temperature in equation (3.7).

$$\frac{\partial T}{\partial t} = \frac{\cancel{T_0}}{\cancel{n}+1}\left(\frac{n+1}{T_0}\left(\overline{T}-T_0\right)+1\right)^{\frac{1}{n+1}-1} \cdot \frac{n+\cancel{1}}{\cancel{T_0}}\frac{\partial \overline{T}}{\partial t}$$

$$\frac{\partial T}{\partial t} = \left(\frac{n+1}{T_0}\left(\overline{T}-T_0\right)+1\right)^{-\frac{n}{n+1}} \frac{\partial \overline{T}}{\partial t}$$

Substituting this partial derivative into equation (3.4) yields the result shown below. Note that the terms in front of the partial derivative are just constants and can be grouped together to form a single constant..

$$\underbrace{C\left(\frac{n+1}{T_0}\left(\overline{T}-T_0\right)+1\right)^{-\frac{n}{n+1}}}_{\overline{C}} \frac{\partial \overline{T}}{\partial t} = \kappa\left(T_o\right)\nabla^2\overline{T}+H$$

After combining the constants to a single term, the result is the desired linear heat flow equation written in terms of the transformed temperature. The equation is shown below as equation (3.8). Note that H, the Joule heating term, is an explicit function of absolute temperature.

$$\overline{C}\frac{\partial \overline{T}}{\partial t} = k(T_o)\nabla^2\overline{T}+H(T) \qquad (3.8)$$

To transform the equation into a lumped thermal equation, integrate the linear differential heat flow equation around a unit volume representing a single device.

$$\int_V \overline{C}\frac{\partial \overline{T}}{\partial t}dV = \int_V k(T_o)\nabla\cdot\nabla\overline{T}dV + \int_V HdV$$

The unit volume is taken to be a rectangular prism so that V is the volume of that prism and S_f is the area of each of the 6 surfaces. Apply the divergence theorem to the second term to transform the volume integral into a surface integral. The divergence theorem and the integral equation that results from its application are shown below in equations (3.9) and (3.10), respectively.

$$\int_V \nabla \cdot \vec{F} dV = \int_{\partial V} \vec{F} \cdot d\vec{S} \tag{3.9}$$

$$\int_V \overline{C} \frac{\partial \overline{T}}{\partial t} dV = \int_S k(T_o) \nabla \overline{T} \cdot d\vec{S} + \int_V H dV \tag{3.10}$$

To compute the integrals in equation (3.10), there are some assumptions and considerations to note. Assume that time and spatial variations in temperature are constant in the volume and on the surface of the prism so that $\frac{\partial T}{\partial t} = \frac{\Delta T}{\Delta t}$, $\frac{\partial T}{\partial x} = \frac{\Delta T}{\Delta x}$, $\frac{\partial T}{\partial y} = \frac{\Delta T}{\Delta y}$, and $\frac{\partial T}{\partial z} = \frac{\Delta T}{\Delta z}$. These terms can be removed from the integral. Also note that heat flows in the direction of decreasing temperature. Since $-k\nabla \overline{T}$ is the heat flux through a surface, incorporate a minus sign for the temperature gradient for each surface integral to ensure that heat flows down the temperature gradient. Finally, remember that $\overline{C}$ and $k(T_o)$ are constants and can be removed from the integrals. The final solution after substitution and integration is given in equation (3.11). Note that $\frac{\Delta T}{\Delta l_f}$ is used to denote the temperature gradient through a surface S_f.

$$\int_V \overline{C}\frac{\Delta\overline{T}}{\Delta t}dV = \sum_{f=1}^{6}\int_{S_f} -k(T_o)\frac{\Delta T}{\Delta l_f}dS_f + \int_V HdV$$

$$\overline{C}\frac{\Delta\overline{T}}{\Delta t}\int_V dV = -\sum_{f=1}^{6}k(T_o)\frac{\Delta T}{\Delta l_f}\int_{S_f} dS_f + \int_V HdV$$

$$\overline{C}V\frac{\Delta\overline{T}}{\Delta t} + \sum_{f=1}^{6}k(T_o)\frac{S_f}{\Delta l_f}\Delta\overline{T_f} = \int_V HdV \tag{3.11}$$

Equation (3.11) is the linear solution to the heat flow equation for a single device. There are 6 faces or surfaces on the rectangular prism, and Δl_f is the distance the heat travels from the center of the device under consideration through a surface to the center of another device. The area of the surface the heat flows through is S_f and $\Delta\overline{T_f}$ is the transformed temperature difference between the centers of the 2 devices. Finally, $\Delta\overline{T}$ is the transformed temperature variation at the center of the device under consideration and $\int_V HdV$ is the Joule heating of the device. To find the full chip heating, formulate this equation for each device on the chip and then solve simultaneously the resulting linear system.

3D Thermal Network

The beauty of equation (3.11) is that it resembles the KCL equations that can be written for an electrical circuit. Consider a node *A* in an electrical circuit. Let there be 6 resistors connecting node A to 6 other nodes, numbered 1 through 6. Consider a current source I feeding the node and a capacitance C connected to the node from ground. Denote the node voltages as V_A for node *A*, and V_1 through V_6 for the neighboring nodes. Kirchoff’s Current Law (KCL) requires that the current into node

A from the current source *I* be equal to the sum of the currents out of the node through the resistors and capacitor. The KCL equation is shown in equation (3.12).

$$C\frac{\Delta V_A}{\Delta t}+\sum_{i=1}^{6}\frac{V_A-V_i}{R_i}=I \tag{3.12}$$

Comparing KCL equation (3.12) with lumped thermal equation (3.11) reveals that the thermal equation looks just like the electrical equation with *T* and *V* interchanged and *H* and *I* interchanged. The conclusion is simple: the thermal equation for the device reveals a thermal configuration of capacitances, resistances, and heat sources that is equivalent to an electrical configuration of electrical capacitances, resistances, and current sources for an electrical node. The equivalent thermal capacitance and resistance is given by equations (3.13) and (3.14).

$$C^{th}=\overline{C}V \tag{3.13}$$

$$R_i^{th}=\frac{\Delta l_i}{k(T_o)S_i} \tag{3.14}$$

For each device, the single thermal capacitance is connected from the midpoint of the device to ground. The six thermal resistances are connected between the midpoints of the device and its six nearest neighbors, and the Joule heat source feeds the device node from ground. Since the heat flow equation is written for each device on chip, the result is a network of device nodes sharing thermal connections. A representative 3D thermal network and a single representative device node with its thermal connections are shown in Figure 11.

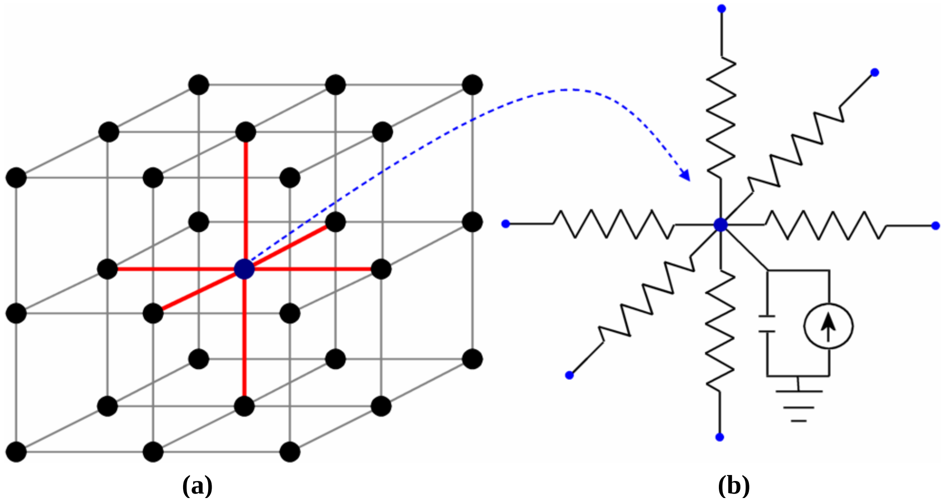


Figure 11: Representative 3D thermal network with each node representing a device: (a) device nodes arranged uniformly and (b) a single device node with its 6 thermal resistances, capacitive connection to ground, and heat source representing Joule heat generated by the device.

Devices on the same layer lie in the *x*-y plane, while devices on different layers are stacked in the z-direction. Thus, the vertical resistances are connected between devices on different layers. Devices on bottom layers have the negative z-component resistance terminated to ground through the substrate and packaging while devices on top layers have the positive z-component resistance terminated to ground through the oxide and packaging. The network is useful because, instead of finding a linear lumped thermal heat flow equation for each device, the equations can be obtained from inspection of the 3D thermal network.

The values of the capacitive and resistive components in the thermal network are found from the architectural makeup of the chip. The thermal and geometrical properties of the materials separating devices are used in equations (3.13) and (3.14) to calculate these components. The value of the heat source that represents the heat generated by the device is obtained from simulations of the operation of the device under given bias conditions.

For this work, steady state is considered so capacitance can be ignored and the KCL type thermal equation for a node (i,j,k) is given in equation (3.15).

$$\frac{\overline{T}_{i,j,k}-\overline{T}_{i\pm 1,j,k}}{R_{i\pm\frac{1}{2},j,k}}+\frac{\overline{T}_{i,j,k}-\overline{T}_{i,j\pm 1,k}}{R_{i,j\pm\frac{1}{2},k}}+\frac{\overline{T}_{i,j,k}-\overline{T}_{i,j,k\pm 1}}{R_{i,j,k\pm\frac{1}{2}}}=H_{i,j,k}(T_{i,j,k}) \qquad (3.15)$$

For *n* devices on the 3D chip, there are $n=n_x\cdot n_y\cdot n_z$ equally spaced devices nodes in the network, where n_x, n_y, and n_z are the number of nodes in the *x*, *y*, and *z* directions. If a device node is denoted *i* and the *n* device nodes are numbered $i=0$ to $i=n-1$, then devices on the first layer are numbered $i=0$ to $i=n_x\cdot n_y-1$, second layer devices are $i=n_x\cdot n_y$ to $i=2\cdot n_x\cdot n_y-1$, and so forth. This allows the nodal equation for the i^{th} device to take the more familiar and useful form shown in equation (3.16).

$$\frac{\overline{T}_{i}-\overline{T}_{i\pm 1}}{R_{\pm x}}+\frac{\overline{T}_{i}-\overline{T}_{i\pm n_x}}{R_{\pm y}}+\frac{\overline{T}_{i}-\overline{T}_{i\pm n_x\cdot n_y}}{R_{\pm y}}=H_i(T_i) \qquad (3.16)$$

To calculate node temperature, write an equation of this form for each device node in the thermal network. Note the equations are coupled due to the thermal resistances connecting the nodes. Solve simultaneously the *n* equations for transformed node temperature $\overline{T_i}$. The absolute device node temperatures are then found using equation (3.7) that relates absolute temperature to transformed temperature. For silicon, $n=-\frac{4}{3}$, and the transformation back to absolute temperature is simplified to equation (3.17) below.

$$T=T_0\left(1-\frac{\left(\overline{T}-T_0\right)}{3T_0}\right)^{-3} \qquad (3.17)$$

The *n* equations can be written in matrix form since they form a linear system of coupled equations. The matrix equation is written as $G\vec{\vec{T}} = \vec{H}$, where *G* is a conductance matrix of dimensions $n \times n$, $\vec{\vec{T}}$ is an $n \times 1$ vector of transformed temperatures for each node, and $\vec{H}$ is an $n \times 1$ vector of heat sources for each device node. For the conductance matrix and the heat vector, the entries are known values, since the conductance seen by a device is calculated from the layout of the chip, and the Joule heat generated by a device is known from device modeling. In the conductance matrix, each of the *n* rows will have *n* entries. It is evident that the off-diagonal elements, $G_{i,j}$, $i \neq j$, are minus the conductance between nodes *i* and *j* while the diagonal elements, $G_{i,j}$, $i = j$, are the total conductance at the node.

In summary, each device on a 3D chip has been represented in a thermal network with a KCL type thermal coupled equation for each device. The equation is a function of the temperature of the device, the temperature of neighboring devices, the thermal conductance between that device and its six nearest neighbors, and the Joule heat generated by the device. The conductance and Joule heating can be calculated. The unknown node temperatures are found by solving the linear system of coupled thermal equations for the *n* device nodes.

Reduced 3D Thermal Network

As stated previously, the number of devices on the 3D chip determines the number of nodes in the thermal network and the number of equations in the thermal matrix equation that must be solved. In general, a 2D chip can contain over 40 million devices. It follows that a 3D chip could contain at least 40 million devices on each layer. As the number of layers increases, the number of total devices on the chip, and thus the number of equations to be solved, can increase dramatically. Therefore, it is useful to find a reduced thermal network with fewer nodes in order to reduce the number of equations to be solved.

The concept of the reduced network is to divide groups of device nodes into blocks, each containing the same amount of devices in the block. The goal is to find a simpler representation of the block of device nodes that has the same thermal characteristics as the original block of devices, but less nodes—and less equations—to solve. For this work, the aim is to reduce a block of device nodes to 6 nodes. The 6 nodes contain thermal resistances and heat sources in such a fashion that preserves the thermal characteristics of the original block.

The process of size reduction begins with a single group of nodes. Consider a cubic subblock of n nodes of dimensions *b* by *b* by *b* so that there are $n = b^3$ nodes in the block. Introduce six new nodes, lettered A through F that are half the resistance away from each of the boundary nodes. A 2D view of the division of a subblock is shown in Figure 12.

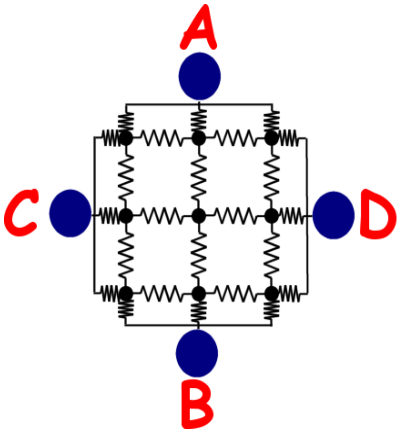


Figure 12: 2D view of a group of device nodes grouped into a single subblock. Additional lettered nodes are introduced half the resistance away from the boundary nodes. The 3D picture would show the 27 device nodes and additional nodes E and F below and above, respectively.

The goal of the reduction is to find a thermal representation of the interior nodes using only the letter nodes. Thus, the letter nodes will have thermal resistances and Joule heat sources connected in some fashion such that the original thermal properties of the interior nodes are the same as the thermal properties for the letter nodes. As such, the next step to reducing the size of the system is to find the thermal resistances and Joule heat sources for the letter nodes.

To find the thermal connections between the lettered nodes, write the KCL-type thermal equations for the 6 lettered nodes and the n interior device nodes. Write the resulting coupled linear system in matrix form. The resulting *G* matrix is $(n+6)\times(n+6)$ and the $\overrightarrow{\overrightarrow{T}}$ and $\overrightarrow{H}$ vectors are $(n+6)\times 1$, as shown in equation (3.18).

$$\begin{pmatrix} G_{A,A} & \cdots & G_{A,F} & G_{A,1} & \cdots & G_{A,n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ G_{F,A} & \cdots & G_{F,F} & G_{F,1} & \cdots & G_{F,n} \\ G_{1,A} & \cdots & G_{1,F} & G_{1,1} & \cdots & G_{1,n} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ G_{n,A} & \cdots & G_{n,F} & G_{n,1} & \cdots & G_{n,n} \end{pmatrix} \begin{pmatrix} \overline{T_A} \\ \vdots \\ \overline{T_F} \\ \overline{T_1} \\ \vdots \\ \overline{T_n} \end{pmatrix} = \begin{pmatrix} H_A \\ \vdots \\ H_F \\ H_1 \\ \vdots \\ H_n \end{pmatrix} \qquad (3.18)$$

Since what is desired is the conductance between the lettered nodes only, what is necessary is to somehow find the matrix equation for the letter node system in terms of what is known about the interior nodes and their connection to the letter nodes. To do this, solve the coupled equations for the letter node temperatures.

