DESIGN, LAYOUT AND PLACEMENT OF ON-CHIP DECOUPLING CAPACITORS IN IP BLOCKS

by

JESSE CHIA

B.A.Sc., University of British Columbia, 2002

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

THE FACULTY OF GRADUATE STUDIES
ELECTRICAL ENGINEERING

The University of British Columbia

December 2004

ABSTRACT

In today's deep submicron technologies, a number of new signal integrity issues have arisen due to increased resistance, coupling capacitance and inductance in the metal interconnect. One of the main concerns is power supply noise in the form of *IR* drop and *Ldi/dt* effects as clock frequency continues to increase. As supply voltages scale down with technology, and the level of current and its rate of change increases, the noise levels on the supply are increasing to detrimental effect. Power supply noise can increase the delay of gates and even cause them to function improperly.

This thesis addresses the use of *standard-cell* decoupling capacitors (decaps) to reduce power supply noise in IP (intellectual property) blocks that are designed using standard cell layout tools. First, we study the performance of decaps, implemented using MOS transistors, as a function of frequency. A number of observations are made, both qualitatively and quantitatively, about the behaviour of NMOS and PMOS transistors when used as decoupling capacitors.

In addition, a number of equations are derived to characterize the frequency and time-domain behaviour of decaps. Using these equations and SPICE simulations, a number of recommendations are made as to how decaps should be laid out. The issues of where to place decaps and how much to place is also addressed using a commercial power grid analysis tool. General design guidelines are developed based on the results to keep noise below the budgeted amount during the layout of an IP block.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

First of, I would like to thank my academic and industry supervisors Dr. Resve Saleh and Dr. Karim Arabi for their encouragement, technical advice, and financial and moral support.

I express my appreciation for the support of the engineers at PMC-Sierra, without whose help, this research would not have been possible. Specifically, I would like to thank Andy Hung for his valuable advice and knowledge on Apache Redhawk® and placement-and-routing, Daisy Chen for her help on cell design, extraction and verification. Additionally I would also like to thank Mamta Bansal, Jianming Chen, Jurgen Hsissen and (former PMC-Sierra engineer, now SoC-lab researcher) Brad Quinton.

These last couple of years, I have gotten to know new friends and old friends better. I shall remember the friendship and support of Andy Kuo, Andy Yan, Neda Nouri, Roberto Rosales, Dr. Peter Lawrence and Dr. Luis Linares.

I greatly appreciate the financial support provided by the Natural Science and Engineering Council of Canada (NSERC), and PMC-Sierra. I would also like to thank Canadian Microelectronics Corporation (CMC) for providing the CAD tools.

Finally, I would especially like to thank my friends and family for their kindness and continual support.

Chapter 1

Introduction

1.1 Motivation

As integrated circuits (ICs) scale according to Moore's Law, there is a doubling of the number of transistors on a chip per unit area with each successive technology generation.[1] The capabilities of the IC industry have reached the point where a single chip with one billion transistors is feasible, and an announcement of such a chip is expected to occur in the next few years. To handle this level of complexity, a new chip design methodology has been adopted where reusable blocks are used to rapidly assemble a large design in a relatively short amount of time. This emerging methodology is referred to as System-on-chip (SoC) and the reusable components are referred to as intellectual property (IP) blocks.

Typically, IP blocks are represented using a high-level hardware description language such as VHDL [2]. They describe digital logic circuits in a form that can easily be synthesized into gate-level and transistor-level circuits. The descriptions are converted to an IC layout using an ASIC design flow that employs standard cells that are maintained in a library. This flow has been used successfully since the 1980's. However, recently a number of problems associated with technology scaling into deep submicron (DSM) have arisen to disrupt this flow, as described below.

Around 1988, the channel length of the MOS transistor was nominally 1μm. It has been scaled by a factor of roughly 0.7 every 2 - 3 years since that time, with a corresponding increase in the number of transistors per unit area. Of course, having more transistors to connect together creates a need for more levels of interconnect. Today the industry has reached, and in some cases exceeded, 8-10 metal layers in advanced processes. When mainstream technology reached 350nm around 1995, new problems were encountered on-chip due to interconnect. Wires were scaled in the same way as transistors and this increased their resistance. Furthermore, the wires were squeezed closer together and this increased capacitive coupling between wires. These two trends introduced a set of problems that are collectively referred to as *signal-integrity issues*. The list of issues in this category includes [3]: interconnect *RC* delay dominating gate delay; signals that are capacitively-coupled leading to delay variations and noise injection; power-supply noise increasing due to voltage drops from resistance effects.

In this thesis, the issues related to the integrity of the on-chip power distribution system are addressed. Today, the power grid is much more complex to design than in years past and has therefore been listed as one of the long term *Grand Challenges* identified by the International Technology Roadmap for Semiconductors (ITRS) [4]. Currently, designers of the power distribution system are faced with two main issues. First, as power lines become thinner, their resistance increases and with it, power-supply noise. Power-supply noise occurs when logic gates switch and charge flows through the power grid to deliver current to the gates. This current creates a voltage drop due to resistance along the power lines known as *IR*-drop. This is illustrated in Figure 1.1. Second, there are inductances associated with the connections from the chip to the external supply voltage, also shown in Figure 1.1. This inductance, together with a high rate of change of current with respect to time, introduces another source of voltage drop known as $L\,{}^{dI}\!/_{dt}$

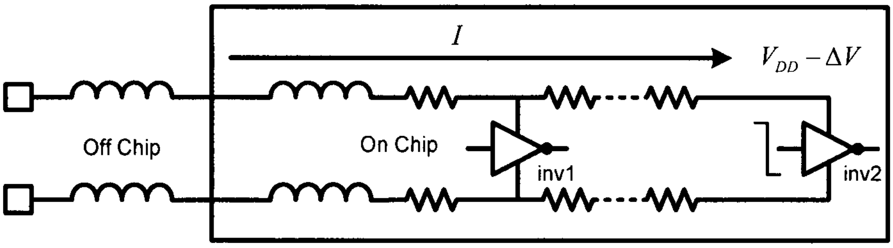


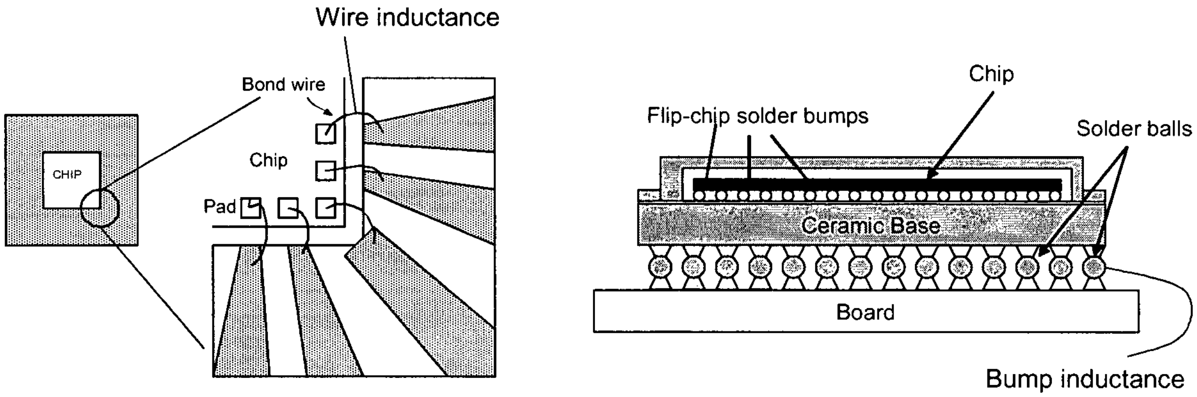
Figure 1.1 Sources of voltage drops in power supply.

Together, the total voltage drop at any point in the power grid is given by:

$$\Delta V = IR + L\frac{dI}{dt} \tag{1.1}$$

Each of these effects is considered separately starting with *IR*-drop. The basic concept of *IR*-drop can be understood by again examining Figure 1.1, which depicts two large buffers connected to a resistive power supply. Initially, all voltage levels in the power grid are at V_{DD}. As the second driver, inv2, begins to switch, the demand for current from the power grid stresses the grid. Specifically, the wire resistance creates voltage drops that increase as the current moves from the external supply towards inv2. The voltage remains relatively high near the V_{DD} connections at the periphery of chip, and drops by ΔV at the connection to inv2. In practice, *IR*-drop is caused by simultaneous switching of clock buffers, bus drivers, memory decoder drivers, etc. These simultaneous switching events can occur anywhere on the chip and, therefore, all regions are susceptible to *IR*-drop. The ground grid is subject to the same type of problem when the outputs switch low, except that the voltage of the ground line will increase in value. This is sometimes referred to as *ground bounce*.

The second source of voltage in Eqn. (1.1) drop is due to inductance. This inductance is due to either the bonding wire from the chip to the external supply, or the solder bump used in ball-grid array packaging technology. These two cases are illustrated in Figure 1.2. When dc current flows through these wires, there is no voltage-drop across the inductance. However, when the current is changing with respect to time, a voltage drop occurs. The higher the *di/dt* value, the larger the voltage drop, as seen in Eqn. (1.1). Today, the value of *di/dt* continues to increase as on-chip switching gets faster.



(a) Dual-inline Packaging

(b) Ball Grid Array Packaging

Figure 1.2 External Sources of Inductance in Power Grid

Typically, a noise budget of roughly 10% of V_{DD} is set for the power grid voltage fluctuations. That is, the design must operate correctly even if the difference between V_{DD} and G_{ND} is 10% smaller than the nominal value.

This thesis is focused on the issue of power supply noise and, in particular, its reduction through the use of *standard-cell* decoupling capacitors. Decoupling capacitors, or decaps, are effective at mitigating voltage swings on the power grid. As shown in Figure 1.3, decaps are essentially capacitors that store reservoirs of charge that act as voltage sources near the switching devices.

The decaps provide the charge needed for initial current flow when devices switch, and are replenished later in the cycle by charge delivered from the supply. Decaps can be placed in the open areas of the chip between IP blocks (known as global decap) and within the IP blocks themselves. The concern here is with the on-chip decaps within an IP block as oppose to 'global' on-chip decaps found between IP blocks or off-chip decaps ([5]-[9]).

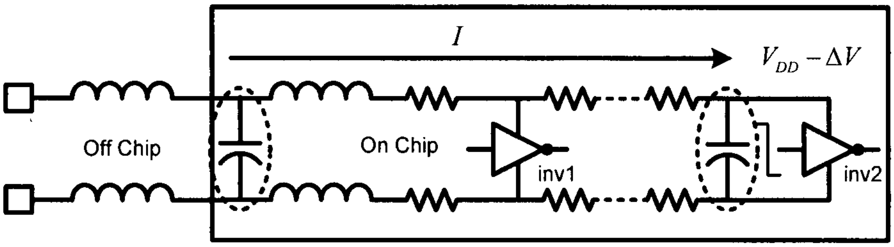


Figure 1.3: Use of decoupling capacitors to Reduce Voltage Fluctuations

1.2 Decoupling capacitors

The most effective way to implement on-chip *standard-cell* decaps is through the use of a MOS transistor connected between V_{DD} and G_{ND}, as shown in Figure 1.4. This is because the thin-oxide of the gate delivers a higher capacitance for a given area of silicon than any other oxide available in the standard MOS fabrication process. For this device, the total capacitance is roughly WLC_{ox}, where W is the transistor width, L is the transistor length, and C_{ox} is the oxide capacitance per unit area.

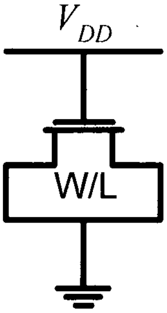


Figure 1.4 Decoupling capacitors Implemented Using an NMOS Transistor

While the desired capacitive behaviour holds at low frequencies, decaps are now being used at higher and higher frequencies. Therefore, there is a need to study the performance of decaps as the frequency increases. Since decaps are usually built using MOS transistors, the high-frequency behaviour of MOS transistors will be investigated. Previous work on modeling the high-frequency response of MOSFETs include [10]-[12]. Unfortunately, it will be shown that at higher frequencies the effective capacitance of the MOS transistor drops off significantly if the channel length of the device is too large. In effect, the decap cannot deliver the needed charge and the *IR*-drop will be larger than expected.

Based on frequency and time-domain analysis, the layout requirements of a decap can be determined. In particular, a single MOS device with a large *L* may not have the desired frequency response. Instead, the same target capacitance can be obtained with a number of parallel devices that, when combined, deliver better performance. For example, a target capacitance obtained with a given *W* and *L* can also be realized with three transistors of size *W* and *L*/3 that are connected in parallel. Also, PMOS devices can be used in parallel with NMOS devices to implement the desired decap value. Thus, different approaches to standard cell layout of decaps using NMOS and/or PMOS devices must be explored.

For a standard cell layout of an IP block, identifying the proper location and size of decaps is important. Today, the insertion of decaps is carried out after initial standard cell placement is completed. The general rule-of-thumb is that the amount of decoupling capacitance is usually ten times the amount of switching capacitance [13]. As the frequencies continue to increase, decoupling capacitor placement must be considered earlier in the design process. For example, decoupling capacitors may be placed either in empty areas of the layout, near the center of the block, or near the solder bumps and V_{DD} pins. However, if decoupling capacitors are added far away from large noise violations, they will do little good. The goal here is to try various schemes for placement and sizing of decaps to determine which method is most effective.

Most of the work reported here pertains to the 180nm, 130nm and 90nm technology nodes. Modeling equations will be developed that are suitable for these technologies. Layout approaches for standard cell decaps will be described. Different placements will be explored to assess their ability to eliminate *IR*-drop hot-spots. Finally, analysis is carried out with a commercial power grid analysis tool to evaluate different options using typical IP blocks from industry.

1.3 Research Objectives

The research objectives of this work are as follows:

- Study the performance of decaps using frequency-domain and time-domain analysis
- Develop useful modeling equations for design purposes
- Determine decap layout techniques to improve decap performance based on frequency and time-domain analysis
- Develop heuristics for location and size of decaps for IP blocks in standard cell layouts

1.4 Organization of the Thesis

The rest of this thesis is organized as follows. Chapter 2 provides the necessary background for supply noise, power grid design, IP block design using standard cells and additional issues surrounding decoupling capacitor design.

Chapter 3 explores the high-frequency issues associated with decap performance. It develops simplified equations to model the behaviour of decaps in the frequency domain. These equations can be used to understand the effect of scaling on high-frequency performance of decaps and provide first-order sizing information.

Chapter 4 explores the time-domain behaviour of decaps. Initially, the results of Chapter 3 are validated using transient analysis. Then, a simple equation for the required decoupling capacitance for a given buffer size is derived using charge-sharing. Next, various options for standard-cell layout of decaps are explored to optimize overall performance.

Chapter 5 addresses the issue of decap design for IP blocks constructed from standard cells. In this chapter, the questions of where to place the decaps and how much to use are explored using a commercial simulation tool, and a standard cell placement and routing CAD flow.

Chapter 6 summarizes the results of the thesis and provides directions for future research.

Chapter 2

Background

2.1 Introduction

This chapter begins with an overview of problems associated with power supply noise and its detrimental effects on IC logic circuit timing and functionality. This is followed by a description of methods to reduce power supply noise; in particular, the use of decoupling capacitors. Then, a brief description is provided on how typical power grids are designed. An overview of the standard cell layout approach is also described. Next, an illustration is given of the type of analysis necessary to identify block-level and chip-level noise in integrated circuits. Finally, some of the design issues associated with decoupling capacitors are highlighted to motivate the topics in the rest of the thesis.

2.2 Power Supply Noise

Ideally, a circuit's power supply should be constant and equal to the desired value of V_{DD} or G_{ND}. However, noise due to a variety of effects tends to disturb the supply voltage from its ideal level. For example, noise can be induced by the switching of digital circuits that demand current from the power supply lines. Power-supply noise refers to unwanted time varying voltage levels in the power-supply lines that act to reduce the performance of an integrated circuit. Figure 2.1 displays two parasitic elements that contribute to power supply noise: resistance in the metal lines and inductance associated with bond wire or solder bump connections of the chip. Current through the resistance creates one form of noise known as *IR*-drop, while the *rate-of-change* of current through the inductive parasitics creates a second form of noise, *Ldi/dt*, sometimes called "L-di-by-dt" noise or delta-I noise. In the figure, there is also a representation of the type of voltage variation that one may observe on the supply line versus time. The switching of the near-end (inv1) and far-end (inv2) inverters causes the voltage at the far-end of the supply line to droop below the ideal level of V_{DD}. The graph in the figure shows that the noise caused by the *far*-end inverter causes a bigger droop than the *near*-end inverter.

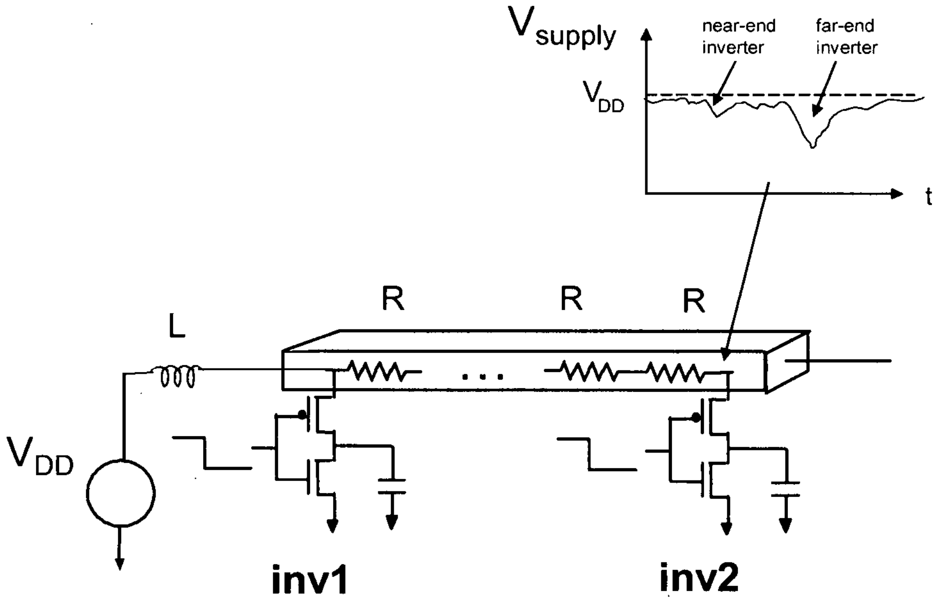


Figure 2.1: A noisy power supply line.

For digital circuits, this variation in the supply voltage results in undesirable effects such as increased gate and clock delays [14] and a potential loss of functionality. For example, the gate delay of a MOS inverter depends on its drive capability, which in turn is controlled by a term called the gate overdrive, $V_{DD} - V_T$, where V_T is the device threshold voltage. If V_{DD} is reduced in value due to noise, there is a corresponding reduction in the drive, which then increases the delay. Of course, if the supply voltage is reduced by too great a value, there may be a complete loss of functionality of the gate during the duration of the noise surge. If it takes too long for the signal to toggle or if an extremely low V_{DD} (or high V_{SS}) prevents the signal from toggling, then the incorrect value will be generated at the gate output and become latched in a flip-flop, cause the logic circuit to operate improperly.

