On-Chip Power Supply Noise:

Scaling, Suppression and Detection

by

Tasreen Karim

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2012

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Tasreen Karim

Abstract

Design metrics such as area, timing and power are generally considered as the primary criteria in the design of modern day circuits, however, the minimization of power supply noise, among other noise sources, is appreciably more important since not only can it cause a degradation in these parameters but can cause entire chips to fail. Ensuring the integrity of the power supply voltage in the power distribution network of a chip is therefore crucial to both building reliable circuits as well as preventing circuit performance degradation. Power supply noise concerns, predicted over two decades ago, continue to draw significant attention, and with present CMOS technology projected to keep on scaling, it is shown in this work that these issues are not expected to diminish.

This research also considers the management and on-chip detection of power supply noise. There are various methods of managing power supply noise, with the use of decoupling capacitors being the most common technique for suppressing the noise. An in-depth analysis of decap structures including scaling effects is presented in this work with corroborating silicon results. The applicability of various decaps for given design constraints is provided. It is shown that MOS-metal hybrid structures can provide a significant increase in capacitance per unit area compared to traditional structures and will continue to be an important structure as technology continues to scale. Noise suppression by means of current shifting within the clock period of an ALU block is further shown to be an additional method of reducing the minimum voltage observed on its associated supply. A simple, and area and power efficient technique for on-chip supply noise detection is also proposed.

.

Acknowledgements

As the alchemist has his crucible, so have I been blessed with my crucible that has been uncovered through the world of loving, kind and giving people that I have been surrounded by on this journey. No words are sufficient to describe the immense gratitude I have for you all and while I cannot possibly name everyone here please know that your kindness too is greatly appreciated.

I am grateful for the unconditional love and abundance that has carried me through this entire journey of growth. The love from my husband Karim that is with me every step I take and has touched and awakened in me great faith and belief in myself. His incredible ability to excel at anything with ease is a great source of inspiration to me and his generosity and respect for even the smallest of beings a reminder of how to live a great life. He makes every aspect of my world joyful beyond measure. I am also stupendously thankful to my parents who have always provided me with the feeling of certainty that I need to venture out and do all the things I choose. To my father Kabir for his unconditional love and for the sense of freedom this brings. I am thankful to my mother Farida for her incredible caring and generosity that she shares and continues to teach me every day. I am thankful to my sister Shafina for always giving every ounce of herself no matter how big or small the need. I am thankful to both my miracles, my daughters Hana and Quesara, for it is their zest for life that reminds me how beautiful life is. Their delightful smiles and warmth bring more bliss than I had ever imagined possible and I thank them for the person they have helped me become. I am also deeply thankful to all my in-laws, especially my mother-in-law Dr. Dilshad Karim, a truly amazing woman, and great role model to me, and my father-in-law Dr. Sallaudin Karim, for all the peace he brings.

Of course I would not be here was it not for the grace and profound generosity of both my supervisors. I am thankful to Professor Sachdev whose truly kind and caring nature has enabled the

forging of the various facets of my experience to be as beautiful as they are. I am thankful for his persistence and desire in making this degree what it is today and for never ceasing to have faith in me along the way. I am always touched by his generosity and his finding of a way to give of himself in any situation. His ability to bring people together has enabled the creation of a truly magnificent research team which I am thankful for, for my every day spent with them is a cherished pleasure. I am also immensely thankful to Professor Opal who has been a remarkable example of excellence to me and has really helped me experience the true pleasures of doing research. His undivided attention and care put into all the work presented to him has touched me greatly and shown me how to savor and enjoy life in the moment. I have immensely enjoyed all our discussions and they have played a large role in how the research has progressed. His respect and value for others is yet another valued lesson I have learned from him over the years. I am, again, truly grateful to both my supervisors for all that they have shared with me. Their desire to persistently give has helped me grow in many ways and has changed my life in ways that cannot be described.

I would also like to take this opportunity to immensely thank my PhD Committee Members. Your time, sharing of your knowledge and feedback is very much appreciated in making both myself and the work richer. Thank you to Professor Barby for being an inspiring leader, to Professor Nairn for sharing of your immense knowledge and Professor Martin for all the effort and caring you put into understanding this work that is outside your field. A sincere and special thank you to Professor Harjani also for contributing your time to this endeavor and for making the trip all the way to Waterloo. I have been an admirer of your work and I am immensely thankful and honored to have you serve on my Committee.

I am further very thankful and privileged to have been given the opportunity to meet some great minds in the field of power supply noise at Qualcomm Incorporated. Your generosity in sharing of

your time and knowledge touched me greatly and is appreciated immensely. My sincere thanks to my husband for the introduction to this team.

I would like to thank every person in our outstanding research team. A special thanks to Pierce Chuang, an example of a truly wonderful human being. One need not even ask for him to lend a helping hand and he does so with as much willingness, care and meticulousness as for his own work. My sincere gratitude goes to him for his kindness and selflessness and for helping me with my research and truly making my experience a joyous one. Similar thanks go to David Rennie who also helped me immensely with my research and also to have faith in myself. His generosity will always be one I will remember to pay forward. My deep thanks also go to Jaspal Singh Shah for great conversations and for also lending a helping hand whenever needed. Tahseen Shakir's generosity has also not gone unnoticed and I thank him for playing a large part in creating a wonderful, open and giving culture within the group. My thanks also to Adam Neale for inspiring and enjoyable conversations, and to David Li for the many fun moments he brought to the group. My sincere thanks also go to Hassan and Hossein Sarbishaei, Maofeng Yang, Daniel Febrowski and Dylan Bespalko for their assistance with my research. A heartfelt thank you to Sara Attar also for giving so kindly of yourself and for inviting me into your lab for testing of my chip. I also thank Nitin Mohan for his helpful thoughts on my research over email. Thank you also to Shah Jahinuzzaman, Danny Tsuei, Manuchehr Ibrahimi, Qing Li, Shahab Ardalan and Sumanjit Singh Lubana for all your contributions to my experience.

My acknowledgements would not be complete without thanking all the administrative and support staff at the University of Waterloo, especially Phil Regier. His ability to solve any computer related problem on any day at any time has been an immense source for gratitude during my work. And thank you to Mary McColl, Deborah Perchaluk, Annette Dietrich, Lisa Hendel, Susan King, Wendy Boles,

Diana Macfarlane, Karen Schooley, Chris Schroeder, Betty Slowinski, Lisa Szepaniak, Kathy Holston and Irena Baltaduonis for your friendship and making my experience here richer.

I am grateful beyond words for the experiences I have had at Hildegard Marsden Nursery - a true well of smiles and love on-campus. I am thankful to all of my children's teachers and especially to Rita, Rebecca, Marcia, Anzhela, Melissa, Shannon, Leanne, Lora, Heidi, Tanya, Stacey, all the support staff and especially Karen Hinnigan, for the love they have shared with our family and our children and for the peace of mind they provide us with while we venture into our journeys of research.

I am awed by and grateful for all the role models that have appeared in my life during the course of my degree and life, and for the miracles they have brought with them.

Lastly, I am thankful to everyone for the opportunities that they have blessed me with to contribute in return to them. I am especially thankful for the opportunities to give to my family, co-workers, my HMN family and the Ismaili Community. The pleasure has been all mine for it is really in giving of ourselves that we receive.

May you all find your crucibles.

This work is dedicated to all the kindred souls that read it…

may your lives be richer in some way for the words your eyes fall upon….

Table of Contents

List of Figures

List of Tables

The Journey

As a child I existed
My body fed
And so my soul
Though bathed in Love I did not see
A time for messages to receive
Then I walked down a road that led to a place
Pulled by a powerful force unknown
Through the web I wandered
With excitement and joy
With anger and sadness
Mostly confusion yet desire
Now as I approach the end of the journey
I see there is no end at all
Into life I do not go
For inside life I already am
From the earth I receive
And to the earth I return
Part of the whole
The eyes don't see
Surrounded by Love my own
No one purpose
No continuity
Simply Giving
Simply Love
Conversations I may have
Names may come my way
Yet
Simply Giving
Simply Love
Simply Abundance
Simply Joy
And
Simply Now

Chapter 1

Introduction

"if the circuit fails, it does not matter how small the circuit is, how fast it runs, or how little power it consumes"

… Howard H. Chen and J. Scott Neely (IBM)

This chapter provides a context for and definition of power supply noise. It then introduces the problem of supply noise and briefly looks at how the evolution of technology affects the noise. It concludes with detrimental effects of supply noise on various circuits and stages of the design flow.

In the *analog* circuit domain, the noise sources of interest are generally those that result from the physical properties of silicon devices and include shot noise, thermal noise, flicker noise, and burst or popcorn noise [1]. The popularity of *digital* circuits, on the other hand, is largely due to their associated noise immunity to such physical noise sources. However, the high gain of digital circuits results in an alternate class of noise sources, which can be considered as "man-made" sources [2]. These noise sources are orders of magnitude larger than the physical noise sources described and include leakage noise, charge-sharing noise, reflection noise, and power supply noise [2].

Leakage current in Complementary Metal-Oxide-Semiconductor (CMOS) Field-Effect Transistors (FETs) can contribute up to 50% of the total chip current in sub-100 nm technologies [3] and is primarily a function of the device threshold voltage and temperature. As the device threshold voltages are lowered,

an increase in leakage currents is observed, resulting in a reduced noise margin available for other sources. Charge-sharing noise is mainly caused by the redistribution of charge between various capacitive nodes within the circuitry resulting in a degradation of signal integrity. Sometimes the charge sharing problem can be mitigated by means of keeper devices that restore signal levels or by employing other circuit techniques. Cross-talk noise occurs when inductive and/or capacitive coupling is present between neighboring signal lines on a chip. When a change in signal occurs on one line, these changes are transferred onto adjacent lines via these coupling effects which are observed as noise on the victim lines. As the lengths over which signals are routed increase, crosstalk noise increases, making this source of noise an important contributor in state-of-the-art chips. Another noise source seen in high frequency CMOS circuits is reflection noise. This noise occurs at each impedance discontinuity on a transmission line and essentially occurs when the signal on a line is reflected back on itself due to impedance mismatches on the line [4]. The various noise sources described thus far, like on any other signal, can cause fluctuations on the power supply signals, however, the term 'power supply noise' or 'supply noise' is typically used to describe *switching noise*, which is the primary source of the supply fluctuations. Switching noise is correlated to the switching activity and current consumption of a circuit in the presence of inductive, capacitive and resistive parasitics along the power grid.

The time-domain behavior of power supplies can be abstracted into two categories: supplies with DC noise, and supplies with pulse noise [2], as illustrated in Figure 1.1. DC noise is a result of a change in the steady-state level of the supply, with leakage noise being an example of such a noise source since it causes an approximately constant current draw from the supply. With pulse noise, the variation in supply is instigated by an on-chip event and the noise dissipates once the event has occurred, returning the supply to its steady state value. Charge-sharing noise, cross-talk noise and power supply noise are examples of pulse noise, where the noise event is determined by a changing signal on the chip. DC noise can be

limited during the design stage and the power supply grids of present day microprocessors are commonly designed to have sub-mΩ impedence targets [5] to minimize this noise. Pulse noise, however, can be more of a problem since it can be unpredictable in both time and magnitude.

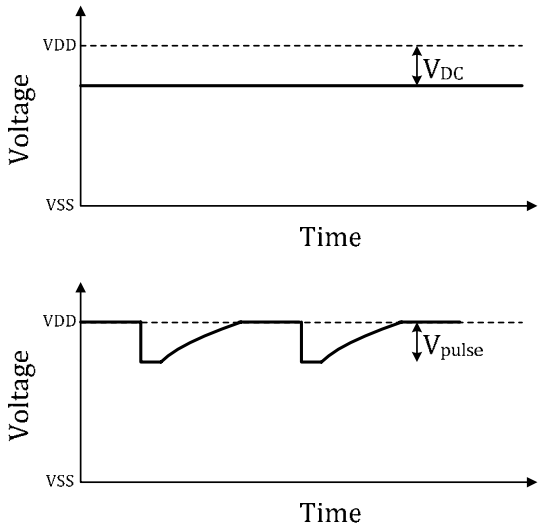


Figure 1.1 Time domain abstractions for supply noise (adapted from [2])

When referring to the noise of a power supply waveform various quantities can be used to identify the noise magnitude. The most common characteristic used to quantify the noise is the maximum droop or peak in the supply voltage from the nominal value. Correspondingly the standard deviation of the waveform can also be used as a measure of the deviation from the nominal value. The peak-to-peak and average supply voltage [6] can further be quoted as a measure of noise. In the work presented here, the term supply noise is used to refer to the maximum droop in voltage from the nominal value.

1.1 The Power Supply Noise Problem

As stated, switching noise is a function of the switching activity and current consumption of a circuit in the presence of inductive, capacitive and resistive parasitics along the power grid. Consider the simplified depiction of a power delivery system illustrated in Figure 1.2. Clearly this is an over-simplification of the

entire power distribution system of a chip, however, the model is used here to present the fundamental effects taking place. The variable current source represents the current draw *I(t)* of a switching circuit as a function of time. The ideal power source provides the nominal supply and ground voltages to the circuit through finite, lumped supply grid parasitic resistances *R* and inductances *L,* assumed to be equal in both the power and ground path for convenience, that result from both on-chip interconnect and off-chip packaging components. Capacitance *C* represents on- and off-chip capacitance between the supply rails including circuit capacitance and any added decoupling capacitance.

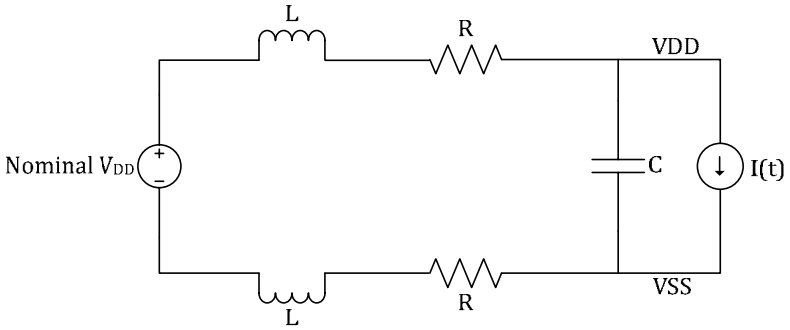


Figure 1.2 Simplified depiction of a power delivery system

There are three causes that result in the voltage *VDD* and *VSS* deviating from their nominal values. Firstly, as the current draw increases, a dip in *VDD*, ΔV_R, and corresponding peak in *VSS*, are experienced due to the voltage drops across the parasitic resistances *R*. The magnitude of this *IR noise* is given by

$$\Delta V_R = I(t) \cdot R \tag{1-1}$$

Secondly, current transients across the parasitic inductances *L* result in *dI/dt noise*, also known as *delta-I noise*, which owes its names to the relation

$$\Delta V_L = L \cdot \frac{dI(t)}{dt} \tag{1-2}$$

Lastly, the parasitic inductance and capacitance form an LC-tank circuit which when excited by repeated switching at the resonant frequency causes the system to oscillate. When insufficient resistance is present

in the system to dampen the oscillations, this *resonance* can be a persistent noise source making it problematic since it affects circuits over a relative long period of time.

Based on these relationships, the faster the circuits switch and/or the more current they draw in the presence of the inductive and resistive effects, the larger the noise seen on the supply lines. Any noise on the power supply can be detrimental as it can couple onto numerous evaluation nodes of a circuit. If the power supply voltage droops too low, performance and functionality of the circuits can be compromised. Conversely, large overshoots can have circuit reliability implications, such as, electromigration problems and degradation of the device gate oxide layer. The overshoots and undershoots present on noisy power supplies not only affect the actively switching circuits but also present potential problems for non-switching circuits and analog circuitry that receive power from the same supply. Thus, the power supply delivered to a circuit must maximize performance while maintaining reliable operation.

1.2 Evolution of Technology

Here we look at how technology scaling has affected power supply noise over the years and how it is expected to affect this noise source in times to come. Since the invention of the transistor in 1947 by John Bardeen of Bell Labs, and the co-inception of the first monolithic integrated circuit (IC) in 1958 by Jack Kilby of Texas Instruments and 1959 by Robert Noyce of Fairchild Semiconductor, "silicon chip" technology has advanced at an astounding rate. As Gordon Moore, of Fairchild Semiconductor at the time, noticed and predicted in his well known 1965 paper, the number of transistors placed inexpensively on a single chip has doubled initially approximately every year and since 1975 every two years, and according to the 2010 International Technology Roadmap for Semiconductors (ITRS2010) [7] the number of transistors on a single chip will continue to increase over the next decade. While ICs of various sizes are ubiquitous in modern society, Intel's 10-core Xeon Westmere-EX boasts the highest transistor count

in a commercially available CPU today with over 2.5 billion transistors. Figure 1.3 illustrates the trend in transistor count of various Intel Central Processing Units (CPUs) over the years. As a result of scaling transistor dimensions, transistor speeds and correspondingly clock frequencies have also increased until recently where the advent of multi-core processors has led to the clock frequencies being held approximately steady as seen in Figure 1.4. One reason for the move to multi-core from single-core is the exponentially increasing power consumption with frequency making multi-core processors more power efficient than faster single-core processors. In other words, a dual-core processor consumes less power than a single-core processor with twice the speed.

With increasing transistor densities and increasing chip dimensions, the current densities and overall current consumption of chips has also seen a general increase over the years as shown for various Intel CPUs in Figure 1.5. Even with a shift towards multi-core processors and even if the power is kept within existing levels, the projected decrease in supply voltage to be described, will necessarily result in an increase in chip currents.

Scaling of the nominal power supply voltage level for a given technology has followed somewhat of an unpredictable trend over the years. The scaling of transistors, first described by Dennard *et al.* [8], known as *ideal scaling*, or constant electric field scaling, assumes a scaling factor of $S>1$ where all transistor dimensions, including the device width and length, oxide thickness and junction depth, scale uniformly as $1/S$, the supply voltages scale as $1/S$ and the doping concentrations scale as S. This method of scaling maintains the electric field across the transistors thus ensuring proportional scaling of the transistor current-voltage characteristics, leading to greater device density, higher performance and reduced power consumption in the scaled devices. Scaling the supply voltage arbitrarily is, however, not always a feasible option since adapting existing external components to the new voltages can be significantly expensive [9]. Therefore, an alternate form of scaling, *fixed-voltage scaling*, was solely used up to the 0.5

μm technology node [9]. In this type of scaling, the supply voltage is kept constant while the dimensions are scaled by 1/*S*. Table 1-I summarizes the scaling factors for various device and circuit parameters for both ideal and fixed-voltage scaling in short channel, velocity saturated devices [9]. As is evident, a key disadvantage of fixed-voltage scaling of velocity saturated devices is that the performance gain achieved, which is the same as in the ideal scaling case, comes with a major power penalty. It can be noted that in the case of long channel devices, keeping the supply voltage constant while scaling the dimensions would give a greater performance advantage compared to their short channel counterparts. All modern day CMOS devices, however, are affected by short channel effects. Furthermore, oxide reliability constraints prevent the voltage from being held constant while scaling the oxide thickness which has gone from about 100 nm in the 70's to less than 1 nm in current technologies [5].

Another disadvantage of ideal scaling is that some intrinsic device voltages such as the bandgap of silicon and built-in *p-n* junction potential cannot be scaled. In addition, the threshold voltage of the devices cannot be aggressively scaled since lower threshold voltages pay a penalty in terms of device leakage currents. Therefore *general scaling* is followed in practice where the dimensions are scaled by the factor 1/*S* while the supply voltages are scaled by an alternate factor, $1/S_V$, commensurate with the limitations associated with voltage scaling. The trend in supply voltage for various Intel processors is given in Figure 1.6 (a) and that predicted for future technologies according to the ITRS [7] is given in Figure 1.6 (b).

With the aforementioned increase in chip current densities, the deviations from nominal supply voltage levels according to (1-1) and (1-2) is expected to increase with scaling. In addition, the decrease in the nominal supply voltage level itself is expected to result in a further degradation of signal to noise ratio.