Start by pre-multiplying by the inverse of the *G* matrix to realize the matrix equation, $\overrightarrow{\overrightarrow{T}} = G^{-1}\overrightarrow{H}$. The inverse of the *G* matrix is a matrix of resistances, denoted *R*, that can be divided into 3 smaller matrices, as shown in equation (3.19). From here, the ever important equation (3.20) can be written. This equation reveals the letter node temperatures in terms of components of the resistance matrix *R*, the Joule heat sources for the letter nodes, and the joule heat sources of the interior nodes.

$$\begin{pmatrix} \overline{T_A} \\ \vdots \\ \overline{T_F} \\ \overline{T_1} \\ \vdots \\ \overline{T_n} \end{pmatrix} = \begin{pmatrix} \begin{bmatrix} R_{1,1} & \cdots & R_{1,6} \\ \vdots & \ddots & \vdots \\ R_{6,1} & \cdots & R_{6,6} \end{bmatrix} & \begin{bmatrix} R_{1,7} & \cdots & R_{1,n+6} \\ \vdots & \ddots & \vdots \\ R_{6,7} & \cdots & R_{6,n}+6 \end{bmatrix} \\ \begin{bmatrix} R_{7,1} & \cdots & R_{7,6} & R_{7,7} & \cdots & R_{7,n+6} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ R_{n+6,1} & \cdots & R_{n+6,6} & R_{n+6,7} & \cdots & R_{n+6,n+6} \end{bmatrix} \end{pmatrix} \begin{pmatrix} H_A \\ \vdots \\ H_F \\ H_1 \\ \vdots \\ H_F \end{pmatrix} \tag{3.19}$$

$$\underbrace{\begin{bmatrix} \overline{T_A} \\ \vdots \\ \overline{T_F} \end{bmatrix}}_{\mathrm{T_L}} = \underbrace{\begin{bmatrix} R_{1,1} & \cdots & R_{1,6} \\ \vdots & \ddots & \vdots \\ R_{6,1} & \cdots & R_{6,6} \end{bmatrix}}_{\mathrm{R_L}} \underbrace{\begin{bmatrix} H_A \\ \vdots \\ H_F \end{bmatrix}}_{\mathrm{H_L}} + \overbrace{\underbrace{\begin{bmatrix} R_{1,7} & \cdots & R_{1,n+6} \\ \vdots & \ddots & \vdots \\ R_{6,7} & \cdots & R_{6,n+6} \end{bmatrix}}_{\mathrm{R_{in}}} \begin{bmatrix} H_1 \\ \vdots \\ H_n \end{bmatrix}}^{\mathrm{T_{in}}} \tag{3.20}$$

Equation (3.20) is important because it gives the letter node temperatures in terms of the known components of the inverse of the original conductance matrix for the whole system and the known Joule heat source values for the interior nodes in the subblock. The only unknown terms are the Joule heating components for the letter nodes.

Since temperature is analogous to voltage and heat generated is analogous to current, equation (3.20) is the same as the KVL equation that can be written for a thevenin equivalent voltage V_s and thevenin equivalent resistance R_s loaded by an input resistance R_{in} (see Figure 13). Because of the correlation between the circuit in Figure 13 and equation (3.20), some generalizations can be made:

- The letter resistance matrix R_s is the thevenin equivalent resistance matrix for the letter nodes, just as R_s is the thevenin equivalent resistance in the electrical circuit.
- The second matrix is like an input resistance matrix. Its product with the matrix of heat generated for each interior device node is the temperature of the interior nodes like V_{in} is the voltage across the input resistance R_{in} .

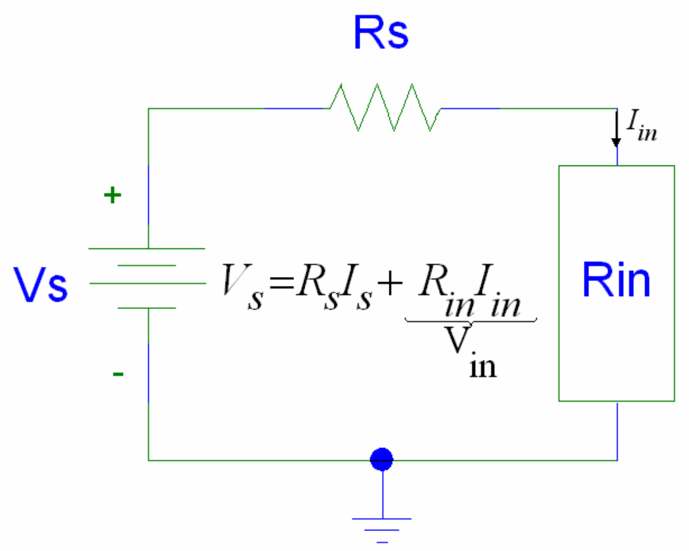


Figure 13: Thevenin equivalent circuit with load resistance *R*$_{in}$

The derivation is completed by considering the inverse of equation (3.20), $R_L^{-1}\overrightarrow{\overline{T_L}} = \overrightarrow{H_L} + R_L^{-1}R_{in}\overrightarrow{H}$. The matrix form is rewritten and shown in equation (3.21).

$$\underbrace{\begin{bmatrix} H_A \\ \vdots \\ H_F \end{bmatrix}}_{\mathrm{H_L}} = \underbrace{\begin{bmatrix} G_{A,A} & \cdots & G_{F,F} \\ \vdots & \ddots & \vdots \\ G_{F,A} & \cdots & G_{F,F} \end{bmatrix}}_{\mathrm{R_L}^{-1}} \underbrace{\begin{bmatrix} \overline{T_A} \\ \vdots \\ \overline{T_F} \end{bmatrix}}_{\mathrm{T_L}} - \underbrace{\underbrace{\begin{bmatrix} G_{A,A} & \cdots & G_{F,F} \\ \vdots & \ddots & \vdots \\ G_{F,A} & \cdots & G_{F,F} \end{bmatrix}}_{\mathrm{R_L}^{-1}} \underbrace{\begin{bmatrix} R_{1,5} & \cdots & R_{1,n} \\ \vdots & \ddots & \vdots \\ R_{6,5} & \cdots & R_{6,n} \end{bmatrix} \begin{bmatrix} H_1 \\ \vdots \\ H_n \end{bmatrix}}_{\mathrm{T_{in}}}}_{\mathrm{H_{in}}} \qquad (3.21)$$

The significance of equation (3.21) is evident. There is a conductance between each of the letter nodes and a heat generation source feeding each letter node. The connections form the octahedron shown in Figure 14. The conductance between each node is contained in the inverse letter resistance matrix R_L^{-1} and the heat source feeding each letter node is $\overrightarrow{H_{in}} = R_L^{-1} R_{in} \overrightarrow{H} = R_L^{-1} T_{in}$.

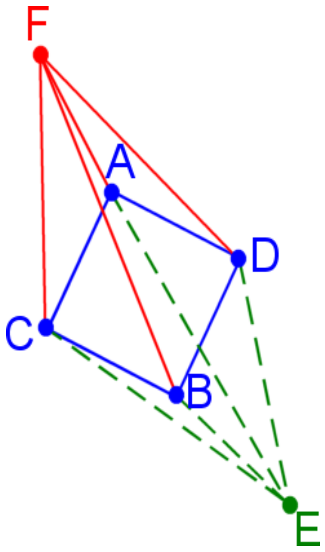


Figure 14: Octahedron formed by the thermal connection of the 6 letter nodes. Resistances are present between a given node and each of the other 5 nodes. For clarity, the resistance between nodes AB, CD, and EF and the Joule heat source that feeds each letter node is not shown.

The complete reduction of the full thermal network for all the nodes on chip follows logically from the reduction of a single block. All the device nodes in the original network are each grouped into identical subblocks. The subblocks are then replaced by the octahedron configuration characteristic of the sublock. The result is a network of octahedrons where the node temperature for each node on the octahedron is found by writing the KCL type thermal equations and solving the new system. An example of the reduction of a 3D chip with 108 nodes across 3 layers grouped into 4 cubic subblocks is shown in Figure 15. The subsequent reduction to a reduced network of 4 octahedrons is shown in Figure 16.

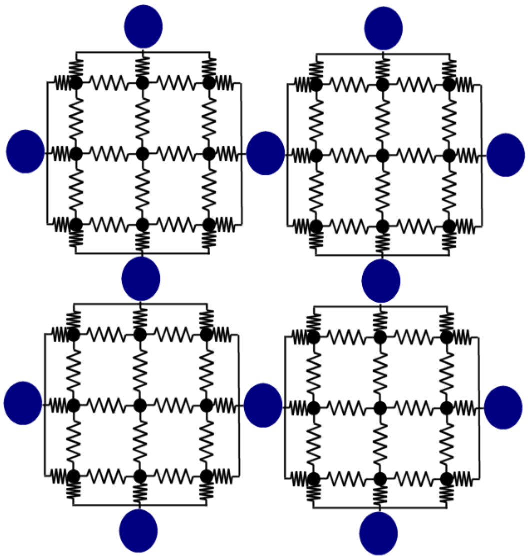


Figure 15: 2D view of 108 devices grouped into 4 subblocks of 9 by 9 by 9. The 3D view shows all 3 layers of device nodes and 4 additional nodes above and below the subblocks.

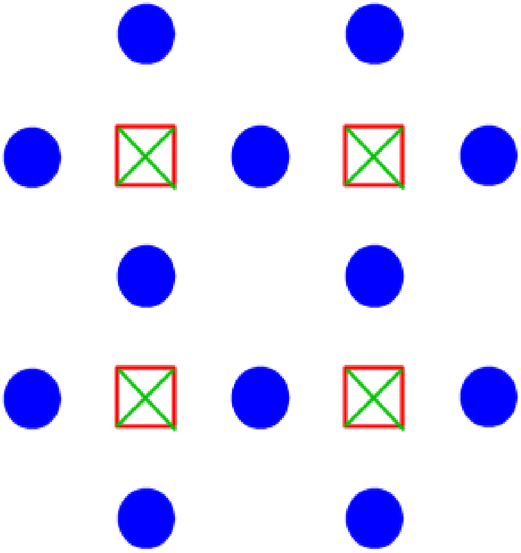


Figure 16: Reduced network of octahedrons that represent 108 device nodes grouped into 4 subblocks of 9 by 9 by 9. The blue nodes represent the 4 nodes in the plane of the page. The red squares are the nodes in the –z direction and the green Xs are the nodes in the +z direction. Each octahedron has thermal resistances and Joule heat sources, not shown for clarity.

Looking at the 2 figures, the benefit of the size reduction is evident. The original network contained 108 nodes, resulting in the necessity to solve 108 coupled thermal equations for 108 node temperatures. The reduced network contains only 20 nodes, thus there are only 20 coupled equations to write and 20 node temperatures to solve for by solving the linear system of equations.

In short, the importance of finding a reduced network is that a block of devices is reduced to 6 nodes. The 6 nodes contain thermal resistance connections that form an octahedron. The octahedron has the same thermal characteristics as the original block of devices and the original thermal network is represented by the reduced network of octahedrons. Each subblock of devices on the chip is replaced by such an octahedron and the resulting reduced thermal network can then be solved. Therefore, instead of solving an exhaustive number of equations for the original network, the number of equations has been reduced to 6 equations for each octahedral block in the reduced network. The steps for finding the thermal conductance and Joule heat source for each node in the reduced network are summarized as follows.

1. Decide on the size of the cubic subblock. For this work, subblocks of 22 by 22 by 22 were chosen.
2. Introduce 6 new nodes half the resistance away from the boundary nodes.
3. Find the conductance matrix *G* that emerges from writing the KCL-type thermal equations for the 6 letter nodes and the interior nodes.
4. Find $R = G^{-1}$.
5. Divide *R* into 2 smaller matrices:
    a. R_L is rows 1 through 6 and columns 1 through 6 of *R*
    b. R_{in} is rows 1 through 6 and columns 7 through $n+6$ of *R*

6. Find the conductance matrix for the letter nodes by computing

   $G_L = R_L^{-1}$

7. Find the Joule heat source vector for the letter nodes by computing

   $H_{in} = G_L R_{in}$ (let the interior nodes have unity Joule heating—the significance of which will be explained later).

Thermal Resistance in 3D Thermal Network

Determining the value of the thermal resistances in the thermal network is a crucial component for calculating temperature on chip. The thermal resistances couple one device with its neighbors and, therefore, hold a lot of information crucial to the thermal network. As such, in order to develop an accurate thermal model of the chip to use for simulation, accurate resistance calculations are a necessity.

Recalling equation (3.14) for thermal resistance, it is evident that there are two components that must be known in order to calculate thermal resistance: material conductivity and dimensions. Heat is generated internally within each device on chip and travels across chip to neighboring devices. The thermal conductivity at room temperature is needed for each of the materials that the heat will travel through, since each of those materials will offer a different resistance to heat flow. Which materials the heat travels through depends on the fabrication process and the direction of heat flow being considered. The distance the heat flows (thickness) and the area it flows through depends on the fabrication process, design structure, dimensions of the device, and layout of the devices on the chip.

Materials

The design structure of a 3D chip along with knowledge of basic fabrication processes allows determination of the materials used for 3D fabrication. These materials include silicon, silicon dioxide for insulation, copper or aluminum for metal lines, and polysilicon for interconnect lines and the gate electrode of a transistor. Heat spreading in the vertical direction from a device on one layer to one above or below it would encounter resistance caused by the insulator separating the 2 devices, any metal layers between the device layers, the polysilicon gate and gate oxide within a device, and the silicon island in which the neighboring device or block is fabricated. Heat spreading in the plane on a single layer would see resistance caused by the insulator and silicon islands separating the two devices. The total resistance in a given direction is the sum of the resistance offered by each material the heat passes through. This resistance, as stated before, depends in part on the thermal conductivity of the material. The thermal conductivity at room temperature for common fabrication materials is given in Table 1.

Material	**Thermal Conductivity (** $\boldsymbol{\frac{W}{m \cdot K}}$ **)**
Silicon	1.5×10^{2}
Aluminum	2.1×10^{2}
Copper	4.0×10^{2}
Polycrystalline Silicon	1.25×10^{2}
Silicon Dioxide SiO_2	1.04

Table 1: Thermal Conductivity at room temperature for materials used in 3-D chip fabrication.

Architecture and Dimensions

In the heat flow theory section it was shown that thermal resistance is measured from the midpoint of one device to the midpoint of a neighboring device. The thickness Δl is the distance from the midpoint of the one device node to the midpoint of the neighbor and the area *S* is the area the heat flows through. To know thickness and area dimensions, it is necessary to know the spacing between devices on the chip for all three dimensions. Since devices are made of different materials, each offering a different thermal conductivity and thus a different resistance, it is further necessary to know the dimensions of the different materials used to make the devices as well as those used to separate devices.

For this model, the chip block layer structure shown in Figure 1 is adopted. In this case, devices on each layer are interconnected to form blocks of circuits. The 3D chip is then constructed by vertically interconnecting the different blocks. For this work, the assumption is made that devices are equally distributed on a given layer and comprise a planar area of $2\mu\text{m} \times 2\mu\text{m}$. The midpoint of the device is the middle of the channel region. The spacing in the vertical direction is given by the thickness of the materials separating the devices. Dimensions for the silicon and insulating layer thicknesses were taken from research reports on 3D fabrication detailed in chapter 2. Fabrication specific parameters such as metal, polysilicon, and gate oxide thickness were taken from data given for a generic 2 μm fabrication process. The compiled layer thicknesses are provided in Table 2.

