Strategies to improve power delivery to CMOS circuitry using localized power decoupling

By

Ophelia Mohaimen
Rezwana Habib Mustazir
Laila Sharmin Sraboni

Submitted to the

Department of Electrical and Electronic Engineering
Faculty of Sciences and Engineering
East West University

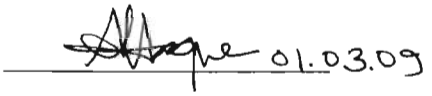
in partial fulfillment of the requirements for the degree of
Bachelor of Science in Electrical and Electronic Engineering
(B.Sc. in EEE)

Fall, 2008

Approved By

Thesis Advisor

Md. Ishfaqur Raza



Chairperson

Anisul Haque

ABSTRACT

As current travels all the way from the power supply to the silicon, it sees the current path in each of the levels as resistive and inductive drops, thus voltage deteriorates. Power droop in the silicon is a major cause for system performance degradation. Higher frequency of operation and reduced power levels are limiting the timing and voltage budget, which is designed in circuits to account for system noise, which includes voltage drooping due to inductive losses. Novel techniques are evolved to compensate for these losses at all levels, starting from motherboard, package, socket and finally down to the silicon level.

Due to lack of available space, design constraints and fabrication difficulties, decoupling at the die level is very limited. Though available decoupling techniques exist for the board, package and socket, yet their response time is slow and in some cases worthless.

Here a proposal is made to provide for decoupling at the CMOS levels, right where the power is needed. This work merges the advanced DRAM technologies with that of CMOS to create decoupling in silicon without taking any extra rooms for it. Simulation of sub 100 nm multi-metal layer circuit demonstrates the advantage of proposed localized decoupling.

ACKNOWLEDGEMENTS

We would like to express our deep gratitude and appreciation to our supervisor Md. Ishfaqur Raza, PhD, Associate Professor of the Electrical And Electronics Engineering, Department of East West University, for his excellent guidance and encouragement. The thesis completion is a result of his unparallel support and care.

Special thanks to Dr. Anisul Haque, Dean of Science Faculty East West University for his guidelines, support and care.

We would like to thank S. M. Salahuddin, Research Lecturer of the EEE Department of East West University, for his generous assistance.

We also thank all the respected teachers of EEE department of East West University for their continuous inspiration and support.

We are thankful to M.A Jobayer Bin Bakkre, the Lab-in-Charge of VLSI Lab and all other Lab-in-charges of Electrical, Electronics and Tele-Communications Labs.

We are grateful to our family members and classmates for being there when we needed their help.

Above everything we thank our Almighty, Allah, for giving us the ability, scope, patience and energy to complete this commitment efficiently.

AUTHORIZATION PAGE

We hereby declare that we are the sole author of this thesis. We authorize East West University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Ophelia 24th Dec '08

Rezwana 24th Dec 2008

Laila Sharmin Sraboni 24th Dec '08

Ophelia Mohaimen
Rezwana Habib Mustazir
Laila Sharmin Sraboni

We further authorize East West University to reproduce this thesis by photocopy or other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Ophelia 24th Dec '08

Rezwana. 24th Dec 2008

Laila Sharmin Sraboni 24th Dec '08

Ophelia Mohaimen
Rezwana Habib Mustazir
Laila Sharmin Sraboni

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

1. INTRODUCTION

1.1. Power supply to a system

The working of the motherboard involves a complex set of power distribution among different levels and components, both simultaneous and separate. Various systems, circuitry and technology offer design options for the current path starting from power plane down to the transistor level. The power delivery network is a critical design choice, comprising of three steps. These are establishing a target impedance, proper system-level decoupling network and selecting the right Voltage Regulator Module.

The Power Supply Unit (PSU) mounted on the motherboard, usually converts 100-120 V (North America and Japan) or 220-240 V (Europe, Africa, Asia and Australia) AC mains supply from the wall outlet to usable low-voltage DC power for the internal components of the motherboard. It has the typical shape of a square metal box, often known as the silver box consisting of diodes, transistors, transformers and capacitors, as well as heat sinks and fans [1].

These are rated based on their maximum output power which typically ranges from 300 W to 500 W. The PSU can be of two types-linear and switched mode. The linear one usually uses a transformer to convert the voltage from the wall outlet to a different, usually a lower voltage; then a rectifier is used to convert it to DC and a capacitor smoothens it. The latter one (SMPS) switches a power transistor between saturation and cutoff with a variable duty cycle supplying the desired output voltage. The resulting rectangular waveform is low-pass filtered with an inductor and capacitor. Though it is smaller and dissipates less energy, it offers a lot more complexity. Yet it is a better option while considering efficiency and heat dissipation. Power delivery network in the motherboard is maintained by steps of conversion from the mains to usable voltage by the chip [1].

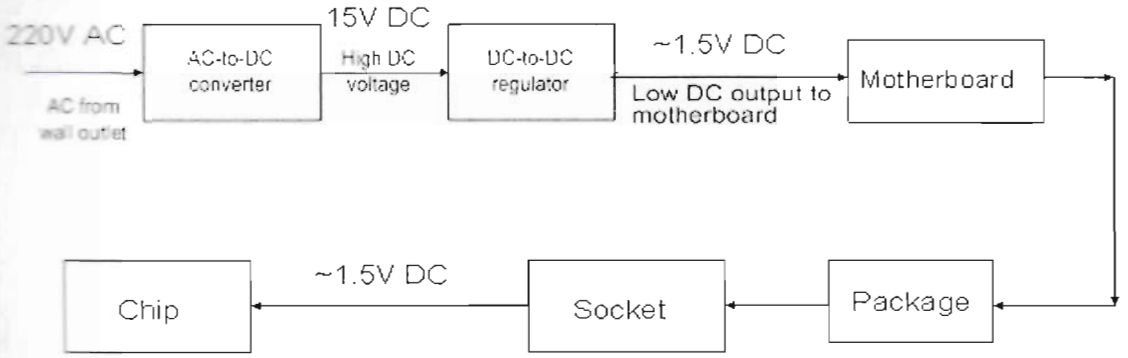


Figure 1: Block diagram of working principle of the PSU

The diagram above shows a flowchart, detailing all the steps of the working of PSU. The Silver Box in the PSU converts 220V or 110V AC to a low-voltage DC, usually to 5V, 12V or 15V. Then this DC is fed into Voltage Regulator Module (VRM) which reduces the voltage to even smaller voltage required for the CPU. The VRM has a high tolerance and cuts down undesirable peaks and troughs. These are actually buck regulators, driven by feedback circuit and helps in maintaining a smooth voltage output. These can again be categorized into Inductor-based VRM, Charge-pump VRM and Linear VRM, depending on the circuit-design. A good VRM design can save up to 18% power. As soon as the Motherboard is powered, current runs through Power Planes in the Printed Circuit Board (PCB). Since PCBs handle with the maximum power among the next components, planes being wider than traces, offer low resistance path. Current gets divided as it travels along the PCB through the Package and Socket, finally reaching the Chip.

1.2. Degradation

After being converted to the required voltage, power is now supplied to the board, package, and socket and eventually to the silicon. The wide power planes in the board offer less resistive drop. The package power distribution design has undergone several changes of pad design from pin grid array to ball grid array and eventually to land grid array. Each of these changes has offered less drop and space but has increased performance speed. Yet, the degradation could not be brought down to zero. The pad layout which determines the placement of traces and vias, the length and width of traces

and vias and the ways in which they are connected to the power and ground planes—all these determine ESL.

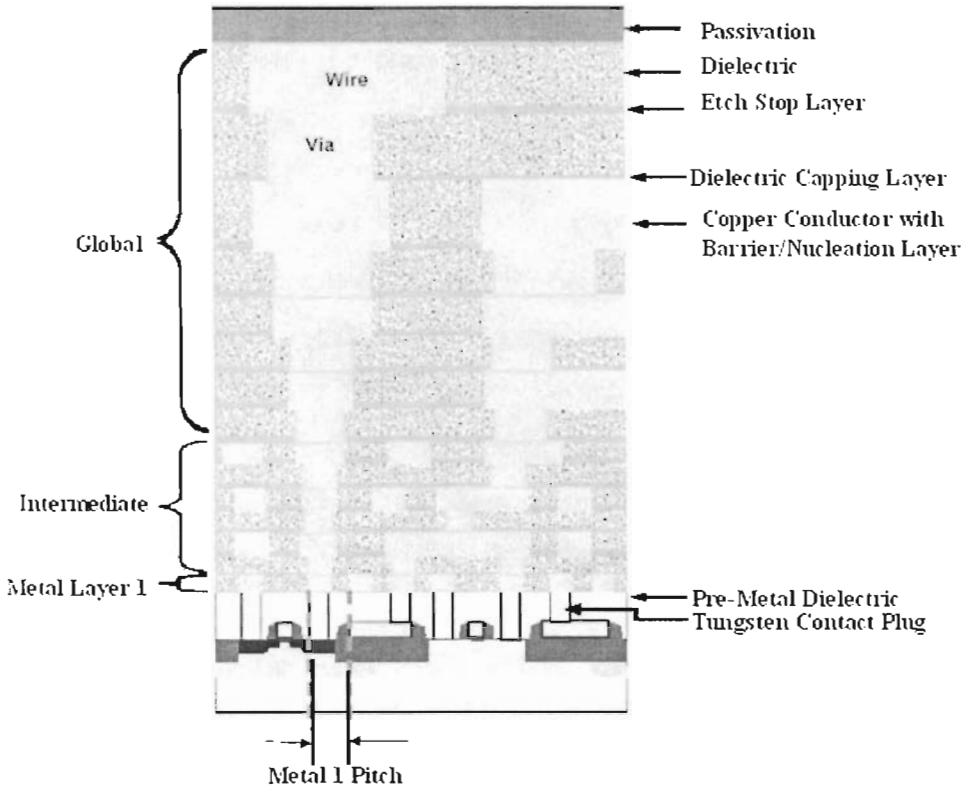


Figure 2: The power path through metal layer, vias and interconnects (reference not available)

Though optimum designs keep ESR and ESL low, power and voltage are degraded all the way from power plane to transistors. The sockets are also placed such that they offer the least resistive and inductive drop. Even in the silicon level, care is taken in bringing down this drop. The power lines are made to travel through the upper metal layers since they have wider and thicker metal traces. The thick dielectric in the upper layers keeps the line-to-line capacitance low. In these layers, the wider via with greater pitch reduces inductive effect. Smaller current loop reduces the ESL contribution. The power traces are kept alternate with the ground in the same metal layer. But, for the second metal layer, the power traces run orthogonally to the first so as to further reduce capacitive coupling between the two. Yet again, the design needs to be kept as simple as possible to minimize the interconnect delay. Even after all these, the droop in power can not be totally made negligible.

Then the interest falls on how to nullify this minimum delay. Large capacitors, storing charge when they get hold of the supply voltage are then made to discharge to smoothen the droop. On-die white area capacitances also help. But both these take time to arrive. By this time, the transistors to switch first see an amount of voltage insufficient to turn them on. Thus the voltage degradation as current runs all the way from the dc supply to the silicon becomes an important issue.

Different technologies are available to help the system provide the necessary power to the circuit. For motherboards and packages, low inductance capacitors are available which are easily integrated into a system. Advanced decoupling technology, embodied in devices such as LICA (Low Inductance Capacitor Array), Monolithic Capacitors (e.g.-X7R,X5R) , Super Capacitors have extremely low internal resistance or ESR and ESL, high efficiency (up to 97-98%), high output power, extremely low heating levels, and improved safety. Currently, the elements of monolithic ceramic capacitors are joined to each other by solder layers and are stacked on each other to minimize the value of equivalent series resistance and equivalent series inductance [2].

Though these are efficient enough to provide power for different stages of power supply, yet the delay in their response time makes it necessary for a nearby decoupling source.

1.3. Current decoupling technology

1.3.1. Capacitance in Motherboard

When a CMOS switches simultaneously and draws short-circuit leakage current, then an undesirable voltage droop causes. The resistive loss and the inductive drop in motherboard, package and socket lower the source voltage which becomes a major issue as it reaches the CMOS inverters. Better designs allow least resistive loss by having power and ground planes in motherboard, greater number of via in package and solder balls, flipped-chip etc. in socket. But due to the space constrains extra capacitance cannot be added, and the increment in supplied voltage level will cause the

heating effect which will create another major problem. In recent days, to minimize the decoupling voltage droops, several processes of providing necessary power are being used. As a consequence, capacitors are being positioned by using the free spaces of motherboard effectively. In nanotechnology, the size of the package is very small compared to the motherboard. On the other hand, for the protection issue of the chip, the size of the motherboard cannot be reduced. So the unused spaces are used to place capacitors and an efficient floor plan allows having large capacitances in motherboard [3].

Capacitances in motherboard are responsible to provide additional power during power starvation. According to the technology adopted in recent days and considering the space constraint of system, these motherboard capacitances are a better idea to minimize the power starvation and can help to gain the desired voltage level but still the voltage droop in the package and socket level exists. So in spite of having large capacitance in motherboard the response time is very slow in CMOS level.

1.3.2. On-die White area Capacitance

On-die White area Capacitance is another approach used in recent technology to compensate the power starvation of CMOS to some extent. On-die capacitances interact with the inductance and resistance of the power distribution network to supply electrical charge. At high frequencies, the supply of decoupling charge is highly localized, and the effective decoupling capacitance is determined primarily by power grid inductance.

The power distribution network is required to have a low impedance resonance-free profile over a wide frequency range. This target impedance is achieved by decoupling capacitances at the board, package, and die levels. At high operating frequencies, the on-die elements must be treated as a distributed system. The on-die power grid exhibits both a resistive and inductive nature [4].

The power grid is typically uniform and symmetric in the X-Y directions on a die capacitances supply locally stored charge for fast current transients. In most of the cases the capacitances are of two types: symbiotic and intentional. Symbiotic decoupling capacitance is provided by existing devices and interconnects within idle logic circuitry. Intentional capacitance is provided by specially designed whose primary role is decoupling. The sum of both types is the on-die decoupling capacitance.

The on-die power network components are greatly dependent upon the high frequency behavior of the impedance; typically about above hundreds of MHz. The complex power distribution network can be approximately calculated by a lumped effective capacitance [5].

The module comprises a plurality of conductive power planes that are separated from a plurality of conductive ground planes by layers of dielectric material. The power planes each have opposite extending tabs that are offset from similar tabs extending from the ground planes and which are coupled together by layers of conductive material. Likewise, the tabs of the ground planes are coupled together by additional layers of conductive material. The corresponding power and ground planes are also coupled together by vias located throughout the module. The conductive layers couple both sides of the corresponding conductive planes and provide contact pads for further assembly to a semiconductive die. The module is attached to the semiconductive die by a plurality of gold bumps which are formed on the top surface of the die.

1.4. Proposition

1.4.1. Location of Decoupling

Typically, the capacitors are used as instant and temporary power suppliers in CMOS devices during the switching of all the transistors at the same time. As the simultaneous switching of billions of transistors draw a significantly large current, the supply voltage suffers from a power insufficiency. Decoupling capacitors are very innovative

concepts to act as miniature power stations for the instant supply of charge to keep the voltage supply constant at this critical period. But sometimes these decoupling capacitors are not efficient enough to serve the intention and designers have to focus on the utilization of the free spaces of motherboard effectively.