A number of different approaches can be used to reduce the amount of supply noise generated in a circuit. A short list of the possible options are as follows:

- Increase the number of V_{DD} or V_{SS}[1] pins to reduce the average resistance between pin and gate.
- Re-route power and ground to deliver more current at a reduced impedance level [15]
- Re-design the chip floorplan to spread out noisy-blocks [17]
- Re-design the logic or clock circuits and their timing to reduce instantaneous current demands
- Use decoupling capacitors wherever there are high-levels of noise [16]-[24].
- Use voltage regulators

Typically, all of these options are used in one form or another to reduce noise. This thesis focuses on the use of decoupling capacitors.

[1] Note V_{SS} and G_{ND} are used interchangeably in this thesis.

2.3 Layout of Power Grids

Before getting into the details of decoupling capacitor design, the manner in which power grids are usually designed is given to provide a context for the use of decoupling capacitors. The role of the power and ground grids are to deliver the necessary voltage and current to all the blocks in the design. In a digital design, this includes IP blocks, memory, clocks (including PLLs), I/O, etc.

Power and ground lines are routed on the various layers of metal available in a given CMOS technology. Today, there may be 8 layers of metal, or more. Each layer is routed in an orthogonal direction to the layer above and below it. This is illustrated for Metal4 and Metal5 in Figure 2.2 with vias indicated to connect the two layers to each other. Other even and odd-numbered layers are oriented in the same way as Metal 7 and Metal 8, respectively. The resulting pattern of the power distribution system is a grid, hence the name *power grid.*

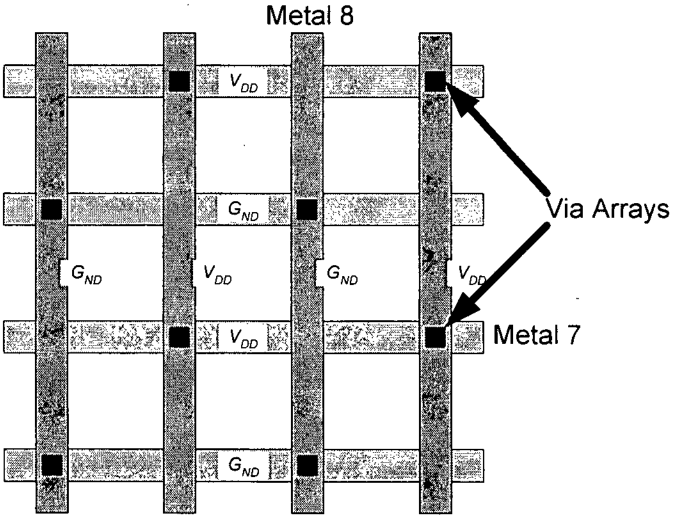


Figure 2.2 Orthogonal Routing of Each Layer of a Power Grid

Eventually, the metal lines must be connected to the PMOS and NMOS transistors. This is typically done on Metal 1. The exact nature of the metal routing at the lowest layers depends on the design style. In the next section, V_{DD} and G_{ND} routing for standard cell design will be described. The width and spacing of the metal lines at the higher levels are determined by the number of major trunks needed, the number of other global signals such as clocks, busses, etc. that must be routed, and the current demands of the various blocks in the design. While the grid-like structure represents one way of designing the power system, a number of variations are possible.

As mentioned earlier, careful layout of the power lines is one method to reduce noise. There are a variety of ways to accomplish this. First note that each segment of the power distribution system introduces some resistance, capacitance and inductance. The concern here is with resistance since it leads directly to *IR*-drop. The resistance of a segment is given by:

$$R = \frac{\rho L}{A} = \frac{\rho L}{TW}$$

where ρ is the resistivity of the metal, L is the length, T is the thickness and W is the width (A is the cross-sectional area). Clearly, running shorter and wider wires will reduce the resistance. This is illustrated in Figure 2.3. In the figure on the left, V_{DD} is being routed across two blocks in the manner shown. The far end of the wires will experience larger *IR*-drop. This can be reduced by using routing the trunks so that the distance from block to trunk is reduced as shown in the diagram on the right.

Usually, the different routing approaches are not enough to eliminate all *IR*-drop so on-chip decoupling capacitors are required to further alleviate power-grid noise. Previous work on the modelling and design of power grids can be found in [25]-[33].

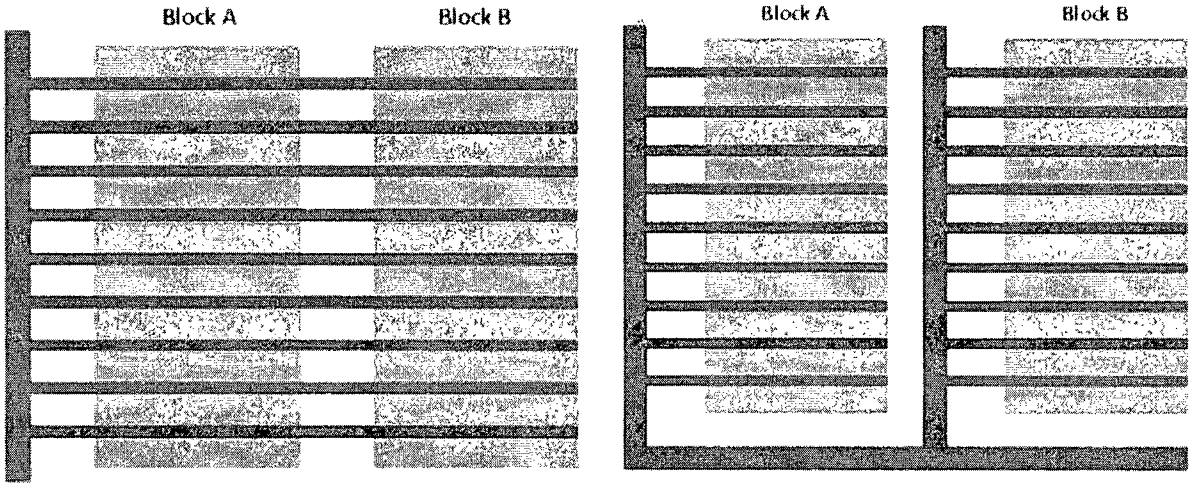


Figure 2.3: Reducing *IR*-drop. Block B on the right experiences less noise due to the added power 'trunk'.

2.4 Standard-Cell Placement

IP blocks are specified in a register-transfer-level (RTL) language such as VHDL or Verilog. This representation is converted to a layout using an ASIC design flow. A simplified version of this flow is represented in Figure 2.4. The RTL is synthesized into gates and then a gate-level simulation is performed. This comprises the front-end of the design process. Of particular interest here is the so-called back-end flow where the physical design is carried out. In this phase, the gates are replaced by elements from a standard cell library. These standard cells are placed and then routed. Once these two steps are completed, a timing verification is performed to ensure that the timing specifications are met. Finally, certain physical verification steps are carried out such as design-rule checking, layout vs. schematic, *IR*-drop analysis, antenna violation checks and signal coupling analysis. Once all these steps complete and show no violations, the layout is ready for the tape-out process and the electronic data is sent for fabrication[1].

[1] Note that there are many iterative loops in the flow that are not shown in the figure. If any of the steps cannot be implemented, the designer must go back to an earlier step and try again until the design can be taken from beginning to end without any violations.

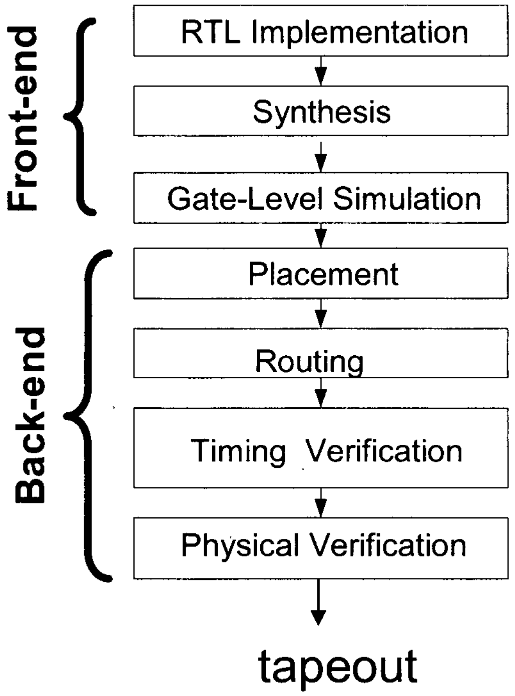


Figure 2.4 Conventional ASIC Flow

Examples of the detailed layout of two standard cells, a NAND gate and a NOR gate, are shown in Figure 2.5. The V_{DD} metal line runs across the top of each cell while the G_{ND} metal line runs across the bottom of each cell. The PMOS devices are in the upper portion (since n-well regions are defined here) while the NMOS devices are in the lower portion (where p-wells are defined). Contacts are placed as necessary between metal, polysilicon and diffusion areas, and for well contacts. Note that the cells are of fixed height and variable width.

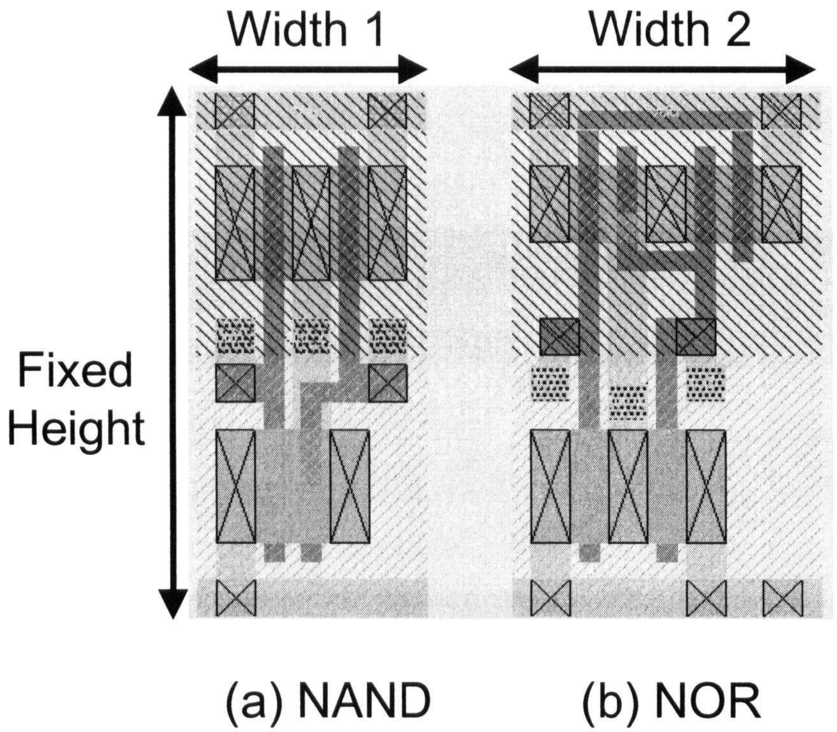


Figure 2.5 Two Standard-Cell Layout Examples

During the placement operation, standard cells are automatically placed in rows. Since each cell has the same height but differing widths, when cells are placed adjacent to one another, the power and ground lines are connected by abutment. This is also true of the n-well and p-well regions. The routing of signal lines is carried out after placement and may not always be possible to complete due to wire congestion in certain areas of the design. A depiction of a standard cell layout is provided in Figure 2.6 along with a detailed representation of portions of the first two rows of the layout. The top-left most portions have the actual layout, the sections to the immediate right have cell name representations, and the remaining cells are shown simply as bounding boxes.

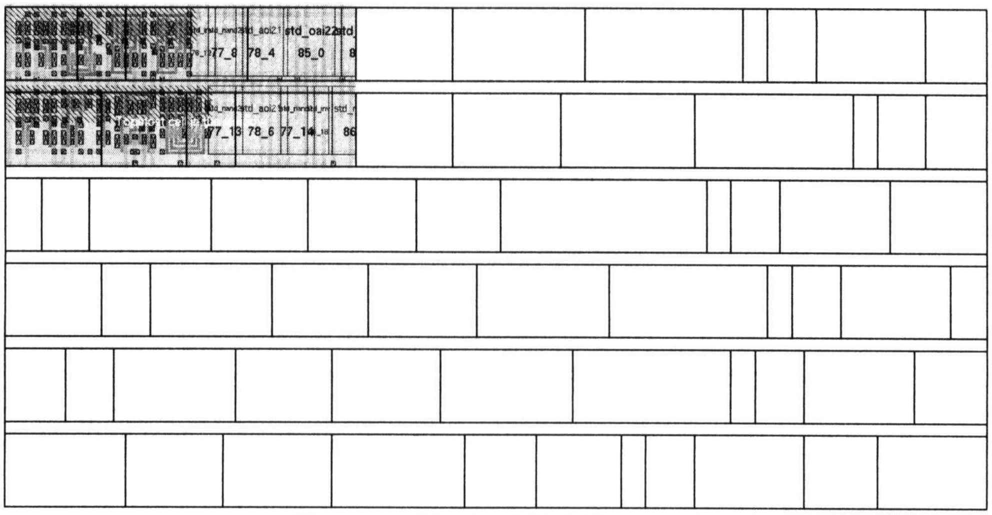


Figure 2.6 Standard Cell Placement

Because the total area of the cells when placed may not equal the area of the entire block, there will be some empty spaces between cells. These empty areas are ideal candidates for the placement of decap cells. As with normal standard-cells, decap cells must adhere to the rules of layout and placement as described above, that is, any decaps that are needed must be laid-out with a fixed height and variable width, and must satisfy all design rules of cells placed adjacent to them.

Now consider the plot shown in Figure 2.7. This is an *IR*-drop plot from a commercial tool showing violations of the supply noise budget for an industry-placed block. It was obtained by representing the power grid by its corresponding resistances and capacitances, and then adding current sources in place of the logic gates in the circuit. An analysis was performed to determine the voltage at each point in the grid due to current drawn by the gates. The voltage levels were then colour-coded to easily identify the "hot-spots" for *IR*-drop. The red areas represent cells that violate the 10% noise margins, all other colours signify cells with noise between 0-10% and black spots indicate empty spaces between cells. These problem areas must be fixed through the proper sizing and placement of decoupling capacitor cells before the IP block can be safely used.

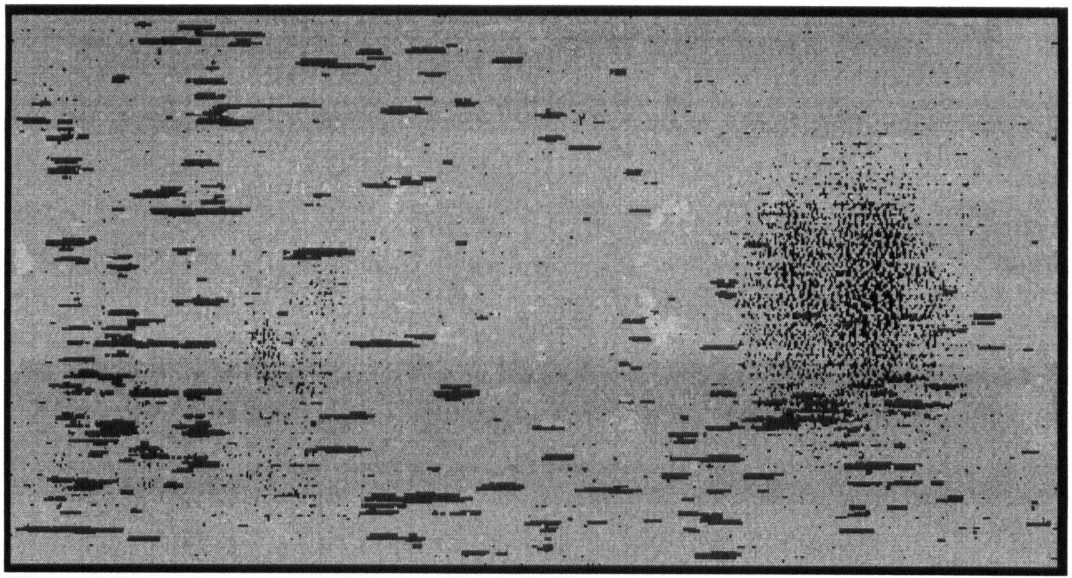


Figure 2.7: A noise-contour plot of an IP block. The red areas represent *IR*-drop 'hot-spots'.

2.5 Decoupling Capacitors

To better understand how decoupling capacitors reduce *IR*-drop, consider Figure 2.8a which depicts a row of logic-level standard-cells with a hypothetical segment of V_{DD} that violates the noise margin in the vicinity of the inverter framed by the dotted lines. The black boxes between these gates represent unused areas of the rows (just like the black dots in the previous diagram). Placement of the decoupling capacitors of sufficient size in these areas would prevent the violation of the noise margin around the inverter (Figure 2.8b).

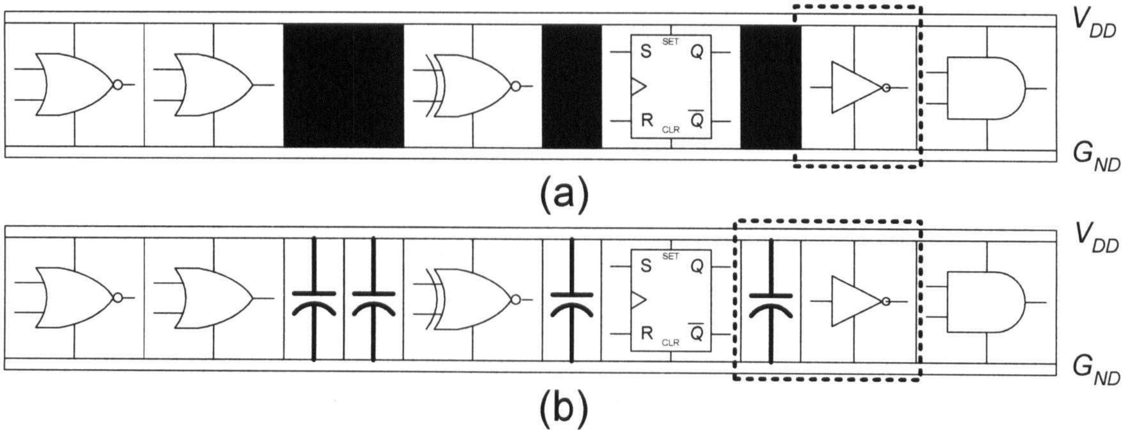


Figure 2.8: An example of a row of standard cells (a) without decaps and (b) with decaps

To illustrate this further, consider the inverter and capacitor framed by the dotted lines in Figure 2.8b which is redrawn in schematic form in Figure 2.9. This diagram also represents the supply voltage and resistance of the metal segments leading up to the inverter but ignores all other elements connected to it. The decoupling capacitor placed in close proximity to the switching inverter acts as a reservoir of charge that provides instantaneous current needed when the inverter switches. When the input of the inverter switches from logic '1' to '0' it will draw charge primarily from the capacitor instead of from the power source, V_{DD}, through the resistor, R, thus preventing IR-drop.