TABLE 1-I SCALING FACTORS FOR CMOS DEVICE AND CIRCUIT PARAMETERS

PARAMETER	SCALING FACTOR		
	Ideal or Constant Electric Field Scaling [8]	Fixed-Voltage Scaling	General Scaling
Device Dimensions (L, W, t_{ox})*	1/S	1/S	1/S
Supply Voltage	1/S	1	1/ S_V
Saturation Current	1/S	S	S/ S_V^2
Current Density (I_{sat}/Area)	S	S^3	S^3/ S_V^2
Intrinsic Delay ($R_{on}C_{gate}$)*	1/S	$1/S^2$	S_V/S^2
Power Per Device	$1/S^2$	S	S/ S_V
Power Density (Power/Area)	1	S^3	S^3/ S_V

* L = transistor length, W = transistor width, t_{ox} = transistor gate dielectric thickness, R_{on} = device on resistance, C_{gate} = gate capacitance

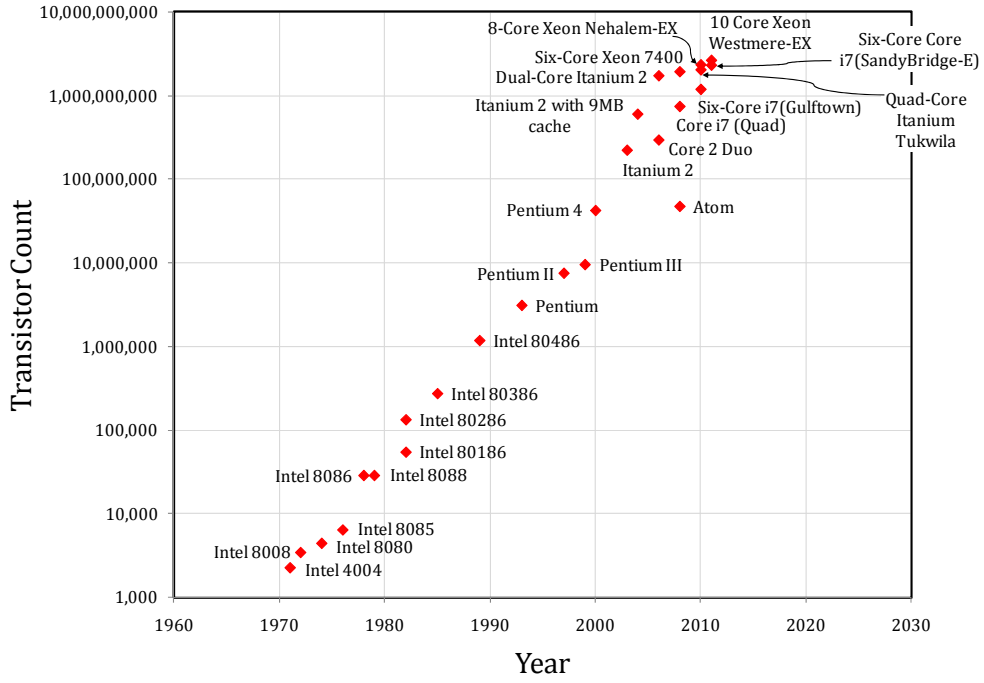


Figure 1.3 Trend in transistor count of various Intel CPUs

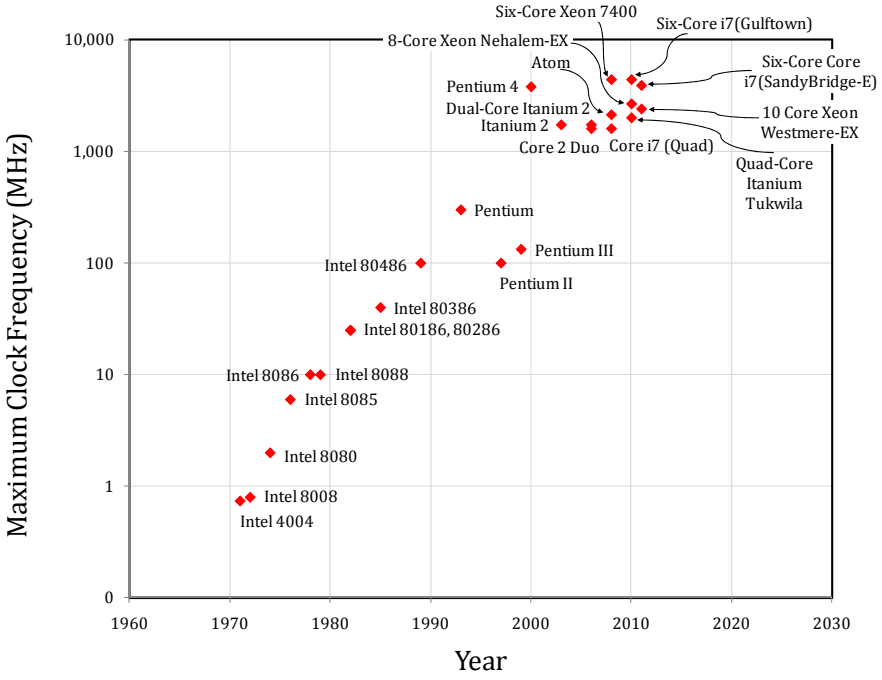


Figure 1.4 Trend in maximum clock frequency of various Intel CPUs

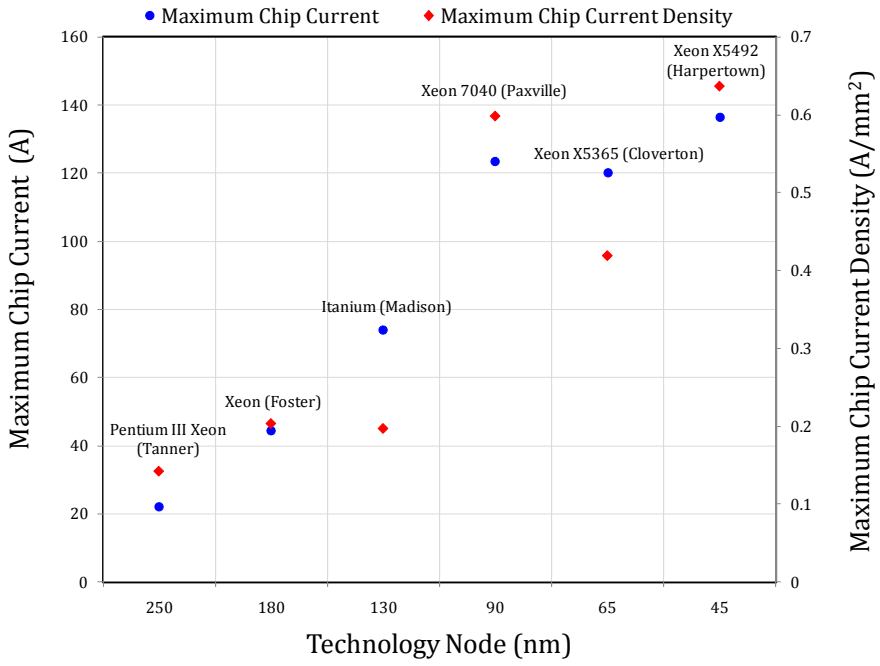


Figure 1.5 Trend in maximum current draw and density of various Intel CPUs

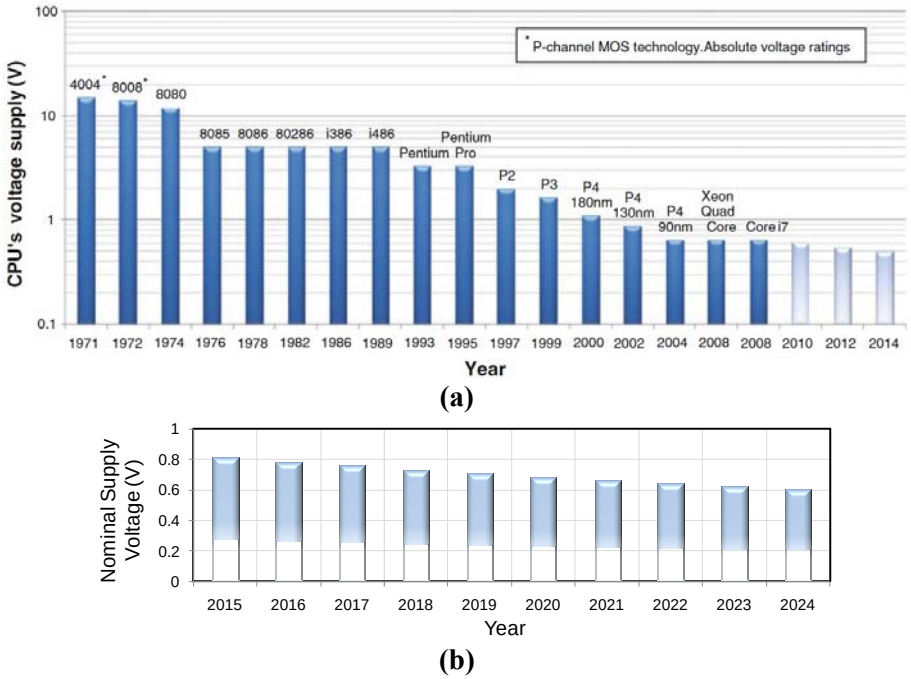


Figure 1.6 Trend is power supply voltage for (a) various Intel processors [10], and (b) predicted for future CMOS technologies [7]

1.3 Detrimental Effects of Power Supply Noise

Switching noise is of concern not only in mixed signal and sampled-analog circuits, but also in purely digital systems where it can cause both logic and timing errors [11]. However, because analog circuits normally require stable bias points they can be more sensitive to power supply noise since variations in the supply can change these bias points and thus affect important performance parameters such as gain and linearity of the circuits. For digital logic, the situation is less severe, but important nonetheless, since the supply noise can exceed the noise margin provided.

When dealing with power supply noise, its suppression, or mitigation, is of immense importance in managing supply noise as it prevents the noise from reaching sensitive circuits. This mitigation can be passive in that it suppresses fluctuations continuously over time or it can be active in that the suppression

is activated at certain instances in time when the fluctuations become intolerable. Another area of importance is simply the detection of supply fluctuations to obtain knowledge of the supply noise amplitude and frequency. Modeling of noise generation processes for large circuits prior to their fabrication is generally prohibitively expensive with respect to time and CPU resources due to the large number of components present in the supply grid. Post manufacture supply noise data can thus be very beneficial. The supply waveforms must be obtained using on-chip measurement circuits since at high frequencies, it is extremely difficult to obtain this information via probing of pads using test equipment [12][13]. The logic used to drive the supply signals off-chip, such as buffers in the pads, and parasitics of the probes, tend to suppress the variations in the supply signal thus resulting in distorted measured signals. It is therefore necessary to determine the signal waveforms before driving them off-chip. Following are examples of some undesirable effects of supply noise, where measurement and mitigation of the noise could be beneficial.

1.3.1 Increased Gate Delay and Logic Errors

In modern chips there are three primary concerns raised by power supply noise. The first is '*flop-to-flop' gate delay*. For most types of CMOS logic, the critical path delay is inversely proportional to the supply voltage [14][15] and a dip in supply voltage can thus increase the delay of the logic block. Gate delay has become significantly more sensitive to supply voltage variations with scaling [16][17]. In a 0.13 μm technology, a 10% voltage variation has been shown to result in a 30% variation in the delay of typical gates [17], and in a 90 nm technology, a 1% voltage variation has been shown to cause approximately a 4% variation in delay [16]. Large delays can further result in logic errors. Consider the block diagram in Figure 1.7 comprising a logic block between two flip-flop circuits. The first flip-flop provides an input to the logic block at a clock edge and the second flip-flop stores the output from the logic block at the subsequent clock edge. If the delay of the logic block is larger than the clock period, the

second flip-flop will be triggered before the logic block has completed its evaluation, thus resulting in a potential error in the latched logic level.

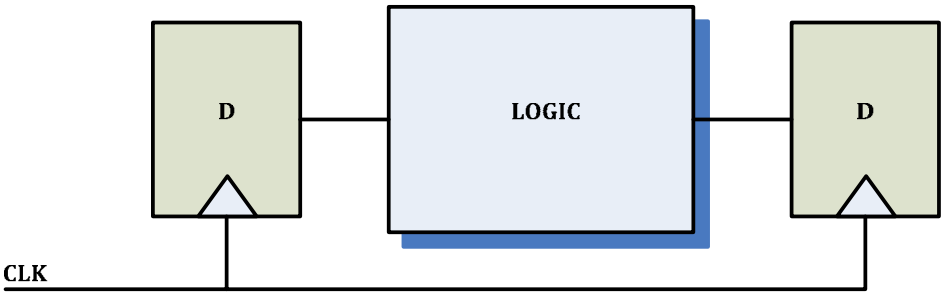


Figure 1.7 Digital block diagram to illustrate delay errors

1.3.2 Instability of On-Chip Storage Elements

Secondly, on-chip storage elements such as flip-flops, latches, registers, and static random-access memories (SRAMs), have stringent limits on the minimum supply voltage, V_{min}, they can experience. Below this voltage, errors in the stored data can occur. Consider, for example, and SRAM cell. As CMOS technology scales, maintaining the stability of dense static random-access memories (SRAMs) continues to be a challenge. The stability of an SRAM cell essentially refers to the cell's ability to hold on to its data in the presence of varying conditions during a data read cycle [18], as it is during this cycle that the cell is most vulnerable to noise. Power supply noise is one such source of variation that can adversely affect the stability of a cell [18]. As illustrated in Figure 1.8, an SRAM cell typically employs two back to back inverters to store a digital state. During a read operation, the word line (WL) devices M5 and M6 are turned ON connecting the precharged bit lines, *BL* and $\overline{BL}$, to nodes Q and $\overline{Q}$ respectively. If a "0" is stored on node $\overline{Q}$, noise on WL originating from the power supply can cause the resistance of transistor M_5 to vary such that either $\overline{BL}$ is not sufficiently discharged or the high value of $\overline{BL}$ overwrites the "0"

stored on node $\overline{Q}$. Similarly, noise on $\overline{BL}$ can cause node $\overline{Q}$ to flip. Errors in writing data to an SRAM cell can similarly occur as a result of power supply noise [19].

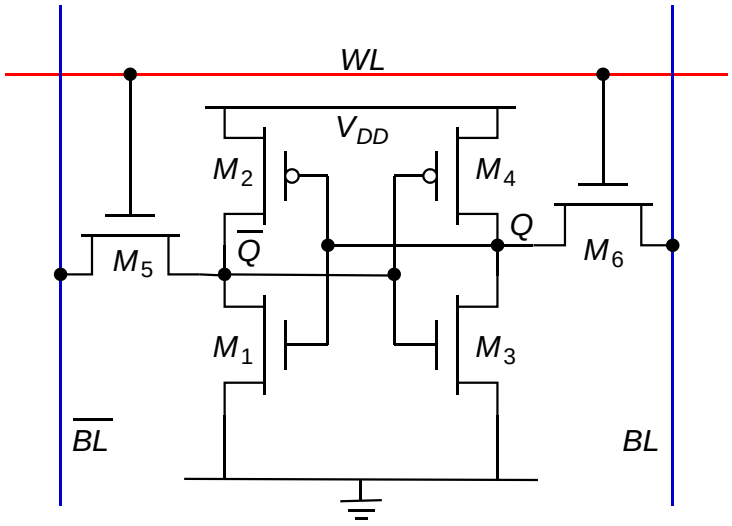


Figure 1.8 An SRAM cell [9]

1.3.3 Timing Signal Jitter

A third undesirable effect of power supply noise is its effect on the timing signals present on a chip. Noise on the supply of circuits that generate these timing signals, such as digital clock generation circuits and phase-locked loops (PLLs), results in timing jitter and can adversely affect circuits making use of the timing signals generated. Work done in the 0.35 μm and 0.18 μm technology nodes has shown that the amount of jitter introduced in the timing signals is, not surprisingly, linearly proportional to the standard deviation of the power supply noise, as well as both a function of instantaneous and past, integrated behavior of the power supply noise [20]. Digital clock generation circuits (which essentially use a combination of inverters, NAND and NOR gates) are often used to supply signals to analog-to-digital

converters (ADCs) and in high-speed ADCs noisy power supplies can significantly restrict their jitter budgets. PLLs are also integrated into many ICs for frequency multiplication and skew reduction, and comprise voltage-controlled oscillators (VCOs) for generating the clock signals. These circuits have limited power supply rejection and thus changes in the power supply are reflected as timing jitter in the generated clock signals. Referring again to Figure 1.7, jitter in the timing of the clock signal can cause the second flip-flop to trigger before the logic block has completed evaluation. In addition, if the power supply noise extends over a number of clock cycles, each clock edge can deviate progressively more from the ideal edge location which can result in synchronization errors between different clock domains.

1.3.4 Testing, Debug and Reliability Implications

In addition to the above three major effects, power supply noise can also have testing, debug and reliability implications. Traditional digital test methods such as automatic test pattern generation (ATPG) and built-in self test (BIST) generally deal with the digital domain and are unable to address the more analog issues of high-speed design. Knowledge of the supply noise magnitude can be useful in the testing and debug of manufactured chips in correlating power supply fluctuations to logic or delay errors. The debug stage of testing can be extremely time consuming and any reduction in this time can reduce the overall time-to-market of products with the associated financial benefits.

Furthermore, during testing, it is not always practical to determine the input patterns that will induce the worst case switching noise since the complexity of the solution space is exponentially proportional to the number of primary inputs, quickly causing the effort to become futile [21]. Statistical methods have been proposed in the past where switching activity [22] and power dissipation [23][24] have been estimated, and methods that estimate near worst case noise have also been proposed [21]. However, these methods, although useful in predicting noise, inevitably suffer from simplification errors resulting in the possibility of greater than expected noise during operation. Thus, a low resource, active noise suppression

technique would be beneficial in such situations where the suppression can be activated in events that the supply noise exceeds expected levels. At-speed delay tests are also often performed during testing to detect small manufacturing defects that reduce circuit speed without actually causing functional failure [25]. When generating the test patterns for such tests, a random fill of don't care bits is often used to increase the odds of detecting defects. However the use of a random fill can lead to large power supply noise that is drastically overestimated [26]. In this case a power supply suppression technique can be very useful in reducing the supply noise for testing purposes.

Reliability problems can further result when the power supply noise causes the supply to be higher than anticipated. These overshoots can cause issues such as electromigration [9] problems and hot carrier effects (HCE) [27]. Hot carrier effects result from high applied voltages in the drain region of transistors which lead to a gradual degradation of transistor characteristics [28]. The damage is commonly accepted to be due to trapping of carriers in the gate dielectric or the creation of interface states at the silicon-dielectric interface [28]. The degradation in the quality of the gate oxide and its interface with the silicon results in a shift in threshold voltage and subsequently a decrease in the drain currents of the device. Another phenomenon that affects the reliability of the devices is negative bias temperature instability (NBTI) [27]. This effect occurs in PMOS devices and is a result of stressing the gate with negative voltages, which can occur on the ground supply lines. Again, it is commonly accepted that these voltages result in the creation of interface states that result in an increase in the threshold voltage and subsequent reduction in drain currents. Signal overshoots may not result in failure during testing of devices however, as seen above, can have significant reliability implications [12]. Measured supply noise signals can be useful in determining the frequency and magnitude of these overshoots thus allowing important predictions to be made with respect to the lifetime of the circuits [12].

1.3.5 Floor-planning and Supply Distribution Network Design

Power supply noise must also be considered during the floor-planning stage of design. Appropriate floor-planning [5][29] can reduce power supply noise by placing noisy blocks far away from each other and closer to the power pins, as the closer they are to the pins the less noise they will generate. Global on-chip power distribution networks are also generally designed during the early steps of a design cycle and necessary resources are allocated at this time. Although various methods are available to assist designers in identifying the noise *hot spots* on a chip, they can require knowledge of designs such as current consumption patterns which may not be available at the pre-design stage. Incorporating additional resources at later stages of the design process can be difficult and prohibitively expensive to implement especially in the microprocessor market where the sheer volumes make adding or removing even a single capacitor significantly expensive [5]. Power distribution networks are thus generally over-designed [30][31]. As technology continues to scale, and chip power consumption increases in future processor generation, this worst case design strategy will likely not be sustainable. Thus, as technology scales there will be an increased need for more efficient noise suppression techniques.

1.4 Thesis Scope and Organization

This thesis deals with various aspects of power supply noise. An overview of the power distribution network of a chip is first provided in Chapter 2. Various details of on-chip power supply noise are considered in Chapter 3, including a closer look at the scaling of on-chip supply noise. Chapter 4 deals with the area of power supply noise management. Here, passive decoupling capacitors in particular are considered in depth. In Chapter 5, active supply noise management techniques are considered and a noise suppression technique based on current shifting is presented. A brief analysis of the switched capacitor technique is also presented. Chapter 6 then looks at supply noise detection where an on-chip supply

measurement method is proposed. Lastly, a summary of the work presented and future work suggested in the field is provided in Chapter 7. Sections where significant contribution has been made to the field have been identified by the symbol 🏛.

Chapter 2

The Power Distribution Network

"The crux of the problem in designing a power grid is that there are many unknowns until the very end of the design cycle…[and] decisions ... have to be made ... when ...the chip design has not even begun."

… Abhijit Dharchoudhury *et al.* (Motorola)

This chapter provides an overview of the complete power distribution network from the power source to the on-chip IC. The electrical properties of the power distribution system are also described and a simplified model of the network provided. A brief introduction to designing power distribution networks is further included.

2.1 Physical Structure of Power Distribution Networks

The physical power distribution network spans several levels of hierarchy of an integrated circuit (IC) system as illustrated in Figure 2.1. The power supply enters the system via a voltage regulator module (VRM) which is essentially a DC-DC converter which converts an input DC voltage to the nominal supply voltage level (*VDD*) required by the IC chip. The printed circuit board (PCB), which physically

supports the various components of the system, also provides a conductive path for the supply between the VRM and the IC package via power and ground planes. The IC package is connected to the PCB by an array of contacts (e.g. ball grid array or pin grid array) which provide a power path between the PCB and power and ground planes within the IC. The IC package power planes are connected to the chip power distribution network using either a flip-chip or bondwire bonding technology (flip-chip shown in figure). Decoupling capacitors (decaps) are placed on-board, on-package and on-chip for reasons explained in a later section.