In this paper, we have proposed to allocate the decoupling capacitors next to the transistors level using DRAM technology by "Trench Approach". In this approach, a narrow trench which is used only for isolation purpose, adjacent to the CMOS would act as a decoupling capacitor.[1]

1.4.2. Use of DRAM technology

After deciding the position of decoupling, we then set to insert a capacitor in CMOS, but the space constraint becomes a major issue. The problem was solved by switching to Trench DRAM technology. The current aspect ratio of depth is to width is 50:1 enables us in providing a capacitance with the available dimensions. Digging a Trench of 126nm deep in the Silicon Trench Isolator (STI) which is usually 250nm for a 90nm process technology and between two inverters placed back to back does not pose that much difficulty. Thus the merging of DRAM trench in the STI to provide local power proves useful [6].

2. THEORY

2.1. Voltage droops in a real system:

The present invention relates to integrated circuits more particularly to have a stable power supply. Advances in integrated circuit technology have resulted in higher density of devices, faster operating speeds, higher currents, and lower operating voltages. It is necessary to reduce noise on the power supply for optimum operation at maximum speed.

With chips operating faster and supply voltages going lower, noise is becoming an increasing problem for chip designers that cause logical errors and severe timing slowdowns. On-chip noise arises from parasitics inherent in the pins connecting chips to packages. When CMOS devices remain in static state, power consumption is minimal. CMOS dissipates power only when it switches from high to low or vice versa. However, there is a small amount of inductances in the pins connecting the chip to the package. These inductances cause large changes in voltage by creating a reverse EMF followed by the equation:

$$V = -\frac{Ldi}{dt}$$

Moreover, IR drop has a significant effect in the circuit. It occurs in the main supply voltage (V_{DD}) along the power rails between the power pads, the ground pads, channel length and the logic components of a chip.

The simultaneous switching of billions of transistors draws huge currents from the system DC power supply. So, there might be a significant drop on chip when the chip demands current. This inductive drop and Ri drop is sometimes so significant that virtually all the charge used to supply this current comes from chip capacitance. At this time, the speed, at which system is operated, is insufficient to drive the device effectively. Inherent parasitic components and cross-coupled effects of the system are held responsible for a short time delay in delivering power and slower response time to

the circuits. A fast increase in the current can also cause a drop in the supply voltage, since the high rate of change in current is through the package inductance. At this time, the transistors suffer from power starvation followed by voltage droop.

Voltage droop is the loss in output voltage from a device as it tries to drive a load. Although small amounts of voltage droop are unavoidable, there are times when voltage droop is excessive and can are hazardous for unexpected and unwanted outputs in a circuit. For example: simultaneous switching results in sudden flow of current through inductances on to the board that develops a voltage. This creates a potential difference between the board ground and the device, developing a low voltage signal greater than ground level, called "ground bounce". Ground bounce can cause an output low to be seen as a high by other devices on the board.

To reduce the droop, decoupling capacitors are used. Though loss can be compensated in CMOS using decoupling capacitances and minimizing the short circuit current, it can be encountered by reducing the channel resistivity and increasing channel mobility.

2.1.1. Resistive and Inductive Drop

The voltage droop is achieved due to the inductive and resistive circuitry inherent to the mother board, package, socket and Silicon die. These voltage droop can be characterized by the following equation:

$$\Delta V = i(R + j\omega L)$$

However, the impedance faced by the incoming current is reduced by using a power plane rather than through wires and vias.

Decoupling capacitors are used to compensate the inductive drop in any circuitry. Usually when the value of capacitive drop is identical to the value of inductive drops, this noise is minimized. On-chip decoupling is an effective idea but it consumes precious silicon property and leakage current.

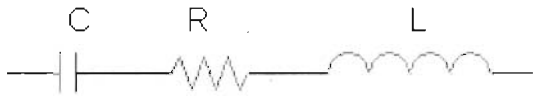


Figure 3: RLC equivalence of the circuit

For the above circuit, we can write, impedance,

$$Z = (R + j\omega L) + \frac{1}{j\omega C}$$

But we have already seen that compensation occurs under this condition,

$$| j\omega L | = | \frac{1}{j\omega C} |$$

That finally yields, frequency,

$$f = \frac{1}{\sqrt{LC}}$$

Here, this is the resonant frequency of the system.

When the switching frequency of a load approaches to the self resonant frequency of the power grid, the voltage drop caused by the RLC system, increases. This phenomenon is known as resonance.

The switching frequency of the current load must not be identical to the resonant frequency of the transistor. It has to be either less or greater than the switching frequency of the system; for at this circumstance, as a result, the voltage response of the overall system oscillates.

The above equation also shows a frequency dependency which is illustrated in the following diagram:

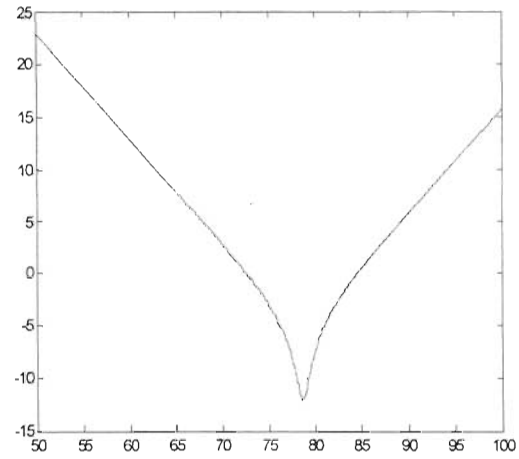


Figure 4: Impedance Vs Frequency plot (Using data from MURATA)

This diagram illustrates that the resonant frequency is between 75 to 80 MHz where the impedance is minimal. That is,

$$Z = R$$

So, the drop occurs during this time is the resistive drop due to the resistivity of wires and vias of the chip.

However, as frequency increases, the planes' characteristics become much more complicated. More precisely, a pair of planes forms a parallel- plate act as transmission-line system rather than acting as capacitors.

In overall view, it might appear that higher operating frequencies are favorable for significant output performance. But unfortunately, the complexity of skin depth arouses in this situation that increases the resistivity of the circuit and the circuit encounters high impedance.

2.1.2. Vdc-IZ Drop

It is essential to keep the impedance low for low voltage fluctuation. As the biasing voltage is shared through the circuit it is necessary to keep it constant for better performance of the system.

The RLC drop that occurs in the transistor level are due to inter connect, vias, wires and the resistivity of the metallic path. The inductive loops created due to the metal layers are the reasons for the inductive drop, Ldi/dt, in the circuit. Vias and wires are nothing but metallic conductors and the channel length of the integrated circuits cause the resistive obstacle of the system. Copper (Cu) is used as the interconnect material with a conductivity of (1:72 '-¢cm) 1. The summations of drops that occur due to such impedance are IZ drops. All these drops add up to a significant number that is responsible for the power starvation of the system supply. For Example: If the system power supply is 5V and the total IZ drop till the transistor level is 1.7V then the voltage shared by the system would be 3.3V, but the system was designed for 5V power supply. Thus the system suffers from power starvation.

To reduce this, power distribution grids with multiple supply voltages are offered. In such designs, multiple supply voltages and multiple grounds are presented through one power distribution grids.

2.2. Signal transmission

2.2.1. Higher frequency trend requires sharper edge

With the advent of technology, the demand for faster RAM, processors, motherboards, etc. has significantly increased with the frequency of signal transmission. Higher frequency in digital signal transmission is nothing but faster switching of 0s and 1s. As the frequency increases, the transitional rate between the 0s and 1s has to cope. The vertical transitions are ideally meant to take no time but in real cases, take very little or negligible time to switch from 0s to 1s or vice versa. When the frequency is increased, the switching rate increases. If the transitional time is comparable with the time the voltage is kept constant, the device can read an error. This becomes more complicated if it deals with a particular type of logic where the result is quite crucial [8].

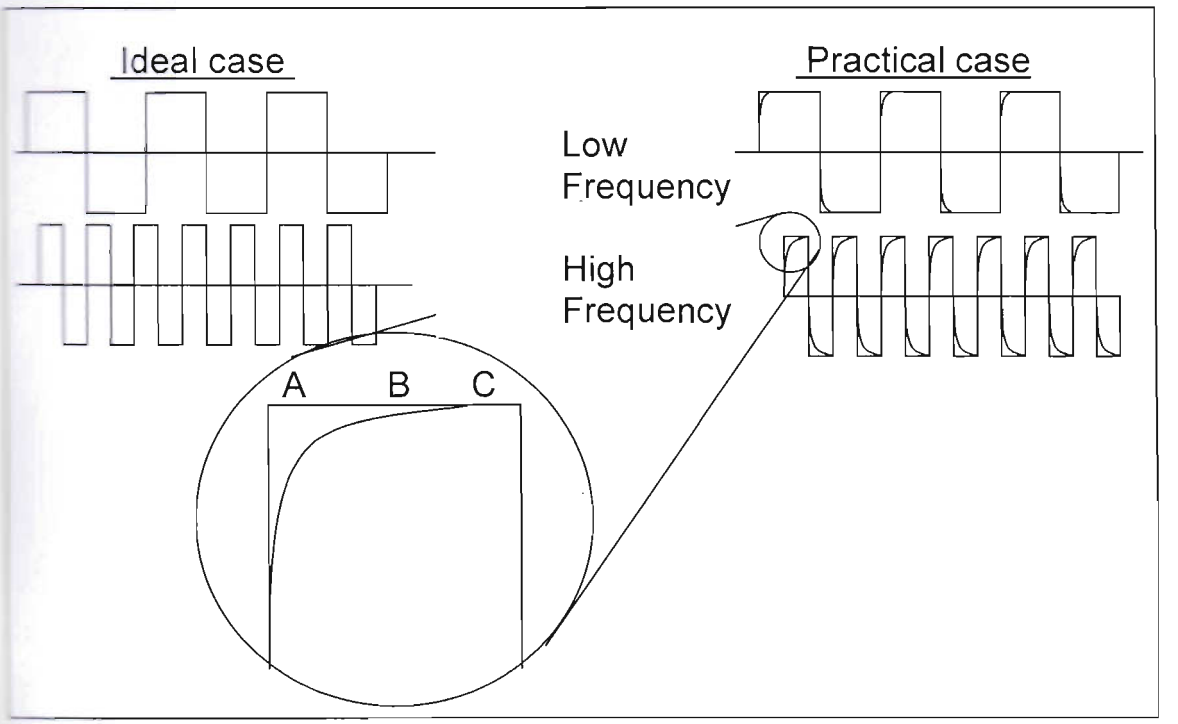


Figure 5: Ideal and real-life digital signal of low and high frequency

For example, if the system is designed to take the reading at the falling edge of the signal, it would read correctly; otherwise not. If it reads at B (as seen in the above figure), it would read a 1 as a 0. Thus transitional time must be very small compared to static and sharper edge becomes mandatory. The transient time in rise and fall time can also be controlled so as to add sophistication to the desired one. This can be done by adjusting the channel length of the PMOS and NMOS, between which the capacitor is connected.

2.2.2. Voltage droop depends on harmonics content of the signal:

When traces and vias are modeled with elementary circuit components, it becomes an RLC circuit. Series RLC circuits with equivalent inductive and capacitive reactances resonate. These circuits create undamped oscillations in voltage and current and produce a high peak at a particular frequency, discriminating others. The highest peak occurs at fundamental frequency.

At odd multiples of this frequency, smaller peaks occur. Since the frequency multiplies, frequency-related problems show up and degrade circuit performance.

Thus at different frequency, different droop is often observed in a system, leading to high instability. Nowadays, many tools are available to analyze and simulate the harmonics of a signal, at various frequency and scale.

2.3. Reactive element in a circuit

In theoretical concept, the capacitor and the resistor are considered as ideal or perfect devices that only contribute to the capacitance and resistance to the circuit. As physical devices are connected to a circuit through conductive leads and paths they contain inherent, usually unwanted, inductance. This means that capacitors are physically confined by some inductive components in addition to their other properties. These are counted as the "reactive components" of the circuit.These reactive components are acknowledged by Equivalent Series Resistance (ESR) and Equivalent Series Inductance (ESL). For example: For a capacitor, ESL is the reactive component in a circuit. This value of ESR and ESL is very small for a single transistor. But in a CMOS circuit where billions of transistors operate at Radio Frequency the concern of ESR and ESL become a very important issue.

An easy way to deal with these inherent inductances in circuit analysis is using lumped element models. Each physical component is expressed as a combination of an ideal component and a small inductor in series carrying an equivalent value to the inductance present in the non-ideal physical device.

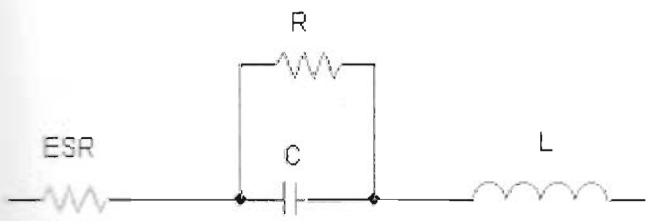


Figure 6: Equivalent series resistance and inductance of a capacitor (PSpice)

2.3.1. Drop from inductance

Typically, lumped modeling is applied for transmission lines but it will help us understand the behavior of inductive loop. Similar metal layers in package, power-ground pin loop or via in socket make up inductive loops. The ground wires are the metal layer return paths of the circuit. The nonlinear variation of inductance with circuit length is a result of inductive coupling among circuit segments.

The power supply is located some distance away from the transistors it drives. When the logic pulse arrives, switching occurs in transistors which pull the energy along the entire path length from power supply down to the transistor level. Resistive losses in the form of heat and losses due to the current changes to the path inductances are serious issues. Moreover, if the distance between the forward and return current path is smaller than the loop length the inductive coupling is negligible.

The decrease in the RC time constant of a circuit will cause the switching current to race quicker, resulting in a large di/dt. From one perspective this is good as it will increase the switching speed, which is the ultimate goal. However, the faster switching speed translates into higher harmonics in the signal. In time domain the large di/dt translates into large voltage drop across any inductance that may lie in the path of the current.[9]

2.3.2. Power from capacitance

The coupling of energy from one line to another via produces capacitive coupling or cross talk provided they have potential difference in them. The circuit element that represents this transfer of energy is followed by these equations:

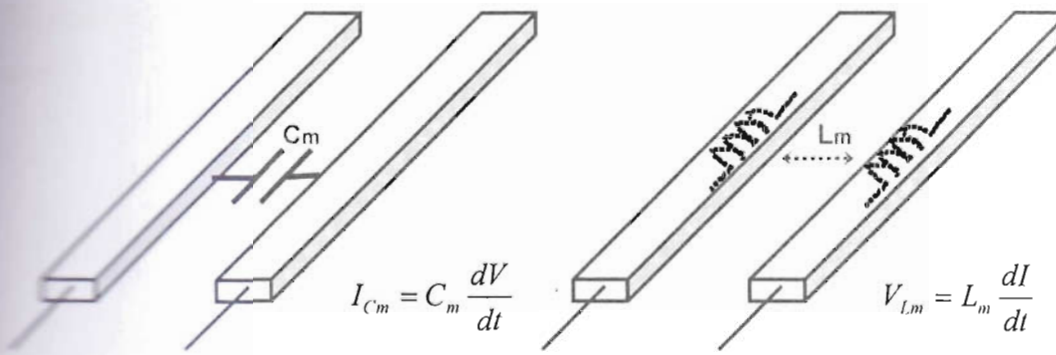


Figure 7: Capacitive coupling due to adjacent via and traces

The mutual inductance will induce current on the victim line opposite of the driving current (Lenz's Law) and the mutual capacitance will pass current through the mutual capacitance that flows in both directions on the victim line.