Layer	**Thickness (μm)**
Bulk Silicon Substrate	300
Insulator (SiO_2)	0.5
Silicon Island	80
Metal	0.6
Polysilicon	0.3
Gate Oxide	7.5×10^{-3}

Table 2: Material thicknesses for a 3D chip.

The computation of thermal resistance is simple. The resistance from one node to one of its 6 nearest neighbors is the sum of the resistances of each material heat passes through. Depending on which of the 6 resistances is being computed, the thicknesses and areas are determined by either the spacing between devices or the thickness of the material under consideration. For example, consider the vertical resistance from node (i, j, k) to node ($i, j, k+1$), denoted R_{z^+} since it is the resistance from node (i, j, k) to the neighbor on the layer directly above (positive z-direction). According to the structure adopted for this work, the materials separating the midpoints of the two devices are the gate oxide, the polysilicon gate electrode, an insulating oxide, metal (minimum of one layer), and the silicon that forms the island for device ($i, j, k+1$). The total resistance R_{z^+} is given by the sum of the resistances of each of the layers since the layers are in series. The resistance of each layer is computed using equation (3.14) where the area is $A = 2\mu m \times 2\mu m = 4\mu m^2$, the thickness Δl is given by the thickness of each of the layers as defined in Table 2, and thermal conductivity of each layer is given by the appropriate value from Table 1. For this work, the resistances between device nodes on the same layer in both the x- and y-directions are found to be $R_x = R_y = 50\frac{K}{W}$. The resistance between device nodes on separate layers is found to be $R_z = 5\times10^5 \frac{K}{W}$.

Chip Packaging

The packaging for the chip serves the important role of assisting in the dissipation of heat from the chip to the ambient. Good packages facilitate rapid thermal dissipation to hinder the increase of chip temperature due to heat building up on chip. To accurately model a 3D chip using a thermal network, the characteristics of chip packages need to be considered so that heat dissipation is adequately modeled and accurate temperature calculations can be made.

General packaging technologies resemble the package shown in Figure 17. The die is attached to a pad using an adhesive. The entire fixture is encased using a molding compound and wires are connected from the pins on the die to lead frames protruding out of the package for off-chip connection. More sophisticated packaging schemes are adopted for computer processors such as a Pentium processor. These chips require better packages and often employ an integrated heat spreader and heat sink to increase thermal conductivity, thereby reducing the package resistance. The components are arranged using flip chip technology so that the backside of the die is exposed. The heat spreader is integrated into the processor package and the heat sink is attached directly to the die of the processor during manufacturing. Since the heat sink makes a good thermal contact with the die and also offers a larger surface area for better heat dissipation, it can greatly increase thermal conductivity and facilitate cooling [37]. The Intel flip chip package with integrated heat spreader and heat sink is shown in Figure 18.

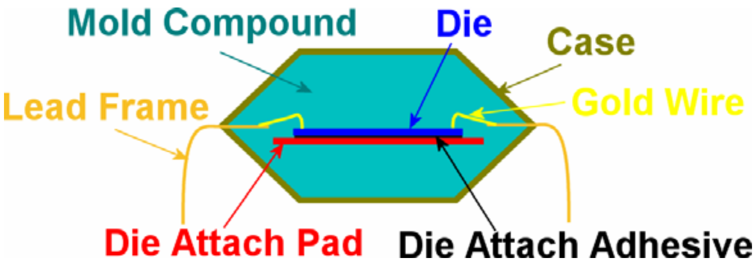


Figure 17: Components of a basic packaging technology.

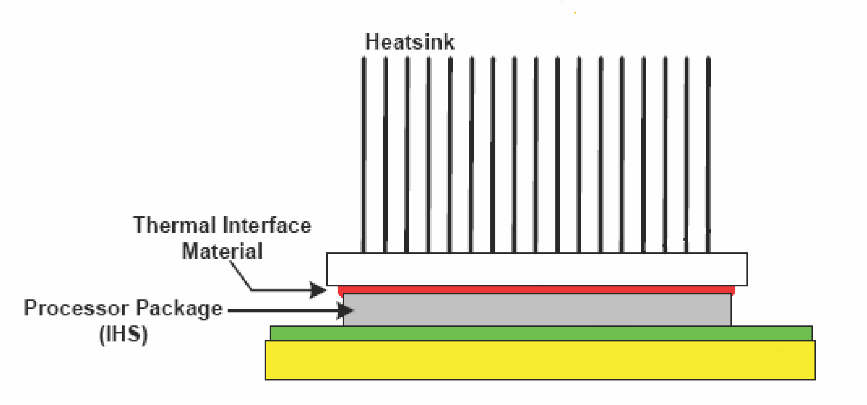


Figure 18: Intel Flip Chip package (FCPGA2). The backside of the die is left exposed inside the processor package, which includes an integrated heat spreader. A heat sink is attached to the exposed die for better heat dissipation.

Common components used in packages include alumina for the casing, silica filler for the molding compound, and silver filled glass or silver filled epoxy for the die attach adhesive. The dimensions and thermal conductivity for each of these materials should be accounted for when computing package thermal resistance. The thermal conductivity of different materials used for packing is shown in Table 3.

Package thermal resistance is important because some device nodes on the chip are at the edges of the chip. As such, they do not have all 6 thermal resistances, but instead are bounded by the packaging. To accurately model chip temperature, when forming the thermal network, the package thermal resistance should be accounted for when writing the KCL-type thermal node equations for devices nodes at the edges of the chip. For this work, package resistance is $R_p = 2.5\times 10^5 \frac{K}{W}$.

Material	**Use**	**Thermal Conductivity** $(\frac{W}{m\cdot K})$
Alumina	Casing	18
Copper Alloy	Lead Frames	160
Silver Filled Glass	Die Attach Material	270
Silver Filled Adhesive	Die Attach Material	2.5
Silver Filled Epoxy	Die Attach Material	1.6
Silica Filler	Molding Compound	0.6

Table 3: Thermal conductivity of common materials used to package integrated circuits [38].

Full Chip Temperature Calculation

Thus far, it has been shown that the thermal characteristics of an integrated circuit can be modeled using a reduced thermal network of device nodes evenly spaced across chip. The nodes each have thermal resistances connected to neighboring device nodes, and thermal capacitances and heat sources connected to ground. KCL-type thermal nodal equations can be written for each node, where temperature is analogous to electrical voltage and heat is analogous to current. Temperature at each node is found by solving simultaneously the coupled KCL-type thermal equations.

It is desirable to account for the specific heat generation characteristics for each device on the chip since each device on chip doesn’t behave the same. Some devices will switch more frequently then others, thereby generating more heat. These differences must be accounted for to accurately calculate temperature for a node in the network. This necessitates convergence of the solution of the linear matrix equation for the full chip with the individual device level equations for each device.

The methodology presented in this chapter employs 2 separate solvers to calculate chip temperature as a function of position. The device solver finds solutions to the quantum and semiconductor equations for a representative device to obtain characteristics of the device operating under specified bias conditions. The characteristic of importance for this work is the Joule heating for the device. The network solver finds the solution to the linear matrix equation of KCL-type thermal equations that account for the thermal connections of devices on chip. It takes into account the specific device activity level for each device on the chip under consideration. The complete solution involves iterative convergence of both the device and chip level solvers, thereby establishing the desired link between device and full chip heating.

The basic approach to finding full chip temperature is as follows:

1. Solve the device equations for Joule heating for a representative device on chip. The representative device is the one at the mean temperature of all devices on chip.
2. Solve the thermal nodal equations for the full chip. The Joule heating (heat source) for each node is obtained by mapping the Joule heating for the representative device to each device on the chip using a Monte Carlo application that incorporates activity level for each device.
3. Find the new representative mean temperature.
4. Repeat steps 1-3 until mean temperature is constant between iterations.

To implement the process described above, the chip under consideration must be analyzed for the activity level for each device on the chip. This activity level is then incorporated into the heat source for each node in the thermal network. The following subsections outline the setup of the chip for which temperature is to be calculated, the methodology for obtaining the device activity levels, the actual operation of the device and thermal network solvers, and finally the complete algorithm for obtaining full chip temperature at the resolution of a single device on chip.

3D Chip Configuration and Activity Profile

The 3D chip under consideration in this thesis is a 5-layer stack of Pentium III chips. The Pentium III contains over 40 million 0.18 μm devices so that the entire stack contains over 200 million devices. The chip is divided into functional blocks such as clock, cache, fetch, etc. so that the power consumed in each block of the chip can be used to figure device activity for each device in the block. The Pentium III chip divided into functional blocks is shown in Figure 19.

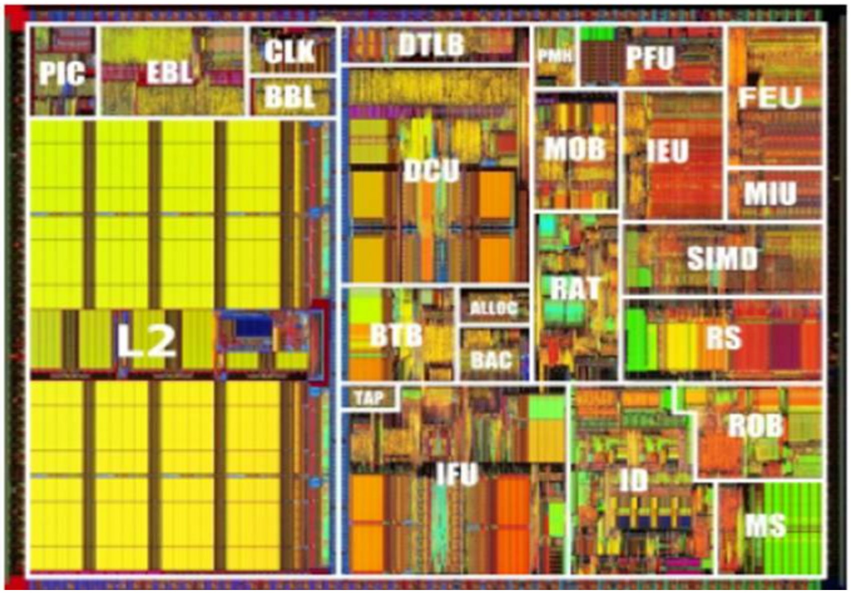


Figure 19: Pentium III chip divided into functional blocks. The chip is stacked to form a 5-layer 3D chip for which local temperature is to be calculated.

The device nodes in the thermal network are assumed to be equally spaced. Thus, some device nodes may lie in the same functional block while others will lie in a different functional block. Different functional blocks serve different functions and may operate more or less frequently, generating more or less heat. This results in different Joule heating for different devices on the chip. For this work, one focus is on finding a way to accurately model the differences in device activity for different blocks without solving the device equations for Joule heating for each device on each layer for the 3D chip. Instead, it is more favorable and computationally efficient to solve for Joule heating for one device and use this value to statistically generate Joule heating for other devices. To accomplish this, power per unit area for each functional block is used to generate probabilities for finding devices on or off within each functional block. Joule heating for the representative device is then statistically scaled for each node in the network based on this probability.

The first step to determining an activity level profile is to simplify the functional blocks. The numerous blocks in Figure 19 are combined to form the 10 functional units shown in Figure 20. The locations of the boundaries are also shown. The ten units are: bus interface unit, clock, L1 cache, L2 cache, fetch, memory order buffer, register alias table, decode unit, execute unit, and issue logic.

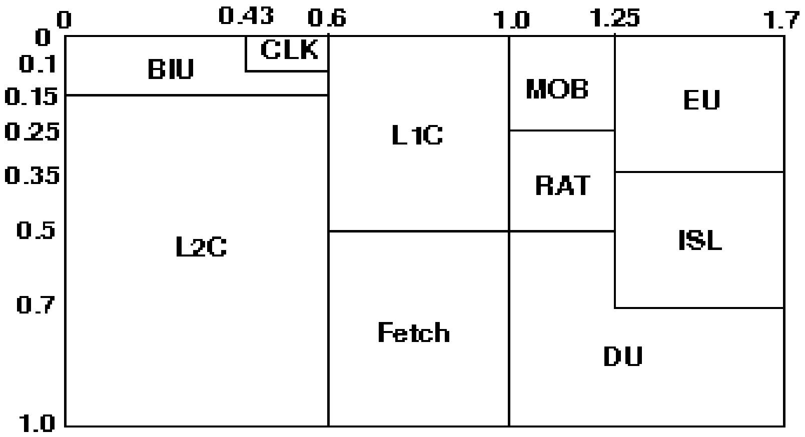


Figure 20: Pentium III simplified functional units. Various units are combined according to application to reduce the number of units.

A measure of the probability of finding an active device is needed in order to find the Joule heating for a given device. To find this, the area and percentage of total power consumed for each block are needed. The area of each block is obtained by measurement while the percentage of power consumed (relative to total chip power consumed) is calculated using data provided in [39]-[40]. Since various blocks are grouped, the representative percentage of total power consumed in each block is a collection of the values for the original blocks that were combined. Power per unit area is calculated for each functional unit by taking the ratio of the percentage of power consumed in each block to the area of each block. The value is normalized by that for the clock in order to give the percentage of power consumed per unit per area relative to the clock. Since the clock is always active, this normalized power per unit area represents the likelihood of finding an active device in the given functional unit relative to devices in the clock that are always active. The area, power percentage, and normalized power per unit area for the each of the ten functional units is provided in Table 4.

Functional Unit	Area	Percentage of Chip Power	Normalized Percentage Power Per Unit Area
Bus Interface Unit (BUI)	4.3	5.9	0.27
L2 Data Cache (L2C)	29.8	8.8	0.05
Fetch	12.5	16.9	0.26
Clock (CLK)	1.0	5.2	1.0
L1 Data Cache (L1C)	12.5	9.8	0.15
Memory Order Buffer (MOB)	3.3	4.7	0.28
Register Alias Table (RAT)	3.3	4.7	0.28
Execution Unit (EU)	9.5	13.0	0.26
Issue Logic (ISL)	9.5	14.1	0.29
Decode Unit (DU)	14.6	17.2	0.23

Table 4: Activity profile of devices in various functional units.

Now that the profile for the device activity on the full chip has been obtained, a map for statistically determining the Joule heating for each node in the thermal network can be established. Given the Joule heating of the representative device, the Joule heating for each device on chip can be found by weighting the representative Joule heating by a factor that represents the activity level of the device under consideration. The weighted Joule heating can then be used as the heat source in the thermal network for the device under consideration. Thus, a method for transforming the normalized power per unit area to a weighting coefficient for each device in the unit is needed.

To map the representative Joule heating to the full chip network, a Monte Carlo statistical approach is taken. Consider a single functional unit. The normalized power per unit area value is the probability that a device in the functional unit is active. However, every device in the entire unit may not always be in the on-state or have been in the on-state for as long as another device. Thus, the heat generated by each device individually may be different. What is desired is a unique weighting factor for each device in the unit that can be used to scale the representative Joule heating. The approach is to use a probability density function. Let the normalized power per unit area be the probability of a device being in the on-state within the functional unit. The probability of the device being in the off-state is then the complement of the normalized power per unit area. Since there are only 2 states, let the density function have a domain [0, 1], where random variables in the range [0, 0.5] represent the off-state and those in the range [0.5, 1] represent the on-state. A representative density function for the fetch unit is shown in Figure 21.

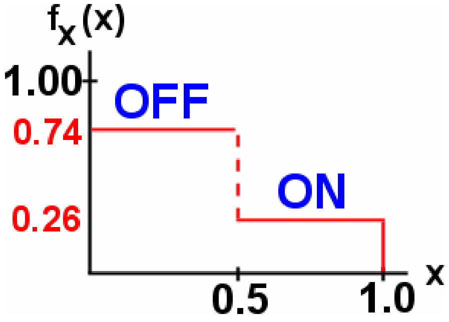


Figure 21: Probability density function for a device in fetch unit. The device state (on or off) is represented by a number in [0, 1] and the probability of being in the state is given by the normalized power per unit area.