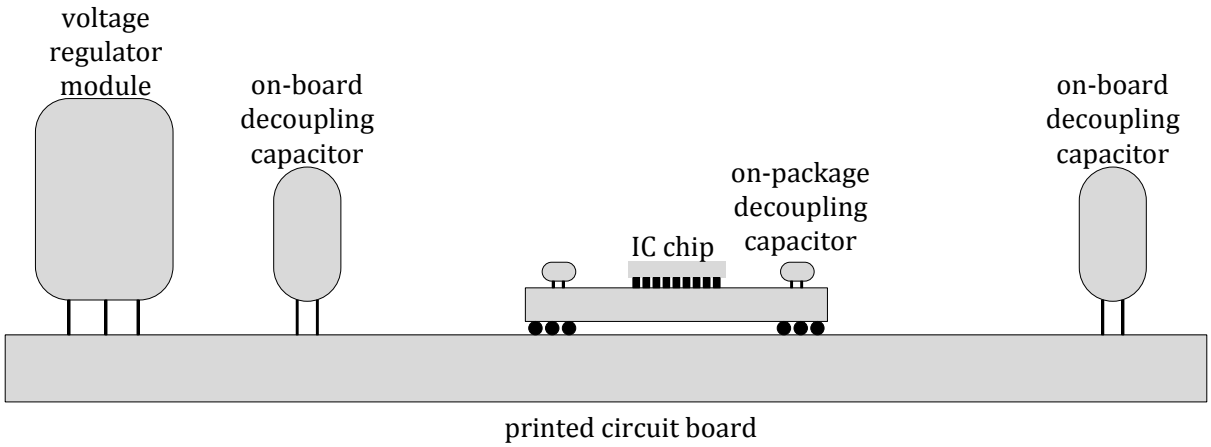
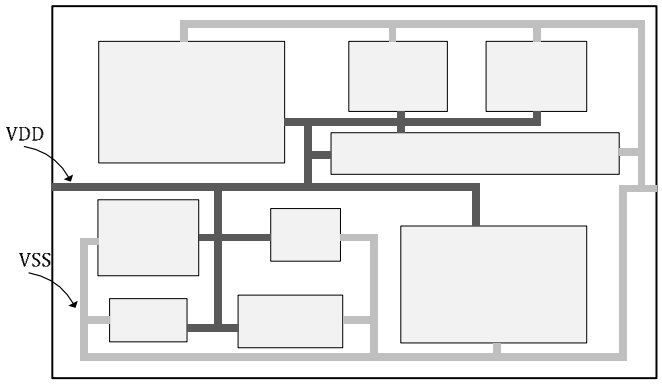


Figure 2.1 Power distribution network of an integrated circuit system

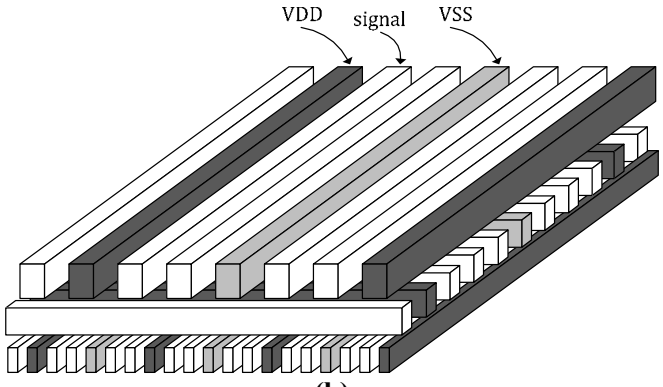
2.1.1 On-Chip Power Distribution Grid

There are a number of styles of on-chip power distribution networks, sometimes with more than one style used across a chip [32]. These routing styles range from an essentially ad hoc style, to grid-like structures, to power and ground planes. With the ad hoc style [33], local blocks are connected to the I/O pads by dedicated power/ground lines as illustrated in Figure 2.2 (a). The grid-like structure, however, is used in most modern high power ICs [34][35][36]. In this structure, illustrated in Figure 2.2 (b), multiple layers are used for the power grid where the power and ground lines are interdigitated within each layer

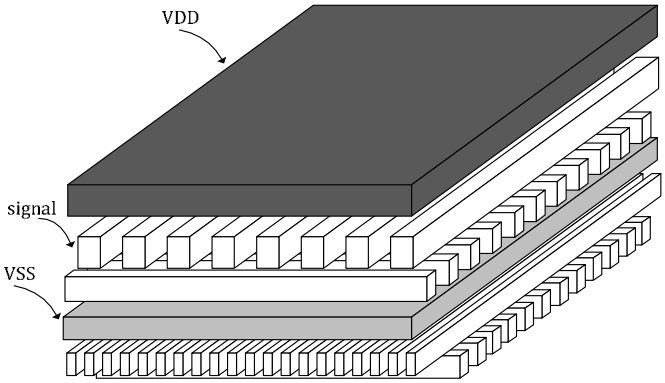
and orthogonal to each other from one layer to the next. The popularity of this design comes from the fact that the area between the power and ground line can be used for signal routing, and multiple redundant paths exist to the various circuits on the chip, which makes the power supply level less sensitive to the current requirements of individual circuit blocks. A failure in power delivery to one block also does not prevent power from being delivered to other circuit blocks. Furthermore, the grid structure also supports the use of dual supply voltages, which are commonly used in high-performance chips to reduce power consumption [32]. The grid structure allows the individual supply lines to easily be connected to the appropriate supply voltage. While power planes, as illustrated in Figure 2.2 (c), have been used in the past [34] and provide a low impedance path for the current, they undesirably result in entire metal layers being unavailable for signal routing.



(a)



(b)



(c)

Figure 2.2 On-chip power distribution network styles

2.2 Electrical Properties of Power Distribution Networks

2.2.1 Inductive Properties

The inductive properties play an important role in determining the overall behavior of a power distribution system. The classical definition of inductance [37] is a quantity that relates the change in magnetic field associated with a current flowing through a conductor to the voltage drop induced across the conductor. It represents the ability of the conductor to store energy in the form of a magnetic field. A conductive coil is generally used to explain the concept of inductance where its self inductance relates the change in its magnetic flux lines to a voltage drop across itself, and its mutual inductance relates the change in its magnetic flux lines to a voltage drop in an adjacent conductor.

The presence of current flow and corresponding magnetic fields associated with various traces and components present on-board and on-chip exhibit inductive properties. Due to the differing physical implementations of the various components, the electrical properties of components vary significantly

over the network. For example, the inductance of the board-level power distribution network is large compared to the inductance of the on-chip power distribution network which is relatively low. The inductance of the package-level power distribution falls somewhere in between the two [32].

2.2.2 Resistive Properties

The resistive properties of the various levels of abstraction in the power distribution network follow a somewhat reverse trend compared to the inductance [32] . The board-level power distribution network resistance is relatively low since the conductors are in the form of planes with relatively large thickness. The corresponding on-chip resistances, however, are relatively large due to the comparatively smaller dimensions of the on-chip power distribution network.

2.2.3 Capacitive Properties

In addition to off-chip capacitance in the system, on-chip circuitry also contributes to the capacitance present between supply rails [4]. These intrinsic capacitors include *n*-well capacitors present between *n*-well substrate contacts and substrate contacts on the *p*-type substrate, as well as circuit capacitors that are a result of circuit load capacitance that appear between the supply and ground. As illustrated in Figure 2.3 the *n*-well capacitor is a reveres-biased *p*-*n* junction capacitor between the *n*-well and *p*-substrate, where the contributed capacitance is a function of the area, perimeter and depth of the *n*-well.

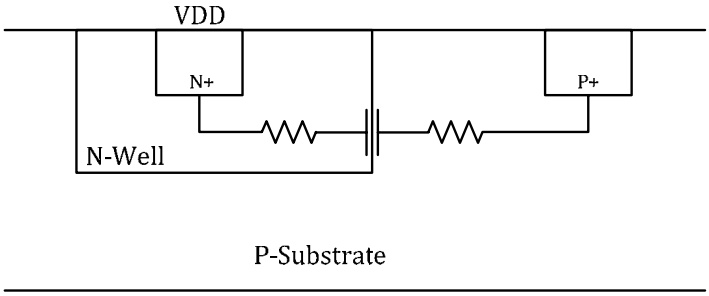


Figure 2.3 N-well junction decoupling capacitance

Figure 2.4 illustrates an example of circuit capacitors that are present due to circuitry that is not switching on a chip. In the case illustrated where the input is 1 and output is 0, the NMOS transistor will be turned ON connecting C_P from *VDD* to *VSS*. The capacitance C_P thus provides decoupling capacitance to other circuits that are switching. In the case illustrated where the input is 0 and the output is 1, the PMOS transistor will be turned ON connecting C_N from *VDD* to *VSS*. In this latter case, the capacitor C_N thus provides decoupling capacitance to the other switching circuits.

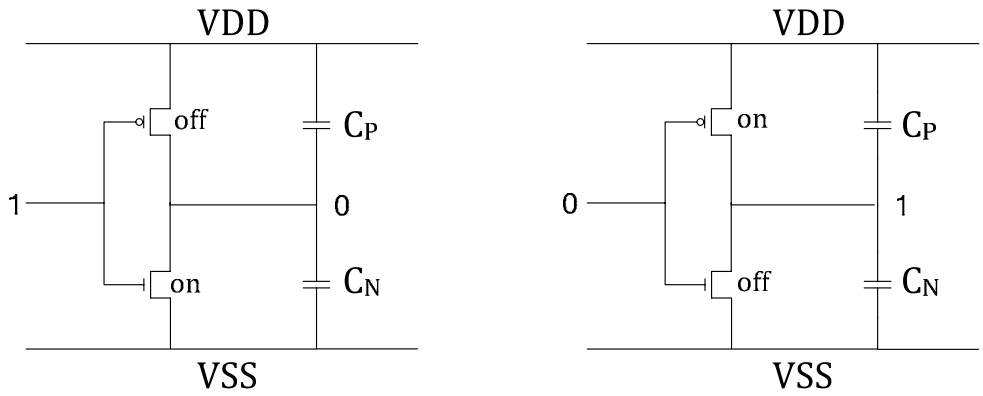


Figure 2.4 Circuit intrinsic decoupling capacitance

The total decoupling capacitance from non-switching circuits can be estimated as [4]:

$$C_{non} = \frac{P}{\left(V^2 f\right)} \cdot \frac{\left(1-\alpha\right)}{\alpha} \tag{2-1}$$

where C_{non} is the equivalent capacitance of the circuitry that is not switching, P is the power of the circuit, V is the supply voltage, f is the frequency of operation, and α is the switching factor which quantifies how often a circuit switches in a given cycle.

2.2.4 Power Distribution Network Design

The primary objective in designing a power distribution network is to maintain the grid impedance below a given level. The overall target impedance of the grid Z_{target} is calculated based on [38]

$$Z_{target} = \frac{VDD \times ripple}{I_{\max}} \tag{2-2}$$

where *VDD* is the nominal supply voltage level and I_{max} the maximum chip current draw. The term ripple refers to the percentage maximum noise or ripple allowed in the supply level. Thus, for a chip with a 1 V supply, maximum current draw of 100 A, and maximum allowable ripple of 10%, the target impedance is an alarmingly low 1 mΩ. Further complicating the issue is that the target impedance at each level of the network must be met over the relevant current transient times [39]. The on-chip current transient times are typically around 10% of the clock frequency.

2.2.4.1 Decoupling Capacitance

Given the inductive and resistive properties of supply networks, additional "intentional" capacitance, in the form of decaps, must be added to the system in order to bring its overall impedance down. The decaps essentially provide reservoirs of charge with varying response times to support surges in current draw by the IC that cannot be met due to voltage drops caused by the inductive and resistive components present. For practical circuits, the capacitance required is relatively large and the inductance requirements on the capacitor stringent, resulting in a prohibitively expensive on-chip solution. A more cost effective method of implementing this capacitance, as explained in [32], is to use a hierarchical placement scheme where significant capacitance is placed off-chip with the various capacitors targeting different frequency ranges. Each stage of decoupling capacitors determines the impedance characteristics for a given range of frequencies. Outside the particular frequency range, the corresponding capacitors have little influence on the particular impedance characteristics. The board-level decaps (electrolytic bulk capacitors) target low frequency noise and provide charge storage for current transients faster than the VRM can respond to. Successively closer decaps target progressively higher frequencies, thus the package-level (ceramic capacitors) and chip-level decaps provide charge with response times that their preceding level of supply

network are unable to respond to. Each tier of packaging hierarchy therefore has its associated decaps as illustrated in Figure 2.1. Even with the hierarchical placement, on-chip decaps still occupy more than 20% of the chip area in modern high performance chips [35].

2.2.5 Electrical Model

The various components of the power distribution network of an IC system can be represented electrically by the simplified model [32] shown in Figure 2.5, with the various components identified in Table 2-I. As can be seen, at low frequencies, the current loop encompasses the entire system. With each progressively higher range of frequencies, the current loops correspondingly become smaller.

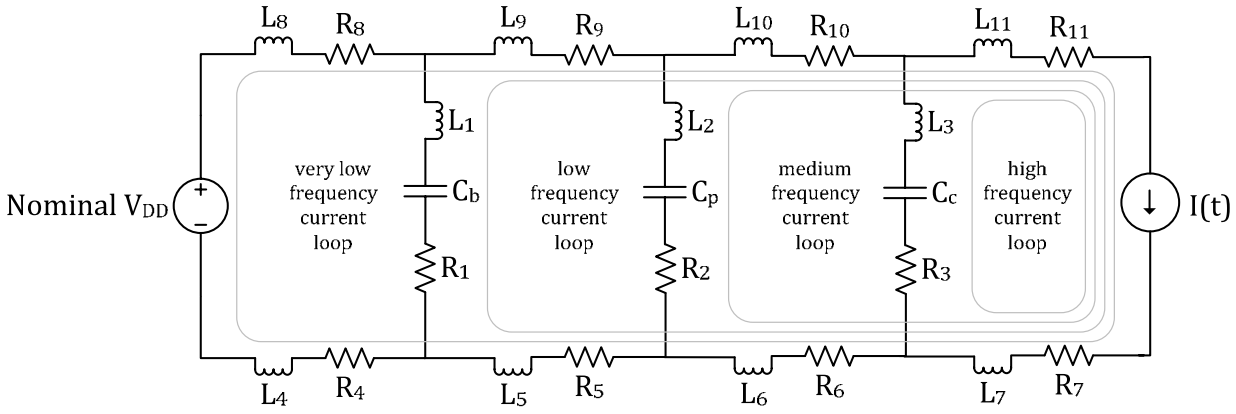


Figure 2.5 Simplified electrical model of the power distribution network of an IC system

TABLE 2-I VARIOUS COMPONENTS IN SIMPLIFIED ELECTRICAL MODEL OF FIGURE 2.5

SYMBOL	DESCRIPTION
C_b	total board-level capacitance including decoupling capacitance
C_p	total package-level capacitance including decoupling capacitance
C_c	total chip-level capacitance including decoupling capacitance
R_1	series resistance of conductors associated with C_b
R_2	series resistance of conductors associated with C_p
R_3	series resistance of conductors associated with C_c
L_1	series inductance of conductors associated with C_b
L_2	series inductance of conductors associated with C_p
L_3	series inductance of conductors associated with C_c
	In the ground path:
R_4	series resistance of regulator + resistance of board conductors upstream† of C_b
R_5	resistance of board conductors downstream† of C_b + resistance of package conductors upstream of C_p
R_6	resistance of package conductors downstream of C_p + resistance of chip conductors upstream of C_c
R_7	resistance of chip conductors downstream of C_c
L_4	series inductance of regulator + inductance of board conductors upstream of C_b
L_5	inductance of board conductors downstream of C_b + inductance of package conductors upstream of C_p
L_6	inductance of package conductors downstream of C_p + inductance of chip conductors upstream of C_c
L_7	inductance of chip conductors downstream of C_c
	In the power path:
R_8	series resistance of regulator + resistance of board conductors upstream of C_b
R_9	resistance of board conductors downstream of C_b + resistance of package conductors upstream of C_p
R_{10}	resistance of package conductors downstream of C_p + resistance of chip conductors upstream of C_c
R_{11}	resistance of chip conductors downstream of C_c
L_8	series inductance of regulator + inductance of board conductors upstream of C_b
L_9	inductance of board conductors downstream of C_b + inductance of package conductors upstream of C_p
L_{10}	inductance of package conductors downstream of C_p + inductance of chip conductors upstream of C_c
L_{11}	inductance of chip conductors downstream of C_c

† with respect to flow of power from supply to IC

Chapter 3

On-Chip Power Supply Noise

"... it will take a lifetime to search the entire solution space for even a moderately complex system"

… Shiyou Zhao and Kaushik Roy (Purdue University)

This chapter considers on-chip power supply noise in more detail and introduces the concept of localized supply noise. It briefly considers on-chip supply noise models and then presents a detailed scaling analysis showing how supply noise is expected to vary with progressing CMOS technology.

3.1 Modeling Limitations

While the lumped model used thus far is useful in designing the overall power distribution network, it cannot be used to accurately describe the behaviour of the entire system. The model is especially inadequate to describe the on-chip power distribution grid since a complex IC typically comprises millions of line segments. Furthermore, the various decoupling capacitances are implemented by a collection of smaller capacitance components both on- and off-chip and the closest capacitors to a circuit with a current surge are the ones that will have the smallest impedance and will respond immediately to

the current needs of the circuit. The presence of this massively large number of components and nodes makes the modeling and analysis of power supply noise effects immensely challenging [40].

3.2 Concept of "Localization"

While many questions are presently not fully answered about the behaviour of the power supply system, a widely accepted property of the on-chip grid is that the droops in supply voltage level occur within localized areas [17][25] as illustrated in Figure 3.1. Some recent work done at Intel [17], where a relatively detailed full-die model of an industrial microprocessor was studied, showed that high frequency noise tends to remain extremely localized on the chip within a radius of a few micrometers. This is attributed to the higher energy involved in larger loops, the large number of vias present on a power rail that provide numerous power paths, and the device, wire and package parasitcs dissipating the high frequency energy. While it has been shown that the influence of a switching block falls off rapidly with the distance from the noise source, the cumulative effect of closely neighboring switching blocks can add up to produce significant noise in certain regions [41]. As a result of these individual and cumulative noise effects, there are localized areas on a chip, sometimes referred to as noise *hot spots*, where the voltage can droop below tolerable levels [42].

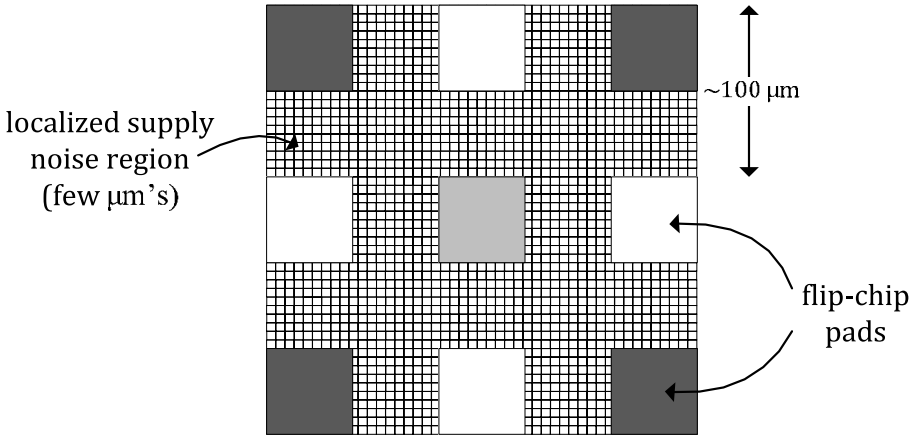


Figure 3.1 Illustration of localized areas of power supply noise within a chip

3.3 A Closer Look at Supply Noise Components

3.3.1 dI/dt Noise

The general definition of inductance based on a closed current loop [32] is highly impractical when applied to circuit elements. This led to the relationship:

$$\Delta V_L = L \cdot \frac{dI(t)}{dt} \tag{1-2}$$

This definition is based on the segmentization of the current loop and the inductance here is technically a *partial* self-inductance of a particular conducting segment [32].

In terms of the on-chip power supply noise attributed to this inductive voltage drop, the dI/dt noise is generally attributed to the package inductance. However, there have been some recent differing opinions on whether the on-chip grid inductance contributes to the noise significantly [17][41][43]. Intel's detailed model [17], showed that on-chip inductance does not in fact contribute significantly to the dip in supply level. Mezhiba [43], models the grid inductance elegantly for the grid-like structure using a relationship for the grid sheet inductance $L_{\square}$ where

$$L_{\square} \propto P \tag{3-1}$$

where P is the pitch of the global metal lines forming the grid in meters. Similar to sheet resistance, the sheet inductance is shown to increase with the length and decrease with the width of the grid.

As described earlier, the two primary IC-to-package bonding technologies are wire-bonding technology and controlled collapse chip connection (C4) flip-chip technology [44]. In wire-bonding technology, the I/O pads are placed around the periphery of a chip, as illustrated in Figure 3.2 (a), and a bond wire is used to connect each pad to a corresponding pad on the package. In flip-chip technology, the I/O pads are connected to the package via solder bumps placed on top of the I/O pads. The pads therefore do not need

to be at the periphery of the chip and are generally placed across the entire area of the die as illustrated in Figure 3.2 (b). The inductance of a typical bond wire is on the order of a few nH and that of a typical flip-chip bond is on the order of a few tens of pH [45]. Given that a larger number of pads can be placed on a chip with flip-chip technology and a significant portion of pads are allocated to the supply rails, and the inductance per bond is low there is great incentive in using this technology which has resulted in it being the predominant bonding mechanism in high performance chips.

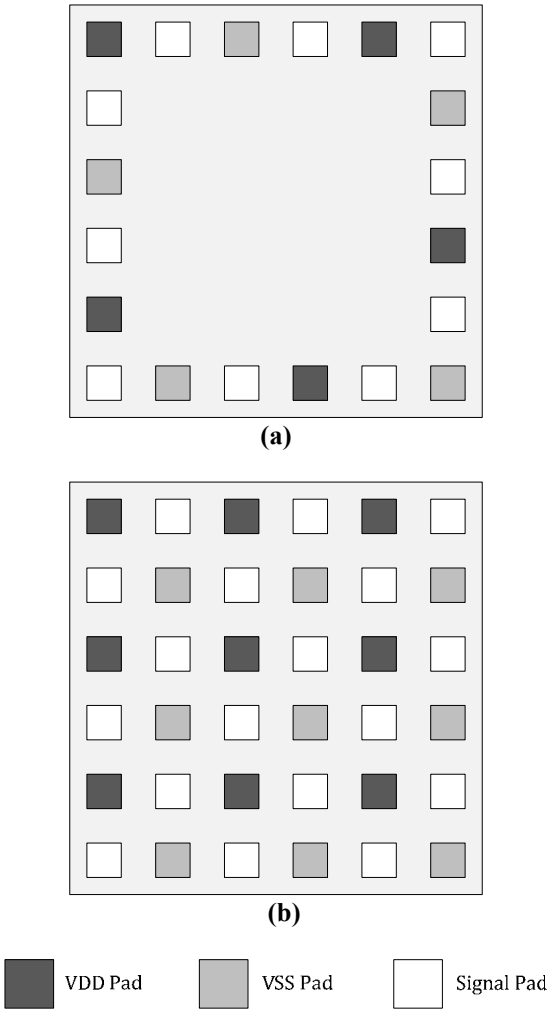


Figure 3.2 Pad layout for (a) wire-bonding, and (b) C4 flip-chip, technologies

3.3.2 IR Noise

As stated earlier, the IR noise

$$\Delta V_R = I(t) \cdot R \tag{1-1}$$

is dominated by the resistance of the on-chip supply grid since the power planes at the board level and package level serve to efficiently minimize their supply series resistance. The top most metal layers in a process are generally thicker than the lower layers and are typically mostly allocated to power supply routing. The resistance of these layers thus dominates the on-chip grid resistance experienced [46]. The use of multiple layers for the on-chip grid and multiple parallel supply lines as shown earlier also serves to reduce this resistance to some degree [46].