In typical CMOS devices, the potential difference between the V_{dd} and the ground is very high and air between them act as a very good dielectric. So they turn into a capacitance.

The metals layers in PCB have Si inside them which also act as dielectric and hence there is a good capacitive effect inside the transistors. The dielectric constant of Si is 11.9 which is quite a bigger value.

Better system designs target low resistive loss by creating large power and ground planes in motherboard and packages, to reduce inductance and resistance. It includes large number of vias in packages and connectors. A large section of solder ball and pin grid arrays in packages are devoted to ground and power to reduce the average resistive and inductive impedance.

2.3.3. ESR, ESL

A theoretically perfect capacitor would be loss less and have an ESR of zero. But in single capacitance the value of these ESR and ESL are very small. These are not fixed values and they are the inherent properties of any capacitor. Typically, the values of

ESR and ESL are in several μ-Ohms and they cannot be measured by normal Ohm-meter.

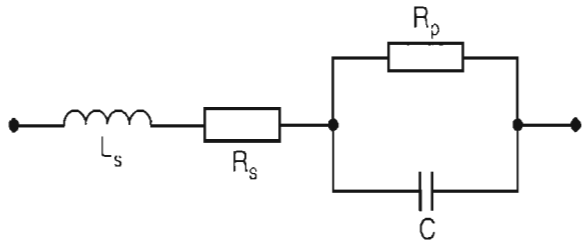


Figure 8: ESR and ESL content of circuit elements (PSpice)

It is critical to minimize the equivalent series resistance and inductance (ESR, ESL) of the DRAM capacitors. Large pitch between metal layers connecting the capacitor to the V_{dd} line will results in a high ESL, while too small a cross section of the trace will result in high ESR. The ESL and ESR of decoupling capacitor show significant degradation with frequency. At higher operating frequencies, the ESR of decoupling capacitor increases due to skin effect. The equivalent impedance of the capacitor is

$$|Z| = \sqrt{R_{esr}^2 + (\omega L_{esl} + \frac{1}{\omega C_c})^2}$$

Where, $\omega = 2\pi f$.

The impedance of the capacitor is minimal at resonance.

The operating frequency range of the capacitor should fell near the resonant frequency.

At high frequency, skin effect becomes significant. Most of the high frequency currents are crowded within the skin depth of the wire. For systems operating at GHz frequencies, the significant components that define the signal transition edges can be well into the 30-40 GHz. In this paper, skin depth is calculated for 40GHz for the realization of high frequency effect.

$$\delta = \frac{1}{\sqrt{\pi f \mu \sigma}} = \sqrt{\frac{1.7 \times 10^{-8}}{\pi \times 40 \times 10^{9} \times 4\pi \times 10^{-7}}} = .33\,\mu m$$

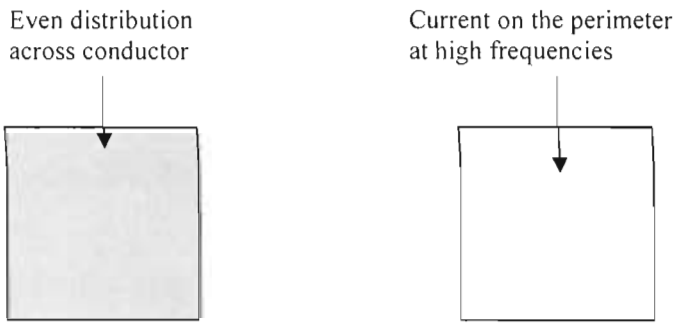


Figure 9: Skin depth constraint at high frequency

The Murata Software helped us comprehend that at lower frequencies the capacitive component is significant ($1/j\omega C$) and in higher frequency trend the inductive components prevail. For a particular frequency, there is no capacitive component or inductive component and impedance is equivalent to only resistive part .Typically, we try to operate our devices at this frequency.

3. TRENDS IN TECHNOLOGY

3.1. DRAM

Dynamic random access memory (DRAM) is a type of random access memory that stores each bit of data in each elementary DRAM cell. A cell contains a single MOS transistor and a storage capacitor. This charge, however, leaks off the capacitor due to the sub-threshold current of the cell transistor; the information eventually fades unless the capacitor charge is refreshed periodically.

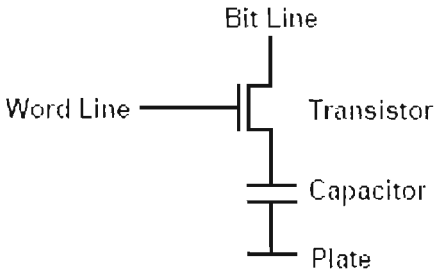


Figure 10: DRAM cell placement

The memory cell is written to by placing a "1" or "0" charge into the capacitor cell. This is done during a write cycle by opening the cell transistor (gate to power supply or V_{CC}) and presenting either V_{CC} or 0V (ground) at the capacitor. The word line (gate of the transistor) is then held at ground to isolate the capacitor charge.

A great deal of design effort has been made to shrink the cell area, particularly, the size of the DRAM capacitor. As memory density increases, the cell size must decrease.

Designers have managed to shrink overall cell size. However, due to factors such as noise sensitivity and speed, it has been a challenge to reduce the capacitance.

Different DRAM elementary cell includes Planar, Trench, Stack, Cup, Crown etc technologies. Compared to all other techniques mentioned above, the trench cell offers

a large storage capacitance, a mask less self-aligned contact, comparatively low dielectrics, simpler circuitry and last, but not the least, it consumes the least space [10].

A rod-like metallic structure sinks down the trench. Rest is filled with dielectric. New innovations including Checkerboard (CKB), hemispherical silicon grains (HSG) combined with the use of a bottle-shaped trench, increase the surface area of the trench capacitor and thus the capacitance. The surface area can be further increased by making the wall of the trench rough.

According to ITRS, 2007 Edition, MPU/ ASIC Metal1 half pitch falls to 12nm in 2022 from 68nm in 2007. The DRAM half pitch falls from 65nm, in 2007 to 11nm in 2022. This paper says that the fabrication of DRAM has undergone and would continue to do so in near term years:

1. DRAM ½ pitch is 65nm and would decrease further in near term years.
2. The aspect ratios of the spaces between adjacent gates in DRAMs are expected to be greater then 16:1 by 2007 and will increase thereafter
3. Low-κ dielectrics will be required in DRAMs to reduce capacitance in the layer incorporating bit lines
4. ALD is needed for high aspect ratio in DRAM
5. DRAM technology moves to high k metal gates

Compared with other memory ICs, DRAMs suffer from a speed problem. The on-chip circuitry required to read the data from each cell is inherently slow. As such, DRAM speeds have not kept pace with the increased clock speed of CPUs.

To face this speed discrepancy, DRAMs have branched into many sub-categories. Each features a variation of system interface circuitry with the intent of enhancing performance. Furthermore, each design attempts to answer needs of specific applications.

3.2. Semiconductor trend

Describing a long term trend of the history of computing, Intel co-founder Gordon Moore had made a revolutionary prediction in 1965. His prediction, popularly known as Moore's Law, states that the number of transistors on a chip will double about every two years.

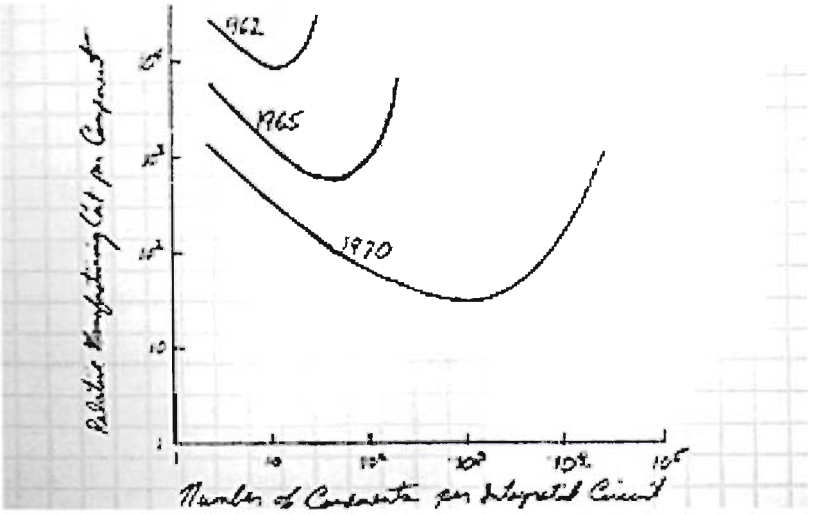


Figure 11: Gordon Moore's original graph from 1965

Though Moore's law was made from an observation, as it got more widely accepted the industry followed it as a goal. In spite of the falling consumer price, the industry kept on adopting new expensive methodologies to pace up with technological advancement. From four transistors in 1954's transistor radio to 820million transistors in Quad-Core Intel® Xeon® processor of the year 2007, manufacturers have accelerated the change in semiconductors. Computer industry technology road maps predicts (as of 2001) that Moore's law will continue for another several chip generations [11].

3.2.1. Higher frequency operation

Another term that paced up advancement in semiconductors is frequency which is the number of occurrences of a repeating event per unit time. The standard unit of frequency is the hertz, abbreviated Hz. Larger units of frequency include the kilohertz, megahertz (MHz), gigahertz (GHz) and terahertz (THz) [1].

The whole motherboard of a personal computer is run by a central clock known as the system clock. Frequency of the system clock is multiplied many times by the PLL (phase locked loop) of a chip and fed into a microprocessor, its adjoining units and the cache. There is usually a single PLL in a chip and this frequency of operation is maintained in the working of the whole processor chip. The higher the frequency the faster the work is done by the processor [8] .

The Altair 8800 was the first successful home and personal computer in 1975. Since then, manufacturers developed microprocessors with increased functionality and the advancement in frequency came in parallel. 60MHz processors in 1993 have accelerated to more than 3GHz in 2007. The manufacturers have coped up with problems associated with high frequency operation such as resonance, overheating, skin depth etc. Some of the common solutions are damped circuits, switching off part of the processor and robust designs [11].

3.2.2. Scaling down

In trying to pace up with Moore's law and increased density of transistors, their size has gone down at an amazing rate. 1 micrometer (μm) process introduced in 1987 has shifted to 0.3 μm in 1997 and accelerated to 45nm in 2007. This aggressive scaling has brought in new technologies, innovative techniques, fabrication processes and even an extension of particle physics. The benefits of scaling include improvement of packing density, faster speed and lower power dissipation [11].

The process technology is actually defined by the channel length of the MOSFETs used in that process. For instance, a 45nm process incorporates MOS of channel length 45nm; rest of the dimensions is scaled down accordingly. The R&D of the MOS industry is even a step ahead of what they produce. When they market a 45nm process, they would actually be working on a smaller scale [12].

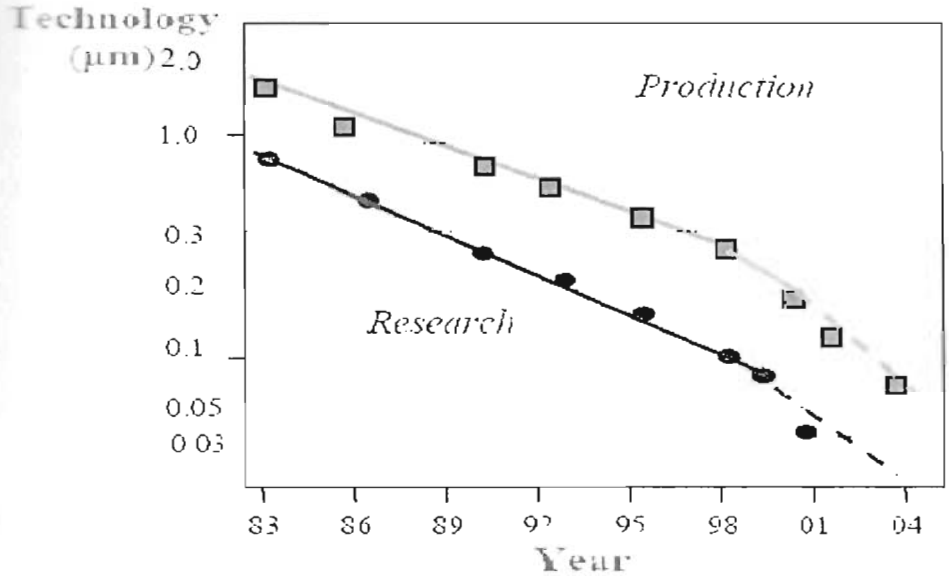


Figure 12: Difference in R&D and production

A key concept in scaling is that the various Structural parameters of the MOSFET should be scaled such that the device functions properly. If short-channel MOS are not scaled proportionately, unintentional electrostatic interactions may occur between the source and drain known as DIBL (Drain Induced Barrier Lowering) effect [13].

Table 1: Scaling factors of different Structural Parameters

Structural parameters	Scaling Factor
Surface Dimension	1/K
Vertical dimensions	1/K
Current, voltage	1/K
Capacitance	K
Circuit delay time	1/K
Power dissipation	$1/K^2$

There are many problems associated with this scaling. As devices are scaled down, internal electric field is to be kept constant for each generation chips. Hot carrier effect becomes more prominent with shrinking. However, the use of high-k materials provides a solution to that.

Yet, MOS are being scaled down, trend going faster after each generation. In 1995, the 0.5 micrometer process had three layers of interconnect for a 120 MHz signal. 0.18 micrometer process of the year 2000 had seven layers for 500MHz signal. In the next two years, devices were scaled down to 0.12 micrometer process and the frequency rose to 1500 MHz. this trend continued and devices were scaled down faster than before while keeping down voltage and space; and increasing layers of interconnect and frequency [11].

3.2.3. Decoupling technology

Small deviations from the ideal behavior of a device can become significant when circuits are operating under 'fast corner' conditions, i.e. high frequency, high current, or temperature extremes. However, all these low ESR/ESL technology can only serve to compensate the power droop due to motherboard, package, and connector level impedance. These new decoupling technologies are not enough to the increasing frequencies and large current switching. Challenge remains in trying to provide for decoupling at the integrated circuit level.

Different technologies are available to help the system provide the necessary power to the circuit. For motherboards and packages, low inductance capacitors are available which are easily integrated into a system. Advanced decoupling technology, embodied in devices such as LICA (Low Inductance Capacitor Array), Monolithic Capacitors (e.g.-X7R,X5R) , Super Capacitors have extremely low internal resistance or ESR and ESL, high efficiency (up to 97-98%), high output power, extremely low heating levels, and improved safety. Currently, the elements of monolithic ceramic capacitors are joined to each other by solder layers and are stacked on each other to minimize the value of equivalent series resistance and equivalent series inductance.