The weighting coefficients for each device are found by transforming the density function for the functional unit to a uniform random variable. This way, the generation of a uniform random variable *r* will result in transformed values in the range [0, 0.5] according to the probability of a device being in the off-state. Likewise, values in the range [0.5, 1] will be transformed according to the probability of a device being in the on-state within that unit. Thus, for each node in the thermal network, given its location on the chip, a random variable *r* can be generated and transformed to a Joule heating weighting factor. For example, for the fetch unit, a generated uniform random variable *r* will be transformed to a weighting coefficient in the range [0, 0.5] (off-state) 74 out of 100 times. The coefficient will be in the range

[0.5, 1] (on-state) 26 out of 100 times. Thus, the coefficient represents the “activity” of the device numerically and can be used to scale the representative Joule heating, obtained from the device solver, to a measure of Joule heating for the device under consideration.

Device Performance Model

The device performance model used in this work is the same one used in [35]. The model solves quantum and device equations such as the Schrödinger, Poisson, electron and hole current continuity, and lattice heat flow equations. The goal is to obtain the non-isothermal device characteristics such as potential, electron and hole concentration, electric field, Fermi level, lattice temperature, and wave functions for the device.

The device model is employed to calculate the Joule heating, $H = -J \bullet \nabla \phi$, for a representative device on the chip. The inputs to the solver include lattice temperature, device dimension, bias voltage, device type, and doping profile. For this work, the representative device is chosen just as in [35] to be an n-channel MOSFET of device gate length 0.13 μm and device width of 0.4 μm. The gate-source and drain-source biases are selected to be 1.5 V. The lattice temperature of the representative device is selected to be the average temperature of all devices on chip. Since chip temperature changes as the chip is operated, this is the parameter that is iteratively updated in the full-chip temperature algorithm.

Thermal Network Solver

The thermal network solver is responsible for connecting the device level performance to full chip heating. It performs the Monte Carlo statistical map and solves the linear matrix equation $G\overset{\rightrightarrows}{T} = \overrightarrow{H}$. The output is a vector containing the temperature at each node in the network and the mean temperature of all the nodes in the network. The mean temperature can be fed back to the device simulator for further iteration while the temperature map is representative of the full-chip temperature at the simulated representative average temperature.

The solver takes as input the number of nodes in the original 3D network, the thermal resistances calculated in this chapter, and the Joule heating of a representative device in the network. Thus, the calculations performed this far are hard-coded into the solver.

After storing the resistance values, the solver performs the Monte Carlo method to statistically obtain the Joule heating, and thus heat source value, for each node in the network. After storing these values as the *H* vector in the matrix equation, the conductance matrix is constructed.

The conductance matrix is constructed as outlined in the section on the 3D thermal network in this chapter. For convenience, its construction is noted again. Each entry (i, j) in the matrix is minus the conductance that exists between nodes *i* and *j*. For the diagonal, $i = j$, the entry is given by the total conductance seen at the node under consideration. Because not all nodes in the network are interconnected thermally, the conductance matrix is sparse and need not be stored completely in memory. Instead, only the non-zero entries are remembered.

Having found conductance and heat sources for each node in the network, the linear equation can be solved. For this work, the bilateral conjugate gradient method is chosen to iteratively solve the equation for temperature at each node. This method is chosen because it is deemed to yield the fastest convergence to the solution.

Complete Simulation Algorithm

Now that the specifics of the thermal resistance, chip layout, device activity profiles, and the device and thermal solvers have been established, the complete algorithm for obtaining full chip temperature can be thoroughly explained. Although the basic algorithm was given at the start of this section, simplifications can be made to ease the simulation time and complexity of the basic algorithm. While the concept of the basic algorithm is still valid—solve for temperature by iteratively solving for device and full chip mean temperature convergence—certain simplifications can be made that preserve this convergence with less computation time and effort.

The complete algorithm takes advantage of the following fact: the matrix equation to be solved, $G\overrightarrow{\overline{T}} = \overrightarrow{H}$, is a linear equation. Thus, any scaling done to the heat component on the right hand side is equally applied to components on the left hand side. Since the conductance matrix is a matrix of constants (the values only change when the architecture of the chip changes), any scalar applied to the left hand side can be considered to affect only the entries of the temperature vector. This linearity is exploited to allow for solving of the thermal device solver only once. If the temperature response can be found for a unit heat source input, then any solution to the thermal network can be found from this unit temperature response.

The unit heat response concept is explained as follows. All heating sources in the thermal network are taken to have unity strength weighted by the device activity profile weights generated using the Monte Carlo methodology. These sources make up the unit heat input H_0. If the thermal network is then solved, the calculated node temperatures form the unit temperature response, $\overrightarrow{T_0}$. Simply stated, the node temperatures show how the thermal connections and device activity affect chip temperature for a unitary set of heat generation sources.

According to the basic algorithm, after solving the thermal network, the mean node temperature is used to find a new representative Joule heating on the device level. The thermal network is then solved again with the new Joule heating for new node temperatures. This process repeats until the old and new device temperatures converge. However, the availability of the unit temperature response eliminates the need to repeatedly solve the thermal network for each new representative Joule heating. Instead, since the individual device node activity has already been accounted for in the unit response, this new Joule heating is simply a scaling factor to the unit heat source input H_0. As such, because of linearity of the system, the new temperatures that would result from solving the network again are just the unit temperature response values scaled by this same scalar. Thus, repeatedly solving the thermal network has been replaced by solving the network only once and then scaling node temperature for each new representative Joule heating value obtained. The simulation is complete when node temperatures become constant with iteration.

The only thing left is the initial conditions to begin the algorithm. It was already stated that the initial heat sources are unit sources weighted by the device activity coefficient for each device on chip. The initial temperature for the device solver is the ambient temperature (300 K) since devices will reside at the ambient temperature while the chip is not being operated. Once the chip is turned on, the devices will begin to switch and the chip will heat up.

The complete algorithm for modeling chip temperature as a function of position is given in the list that follows and depicted in the flowchart in Figure 22.

1. Divide the chip into equally spaced device nodes.
2. Calculate thermal resistance for the chip package and the x-, y-, and z-directions.
3. Divide the chip into functional units and obtain the area, percentage power, and normalized power per unit area for each unit.
4. Calculate unit temperature response $\overrightarrow{T_0}$ for a unit heat source input H_0
5. Obtain mean node temperature from the unit temperature vector.
6. Simulate device at room temperature to obtain initial representative Joule heating H^i.
7. Compute new node temperatures by scaling the unit temperature response by the initial Joule heating: $\vec{T}^{k+1} = H^i \overrightarrow{T_0}$.
8. Calculate the new mean node temperature.
9. Simulate the representative device at the new average temperature to obtain the new representative Joule heating, H^k.
10. Compute the new node temperatures by scaling the unit response by the representative Joule heating: $\vec{T}^{k+1} = H^k \overrightarrow{T_0}$

11. Repeat steps 8-11 until the difference in T^{k+1} and T^{k} is sufficiently small.

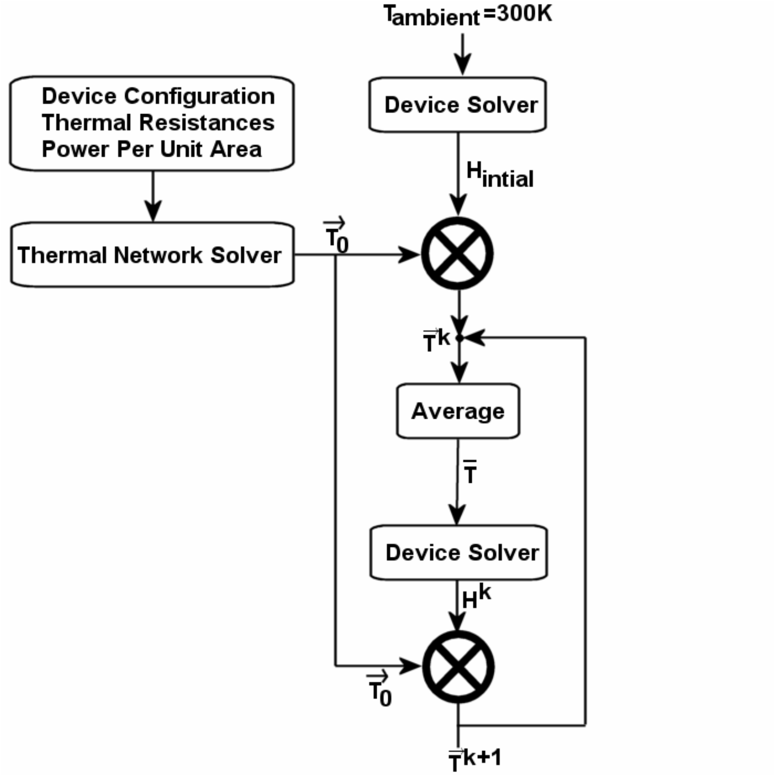


Figure 22: Flowchart for complete temperature simulation algorithm.

Results and Analysis

The results from application of the algorithm presented in the previous section are expected to be in accordance with the claims made thus far. In general, the results are expected to show that varying device activity leads to local device heating, resulting in a local increase in temperature. More specifically, the results are expected to show that hot spots can develop on chip and increasing the number of layers on chip results in higher chip temperatures. Finally, the results should offer insight into possible solutions for reducing chip temperature.

For the hot spots, the temperature at some locations is expected to be so high that it becomes detrimental to device and circuit operation. With respect to device activity, consider the activity profile shown in Table 4 for the Pentium III structure that comprises each layer of the simulated 3D chip. There is a trade off between functional unit area and power per unit area that affects temperature of devices within a given functional unit. For example, the clock unit has the largest power per unit area and thus devices in this unit have the greatest likelihood of being in the on-state generating more heat. However, the clock also has the smallest area, so there are fewer devices within this unit. The trade-off is how much will heat generated in this unit affect temperature at that location and neighboring areas since the area is small. In general, one would expect that the L2 cache will operate at lower temperatures because of the large area it comprises and the low power per unit area it exhibits. The clock unit, since it has the highest power per unit area, is expected to have the highest temperature, even though it has the smallest area. It will be interesting to see if the operation of the other units can collectively create higher temperatures then those within the clock.

Finally, for multiple layers, the increase in device and power density combined with the low thermal conductivity of insulating layers is expected cause even greater temperature gradients, and thus even higher local temperature. Upper layers are expected to operate at higher temperatures then layers below because of their isolation from the substrate and package cooling components like the heat sink.

In this section, temperature contours for the stacked 5-layer Pentium III 3D chip are presented. In addition, analysis of minimum and maximum temperature for chips with different numbers of layers is performed. Results are also presented for an alternate chip layout.

5-Layer Stacked Pentium III

The maximum, minimum, and average temperatures for each layer of the 5-layer Pentium 3D chip are shown in Figure 23. The temperature for each functional unit for each layer is given in Table 5. The temperatures are obtained from contour plots of isothermal lines on chip. The devices within the unit have temperatures in the range [T-ΔT, T+ΔT], for ΔT provided in the table. Because of the vertical symmetry, the 4th and 5th layers have the same temperature map as the first and second layers, respectively. The maximum chip temperature is 404.2 K, which is 104.2 K above the ambient. The minimum is 329.4 K, or 29.4 K above the ambient. The highest temperatures on each layer occur in the clock and issue logic units while the lowest occur in the L2 cache.

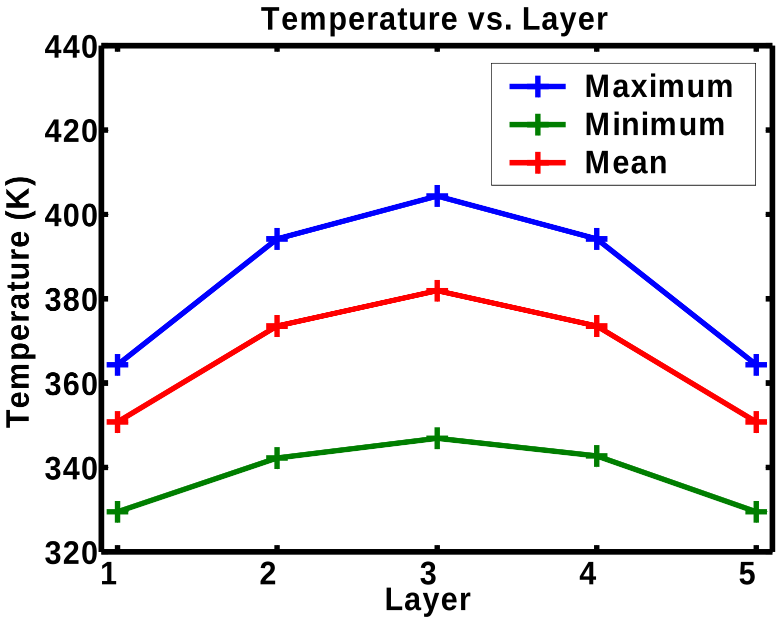


Figure 23: Minimum, maximum, and mean layer temperature for 5-stack Pentium III 3D chip. The maximum temperatures occur in the clock and issue logic units while the minimum temperatures occur in the L2 cache.

Functional Unit	Layer 1 Temperature	ΔT	Layer 2 Temperature	ΔT	Layer 3 Temperature	ΔT
BIU	355	10.5	380	15.6	389	17.1
L2C	336	3.5	352	5.2	357	5.7
Fetch	355	3.5	380	5.2	394	5.7
Clock	361	3.5	389	5.2	399	5.7
L1C	342	7.0	361	10.4	368	11.4
MOB	355	3.5	385	5.2	394	5.7
RAT	355	3.5	385	5.2	394	5.7
EU	358	7.0	385	10.4	394	11.4
ISL	361	3.5	389	5.2	399	5.7
DU	352	3.5	375	5.2	389	5.7

Table 5: Temperature for each unit on each layer of the 5-layer stacked 3D chip. Devices within the unit operate at temperatures in the range [T-ΔT, T+ΔT]. All units are in Kelvins.

The temperature results for the three layers yield expected results. First, the presence of temperature gradients is evident. Because of the different device activity in the different functional units, temperature is not constant across each layer. Instead, it is varying in accordance to the device activity within a functional unit. Second, the affect of multiple layers is evident. The increase in the number of layers, and thus devices, causes more heat to be generated; the low conductivity of the insulating layers results in high resistance to heat dissipation for upper layers. Thus, temperatures on upper layers are higher.

From the temperature table it is also evident that the L2 cache is the coolest region. This makes sense because the goal of cache systems is to decrease access to off chip memory. The L1 cache is the primary cache and is used most often. An efficient pipeline will limit access to the L2 cache, resulting in an even smaller access to off-chip memory. Since the L2 cache is accessed the least (devices in the unit are on the least), despite its area, it remains the coolest region on chip.

The clock and issue units operate at the highest temperature for each layer. This suggests that there is a trade-off between heat from neighboring devices, activity within a unit, and area of the unit. The clock has the highest power per unit area, but it occupies a smaller area and is placed near a significantly cooler L2 cache. On the other hand, the issue unit has the second highest power per unit area, an area larger then the clock, and is placed near other units that are not as cool as the L2 cache. The clock's placement next to the L2 cache might help keep temperature down within the clock. The presence of the issue unit near other units that are also heating up significantly must be enough to result in temperatures comparable to the clock even though device activity within the unit is lower then for the clock. Thus, a layout suggestion is to find a balance between interconnect length and unit placement so that hotter units can be placed closer to cooler units to help stabilize temperature increases. Concentrating units with medium heat generation as compared to the cache and clock might allow for temperatures just as hot as the clock even though device activity is less.