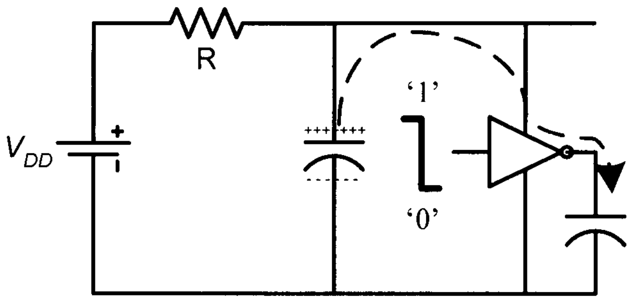


Figure 2.9: A decoupling capacitor used to reduce noise.

Decoupling capacitors are constructed using MOSFETs with the gate connected to one power line and the source and drain to the other. MOSFET's happen to be very convenient structures to use as decoupling capacitors since they do not require new processes and possess thin oxides that provide substantial capacitance. In the case of NMOS decaps, the gate is connected to V_{DD} with source and drain connected to G_{ND}. For PMOS decaps, it is the other way around with the gate connected to G_{ND} and the drain/source connected to V_{DD}. In standard cells, it is possible to make decaps using both types of transistors, as shown in Figure 2.10. This is because the upper half of the cell is usually reserved for PMOS devices while the lower half is for NMOS devices.

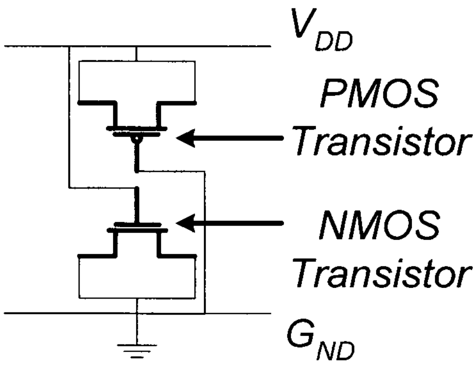


Figure 2.10: A pair of MOSFETs used as decoupling capacitors.

One possible standard cell layout of the decap above is illustrated in Figure 2.11. As expected, the NMOS decap occupies the bottom half of the cell with the PMOS decap in the upper half. The polysilicon gates are laid on top of the channel region of the MOS transistor to create the thin-oxide capacitor. Note that the cells are fixed height so that the size of the decaps are determined by the channel length of the device (horizontal dimension in the figure). This implies that large decaps will require very long channels.

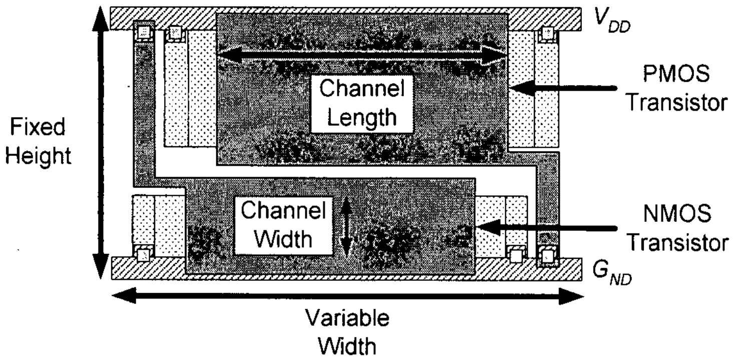


Figure 2.11: An example decoupling capacitor for standard cell layouts.

The problem with decaps laid out in this fashion is that they have poor performance at high frequencies. Therefore, the frequency response of MOS transistors with varying channel lengths must be investigated and the most effective value for layout purposes be determined. This is the main purpose of the next chapter of this thesis.

Chapter 3

High-Frequency Response of Decoupling Capacitors

3.1 Introduction

At first, the use of MOS transistors as decoupling capacitors appears to be a straightforward solution to the CMOS decap design problem. However, non-idealities such as nonlinear resistive and capacitive effects and high-frequency response characteristics associated with the MOS device must be taken into account in order to achieve the desired performance. This is because, as will be shown, the MOSFET decap is neither an ideal capacitor nor does it hold its capacitive value at high frequencies. This chapter focuses on frequency-domain analysis and some first-order design issues for MOS decaps. It also examines technology trends to determine whether the results will hold for upcoming CMOS generations.

To begin, the MOSFET and its characteristics in the vicinity of the channel region will be analyzed. Ideally, the transistor should have a well-defined capacitance value that is determined by the width, length and oxide thickness of the device. In an ideal capacitor, the charge on the plates always equals CV. If the voltage were to change quickly by ΔV, then the change in charge ΔQ will be $C\Delta V$ and the relationship $Q = CV$ is maintained. A model that treats a MOS transistor in this manner is known as a quasi-static model (QS). This type of assumption is quite valid for most types of analyses carried out in programs such as HSPICE, especially for time-domain analysis of digital circuits with short channel lengths [34].

The actual behaviour of a MOSFET differs from the idealized behaviour above, particularly in the case of decaps where the required area to realize large capacitance values often results in the use of long channel devices. First, the channel of the MOSFET exhibits parasitic resistance that degrades transient performance as shown in Figure 3.1. [34]

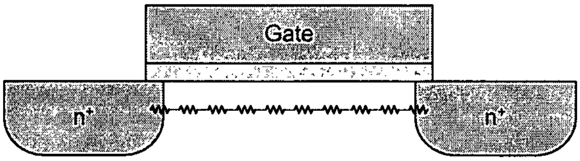


Figure 3.1: Channel Resistance.

Second, a non-ideality is encountered during high-frequency operation. When the voltage at the gate terminal changes too rapidly, the channel charge cannot respond quickly enough to match the charge on the top plate of the capacitor (see Figure 3.2). In the MOSFET, the change in charge is not instantaneous since mobile carriers from the source and drain cannot redistribute themselves at the rate determined by the gate voltage. This process is depicted in Figure 3.2. The response time of the channel charge is controlled by the *transit time* of the device which is proportional to L^2. This is particularly troublesome for decaps with long channels operating at high frequencies, since the effective capacitance is greatly reduced as the frequency increases.

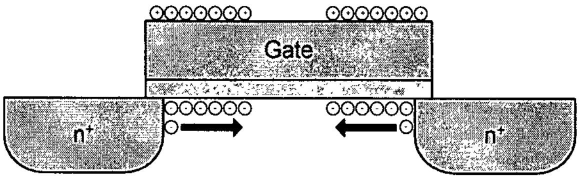


Figure 3.2: Transit time effect reduces effective capacitance.

MOSFET models that capture this type of behaviour are known as non-quasi-static (NQS) models. As mentioned earlier, most models in use today are typically *quasi*-static [34]; that is, the charge below the gate is formed immediately when voltage is applied. However, as will be shown in this chapter, the NQS model is essential when simulating decaps used in IC's running in the gigahertz range of operation. One additional issue to resolve is that the gate capacitance and channel resistance are really distributed between the source and drain. This distributed nature of the channel can be modeled as a lumped *RC* circuit where both the resistance and capacitance are functions of frequency, as shown in Figure 3.3. The extraction of these parasitics is discussed in the next section.

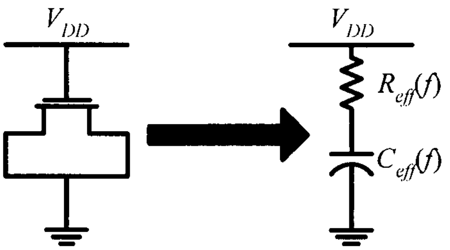


Figure 3.3: Modeling the decap as series *RC*.

3.2 Review of AC Analysis of *RC* Circuits

Much of the analysis in this chapter will be carried out in the frequency domain, so a brief tutorial on the basics of AC analysis is required beginning with a short review of *RC* circuits such as those shown below in Figure 3.4. Assume that the input to each circuit is a sinusoidal voltage with a frequency *f*. The operating frequency is usually represented as $\omega = 2\pi f$ in units of radians per second. For each of the circuits, the current in the loop can be computed as $I = \frac{V}{Z}$.

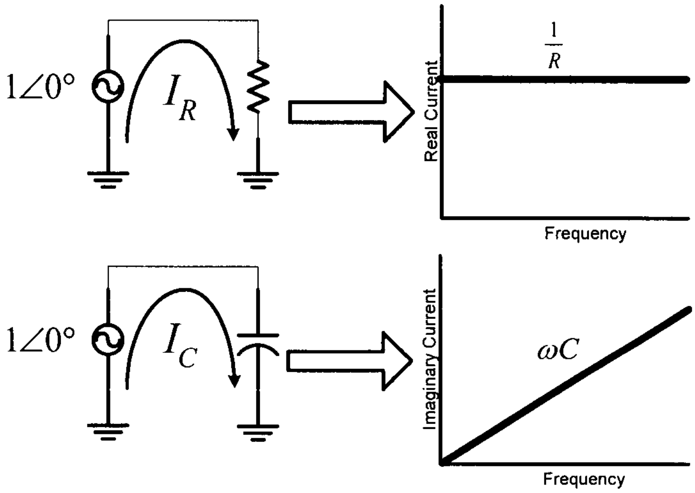


Figure 3.4: Resistor and Capacitor Circuit Analysis

The impedance of the resistor is, of course, *R*, while the impedance of a capacitor is *1/jωC*. Therefore, the currents through these two circuits (assuming $V = 1\angle 0°$ where the notation $M\angle\theta$ refers to a sinusoidal signal with magnitude *M* and phase *θ*) is:

$$I_R = \frac{V}{Z_R} = \frac{1}{R}$$

$$I_C = \frac{V}{Z_C} = \frac{1}{\frac{1}{j\omega C}} = j\omega C$$

Notice that the resistor's current is purely real and independent of frequency while the capacitor's current is completely imaginary and linearly dependent on frequency. These two properties are shown in the corresponding plots of the real and imaginary currents in Figure 3.4.

In Figure 3.5, a $1\angle 0°\text{V}$ AC voltage source is applied to a series *RC* circuit based on Figure 3.3. The full current equation is:

$$I_{RC} = \frac{V}{R_{eff} + \frac{1}{j\omega C_{eff}}}$$

Now setting $V = 1\angle 0°\text{V}$ as in Figure 3.5:

$$I_{RC}=\frac{1}{R_{eff}+\frac{1}{j\omega C_{eff}}}=\left(\frac{\omega^2 R_{eff}C_{eff}^2}{\omega^2 R_{eff}^2 C_{eff}^2+1}\right)+j\left(\frac{\omega C_{eff}}{\omega^2 R_{eff}^2 C_{eff}^2+1}\right)$$

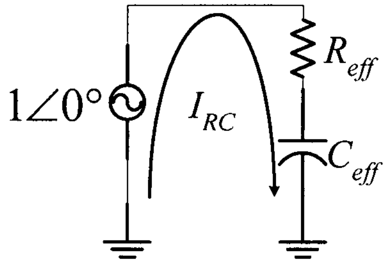


Figure 3.5: *RC* circuit.

Therefore, the real, imaginary and magnitude components are:

$$\mathrm{Re}(I_{RC})=\frac{\omega^2 R_{eff}C_{eff}^2}{\omega^2 R_{eff}^2 C_{eff}^2+1} \tag{3.1}$$

$$\mathrm{Im}(I_{RC})=\frac{\omega C_{eff}}{\omega^2 R_{eff}^2 C_{eff}^2+1} \tag{3.2}$$

$$\mathrm{Mag}(I_{RC})=\sqrt{\left(\frac{\omega^2 R_{eff}C_{eff}^2}{\omega^2 R_{eff}^2 C_{eff}^2+1}\right)^2+\left(\frac{\omega C_{eff}}{\omega^2 R_{eff}^2 C_{eff}^2+1}\right)^2} \tag{3.3}$$

Setting $R_{eff}=1\mathrm{k}\Omega$ and $C_{eff}=50\mathrm{fF}$, SPICE can be used to plot the three equations as a function of frequency as shown in the graph in Figure 3.6.

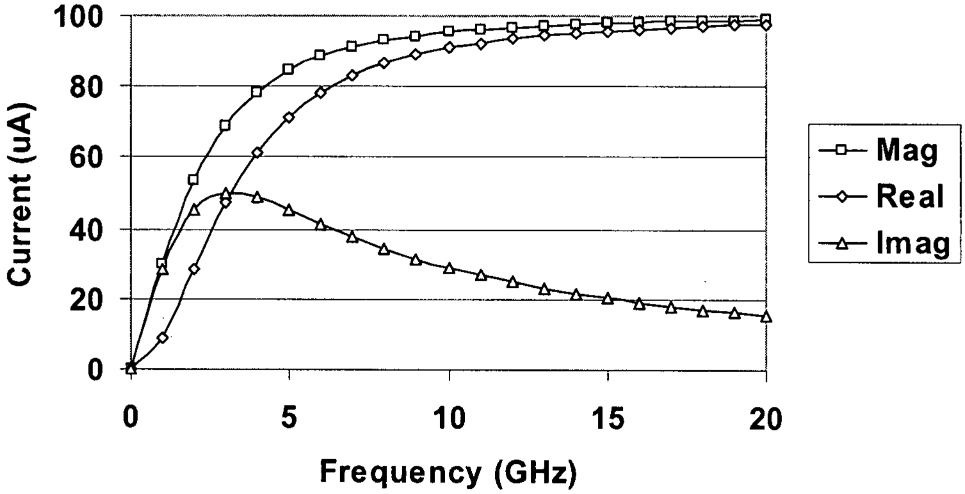


Figure 3.6: The components of the current through the *RC* circuit.

Because it is straightforward to plot the AC currents of a MOSFET in SPICE, it is possible to obtain plots of the effective resistance and capacitance values by using Eqns. (3.1)-(3.3). First the resistance is derived by dividing the real component by the square of the magnitude:

$$R_{eff} = \frac{\mathrm{Re}(I_{RC})}{\mathrm{Mag}^2(I_{RC})} \tag{3.4}$$

Next, the capacitance is derived by taking the square of the magnitude and dividing it by the imaginary component, the frequency and the voltage:

$$C_{eff} = \frac{\mathrm{Mag}^2(I_{RC})}{2\pi f\,\mathrm{Im}(I_{RC})} \tag{3.5}$$

These two equations will allow one to extract the series resistance and capacitance of the MOSFET from a SPICE simulation as a function of frequency.

3.3 *R_{eff}* and *C_{eff}* Derivation

The DC ($f = 0$) resistance and capacitance of the MOSFET are now derived. First, looking at the channel resistance, it might be tempting to derive the equation by using g_m since the gate voltage is varying and this will affect the drain-to-source current. However, using g_m creates a problem as follows:

$$I_{DS}\big|_{lin} = \mu C_{OX}\frac{W}{L}\left[(V_{GS}-V_T)V_{DS} - \tfrac{V_{DS}^2}{2}\right]$$
$$g_m = \frac{\partial I_{DS}}{\partial V_{GS}} = \mu C_{OX}\frac{W}{L}V_{DS}$$
$$R_{eff} = \frac{1}{g_m} = \frac{L}{\mu C_{OX} W V_{DS}}$$

As can be seen, when drain and source are grounded together as in an NMOS decap, $V_{DS} = 0$ and R_{eff} becomes infinite. While this type of equation is derived in a number of textbooks, it is not suitable for use when working with decaps. A more appropriate method for the channel resistance calculation would be to base it on the output resistance, r_0, which is related to g_{ds}:

$$g_{ds} = \frac{1}{r_0} = \frac{\partial I_{DS}}{\partial V_{DS}} = \mu C_{OX} \frac{W}{L}\left(V_{GS} - V_T\right) - V_{DS}$$

$$R_{eff} = r_0\big|_{V_{DS}=0} = \frac{L}{\mu C_{OX} W\left(V_{GS} - V_T\right)}$$

However, this equation is not quite accurate either! From [34], the expression for the channel resistance requires a pre-multiplier as follows[1]:

$$R_{eff} = \left(\frac{3\alpha^3 + 15\alpha^2 + 10\alpha + 2}{10(1+\alpha)(1+2\alpha)^2}\right)\frac{L}{\mu C_{OX} W (V_{GS} - V_T)}$$

where α is:

$$\alpha = \begin{cases} 1 & \text{if} \quad V_{DS} < V_{DS,sat} \quad (\text{linear}) \\ 0 & \text{if} \quad V_{DS} > V_{DS,sat} \quad (\text{saturation}) \end{cases}$$

Since the device is in the linear region, $\alpha = 1$. For this special case, the substitution of this value produces the effective resistance of a decoupling capacitor:

$$R_{eff} = \frac{1}{6}\frac{L}{\mu C_{OX} W (V_{GS} - V_T)} \qquad (3.6)$$

This new equation for R_{eff} is reasonably accurate, as will be shown later.

The expression for capacitance is more straight-forward. The first-order expression for the sum of the thin-oxide and overlap capacitance is given by [15]:

[1] The details of this modification are beyond the scope of this thesis but are well described in [34].

$$C_{eff} = C_A + C_F = C_{ox}WL + 2C_fW \tag{3.7}$$

where C_{ox} represents the oxide capacitance per unit area and C_f is the fringing and overlap capacitances per unit width of the channel. The fringing capacitance is associated with the edges of the polysilicon gate as shown in Figure 3.7.

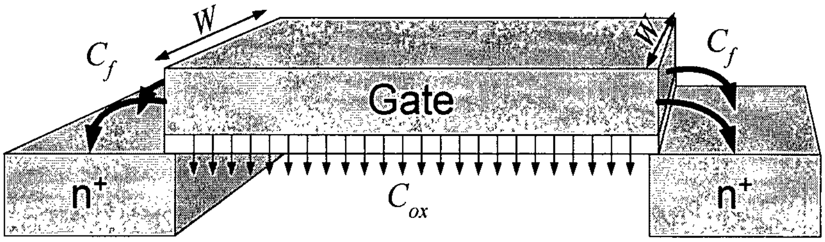


Figure 3.7: The gate capacitances of the decoupling capacitor.

The complete behaviour of the decap is modeled with the circuit in Figure 3.8.

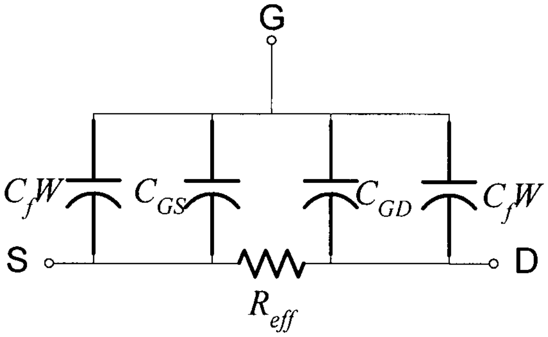


Figure 3.8: Basic RC Model for Decaps

3.4 Problems using the Quasi-Static Models in SPICE

The equations of the previous section are only suitable for DC scenarios ($f = 0$). In reality, the resistance and capacitance values change with frequency, a behaviour not found in the quasi-static models as will be seen shortly. Since the transistor cannot respond quickly enough at high frequencies, the resistance and capacitance will be frequency-dependent.

Why do the resistance and capacitance drop with frequency? Referring back to Figure 3.2, an intuitive first-order explanation is that, at low frequencies, all the mobile carriers are able to respond fast enough to provide negative charge on the bottom plate of the channel. However, at higher frequencies, a smaller percentage of the needed charge is available due to the transit time limitations of the device. Effectively, the shorter distance traveled by the charge results in a smaller resistance and smaller capacitance.