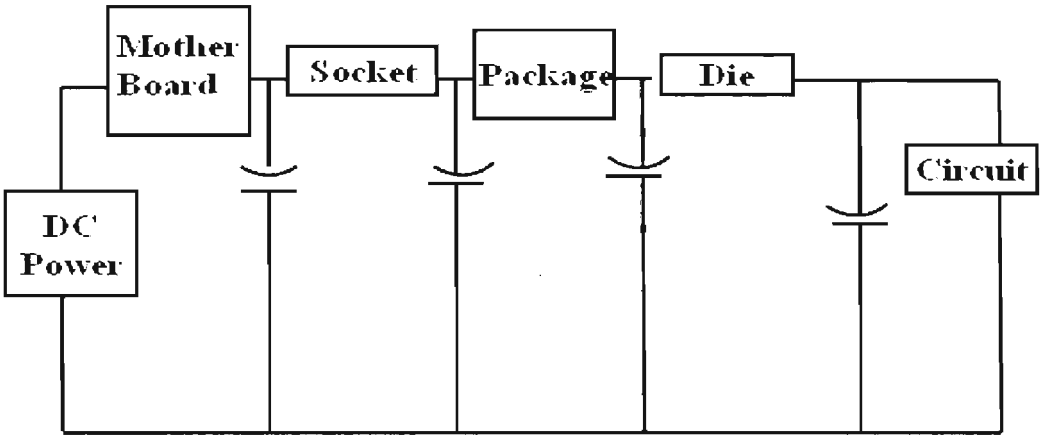


Figure 13: System level power delivery decoupling solution.

3.3. Fabrication Process

To improve the power delivery to the inverter level circuitry we had focused over two fabrication processes. First of all to introduce a capacitor between the p-MOS and n-MOS we have to go through the CMOS fabrication process and then as we are proposing the capacitor will be deposited in the inverter according to the process that is followed in case of fabricating the DRAM, so the DRAM fabrication process has to be included here. Thus the combination of these CMOS and DRAM fabrication process will be the complete proposed process that could be followed for creating the localized decoupling.

To improve the power delivery to the inverter level circuitry two fabrication processes have been taken under consideration. First to introduce a capacitor between the p-MOS and n-MOS CMOS fabrication process is considered and then DRAM fabrication process has been included to work with the details of the DRAM trench. Thus the combination of these CMOS and DRAM fabrication process will be the complete proposed process that could be followed for creating the localized decoupling.

3.3.1. CMOS Fabrication

CMOS fabrication process has several steps and all those steps are again divided into several phases. Again different fabrication companies follow different methods accordingly. Here in our proposal we will focus on the major fabrication phases that are widely used. For basic processing step we will look up in layering, patterning, doping, and masking for n-well CMOS fabrication process.

Basic Processing Steps:

Lithography:

Lithography is the process used to transfer a pattern to a layer on the chip. Each processing step requires that certain areas are defined on chip by appropriate masks. Consequently, the integrated circuit may be viewed as a set of patterned layers of doped silicon, polysilicon, metal and insulating silicon dioxide. In general, a layer must be patterned before the next layer of material is applied on chip [15].

Thermal Oxidation:

The sequence starts with the thermal oxidation of the silicon surface, by which an oxide layer of about 1 micrometer (1000 nm) thickness, for example, is created on the substrate. This oxide layer is also called field oxide.

Deposition of Photoresist:

The entire oxide surface is covered with a layer of photoresist, which is a light-sensitive, acid-resistant organic polymer, initially insoluble in the developing solution. The photoresist material is exposed to ultraviolet (UV) light, the exposed areas become soluble so that then they are no longer resistant to etching solvents [16].

Mount Mask above Si:

To selectively expose the photoresist, some of the areas are covered on the surface with a mask during exposure. Thus, when the structure with the mask on top is exposed to UV light, areas which are covered by the opaque features on the mask are shielded. In

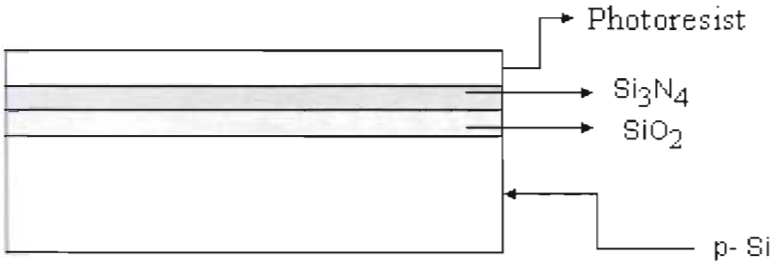
the areas where the UV light strikes the photoresist, it is "exposed" and becomes soluble in certain solutes.

Etching:

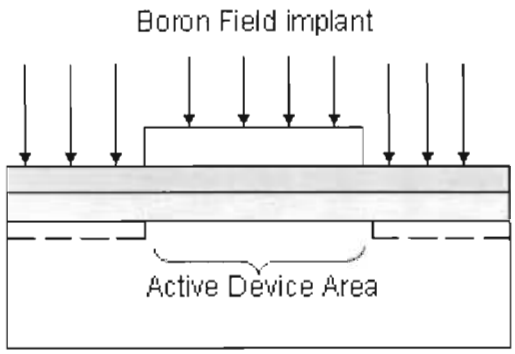
Now, the silicon dioxide regions which are not covered by hardened photoresist can be etched away either by using a chemical solvent (HF acid) or by using a dry etch (plasma etch) process. At the end of this step, an oxide window is obtained that reaches down to the silicon surface. The remaining (unexposed part) photoresist can be stripped from the silicon dioxide (SiO_2) surface by using another solvent, leaving the patterned silicon dioxide feature on the surface [17].

Diffusion:

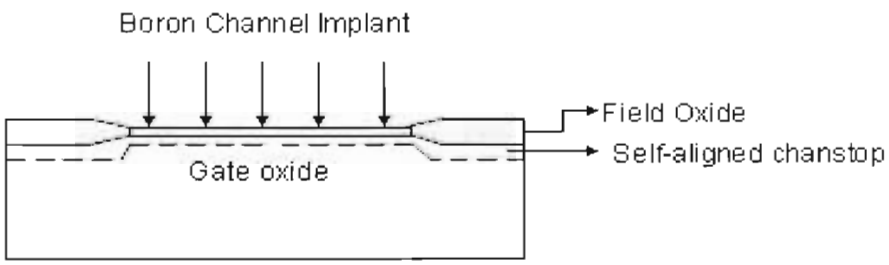
Several diffusion processes are adopted during different deposition phases of CMOS fabrication process. Thermal diffusion is one of the main. Thermal diffusion of dopants in furnace, where first of all oxidization is done and the windows are opened in the oxide using the photolithography and etching steps described earlier. The dopants are gradually transported from the high concentration region near the surface into the substrate. The diffusivity of dopants in solid has a strong dependence on the temperature [10].



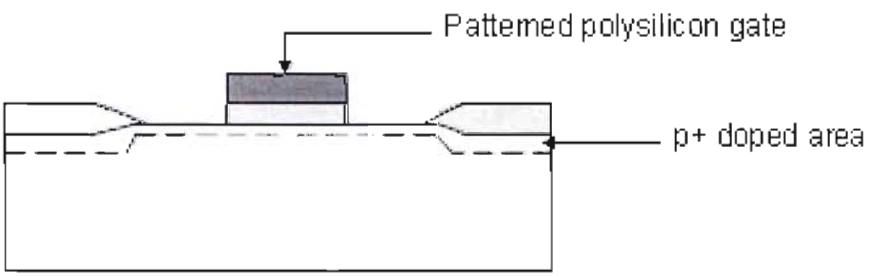
Formation of SiO_2, Si_3N_4, and photoresist layer.



Boron implant



Formation of Field Oxide



Formation of gate

Figure 14: Fabrication steps

Grow the gate oxide:

The composite nitride-oxide layer over the active device area is removed and a thin gate oxide layer (<10nm) is grown. Adjustment of threshold voltage is another key step here. Two kinds of *n*-channel devices :

- For enhancement-mode *n*-channel device boron ions are implanted in the channel region to increase the threshold voltage (~ +0.5V).
- For depletion-mode *n*-channel device arsenic ions are implanted in the channel region to decrease the threshold voltage (~ -0.5V).

Formation of gate:

A polysilicon is deposited and heavily doped by diffusion or implantation of phosphorus (for gate length >3μm). For smaller devices a composite layer of metal silicide and polysilicon named polycide can be used. The use of polycide as a gate material of MOSFET can reduce the sheet resistance (~ 1Ω/m). The gate silicon is properly etched which exposes the bare silicon surface on which the source and drain junctions are to be formed [18].

Formation of the source and drain:

After gate is patterned it will act as a mask for the arsenic implantation to form source and drain. The entire silicon surface is then doped with a high concentration of impurities, either through diffusion or ion implantation. This source and drain are self-aligned with respect to the gate. To minimize lateral diffusion, low temperature fabrication processes are used. In that case the parasitic gate-drain and gate-source coupling capacitances can be much smaller than the gate-channel capacitance [19].

Metallization:

The metallization process followed by a phosphorus-doped oxide (P-glass) is deposited over the entire wafer followed by the flow of heat that is through evaporation to form a smoother surface. Contact windows and interconnection pattern are selectively defined and etched. A metal layer is deposited and patterned. The gate contact is usually made outside the active device area to avoid damage to the thin gate oxide [20].

After the interconnection metallization is completed, a protective overcoat of Si_3N_4 is deposited. Then the individual integrated circuits can be separated by sawing and breaking the wafer. This phase of device fabrication is called back-end processing.

The fabrication of semiconductor devices requires several such pattern transfers to be performed on silicon dioxide, polysilicon, and metal.

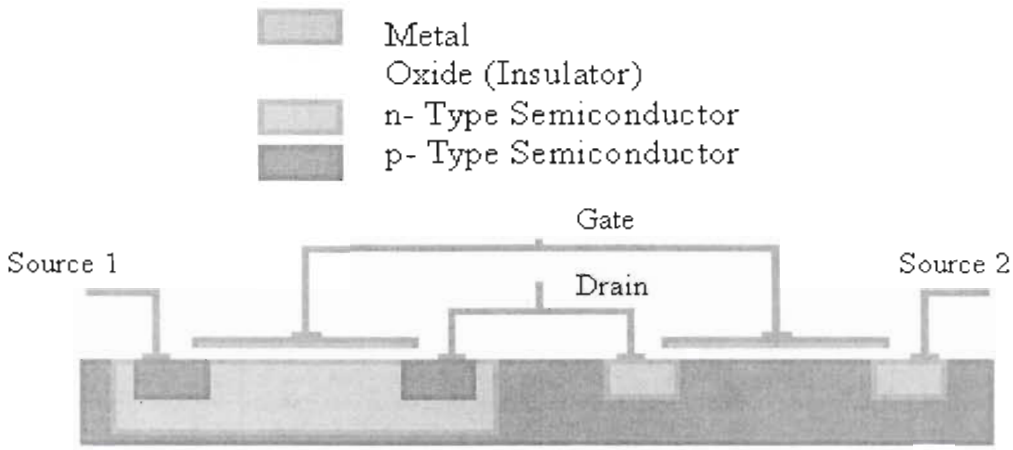


Figure 15: n-well CMOS integrated circuit

Fabrication sequence of n-well CMOS integrated circuits:

N-Well Implant and Drive-in Diffusion

The n-well CMOS process starts with a moderately doped (impurity concentration ~1016/cm3) p-type silicon substrate. Then, an initial thick "field" oxide layer (5000A) is grown on the entire surface. The first lithographic mask defines the n-well region. Donor atoms, usually phosphorus, are implanted through this window in the oxide. Once the n-well is created, the active areas of the nMOS and pMOS transistors can be defined.

Thick Field Oxide Growth

Following the creation of the n-well region, a thick field oxide is grown in the areas surrounding the transistor active regions, and a thin gate oxide is grown on top of the active regions. The thickness and the quality of the gate oxide are two of the most critical fabrication parameters, since they strongly affect the operational characteristics of the MOS transistor, as well as the long-term reliability of the CMOS.

Before the next process step that is deposition of polysilicon layer, by masking the wafer, the channel regions can be separately implanted. This extra step will allow adjustment of the threshold voltage for both the V_{tn} and V_{tp}.

Deposit Gate Poly Si Layer and Etch

The polysilicon layer (3000A) is deposited using chemical vapor deposition (CVD) and patterned by dry plasma etching. The created polysilicon lines will function as the gate electrodes of the nMOS and the pMOS transistors and theirs interconnects [21].

Implant n-Channel n+ Regions

Using a set of two masks, the n+ and p+ Source and Drain regions are implanted into the substrate and into the n- well, respectively. The ohmic contacts to the substrate and to the n-well are implanted in this process step. If a doped silicon region is partially doped to >1018/cm^3, then metal contacts to that volume are almost always ohmic (no Shottkey Barrier effect). The possibility of a Shottkey Barrier effect is always a problem, and to avoid this care must be made of the selection of doping and metal contacts. The gate polysilicon will act as mask and thus alignment of gate and source and drain is automatic (which is described as self-aligned process).

Implant n-Channel p+ Regions

After the p^+ source and drain regions of p-MOS transistor is defined and masked, p-type impurity atoms of Boron (B) is implanted. Here the polysilicon layer protects transistor channel regions from the boron dopants. Moderate temperature drives the impurities deeper into the substrate. Some repairing can be made for some crystal structure damage. The use of lateral diffusion under the gate can cause the overlap in capacitance. The deposition of insulated layer using CVD technique causes non-planar surface which becomes an issue in metallization phase.

Deposition of CVD SiO_2 Layer

An insulating silicon dioxide layer is deposited over the entire wafer using CVD (5000A). This is for passivation, the protection of all the active components from contamination. The contacts are defined and etched away to expose the silicon or polysilicon contact windows. These contact windows are necessary to complete the circuit interconnections using the metal layer, which is patterned in the next step [22].

Chemical Vapor Deposition, where reactive gases collide above the wafer, and chemical reaction products then fall onto the wafer creating a new layer. Abrupt steps are smoothed over with deposited insulating layer.

Open Contact Cuts

Before metallization the open contact cuts in the insulating layers is made to establish the contact area. The most important part is to make the contacts to polysilicon outside of the gate region. Here the metal spikes through the polysilicon and the thin gate oxide is avoided through using the contact mask.

Deposition of Aluminum (Al) layer as metal contact

The semiconductor devices have to be connected with each other and with the IC package through metallization. Metal film (aluminum, >5000A) is deposited over the entire chip surface using metal evaporation or sputtering, and the metal lines are patterned through etching. Since the wafer surface is non-planar, the quality and the integrity of the metal lines created in this step are very critical and are ultimately essential for circuit reliability.

Since the metal connects two separate devices, it is called Local Interconnect. The connection of adjacent devices is often called LI-1, as being the lowest level of interconnection.

Final Circuit - CMOS Inverter

The final step is to deposit a full SiO2 passivation layer (5000A), for protection, over the chip, except for wire-bonding pad areas. If the wafer will be stored for some months, a final thin blanket layer of Si_3N_4 may be applied to prevent penetration by water vapor. Completed FEOL wafers are sometimes stored for more than a year before processing in a BEOL factory [13].