Finally, from [41], while the maximum recommended operating temperature differs for different types of Pentium III processors, the biggest recommended operating temperature is 90 °C (363 K). From the temperature table, it is obvious that while only devices inside the clock on layer 1 begin to reach this limit, devices on the second layer outside of the L2 cache have eclipsed this value. Furthermore, all of the devices on the third layer are operating beyond the recommended maximum temperature. This is very problematic as device behavior becomes unpredictable at these temperatures, resulting in potential timing errors and device or complete chip failure.

Minimum and Maximum Temperature

It is also of interest to observe local heating for different numbers of layers to show that as the number of layers increases, so does chip temperature. Figure 24 shows the minimum and maximum chip temperatures for 1, 2, 3, 4, and 5-layer stacked Pentium III 3D chips.

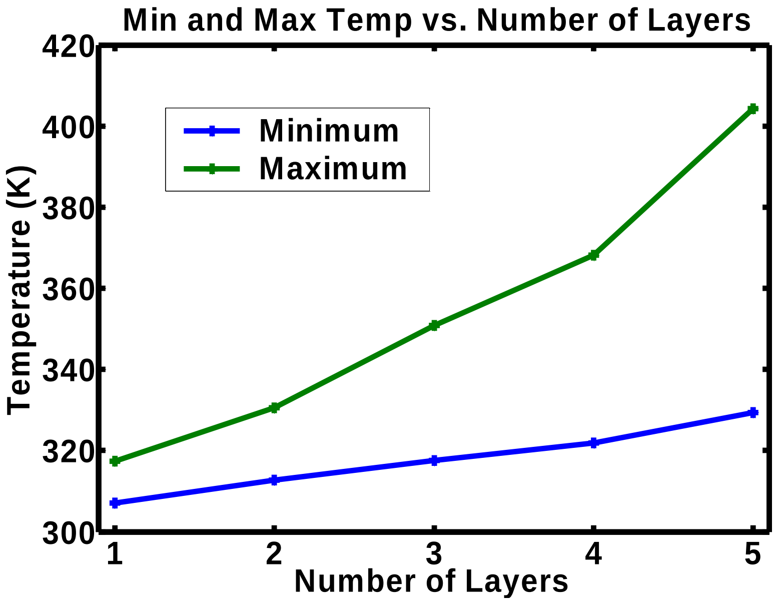


Figure 24: Minimum and maximum chip temperature versus number of stacked Pentium III layers on a 3D chip.

The figure shows that as the number of layers increase, the minimum and maximum hot spot temperatures increase. For both minimum and maximum, the increase starts out as a linear increase. However, for the 5-layer stack, the trend is broken. The 4- and 5-layer stacks both reach maximum temperature exceeding the acceptable maximum as defined by Intel. The graph basically shows that for 3D chips (more then one layer), hot-spots temperatures are increasingly more important. For a planar circuit (1-layer), minimum and maximum temperatures do rise above the ambient, but for a typical activity profile, temperatures do not reach an unacceptable limit. However, as the number of layers increases, even typical operation yields unacceptably high maximum temperatures. Temperature increases are potentially worse for high performance, high activity chips and chips that are being overclocked.

Alternate Chip Layout

After observing the existence of a trade-off between functional unit location, area, and heat generated within neighboring units, an alternate chip layout was simulated in an effort to see if an alternate layout could help reduce temperature on the 3D Pentium III 5-layer stack. Instead of stacking all 5 layers repeatedly, for this simulation, the 4th and 5th layers were each rotated 180 degrees before being stacked. The aim is to see if rearranging functional unit location helps reduce device temperature. The restructuring pits the cooler L2 cache on layers 4 and 5 near the hotter issue unit on layer 3. Similarly, the hotter clock unit on layer 3 is positioned near the cooler decode units on layers 4 and 5. The arrangement should help to provide a path for heat flow from the hotter units to the cooler ones to result in lower temperatures.

The results of the simulation are promising. For each layer, device temperatures decreased within each functional unit except for the L2 cache, which experienced a marginal increase. This result means that by just rearranging the units to put cooler units closer to hotter units has resulted in overall lower device temperature for all the layers of the 3D chip.

A plot of the maximum, minimum, and mean temperatures for each layer of the alternate 5-layer stack is shown in Figure 25. Comparing to Figure 23, it is clear that the maximum and mean temperatures for all 5 layers of the alternate layout decreased while the minimum temperature increased only slightly. The average decrease in maximum temperature is 11.0 K per layer and the average decrease in mean device temperature is 10.7 K per layer. On the other hand, the average increase per layer in minimum temperature is only 3.3 K. Thus, using the alternate layout has resulted in a decrease in mean and maximum temperature on each layer with only a small increase in the minimum temperature on each layer.

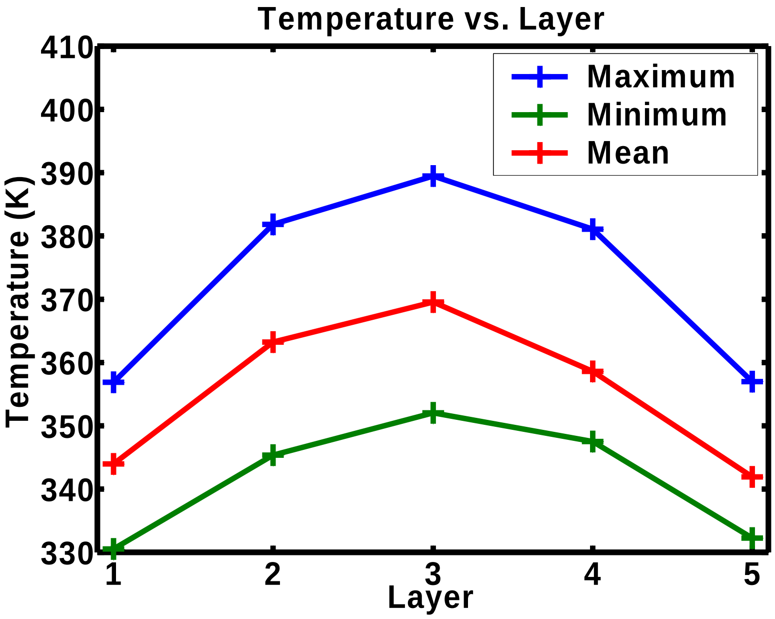


Figure 25: Minimum, maximum, and mean layer temperatures for alternate 3D chip layout in which layers 4 and 5 are rotated 180 degrees from the original 5-layer chip.

The simulation suggests another potential solution to the hot spot problem. In addition to placing cooler units near hotter units on the same layer as suggested previously in this work, another solution is to do use the alternate placement approach but on different layers. The L2 cache is a cool unit that covers a lot of chip area. Its area is comparable to the area of the memory order buffer, register alias table, execute, and issue logic units. These were the units that contributed to the heating of the issue logic to temperatures comparable to the clock. By positioning the L2 cache on layers 4 and 5 vertically above these units, the result is that these areas are now much cooler with minimum increase in L2 cache temperature. The implication is that layout is an important consideration for maintaining acceptable chip temperatures. A simulation such as the one performed in this section could be performed prior to chip fabrication and used to help decide on the best temperature aware layout prior to chip fabrication.

Chapter 4: Temperature Experiments

Much has been stated in this thesis about the detrimental effects of high temperatures on device and circuit chip behavior. It has been stated and shown through simulation that temperature is not constant across a chip, but instead, it varies according to the level of activity of the devices on the chip. The purpose of this chapter is to present results from experiments that show that temperature differences can affect circuit behavior and that different device activity does cause temperature gradients to exist on chip.

The first section focuses on the effects of increased temperature on frequency of oscillation in a ring oscillator circuit. Ring oscillators operating under different temperature conditions are simulated and frequency is observed. The goal is to show that as temperature increases, oscillation frequency decreases. This is problematic because it shows that identical circuits placed at different locations on a chip will operate differently if the temperatures are different at the two locations.

The second section focuses on measuring temperature at various locations on a chip. In order to experimentally validate the claim that varying levels of device activity lead to differences in temperature across chip, a method for measuring localized chip temperature is needed. The section documents the design and fabrication of a network of temperature sensors that can be used to measure temperature at various points on a chip. Results from testing the design are also presented.

The final section focuses on showing that varying levels of device activity across a chip can lead to temperature differences across chip. To obtain a chip with varying levels of device activity across the chip, individually controlled blocks of devices were fabricated on a chip. The blocks are used to selectively heat different areas of the chip. The temperature sensor detailed and presented in section two is used to measure the temperature at the different areas. The goal is to show experimentally that the blocks of circuitry, operating at different activity levels, can cause chip temperature to increase locally, creating hot spots. The implication is that, for more elaborate circuits generating more heat, the temperature increases are large, resulting in detrimentally high temperature that can cause local, or even worse, global failure.

Effect of Temperature on Frequency

Device and circuit uniformity is a major concern for 3D integration. It is of the utmost important to have consistency in device and circuit performance across chip. If identical devices and circuits do not behave the same across a chip or across a wafer, then poor chip operation and circuit yield can result. This experiment is intended to show that differences in operating temperature for a ring oscillator can affect the oscillation frequency of the circuit.

For digital circuits, timing is very important. The clock used to drive the circuit must be reliable and must operate to the specifications of the designer. However, since the clock is a circuit that is generally always on and always switching very quickly, temperature is a major concern. With the quick switching of the circuit, devices never have a chance to cool before being switched again. The result is an increase in temperature of the circuit. For a clock circuit, this can affect the frequency of clock, thereby causing timing errors for all circuits driven by the clock.

In this experiment, a 31-stage ring oscillator is used as a representative of an elementary clock circuit. The circuit is shown in Figure 26. The ring oscillator was designed in SPICE using the AMI C5N process. The n-channel MOSFETs each had a drawn gate length of 0.6 μm and a width of 1.5 μm. The p-channel MOSFETs each had a drawn gate length of 0.6 μm and a width of 4.5 μm. The SpectreS environment was used to run transient simulations for 40 ns at various temperatures. Frequency of oscillation was measured from the transient output curves. Since oscillation frequency is proportional to current and current is inversely proportional to temperature, the simulations are expected to show an inverse relationship between frequency and temperature.

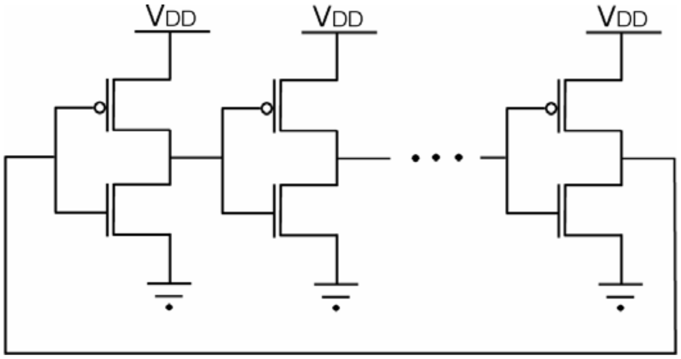


Figure 26: A 31 stage ring oscillator.

Operating temperature conditions were chosen based on temperature at the location of the ring oscillator. Temperatures were obtained from simulations of the 5-stack Pentium III 3D chip presented in chapter 3.

Frequency versus Temperature Simulation

The first simulation performed was an observation of oscillation frequency for a ring oscillator over a range of operating temperatures. Operating temperatures were selected from the range 300 K to 432 K. Since oscillation frequency is proportional to current and current is inversely proportional to temperature, the simulations are expected to show an inverse relationship between frequency and temperature. The results of the simulation are shown in Figure 27. The results do indeed show an inverse relationship between frequency and temperature. This is significant because it shows that frequency of a clock circuit is affected by temperature. If a circuit is expected to operate at a given frequency, depending on the temperature changes in the area where the circuit is fabricated, the circuit frequency might change. In some cases, the frequency change might be significant enough to severely affect overall chip performance.

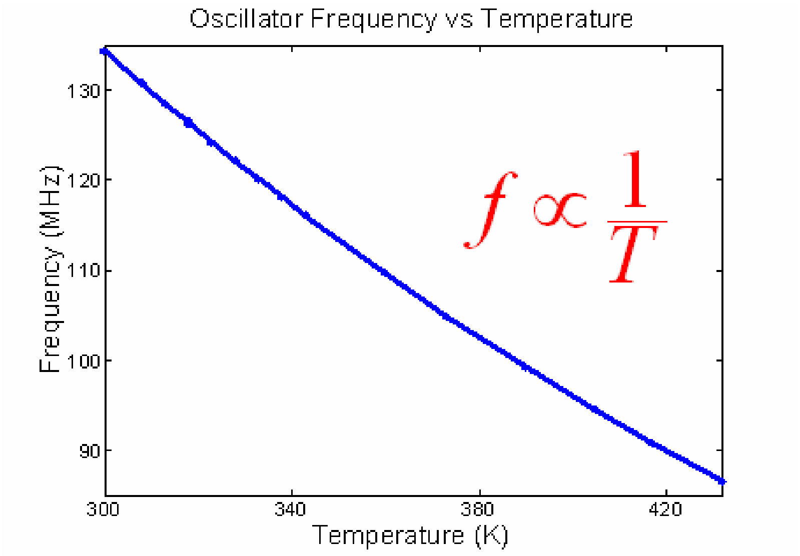


Figure 27: Ring Oscillator frequency versus temperature. The results show an inverse relationship between frequency and temperature.

Frequency Variations on Same Layer

The second simulation involved placing the ring oscillator at the hottest location (clock unit) and the coolest location (L2 cache unit) on the same layer of the 3D chip. Oscillation frequency is expected to vary at these 2 locations because the temperature is different. The operating temperatures used in this simulation obtained were 335 K for the L2 cache unit and 373 K for the clock unit. Plots of the output voltages versus time are shown in Figure 28. The results again show that frequency of oscillation is different at the 2 locations, validating the claim that temperature differences on the same layer can have adverse effects on circuit operation. The desired oscillation frequency is 134.2 MHz obtained at room temperature. However, in the two different locations on the same layer of the chip, the actual operating temperatures are 335 K and 373 K respectively and the frequency of oscillation drops to 119.2 and 104.9 MHz respectively.

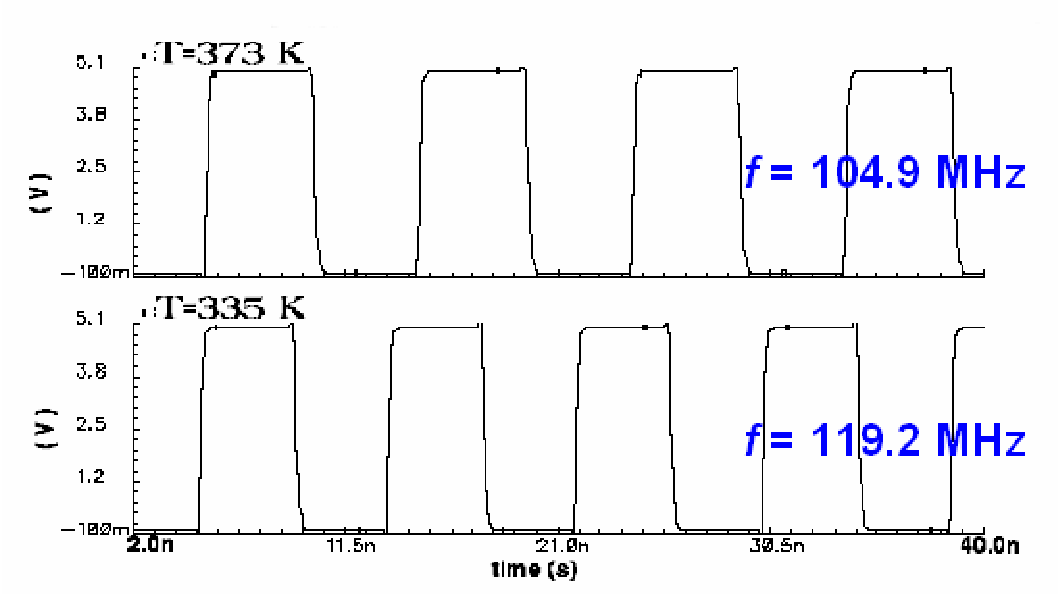


Figure 28: Output voltage plot for a 31-stage ring oscillator. The representative temperatures are found in 2 different locations on the same layer of a 3D circuit. The plots show the expected result that oscillation frequency differs for an oscillator places on the same layer at 2 different locations.