It is important that this effect be modeled in circuit simulators to obtain accurate results. Quasi-static models are not suitable for high-frequency analysis of MOS circuits, in general, and decaps, in particular. To illustrate the possible errors in the model, BSIM version 3v3 (BSIM3v3) was used in quasi-static mode to generate the frequency response of a number of decaps with differing channel lengths. The experimental setup in Figure 3.9 was used to extract the decap's resistance and capacitance. An NMOS decap was simulated with a fixed gate area of $400\lambda^2$ and lengths of 2λ, 4λ, 8λ, $16\,\lambda$ and 32λ, where $\lambda = 90\text{nm}$ for 180nm technology.

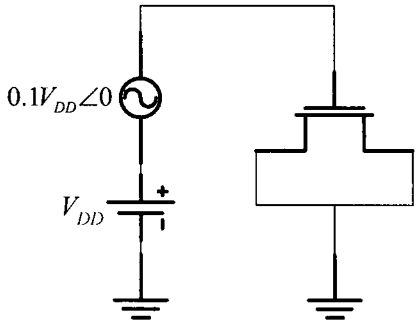


Figure 3.9: Circuit setup to extract effective Resistance and Capacitance.

Current plots were generated with HSPICE using 180nm TSMC technology and the BSIM3v3 model. Eqns. (3.4) and (3.5) and the real, imaginary and magnitude of the gate current were used to plot the effective resistance and capacitance as a function of frequency. These plots are shown in Figure 3.10 and Figure 3.11. Note that for the QS model, the resistance and capacitance plots are ‘flat’ and therefore independent of frequency. This is not correct. The plots should show both the resistance and capacitance values decreasing as frequency increases but the QS model does not account for this behaviour.

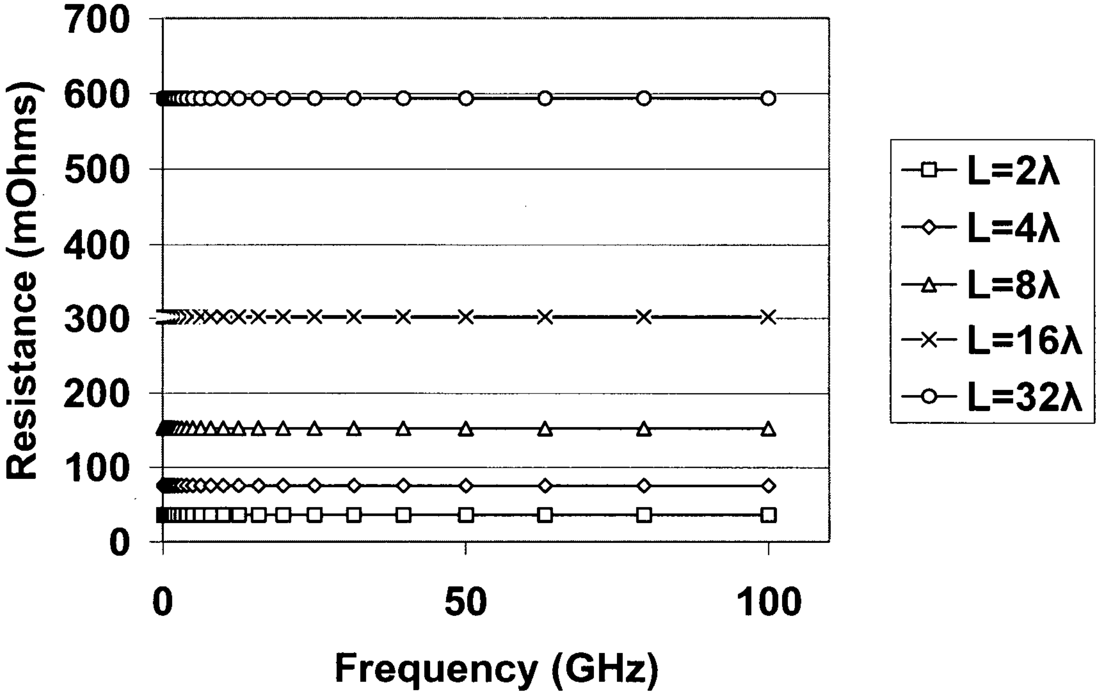


Figure 3.10: Resistance results for QS 180nm NMOS using SPICE.

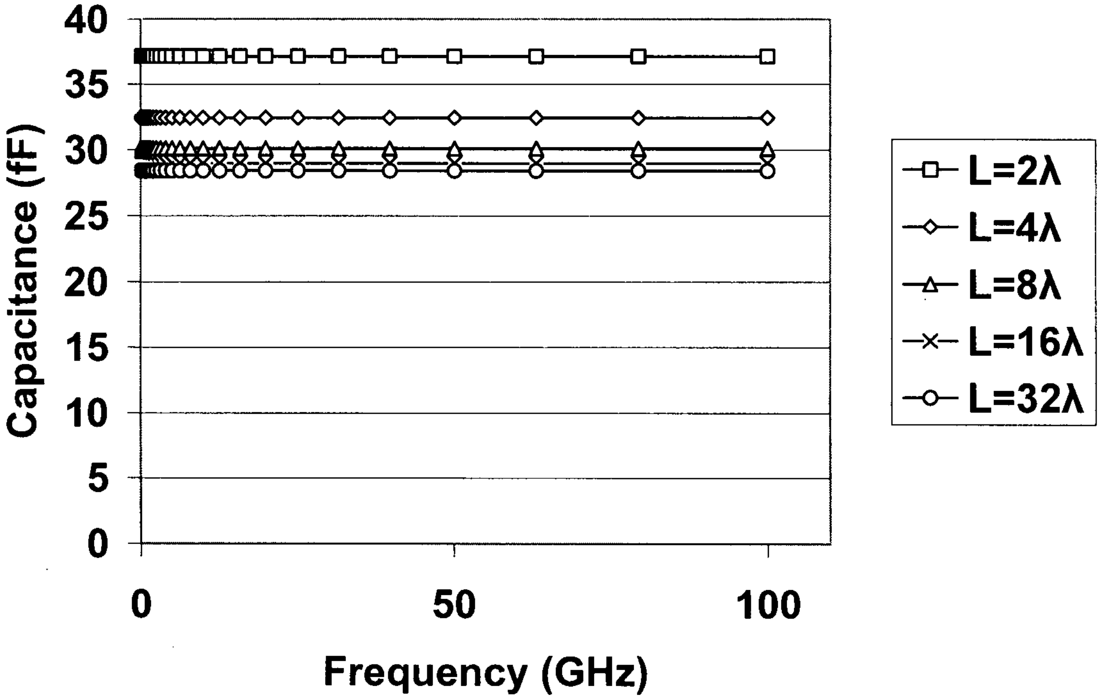


Figure 3.11: Capacitance results for QS 180nm NMOS using SPICE.

To further illustrate, the resistance and capacitance at $f=0$ are provided in Table 3.1. The results in columns 2 and 4 are computed using Eqns. (3.6) and (3.7) whereas columns 3 and 5 are derived from Figure 3.10 and Figure 3.11 at $f=0$. Clearly, there is a discrepancy with the resistance value but the capacitance values are fairly accurate. The discrepancies will be solved with the analysis of the NQS model in the next section.

Table 3.1: Predicted and extracted resistance and capacitance numbers for NMOS.

NMOS	**Resistance (Ω)**		**Capacitance (fF)**	
Length	R_{eff} Eqn. (3.6)	**HSPICE**	C_{eff} Eqn. (3.7)	**HSPICE**
2	5	0.037	37	37
4	21	0.076	33	32
8	86	0.152	31	30
16	343	0.302	30	29
32	1373	0.594	29	28

3.5 The Non-quasi-static (NQS) model

Recent BSIM models in SPICE capture the transit time effect of long channel devices called the non-quasi-static (NQS) effect. First the qualities of the NQS model will be demonstrated. To run SPICE in NQS mode, the variable NQSMOD is set to '1'. Using BSIM3v3's NQS model (as described in [34]) for TSMC's 180nm technology and the experimental setup in Figure 3.9, the effective NQS resistance and capacitance are plotted up to 100GHz[1] in Figure 3.12 and Figure 3.13. Notice this time that:

[1] 100GHz is an extremely high and unrealistic value for noise analysis but is used for the purposes of model development and understanding.

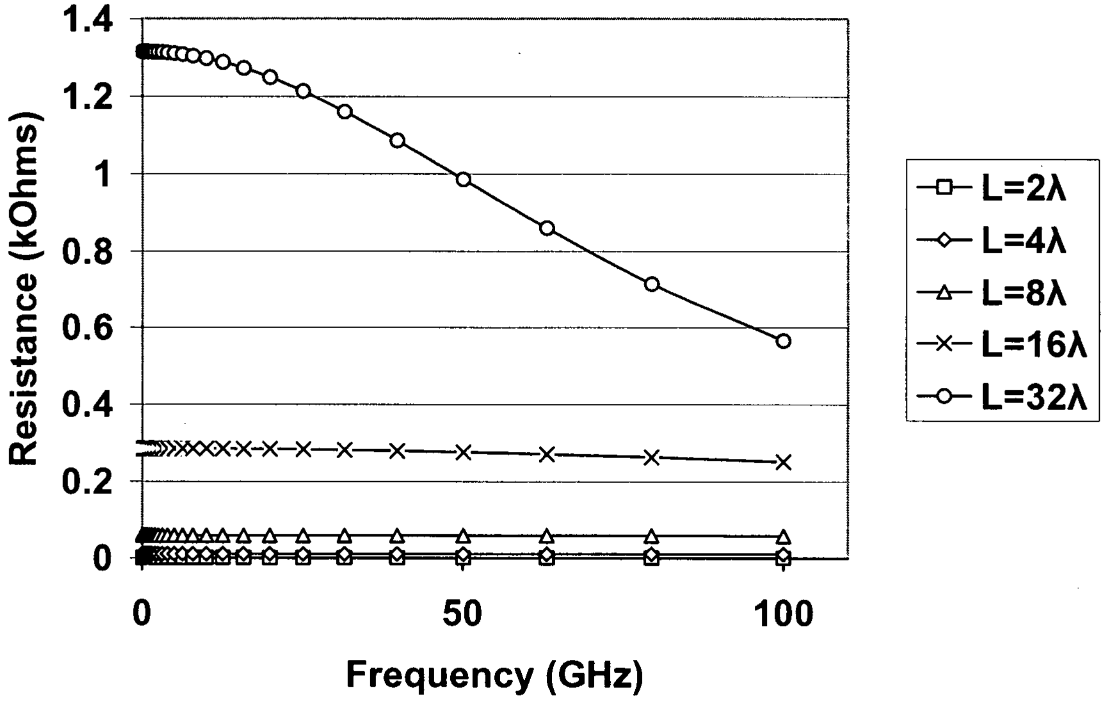


Figure 3.12: NMOS Resistance extracted from SPICE.

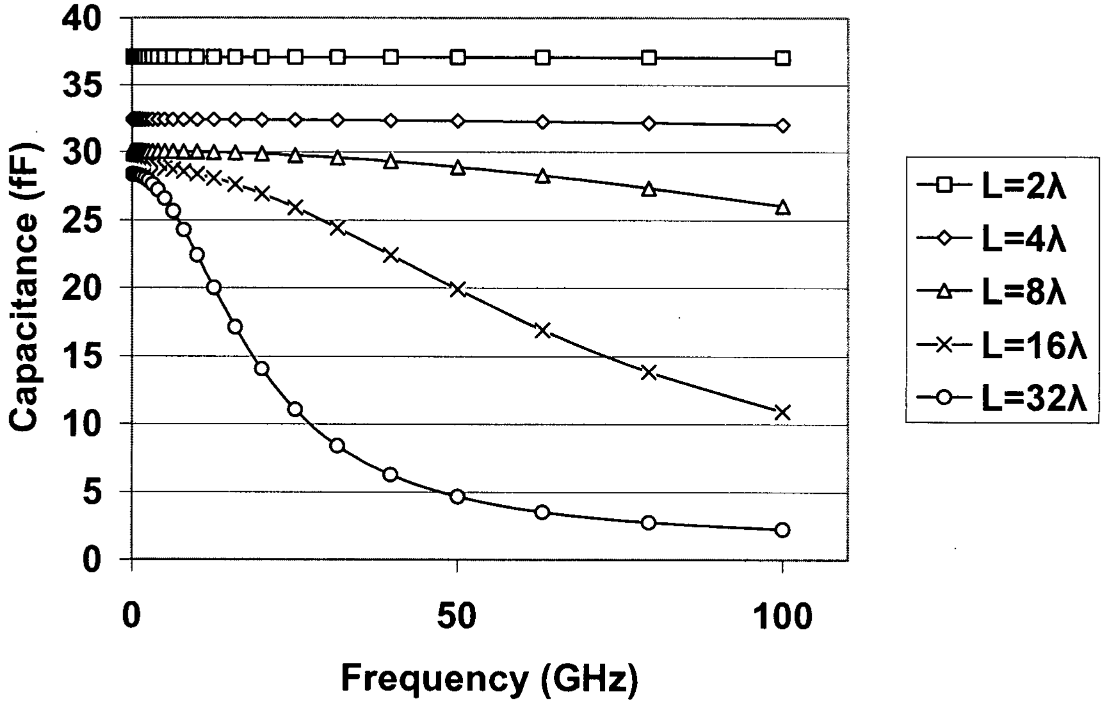


Figure 3.13: NMOS Capacitance extracted from SPICE.

- The effective resistance and capacitance *are* dependent on frequency.
- That effective resistance is in the range of Ω to kΩ and more closely matches the values calculated in Table 3.1.
- That for low frequencies, the capacitance values match those of the QS model.

The main reason for this frequency dependence is the finite response time of the channel charge as characterized by the transit time. A new expression for transit time is now derived for a MOS device connected as a decap. The transit time is the time it takes for a mobile carrier to travel the length of the channel. Its derivation is typically carried out by computing the maximum operating frequency of the device, as follows:

$$f_o = \frac{\omega_o}{2\pi} = \frac{g_{ds}}{2\pi C_G} = \frac{\frac{W}{L}\mu_n C_{ox}\left(V_{gs} - V_T\right)}{2\pi(WLC_{ox} + 2C_f W)}$$

To simplify the expressions, assume that $WLC_{ox} << 2C_f W$:

$$f_o \cong \frac{\mu_n\left(V_{gs} - V_T\right)}{2\pi L^2}$$

$$\tau_{tr} = \frac{1}{\omega_o} = \frac{L^2}{\mu_n\left(V_{gs} - V_T\right)}$$

The form of this equation makes intuitive sense. Using basic physics:

$$\text{velocity} = \frac{\text{distance}}{\text{time}}$$

$$\therefore t \propto \frac{d}{v} = \frac{L}{\mu E} = \frac{L^2}{\mu\left(V_{GS} - V_T\right)}$$

Again, from [34], there is a pre-multiplier similar to Eqn. (3.6):

$$\tau_{tr} = \left(\frac{3\alpha^3 + 15\alpha^2 + 10\alpha + 2}{10(1+\alpha)(1+2\alpha)^2}\right)\frac{L^2}{\mu_n\left(V_{gs} - V_T\right)}$$

$$\tau_{tr} = \frac{1}{6} \frac{L^2}{\mu_n \left(V_{gs} - V_T\right)} \tag{3.8}$$

For a 180nm technology, one can use this equation[1] to compute the value of f_o as shown in Table 3.2.

Table 3.2: Cutoff frequency for NMOS transistors of various lengths.

Length ($\lambda = 90\text{nm}$)	**f_o (GHz)**
2	179.1
4	44.8
8	11.2
16	2.8
32	0.7

Ideally, one would like to derive equations capturing the frequency-dependent behaviour of R_{eff} and C_{eff} of Figure 3.12 and Figure 3.13, respectively, that somehow involve the transit time. However, this is not straight-forward. Therefore, empirical models based on the equations given in [34] have been developed.

The form of the two empirical models is as follows:

$$R_{eff} = \frac{R_{eff,0}}{1 + \left(\frac{\omega}{\omega_0}\right)^2} \tag{3.9}$$

$$C_{eff} = \frac{C_{eff,0}}{1 + \left(\frac{\omega}{\omega_0}\right)^2} \tag{3.10}$$

where $R_{eff,0}$ and $C_{eff,0}$ are the values at $f = 0$ in Eqns. (3.4) – (3.5) and ω_o is given by:

$$\omega_o = \frac{\mu_n \left(V_{gs} - V_T\right)}{L^2}$$

[1] A further modification may be needed for the transit time. Since both source and drain are grounded, the maximum distance traveled by a carrier should be reduced by half.

The plots of these equations are shown in Figure 3.14 and Figure 3.15. The thick dotted lines represent the results of plotting Eqns. (3.9) and (3.10) superimposed on the plots from Figure 3.12 and Figure 3.13. These equations have similar general behaviour to the plots extracted from HSPICE, but not the same values. This is due to the actual value of the transit time in HSPICE compared to the one in Eqn. (3.8). Some adjustments are therefore needed.

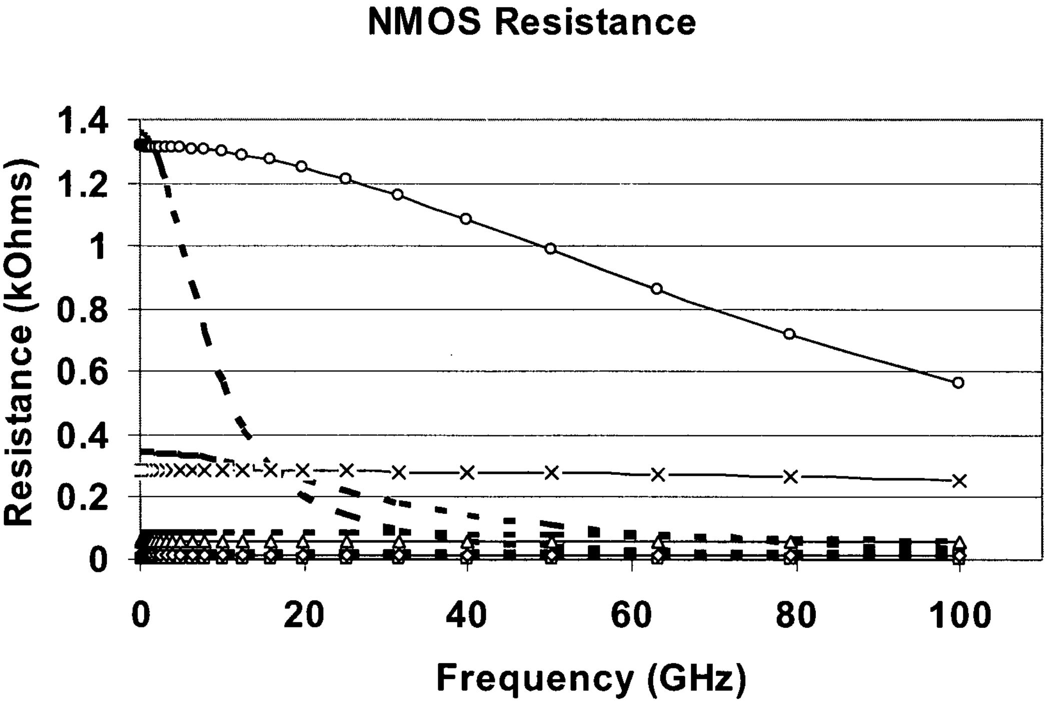


Figure 3.14: NMOS resistance calculated.