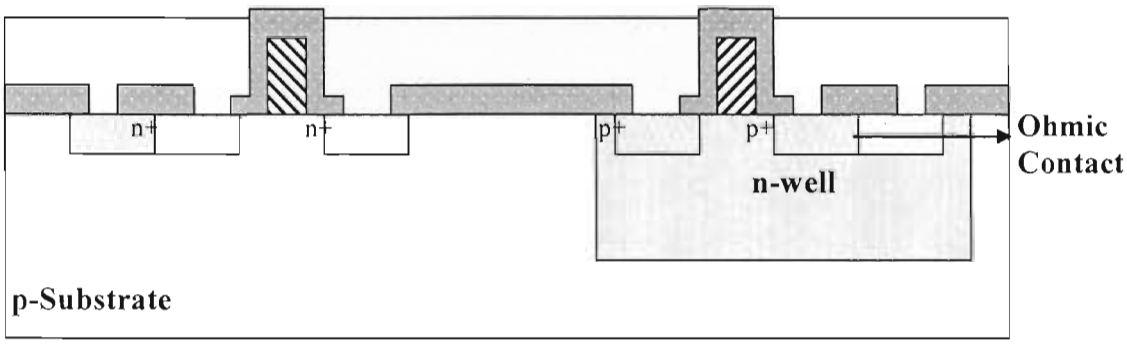


Figure 16: CMOS inverter circuit

3.3.2. DRAM fabrication

There are several types of DRAM fabrication process depending on the types of DRAM. As we discussed earlier that mostly stack and trench type DRAM fabrication and implementation is popular. The fabrication process includes diffusing, layering, patterning, etching, CVD (Chemical Vapor Deposition), ion implementation process and doping like the CMOS formation. But the main problem occurs as the fabrication of DRAM has to be done parallel to the CMOS creating trench or stack [23] . The formation process can be shown through the diagrams of different phases:

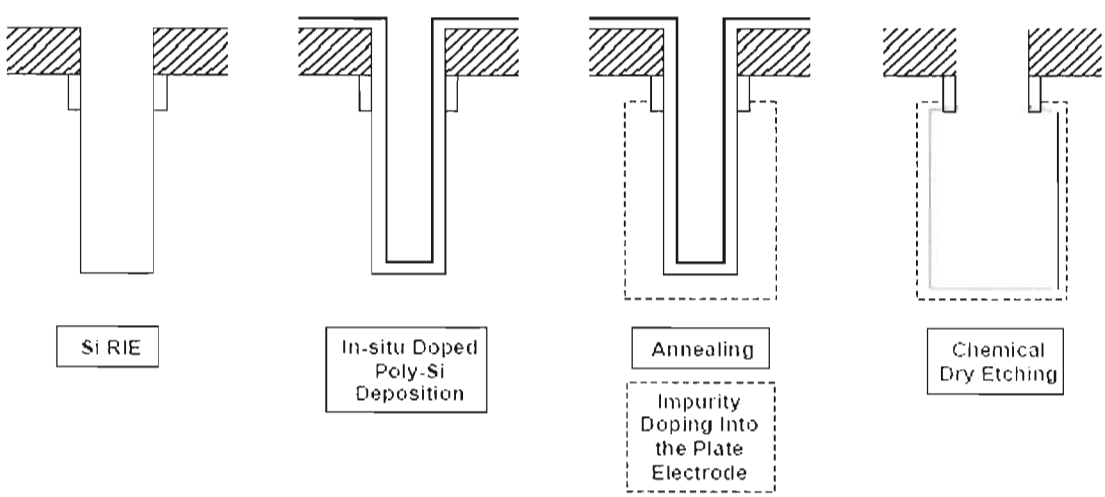


Figure 17: Fabrication process of DRAM

Most leading DRAM manufacturers are working on 1Gbit cells. To reach the nanotechnology, their goal is to decrease the size of the cell as small as possible without compromising the value of the capacitor.

Two types of 1Gbit cell developments are manufactured according to the available technology. First one is through improvement in the trench capacitor concept by creating a bottle-shape trench design. And the second on is through improvement in the stack concept with a vertical and circular capacitor.

The fabrication steps are briefed as: first well formation steps are done where a p-well is formed. In order to form the p-well Boron atoms are implanted with high acceleration energy of about 180KeV. Followed by anti punch through implant, Boron was implanted with 80Kev acceleration energy. The implantation dose changes in order to vary the different boron concentration required for the p-well. Then a bottom oxide layer is formed on the p-substrate and two semi conductor layers are formed one by one. The first semi conductor layer is formed on the bottom oxide layer and a second semiconductor layer is formed on the first semiconductor layer. Then an electrode stack structure is formed on the second semiconductor layer. Afterwards, electrode stack structure, the second semiconductor layer and the first semiconductor layer are patterned. Subsequently an insulation layer is formed at two sides of the first semiconductor layer and the second semiconductor layer. The height of the insulation layer is greater than that of the first semiconductor layer. After, a doping layer is formed on the insulation layer. Then the dielectric material with a higher dielectric constant has to be deposited here accordingly to create the capacitive effect which will store the charge to store the data in the form of high and low. The formation of dram ends up with the creation of contact through which the capacitor will be charged and discharged [23].

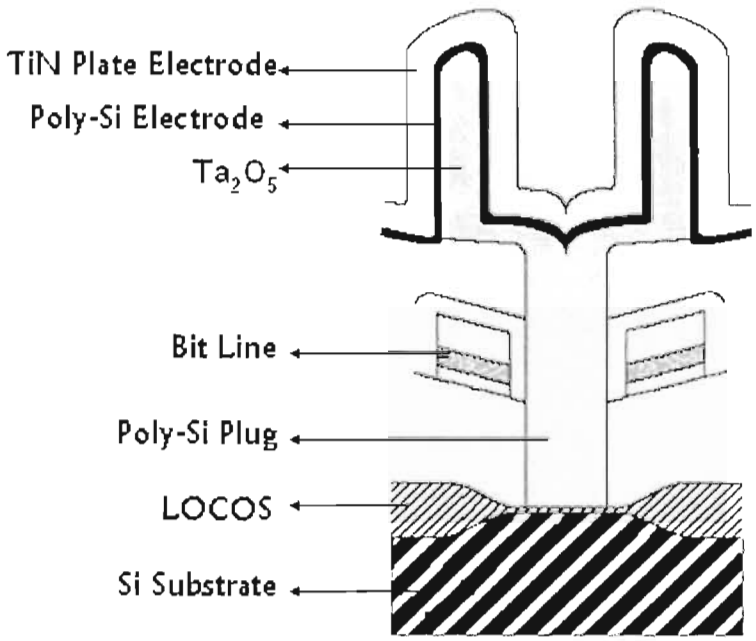


Figure 18: Structure of DRAM

The above figure illustrates the structure of a DRAM following the recent fabrication process. Here polisilicon and titanium-nitride electrode has been used and Tantalum pentoxide which is also known as tantalum(V) oxide used as dielectric element between two electrodes.

4. PROPOSED LOCALIZED DECOUPLING

4.1. CMOS inverter

CMOS stands for Complementary metal–oxide–semiconductor and is the short for complementary MOSFET. A MOSFET is the most common electronic device, where FET stands for Field –Effect Transistor. It is used to switch or amplify digital and analog signals. The word "complementary" refers to the presence of complementary pairs of n-type and p-type MOSFETs [1].

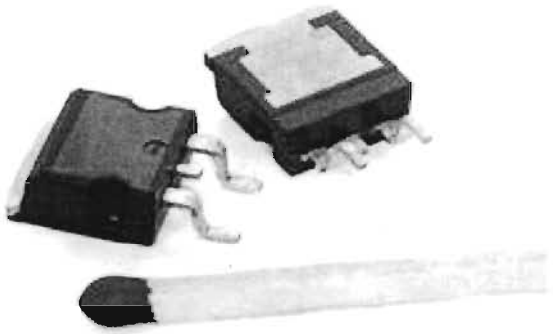


Figure 19: Typical size of CMOS as compared with a matchstick (courtesy: Wikipedia)

Inversion of Input:

The inversion of signals implemented in digital circuitry of computers can be done in a number of ways. But, CMOS uses the least space and power in doing so. This promoted scaling down a lot. When the input is low, the output is high and vice versa.

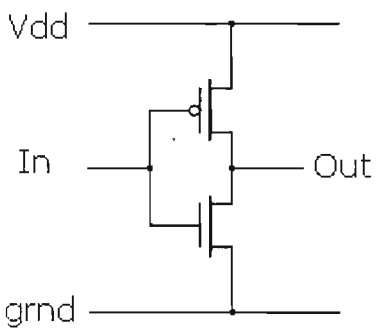


Figure 20: CMOS inverter

When a HIGH is given to "In", the PMOS (upper transistor) poses a high resistivity and the supply voltage Vdd is blocked from the output. With this HIGH, the NMOS offers a less resistance, draining whatever was in output. Thus, the output is LOW, inverting the HIGH input.

A Brief Insight:

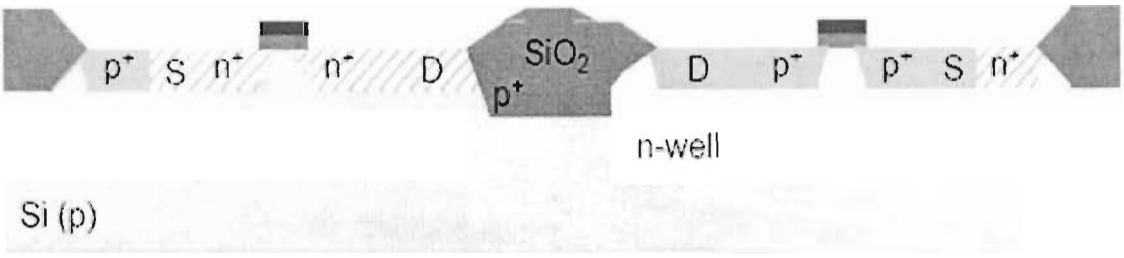


Figure 21: Cross-section of CMOS

The CMOS is usually built on a p-substrate and an n-well is dug for a NMOS. The Source of the PMOS is connected to the supply and the source of the NMOS is grounded.

Both the MOS have diffused regions for Drain and Source, p^+ diffusion in PMOS and n^+ in NMOS. The PMOS has a p-type channel, over which lies a thin pad of dielectric. Thin, due to the scaling, but thick enough to stop hot carrier effect which is the escaping of electrons through the dielectric under the influence of electric field. A metallic gate is built over it to bias the device for activation. Biasing current is very crucial, the device is cut-off if bias is less, pinched off if more. The channel width and length are also crucial since it has to deal with the mobility of carriers and process technology. The PMOS needs to have twice the channel width as that of NMOS to have the same mobility [13].

The CMOS dissipates power only when it switches. The short circuit current flows momentarily through the two MOS when both are on. This is denoted as the dynamic power dissipation. When the CMOS is on, it dissipates static power, which according to

the following graph is much lower than that of dynamic power. Low power design can save a lot of power.

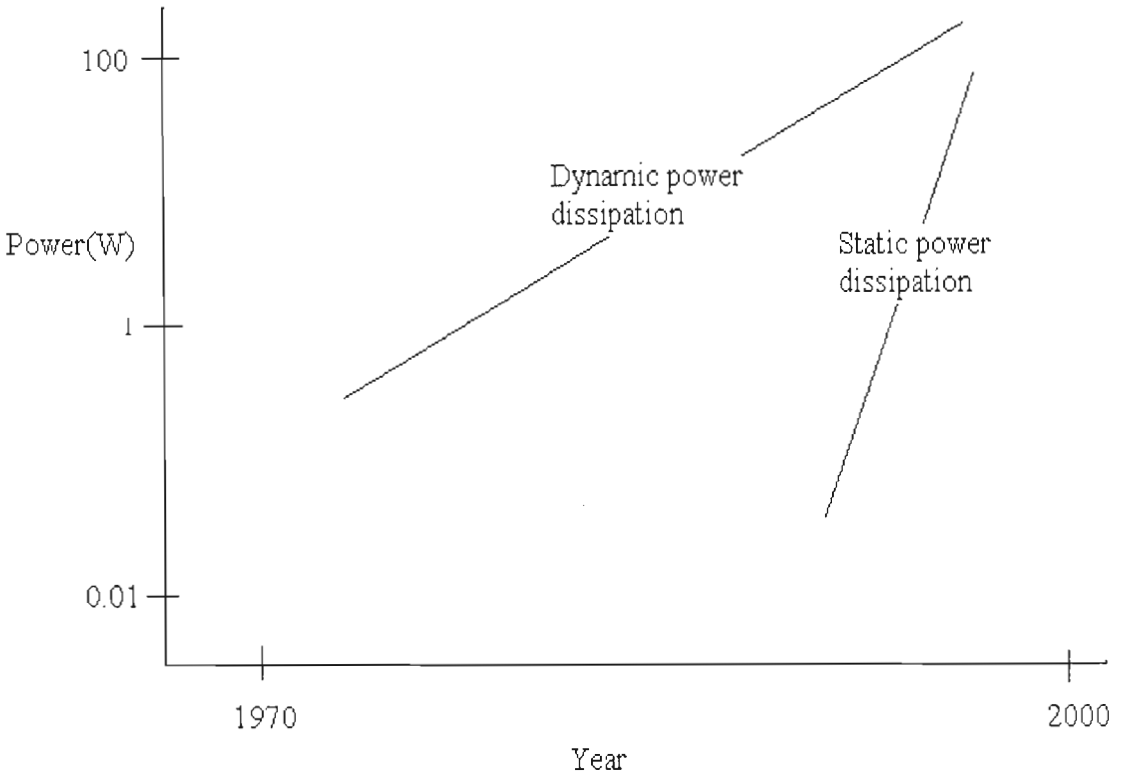


Figure 22: Dynamic and static power trends

The PMOS of two CMOS positioned side by side is separated by a region which is the Shallow Trench Isolation (STI). It provides the two MOS to be put closer together, avoiding any unexpected interactions and allows a higher breakdown voltage. A 90nm process has typical trench width of 140nm and depth of 400nm.

4.2. Why and How Transistors Starved Off Power

Capacitors are thin conducting plates, usually made of metal, separated by a layer of dielectric and stacked or rolled to form a compact device. The general objective of the capacitors is to store charge during "On state" and discharge the same charge at "Off - state". Utilizing this concept, capacitors are served as small suppliers of charge when the circuit suffers from short of supply voltage. [1]

In CMOS technology, the perception of decoupling capacitors arouse when the scaling down of the devices came to the field. It became a severe issue to operate the devices into a reduced scale providing a constant voltage to its transistors during switching. This affair was recognized in the year 2001 at International Technology Roadmap for Semiconductors (ITRS) as one of the difficult challenges.

In a switching sub-circuit, like CMOS, switching noise must be suppressed. When the switching occurs, all the transistors draw a significantly large current and act as a load to the applied voltage source. These are due to the inductances of chip, package, motherboard and eventually Silicon itself. It encounters by lowering the voltage that the voltage supply offers. At this time, the transistors not only suffer from deficiency of voltage, every other transistor that shares the voltage undergo through these circumstances. Though the voltage restores to its normal state, a temporary reduction in voltage disturbs other chips [13].

To decouple others from the effect of the sudden current demand, a decoupling capacitor can be placed between the supply voltage and its reference (ground). For example: in four metal CMOS technology, a fabricated test chip of *300MHz CMOS RISC Microprocessor* with the super H architecture is of 0.35-um channel length and on-chip decoupling capacitance is 160nF [24].