3.3.3 Resonance

Resonance is a component of supply noise that is gaining increasing importance in modern chips and is a noise source that must be factored into the voltage regulation specification of a chip. As introduced earlier, the on-chip capacitance, both inherent to the circuits and added decap, combine with the package inductance to form an LC-tank circuit [47][48][49]. When excited by switching at the resonant frequency, this circuit can cause significant oscillations in the supply voltage. When insufficient resistance is present in the system to damp the oscillations, this noise can be persistent over multiple clock cycles and detrimental to the chip. It has been suggested that additional intentional resistance should be added to a system where resonance is a problem [50]. Appropriate sizing of sleep transistors has also been shown to be a means of controlling on-chip resonance [51].

Thus, as opposed to high frequency noise, this component of noise, referred to as mid-frequency noise [39], is a global chip noise. The frequency of this noise source

$$freq_{res} = \frac{1}{\sqrt{L \cdot C}} \tag{3-2}$$

depends on the size of the inductance L and capacitance C in a system and has been shown to range anywhere from a few MHz to a few hundred MHz [17][39][52]. Much of the design effort is expended on dealing with resonance in modern chips[53].

Another phenomenon recently discovered [17] is the presence of resonance within more localized areas of a chip acting as "mini die". These areas, ~1000 μm in radius, exhibit resonance in the 1 to 2 GHz. This is attributed to the possible resistive isolation of capacitive pockets interacting with localized packaging inductors.

3.3.4 Characteristics of Supply Waveforms

Given all the above factors that affect the power distribution network, what does the on-chip power supply look like? Figure 3.3 illustrates the supply waveform of an Intel microprocessor chip [13]. The various resonant frequency components can be seen with their sources identified. Also shown are a positive dI/dt event, which causes a voltage dip, and a negative dI/dt event, which causes a voltage peak.

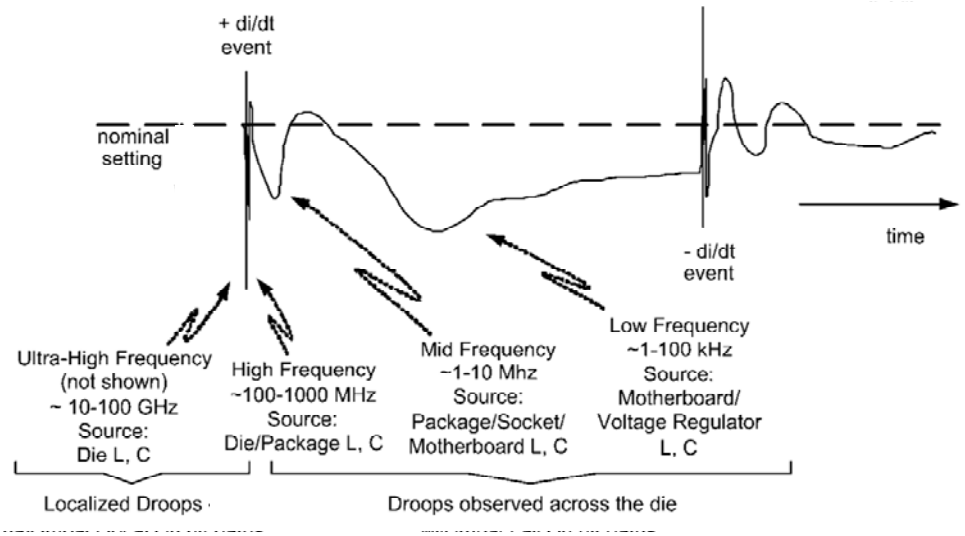


Figure 3.3 Power supply waveform for an Intel microprocessor [13]

The local power supply noise budgets for a chip can vary depending on the application, however, they are typically in the range of 10-15% of the nominal value [36][54]. The total acceptable droop in core power supply magnitude for an Intel i7 microprocessor chip, for example, after implementation of decaps, is 150 mV on a 0.8 V supply with an acceptable overshoot of 50 mV for 25 µs [55]. In other words, for this processor, the supply voltage can droop by approximately 19% and increase by approximately 6% for a 25 µs period while allowing the performance and reliability to remain within acceptable limits.

3.4 Supply Noise Models and Estimation

As mentioned earlier, it is extremely difficult to simulate the entire power supply grid and a large amount of research [17] [21] [36][39][40][53][56][57][58] [59] is conducted into modeling and estimating the magnitude of power supply noise and its spatial and temporal distribution in chip. Nassif *et al.* provide an analytical model [40] in the time domain based on the simplified electrical model for a power supply network as shown in Figure 3.4 assuming a triangular current draw pattern for the current draw *I(t)* as is generally done to model the transistor switching currents. Here L is the grid series inductance, R_g the grid series resistance, C_{eff} the combined effective capacitance of both the on-chip circuitry and any 'intentional' added decoupling capacitance, and R_{eff} the associated effective resistance in series with C_{eff}.

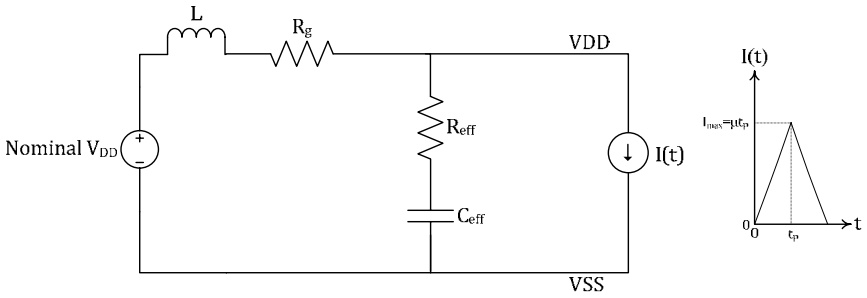


Figure 3.4 Simple model of a power distribution network for estimation of supply noise [40]

Only the over-damped case is considered since if the system is not over-damped the problem becomes one of managing oscillations rather than simply quantifying the deviations experienced by *VDD*. The time domain expression for the maximum droop in *VDD*, ΔV_{max}, is

$$\Delta V_{\max} = \mu\left(L + R_g t_p - C_{eff} R_g{}^2\right) - \left[\left(\left(\exp\frac{-\left(\left((R_g + R_{eff})\cdot C_{eff}\right) + \sqrt{\left((R_g + R_{eff})\cdot C_{eff}\right)^2 - 4LC_{eff}}\right)t}{2C_{eff}L}\right)\right.$$

$$\left.+\left(\exp\frac{-\left(\left((R_g + R_{eff})\cdot C_{eff}\right) - \sqrt{\left((R_g + R_{eff})\cdot C_{eff}\right)^2 - 4LC_{eff}}\right)t}{2C_{eff}L}\right)\right)\cdot\frac{\mu\left(L - C_{eff}R_g{}^2\right)}{2}\Bigg]$$

$$-\left[\left(\left(\exp\frac{-\left(\left((R_g + R_{eff})\cdot C_{eff}\right) + \sqrt{\left((R_g + R_{eff})\cdot C_{eff}\right)^2 - 4LC_{eff}}\right)t}{2C_{eff}L}\right)\right.\right.$$

$$\left.\left.-\left(\exp\frac{-\left(\left((R_g + R_{eff})\cdot C_{eff}\right) - \sqrt{\left((R_g + R_{eff})\cdot C_{eff}\right)^2 - 4LC_{eff}}\right)t}{2C_{eff}L}\right)\right)$$

$$\left.\frac{\mu\cdot C_{eff}}{2\sqrt{\left((R_g + R_{eff})\cdot C_{eff}\right)^2 - 4LC_{eff}}}\left(\left(R_g + R_{eff}\right)\cdot C_{eff}\right)R_g{}^2 - L\left(3R_g - R_{eff}\right)\right] \tag{3-3}$$

Needless to say, even a very simplified electrical model leads to a rather complicated expression for ΔV_{max} when cumulative effects taking place are considered.

Larsson and others [58][61][62][63] found that the triangular waveform overestimated the noise and further considered the effect of ground bounce and velocity saturation on the current drawn by the switching circuitry. However, given the challenges in modeling the supply noise, majority of the methods

[4][17] [21] [36][53][56][57] [59][60] used to estimate the noise involve extracting current patterns and other parameters directly from actual circuitry and building a model around this information.

3.5 Scaling of Power Supply Noise

Given the complexity of existing supply noise models and estimation methods, gaining insight into the effect of scaling on supply noise is challenging. To simplify the analysis, each of the components known to affect scaling is typically considered in isolation. For example, a number of scaling analyses have been conducted on individual noise types, i.e. either dI/dt or IR noise, and for a particular packaging technology, i.e. flip-chip or wire-bonding, and with various assumptions such as fixed die size, constant top metal thickness, etc. Song *et al.* [46], for instance, consider IR noise only, Bakoglu [47] and Larsson [58] only dI/dt noise, and Mezhiba *et al.* [43] only on-chip supply noise scaling with flip-chip technology. These analyses are based on ideal voltage scaling, however, modern scaling no longer follows this trend with respect not only to voltage but other parameters such as switching frequency and pad pitch, as well. Our objective here is to provide a more in-depth assessment of power supply noise scaling laws considering modern scaling trends and practices especially the stagnation of clock frequencies with the advent of multi-core architectures, as well as to provide insight into the effect of scaled decoupling capacitors on the supply noise.

Only on-chip high frequency localized noise is considered here since it is this noise that directly affects the functionality and performance of the corresponding circuits in the vicinity of the noisy supply. It is also this noise constituent that is susceptible to CMOS technology scaling effects, with the assumption that it has the largest magnitude and off-chip capacitances sufficiently minimize the lower frequency components. The simplified model presented in Figure 3.4 is used to represent a ‘localized’ area as described previously. The parasitics in the ground path are neglected as is typically done since the ground

bounce is generally less than the supply noise as a result of the large common substrate that is part of the ground path. Typical values for each of the components in the model are given in Table 3-I.

TABLE 3-I TYPICAL VALUES FOR COMPONENTS IN LOCALIZED NOISE MODEL

COMPONENT	RANGE	ASSUMPTIONS
R_g	0.02 to 2 Ω	Both flip-chip and wire-bonding technology considered, average chip size of modern intel microprocessors used (150 mm^2), top metal sheet resistance of 65 nm used (0.0227 Ω/□), 65 nm pad pitch (130 μm)
R_{eff}	0.1 to 10 Ω	Typical [40]
C_{eff}	C_{non}: 3.5 to 7 pF C_{decap}: 0.5 to 1 pF	32 bit and 64 bit adder assumed , decap ~15% of 'active' area (i.e. excluding diffusions and routing)
I_{max}	55 mA	Based on adder peak current
L	5 to 90 pH	Average chip size of modern Intel microprocessors used (150 mm^2), a quarter of total wire-bonded pads allocated to *VDD*, 65 nm pad pitch (130 μm), L/flip-chip 20 pH, L/wire-bond 4 nH, 4 *VDD* pads per flip-chip localized area
t_p	50 to 200 ps	Clock frequency 500 MHz to 2 GHz

Resonance effects are, again, neglected since this is a global chip effect and, as explained, a consequence of the cumulative capacitance across the chip. The resonance *frequency* is a function of the total chip capacitance and while the chip capacitance scales with technology, the *magnitude* of the resonance noise is determined by the extent of damping present, which is a design parameter used to manage resonance rather than an inherent parameter tied to scaling laws. Thus, the resonance effects must be appropriately dealt with in the design of the power grid or its contributions to the overall noise budgets considered separately to the scaling analysis presented here. Each of the parameters in the localized noise model is individually considered henceforth.

3.5.1 Resistance, R_g

It has been widely accepted that the on-chip resistance dominates the IR supply noise as explained earlier [43]. In line with this observation, the effect of scaling of only the on-chip supply grid is

considered here. Previous work done by Song *et al.* [46] also showed that there is a relatively weak correlation to the number of metal layers used for the supply grid since the top thick global metal layers dominate the resistive grid properties accounting for 99% of the resistive voltage drop. They also determined that metal contact resistance is orders of magnitude smaller than the resistive voltage drop across the metal traces and can safely be neglected. Based on these observations, only the top global metal layer is considered in this analysis.

In the case of flip-chip technology, since power pads cover the entire chip area, each circuit block is assumed to draw power only from its closest surrounding pads [43][64]. Each area within which a group of circuit blocks all draw power from the same pads can be thought of as a power "micro die" as illustrated in Figure 3.5, with the assumption that a given circuit will draw most of its power at least from its immediately surrounding four pads. Clearly, for larger circuit blocks, power would be drawn from a larger group of pads forming a larger power micro die. The resistive effects of the power distribution therefore are not a function of the overall die size with flip-chip technology.

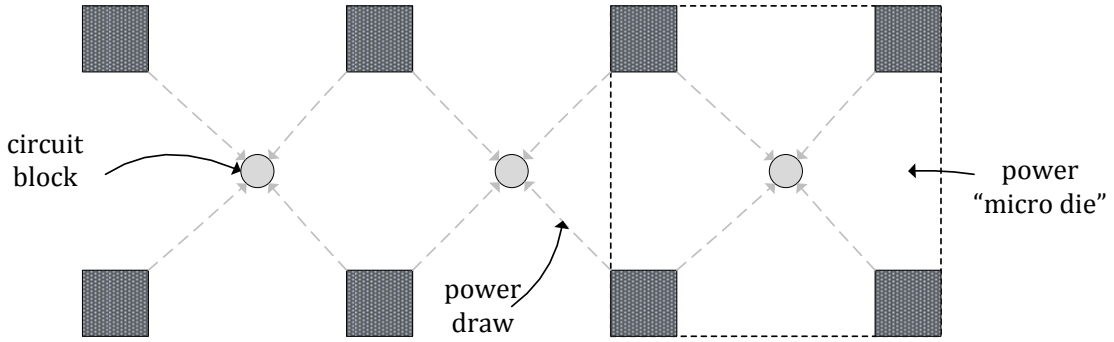


Figure 3.5 Power draw in a local chip area with flip-chip packaging technology

Figure 3.6 (a) further illustrates a group of pads in a flip-chip technology chip with a global supply grid, and Figure 3.6 (b) a scaled down version of the same. Consider the resistive path from the circuit block shown to the closest power pad. The resistance of the path is given by the grid resistance

$$R \propto \frac{\rho}{H} . \frac{L_{grid}}{W_{grid}} \tag{3-4}$$

where ρ is the metal resistivity, L_{grid} the length of the global power grid, W_{grid} its width, and H the thickness of the global metal layer. As can be seen, the *ratio* of L_{grid}/W_{grid} remains constant from one technology to the next, and the grid resistance is a function of only the global metal thickness, assuming the metal resistivity does not change. Due to the stringent constraints on the supply network impedance and interconnect delay considerations, global metal thickness and correspondingly the global line pitch has not always been scaled in proportion to the local line pitch through several preceding technology generations [43]. There are two main types of interconnect scaling [65]. In the first type of scaling, the thickness and minimum line pitch of the top metal, which is used primarily for global power distribution, is kept constant. In the second type of scaling, the top metal thickness and minimum pitch are scaled down in proportion to the local interconnect. While the aspect ratio has remained, and is expected to remain, approximately constant, the thickness of the global metal has followed a somewhat unique scaling trend and the thickness is expected to fall somewhere between the ideally scaled thickness and a constant 2 μm, over the next decade [7]. A unique scaling factor S_{mp} is thus assigned to the global metal pitch. The metal pitch itself does not affect the overall grid dimensions L_{grid} and W_{grid}, however again, is related to the metal thickness. Thus, following from (3-4), the on-chip supply series resistance affecting IR noise, R_g, is expected to scale in proportion to

$$R_g \propto S_{mp} \tag{3-5}$$

This trend is similar to that obtained by Mezhiba *et al.* [43], except here the presence of a unique scaling factor is identified.

In the case of wire-bonding technology where the pads are around the periphery of the chip, the ratio of L_{grid} to W_{grid} is similarly expected to remain the same as the chip dimensions increase, assuming an

equivalent increase in chip size in both directions. The resistance is thus expected to also scale according to (3-5) for wire-bonding technology as well.

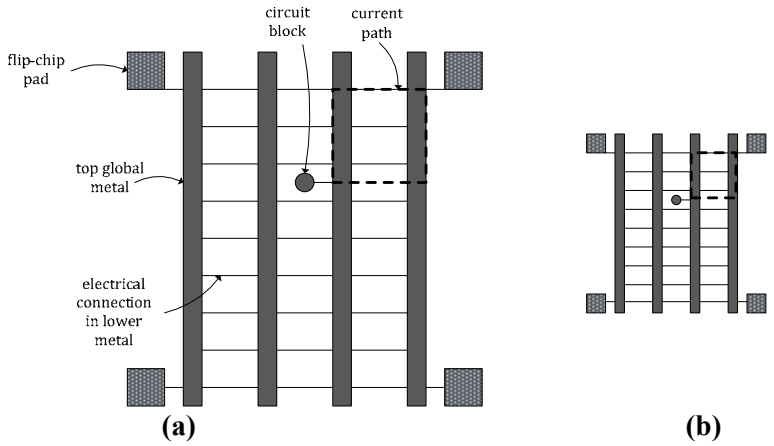


Figure 3.6 (a) Group of pads and global power grid in a flip-technology chip, and (b) its scaled down version

3.5.2 Inductance, L

The scaling of both package inductance L_{pkg} and on-chip inductance L_{chip} is considered here, where the total inductance L in the model represents the sum of the two. According to (3-1) [43], the on-chip grid sheet inductance $L_{\square}$ scales in proportion to the pitch of the global metal layer, thus scaling as

$$L_{\square} \propto \frac{1}{S_{mp}} \tag{3-6}$$

The on-chip effective chip inductance L_{chip}

$$L_{chip} = L_{\square} \frac{L_{grid}}{W_{grid}} \propto \frac{1}{S_{mp}} \tag{3-7}$$

is expected to scale with the same trend as $L_{\square}$ since the ratio of L_{grid}/W_{grid} for both flip-chip and wire-bonded chips is similarly assumed to remain constant.

The effective package inductance L_{pkg} is, again, a function of packaging technology as well as the number of pads available for power connections. Again, considering the power micro die in Figure 3.5 for flip-chip technology, the number of pads a particular circuit draws power from is expected to remain constant with scaling. In practice, since the pad pitch does not scale at the same rate as the minimum feature dimensions there may be a relatively small sub-set of cases where the number of pads the chip draws power from may change. In most cases, however, assuming no significant improvements in flip-chip inductance per bond, the packaging inductance per localized chip area is expected to stay constant with scaling for chips packaged with flip-chip technology.

For wire-bonding technology, on the other hand, all the on-chip circuits collectively draw power from all the power pads on the pad ring. Therefore, as the pad pitch and chip size scale, so does the number of pads and effective packaging inductance. As Bakoglu [47] assumed in his scaling analysis, the pad pitch scaled as 1/S in older technologies, however, modern scaling has been shown to follow a slower trend of pad pitch scaling of $1/\sqrt{S}$ [43]. Therefore, assuming that each chip dimension scales equally as S_C, the number of pads P_N in a wire-bonded chip scales as

$$P_N \propto \sqrt{S}.S_C \tag{3-8}$$

Assuming the proportion of pads allocated to the power supply remains constant, the effective packaging inductance L_{pkg} for a wire-bonded chip is expected to scale as

$$L_{pkg} \propto \frac{1}{\sqrt{S}.S_C} \tag{3-9}$$

3.5.3 Peak Current Consumption, I_{max}

Previous scaling analyses [46][47][43][58] assume ideal scaling [8] in their derivation of the scaling laws. As discussed earlier, modern scaling laws no longer follow ideal scaling laws. Consider how the

transistor saturation current relation scales with general scaling where the voltage is scaled by the factor S_V

$$I_{sat} = \frac{1}{2}\mu \frac{\varepsilon_{ox}}{(t_{ox}/S)} \frac{(W/S)}{(L/S)} \left(\frac{V_{GS} - V_t}{S_V} \right)^2 \propto \frac{S}{{S_V}^2} \tag{3-10}$$

The current I_{max} is assumed to be directly related to the saturation current for a given technology. Determination of the effect of circuit current consumption on the supply noise scaling trend is based on an assumption of the size of the localized noise areas on a chip. As discussed earlier, the actual power grid is a distributed system of inductances, resistances and capacitances, and there are localized current loops present that result in the localization of supply noise to within a few microns [17]. If the size of a localized noise region on a chip is assumed to remain constant with scaling, the current consumption within the given area will increase with scaling according to

$$I_{\max} \propto \frac{S^3}{{S_V}^2} \tag{3-11}$$

However, if the assumption is made that the size of the localized area also scales commensurate with the minimum feature size of the technology, then the current consumption of the localized area will actually decrease in the case of ideal scaling according to

$$I_{\max} \propto \frac{S}{{S_V}^2} \tag{3-12}$$

Since, there clearly has been no observed decrease in supply noise with scaling, it can be assumed that the size of the localized noise region does not in fact scale at the same rate as the minimum features on a chip or the pad pitch. Arledge *et al.* [64] and Mezhiba *et al.* [43] make the assumption that the localized noise region is circular in shape with a radius of half the pad pitch since each pad would service the enclosed area. The current is thus assumed to be related to the pad pitch and thus scale as 1/S. The magnitude of

the pad pitch however is about two orders of magnitude greater than the size of the localized regions observed by Pant *et al.* [17] on-chip, and the area is thus assumed not to be directly related to the pad pitch here. Both Bakolu [47] and Larsson [58] assume the noise area remains constant with scaling. Given the relative size of the observed localized regions, as well as remaining conservative in the analysis and assuming a worst case noise scaling trend, the size of the localized region is similarly assumed to remain constant with scaling. The current per localized area is thus expected to follow the trend in (3-11). Since the size of the localized area is assumed to be unrelated to the pad pitch in a flip-chip packaged chip, it is similarly assumed to be independent of the chip size in the case of a wire-bonded chip, leading to a similar current scaling trend for the latter.