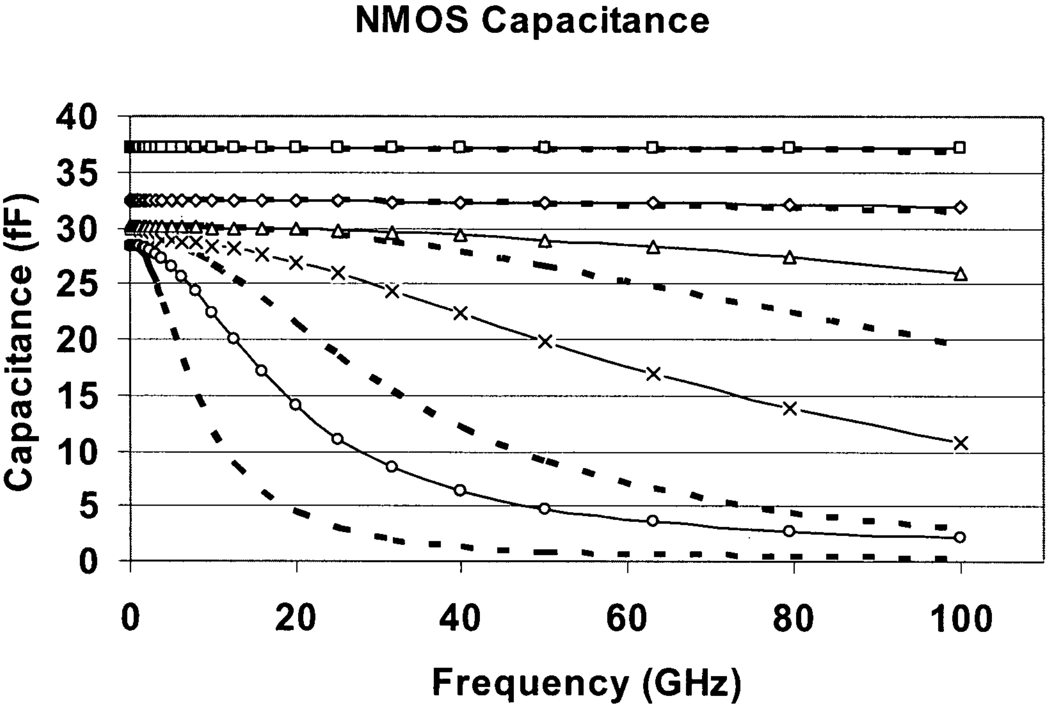


Figure 3.15: NMOS capacitance calculated.

A fitting parameter, β, is now introduced to adjust the transit time so that the curves will match:

$$R_{eff} = \frac{R_{eff,0}}{1+\left(\frac{\omega}{\beta\omega_0}\right)^2}$$

$$C_{eff} = \frac{C_{eff,0}}{1+\left(\frac{\omega}{\beta\omega_0}\right)^2}$$

Using these equations, the frequency response of both NMOS and PMOS devices were analyzed in HSPICE and, after curve fitting, $\beta = 2$ is obtained for NMOS devices and $\beta = 10$ for PMOS devices. The results are shown in Figure 3.16 and Figure 3.17.

Some general observations on the performance of NMOS and PMOS decaps can now be made. First, from Figure 3.16 and Figure 3.17, it is clear that NMOS is superior to PMOS in its high-frequency behaviour. This is due to the fact that the transit time is inversely proportional to the mobility according to Eqn. (3.8). Since the PMOS device has a lower mobility, by a factor of almost 4, it will have a higher transit time (see Eqn. (3.8)), and hence a lower ω_0. This implies that the PMOS device must have half the channel length of an NMOS device to obtain the same performance. Therefore, NMOS devices should typically be used to implement decoupling capacitor whenever there is a choice. Second, for the NMOS decaps, a length of 16λ can be used up to about 20GHz without seeing much degradation in performance. This corresponds to roughly 8λ for PMOS decaps. These are good starting points for decap sizing in 180nm technology.

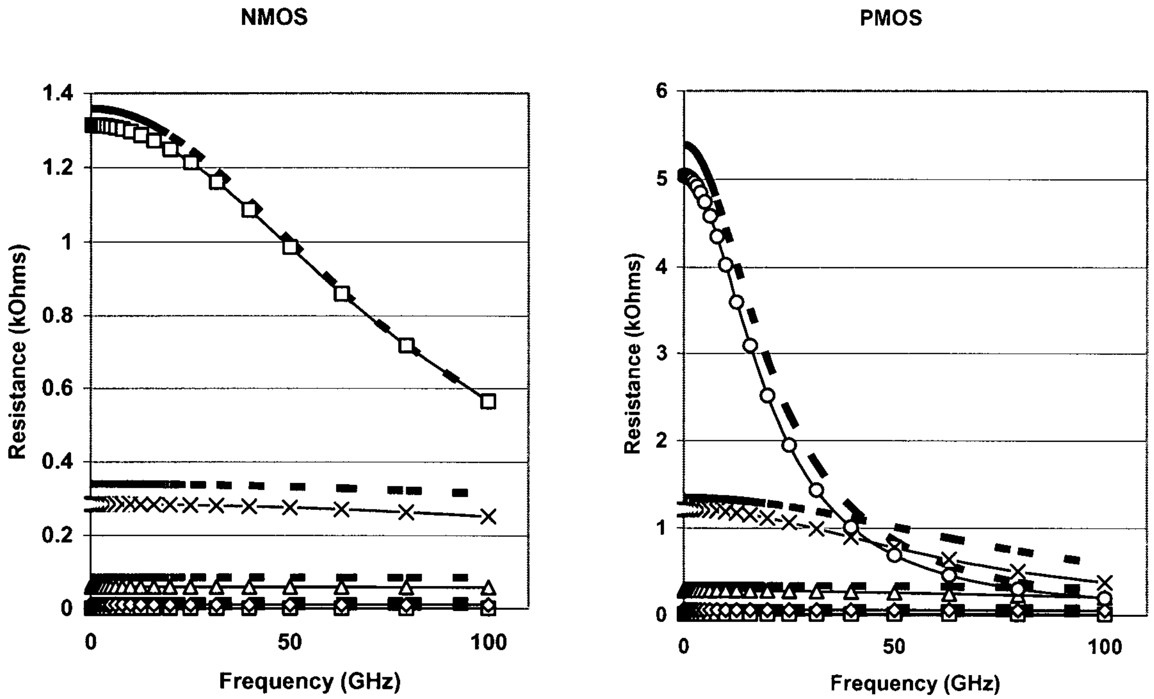


Figure 3.16: Curve fitting for 180nm resistance.

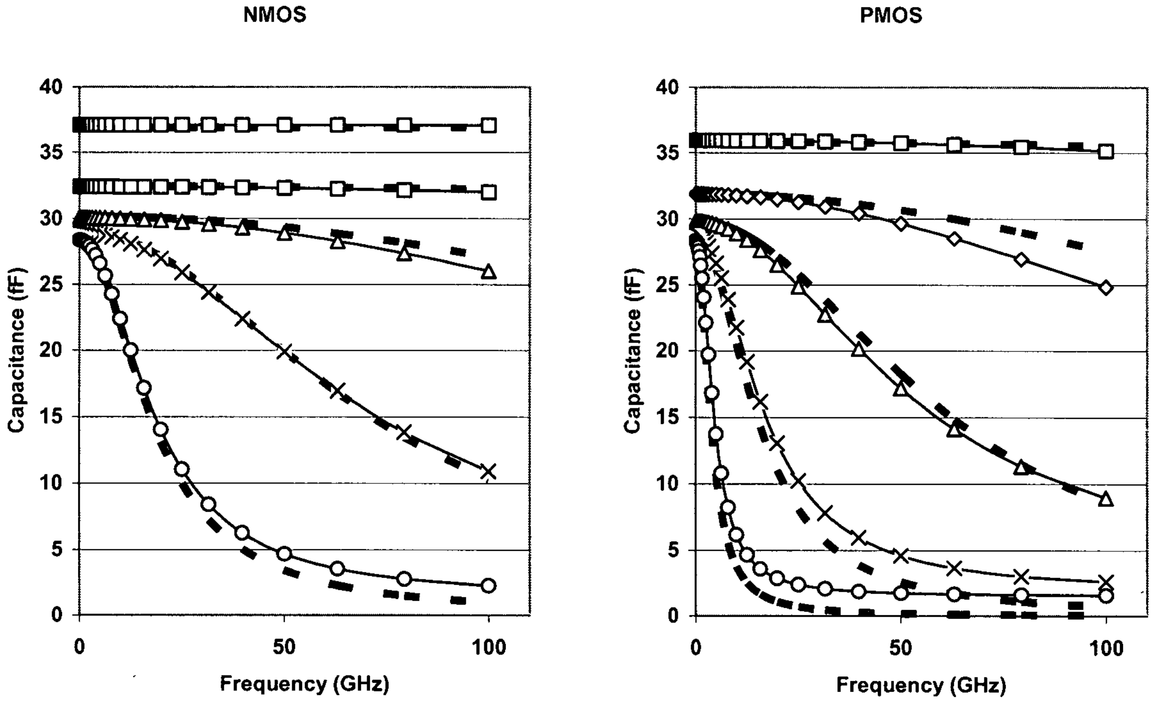


Figure 3.17: Curve fitting for 180nm capacitance.

3.6 Technology Scaling Trends for Decaps

The results shown so far are associated with 180nm technology. Since, as of this writing, the industry is beginning to adopt 130nm and will switch to 90nm in the next few years, it is worthwhile to examine whether the results obtained thus far still hold as technology scales. Using normalized values for the x- and y-axes, the effective resistance and capacitance is plotted for both NMOS and PMOS transistors using 180nm, 130nm and 90nm technology nodes and TSMC parameters. The results are shown in Figure 3.18 and Figure 3.19.

The most notable conclusion one can make about the trend is that the performance of a decoupling capacitor appears to *improve* with scaling. This is simply due to the fact that as technology scales, the value of λ is reduced and therefore the channel length is reduced. For example, if $L = 16\lambda$ in a 180nm technology, then the same channel length is obtained by setting $L = 32\lambda$ in a 90nm technology. The results for these two graphs are almost identical. In fact, Eqn. (3.8) tells us that this is expected, with the only difference being due to the term $V_{GS} - V_T$. Hence, it is the physical channel length that controls the performance of the decaps. The comparison of NMOS to PMOS decaps as technology scales also demonstrates that NMOS is better than PMOS.

The results obtained here are based on frequency-domain analysis which is based on linearized small-signal models. This is useful in obtaining first-order information on decaps. However, to properly assess their performance, one needs to switch to time-domain analysis where nonlinear effects are taken into account. One can also validate the results obtained in the frequency domain with time-domain simulation. This is the topic of the next chapter.

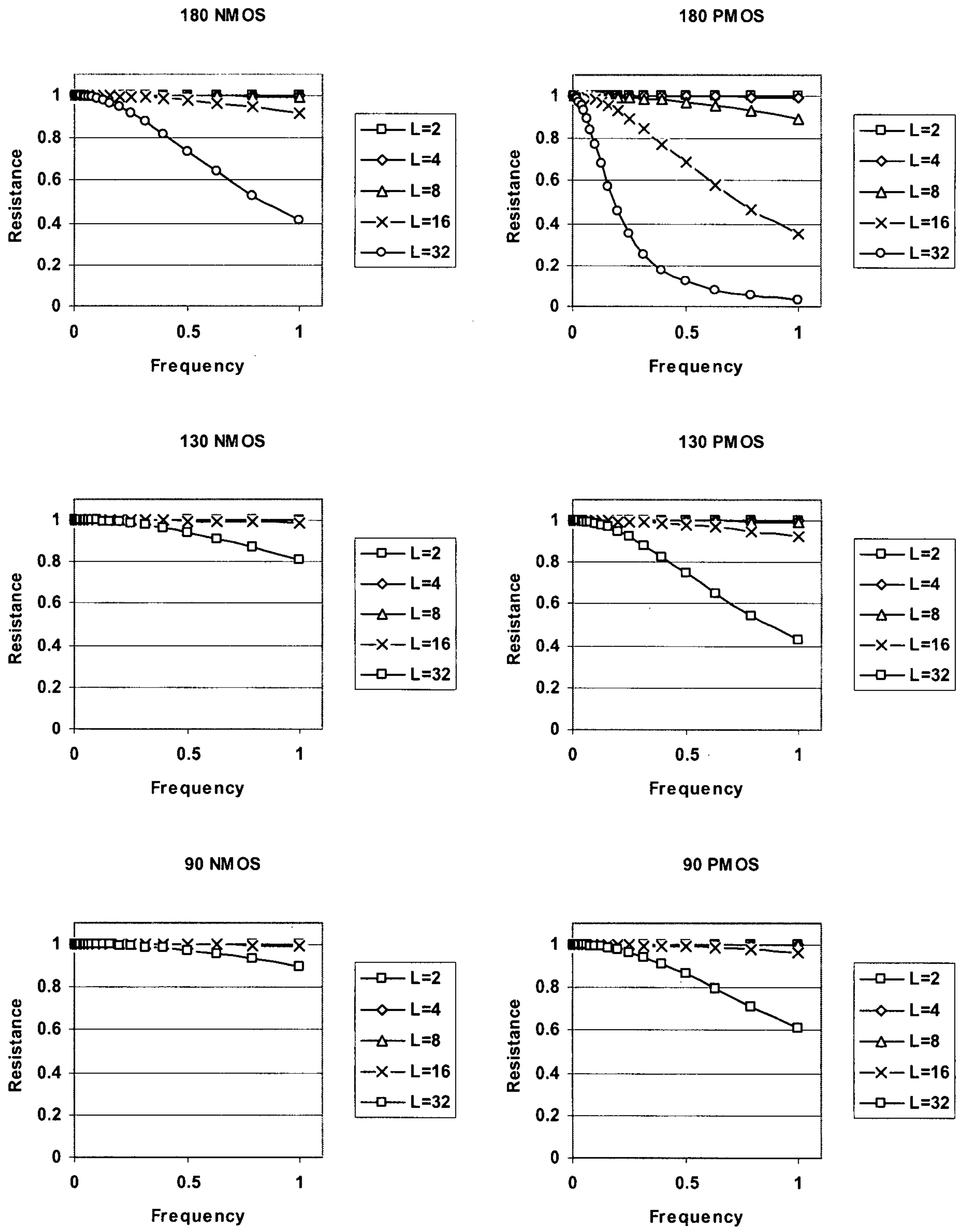


Figure 3.18: Effective Resistance for NMOS and PMOS at 180nm, 130nm, and 90nm.

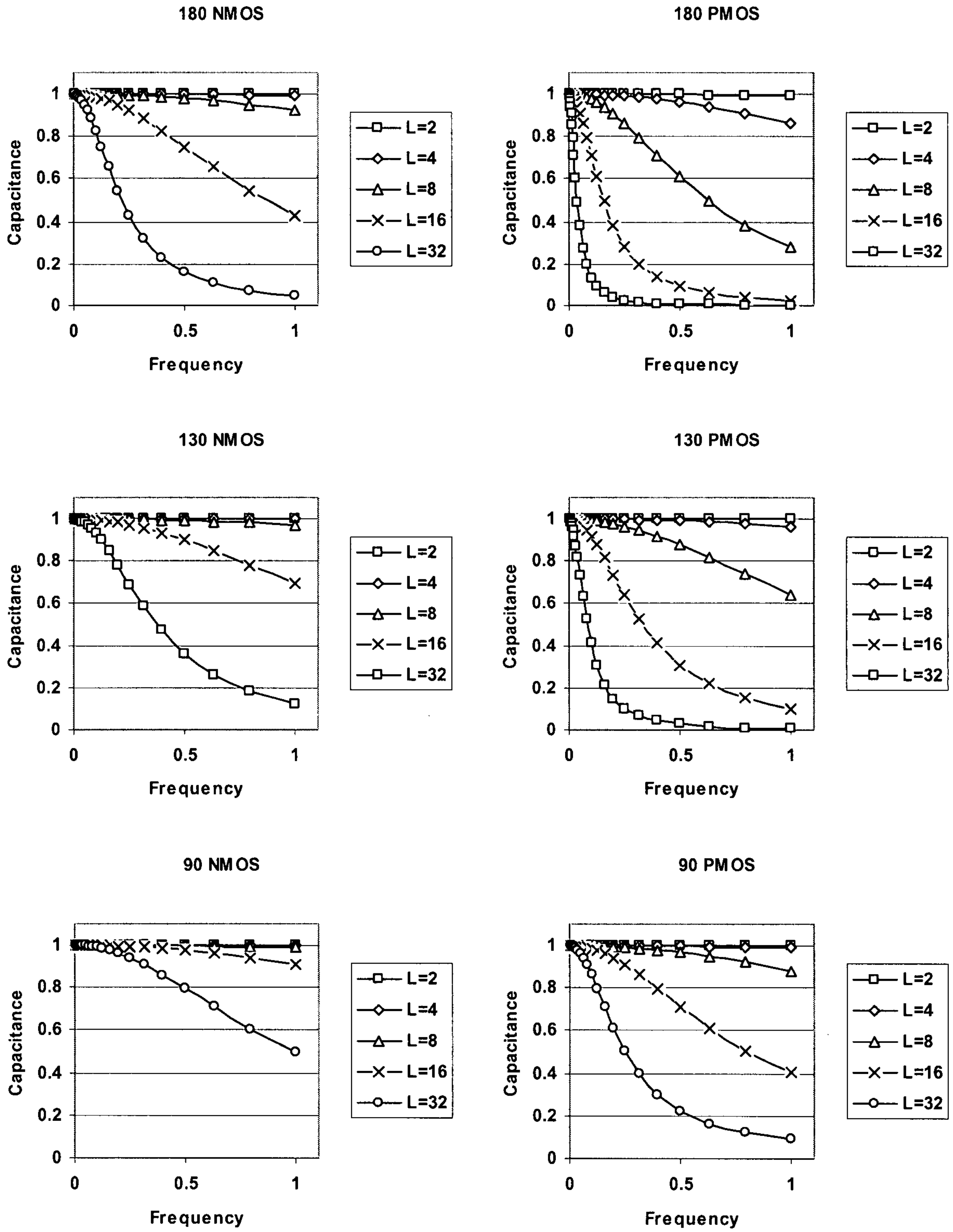


Figure 3.19: Effective Capacitance for NMOS and PMOS at 180nm, 130nm, and 90nm.

Chapter 4

Transient Response and Layout of Decoupling Capacitors

4.1 Introduction

The objective of this chapter is to analyze decoupling capacitors in the time domain and determine the appropriate sizing of decaps for layout purposes. In addition, the validity of the *RC* values extracted from the previous section are demonstrated using transient analysis. There are essentially two issues to address in terms of layout. First, the use of all NMOS, or all PMOS or a combination of NMOS and PMOS is explored. Second, the effect of area constraints or decap value constraints on the standard cell layouts is investigated. A methodology is developed for determining the optimized layout of decaps for noise reduction over a certain frequency range.