4.3. Localized Decoupling

When the billions of transistors switch simultaneously, the capacitor initially supplies the required current and thus, the voltage droops are reduced for drawing instant current from capacitor. The location of the decoupling available in PCB, package and die in on-going technology can be shown through the diagram shown [2]:

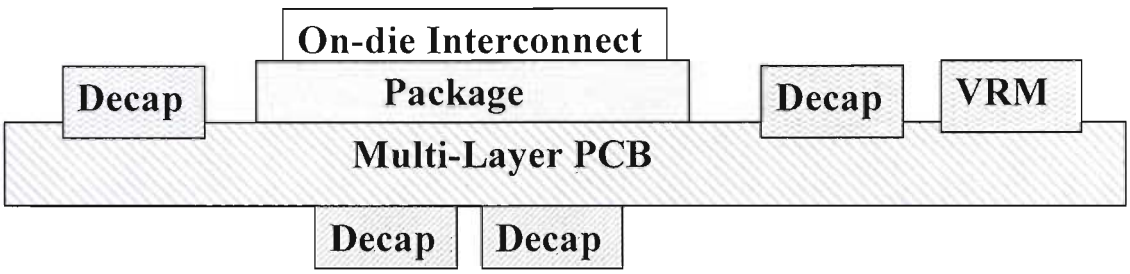


Figure 23: The Location of Decoupling Capacitors

So far we have seen that, the capacitance between power and ground distribution network, like CMOS, refers to decoupling capacitors or *decap* which act as local charge storage and help in mitigating the voltage drop at supply points. Unfortunately, this decoupling capacitance is sometimes not enough to limit the voltage droop. So the designers often add intentional decoupling capacitance structures on strategic locations.[1]

In this paper, we have proposed to allocate these decoupling capacitors next to the transistor level within the STI using the concept of DRAM technology followed by "trench" approaches. In this approach, a narrow ditch is dug in the DRAM die. Similar to this ditch, there will be trench consisting of two electrodes and filled by low k-dielectric, adjacent to the transistors in the CMOS that will supply sufficient charge at the time of switching and overcome the consequences of voltage droop.

4.3.1. Quick response

If the decoupling is far away from the location of power starvation, the small charge supplied by each of them is effectively reduced by the loss in the power path. The supplied power then suffers from interconnect delay due to metal wires and vias and arrives late at the transistor level. Even for on-die white area capacitances, it takes quite long time to arrive to the CMOS circuitry and compensate the droop. Thus a decoupling at the location of our proposal, right to the door step of transistor has become mandatory to mitigate droop and to respond quickly.

4.3.2. Rise time and fall time:

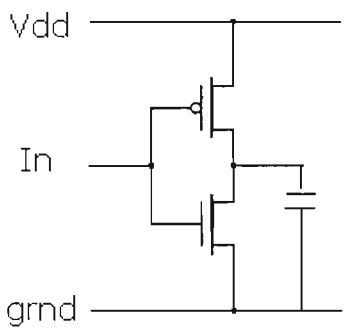


Figure 24: Charging and discharging of capacitor

The rise time depends on the channel width and mobility of charge carriers of the PMOS while that of the fall time depends on those of the NMOS.

As depicted in the above diagram, current comes from Vdd to the capacitor through the PMOS. The capacitor charges itself while the CMOS is on, i.e. at the positive high cycle of the square clock pulse. During the transient time dynamic power dissipates as current flows momentarily through both the MOS. The capacitor discharges during the low of the clock pulse. Thus the capacitor charges as current flows through the channel of the PMOS and discharges as it flows through that of the NMOS. Due to variation of mobility of charge carriers through the channel, their channel width is varied so as to keep the time of flow constant. This causes a change in the charging and discharging time of the capacitor and hence the clock pulse.

4.4. Use of Dram technology

4.4.1. Space constraints

Here, in this case we are concerned about the space constraint of the current technology. But the given proposal will not contradict with the space constraint. The proposed decoupling will not take a large lateral area. To introduce a capacitance in

CMOS circuitry to improve the power delivery, our main strategy was to introduce a capacitive element without making the chip bulky or without disturbing the Moore's goal. Here we have used the STI (Silicon Trench Isolator) area which is only used to create isolation between p-MOS and n-MOS. In our proposal we are just replacing the STI by using any dielectric material which has a higher dielectric constant to get the maximum capacitive effect [25]. According to recent technology the depth of the STI is about 250nm and we are proposing to deposit the dielectric with a depth of 126nm. So by no means space can be a problem for us or the proposed capacitance will not create any specious problem to a CMOS.

4.4.2. Fabrication constraint

The proposal introduces a DRAM capacitor within the CMOS inverter to provide the power when the inverter is in power crisis. Though both the processes are mature and already have fabrication set up, running them simultaneously is a difficult task. The fabrication process of both the CMOS inverter and DRAM is not the same but according to the proposal the fabrication process of CMOS and DRAM has to be done in parallel [26]. So the synchronization of the different fabrication phases of DRAM and CMOS needs a special concern.

Due to nano technology of recent days space constrains is always a special concern and all the fabrication phases have to be accurate to get the proper operation of each individual chip. The local decoupling will be created through digging trench or making stack inside the isolating material. While using high-K-dielectric material, the proper attachment of the dielectric with the immediate contact material or insulator has to be considered as well. The deposition process of the dielectric material has to be uniform. Besides, the uniform deposition of the polysilicon rod inside the trench needs a special consideration as well [27].

To design the whole system we didn't consider the roughness, dishing and trapezoidal effect of the top surface of metal. But due to the sophisticated fabrication process being

required, special phase has to be adopted concerning the roughness of the metal surface.

4.5. Capacitance enhancement

The main objective is to get the maximized level of capacitance using the local decoupling in the CMOS inverter. To obtain the higher capacitive effect several strategies could be adopted. Connecting the local decoupling between V_{dd} and ground and use of a high K-dielectric element in between the metal layers can be such strategies.

4.5.1. Between Vdd and ground

The connection of the proposed localized decupling is done in between V_{dd} and ground. Thus the capacitance will have the maximum voltage difference between the two metal plates which will enhance the value of capacitance. This way of connection will ensure the maximum charge storage in case of localized decoupling to deliver the maximum power during the power starvation of a CMOS inverter [28].

The connection with Vdd will also ensure that the addition of decoupling will not affect the start up speed of the chip. But still the power path between the V_{dd} and the metal plate of the decoupling can have some contribution in inductive droop which can not be avoided.

4.5.2. Use of high –k dielectric in between metal layers

As our proposal followed here using DRAM trench technology, which does not use any high-κ-dielectric material as the dielectric between the metal plates. But we can use any dielectric material having higher dielectric constant if the interface suite with the metal plate. The use of a material having a higher dielectric constant is also important for having a higher level of capacitance.

$$C = \frac{\varepsilon_r \varepsilon_0 A}{d}$$

According to the above formula, C is directly proportional with the ε_r. Higher dielectric constant can ensure higher level of capacitance. So to get maximized localized decoupling high κ- dielectric material can be used in between the metal layers. For instance material like Hafnium-Di-Oxide can be a better proposal [29].

5. SIMULATION:

5.1. System set-up

The system established using the ADS (Advanced Designed System) Software, simulation software which is widely used in vast industrial purpose. To design the CMOS inverter circuit, BSIM4 p-MOS and n-MOS models are used where all the parameters of p-MOS and n-MOS are fixed by following the 90nm fabrication process. The BSIM4 p-MOS and n-MOS models are built in models of MOSFET in ADS. The proposed capacitor is placed between the two CMOS inverter circuit, as the figure below shows, where both of them meets with the same source (V_{dd}).

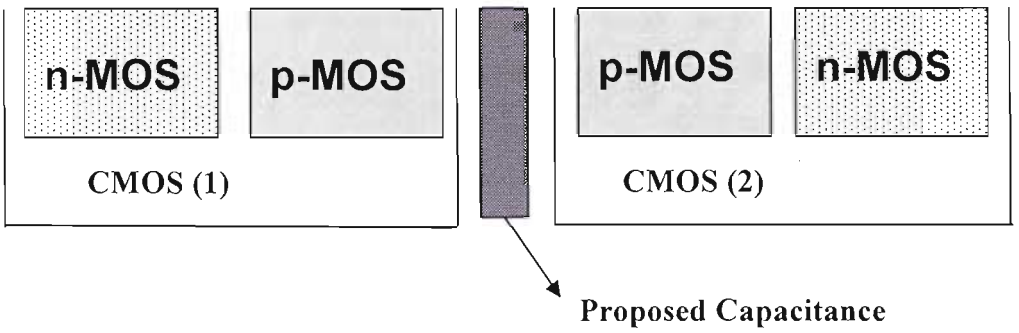


Figure 25: The position of proposed capacitance

As our proposal is to use high-κ-dielectric element in local decoupling capacitor, we can use higher ε_r, but in this case we used the standard dielectric constant of SiO_2 which is used in recent fabrication process. To get further accuracy, numbers of CMOS inverter circuits have been connected in parallel. Here to estimate the possible voltage drop and inductive droop calculation is made and these loss factors are designed through lumped RLC elements in ADS. To calculate the resistance, here the higher frequency skin effect has also considered as it has a significant effect during the higher frequency operation. For the calculation of capacitive loss, only the parallel plate capacitances and inter-wire capacitances are taken under consideration. The distribution of resistance and reactance is done in several levels, like motherboard, package, socket and silicon. In silicon level designing, few metal layers and vias have

been designed as well. Copper is assumed to be the interconnect material of choice in different metal layers. For interconnect designing, we have considered lower dielectric constant to make the interconnect capacitance less which is followed by the ongoing technology. In case of designing the RLC circuit for each step, the resistances and inductances are connected in series and the capacitances are connected in parallel.

To design the local decoupling in spite of considering the cylindrical capacitance, the calculation is done following a parallel plate capacitor. The inherent ESL and ESR of the capacitive effect of the proposed localized decoupling are considered as well. 1V DC voltage is given as the source (V_{dd}) of the whole chip. To generate the input clock and give it to the inverter circuit, odd numbers of inverters are used and connected in a series connection with each other. If a single voltage pulse is given to any of the inverters connected in series with automated clock pulse which would drive the designed CMOS inverters. The importance of using odd numbers of inverter to generate the clock pulse is to get a different output. The frequency of the clock generator is fixed to 3.5GHz which is the most current trend. The transient analysis and simulation of the whole system is done in time domain.

5.2. Impedance estimation according to current trend

According to ITRS for interconnect, 2007 edition "The aspect ratios of the spaces between adjacent gates in DRAMs are expected to be greater than 16:1 by 2007 and will increase thereafter." This will be met in our proposal since the DRAMs would be separated by a CMOS. As depicted in figure A1 of the ITRS paper, the trench height is assumed to be 126nm. The aspect ratio for current DRAM technology is 50:1, which essentially means that the depth: width=50:1. The paper also assumes that the effective dielectric constant can be 2.9, which is easily affordable. All these make the half pitch to be 1.26nm. With simple physics, prediction tells us that the trench would be cylindrical in shape with a hemispherical bottom. The two plates of the capacitor which is a must would be made one, out of the wall and the other, by dipping a metallic rod. These tow metal contact would be separated by a dielectric material. As shown in the

diagram below, L denoted as the half pitch of the trench is half the diameter of the trench width, including outer metal wall. 'y' is shown as dashed line in the diagram, showing the overlapping region of the two plates. 'd' is the average distance between the two plates. Details of the capacitance calculation are as follows. [27].

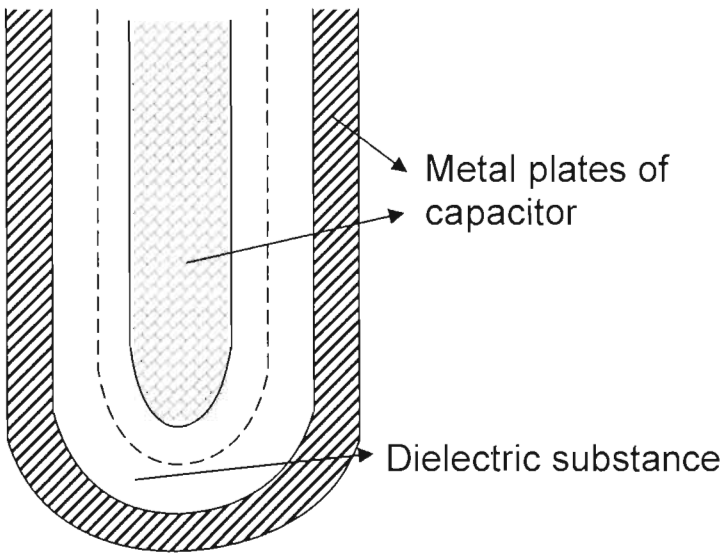


Figure 26: Trench DRAM (simplified)

L=half pitch of the trench

y= overlapping area between the two plates

d=distance between the plates

Using the formula $C=\frac{\varepsilon_r \varepsilon_0 A}{d}, A=Ly$, taking ε_r= 2.9,

$$C = \frac{20 \times (8.854 \times 10^{-12}) \times A}{2 \times (15 \times 10^{-9})}$$

$$A = Ly$$

$$A = L[8 + 8 + \frac{(2 \times \Pi \times r)}{2}]$$

$$A = L[16 + 3.142 \times 1.5 \times 15] \times 10^{-9}$$

$$A = 8.669 \times 10^{-8}(126 + 126 + 1.26)$$

$$A = 2.15 \times 10^{-5}$$

Therefore, $C = 1.296 \times 10^{-7}$ F

6. SIMULATION AND DATA ANALYSIS:

The simulation is done using ADS (Advanced Design System) simulation tool and done using three levels of calculations. Here three types of dielectric material are used having three different dielectric constants. This adaptation of several ε_r helped to observe three different results regarding performance improvement. The results of using low, medium and higher local decoupling in the CMOS inverter level can improve the level of voltage droop which is shown below:

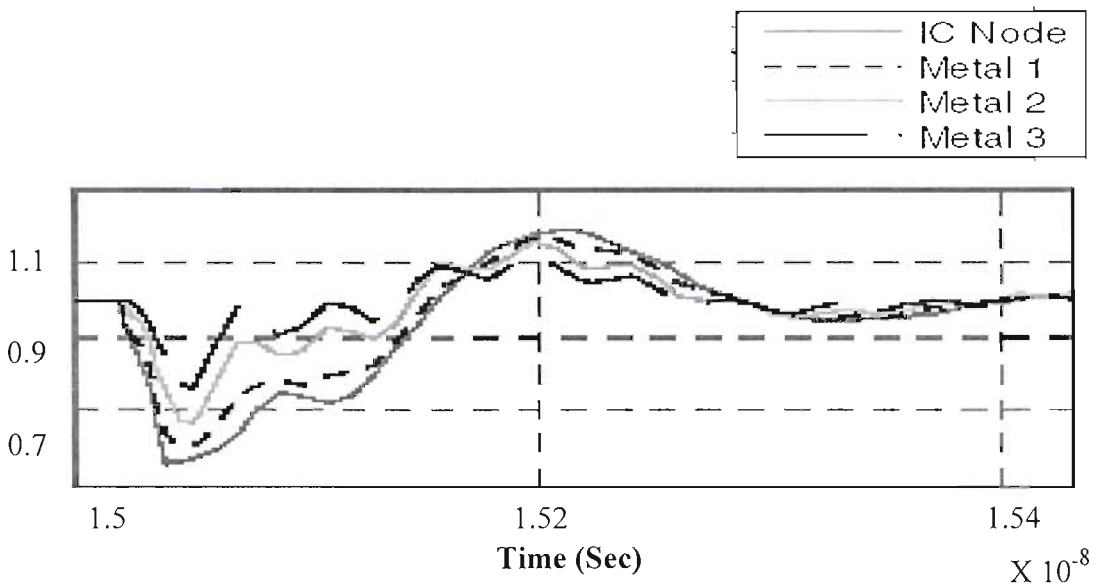


Figure 27: Voltage Droop in the Die with Low Local Decoupling

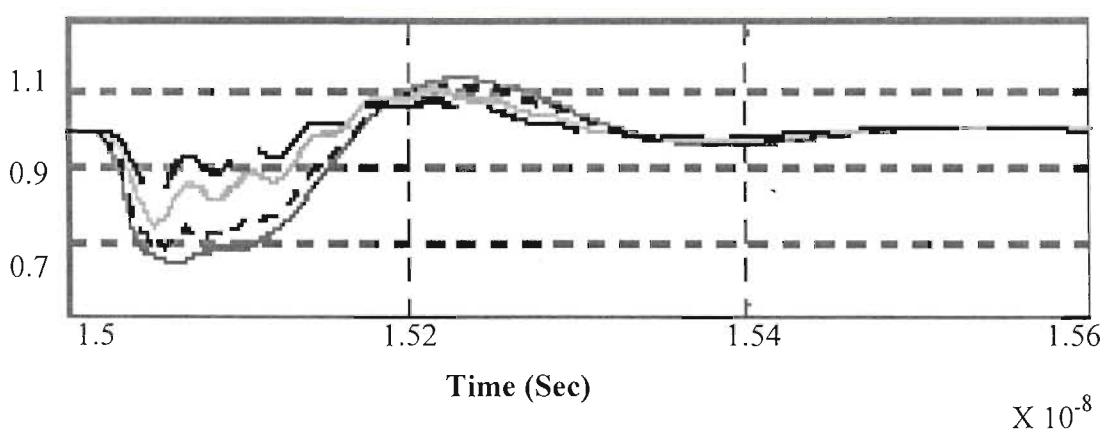


Figure 28: Voltage Droop in the Die with Medium Local Decoupling

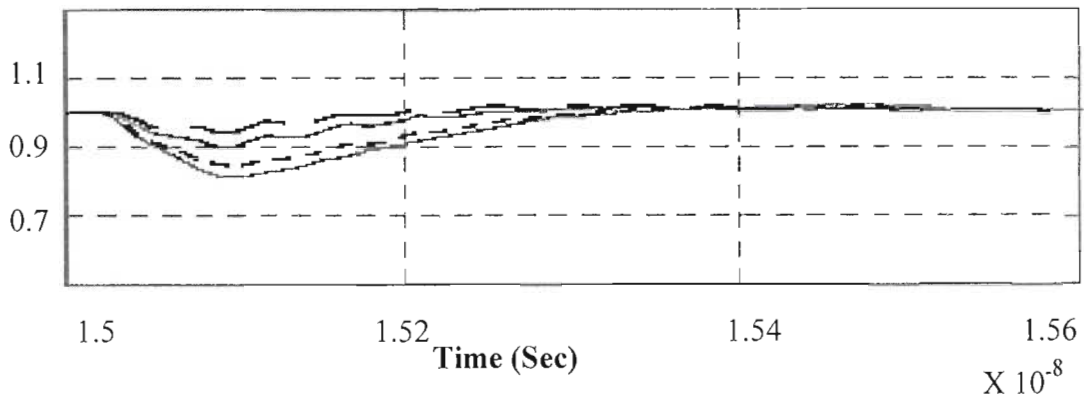


Figure 29: Voltage Droop in the Die with High Local Decoupling

In the above three diagrams, the differences of duration of the voltage drop at each of the three metal layers and the IC node has been compared for performance evaluation. The time taken for the maximum droop is the most for the low local decoupling and the least for the high local decoupling. This also brings about a change in the gradient with respect to the time constant. According to the ADS simulation result, for the low local decoupling the amount of droop is 0.56Volt while for the medium and the high local decoupling the droop is 0.68Volt and 0.81Volt respectively. The voltage droop in metal layer 1, as expected, is the highest. That's why our proposal is to provide required power right at the CMOS circuitry, where the power is required. The above diagrams show the voltage droop in metal layer 1, metal layer 2 and metal layer 3. Here the performance improvement is measured in terms of the extent of the voltage droop for the individual cases as shown in each diagram.

The value of the decoupling capacitor is based on the charge stored by a single DRAM capacitor, which is assumed to be 10 fF. According to the simulation result the low and medium local decoupling can achieve the required V_{dd} level after a few nano second but again it can face a droop and the voltage levels fluctuates. While the high local decoupling is able to minimize the voltage droop very effectively and within less time delay it can achieve the required source level. According to the analysis, this localized decoupling can improve more than 30% performance of the system.

7. DISCUSSION AND CONCLUSION:

Simultaneous switching of billions of transistors draws a huge amount of current. This current is solely responsible to cause voltage droop due to resistive and inductive drop in the power path. Now-a-days power starvation in the Silicon level becomes a major problem which is responsible to degrade the performance of a system. But due to the space constraint in the die, large capacitors cannot be plugged in. Though we have some decoupling in motherboard, socket and package level but still they are quite far away, hence the response is very slow. If we try to have a horizontal decoupling in silicon level, it will require a larger lateral area, which is difficult to achieve. Here our proposal stands for. According to our proposal, devices will have a way to insert a vertical decoupling by creating a trench in the STI region between two inverters. Being right next to the inverter will ensure the maximum voltage compensation with a minimum required time. On the other hand, lower droop introduces lesser noise, so our proposal is successful in case of noise reduction as well. Moreover, as we are introducing a decoupling it will help to minimize ESL and ESR which will reduce the overall impedance. Our proposal provides one capacitor of around 10fF for each pair of transistor to reduce power starvation, noise and impedance. The dimensions of this decoupling are proposed using current trend of CMOS and DRAM, following the 90nm process technology mentioned in ITRS 2007. The merging of these two mature technologies has brought forth a capacitance such that it does not waste any lateral space on die. Besides, this merging of CMOS and DRAM will be cheap to implement as we don't need to develop any new fabrication processes which will ensure the absence of any fabrication constrains as well. The calculations, data analysis and simulation have been verified using ADS (Advanced Design System, by Agilent Tools). Thus, it can be concluded that our proposal holds.

8. FUTURE WORK

The future prospect of our thesis would be to simulate the design taking the same number of metal layers that is being implemented now-a-days. The metal layer design is done using the momentum feature of ADS software. The momentum feature ensures the exact voltage droop due to inherent RLC of metal layers and vias. Besides, here all the parameters are set by following the 45nm process technology and in case of dimension design ITRS 2007 has been taken as reference. The system would be set up such that the power supply is used to run the system through traces and vias experiencing an RLC circuit as it does so. The frequency is fixed at 3.5GHz.

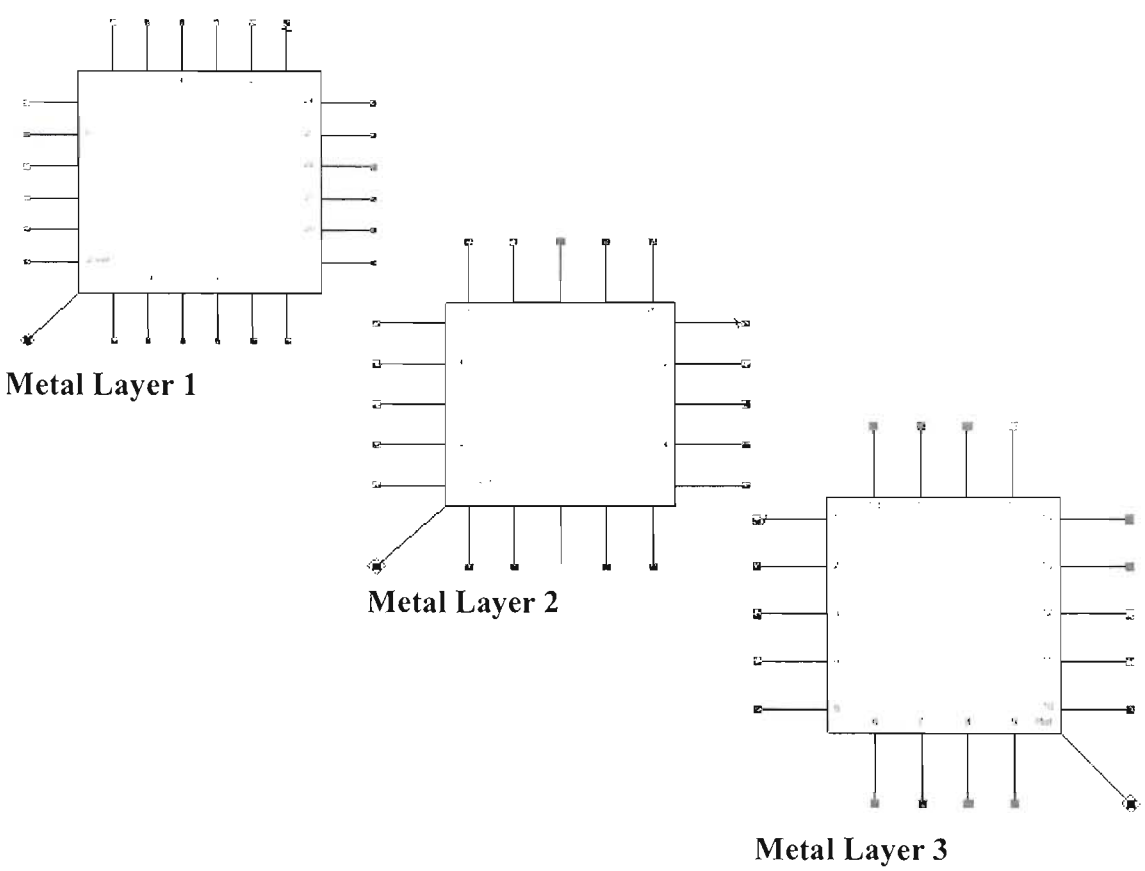


Figure 30: Metal layer designed using ADS

For instance, as the diagram (figure 30) above shows the three metal layers, the Metal Layer 1 consists of 24 pins, of which there are a combination of V_{dd} and G_{nd} (Ground) pins. Among 24 pins the first 20 pins will be connected with the CMOS inverters providing V_{dd} and G_{nd} to the CMOS. Rest of the pins, 21 and 23 are assigned to provide V_{dd} (Source) and 22 and 24 pins are for G_{nd} connectivity with metal layer 2. The metal layer 2 is designed with 20 pins. Among these, the first 18 pins of metal layer 2 will be connected with metal layer 1 and these pins are assigned to supply V_{dd} and G_{nd} to metal layer 1. 19^{th} and 20^{th} pins of metal layer 2 are the source and ground pins respectively which make the connection between metal layer 2 and metal layer 3. Higher level Metal layers from metal layer 3 have the same pin combination and assignation like metal layer 2. All the metal layers have reference pin denoted as ref. in the above figure and this pin is kept grounded.

REFERENCE

[1] http://www.wikipedia.org

[2] Laila S. Sraboni, Ophelia Mohaimen, Rezwana H. Mustazir, and S. M. Salahuddin, and Md Ishfaqur Raza, "*Improved VLSI Circuit Performance using Localized Power Decoupling*"

[3] M. Popovich et al.,"Efficient placement of distributed onchip decoupling capacitors in nano-scale ICs", ICCAD 2007, IEEE/ACM International Conference on Computer Aided- Design, Nov. 2007

[4] Michael Sotman, Avinoam Kolodny, Kolodny Mikhail Popovich, Eby G. Friedman "On-die Decoupling Capacitance: Frequency Domain Analysis of Activity Radius", ISCAS 2006, 0-7803-9390-2/06/$20.00 ©2006 IEEE

[5] R.S. List, B.A. Block, M.T. Bohr, "On-die de-coupling capacitor using bumps or bars and method of making same" US patent 6706584, to Intel Copr., Patent and Trademark Office, 2004

[6] The international technology roadmap for semiconductors: 2007, http://public.itrs.net.

[7] J. Chen and L. He, "Efficient In-Package Decoupling Capacitor Optimization for I/O Power Integrity", IEEE Trans. Computer-Aided Design of Integrated Circuits and Systems, vol. 26, Apr. 2007.

[8] D.W. Bailey, "Clock Distribution", Design of High-Performance Microprocessor Circuits, p261-280, 2000

[9] WO/2005/022597) INTEGRATED CIRCUIT PACKAGE HAVING AN INDUCTANCE LOOP FORMED FROM A MULTI-LOOP CONFIGURATION

[10] S. M. Sze, *physics of semiconductor devices,* 2nd Ed. Wiley Eastern ltd, 1991.

[11] http://www.intel.com

[12] N.H.E. Weste, D. Harris and A. Banerjee, *CMOS VLSI Design: A Circuit and Systems Perspective*, 4th Ed. Wiley Eastern ltd, 1998.

[13] W. Streetman and A. Banerjee, *Solid State Electronic Devices*, 4th International Conference on Solid-State and Integrated Circuit Technology.

[14] Authorized licensed use limited to: IEEE Xplore. Downloaded on November 15, 2008 at 03:02 from IEEE Xplore. Restrictions apply.

[15] S.M. Kang and Y. Leblebic, "*CMOS Digital Integrated Circuit Design - Analysis and Design*", Chapter 2 (Fabrication of MOSFETs).

[16] W. Maly, Atlas of IC Technologies, 1987.

[17] E. Chen, P. Ya-Ling, and W. Tings, "Using effective wet etching technology to improve deep trench shape", IEEE Conference and Workshop on Advanced Semiconductor Manufacturing, 2004. ASMC '04, May 2004.

[18] http://www.imec.be/essderc/ESSDERC2002/PDFs/D26_1.pdf

[19] M. Xiongfei, K. Arabi, R Saleh, "Novel decoupling capacitor designs for sub-90nm CMOS technology", 7th International Symposium on Quality Electronic Design, ISQED '06. pp 6, March 2006.

[20] W Wolf, *Modern VLSI Design System-on-chip Design,* Pearson Education 2005.

[21] S. Akram, M. Brooks, "Method for fabricating semiconductor component with on board capacitor", US Patent 7041537, to Micron Technology, Inc 2006

[22] T. Seidel, J. Dalton, Z. Karim, J. Lindner, M Daulesberg, and Z. Wei, "Advances in atomic level deposition technologies", 8th International Conference on Solid-State and Integrated Circuit Technology, ICSICT '06. Oct. 2006.

[23] T.Y Winarski, "Dielectrics in MOS devices, DRAM capacitors, and inter-metal isolation", IEEE Electrical Insulation Magazine, vol. 17, Nov.-Dec. 2001N. Weste and D. Harris, CMOS VLSI Design: A Circuits and Systems Perspective, 3rd edition, Addison Wesley, 2004.