3.5.4 Current Transient Time, t_p

The current transient time is typically designed to be ~10% of the clock frequency. As described earlier, with the move to multi-core processor architectures, clock frequencies have stagnated with scaling and this trend is expected to continue with the number of cores on the rise. In previous supply noise scaling analyses [43][47][58], the current transient time is assumed to scale as 1/S. In this analysis, however, based on the clock frequencies remaining constant, the current transient time t_p is now assumed to remain constant with scaling technologies.

3.5.5 Capacitance, C_{eff}

The capacitance C_{eff} and its associated series resistance R_{eff} also has an effect on the level of supply noise observed. The capacitance C_{eff} represents the effects of both the on-chip non-switching circuitry, C_{non}, as well as any added intentional decoupling capacitance, C_d. Previous scaling analyses do not take into account the effect of the capacitance scaling on the supply noise level. Again the size of the localized area is assumed to remain constant as is the ratio of the switching circuitry, non-switching circuitry and the

decaps as illustrated in Figure 3.7. Assuming the decaps are MOS-based, the overall parallel plate capacitances C_{non} and C_d depend on how the oxide thickness t_{ox} scales with technology according to

$$C_d, C_{non} = \frac{\varepsilon \cdot A}{t_{ox}} \tag{3-13}$$

where ε is the dielectric permittivity and A the capacitor parallel plate area. The thickness t_{ox} also scales as $1/S$ [8][9], therefore C_{eff} is actually expected to increase with scaling. According to the 2011 ITRS [7], t_{ox} does follow this trend except for a slight increase at the transition from one device type to the next (bulk to SOI then double-gate MOSFETs). In addition, a notable disadvantage of the move towards SOI devices is that the capacitance contribution of non-switching circuitry is expected to diminish since the n-well junction capacitance is no longer present in these devices [4]. The trend in supply noise suppression due to C_{eff}, on the overall supply noise, $-\Delta V_C$, is

$$-\Delta V_C \propto S \tag{3-14}$$

where the negative sign is used here to represent noise suppression as opposed to noise generation. Again, a discontinuity in this trend is expected with the move to SOI devices. As will be seen in the next chapter, the magnitude of R_{eff} is relatively small and the effect of scaling on it is thus neglected in this analysis.

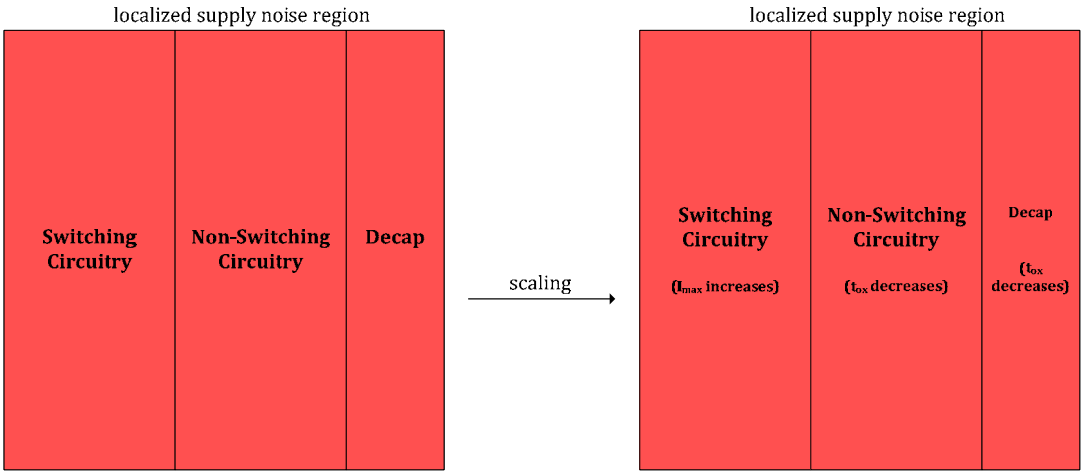


Figure 3.7 Effect of scaling in localized supply noise region

3.5.6 IR and dI/dt Noise

Based on the scaling trends of on-chip resistance and the current draw per localized area, the IR component of the maximum supply noise is expected to scale as

$$\Delta V_R = I_{\max} \cdot Rg \propto \frac{S^3}{S_V{}^2} S_{mp} \tag{3-15}$$

for both flip-chip packaged and wire-bonded chips, where any leakage components are neglected. The dI/dt noise component in flip-chip packaged chips is expected to scale as

$$\Delta V_L = L_{chip} \frac{I_{\max}}{t_p} \propto \frac{1}{S_{mp}} \frac{S^3}{S_V{}^2} \tag{3-16}$$

For wire-bonded chips, the scaling of the package inductance is expected to dominate over the effect of the scaled on-chip inductance [17], and the noise is thus expected to scale as

$$\Delta V_L = L_{pkg} \frac{I_{\max}}{t_p} \propto \frac{1}{\sqrt{S} S_C} \frac{S^3}{S_V{}^2} = \frac{S^{5/2}}{S_C S_V{}^2} \tag{3-17}$$

Again, t_p is not expected to scale with progressing technologies. The signal-to-noise ratio for the IR noise component SNR_R is expected to scale as

$$SNR_R = \frac{VDD}{\Delta V_R} \propto \frac{S_V}{S^3 S_{mp}} \tag{3-18}$$

The dI/dt component signal-to-noise ratio SNR_L in a flip-chip packaged chip is expected to scale as

$$SNR_L = \frac{VDD}{\Delta V_L} \propto \frac{S_{mp} S_V}{S^3} \tag{3-19}$$

and for the wire-bonded case to scale as

$$SNR_L = \frac{VDD}{\Delta V_L} \propto \frac{S_C S_V}{S^{5/2}} \tag{3-20}$$

In order to gain an appreciation for the scaling trends for varying trends in supply voltage and top metal thickness scaling, the best and worst case scaling scenarios for each noise type are determined. For example, in the case of IR noise, the best possible scaling scenario of keeping the top metal thickness constant (S_{mp}=1) and scaling the supply voltage by the same factor as the physical dimensions (S_V=S), the IR noise SNR would still degrade in proportion to $1/S^2$. In the worst case where the top metal thickness is scaled (S_{mp}=S) and the supply voltage held constant (S_V=1), the IR noise SNR would scale in proportion to $1/S^4$. In practise, the scaling trend is likely to lie somewhere in between and closer to $1/S^3$ since the top metal thickness is likely to stay approximately constant and the voltage scaling factor likely to be closer to 1 than to S. Table 3-II summarizes the best and worst case scaling trends for each noise component.

TABLE 3-II BEST AND WORST CASE SCALING TRENDS

NOISE COMPONENT	General	Best Case	Worst Case
		$S_{mp}=1,\ S_V=S$	$S_{mp}=S, S_V=1$
ΔV_R	$\frac{S^3}{S_V^{\,2}} S_{mp}$	S	S^4
SNR_R	$\frac{S_V}{S^3 S_{mp}}$	$\frac{1}{S^2}$	$\frac{1}{S^4}$
		$S_{mp}=S, S_V=S$	$S_{mp}=1, S_V=1$
ΔV_L (flip-chip)	$\frac{1}{S_{mp}} \frac{S^3}{S_V^{\,2}}$	constant	S^4
SNR_L (flip-chip)	$\frac{S_{mp} S_V}{S^3}$	$\frac{1}{S}$	$\frac{1}{S^3}$
		$S_V=S$	$S_V=1$
ΔV_L (wire-bond)	$\frac{S^{5/2}}{S_C S_V^{\,2}}$	$\frac{\sqrt{S}}{S_C}$	$\frac{S^{3/2}}{S_C}$
SNR_L (wire-bond)	$\frac{S_C S_V}{S^{5/2}}$	$\frac{S_C}{S^{3/2}}$	$\frac{S_C}{S^{5/2}}$

3.5.7 Conclusions

Table 3-III summarizes the various scaling trends and also includes trends from previous analyses for comparison. This analysis shows that the IR supply noise component ΔV_R is expected to degrade with scaling, even with the optimistic ideal scaling where S_V=S and keeping the top metal thickness constant (S_{mp}=1) which is also in agreement with the analysis of Song *et al.* [46]. Identification of the unique scaling factors S_V and S_{mp}, in this analysis, however, enable the scaling trend to be determined for given voltage scaling and top metal scaling situations. It should be kept in mind, that the area of the localized area was assumed to remain constant in the analysis presented which may lead to a possible overestimation of the noise with scaling, however, given the relative size of the localized area, this is unlikely. Bakoglu [47] neglects the concept of localized areas and assumes the IR noise is a function of the total chip current that scales with the switching transient (by a factor of S), which leads to a significant overestimation of the noise with scaling. In Bakoglu's flip-chip analysis, the assumption of decreasing resistance with an increase in number of pads leads to an underestimation of the noise with scaling in this case since, as described earlier, the on-chip resistance dominates the IR supply noise component. Furthermore, the ITRS does not predict any increases in global metal thickness over the next decade, the assumption of which again, underestimates the IR supply noise. Mezhiba *et al.*'s [43] assumption of a scaled localized area in relation to the pad pitch further leads to a possible underestimation of the IR noise with scaling.

With respect to the dI/dt supply noise component, in a flip-chip packaged chip, the analysis presented shows that the on-chip inductive component of the noise can be maintained with scaling if the top metal pitch and voltage are aggressively scaled. The metal pitch thus has conflicting effects; keeping it constant helps reduce the IR noise, however, has the reverse effect on dI/dt noise. Given that the voltage is not expected to scale ideally in modern technologies, and the global metal thickness not likely to scale ideally

as well, an increase in the dI/dt component of noise can be expected to increase for both types of chip packaging. For wire-bonded chips, as expected, increasing the chip size can help alleviate the associated dI/dt component of noise. Again here, Bakoglu's [47] and correspondingly Larsson's [58] assumptions in estimating the current draw, lead to an overestimate of the dI/dt noise with scaling for the case of wire-bond packaging. Bakoglu's further assumption that L_{pkg} scales with the chip size as S_C further overestimates the noise, however, Larsson [58] adjusts for this but assumes the pad pitch scales as *1/S* which is no longer applicable to modern technologies. Mezhiba *et al.*'s [43] dI/dt noise trends are the same as those of this analysis for the specific case of ideal scaling, however this is due to the assumptions of a decreasing localized noise area and increasing clock frequencies which negate each other.

In summary, in previous analyses, ideal scaling has been assumed and the effect of stagnating clock frequencies not considered. The concept of the localized area has not always been appropriately considered. The effect of scaling on the suppression of noise by decoupling and non-switching circuitry capacitance has further been neglected, which is expected to scale as S. The power supply noise analysis presented here takes into consideration the more modern device and packaging scaling trends as well as the effect of on-chip capacitance scaling on supply noise.

It should be further noted that since each noise component is considered individually, the scaling trends do not reflect the effect of the relative magnitude of each noise component, but rather a trend for each component. Previous work [40] has shown, however, that the overall supply noise has a strong dependence on R_g followed by C_{eff}. The scaling trends of ΔV_R and $-\Delta V_C$ are thus important components in the scaling of the overall noise. Again, in the most optimistic case of ideal scaling and constant global metal thickness, ΔV_R is expected to increase as S. According to the 2011 ITRS, however, S_V is expected to fall between 1 and S, thus the rate of increase of ΔV_R with scaling is expected to be slightly higher than *S*. Fortunately, the expected increase in C_{ox} results in $-\Delta V_C$ also scaling as S. In other words, the increase in

ΔV_R is partially offset by an increase in $-\Delta V_C$. The effect of an increase in ΔV_L is neglected here since it is assumed that the effect of changes in the magnitude of the on-chip inductance on the overall supply noise is comparatively small. In conclusion, an increase in supply noise is expected with scaling, however, it is expected to scale by a factor smaller than S.

TABLE 3-III SUMMARY OF SUPPLY NOISE SCALING FACTORS

PARA-METER	SCALING RELATIONSHIP		ASSUMPTIONS/COMMENTS
	FLIP-CHIP	WIRE-BONDING	
R_g	$R_g \propto S_{mp}$		Only on-chip single global metal layer is important Size of localized area remains constant with scaling
L	$L_{chip} \propto \frac{1}{S_{mp}}$		S_{mp} is a unique scaling factor for the global metal pitch
	$L_{pkg} = \text{constant}$	$L_{pkg} \propto \frac{1}{\sqrt{S}S_C}$	For flip-chip, circuits draw power from the closest power pads Power wire-bonding, pitch scales as $1/\sqrt{S}$
I_{max}	$I_{max} \propto \frac{S^3}{S_V^2}$		General scaling assumed with unique scaling factor S_V for supply voltage Size of localized area remains constant
t_p	$t_p = \text{constant}$		Frequency will no longer scale
$-\Delta V_C$	$-\Delta V_C \propto S$		Assumes relative size of switching circuitry, non-switching circuitry and decap remains constant
ΔV_R, SNR_R	$\Delta V_R \propto \frac{S^3}{S_V^2} S_{mp}$, $SNR_R \propto \frac{S_V}{S^3 S_{mp}}$		Only on-chip single global metal layer considered
	$\Delta V_R \propto S$, $SNR_R \propto \frac{1}{S^2}$		Song *et al.* [46] Ideal scaling Noise area remains constant with scaling (full chip) Constant top metal thickness
	$\Delta V_R \propto \frac{1}{S}$, $SNR_R = \text{constant}$ (increased global metal thickness)	$\Delta V_R \propto S^3 S_C^2$, $SNR_R \propto \frac{1}{S^4 S_C^2}$ (scaled global metal)	Bakoglu [47] Ideal scaling Localized noise area remains constant with scaling Current transient scales as 1/S Assumes transistor max current drive not used; I_{max} a function of total chip capacitance Sheet resistance scales as S for scaled global metal For increased top metal scenario, assumes resistance drop scales with number of pads and power planes
	$\Delta V_R = \text{constant}$, $SNR_R \propto \frac{1}{S}$ (unscaled global metal) $\Delta V_R \propto S$, $SNR_R \propto \frac{1}{S^2}$ (scaled global metal)	-	Mezhiba *et al.* [43] Ideal scaling Localized area is half-pitch in radius and scales as 1/S
ΔV_L, SNR_L	$\Delta V_L \propto \frac{1}{S_{mp}} \frac{S^3}{S_V^2}$, $SNR_L \propto \frac{S_{mp} S_V}{S^3}$	$\Delta V_L \propto \frac{S^{5/2}}{S_C S_V^2}$, $SNR_L \propto \frac{S_C S_V}{S^{5/2}}$	
	$\Delta V_L = \text{constant}$, $SNR_L \propto \frac{1}{S}$	$\Delta V_L \propto S^3 S_C^3$, $SNR_L \propto \frac{1}{S^4 S_C^3}$	Bakoglu [47] For wire-bonding assumes L_{pkg} scales as S_C. For flip-chipinductance per connection scales as $1/S^2$
	-	$\Delta V_L \propto S S_C$, $SNR_L \propto \frac{1}{S^2 S_C}$	Larsson [58] Same as Bakoglu, except L_{pkg} scales with the number of pads as $1/SS_C$
	$\Delta V_L \propto S$, $SNR_L \propto \frac{1}{S^2}$ (unscaled global metal) $\Delta V_L = \text{constant}$, $SNR_L \propto \frac{1}{S}$ (scaled global metal)	-	Mezhiba *et al.* [43] Ideal scaling Localized area is half-pitch in radius and scales as 1/S Current transient scales as 1/S

Chapter 4

Supply Noise Management

"Do not be satisfied with the stories that come before you. Unfold your own myth."

…Rumi

This chapter considers the topic of supply noise management. A survey of various techniques in the literature is first provided. Various implementations of passive decoupling capacitors are then analyzed in detail including the effect of scaling on their performance.

4.1 Supply Noise Management Techniques

Power supply noise is generally managed through a combination of reduction and isolation techniques which can be grouped into three broad categories [11]. The first involves using circuit techniques that result in less noise being generated. The second entails designing circuits such that their noise immunity is improved. The third and last method suppresses the noise with techniques external to the switching circuits. Figure 4.1 summarizes the various techniques of managing supply noise, each of which is subsequently described.

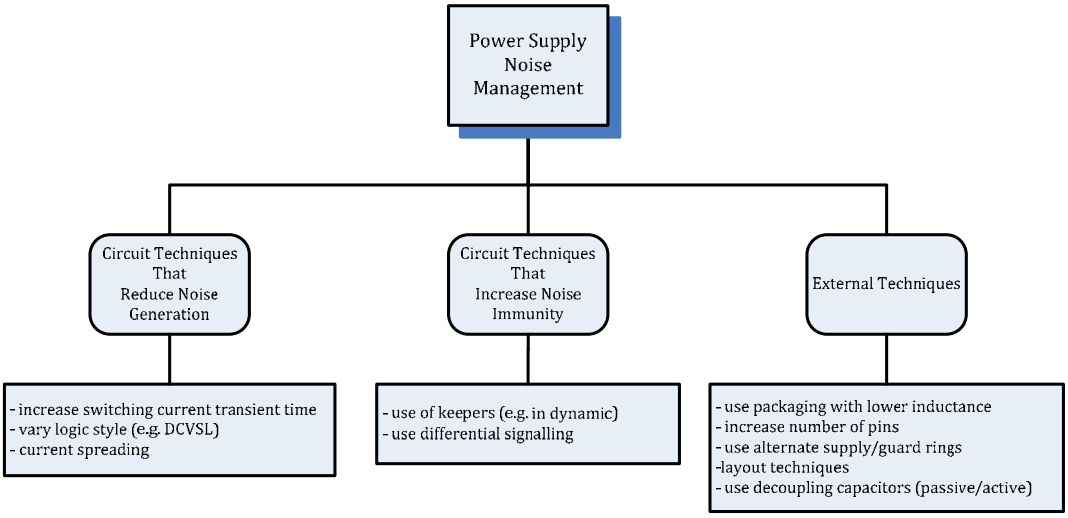


Figure 4.1 Methods of managing supply noise

One circuit technique of reducing noise generation is to increase the current transient time. As discussed earlier, the trend towards multi-core processors is favorable with respect to this objective. In addition, using a logic style such as differential cascade voltage switch logic (DCVSL) can reduce the generation of *dI/dt* noise, since the current draw of this type of logic stays approximately constant during switching. Another technique that has been used is current spreading [66]-[68]. This technique, described later in this chapter, essentially spreads out the switching activity over a longer time period and thus decreases the instantaneous current demands of a circuit.

The immunity of circuits to supply noise can also be increased [11]. For example, the use of a weak keeper in dynamic logic can be used to significantly reduce the effects of noise on logic failures by providing weak feedback to critical nodes. In addition, differential signaling can be used to make noise appear as common mode noise.

In cases where the circuit and noise immunity techniques are not sufficient or practical, other techniques must be used that are external to the switching circuitry. One method of minimizing inductance in the supply path is to use packaging techniques with low inductance. For example, use of C4

flip-chip technology has significantly lowered the package inductance compared to wire-bonding technology. Reducing the inductance of the supply path can reduce both *dI/dt* noise as well as suppress the resonance effects in a system. Another effective method of reducing inductance is increasing the number of pins allocated to the supply. This however, has physical limitations.

Alternately separate supplies can also be used for critical circuitry. Analog and digital domains are typically designed with independent supplies so that the analog circuitry is not affected by the noisy digital supplies. This is often combined with the use of on-chip guard rings which surround sensitive circuits and shield them from noise sources. However, guard rings are found to be ineffective in most modern CMOS technologies due to the low resistive substrates used [69].

A situation where the physical realization of components can have an impact in noise management is in the selection of component sizes. For example, when implementing a decap, whether on- or off-chip, placing several small capacitors in parallel versus a single large device can serve to reduce the effecting parasitic resistance and inductance in series with the capacitors.

The implementation of decoupling capacitors to manage power supply noise is, however, the most common non-circuit noise mitigation technique used. Traditionally, passive decaps have been used, however, recently there have been several active counterparts that have appeared in the literature [70]-[73].

4.1.1 Passive Decoupling Capacitors

In most chips, the intrinsic chip capacitance provides insufficient noise suppression, and passive decaps have been used on- and off-chip for over 40 years, to provide additional suppression. The capacitor is placed close to the noise generating circuit, and acts as a local reservoir of charge such that when current demands of the noise generating circuit are high, the capacitor, by sharing its charge with the noise

generating circuit, provides a large portion of the initial current demand. Decaps area also placed close to sensitive circuitry and similarly provide charge when dips occur in the supply voltage.

Thin oxide gate capacitors are often used to implement passive decaps since the thin gate dielectric layer provides a relatively large capacitance per unit area, although other types of capacitors may be used, as will be seen later in this chapter. An optional fuse or control gate is sometimes inserted in series with decaps in some designs in order to allow the decoupling capacitor to be disconnected from the rest of the circuits in cases where process defects result in short circuits [4].

Modern high performance digital circuit designs include significant amounts of decoupling capacitance to ensure the supply noise remains within specific tolerable limits. Table 4-I indicates the magnitude of the on-chip decoupling capacitance for various commercial ICs. Approximately 20% of the area of such chips can be dedicated toward suppression of the power supply noise to within tolerable limits [9][34].

TABLE 4-I DECOUPLING CAPACITANCE IN MODERN CHIPS

IC	NO. OF TRANSISTORS	CLOCK FREQUENCY	ON_CHIP DECAP
Alpha 21264 Processor [34]	15.2 million	600 MHz	320 nF
IBM S/390 [74]	7.8 million	411 MHz	102 nF
IBM zSeries 900 [75]	-	1 GHz	246 nF
Intel IA-64 [76]	25.4 million	800 MHz	800 nF
SUN ULTRA SPARC III [77]	23 million	1 GHz	176 nF

4.1.2 Active Decoupling Capacitors

Active decoupling capacitors essentially comprise of circuitry that is triggered in the presence of power supply noise, and causes a decrease in the maximum deviation in the supply voltage from its nominal value. The various active supply noise suppression techniques are summarized in Appendix A. A challenge with active on-chip noise mitigation techniques is that the noise event being suppressed must be

predicted prior to its occurrence, thus enabling the noise to be suppressed at the instance that it occurs. In addition, the power supply of the mitigation circuitry itself is exposed to the same noise event that is being mitigated and thus additional quiet supplies are often required. Furthermore, since dips in supply voltage occur at a time when the power draw from the supply is high, it can be challenging to use the same supply to provide additional power for mitigating the dip in supply voltage. The various active decap techniques present in the literature are further described below.