4.2 Estimating Decoupling Capacitance

In order to carry out some initial transient simulations, a value of decoupling capacitance must be chosen that is suitable for the circuit under analysis. The setup in Figure 4.1 shows a rather pessimistic situation where there is no power supply and only a decap provides the needed current that is used by the inverter during the switching process. The value of the required decoupling capacitance can be derived for an inverter with an output capacitance of C_{out}. There is an initial voltage on node V^+ on V_{DD}, and the inverter with an initial logic value of '1' is switched to '0'.

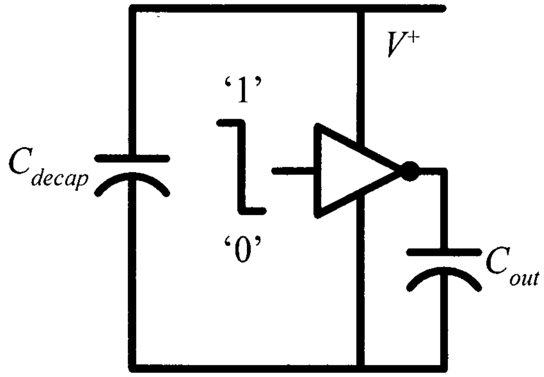


Figure 4.1: Experimental setup to calculate required decap.

If the noise margin was a factor *NM* of V_{DD}, where $NM < 1$, then using charge-sharing equations on this closed system, the required C_{decap} can now be derived as follows:

$$\begin{aligned}
Q_{before} &= Q_{after} \\
C_{decap}V_{DD} &= C_{decap}V^* + C_{out}V^* \\
C_{decap}\left(V_{DD} - V^*\right) &= C_{out}V^* \\
C_{decap} &= \frac{C_{out}V^*}{\left(V_{DD} - V^*\right)} \\
C_{decap} &= \frac{\left(1 - NM\right)C_{out}V_{DD}}{NM \times V_{DD}} \\
C_{decap} &= \frac{\left(1 - NM\right)}{NM}C_{out} \qquad (4.1)
\end{aligned}$$

By selecting different values of *NM*, one can determine the starting value of C_{decap} for a given switching capacitance, C_{out}. In fact, Figure 4.2 plots the ratio between decap capacitance and switching capacitance as a function of the noise margin using Eqn (3.1). As can be seen, the lower the noise margin, the higher the required decap, as expected. For example, if the noise margin is 10% of V_{DD}, then $NM = 0.1$ and $C_{decap} = 9C_{out}$. To add an extra margin of safety, it would be prudent to use a slightly higher value, e.g. $C_{decap} = 10C_{out}$.

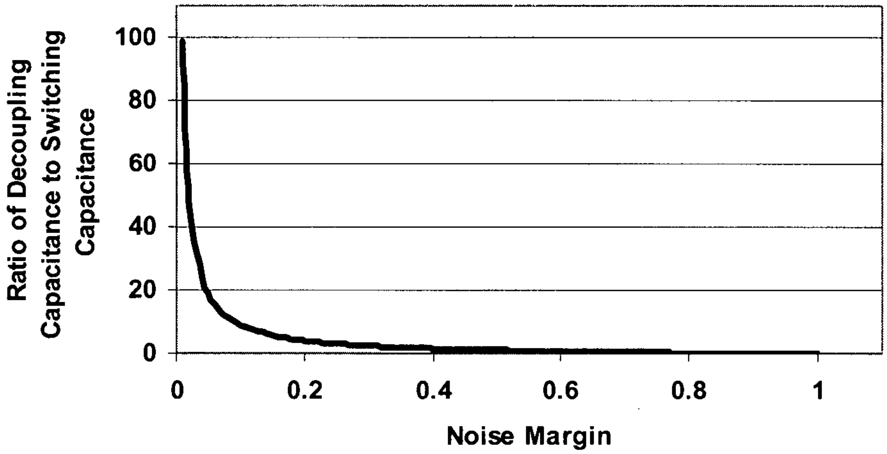


Figure 4.2: Plots of capacitance ratio vs noise margin (NM).

The above result is consistent with the *rule-of-thumb* used in industry for the total amount of decap used in a given design. To illustrate this point, consider the processors displayed in Table 4.1 from [13]. Each of the values of total on-chip decoupling capacitance is about 10 times larger than the total switching capacitance. This is true even though the three processors listed are implemented using three different CMOS technologies and run at three different clock rates. Therefore, this is a good rule to remember, and is validated by the first-order derivation of Eqn. (4.1).

Table 4.1: On-chip Decoupling Capacitance for Alpha processor family [13].

Processor Name	Process Technology	Clock Frequency	Total Switching Capacitance	On-chip Decoupling Capacitance
EV4	0.75um CMOS	200MHz	12.5nF	128nF
EV5	0.5um CMOS	350MHz	13.9nF	160nF
EV6	0.35um CMOS	575MHz	34nF	320nF

4.3 Transient Response

The experimental setup in Figure 4.3 will be used to evaluate the accuracy of the *RC* values that were extracted in the previous section using AC analysis. A 10x inverter[1] is stimulated with a negative step-function as before so that it will draw current from the decoupling capacitor which has been sized to be roughly 10 times larger than C_{out}. On the left is the MOS decap while on the right is the equivalent series *RC* circuit.

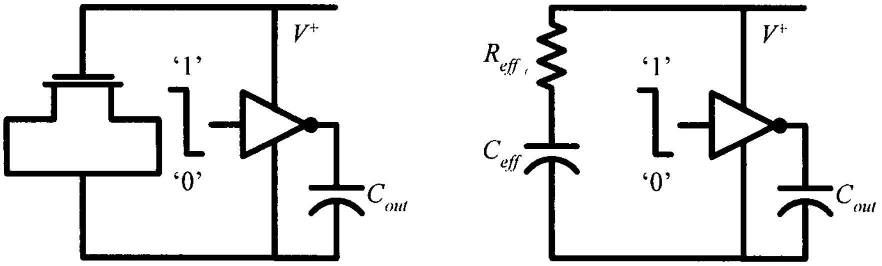


Figure 4.3: Experimental setup.

First, the accuracy of the *RC* values extracted for the devices in the previous chapter with lengths of 2λ, 8λ, and 32λ are tested. The MOSFET and series *RC* response are shown in Figure 4.4. Note that for the lengths of 2λ and 8λ, the transient response for the MOS decap is very similar to the response using the corresponding R_{eff} and C_{eff} for each size but for the length of 32λ, the response of the series *RC* circuit is slightly different from the 32λ device. This provides some validation for the Chapter 3 results that show that a 32λ device is unsuitable as a decoupling capacitor. Although not shown, results for 4λ and 16λ are also close to the MOS decap results. This provides validation for the equations derived in Chapter 3 and the analysis using the NQS model in SPICE.

[1] A 10X inverter is shorthand for an inverter that is 10 times larger than the minimum size inverter, which has a 2λ NMOS device and 4λ PMOS device.

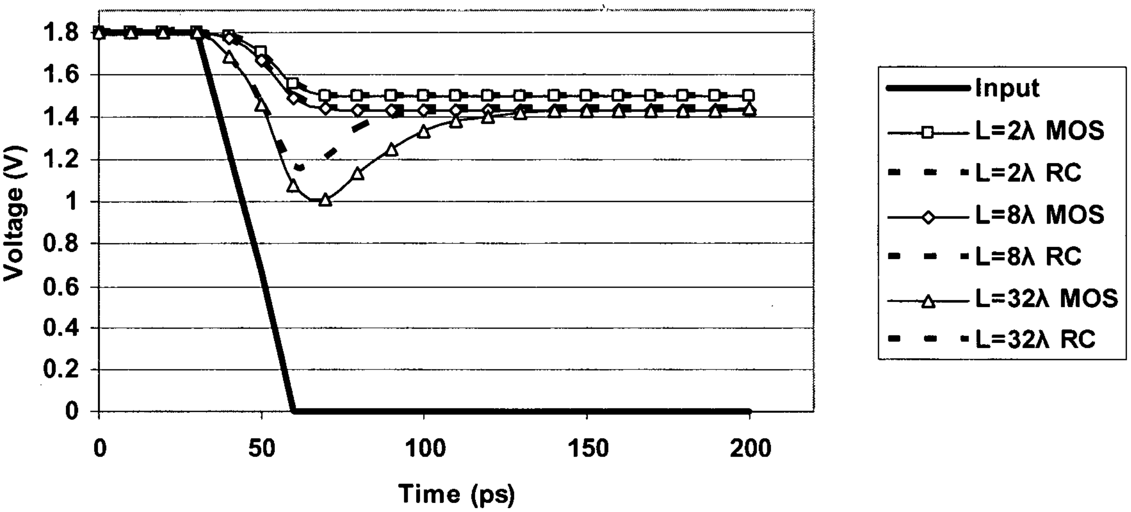


Figure 4.4: 2λ, 8λ and 32λ responses.

Now the selection of a suitable channel-length can be carried out. For this purpose, the transient results for five device sizes are normalized and shown in Figure 4.5. One can see from this plot that the curve with the longest channel but minimal droop lies in the range of 8λ to 16λ, which is also consistent with the results obtained in Chapter 3. The channel-length will be optimized even further in the next few sections.

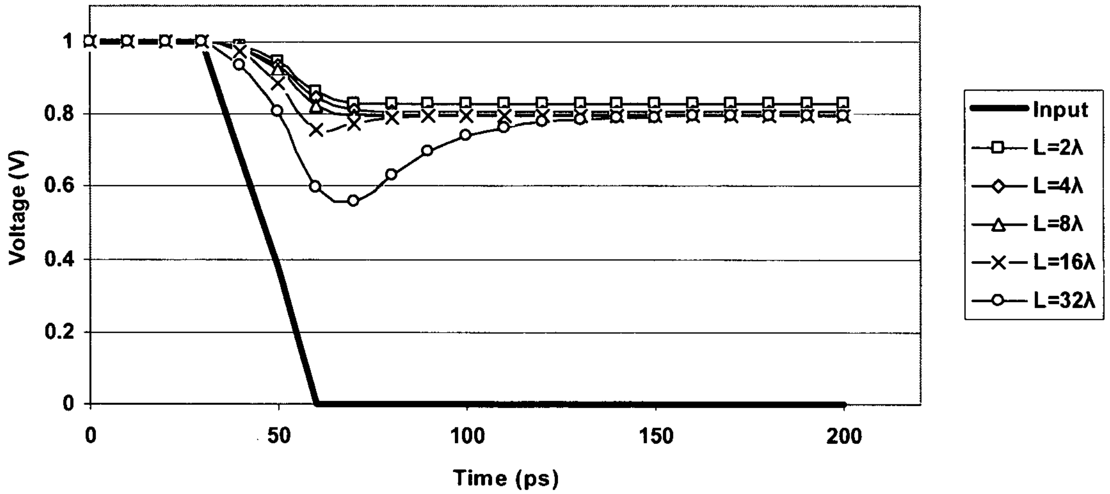


Figure 4.5: Normalized transient responses.

4.4 Standard Cell Decap Layout Schemes

Figure 4.6 displays three basic types of decap layouts using single devices. From left to right, they are: NMOS and PMOS decap (N+P), NMOS-only (N-only) and PMOS-only (P-only). The N+P decap is separated into a p- and n-complex similar to that of a CMOS logic gate. This arrangement is quite common in standard cell layouts since the PMOS devices are situated above NMOS devices. The NMOS-only and PMOS-only decap cells extend from V_{DD} to V_{SS}. These cells require single well regions in the cell but must still satisfy design rules at the interfaces of adjoining cells.

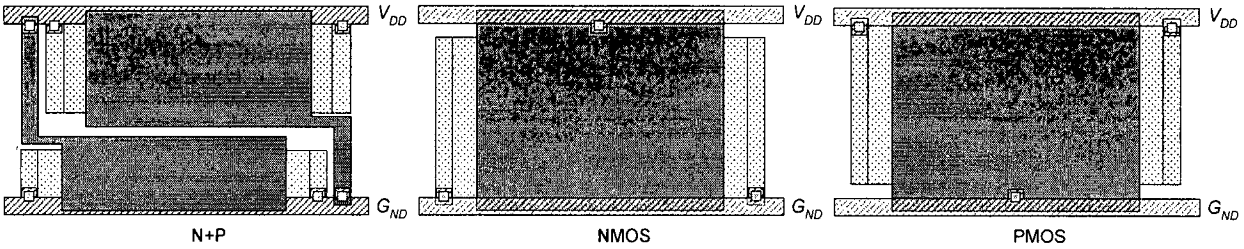


Figure 4.6: Three basic types of decap layouts.

Recall that, in the case of standard cells, the height of the cell is fixed[1]. The only freedom granted to the designer for each cell is the number of divisions or *fingers* of the device. For example, the decap in Figure 4.7 has three NMOS and five PMOS fingers. For this particular cell, the designer may add or remove fingers, but may not change the width or height. The advantage of increasing the number of fingers is the reduction in channel length which improves the high-frequency performance. The disadvantage is the decrease in total capacitance (for a fixed area) or an increase in area (for a fixed capacitance). Each additional finger introduces contacts and source/drain connections that must meet spacing rules.

[1] The total width of the cell is often constrained to be a fixed multiple of a base width as required by placement tools.

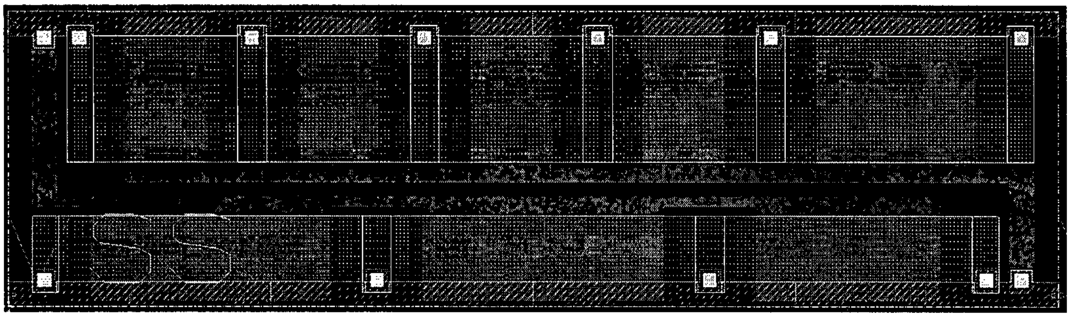


Figure 4.7: Example of N+P decap.

The first question to address in the layout is the selection of the best of the three options from those shown in Figure 4.6. Assume that there is a requirement to deliver a target decap value with a given frequency response. The N-only option would produce the smallest area because it has the best frequency response and would allow the use of the largest channel length. The P-only option would generate the largest area. The N+P option would lie somewhere in between. Therefore, it is most appropriate to select the N-only option for decap layout for standard cells, assuming area is equal.

The detailed analysis required to select the channel length for an N-only decap layout option is now considered. A metric is required to compare the effectiveness of NMOS decaps with the same cell width and different number of fingers. The metric should stress the importance of the charging current while minimizing R_{eff} and maximizing C_{eff}. This objective can be met by maximizing $\mathrm{Mag}(I)$ with a $V = 1\angle 0°\mathrm{V}$ as follows:

$$\begin{aligned} M &= \mathrm{Mag}(I) = \mathrm{Mag}\left(\frac{V}{Z}\right) \\ &= \mathrm{Mag}\left(\frac{1}{Z}\right) = \frac{1}{\sqrt{R_{eff}{}^{2} + \left(\frac{1}{\omega C_{eff}}\right)^{2}}} \end{aligned} \qquad (4.2)$$

Now that a metric defined, its use in designing the decap cells is illustrated. First, a library of cells is obtained, and for each cell width is held constant while different combinations of fingers are simulated. The library of cells listed in Table 4.2 is used and the widths are listed as a multiple[1] of the first cell, Cell 1. These cells will be used to layout the decaps with different styles and number of fingers to determine the optimum layout.

Table 4.2: Results for N-only decap cells optimized for 2GHz noise.

Cell Name	Width (Multiple of Cell 1)
Cell 1	1x
Cell 2	2x
Cell 3	4x
Cell 4	8x
Cell 5	16x
Cell 6	32x

For an N-only decap with the same area as the decap in Figure 4.7 (which happens to be Cell 2) was calculated to be able to support up to nine fingers. Note that with each division, capacitance is reduced due to the added contact but so is resistance as well. Next, for each combination of fingers, the current values are plotted as a function of frequency over a 20GHz range as in Figure 4.8. Notice the following characteristics:

- The current vs. frequency plots for each finger combination generally begins by increasing linearly then flattening at a certain frequency.
- Decaps with fewer fingers have a higher initial current and current slope.
- Currents of decaps with more fingers increase linearly over a larger range of frequencies before flattening.
- The number of fingers to use depends on the noise frequency.

[1] Detailed cell dimensions are proprietary and therefore not shown.

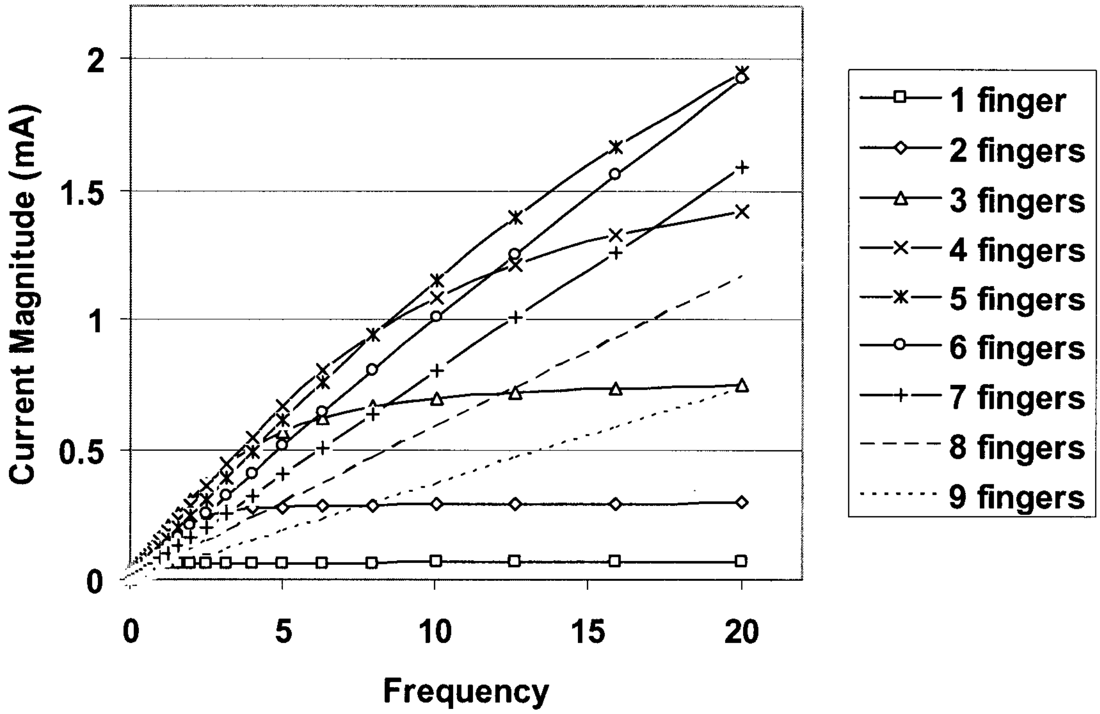


Figure 4.8 Current Profile for Different Number of Fingers

Starting with NMOS-only decaps, measurements from SPICE indicate that for Cell 2, at 2GHz, two fingers would be optimal and at 10GHz, it would be five. The layout for this 10GHz case is shown in Figure 4.9.