Frequency Variation on Different Layers

The final simulation placed the ring oscillator at the same x-y location (inside the clock unit) on each of the 5 different layers of the chip. Because of symmetry, the simulation only needed to be run for the 3 unique temperatures on the first 3 layers (4th and 5th layers are the same as the 1st and 2nd). The 3 operating temperatures used are 373 K, 417 K, and 432 K. The purpose of this simulation is to show that oscillator frequency is not uniform across device layers because the temperature on different layers is different. Even though each layer is identical in layout, the temperature at the same x-y location on different layers is different because of how heat is thermally conducted through the chip and because of the low conductivity of the layer insulator. The results of the simulation are shown in Figure 29.

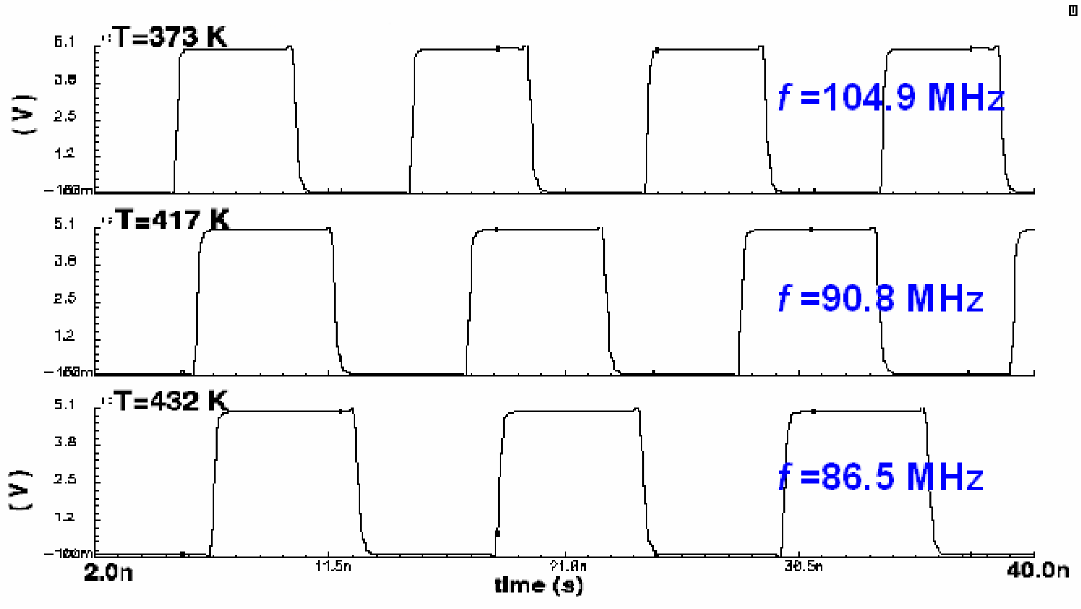


Figure 29: Output voltage plot for transient simulation of 31-stage ring oscillator. The oscillator is placed at the same x-y location on each of the 5 layers of a 3D chip. Because the temperature on each layer is different at that same x-y location, the frequency is different. Temperatures on the upper layers are higher, resulting in lower oscillator frequency.

The results are also in agreement with the expected result: the oscillation frequency decreases with increasing temperature. Since the operating temperature increases for upper layers, the oscillation frequency is lower on the upper layers. If this non-uniformity exists on the same chip, it is expected that the same non-uniformity will exist across the wafer. The result is poor uniformity for different integrated circuits fabricated on the same wafer.

Measuring Local Chip Temperature

In order to experimentally validate the claim that different areas of a chip can generate different amounts of heat and thus operate at different temperatures, an apparatus needs to be designed to measure temperature at different locations on a chip. While concepts such as probing the chip and measuring temperature are possible, a more useful solution is to fabricate chips with the temperature sensing circuitry already integrated onto the chip. This way, as the chip is operated, the measurement circuitry can be monitored to measure temperature at the various locations and no external measurement equipment is necessary. This section focuses on the design and integration of a prototype temperature measurement chip. The steps for fabrication of such a chip are as follows.

1. Design a sensor for measuring temperature.
2. Design and fabricate a chip that uses the sensors to measure temperature at various locations on the chip (Chip #1).
3. Test for functionality, uniformity, and sensitivity to temperature.

Design of Temperature Sensor

The first step to measuring chip temperature is to design a temperature sensor. The sensor should be compact, easy to implement, and adequately sensitive to temperature differences. A basic diode was selected as the temperature sensor to be used because of the simplicity and compactness of the design, and the sensitivity offered. Elaborate sensors, while offering increased sensitivity, would also require more complicated design. Such designs would occupy a large area and also generate more heat. It was desired to have a small sensor whose contributions to local heating are negligible.

The concept behind using a diode as a temperature sensor is evident by observing the basic diode equation (4.1).

$$I_D = I_S \exp\left(\frac{V_D}{V_T}\right) \tag{4.1}$$

Both the reverse saturation current I_S and the thermal voltage V_T are temperature dependent. However, because of the exponential, the variation of thermal voltage with temperature is expected to dominate. This results in an exponential variation of diode current I_D with the inverse of temperature. The exponential relationship provides adequate temperature sensitivity. By heating a representative diode to a known temperature *T*, applying a known voltage V_D across the diode, and measuring the diode current I_D, a table can be compiled that relates diode current to temperature. This way, sensors placed at various locations can be used to measure temperature by measuring the sensor current and using the table to find the temperature that corresponds to the measured current.

The diode sensor was fabricated using the AMI C5N design process offered by MOSIS. Dimensions for the diode were calculated by hand and the layout was done using Cadence. The dimensions were selected by taking into consideration the area of the diode, the desired range of temperature operation, and the desired current. The diode needed to operate at temperatures in the range of 293 K to 400 K and currents on the order of a few hundred microamperes were desired. To achieve these design requirements, the diode was chosen to have n$^+$ and p$^+$ active area dimensions of 10.5 μm x 10.5 μm. The layout of the diode is shown in Figure 30.

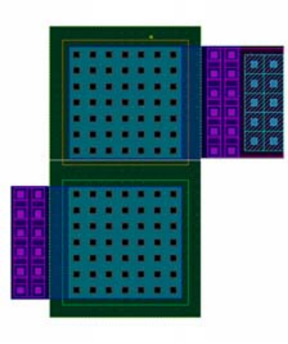


Figure 30: Cadence layout of a diode temperature sensor fabricated through MOSIS.

Design of Temperature Sensor Array

The second step to measuring temperature at various locations is to design a method of placement of sensors at specified locations and develop the circuitry to select a particular diode to measure current through. Because temperature will be measured at multiple locations, an abundance of sensors is needed. However, the sensors should not all operate at the same time because as long as they are on, they are heating up. Thus, what is needed is some type of diode sensor array and control circuitry to select only one diode at a time to turn on and measure current.

The design selected was a simple array of 100 diode sensors, arranged in a 10x10 array. There are 10 column bus lines and 10 row bus lines and each diode has its p-terminal connected to its corresponding column line and its n-terminal connected to its corresponding row bus line. A 2x2 sample array is shown in Figure 31.

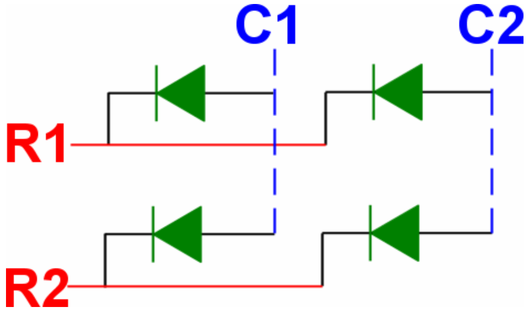


Figure 31: Diode sensors arranged in a 2x2 array. The p-terminal of each device is connected to its corresponding column line and n-terminal of each device is connected to its corresponding row line.

The setup for selecting a diode is shown in Figure 32. Each row is tied to V_{DD} through control circuitry while each column is left open. To select a diode in the array for current measurement, the corresponding row is switched low using control circuitry and a voltage is applied to the column containing the desired device. The result is the selection of only 1 diode out of the 100 because only one diode will have both is n-terminal connected to ground and a voltage applied to the p-terminal. To measure the current through the diode, a resistor is connected to the column line and a voltage is applied. The current through the resistor is the diode current, which can be used in a lookup table to obtain the corresponding diode temperature.

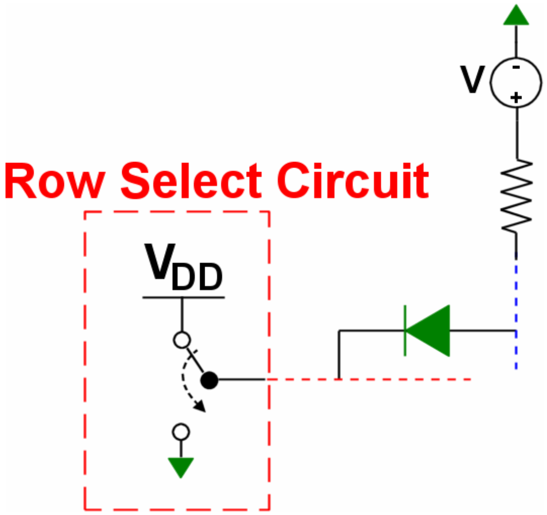
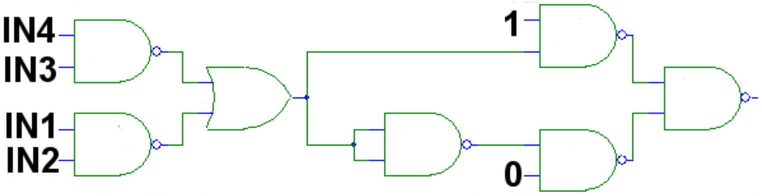


Figure 32: Apparatus for measuring diode current for a particular sensor in the sensor array.

The last part of the design involved designing the row select circuit shown as a block in Figure 32. The circuit behaves like a simple switch so that when a row is selected, the circuit switches the row line from high to low. There are 10 rows in the diode sensor array and in order to select up to 10 different lines, 4 control inputs are needed. If each row is assigned a 4-bit binary number, then the rows will be numbered 0000 to 0110 and the 4 control bits, *ABCD*, can be used to select the row. Figure 33 shows the circuit that will be used to switch the row and its corresponding truth table.



IN1	IN2	IN3	IN4	OUT
X	X	X	X	0
1	1	1	1	1

Figure 33: Circuit schematic and truth table for the row select circuit placed at the beginning of each row of the diode array.

The circuit switches the row line to low when the 4 inputs are high. Otherwise, it outputs low. Thus, given a control input *ABCD*, the 4 inputs must be connected in their inverted or noninverted form to the 4 inputs on the row select circuit for each row. The connection should be done in such a way that the row select circuit for the desired row sees the input $IN_1 IN_2 IN_3 IN_4 = 1111$. For example, since row 2 is assigned the binary number $ABCD = 0001$, the inputs to the row select circuit for row 2 are wired as $IN_1 IN_2 IN_3 IN_4 = \overline{A}\overline{B}\overline{C}D$. This way, the inputs to the row select circuit on row 2 will all be high when the control inputs are $ABCD = 0001$ and row 2 will be switched to low. The inputs to the row select circuits for the other rows are connected using the same method. For any given input, only one row select circuit will switch the row on because only one of them will have all the inputs high.

Fabrication of Chip #1: Temperature Sensing Chip

The entire diode sensor array with the control circuitry was brought together and fabricated as Chip #1. The layout was done using Cadence and the chip was fabricated using the AMI C5N process offered by MOSIS. The chip was integrated in a 40 pin package. Ten of the pins were used to apply voltage to the columns of the sensor array; 4 were used as row selection inputs A, B, C, and D; 2 pins were used for power and ground. The other pins were used to test an individual diode sensor outside of the array, an n-channel MOSFET, and a p-channel MOSFET. The layout of the chip is shown in Figure 34.

The chip was used as a prototype to verify that such an array of diodes could be used to measure temperature at 100 different locations on a chip. Testing the chip consisted of confirming the operation of the array, verifying that the placement of the array did not affect the individual behavior of each diode, and showing that diode current increases as heat is applied to the chip. Testing of each diode sensor should show similar behavior evidenced by comparison of *I-V* characteristic curves. Comparison of the *I-V* curves for the diodes in the array with a diode outside of the array should show that the array does not affect the individual operation of the diodes. Finally, temperature versus current tests should show that diode current increases as heat is applied to the chip.

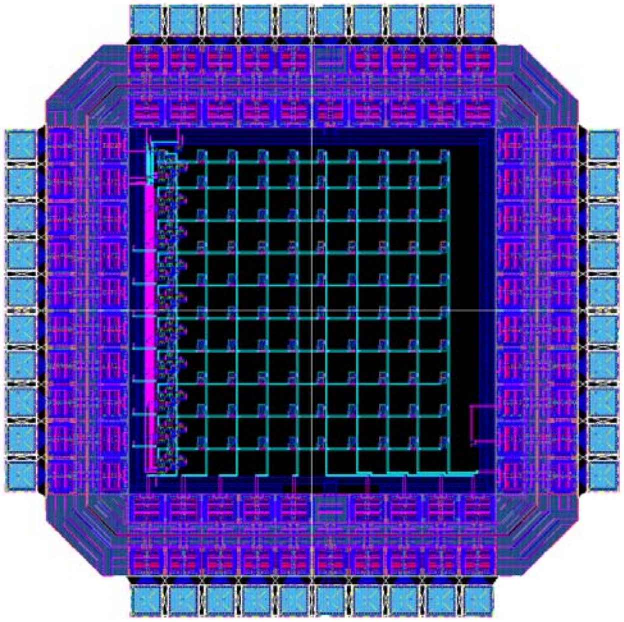


Figure 34: Cadence layout of Chip #1: a 10x10 diode sensor array with control circuitry to select the desired diode.

Functionality Test

Testing of the prototype chip for functionality of the diode sensor array yielded expected results. All 100 diodes in the array were successfully selected and operated. Current in the desired range of hundreds of microamperes was measured through each using the method described previously. A 100 Ω resistor was used and the voltage across the resistor was measured so that the diode current could be calculated using Ohm’s Law. Each of the diodes behaved similarly, as was desired. The *I-V* characteristic of a representative diode sensor in the array is shown in Figure 35. It is worth noting that the diode current is not exponential for all voltages. At high voltage, the depletion region vanishes and the diode behaves as a basic resistor.

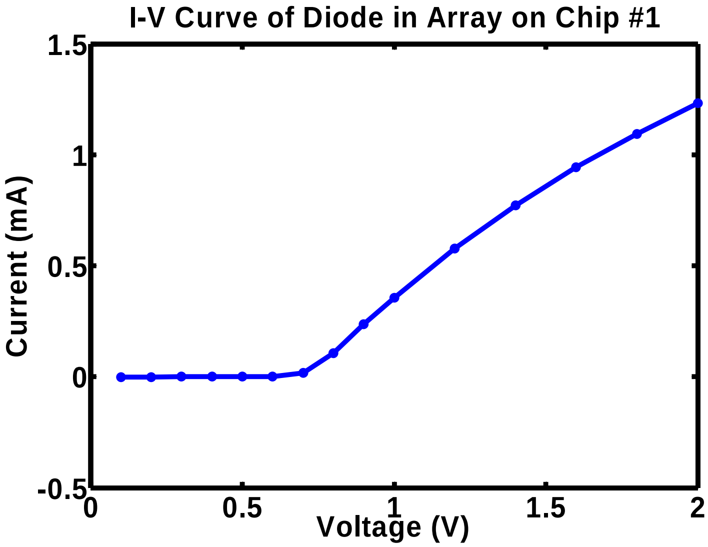


Figure 35: ***I-V*** **characteristic curve for a representative diode in the sensor array on Chip #1.**

Comparison of the *I-V* curve for a single diode fabricated outside of the sensor array to those within the array also yielded expected results. The single diode fabricated outside of the sensor array was tested using a 1 kΩ. The *I-V* curve, shown in Figure 36, is similar to the curves for the diodes in the array. Thus, the row select circuitry and the arrangement of the diodes in the sensor array did not affect the individual operation of the diodes in the array.