4.1.2.1 Switched Capacitor Based Noise Mitigation

In the switched capacitor based mitigation technique [78]-[82], first proposed by Ang *et al.* [70], two or more capacitors are placed in parallel and switched to a series configuration in order to provide a boost in voltage that correspond to dips in the supply voltage. In the parallel configuration, each capacitor charges to the supply voltage. A sensing circuit then determines whether the supply voltage has crossed a certain threshold and, if so, triggers circuitry to switch the capacitors from the parallel configuration to a series configuration. Figure 4.2 (a) and (b) illustrate the parallel and series configurations for the case of two capacitors, respectively, and Figure 4.2 (c) illustrates the implementation of the switched capacitor circuit. In the series configuration, the voltage across the equivalent capacitor doubles causing a spike in the local supply voltage, thus providing charge that dampens the dip in voltage.

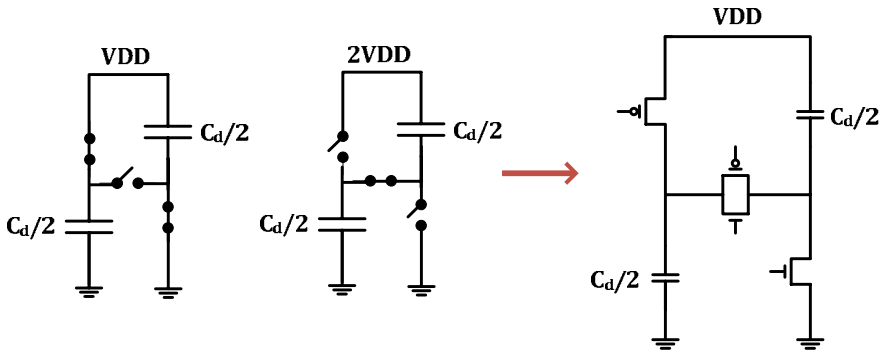


Figure 4.2 Switched capacitor noise suppression technique (a) with capacitors in parallel, (b) with capacitors in series, and (c) circuit implementation

In the parallel configuration, the total charge Q_d on each capacitor C_d in Figure 4.2 is:

$$Q_d = C_d VDD \tag{4-1}$$

When switched to the series configuration, the voltage across the equivalent capacitance becomes:

$$VDD_{switch} = \frac{C_d VDD}{C_{dd} / 2} = 2VDD \tag{4-2}$$

where VDD_{switch} is the voltage across the capacitors when they are in series. The theoretical voltage across the capacitors in the series configuration is twice the voltage across the capacitors in the parallel configuration. In practice the peak in voltage is less than the theoretical voltage due to variations in the supply voltage that the capacitors are charged to in the parallel configuration, switch parasitics, and leakage currents. When the capacitors are switched to the series configuration, the decap impedance is increased due to the presence of the switch.

This noise suppression technique is aimed at suppressing dips in the power supply, in particular. Although peaks can cause a reliability issue, dips are often more immediate concerns since they can cause significant decreases in production yields. A challenge, however, with this mitigation technique is determining the triggering instant of the circuit.

4.1.2.2 Opamp Based Noise Mitigation

Another technique exploits the Miller effect [54][71] for the mitigation of substrate and/or supply noise. The decoupling capacitance is placed in the feedback loop of the opamp and the Miller effect causes this capacitance to appear as a larger capacitance between the supply and ground lines of the chip, as illustrated in Figure 4.3. This larger apparent capacitor between the supply and ground terminals serves to shunt more supply noise compared to a traditional decoupling capacitor. The magnitude of the apparent capacitance is approximately equal to the gain of the opamp multiplied by the feedback capacitance and

this technique can therefore provide significant improvement over the traditional decoupling capacitor technique.

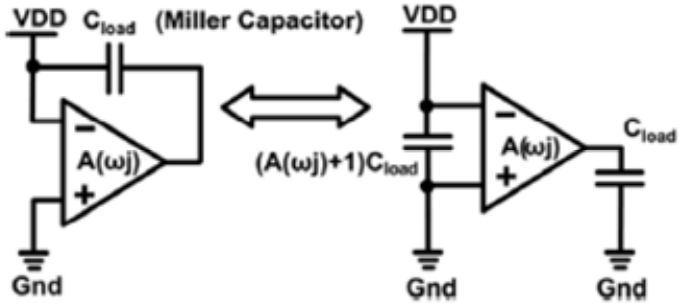


Figure 4.3 Opamp-based noise suppression circuit [54]

4.1.2.3 Feedthrough Based Noise Suppression Technique

This technique uses the concept of feedthrough to produce a boost in the supply voltage by switching the ground terminal of the decap [83], as illustrated in Figure 4.4. A drawback of this approach is the large clock power required to drive the large capacitance.

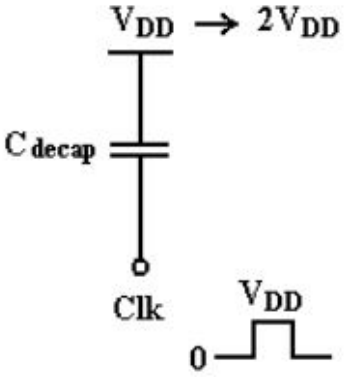


Figure 4.4 Feedthrough based noise suppression technique [83]

4.1.2.4 Other Noise Suppression Techniques

Another noise suppression technique is based on the use of an on-chip transformer to sense changes in current due to the noise [72] as illustrated in Figure 4.5. Based on the sensed current, anti-phase current is generated and injected into the substrate to dampen the noise signal. At 300 MHz, this technique showed an improvement in noise levels of approximately 9% compared to the use of a guard ring.

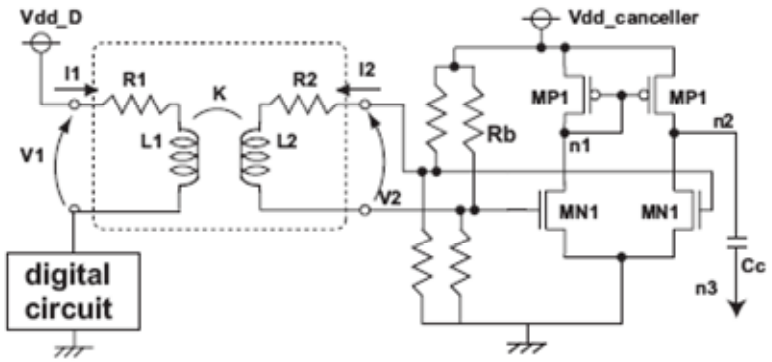


Figure 4.5 Transformer based noise mitigation technique [72]

Yet another technique [73], also aimed at mitigating substrate noise, uses a band-pass filter to first couple the noise to an inverter amplifier which then feeds a comparator followed by a current generator. The current generator similarly produces an anti-phase current and injects this current into the substrate to dampen the noise signal. The major challenge with both the latter techniques described is that they are not easily adaptable to mitigation of noise on the power supply since this would require an external supply with a higher voltage than the nominal supply voltage.

4.2 A Further Look: Analysis and Design of Passive Decaps

As described earlier, the most common technique to suppress power supply noise is to place a relatively large decoupling capacitor, or decap, between the supply rails. On-chip decaps provide the initial current demand of on-chip switching circuitry that cannot be met immediately by the power distribution system. There are several methods of implementing decaps on-chip and each has its limitations. The optimal choice of decap thus depends on circuit constraints and process technology. Further complicating this decision is the availability of increased options within the technology, such as, multi-threshold devices and an increase in available metal layers which provide more ways in which the decaps can be implemented. Furthermore, parasitic resistances associated with the decaps are technology and frequency specific and play a role in determining the overall impedance of the decap. Therefore, the optimal choice of decap implementation is not always evident for a given set of constraints and technology. The objective of this work is to characterize the various decap implementations such that the most favourable decap implementation is apparent for given chip design constraints. This comparison is undertaken using post-layout simulations in a 65 nm CMOS technology. Hybrid decap implementations are further investigated which are shown to provide increased capacitance with no additional cost in area. The effect of scaling is additionally considered on the most optimal structures, and chip measurement results provided for a 90 nm process.

4.2.1 General Decap Model

Figure 4.6 illustrates a general decap model for on-chip decoupling capacitors. The values of the components shown determine the overall decap impedance (or admittance), which is directly related to

the level of noise suppression seen on the supply. The overall decap effective capacitance, C_{eff}, comprises of the area capacitance of the decap, C, a fringe capacitive component, C_{fringe}, and a coupling capacitive component, $C_{coupling}$, as illustrated in Figure 4.7 for a parallel plate capacitor. The capacitance C_{fringe} is present between the sidewalls of the decap and an alternate terminal in any adjacent layers, and $C_{coupling}$ is present between adjacent terminals in the same layer.

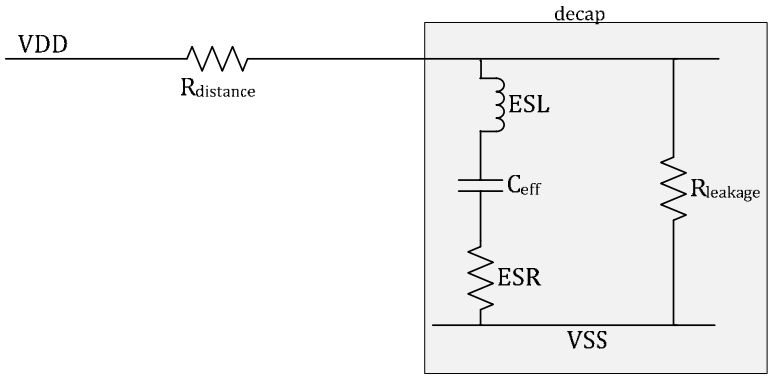


Figure 4.6 General decap model

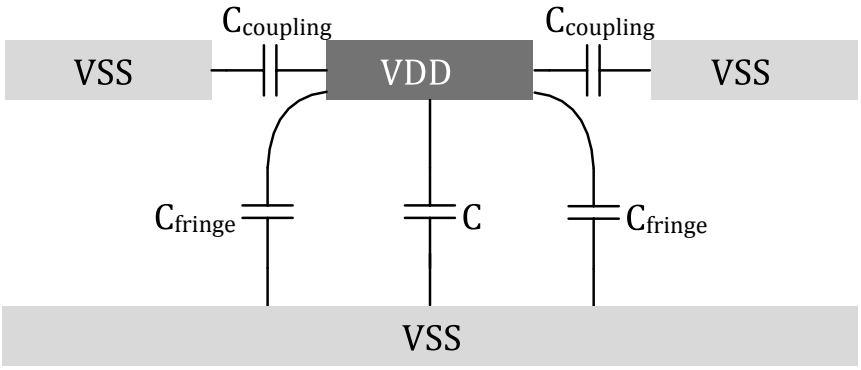


Figure 4.7 Capacitive components of a parallel plate decap

As described previously, the amount of decoupling capacitance on a typical chip can occupy more than 20% of the total chip area [34]. The parallel plate capacitance C is given by

$$C = \frac{\varepsilon \cdot A}{t} \tag{4-3}$$

where ε is the permittivity of the dielectric, A the cross-sectional area of the capacitor and t the dielectric thickness. Therefore, in order to maximize capacitance per unit area of a decap, a thin dielectric (and/or large permittivity) is desirable. A shortcoming of using a thin dielectric, however, is the resulting tunneling leakage current. Decap leakage typically contributes 10-20% of the overall power budget of a chip [54]. Not only does leakage affect the chip's power consumption, but it also reduces the effectiveness of the decap [84] due to the loss of charge. The dielectric leakage effect is represented by means of resistor, $R_{leakage}$ between the supply and ground terminals.

Another parameter of interest in the decap model is the equivalent series resistance (*ESR*). All capacitors exhibit a finite amount of *ESR* that can vary with frequency. The *ESR* is not a physical resistor but rather an "equivalent" resistance that results from the conducting electrodes as well as the insulating dielectric. For the purposes of modeling, the *ESR* of a capacitor is typically represented as a single parasitic element in series with the capacitor. The presence of the *ESR* undesirably increases the impedance between the decap terminals. However, the *ESR* can also play a role in damping oscillations that can result from the *LC* tank formed by the capacitance and parasitic inductances along the power grid [85]. Therefore, it should be kept in mind that where oscillations are present the *ESR* can dissipate energy thus potentially reducing the overall supply noise.

The *ESR* can further have electrostatic discharge (ESD) implications [86]. In modern CMOS technologies the oxide is relatively thin and thus more prone to breakdown as a result of an ESD event. A simple protection scheme is to insert a resistance in series with the capacitor to limit the voltage seen across the dielectric layer [86]. Decaps with an inherent large *ESR* can provide some level of ESD protection, without the need for additional area to implement added resistance.

The quality factor, Q, of a capacitor is often used as a means of quantifying its energy losses in a capacitor and is related to the *ESR* according to the equation

$$Q = \frac{X_C}{ESR} \tag{4-4}$$

where X_c is the reactance in ohms and is given by

$$X_C = \frac{1}{2\pi \cdot f \cdot C} \tag{4-5}$$

where f is the frequency of an AC sinusoidal test signal applied to the capacitor. Normally, a high Q is desirable for minimum decap impedance, however, a low Q value can also be desirable given oscillation damping and ESD considerations.

Like *ESR*, there can also be an equivalent series inductance (*ESL*) associated with decaps. The presence of this additional inductance can further increase the impedance of the decap. Fortunately, the *ESL* of on-chip decaps is small and its effects are typically negligible.

Lastly, the resistance $R_{distance}$, represents the distance the decap is placed away from the supply node. This resistance is important in determining the effectiveness of the decap since it affects the overall impedance provided by the decap structure.

4.2.1 Survey of Decap Types

On-chip decaps can be realized in a variety of ways depending on which process layers and/or devices are used to implement the capacitance. In the analysis that follows, decaps achievable in a standard 65 nm CMOS process are considered. These decaps are essentially MOS-based decaps as well as decaps implemented using the various available metal layers. While MIM capacitors are a non-standard process option, they are frequently offered in various processes and are included in the analysis. Other capacitive structures achievable in non-standard CMOS technologies and not considered in the analysis include poly-insulator-poly (PIP) capacitors and deep trench capacitors (DTCs).

4.2.1.1 MOS Decaps

In a standard CMOS 65 nm technology, the gate oxide layer provides the thinnest dielectric layer and MOS-based decaps are the most area efficient and most commonly implemented structures [84]. One drawback of these decaps, however, is the leakage current increases exponentially with decreasing oxide thickness [87].

NMOS, PMOS and CMOS Decaps

MOS decaps can be implemented as NMOS decaps, PMOS decaps, or a combination of these as CMOS decaps, as illustrated in Figure 4.8. CMOS decaps are commonly used within standard cells since a portion of the area within these cells is typically reserved for PMOS transistors and a portion for NMOS transistors.

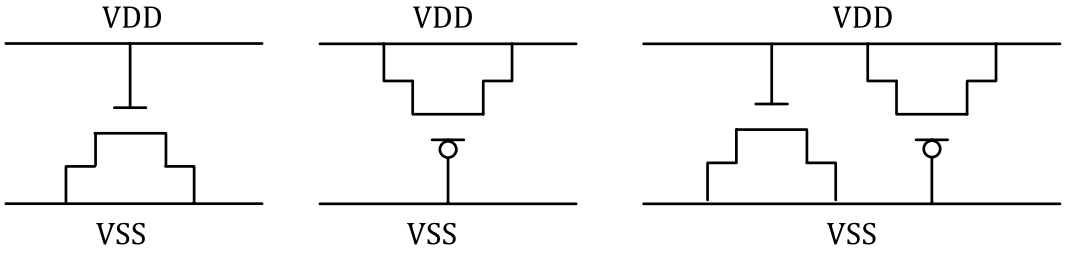
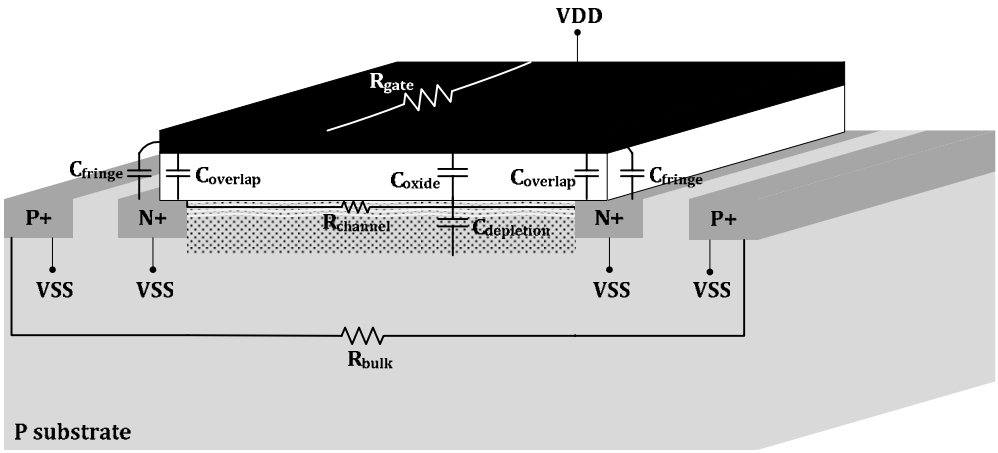
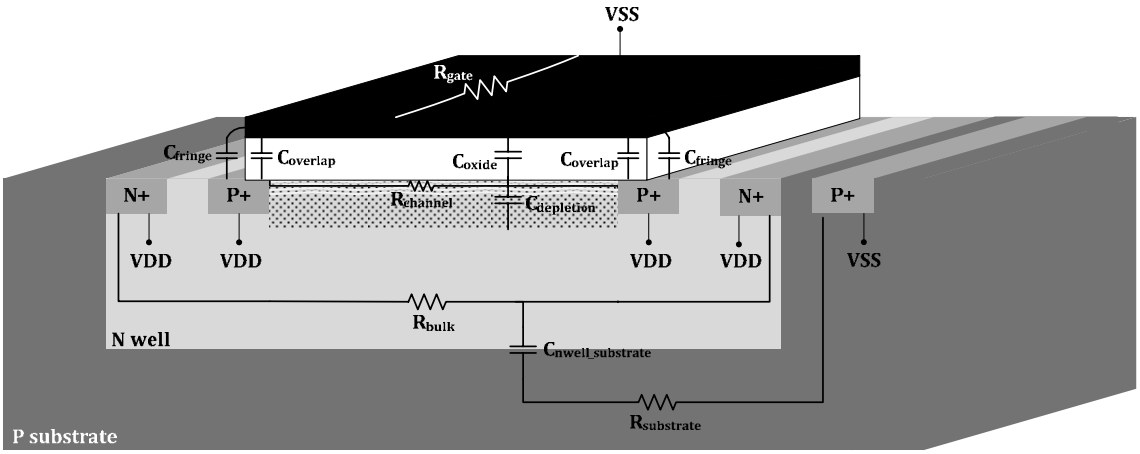


Figure 4.8 Configurations for (a) an NMOS decap, (b) a PMOS decap, and (c) a CMOS decap

Figure 4.9 (a) and (b) illustrate the physical structure of NMOS and PMOS decaps, respectively, with their electrical models overlaid.



(a)



(b)

Figure 4.9 (a) NMOS, and (b) PMOS decap with electrical model overlaid

For an NMOS decap, the *ESR* represents the various resistances as

$$ESR_{NMOS} = R_{gate} + \frac{R_{channel} \cdot R_{bulk}}{R_{channel} + R_{bulk}} + R_{line} + R_{contact} \tag{4-6}$$

where, R_{gate}, $R_{channel}$, R_{bulk}, are the resistance of the gate, channel and bulk, respectively, R_{line} is the combined resistance of any conductors used for routing, and $R_{contact}$ is the contact resistance resulting from any metal semiconductor contacts. The effective capacitance of an NMOS decap, C_{MNOS}, is given by

$$C_{NMOS} = \frac{C_{oxide} \cdot C_{depletion}}{C_{oxide} + C_{depletion}} + C_{overlap} + C_{fringe} \tag{4-7}$$

where, C_{oxide} is the oxide capacitance, $C_{depletion}$ is the channel depletion region capacitance, $C_{overlap}$ is the total capacitance resulting from the overlap of the gate with the source and drain diffusions, and as defined earlier, C_{fringe} represents all the fringe capacitances present in the device. When the voltage across the MOS decap is large and a strong inversion layer is present, the channel shields the depletion region capacitance and $C_{depletion}$ can be neglected resulting in C_{NMOS} becoming

$$C_{NMOS} \approx C_{oxide} + C_{overlap} + C_{fringe} \tag{4-8}$$

MOS decaps therefore behave as a variable capacitors, or varactors, with the capacitance varying with the voltage across it [88].

For a PMOS decaps, the substrate resistance, $R_{substrate}$, and n-well-to-substrate capacitance, $C_{nwell\_substrate}$, further appear between *VDD* and *VSS*. The overall effective capacitance, C_{PMOS}, and *ESR*, ESR_{PMOS}, for a PMOS decap are thus given by

$$ESR_{PMOS} = R_{gate} + \frac{R_{channel} \cdot R_{bulk}}{R_{channel} + R_{bulk}} + R_{line} + R_{contact} + R_{substrate} \tag{4-9}$$

$$C_{PMOS} = \frac{C_{oxide} \cdot C_{depletion}}{C_{oxide} + C_{depletion}} + C_{overalp} + C_{fringe} + C_{nwell\_substrate} \tag{4-10}$$

The capacitances $C_{depletion}$ and $C_{nwell\_substrate}$ are similarly shielded when the transistor is ON and C_{PMOS} becomes

$$C_{PMOS} \approx C_{oxide} + C_{overlap} + C_{fringe} \tag{4-11}$$

Although MOS decaps are often modeled using lumped parameters, J. Rius *et al.* show that a distributed model is more appropriate for capturing high frequency effects due to the variation in channel

resistance at high frequency [84]. This variation in channel resistance results in the *ESR* being a function of frequency.