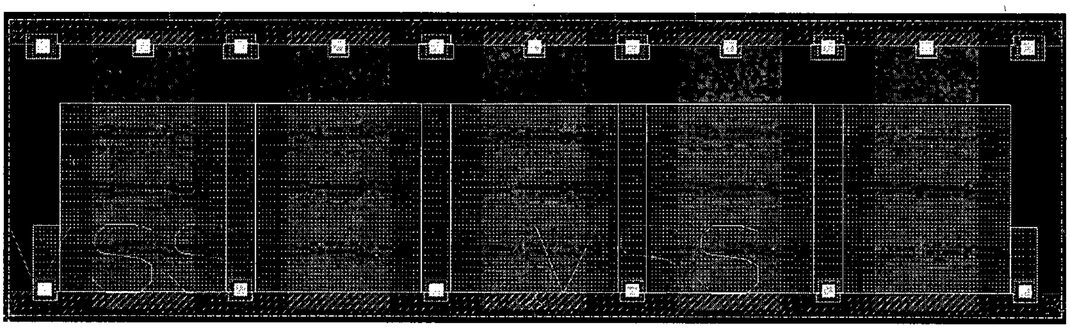


Figure 4.9: Example of NMOS-only decap with (a) two fingers (b) three fingers

Using this method, a complete set of NMOS-only decaps can be laid out for a standard cell library by plotting each decap's current vs. frequency graph for every possible number of fingers as in Figure 4.8. Table 4.3 lists the current values for NMOS-only cells optimized for 2GHz noise and

Table 4.4 lists the 10GHz NMOS-only cells. Each cell was laid-out, extracted with their full parasitics, and simulated to determine their effective resistance and capacitance. Note that, as expected, the NMOS decaps have higher current at their target frequency.

Table 4.3: Results for N-only decap cells optimized for 2GHz noise.

Cell Name	Width (Multiple of Cell 1)	# Fingers (NMOS)	*I* at 2GHz (μA)	*I* at 10GHz (μA)
Cell 1	1x	2	150	154
Cell 2	2x	5	316	660
Cell 3	4x	10	631	1321
Cell 4	8x	20	1263	2641
Cell 5	16x	40	2526	5282
Cell 6	32x	80	5051	10,564

Table 4.4: Results for N-only decap cells optimized for 10GHz noise.

Cell Name	Width (Multiple of Cell 1)	# Fingers (NMOS)	*I* at 2GHz (μA)	*I* at 10GHz (μA)
Cell 1	1x	2	137	523
Cell 2	2x	5	268	1189
Cell 3	4x	10	536	2377
Cell 4	8x	20	1072	4755
Cell 5	16x	40	2143	9509
Cell 6	32x	80	4288	19,018

Some designs use decaps that contain both NMOS and PMOS in keeping with the traditional standard cell layout style. Since there are both n- and p-type transistors, they can be optimized separately using the methodology presented above and then combined. Layouts were created and optimized for 10GHz. The layout is shown in Figure 4.10 and the results are shown in Table 4.5.

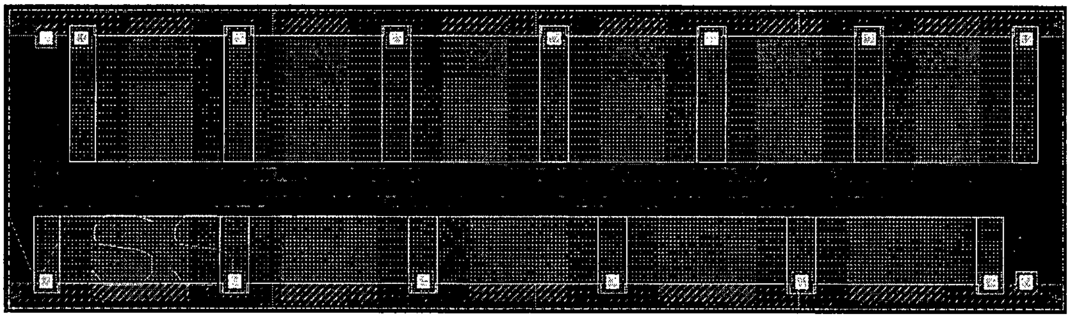


Figure 4.10: A N+P cell optimized for 10GHz frequency range.

Table 4.5: Results for N+P decap cells optimized for 10GHz noise.

Cell Name	Width (Multiple of Cell 1)	# Fingers (N/P)	*I* (μA)
Cell 1	1x	2/3	382
Cell 2	2x	5/6	621
Cell 3	4x	10/12	1243
Cell 4	8x	20/24	2486
Cell 5	16x	40/48	4971
Cell 6	32x	80/96	9942

However, the layout in Figure 4.10 has only one contact feeding an array of parallel devices which is not desirable. One further optimization concerns the number of connections from the polysilicon gate to the power line. For long decaps there will be considerable parasitics between the farthest MOSFET and the power line. This could be reduced by increasing the number of connections or using one of the smaller cells as a base cell that is replicated. The benefits of this are improved frequency response but reduced capacitance due to the space required for the extra connections as shown in Figure 4.11 which uses Cell 1 as a base cell. The currents are displayed in Table 4.6.

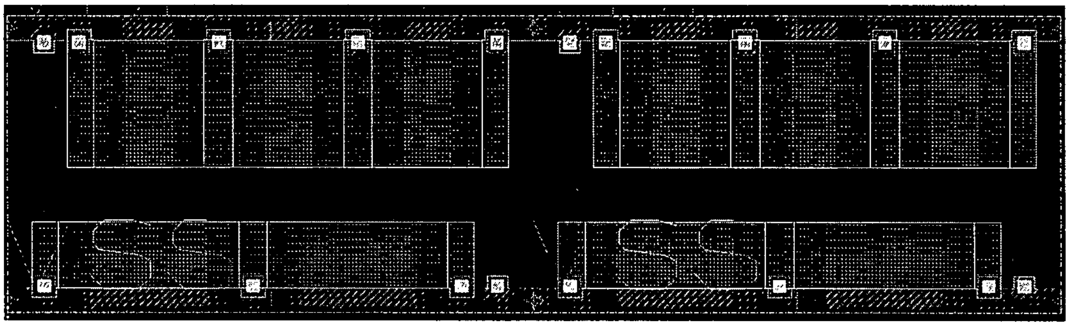


Figure 4.11: A N+P cell optimized for 10GHz range using multiple instances of smaller cells.

Table 4.6: Results for N+P decap cells optimized for 10GHz noise using Cell 1 as base cell.

Cell Name	Width (Multiple of Cell 1)	# Fingers (N/P)	*I* (μA)
Cell 1	1x	2/3	382
Cell 2	2x	4/6	764
Cell 3	4x	8/12	1566
Cell 4	8x	16/24	3053
Cell 5	16x	32/48	6106
Cell 6	32x	64/96	12,212

Comparing the N+P decaps to their counterpart NMOS-only decaps, we see not surprisingly, that the NMOS decaps have better performance because they have greater oxide-area and better performance than PMOS and N+P cells that use Cell 1 as a repeater cell have better performance than N+P decaps with only one connection from gate to power.

4.5 Conclusions

To summarize this chapter, the layout of standard cell decaps should be carried out with N-only devices for maximum frequency response for a given area. The devices should be designed with the channel length determined by either AC analysis or transient simulations, as described in this chapter.

Chapter 5

Standard-Cell Placement of Decoupling Capacitors

5.1 Introduction

This chapter is concerned with decoupling capacitor placement strategies to eliminate *IR*-drop. It first reviews standard-cell placement and its relationship to decap placement. Next it provides background on decoupling capacitance estimation based on power dissipation. Finally, it presents a short example that explores various decap configurations and obtains the minimum decoupling capacitance required to eliminate *IR*-drop.

5.2 Standard-Cell Design Constraints

Before decap cell configurations are attempted, it is useful to review the background of IP block layout. IP blocks are usually laid-out using the standard cell layout approach as described in Section 2.4. When a block is synthesized, placed and routed, there will naturally be empty spaces between cells as shown by the black dots in Figure 2.7 (repeated in Figure 5.1). These spaces are the best places to fill with decaps. Normally, these spaces are filled using a library of decap cells, (such as those introduced in Chapter 4) without affecting the placement of the logic cells. This is a typical solution to the decap placement problem due to its convenience.

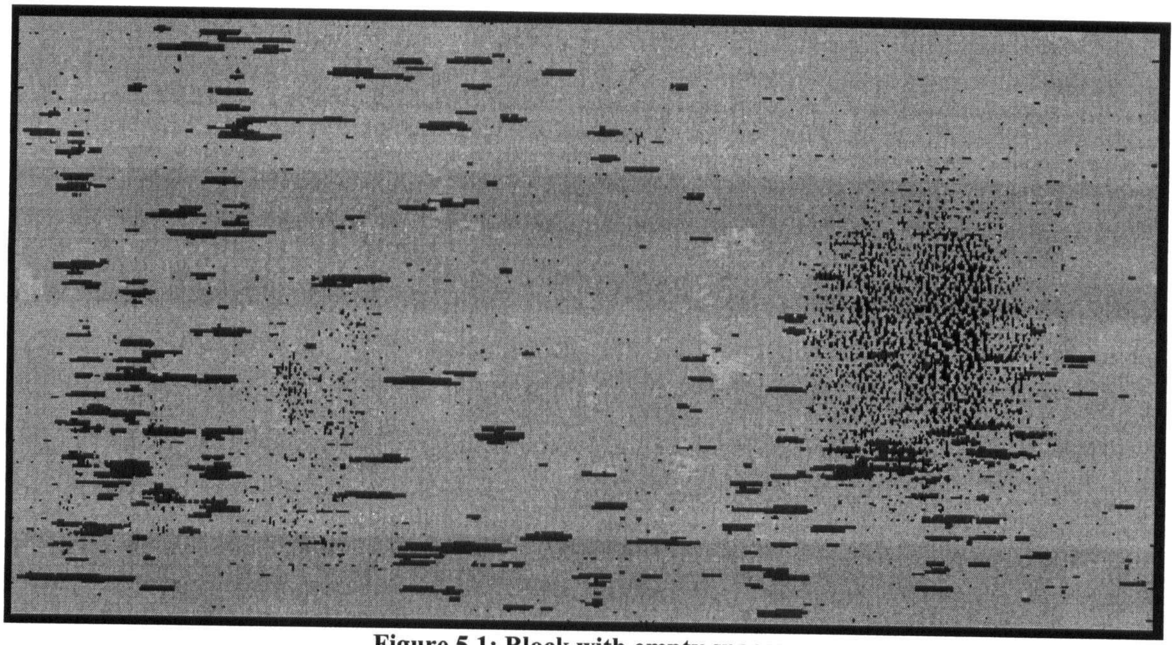


Figure 5.1: Block with empty spaces.

However, this method is not optimal for decap placement because the empty cells may not be located near the areas of high *IR*-drop. If a large amount of decoupling capacitance is needed, some sections of IP blocks may not have enough empty space to provide sufficient capacitance. Moving cells to different rows would be difficult since it would affect routability and timing, but shifting cells along the same row and maintaining relative placement is feasible [19]. A more intuitive approach to decap placement is proposed that takes into account the nature of the IP block and where the high noise levels are expected to occur. Previous work on decap placement include placement using activity [16], global decap placement *between* IP blocks,[17], placement to reduce leakage [18] and standard placement that does *not* affect the relative placement of cells [19].

5.3 Relationship between switched capacitance and total decap

As mentioned in Section 4.2, a 10:1 ratio between decoupling capacitance and switching capacitance if often used as a *rule-of-thumb*. This is the value of the total-chip decoupling capacitance, C_{decap}, to the total-switching capacitance, $C_{switched}$. Here, the interest is in finding out the same ratio for the IP block. Of course, this requires a reasonable estimate of switching capacitance for the block. One method to determine the switching capacitance is through the average power dissipated by the block. Typically, an IP block is simulated with many vectors to produce activity information that is used to estimate power. For example, the dynamic power dissipated can be found using the equation:

$$P_{avg} = \sum_{i=1}^{n} \alpha_i C_i V_{DD}^2 f$$

where i represents a circuit node in the IP block, n is the number of nodes, α_i the switching activity of the node i, C_i the capacitance of the node and f the clock frequency of the design. If V_{DD} and f are the same for all nodes, the equation can be rewritten this as follows:

$$P_{avg} = \sum_{i=1}^{n} \alpha_i C_i V_{DD}^2 f = \left(\sum_{i=1}^{n} \alpha_i C_i \right) V_{DD}^2 f = \alpha C_{total} V_{DD}^2 f$$

Here, α is the overall activity factor for the IP block and C_{total} is the total capacitance of all the nodes in the block. Clearly, the quantity $\alpha \times C_{total} = C_{switched}$; that is, the average switched capacitance is just the total capacitance times the overall activity factor. Then, it follows that:

$$C_{switched} = \frac{P_{avg}}{V_{DD}^2 f}$$

Now one can relate this back to the needed decap amount. As derived in Section 4.2, when targeting a 10% noise threshold, the decoupling capacitance would be set to $C_{decap} = 9C_{switched}$, but because peak power is higher than average power, a more conservative value of $10C_{switched}$ is generally used. Therefore:

$$C_{decap} \approx 10C_{switched} = 10 \times \frac{P_{avg}}{V_{DD}^2 f}$$

This is a very pessimistic result but its usefulness will be evaluated in the next section. The main point here is that, assuming that it is possible to obtain the average power, a first-order estimate of the needed decap amount can be quickly computed. Since IP blocks are reused in many designs, this type of information is readily available. Today, commercial power estimation tools also account power due to leakage and crowbar current. Therefore, a reasonably accurate value for this quantity can be used in the above equation to calculate the total decap value needed in the IP block. Of course, the pre-multiplier may change depending on the findings in the next section.

5.4 Placement Configurations

The next issue is to determine decap locations. For block-level simulation, since solder-bump information is not yet available, average resistance values are applied between Metal 8 and the solder bumps. Different configurations of decoupling capacitor placements are constructed and simulated to determine the optimum decoupling capacitance value. In this section, four configurations are investigated:

1.) All decaps in the center (the power grid typically sags near the center)

2.) All decaps in the corners

3.) Decaps near the hot spots (hot spots refers to *IR*-drop violations, this can be done post-simulation by placing them at *IR*-drop violations but requires an additional simulation to determine *IR*-drop locations)

4.) Decaps evenly distributed throughout IP block (this approach assumes some uniformity in current in the IP block)

An industrial example block (with all pre-placed decap cells removed) is shown in Figure 5.2. Cells are represented as yellow squares. High cell density areas are represented with more solid yellow (such as the large circle in the right half of the block) The block contains approximately 100,000 cell instances each with 1-3 gates per instance and a total switching capacitance of 325pF was reported by Redhawk[1], a power and ground network *IR*-drop simulator that considers both static *and* dynamic noise.

Figure 5.2: Example IP Block

Decap cells were inserted into the standard cell layout according to the strategies listed above (represented by the white areas in Figure 5.3). For fairness, each block was placed with an equal number of cells (all of one type) and thus each have the same area penalty. They were then simulated using Redhawk. The placement was completed as follows:

1.) First, all the existing decap is removed in Synopsys Apollo.

2.) The decap cells were placed in their 'ideal' locations. For example, in the center configuration, all the decaps were placed in one spot, overlapping each other.

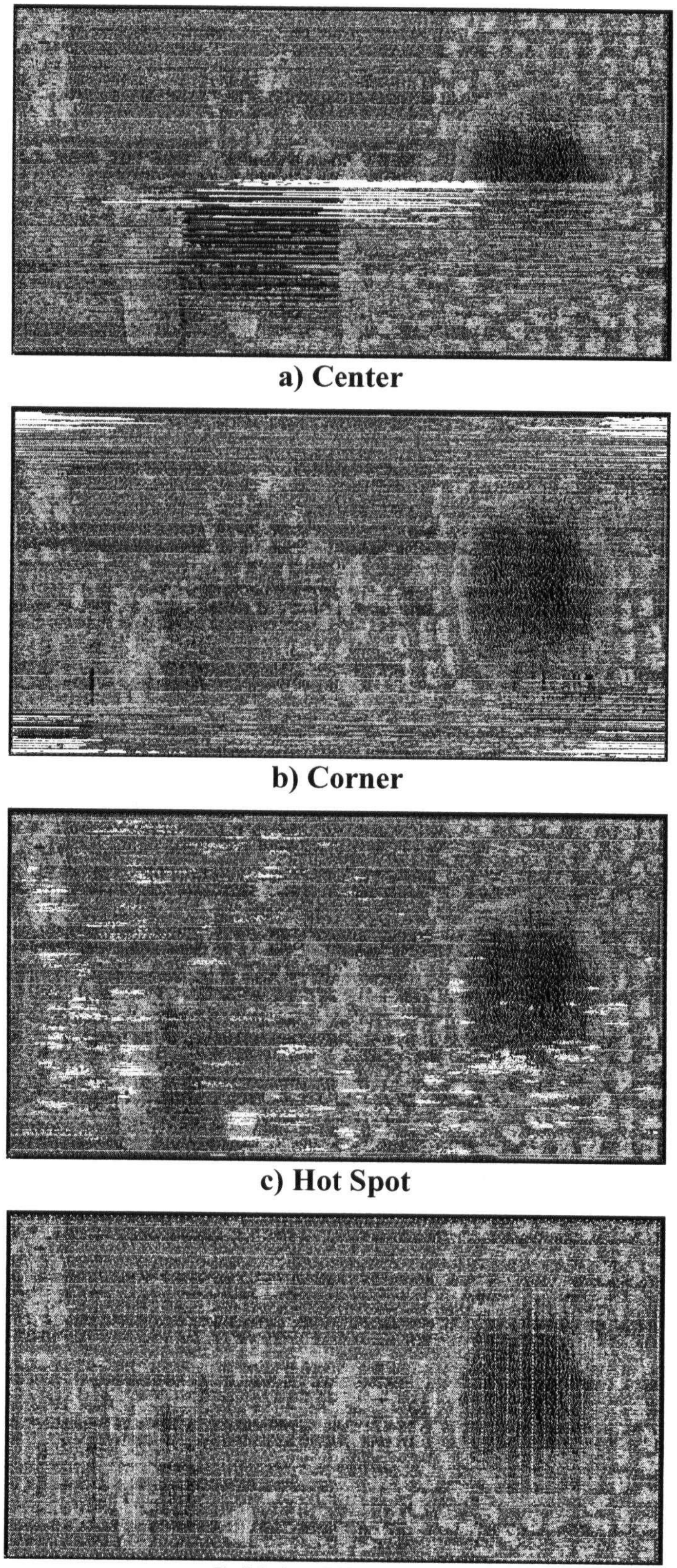
[1] Redhawk is a trademark of Apache, Inc.

3.) ECO placement is then executed where the tool will try to obtain the best placement.

4.) The capacitance per decap is then changed manually (using the binary search method) and the block is simulated in Redhawk until all noise is removed.