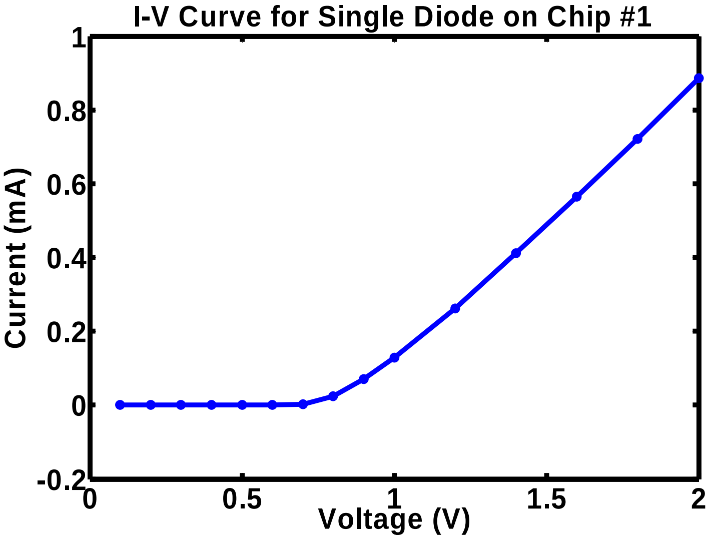


Figure 36: ***I-V*** **characteristic curve for a single diode fabricated outside of the sensor array on Chip #1.**

Temperature Sensitivity Test

The prototype chip was tested to see if applying heat to the chip resulted in an increase in diode current. A current versus temperature test was conducted according to the following experimental setup.

1. Set the control inputs ABCD to 0000 and apply a voltage through a 1 kΩ resistor to column 1 (select the diode in row 1 column 1 of the array).
2. Set the input voltage to 0.995 V to turn the diode on and ensure operation in the exponential region of the *I-V* curve
3. Use a thermometer to measure temperature outside the chip package.
4. Position a lamp directly over the thermometer.
5. Measure initial current and temperature with the light off.
6. Turn the light on.
7. Measure temperature and voltage across the resistor every 30 seconds for 30 minutes.

The light is used to increase the ambient temperature immediately outside of the chip. As the ambient temperature is increased, the temperature of the chip inside the package should increase as well. The diode sensors should respond to this increase with an increase in current.

The test did reveal a sensitivity of the diodes to temperature. As seen in Figure 37, diode current did increase with temperature. However, the relationship was not exponential, as expected. A line seemed to best fit the data, suggesting a more linear relationship. There are a few reasons for this behavior. First, the diode sensors rested inside the packaging on the chip. It is possible that the temperature inside the chip did not increase as much as it did outside, where temperature was actually being measured. Also, the dependence of saturation current on temperature may be more pronounced then previously thought, thereby decreasing the exponential relationship of temperature and current. A much better, more accurate method of heating the chip, perhaps using a furnace, would better quantify the current-temperature relationship.

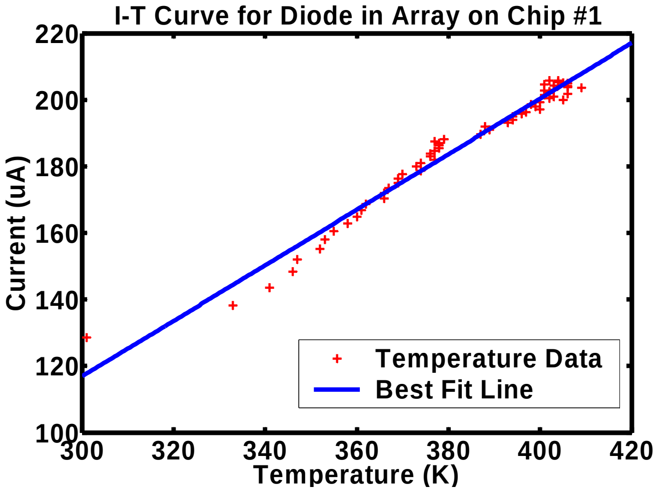


Figure 37: Current vs. Temperature plot for representative diode in sensor array.

Local Chip Heating and Temperature Measurement

It was previously stated that areas of a chip with increased device activity will operate at a higher temperature. Likewise, areas of less device activity will operate at lower temperatures. The tests presented in this section were conducted to experimentally validate these claims. The experiments were intended to show that it is possible to selectively heat a given area of a chip and then measure the temperatures across the chip to see how temperature changes according to the heat generated. To accomplish this task, it was necessary to achieve design requirements as outlined in the list that follows.

1. Design circuitry to selectively heat areas of a chip divided into blocks.

2. Combine selective heating circuitry with the temperature sensor measurement circuitry presented and tested in section 2 of this chapter (Chip #2).
3. Test for functionality and uniformity of the sensor array on Chip #2 and compare to Chip #1.
4. Test for functionality of the entire chip by selectively heating different areas of the chip and measuring temperature at various locations.

Selective Heating Design

The first step in observing localized temperature increase due to device activity is designing the circuitry to selectively heat different areas of the chip. This requires the use of devices to generate heat during operation, grouping of such devices into blocks on the chip, and designing a method for selectively activating the different blocks. The concept is simple: devices generate heat; grouping more devices together generates more heat; spreading groups of devices across the chip in blocks allows for generating heat in selective areas on chip. Temperature within the blocks can then be measured.

Basic n-channel MOSFETs were used as the heat generation devices. MOSFETs were the most obvious choice because of simplicity of fabrication and size. The MOSFETs are packed as close together as design rules allow into blocks of thousands of devices. The blocks are arranged in a 5x4 array. Within the block, each of the MOSFETs has the source grounded and the drain tied to V_{DD}. There is a single block select line in each block that is connected to the gate of every device in the block. This way, a block is easily turned on by applying a voltage to the block select line of the desired block. An example of a single block is shown in Figure 38.

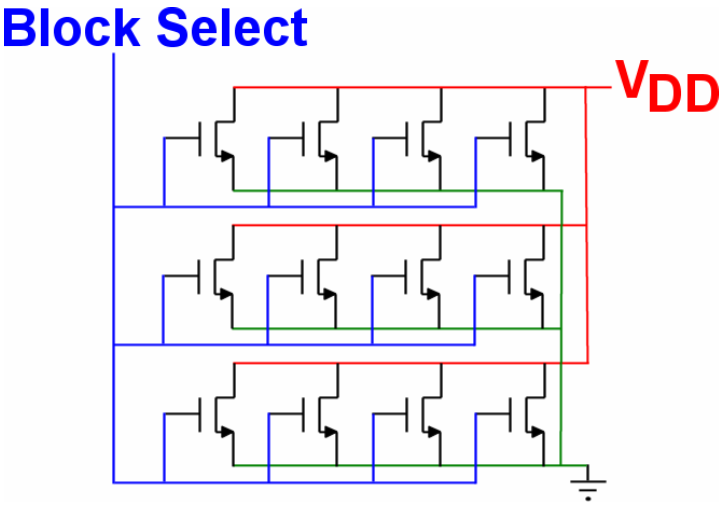


Figure 38: Representative block of n-channel MOSFET devices. The entire block of devices is turned on when a voltage is applied to the block select line. The block facilitates local heating.

Fabrication of Chip #2: Selective Heating and Measurement Chip

To selectively heat a chip and observe local temperature, the selective heating circuit and the temperature sensor array were integrated onto the same chip (Chip #2). This allowed for the creation of a chip divided into 20 functional blocks with 100 temperature sensors spread evenly over the chip. The functional blocks each contain thousands of devices and can be selectively activated. The sensors are used to measure temperature at various locations to observe the effect of heat generation within a block on the temperature at any of the 100 locations in the sensor array.

The chip layout was performed in Cadence using the AMI C5N process. It was fabricated through MOSIS and integrated in a 40 pin package. Ten pins are dedicated to selecting 1 of the 10 columns in the sensor array; 4 are used for the control inputs *A*, *B*, *C*, and *D* to select a row in the sensor array; 2 are used for power and ground; 20 are used to apply a voltage to the 20 different functional blocks. The remaining pins are used to apply a gate and a drain voltage to a single n-channel MOSFET fabricated outside of the functional blocks. The Cadence layout of the chip is shown in Figure 39.

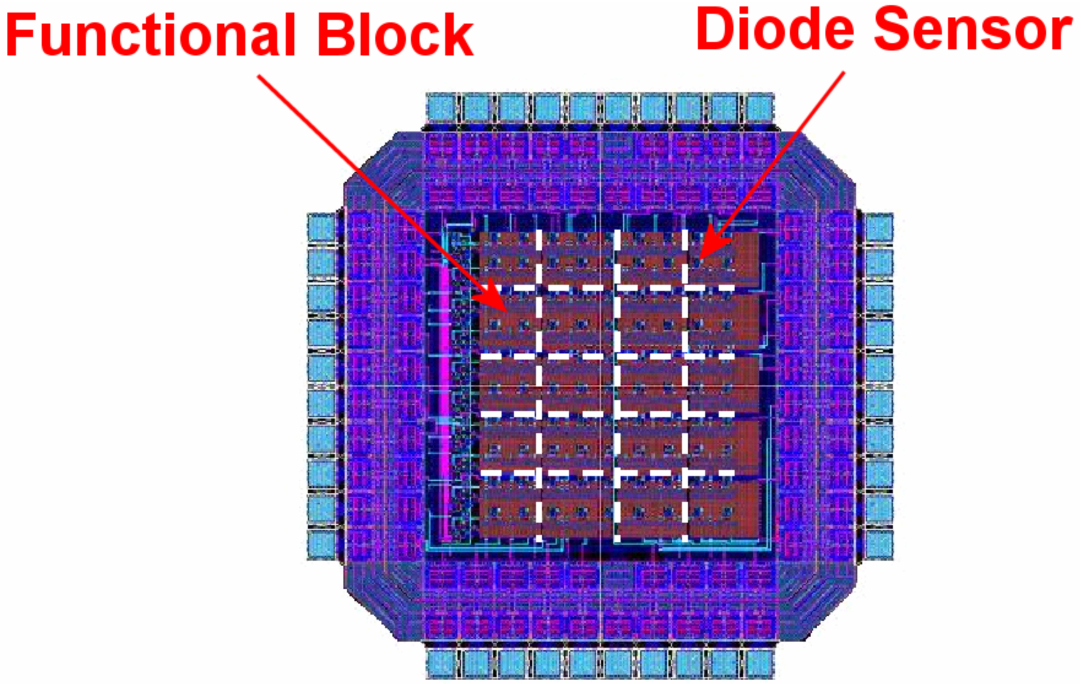


Figure 39: Cadence layout of Chip #2: Selective Heating and Measurement Chip. The chip contains a 10x10 diode temperature sensor array spread across the chip and a 5x4 array of functional blocks. The white dashed lines show the functional blocks.

Functionality Tests

The first tests performed on Chip #2 were functionality and uniformity tests. It was necessary to verify the operation of the diode temperature sensor array and make sure that its behavior was similar to the array fabricated on Chip #1. This guarantees that the fabrication of the functional blocks around the diode did not cause a change in the expected behavior of the diode sensors. In addition, a single n-channel MOSFET fabricated outside of the functional blocks was tested in order to show the *I-V* characteristics for the MOSFETs used in the functional blocks.

To verify the functionality of the diode array, all 100 diodes were individually selected and diode current was measured using a 1 kΩ resistor in accordance with the procedure outlined in section 2 of this chapter. The results showed similar behavior for each of the diodes. The behavior was also similar to that obtained for Chip #1, as desired. The representative *I-V* characteristic curve, shown in Figure 40, is similar to the one shown in Figure 35 for Chip #1. The results show that, since sensor behavior is the same for both chips, the fabrication of functional blocks around the sensor has no adverse effects on sensor operation. Thus, when functional blocks are turned on, the only change in current measured on the diode sensors should be due to changes in temperature cause by heat generated from device activity.

The other functionality test performed on Chip #2 was a verification of device characteristics for the n-channel MOSFETs used in the functional blocks for generating heat to selectively heat the chip. A single n-channel MOSFET was fabricated outside of the diode sensor array and outside of the functional blocks for the purpose of observing representative device characteristics. It is assumed that all the devices on the chip in each of the functional blocks behave similarly to this lone transistor. Drain current versus drain voltage at different gate voltages was measured and the resulting curve is shown in Figure 41. The result is also in agreement with SPICE simulation of the n-channel MOSFET.

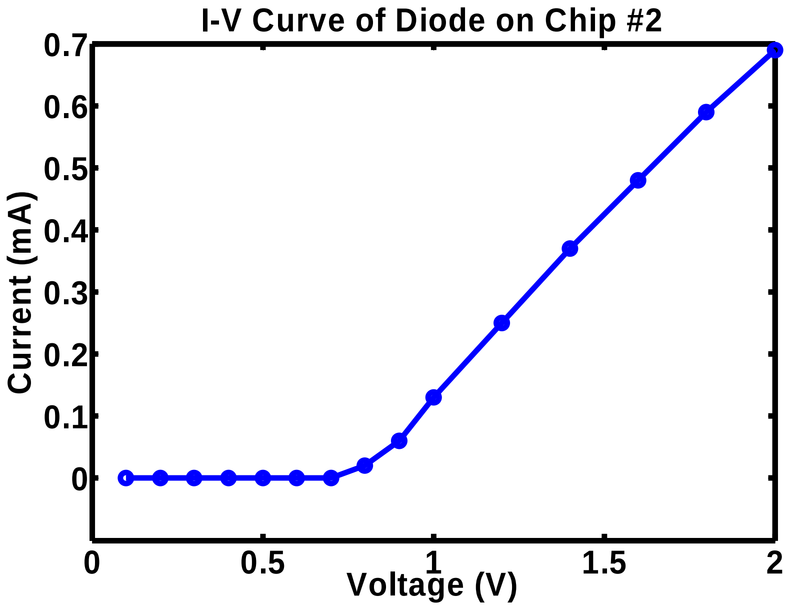


Figure 40: *I-V* characteristic curve for sensor in the sensor array fabricated on Chip #2.

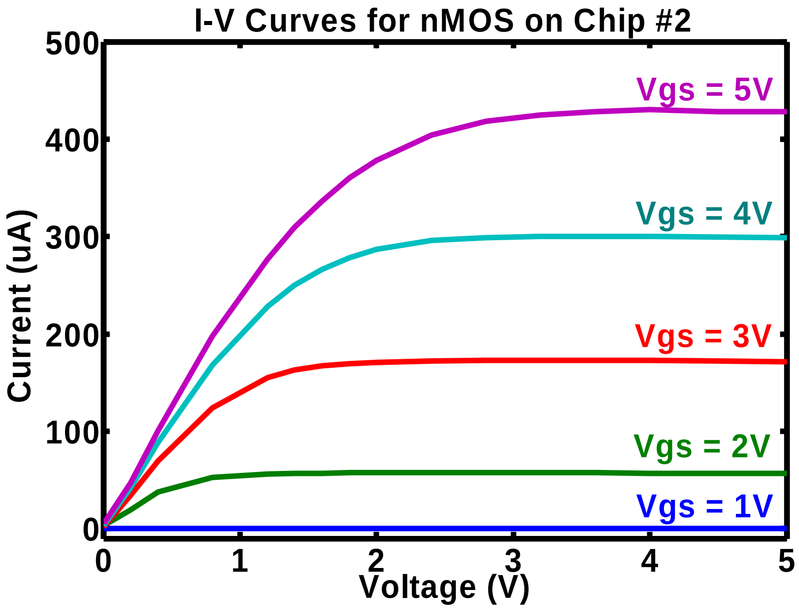


Figure 41: I_D-V_{DS} curve for a single n-channel MOSFET fabricated on Chip #2, outside of the functional blocks and the diode sensor array.