Thick Oxide MOS Decaps

For designs in which leakage currents are an important constraint, thick oxide MOS devices can be used to minimize the leakage through the gate oxide layer [4]. The oxide thickness in these devices is typically three times larger than that of their standard counterparts. A drawback of this implementation is the reduction in capacitance per unit area. Table 4-II provides the relative oxide thickness for thin and thick oxide NMOS and PMOS transistors in a 65 nm CMOS technology.

TABLE 4-II DIELECTRIC THICKNESS FOR STANDARD AND THICK OXIDE MOS DEVICES

DEVICE	OXIDE THICKNESS, T_{OX} (Å)
NMOS	20
PMOS	22
2.5 V thick oxide NMOS	56
2.5 V thick oxide PMOS	59

Variable Threshold Voltage MOS Decaps

The availability of variable threshold voltage devices in modern technologies provides additional transistors with which MOS decaps can be implemented. Low threshold voltage devices have a higher semiconductor doping level and subsequently a lower channel resistance, and thus *ESR*, is expected compared to the higher threshold voltage devices.

Accumulation Mode MOS Decaps

The MOS decaps discussed previously are essentially inversion mode devices. Accumulation mode transistors can also be used to implement decaps and design kits typically model accumulation mode NMOS (A-NMOS) devices for use as varactors. Figure 4.10 illustrates an accumulation mode NMOS decap with its electrical model overlaid.

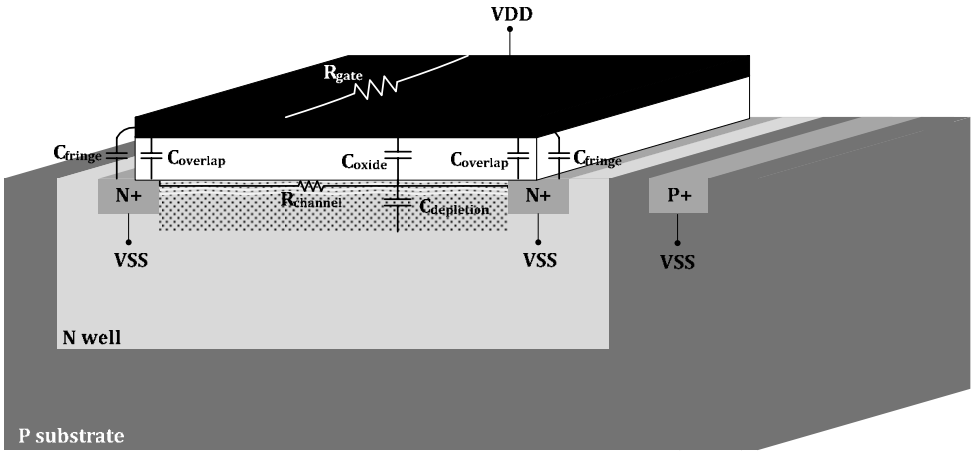


Figure 4.10 Accumulation mode NMOS capacitor with electrical model overlaid

Back-to-Back MOS Decap

A cross-coupled, or back-to-back, MOS decap design has also been discussed in the literature [86] and is illustrated in Figure 4.11. This design inherently provides an additional resistance (the device channel resistance) in series with the capacitance for the purpose of providing local electrostatic discharge (ESD) protection to the decap as well as resonance damping. The oxide breakdown voltage is almost linearly proportional to oxide thickness [89] and for a thin oxide NMOS device in a typical 65 nm technology the oxide breakdown voltage is about 1.2 to 2 V [90]. Since this voltage is close to the operating voltage of 1 V of these devices, the additional resistance inherent in this decap design can provide a degree of protection against oxide breakdown due to supply voltage oscillations, as described earlier. It does, however, trade off an increase in the overall impedance of the structure.

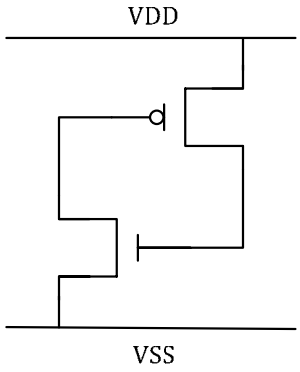


Figure 4.11 Back-to-back decap configuration

The *ESR* for the back-to-back design, ESR_{B2B}, is given by

$$ESR_{B2B} = R_{gate\_np} + \frac{R_{channel\_np} \cdot R_{bulk\_np}}{R_{channel\_np} + R_{bulk\_np}} + R_{line} + R_{contact} + R_{ON\_p} + R_{ON\_n} \qquad (4\text{-}12)$$

where, R_{gate_np}, $R_{channel\_np}$, R_{bulk_np} is the combined resistance of the NMOS and PMOS gate, channel, and bulk layers, respectively, R_{ON_p} is the ON resistance of the PMOS device and R_{ON_n} is the ON resistance of the NMOS device.

Gated Decaps

In gated decaps, a control transistor is placed in series with the decap, as illustrated in Figure 4.12. The transistor here serves to reduce the chip power consumption [91] by enabling the decap to be deactivated when certain circuit blocks are inactive, thus eliminating the associated leakage current. The gate also enables isolation of the decap in the event of shorts between the plates of the capacitor due to process defects. The added channel resistance in series with the capacitor, however, results in an increase in decap impedance. The control transistor thus has conflicting constraints of minimizing its ON resistance and minimizing its gate leakage.

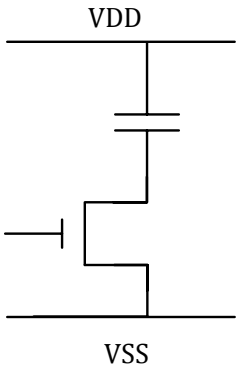


Figure 4.12 Gated decap configuration

The *ESR* for this design, ESR_{gated}, is given by

$$ESR_{gated} = ESR_{decap} + R_{ON\_control} \tag{4-13}$$

where, ESR_{decap} is the *ESR* of the particular decap used to implement the capacitance, and $R_{ON\_control}$ is the ON resistance of the control transistor.

MIM Decap

Some CMOS process options offer MIM decaps where additional steps are introduced into the process specifically for fabricating capacitors. Figure 4.13 illustrates the cross-section of a MIM capacitor fabricated between Metal 7 and Metal 8 of a CMOS process.

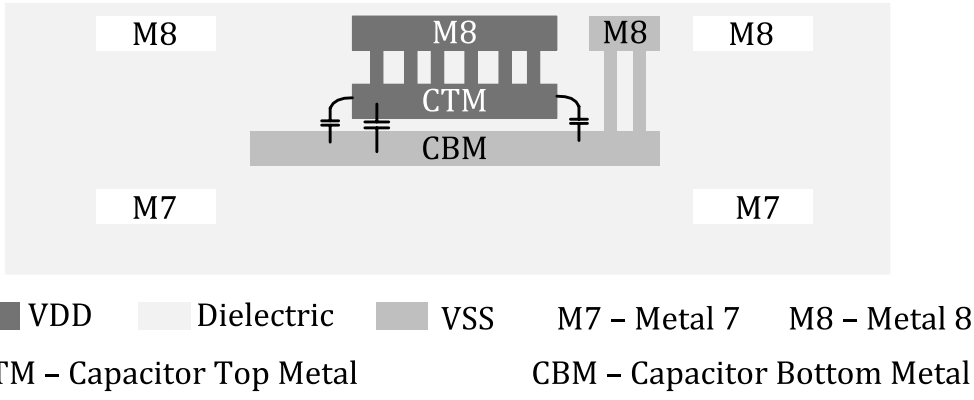


Figure 4.13 Cross-section of a MIM capacitor

The actual MIM decap *ESR*, ESR_{MIM}, and capacitance, C_{MIM}, depend on whether a grounded metal plate is placed underneath the device or if there are no layers present between the device and the substrate. In the case of the grounded metal underneath, an additional oxide is present between the metal and the device. In the case of no layers between the device and the substrate, a relatively small capacitance in series with the substrate parastics is present below the device. Resistance ESR_{MIM}, and capacitance, C_{MIM}, in either case, are given by

$$ESR_{MIM} \approx R_{plates} \tag{4-14}$$

$$C_{MIM} \approx C_{dielectric} + C_{fringe} \tag{4-15}$$

where, R_{plates} is the effective resistance of the dielectric and metal plates, respectively, and $C_{dielectic}$ is the capacitance of the dielectric layer. The inductive parasitics associated with the metal plates are typically modeled by design kits, however their magnitude is relatively small (few pH).

Metal Decaps

The various metal layers available in CMOS fabrication technologies can be used to form lateral and/or vertical capacitors. Lateral capacitors are formed by coupling capacitances between two traces on the same metal layer separated by dielectric. In the CMOS process, lateral capacitors are generally referred to as metal-oxide-metal (MOM) capacitors. MOM capacitors are typically formed as *interdigitated* structures as illustrated in the decap in Figure 4.14, where alternating lines are used to form the two terminals of the capacitor. Quasi-fractal capacitors have also been studied in the literature [92] where the perimeter of the capacitive structures is maximized with respect to area, however the interdigitated structure designed with minimum design rules provides the most lateral coupling capacitance per unit area [93].

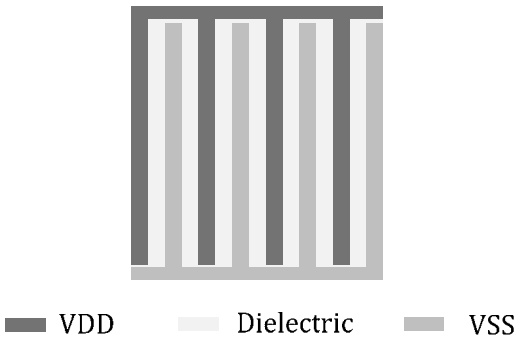


Figure 4.14 Top view of an interdigitated MOM decap

The metal capacitor *ESR*, ESR_{metal}, and capacitance, C_{metal}, are given by

$$ESR_{metal} = R_{plate} \tag{4-16}$$

$$C_{metal} = C_{dielectric} + C_{coupling} \tag{4-17}$$

With respect to the *ESL* of the interdigitiated MOM decap, the size of the unit decaps, is expected to be relatively small, and the *ESL* is negligible per device. As in the case of the supply grid, the interdigitated structure also helps to minimize inductance since the current in adjacent traces flows in opposite directions. The multilayer interdigitated capacitor has been shown to have much less inherent parasitic inductance compared to a simple multi-finger parallel plate capacitor [94].

Other capacitors - PIP, DTC

Other capacitive structures achievable in some CMOS technologies include poly-insulator-poly (PIP) capacitors and deep trench capacitors (DTCs). Both types of capacitors require specialized fabrication processes in addition to the standard CMOS process. PIP capacitors are similar to MIM capacitors however are formed using two polysilicon layers. The DTC process, originally developed for use in embedded DRAMs, provides a capacitor fabricated vertically into the substrate, as illustrated in Figure 4.15. Deep trench decaps have been shown to have a significantly reduced (~ 8 times) area compared to planar decaps with the same capacitance [95]. The cost of implementing DTCs, however, is expensive

due to the additional processing steps required. Neither PIP nor DTC structures are included in this analysis.

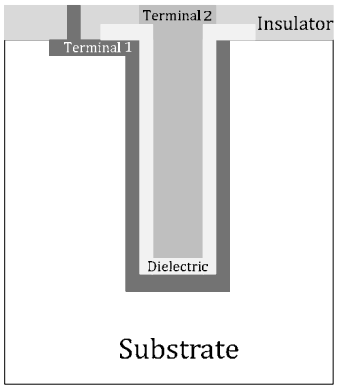


Figure 4.15 Deep trench capacitor (DTC) (not to scale)

4.2.2 Analysis of Decap Types

Table 4-III lists the various decap implementations analyzed. Different multi-layer metal decaps are further considered in a subsequent section. All the analyses are based on post-layout simulations in a 65 nm CMOS technology.

TABLE 4-III LEGEND FOR VARIOUS DECAP CONFIGURATIONS ANALYZED

SYMBOL	DESCRIPTION
NMOS	NMOS decap
NMOS_25	2.5 V thick oxide NMOS capacitor
NMOS_LVT	Low threshold voltage NMOS capacitor
A_NMOS	Accumulation mode NMOS capacitor
PMOS	PMOS capacitor
CMOS	NMOS and PMOS capacitor
MIM	MIM capacitor
GATED	Gated NMOS capacitor
B2B	Back-to-back capacitor
MOM_M1	Interdigitated Metal 1 MOM capacitor

The particular layout implementation of a decap affects its overall capacitance and area. For example, a fixed MOS oxide capacitance can be implemented using a single polysilicon finger or multiple parallel polysilicon fingers. The multiple finger structure will have a larger fringe component compared to the single finger structure, however will also have a larger area due to the minimum spacing requirements between the features. In addition, the frequency response of MOS based decaps has been shown to degrade with increasing channel length due to larger channel resistances at high frequencies [84], which further constrains the finger length. Furthermore, placing multiple fingers in parallel reduces the overall *ESR* and *ESL* of the decaps. Therefore, for the purpose of comparing the area efficiency of the various decap structures, a fixed layout topology is selected for the decap structures considered. Each MOS based decap is designed as a 6x6 array with a gate length of 1 μm. This length provides a practical layout without significantly affecting the frequency response, as shown later in this section. The width of each structure is designed such that each decap has a layout extracted value of C_{eff} equal to ~500 fF at 100 MHz. The MIM decap and MOM_M1 decap are similarly designed to provide a C_{eff} equal to ~500 fF at 100 MHz with the specific dimensions governed by design rules. The area of each structure thus differs based on how area efficient the particular implementation is. The extracted capacitance includes the oxide capacitance as well as any fringe/coupling capacitances that are present and is the overall capacitance of the decap structure. Each decap is thus designed independent of the circuit environment in which it may be used and while the exact magnitude of the suppression is dependent on the frequency components of the specific supply waveform, the degree of suppression is designed to be approximately the same for each decap at 100 MHz independent of the circuit environment. Care was also taken to minimize the area occupied by each decap. The frequency response is studied from 100 MHz to 30 GHz, the range in which the models used are valid. The specific layout parameters for each decap are given in Table 4-IV and the layout for an NMOS decap illustrated in Figure 4.16.

TABLE 4-IV DESIGN PARAMETERS FOR VARIOUS DECAP CONFIGURATIONS ANALYZED

DECAP	LAYOUT CONFIGURATION	UNIT L (µm)	UNIT W (µm)	AREA (µm^2)	LAYOUT EXTRACTED C_{eff} (fF)
NMOS	6x6 array	1	0.83	55	500
NMOS_25	6x6 array	1	2.29	135	500
NMOS_LVT	6x6 array	1	0.82	55	501
A_NMOS	6x6 array	1	0.80	61	498
PMOS	6x6 array	1	0.91	61	501
CMOS	6x6 p-array	1	0.42	75	493
	6x6 n-array	1	0.42		
MIM	7x7 array	2.01	2.01	1343	507
GATED	6x6 array	1	0.83	73	500
	Gate:	0.06	40		
B2B	6x6 p-array	1	0.43	75	500
	6x6 n-array	1	0.43		
MOM_M1	Interdigitated	6.7	0.09	1201	507

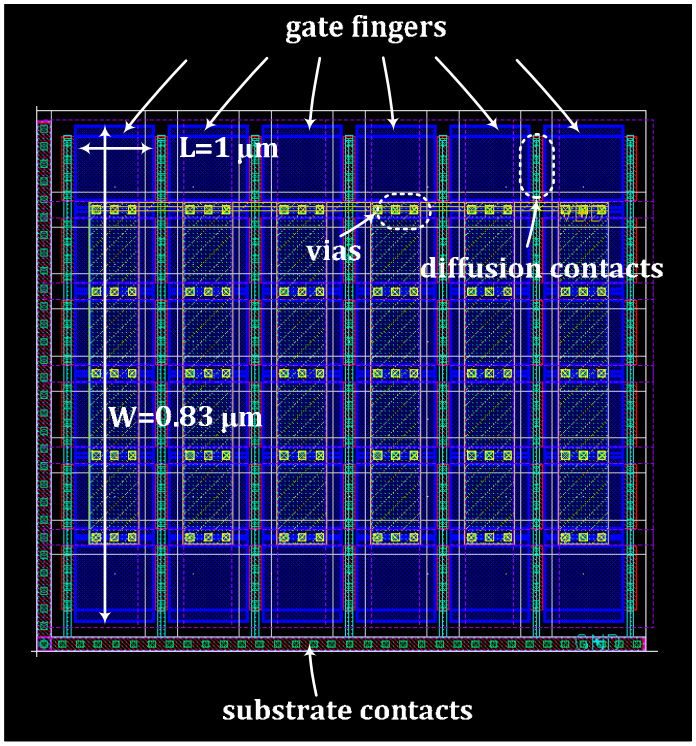


Figure 4.16 Layout of a 65 nm CMOS NMOS decap

4.2.2.1 Effective Capacitance (C_{eff})

The effective capacitance of each decap structure was determined in post-layout simulation by placing a small sinusoidal AC test voltage on a 1 V DC supply across the decap. Since the inductive parasitics of most on-chip decaps are negligible, the *ESL* is neglected in the model. The general decap model (Figure 4.17 (a)) is thus reduced to that shown in Figure 4.17 (b).

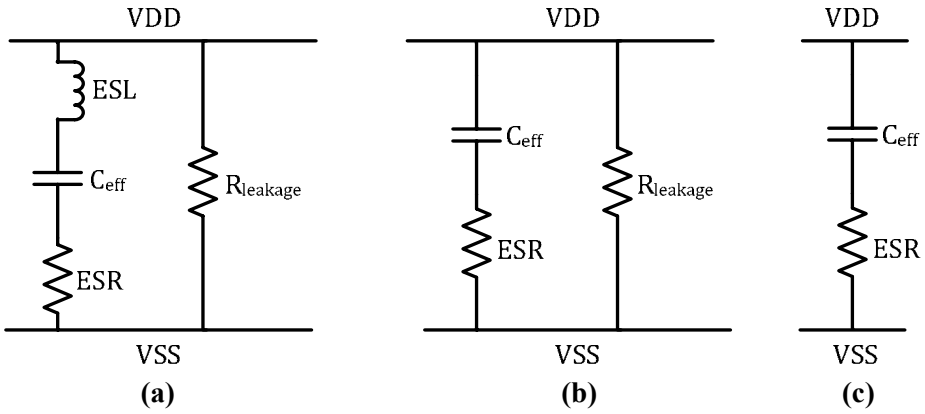


Figure 4.17 (a) Full decap model, (b) simplified decap model with leakage, and (c) simplified decap model without leakage

The decap impedance, *Z,* is thus given by

$$Z = \frac{R_{leakage}\left(ESR + \frac{1}{j \cdot \omega \cdot C_{eff}}\right)}{R_{leakage} + ESR + \frac{1}{j \cdot \omega \cdot C_{eff}}} \tag{4-18}$$

Separating the real and imaginary parts, Z becomes

$$Z = \left(\frac{\frac{R_{leakage} \cdot ESR\left(R_{leakage} + ESR\right)}{\left(R_{leakage} + ESR\right)^2 + \frac{1}{\omega^2 \cdot {C_{eff}}^2}}}{+ \frac{R_{leakage}}{\omega^2 \cdot {C_{eff}}^2 \left(\left(R_{leakage} + ESR\right)^2 + \frac{1}{\omega^2 \cdot {C_{eff}}^2} \right)}} \right)$$

$$+ j\left(\frac{R_{leakage}\left(R_{leakage} + ESR\right)}{\omega \cdot C_{eff}\left(\left(R_{leakage} + ESR\right)^2 + \frac{1}{\omega^2 \cdot {C_{eff}}^2} \right)} + \frac{R_{leakage} \cdot ESR}{\omega \cdot C_{eff}\left(\left(R_{leakage} + ESR\right)^2 + \frac{1}{\omega^2 \cdot {C_{eff}}^2} \right)} \right) \tag{4-19}$$

At frequencies in the range of interest (greater than 100 MHz), Z can be simplified to

$$Z \approx \frac{R_{leakage}}{R_{leakage} + ESR}\left(ESR + \frac{1}{j \cdot \omega \cdot C_{eff}} \right) \tag{4-20}$$

Since $R_{leakage} >> ESR$, Z can be further simplified to

$$Z \approx ESR + \frac{1}{j \cdot \omega \cdot C_{eff}} \tag{4-21}$$

and the model can be simplified to that in Figure 4.17 (c). The capacitance C_{eff} can be determined from

$$C_{eff} = \frac{1}{2 \cdot \pi \cdot f \cdot |Z| \cdot \sin\theta} = \frac{|I|}{2 \cdot \pi \cdot f \cdot |V| \cdot \sin\theta} \tag{4-22}$$

where f is the frequency of the test signal, θ is the phase difference between the voltage across and current through the decap, and $|V|$ and $|I|$ are the magnitudes of the AC voltage and current of the decap, respectively.

Figure 4.18 illustrates the effective capacitance versus frequency for the various decap structures considered. The NMOS decap and PMOS decap have very similar frequency responses for devices with a length of 1 μm in the 65 nm technology. The thick oxide decap and low threshold voltage decap also have similar frequency responses to the standard NMOS and PMOS devices.. The back-to-back design has the poorest response followed by the Metal 1 MOM decap. This is due to the higher series resistance in these

devices (as will be shown), that limits the flow of carriers at high frequencies [84]. The GATED design and MIM decap perform better, although not as well as the standard NMOS and PMOS capacitors.