[24] S. Bobba , T. Thorp, D. Liu, "IC Power Distribution Challenges", Computer Aided Design, Vol 4-8, 2001

[25] N.H.E Weste, D. Harris, "Scaling", CMOS VLSI Design 3rd, 2005

[26] Dally, J. Poulton, "Technology Trends and Digital Systems Engineering" Digital Systems Engineering, 1998

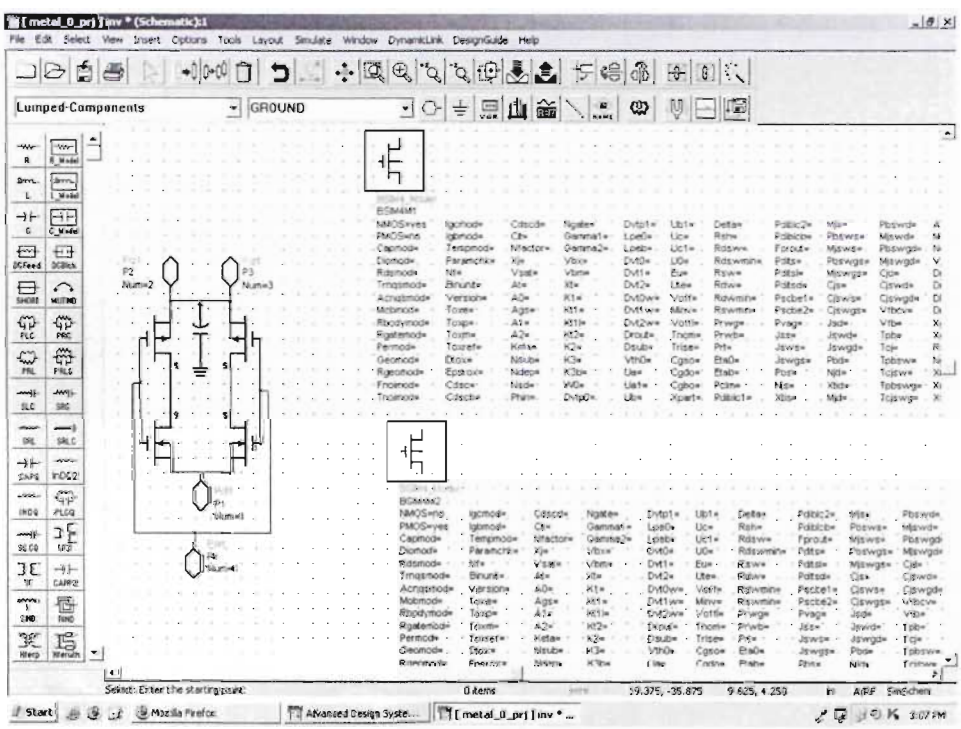
[27] Y. Taur, "CMOS Scaling and Issues in SUB-.25um Systems", Design of High-Performance Microprocessor Circuits, 2000

[28] H. Su, S. S. Sapatnekar, S. R. Nassif, "Optimal Decoupling Capacitor Sizing and Placement for Standard-Cell Layout Designs" IEEE Transaction on Computer-Aided Design of Integrated Circuits and Systems, Vol 22 No 4, 2003.

[29] J. Rius, M. Meiher "A Model for On-Chip Decoupling Capacitor Effectiveness Including Gate Leakage Effects" IEEE 13th Topical Meeting on Electrical Performance of Electronic Packaging.

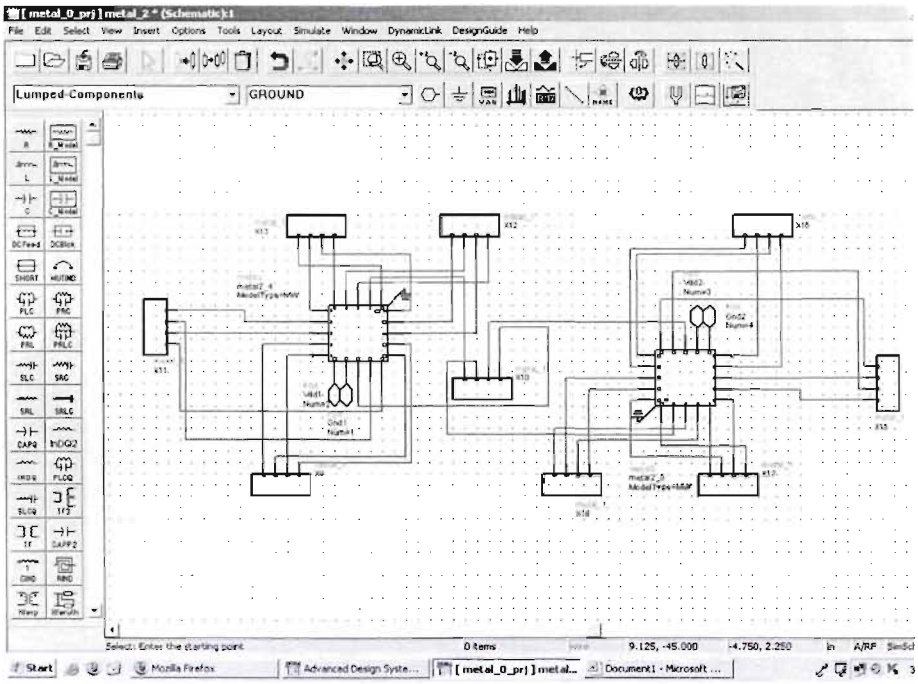
APPENDIX A : ADS SCHEMATICS

To verify the proposal the system was designed using ADS (Advanced Design System), which is a powerful electronic design automation software system by Agilent tools. It offers complete design integration to designers of products such as cellular and portable phones, pagers, wireless networks, radar and satellite communications systems, and high-speed digital serial links. With a complete set of simulation technologies ranging from frequency-, time-, numeric and physical domain simulation to electromagnetic field simulation, ADS lets designers fully characterize and optimize designs. The single, integrated design environment provides system, circuit, and electromagnetic simulators, along with schematic capture, layout, and verification capability -- eliminating the stops and starts associated with changing design tools in mid-cycle.

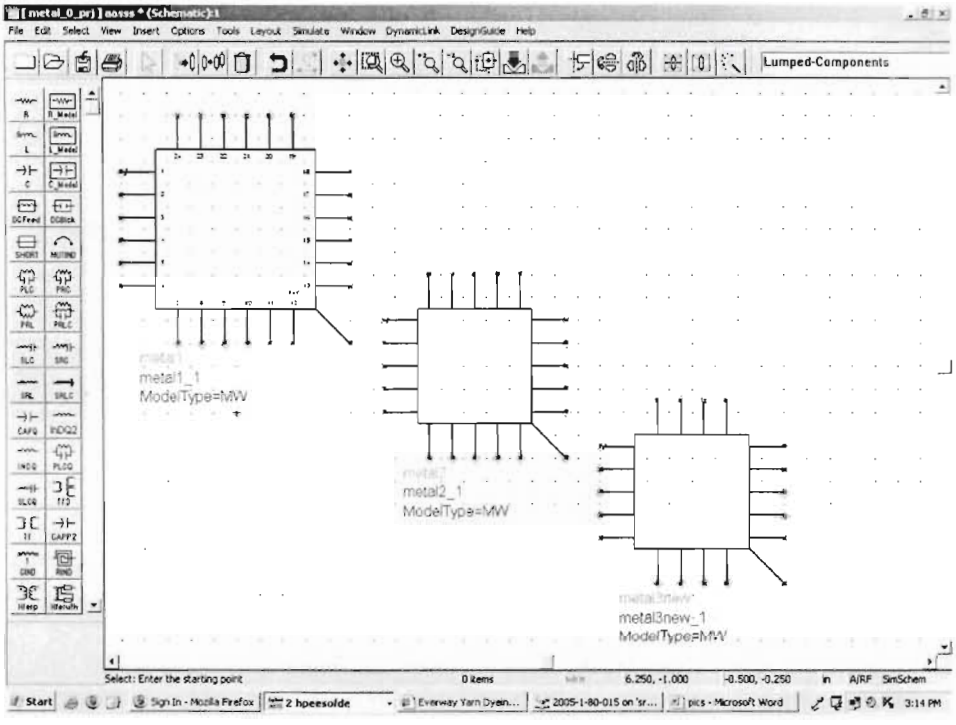


The diagram above is ADS schematics showing two inverters with the proposed capacitor between them. The four ports shown in the figure, two of them are assigned

for the Vdd, one for input clock and one for reference, are connected to the inverters with the first metal layer. The PMOS and NMOS used here to design the system are built in BSIM4 model.



The figure shows another ADS schematics showing the connection between metal layer one and metal layer two. In this figure, two metal layer two is connected with nine metal layer one. Thus the V_{dd} and ground connection passes through the metal layers and vias of a silicon chip.



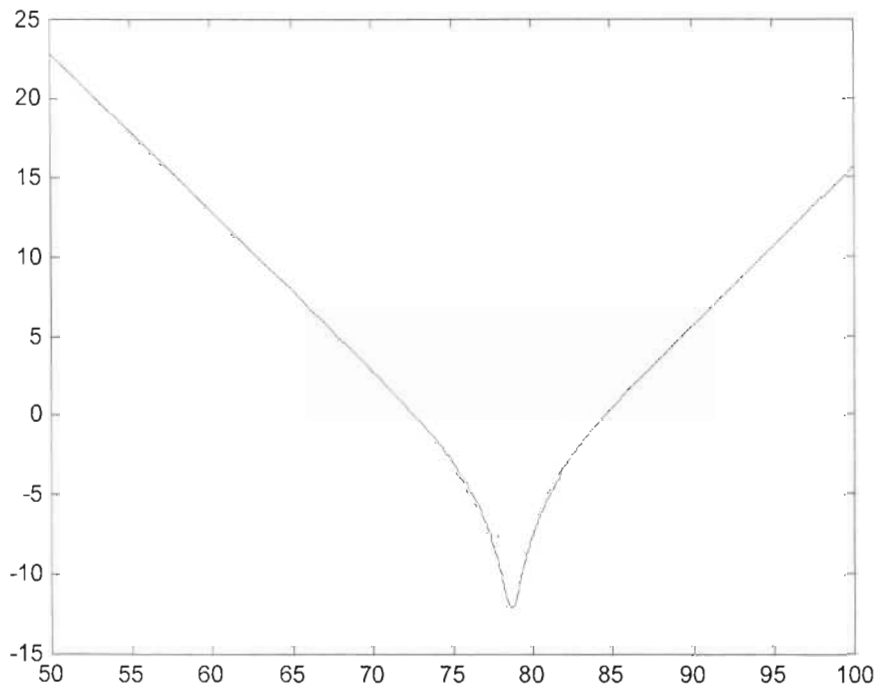
The above ADS schematics shows three metal layers designed using momentum features of ADS.

APPENDIX B: MATLAB

```
r=0.061;
c=8200*10^-12;
l=0.58*10^-9;
f=0.3*10^6;
z=r+(i.*2.*pi.*f.*l)-i./(2.*pi.*f.*c);
s=abs(z)
s =64.6961
```

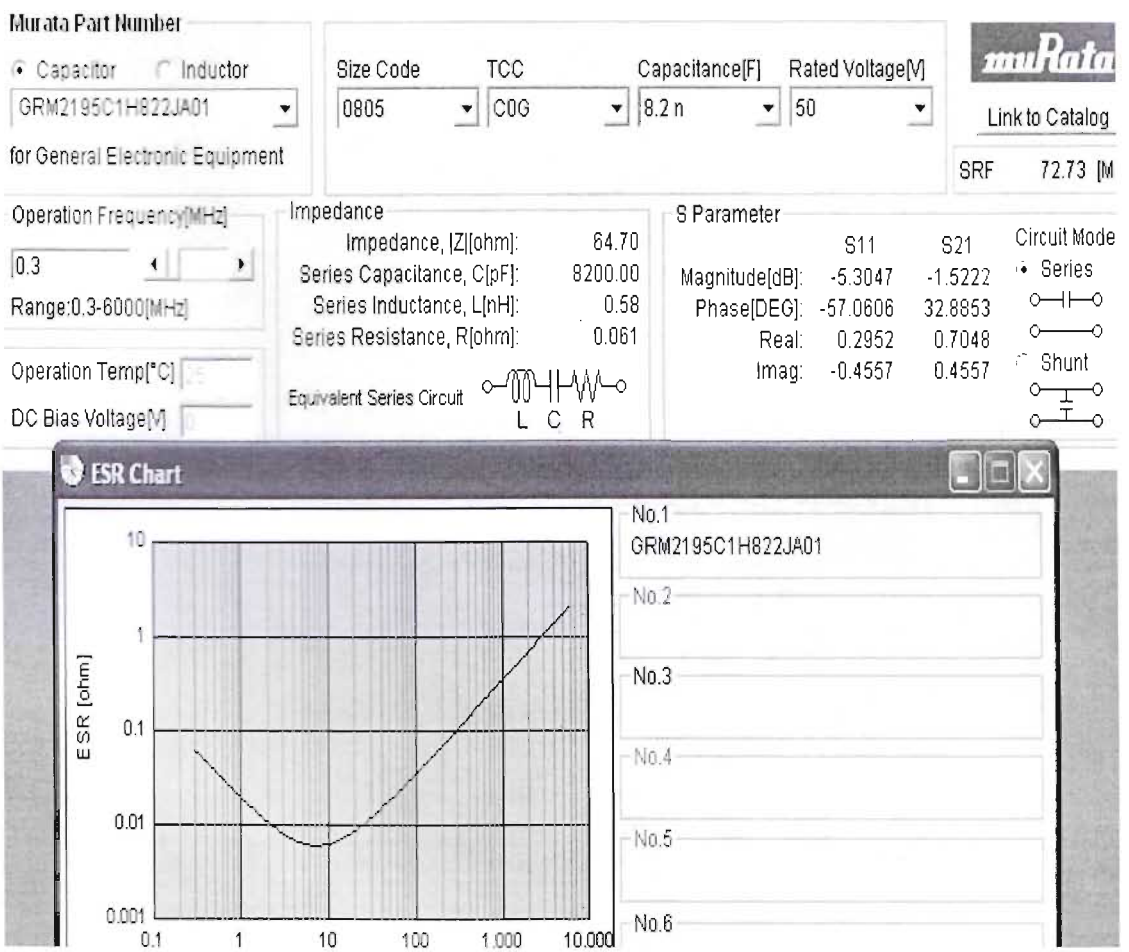
For plot:

```
r=0.061;
c=8200*10^-12;
l=0.58*10^-9;
%ini=10*log10(.1*10^6);
%fin=10*log10(10000*10^6)
f=linspace(.1*10^6,10000*10^6,10000);
z=r+(i.*2.*pi.*f.*l)-i./(2.*pi.*f.*c);
%p=z+(1./t);
s=abs(z);
dB=10*log10(s);
plot(10*log10(f),dB)
```



By using the software of "*muRata*" it was found that at lower frequencies the capacitive component is significant ($1/j\omega C$) and in higher frequency trend the inductive components prevail. For a particular frequency, there is no capacitive component or inductive component and impedance is equivalent to only resistive part .Typically, we try to operate our devices at this frequency.

APPENDIX C: MURATA



This page shows the demonstration of the "*muRata*" software window. Here, for a capacitor named as, "**GRM2195C1H822JA81"** which has a size of 80mil (length) by 50mil (width). Considering its inherent series capacitance as 8.2nF, a rated voltage of 50V being operated at 0.3MHz.frequency, we find an impedance of 64.70Ω