Selective Heating and Temperature Measurement Test

The selective heating and temperature measurement test is the most important test of all tests performed on Chip #2. The results of this test can be used to show the feasibility of selectively heating different areas of a chip and the possibility of measuring localized temperature in order to observe the effects of device activity on local temperature. The experiment conducted consisted of applying a gate voltage to a given functional block of n-channel MOSFETs. Temperature at different locations on the chip is then measured using the sensor array. Observations can be made on how temperature changes at different locations as different functional blocks are turned on or off.

The results of the test, unfortunately, were inconclusive. It was expected that turning on a block of devices would cause in increase in current through diode sensors near that functional block. The current increase is contributed to an increase in temperature inside the functional block and in areas near that block because of heat generated by the devices. The observed behavior, however, was the opposite. When a block was turned on, a decrease in current through diode sensors both near the functional block and far away from it was detected. Since turning on a functional block affected all 100 diode sensor in the same way, the conclusion must be made that the sensitivity to device activity due to turning on a functional block cannot be due to temperature increase because turning on one functional block would not affect temperature across the entire chip equally. Areas closer to the functional block would see a larger increase in temperature then areas far from the device. For large chips, areas far away from the functional block should, realistically, not be affected at all.

The inconclusive results of the experiment lead one to question the problem with the experimental apparatus. It was already experimentally shown through tests on Chip #1 that the sensors in the array were proportionately sensitive to temperature.

Thus, the functional block design must be causing the discrepancies in the results, since it is the only new circuitry added to the chip. Careful viewing of the layout of Chip #2 revealed an error in the functional blocks: the body of the n-channel MOSFETs used in the functional blocks was inadvertently connected to V_{DD} instead of ground. Therefore, the assumption that the behavior of the single n-channel MOSFET was representative of all the devices is incorrect. Furthermore, in some areas on the chip, the substrate is at V_{DD} while in other areas, it is grounded. This is a significant mistake that should be corrected for future measurements.

Another issue with Chip #2 is the maximum current observed through the diodes versus the maximum current attainable through the n-channel MOSFETs. Each row select circuit has the output of a NAND gate connected to the n-terminal of the diodes of that row. The maximum current observed through the single n-channel MOSFET for a gate voltage of $V_{GS} = 5\text{V}$ was about 450 μA, as shown in Figure 41. It follows that the maximum current observable through the diode sensors should not be more then this maximum drawn current. However, currents much higher were observed. This leads to the conclusion that perhaps due to the substrate voltage issue, or perhaps some other reason, current is being sourced into to the diodes. This is problematic and unexplainable given the pins available for measurement and observation.

Further Heating and Measurement

Because of the inconclusiveness of the results for Chip #2, further designs and tests should be formulated to validate the claim that increased device activity leads to increased temperature and thermal coupling between devices leads to differences in how heat is dissipated on a chip. One suggested approach is to redesign Chip #2 with the proper connection of the body of the devices in the functional blocks. This will verify the feasibility of the design of using functional blocks to selectively heat a chip. Another approach is to simply scrap using n-channel MOSFETs in the functional blocks and, instead use an even simpler approach to heat the chip. One design waiting to be tested employs polysilicon resistor chains spread across chip. Since the resistors generate heat, they can be tightly packed around the diode sensors and voltage can be applied to them to generate heat. Temperature changes can then be observed using the diode sensors.

In addition, the experiments in this section and the previous were used to show that the diode sensor array was sensitive to temperature increases. However, for more accurate temperature measurement using the array, a better, more controlled method of heating the chip should be investigated. The lookup table for temperature based on observed current should be accurate, demanding a more controlled way of heating the chip and knowing the precise temperature of the chip when current is observed. One suggestion is to use a furnace that allows for warming of the chip to a precise temperature and then observing the current through the diode sensors.

Chapter 5: Conclusion

The work detailed in this thesis can be categorized as follows: 3D integration, location specific chip temperature calculation, and simulation validation. For the topic of 3D integration, the important results are:

- 3D integration offers increased device density and reduced interconnect lines, resulting in more complex circuits that will also operate faster.
- 3D integration is not without it challenges, most notably the unwanted increase in local, layer, and overall chip temperature because of increased device density.

For location specific temperature calculation, the important results are:

- Heat generation due to increased device activity is a major issue that leads to “hot spots” on a chip.
- Mixed-mode simulation of device and chip level performance models reveals chip temperature as a function of position on chip.
- As the number of layers increases, location specific temperatures exceed the maximum recommended operating temperature.
- Chips with 5 device layers exhibit unacceptable operating temperatures for all devices on the middle layer.
- Functional units with higher device activity should be placed near units with lower activity, since the lower activity units are cooler.

Finally, for simulation validation, the important results are:

- Elementary clock circuits constructed using ring oscillators exhibit an inversely proportional relationship for frequency to temperature.
- 31-stage ring oscillators placed in both the clock and L2 cache units on the same layer of a 3D chip will operate at different frequencies because of differences in temperature in the 2 units. The frequency in the cache is higher then in the clock because the temperature in the cache is lower.
- 31-stage ring oscillators placed in the same functional unit on 3 different layers of a 5-layer 3D chip operate at different frequencies because of temperature differences on the 3-layers. The oscillation frequency is lower for upper layers because temperature is higher on upper layers.
- It is possible to use an array of diode temperature sensors controlled by selection circuitry to experimentally measure chip temperature at various locations on a chip
- Better methods of selectively heating a chip must be adopted in order to adequately test the diode array.

References

[1] Semiconductor Industry Association, “Overall roadmap technology characteristics,” *International Technology Roadmap for Semiconductors 2004 Update*, [online], 10 Jan. 2005, Available: http://www.itrs.net/Common/2004Update/2004_000_ORTC.pdf, 20 Apr. 2005.

[2] David J. Frank *et al.*, “Device scaling limits of Si MOSFETs and their application dependencies,” *Proc. IEEE*, vol. 89, no. 3, Mar. 2001, pp. 259-288.

[3] Kaustav Banerjee *et al.*, “3-D ICs: a novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration,” *Proc. IEEE*, vol. 89, no. 5, May 2001, pp. 602-633.

[4] R. Zhang, K. Roy, and D.B. Janes, “Exploring SOI device structures and interconnect architectures for low-power high performance circuits,” *IEE Proc.on Computer Digital Techniques*, vol. 149, no. 4, July 2002, pp. 137-145.

[5] Sang-Soo Lee and David J. Allstot, “Electrothermal simulation of integrated circuits,” in *IEEE J. Solid-State Circuits*, vol. 28, no. 12, Dec. 1993, pp. 1283-1293.

[6] M. W. Geis, D. C. Flanders, and Henry I. Smith, “Crystallographic orientation of silicon on an amorphous substrate using an artificial surface-relief grating and laser crystallization,” *Applied Physics Lett.*, vol. 35, issue 1, July 1, 1979, pp. 71-74.

[7] Shukri J. Souri *et al.*, “Multiple Si layer ICs: motivation, performance analysis, and design implications,” in *Proc. 37th Conf. Design Automation*, Los Angeles, Calif., June 5-9, 2000, pp. 213-220.

[8] Rongtian Zhang, Kaushik Roy, and David B. Janes, “Architecture and performance of 3-dimensional SOI circuits,” in *1999 IEEE Int. SOI Conf. Proc.*, Rohnert Park, Calif., Oct. 4-7, 1999, pp. 44-45.

[9] Victor W. C. Chan, Phillip C.H. Chan, and Mansun Chan, “3D integrated circuit using large grain polysilicon film,” in *Proc. 6th Int. Conf. Solid-State and Integrated-Circuit Technology*, vol. 1, Shanghai, China, Oct. 22-25, 2001, pp. 58-61.

[10] Mansun Chan, “The potential and realization of multi-layers three-dimensional integrated circuit,” in *Proc. 6th Int. Conf. Solid-State and*

Integrated-Circuit Technology, vol. 1, Shanghai, China, Oct. 22-25, 2001, pp. 405-45.

[11] S.J. Abou-Samra *et al.*, “3D CMOS SOI for high performance computing,” in *1998 Int. Symp. Low Power Electronics and Design*, Monterey, Calif., Aug. 10-12, 1998, pp. 54-58.

[12] K. W. Guarini *et al.*, “Electrical integrity of state-of-the-art 0.13 μm SOI CMOS devices and circuits transferred for three-dimensional integrated circuit fabrication,” in *IEDM Tech. Dig.*, Dec. 8-11, 2002, pp. 943-945.

[13] J. Burns *et al.*, “Three-dimensional integrated circuits for low-power, high bandwidth systems on a chip,” *IEEE Int. Solid-State Circuits Conf. Dig. Technical Papers*, Feb. 5-7, 2001, pp. 268-269; 453.

[14] René P. Zingg *et al.*, “Three-dimensional stacked MOS transistors by localized silicon epitaxial overgrowth,” *IEEE Trans. Electron Devices*, vol. 37, no. 6, June 1990, 1452-1461.

[15] John P. Denton, Sang Woo Pae, and Gerold W. Neudeck, “Vertical integration of submicron MOSFETs in two separate layers of SOI islands formed by silicon epitaxial lateral overgrowth,” in *Proc. 11th Great Lakes Symp. VLSI*, West Lafayette, Ind., 2001, pp. 129-132.

[16] K. Yamazaki *et al.*, “4-layer 3-D IC technologies for parallel signal processing,” in *IEDM Tech. Dig.*, San Francisco, Calif., Dec. 9-12, 1990, pp. 599-602.

[17] Gerold W. Neudeck, “Multiple layers of silicon-on-insulator for nanostructure devices,” *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, vol. 17, issue 17, May 1999, pp. 994-998.

[18] T. Kunio *et al.*, “Three dimensional ICs, having four stacked active device layers,” in *IEDM Tech. Dig.*, Washington, D.C., Dec. 3-6, 1989, pp. 837-840.

[19] Michael L. Alles, “Thin-film SOI emerges,” *IEEE Spectrum*, vol. 34, issue 6, June 1997, pp. 37-45.

[20] Sorin Cristoloveanu and Francis Balestra, “SOI technologies, materials and devices,” in *Int. Semiconductor Conf., 1996*, vol. 1, Sinaia, France, Oct. 9-12, 1996, pp. 3-12.

[21] Richard S. Mullins and Theodore I. Kamins, *Device Electronics for Integrated Circuits*, 3rd ed., New York, N.Y.: John Wiley & Sons Inc., 2003, pp. 443-447.

[22] Anna W. Topol *et al.*, "Enabling technologies for wafer-level bonding of 3D MEMS and integrated circuit structures," in *Proc.ECTC '04*, vol. 1, Las Vegas, Nev., June 1-4, 2004, pp. 931-938.

[23] Gerold W. Neudeck, "Three-dimensional CMOS integration," *IEEE Circuits and Device Mag.*, vol. 6, issue 5, Sept. 1990, pp. 32-38.

[24] Albert W. Wang and Krishna C. Saraswat, "A strategy for modeling of variations due to grain size in polycrystalline thin-film transistors," *IEEE Trans. Electron Devices*, vol. 47, no. 5, May 2000, pp. 1035-1043.

[25] Lance R. Thompson *et al.*, "NMOS device characteristics in electron-beam recrystallized SOI," *IEEE Trans. Electron Devices*, vol. 40, no. 7, July 1993, pp. 1270-1276.

[26] Atsushi Kohno *et al.*, "High performance poly-Si TFTs fabricated using pulsed laser annealing and remote plasma CVD with low temperature processing," *IEEE Trans. on Electron Devices*, vol. 42, no. 2, Feb. 1993, pp. 251-257.

[27] M. A. Crowder, "Low-temperature single-crystal Si TFT's fabricated on Si films processed via sequential lateral solidification," *IEEE Electron Device Lett.*, vol. 19, no. 8, Aug. 1998, pp. 306-308.

[28] Krishna C. Saraswat *et al.*, "3-D ICs: motivation, performance analysis, and technology," *Proc. 26th European Solid-State Circuits Conf.*, [online], 2000, Available: http://www-ee.standford.edu/~kaustav/papers/ESSCIR2000.pdf, Apr. 20, 2005.

[29] Vivek Subramanian and Krishna C. Saraswat, "High-performance germanium-seeded laterally crystallized TFT's for vertical device integration," *IEEE Trans. Electron Devices*, vol. 45, no. 9, Sept. 1998, pp. 1934-1939.

[30] Vivek Subramanian *et al.*, "Low-leakage germanium-seeded laterally-crystallized single-grain 100-nm TFT's for vertical integration applications," *IEEE Device Lett.*, vol. 20, no. 7, July 1999, pp. 341-343.

[31] Hongmei Wang *et al.*, "High frequency performance of large-grain polysilicon-on-insulator MOSFETs," *IEEE Trans. Electron Devices.*, vol. 48, no. 7, July 2001, pp. 1480-1482.

[32] Singh Jagar *et al.*, "Single grain thin-film transistor (TFT) with SOI CMOS performance formed by metal-induced-lateral-crystallization," in *IEDM Tech. Dig.*, Washington, D.C., Dec. 5-8, 1999, pp. 293-296.

[33] Amol R. Joshi and Krishna C. Saraswat, “High performance submicrometer CMOS with metal induced lateral crystallization of amorphous silicon,” *J. Electrochemical Society*, vol. 150, issue 8, Aug. 2003, pp. 443-449.

[34] Jae-Mo Koo *et al.*, “Integrated microchannel cooling for three-dimensional electronic circuit architectures,” *J. Heat Transfer*, vol. 127, Jan. 2005, pp. 49-58.

[35] Akin Akturk, Latise Parker, Neil Goldsman, and George Metze, “Mixed-mode simulation of non-isothermal quantum device operation and full-chip heating,” *Proc. 2003 Int. Semiconductor Device Research Symp.*, Washington, D.C., Dec. 10-12, 2003, pp. 508-509.

[36] K. Skadron *et al.*, “Temperature-aware microarchitecture,” *Proc. 13th Ann. Int. Symp. Computer Architecture*, June 9-11, 2003, pp. 2-13.

[37] Intel Corporation, “Processors,” *Package Type Guide (Desktop Processors)*, [online], Aug. 15, 2004, Available: http://www.intel.com/support/processors/sb/cs-009863.htm, Apr. 20, 2005.

[38] Intel Corporation, “Physical constants of IC package materials,” *Intel Packaging Information*, [online], 2005, Available: http://developer.intel.com/design/packtech/ch_05.pdf, Apr. 20, 2005.

[39] D. Brooks, V. Tiwari, and M. Martonosi, “Wattch: a framework for architectural-level power analysis and optimization,” *Proc. 27th Int. Conf. on Computer Architecture*, vol. 28, issue 2, Vancouver, B.C., June 10-14, 2000, pp. 83-94.

[40] M. Martonosi, D. Brooks, and P. Bose, “Modeling and analyzing CPU power and performance: metrics, methods, and abstractions,” tutorial presented at *SIGMETRICS 2001*, Cambridge, Mass., June 16-20, 2001, Available: http://www.princeton.edu/~mrm/tutorial/Sigmetrics2001_tutorial.pdf, Apr. 20, 2005.

[41] Intel Corporation, “Processor spec finder,” [online], 2005, Available: http://processorfinder.intel.com/scripts/list.asp, Apr. 20, 2005.