5.) In each case, the total area of the block remains the same and some of the original non-decap cells may have been shifted.

Simulation is conducted by first calculating the average power dissipation of the block. Then, the instances with the highest activities are switched for a single-cycle such that the power generated by the switched cells equals that of the average power dissipation. In this manner, one can obtain a reasonable approximation of where the worst *IR*-drop locations exist without having to do full-dynamic vector-based simulations.



a) Center

b) Corner

c) Hot Spot

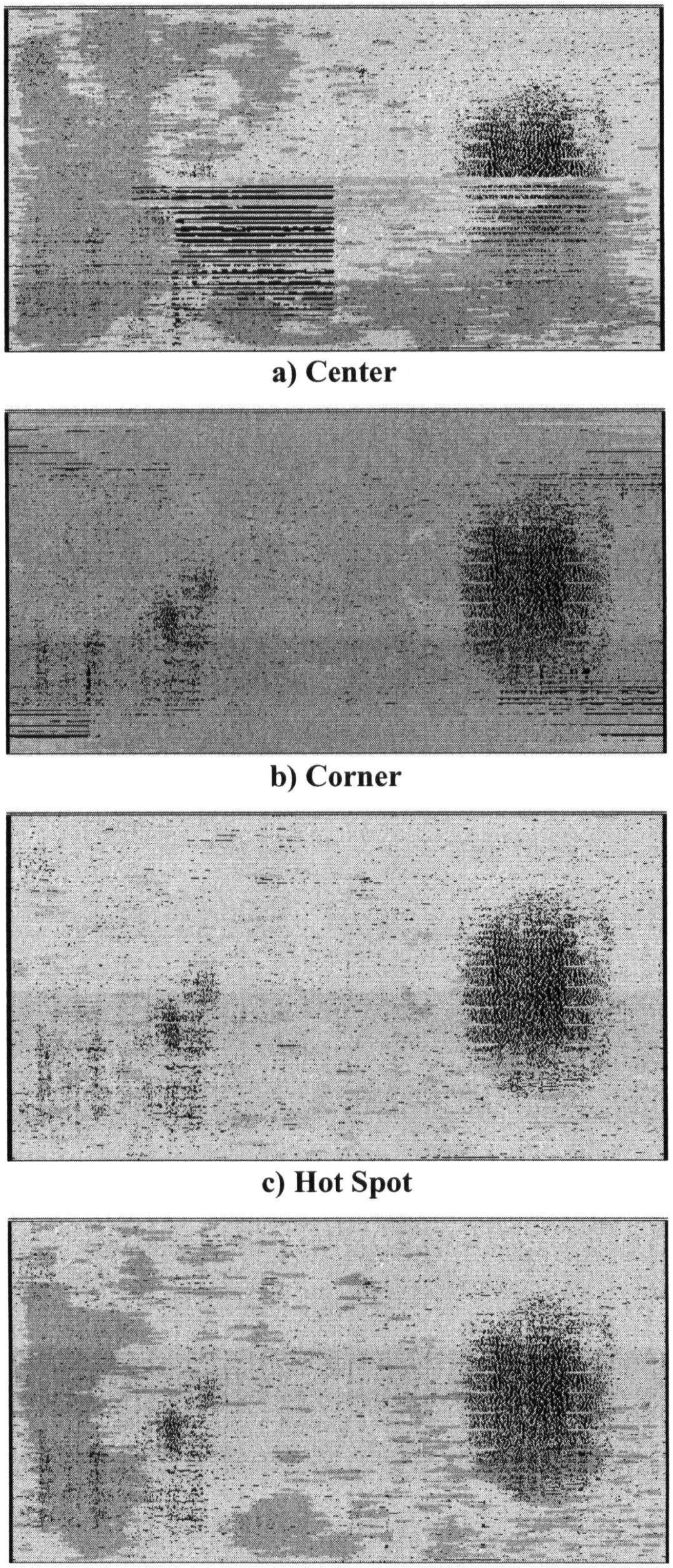
d) Noise Violations

Figure 5.3: Decap Configurations

The minimum decoupling capacitance to reduce *IR*-drop was determined by iteratively changing the decoupling capacitance per cell until the *IR*-drop noise for the design dropped below the noise margin of 10% of V_{DD}. The noise-contour plot for each configuration is shown in Figure 5.4. In these plots, orange represents noise of 6.7%-8.3% of V_{DD}, yellow represents 5%-6.7% of V_{DD} and purple represents the decoupling capacitor cells.

The results of the simulations are shown in Table 5.1 which lists the total decap amount needed to reduce the noise for each case. It is difficult to determine which configuration is the best. The corner configuration may appear to be the best since it requires the least amount of decap, but the hot-spot configuration with only 25% more decap, reduces the noise in all cells to an even lower bracket as evidenced by the 'yellow'. One point that should be mentioned concerning this particular block is that the hot-spot configuration eliminated 99% of the noise violations with much less decoupling capacitance than all of the other configurations. The elimination of the remaining 1% required much more capacitance. Notice that there are no orange cells for the hot-spot configuration in Figure 5.4. Future noise-analysis with more blocks may prove that the hot-spot configuration is indeed the best.

The actual ratio of decoupling capacitance to switched capacitance appears to be in the range of 1.8x–2.25x for this block. This is much smaller than the factor of 10 described earlier. Although some decoupling capacitance exists in the gate and source/drain capacitance of logic cells and the wiring capacitance this alone cannot account for the large difference. Also, in the full-chip designs shown earlier in Table 4.1, all the inter-block regions were filled with decaps, so this is not really a proper validation of the ratio of 10.



a) Center

b) Corner

c) Hot Spot

d) Noise Violations

Figure 5.4: Noise-contour plots

Table 5.1: Optimal value of decoupling capacitance.

Strategy	Total Decap
Center	684pF
Corner	586pF
Evenly Distributed	707pF
Hot Spot	733pF

It appears that a factor of $C_{decap}/C_{switched}$ of less than 2 is sufficient for standard cell designs from this one example block. This is reasonable since it would be difficult to include $10 \times C_{switched}$ inside the given block. The overall area would grow quite large if $10 \times C_{switched}$ were included in the form of decaps (assuming a 10% activity factor). Of course, other blocks should be investigated in order to establish a proper rule of thumb for the $C_{decap}/C_{switched}$ ratio in a standard cell layout.

5.5 Conclusions

The methodology for decap design for standard cells should be carried out as follows. First the frequency range of interest should be determined. Then, based on AC analysis, the suitable value of channel length, *L*, should be determined to deliver the desired decap performance. Next, decap cells should be designed based on a set of target values, with the selected *L* and a corresponding number of fingers, and then stored in a library. Based on this preliminary analysis, during the place-and-route operation, decaps should first be placed in the corners to reduce noise-violations by shifting standard cells to create space in that area and filling it with decap, then they should be near *IR*-drop violations to minimize average noise across all cells, then placed in the center, then uniformly placed and finally, near the *IR*-drop violations. Finally, an *IR*-drop simulation should be carried out and decaps should be added to eliminate any hot-spots in the analysis. In this way, most of the noise problems can be resolved before analysis, and the needed decoupling capacitance may be kept to a minimum. Based on the preliminary analysis given in this chapter, the total decoupling capacitance required should be less than twice the switching capacitance.

Chapter 6

Conclusions and Future Work

6.1 Summary

Power supply noise, due to *IR*-drop and *Ldi/dt* effects, can be reduced by placing decoupling capacitors in the proper locations throughout the power distribution system. In this thesis, we have investigated the use of decoupling capacitors for this purpose in standard cell layouts of IP blocks. This implies that the cells must be of fixed height but can be of variable length. Since circuits are running at higher and higher frequencies, the objective was to understand the response of the decaps when operating a high frequencies and to recommend solutions in terms of layout design for standard cell designs.

This thesis began with an analysis of high-frequency performance associated with decaps. The decap was modeled as a series resistance and capacitance, and new equations were derived to compute these values. It was found that, as the frequency increases, the value of the capacitance drops off and the performance degrades significantly. The controlling factor is the MOS transit time, and therefore a new equation for decaps was derived to include this factor. Empirical equations were also developed to model the behaviour of decaps in the frequency domain. These equations can be used to understand the effect of scaling on high-frequency performance of decaps and provide first-order sizing information.

Next, the time-domain behaviour of decaps was investigated. A simple equation for the required decoupling capacitance for a given buffer size was derived using charge-sharing. Then, various options for standard-cell layout of decaps were explored to optimize overall performance. It was found that all-NMOS decaps in parallel was the best layout approach to obtain the highest performance decaps for standard cell designs. The width of the devices is fixed by the cell height, while the number of fingers and the length of each device can be determined by the desired frequency and transient response.

Finally, the issue of where to place the decaps and how much to place in each location was addressed. Different placements were explored for a number of different IP blocks using a commercial *IR*-drop simulation tool, and a standard cell placement and routing CAD flow. The options were to place decaps near the V_{DD} connections, at the center of the block, uniformly distributed, near the large buffers and at the hot-spot locations. It was found that corner placements provided the best results.

6.2 Contributions in this Thesis

The following list summarizes the contributions:

- Derived decap equations for R_{eff}, C_{eff}, and transit time
- Developed of an empirical model for frequency response of decaps
- Established a relationship between power and decoupling capacitance

- Established a relationship between switched capacitance and decoupling capacitance
- Developed effective standard cell layout for decaps
- Provided preliminary guidelines on placement of decaps in standard cells

6.3 Future Work

There are a number of issues to addresses in the near future regarding decoupling capacitors. First, the topic of global decaps has not been addressed here. While the high-frequency results of Chapter 3 are valid, there may be other issues to address when laying out such decaps. For example, a *waffle*-style layout is often employed and it remains to be seen whether this is optimal. Other options for layout should also be explored. This would require some type of field simulator, such as Synopsys Medici.

Another issue that was not examined here is the problem of thin-oxide gate leakage. This is a known issue in 90nm CMOS technology, but methods of decap design have not be developed to address this issue. For large decaps, one can expect rather large leakage currents. Therefore, a different oxide thickness may be needed for white-space decaps compared to standard cell decaps. It is known that PMOS devices leak less than NMOS devices so perhaps it is better to use PMOS for the 90nm technology node. Another option is use voltage regulators or perhaps capacitance multipliers to address this issue. Certainly, this problem will have to be resolved in the near future.

REFERENCES

[1] S. Thompson, M. Alavi, M. Hussein, P. Jacob, C. Kenyon, P. Moon, M. Prince, S. Sivakumar, S. Tyagi and M. Bohr, "130nm Logic Technology Featuring 60nm Transistors, Low-K Dielectrics, and Cu Interconnects," *Intel Technology Journal*. Vol. 6, Issue. 2, May 2002.

[2] Synopsys, *HDL Compiler for VHDL Reference Manual*, v2002.05

[3] R. Saleh, D. Overhauser, S. Taylor, "Full-Chip Verification of UDSM Designs", *International Conference On Computer-Aided Design*, pp. 453-460, San Jose, CA. Nov. 1998.

[4] ITRS www.itrs.org

[5] Yong-Ju Kim, Jong-Ho Kang, KunWoo-park, Jae-Kyung Wee, Han-Sub Yoon, Dong-Ju Lee, Yong-Tak Kim, Jung-Sik Kee, "A New Circuit Model for Power Plane System Considering Decoupling Capacitances in Multi-Layer Digital Applications," *Electrical Performance of Electronic Packaging*, pp. 187-190, 21-23 Oct. 2002

[6] K. Y. Chen, William D. Brown, Leonard W. Schaper, Simon S. Ang and Hameed A. Naseem, "A Study of the High Frequency Performance of Thin Film Capacitors for Electronic Packaging," *IEEE Transactions on Advanced Packaging, Vol. 23, Issue 2,* pp. 293-302, May 2000

[7] Jonghoon Kim, Baekkyu Choi, Hyungsoo Kim, Woonghwan Ryu, Young-hwan Yun, Seog-heon Hamm, Soo-Hyung Kim and Yong-hee Lee, "Separated Role of On-chip and On-PCB Decoupling Capacitors for Reduction of Radiated Emission on Printed Circuit Board," *IEEE International Symposium Electromagnetic Compatibility, Volume 1*, pp. 531-536, 13-17 Aug. 2001

[8] Todd Takken and David Tuckerman,, "Integral Decoupling Capacitance Reduces Multichip Module Ground Bounce," *Proceedings of the Multi-Chip Module Conference*, pp. 79-84, 15-18 March 1993

[9] Premjeet Chahal, Rao R. Tummala, Mark G. Allen, Member and Madhavan Swaminathan, "Integrated Decoupling Capacitor for MCM-L Technology," *IEEE Transactions on Components, Packaging and Manufacturing Technology Part B: Advanced Packaging, Vol. 21, Issue 2*, pp. 184-193, May 1998

[10] Mansun Chan, Kelvin Y. Hui, Chenming Hu, and Ping K. Ko, Fellow, "A Robust and Physical BSIM3 Non-Quasi-Static Transient and AC Small-Signal Model for Circuit Simulation," *IEEE Transactions on Electron Devices, Vol. 45, Issue 4*, pp. 834-841, April 1998

[11] John R. Hauser, "Bias Sweep Rate Effects on Quasi-Static Capacitance of MOS Capacitors," *IEEE Transactions on Electron Devices, Vol. 44, Issue 6*, pp. 1009-1012, June 1997

[12] Patrik Larsson, "Parasitic Resistance in an MOS Transistor Used as On-Chip Decoupling Capacitance," *IEEE Journal of Solid-State Circuits, Volume 32, Issue 4*, pp. 574-576, April 1997

[13] Jan Rabaey, "Digital Integrated Circuits: A Design Perspective," *Prentice Hall*, 1996

[14] R. Saleh, Z. Hussain, S. Rochel, D. Overhauser, "Clock Verification in the Presence of IR-drop in the Power Distribution Network", *IEEE Transactions on CAD*, Vol. 19, No. 6, June 2000, pp. 635-644.

[15] David A. Hodges, Horace G. Jackson and Resve A. Saleh, "Analysis and Design of Digital Integrated Circuits in Deep Submicron Technology," *McGraw-Hill*, 2004

[16] Howard H. Chen and Stanley E. Schuster, "On-Chip Decoupling Capacitor Optimization for High-Performance VLSI Design," *International Symposium on VLSI technology, Systems, and Applications*, pp. 99-103, 31 May – 2 June 1995

[17] Shiyou Zhao, Kaushik Roy and Cheng-Kok Koh, "Decoupling Capacitance Allocation and Its Application to Power-Supply Noise-Aware Floorplanning," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Volume 21, Issue 1*, pp 81-92, Jan 2002

[18] Howard H. Chen, J Scott Neely, Michael F. Wang, and Gricell Co, "On-Chip Decoupling Capacitor Optimization for Noise and Leakage Reduction," *16th Symposium on Integrated Circuits and Systems Design, 2003. SBCCI 2003. Proceedings*, pp. 8-11, Sept. 2003

[19] Haihua Su, Sachin S. Sapatnekar and Sani R. Nassif, "Optimal Decoupling Capacitor Sizing and Placement for Standard-Cell Layout Designs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, pp. 428-436, 4 April 2003

[20] Sudhakar Bobba, Ibrahim N. Hajj, "Input Vector Generation for Maximum Intrinsic Decoupling Capacitance of VLSI Circuits," *2001 IEEE International Symposium of Circuits and Systems, Volume 5*, pp. 195-198, 6-9 May 2001

[21] Modira Deb Pant, Pankaj Pant and Donald Scott Wills, "On-Chip Decoupling Capacitor Optimization Using Architectural Level Current Signature Prediction," *13th Annual IEEE International ASIC/SOC Conference*, pp. 288-292, Sept. 2000

[22] J. Choi, S. Chun, N. Na and M. Swaminathan and L. Smith, "A Methodology for the Placement and Optimization of Decoupling Capacitors for Gigahertz Systems," *Proceedings of the 13th International Conference on VLSI Design*, pp 156-161, 3-7 Jan 2000

[23] Möndira Deb Pant, Pankaj Pant, and Donald Scott Wills, "On-Chip Decoupling Capacitor Optimization Using Architectural Level Prediction," *Proceedings of the 43rd IEEE Midwest Symposium on Circuits and Systems, Volume 2,* pp 772-775, 8-11 Aug. 2000

[24] LK Wang, and Howard H. Chen, "On-Chip Decoupling Capacitor Design to reduce switching-noise-induced instability in CMOS/SOI VLSI," *IEEE International SOI Conference*, pp. 100-101, 3-5 Oct. 1995

[25] G. Bai, S. Bobba and I.N. Hajj, "Simulation and Optimization of the Power Distribution Network in VLSI Circuits," *IEEE/ACM International Conference on Computer Aided Design*, pp. 481-486, 5-9 Nov. 2000

[26] Haihua Su, Kaushik H. Gala and Sachin S. Sapatnekar, "Fast Analysis and Optimization of Power/Ground Networks," *IEEE/ACM International Conference on Computer Aided Design*, pp. 477-480, 5-9 Nov. 2000

[27] Rajendran Panda, David Blaauw, Rajat Chaudhry, Vladimir Zolotov, Brian Young and Ravi Ramaraju, "Model and Analysis for Combined Package and On-Chip Power Grid Simulation", *International Proceedings of the 2000 Low Power Electronics and Design*, pp. 179-184, 26-27 July 2000

[28] Yi-Min Jiang, Kwang-Ting Cheng and An Chang Deng, "Estimation of Maximum Power Supply Noise for Deep Sub-Micron designs," *1998 International Symposium on Low Power Electronics and Design*, pp. 233-238, 10-12 Aug. 1998

[29] Haihua Su, Kaushik H. Gala, Sachin S. Sapatnekar, "Fast Analysis and Optimization of Power/Ground Networks," *IEEE/ACM International Conference on Computer Aided Design*, pp. 477-480, 5-9 Nov. 2000

[30] Howard H. Chen and J. Scott Neely, "Interconnect and Circuit Modeling Techniques for Full-Chip Power Supply Noise Analysis," *IEEE Transactions on Components, Packaging and Manufacturing Technology, Volume 21, Issue 3,* pp. 209-215, Aug. 1998

[31] Howard H. Chen and David D. Ling, "Power Supply Noise Analysis Methodology for Deep-Submicron VLSI Chip Design," *Proceedings of the Design Automation Conference*, pp. 638-643, 9-13 June 1997

[32] Kenneth L. Shepard, "Design Methodologies for Noise in Digital Integrated Circuits," *Proceedings of the Design Automation Conference*, pp. 94-99, 15-19 June 1998

[33] R. Saleh, M. Benoit, P. McCrorie, "Power Distribution Planning", *Design, Automation and Test in Europe Conference*, Paris, France, pp. 265-270, Feb. 1998.

[34] William Liu, "MOSFET Models for SPICE Simulation including BSIM3v3 and BSIM4," *John Wiley & Sons*, Inc., 2001.

[35] K.M. Cao, WC Lee, W. Liu, X. Jin, P.Su, S.K.H. Fung, J.X. An, B. Yu and C. Hu, "BSIM4 Gate Leakage Model Including Source-Drain Partition," *IEDM Technical Digest of Electron Devices*, pp. 815-818, 10-13 Dec. 2000