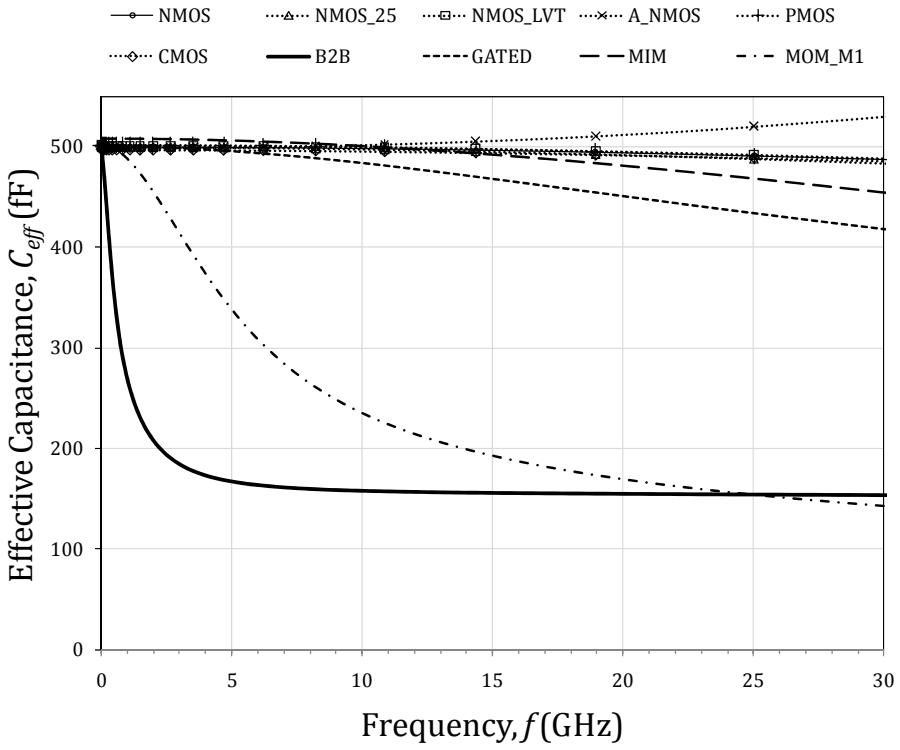


Figure 4.18 Effective capacitance versus frequency for various decaps

Figure 4.19 illustrates the effective capacitance per unit area versus frequency for each structure and thus indicates the *area efficiency* of each decap. The standard and low threshold voltage NMOS decaps are seen to be the most area efficient structures in the frequency range considered. The low threshold voltage device thus provides no significant improvement in capacitance per unit area. The PMOS decap also performs relatively well in terms of area efficiency and using a CMOS decap, as is often done within standard cells, results in a poorer area efficiency compared to using either an individual NMOS or PMOS structure. As expected, the thick oxide NMOS decap has a significantly poorer area efficiency compared

to the standard NMOS design, followed by the back-to-back design, and both MIM and Metal 1 MOM decaps are the most area inefficient designs.

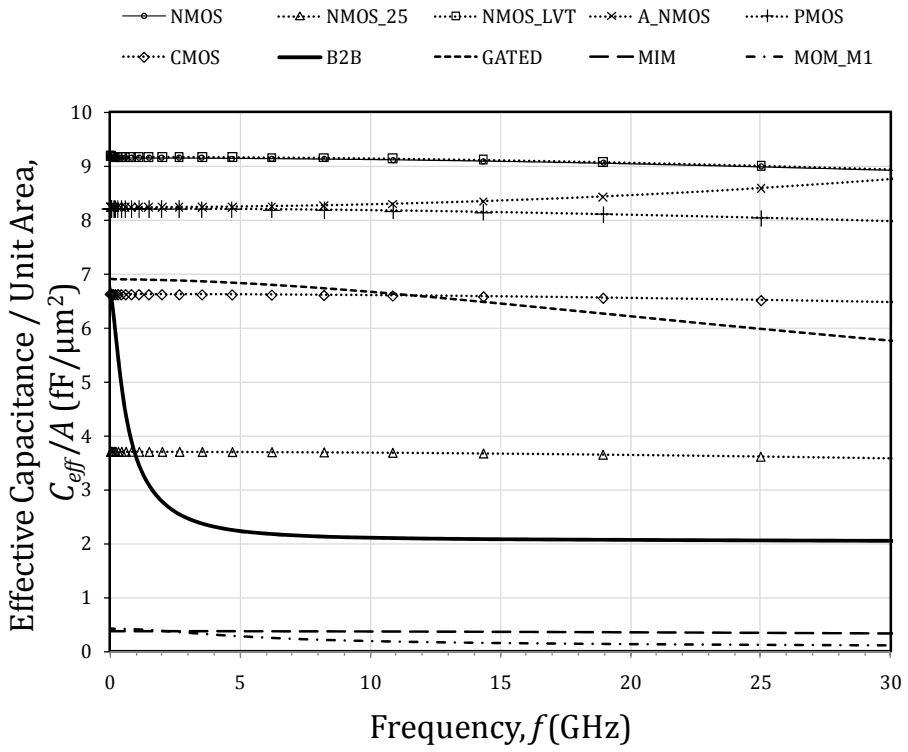


Figure 4.19 Effective capacitance per unit area versus frequency for various decaps

4.2.2.2 Equivalent Series Resistance (ESR)

Following from the analysis used to determine C_{eff}, the *ESR* of each decap structure studied can be obtained from

$$ESR = |Z| \cdot \cos\theta = \frac{|V| \cdot \cos\theta}{|I|} \tag{4-23}$$

Figure 4.20 illustrates the *ESR* of the various decaps versus frequency. The *ESR* remains approximately constant with frequency for most designs. The back-to-back design exhibits the largest *ESR* over most

frequencies as expected from the added channel resistances in series with the oxide capacitances in this design.

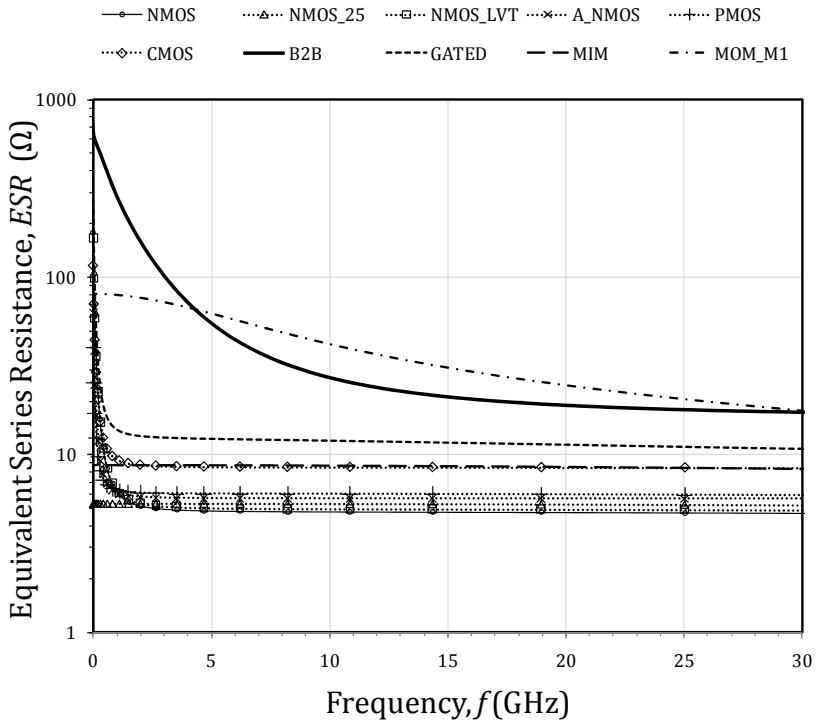


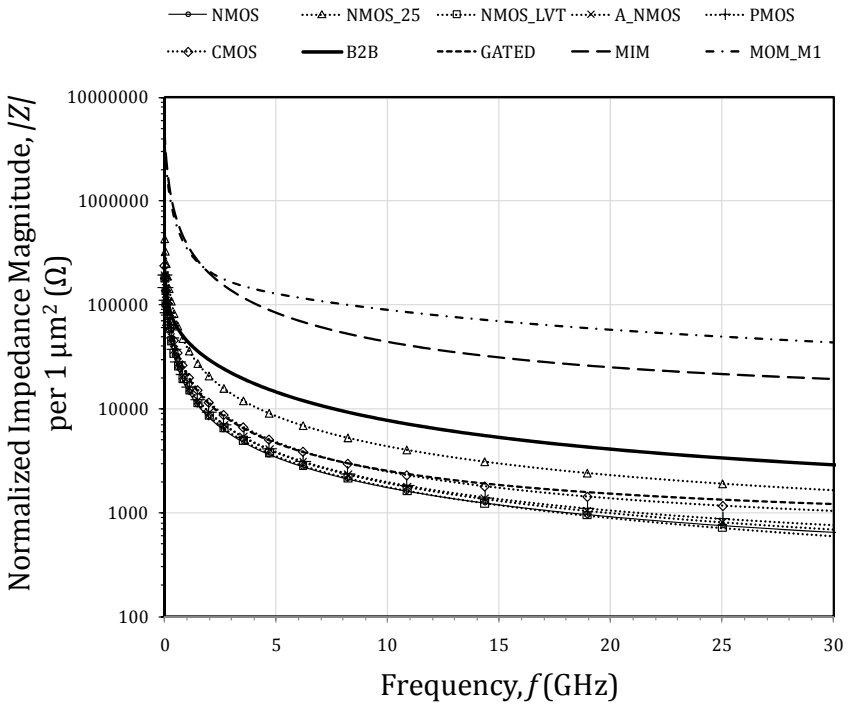
Figure 4.20 Equivalent series resistance for various decaps

TABLE 4-V ESR FOR VARIOUS DECAP CONFIGURATIONS

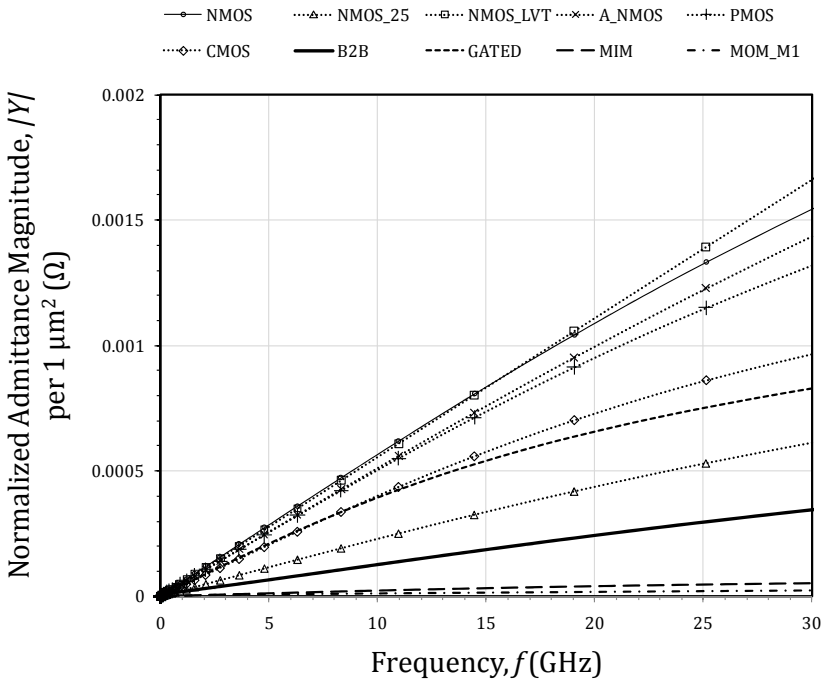
DECAP	ESR		
	900 MHz (kΩ)	10 GHz (kΩ)	30 GHz (kΩ)
NMOS	6.872	4.785	4.309
NMOS_25	5.291	5.275	4.669
NMOS_LVT	6.912	4.961	4.483
A_NMOS	6.345	5.658	5.708
PMOS	6.417	5.994	5.472
CMOS	9.828	8.516	8.107
MIM	8.727	8.666	6.632
GATED	14.41	11.89	8.775
B2B	333.5	25.69	15.9
MOM_M1	79.86	39.52	7.65

4.2.2.3 Impedance and Admittance

The overall impedance of each decap directly determines the extent of supply noise suppression that it provides and is a function of both the *ESR* and C_{eff} according to (4-2). The impedance magnitude normalized to an area of 1 μm^2 and corresponding admittance are plotted in Figure 4.21 (a) and (b), respectively, for each decap configuration considered, with the corresponding phase data provided in Figure 4.22. As can be seen, at higher frequencies (> ~20 GHz), the NMOS_LVT decap shows a slightly improved admittance compared to the NMOS decap and the former. However over most frequencies, the trends in impedance/admittance correspond to those of C_{eff} shown in Figure 4.19 and the C_{eff} data provides a good indication of the noise suppression capability of the decaps due to the relatively low ESR values. As can also be seen, the phase approaches 90º further supporting the dominance of capacitance and negligible inductance. Numerical values of the normalized impedance magnitude are provided for convenience in Table 4-VI at 100 MHz, and 1, 10 and 30 GHz.



(a)



(b)

Figure 4.21 Normalized (a) impedance and (b) admittance magnitude for various decaps

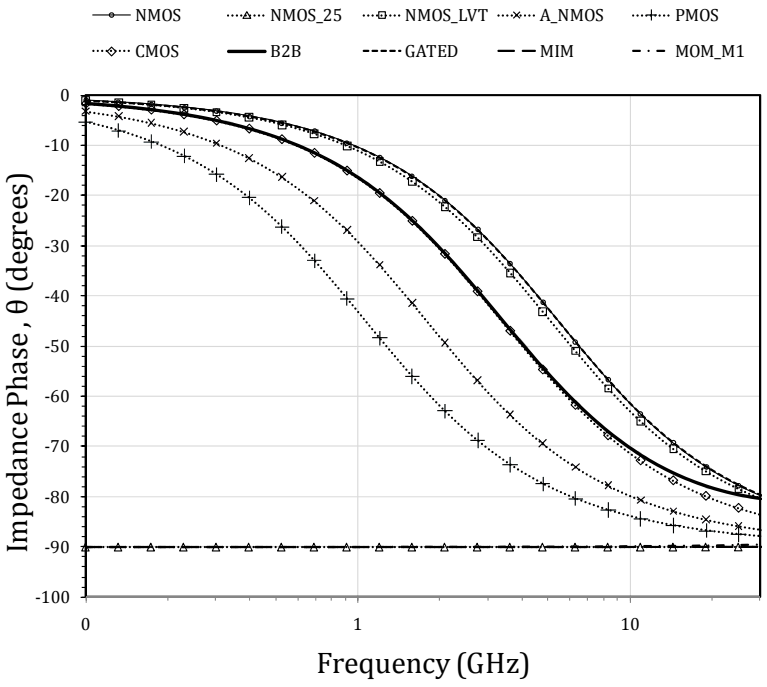


Figure 4.22 Impedance phase for various decaps

TABLE 4-VI NORMALIZED IMPEDANCE MAGNITUDE FOR VARIOUS DECAP CONFIGURATIONS

DECAP	NORMALIZED |Z|			
	100 MHz (kΩ)	900 MHz (kΩ)	10 GHz (kΩ)	30 GHz (kΩ)
NMOS	174	19	1.6	0.60
NMOS_25	430	47	4.0	1.52
NMOS_LVT	180	20	1.6	0.55
A_NMOS	193	21	1.8	0.64
PMOS	194	21	1.8	0.71
CMOS	240	26	2.3	0.98
MIM	4212	462	40.7	18.24
GATED	230	25	2.4	1.15
B2B	244	51	7.2	2.68
MOM_M1	3632	418	86.7	41.21

4.2.2.4 Leakage

Figure 4.23 shows the dielectric leakage currents measured at DC for each of the decap designs compared, with the effective capacitance per unit area at 10 GHz also provided here for convenience. While the data at only a single frequency is given, the relative trends are representative over most of the range of frequencies considered (as seen from Figure 4.19).

The NMOS decap has the highest leakage current (and thus greatest power dissipation) due to the thin oxide thickness and relatively small channel resistance, and the PMOS and accumulation mode NMOS decaps have approximately one third this leakage. In the latter two devices, the smaller leakage is attributed to their slightly larger *ESR* compared to the NMOS device.

The thick oxide NMOS, MIM and Metal 1 MOM decaps have almost no leakage current, with the thick oxide NMOS decap having the largest corresponding area efficiency. The thick oxide NMOS decap is thus the most desirable structure where leakage power is the primary constraint in a design, providing negligible dielectric leakage current at the cost of ~60% in effective capacitance per unit area.

In most designs situations both power and area are of concern, and in this case, both the PMOS and accumulation mode NMOS decaps provide a good trade-off between power and area, with the

accumulation mode NMOS performing slightly better in terms of area efficiency at frequencies > ~10 GHz, and the PMOS decap performing slightly better in terms of leakage. As is evident from Figure 4.23, the PMOS and accumulation mode decaps give a significant reduction in leakage (~70%) with a relatively small reduction in effective capacitance per unit area (~10%) compared to the NMOS decap.

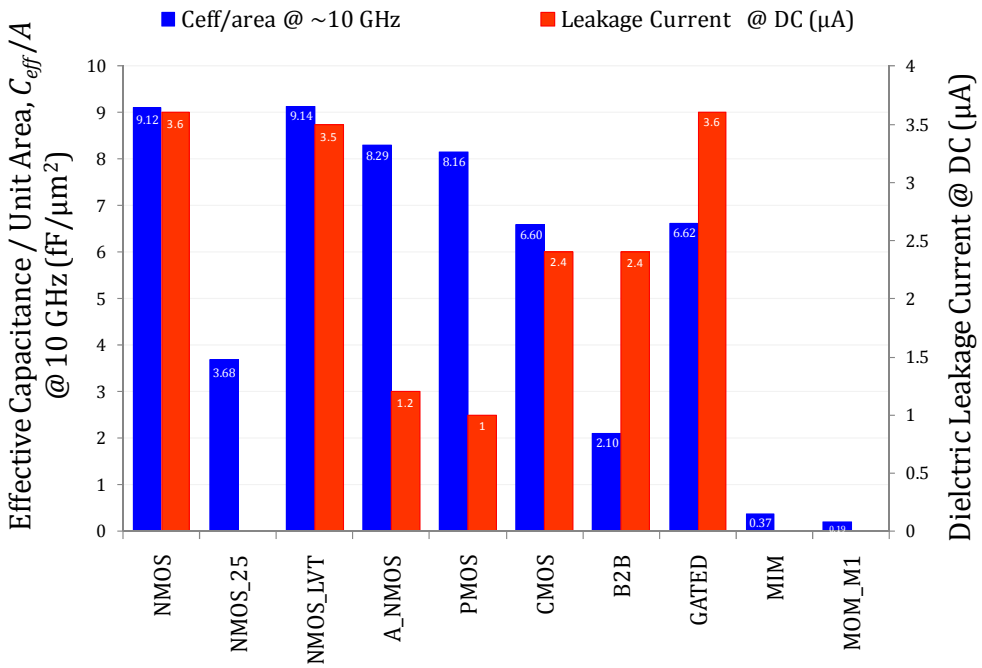


Figure 4.23 Effective capacitance per unit area (@ 10 GHz) and dielectric leakage current (@ DC) for various decap configurations

While this analysis is conducted at the 65 nm technology node, it can be noted that with the advent of high-k materials for use as the gate dielectric for subsequent technologies, the leakage power of decaps can be significantly reduced. The higher permittivity enables the thickness of the dielectric to be increased which exponentially decreases the leakage current of the device. Thus, as technology continues to scale the NMOS decap is likely to be the decap of choice with both minimum area and power.

It should also be noted that, although a decrease in effective capacitance per unit area and an increase in leakage is seen with the gated NMOS decap compared to the standard NMOS decap, the purpose of the

gated structure is to reduce the overall leakage power by modulating the gate to be closed when the decap is not in use. Thus, depending on the circuitry with which the gated decap is used, this decap can provide a desirable alternative to the standard NMOS, or PMOS or accumulation mode NMOS decap structures.

4.2.2.5 Capacitance-Voltage(C-V) Response

MOS based capacitors behave as varactors due to the variation in channel charge with varying voltage which leads to varying degrees of shielding of the depletion layer capacitance in these devices [88]. Figure 4.24 illustrates the *C-V* characteristics for the various decap configurations measured at 100 MHz. As expected, the MIM and Metal 1 MOM decaps provide the most stable capacitance over varying voltages. The supply voltage on most chips, however, does not typically vary more than 15% [54], and the region of interest in the *C-V* curve is indicated by the shaded region. In this region none of the decap structures exhibit a significant variation in capacitance.

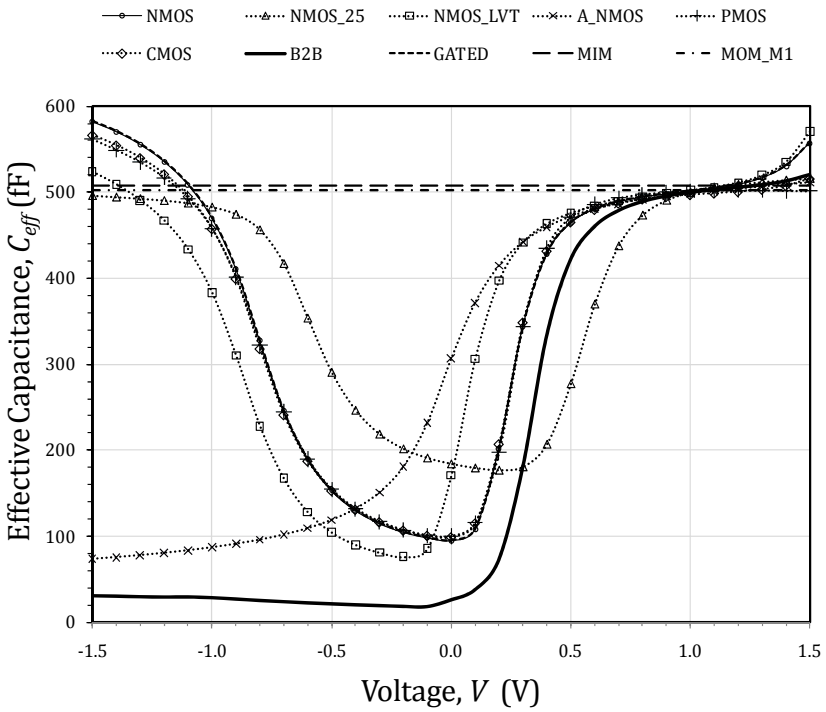


Figure 4.24 Capacitance-voltage plots for various decaps measured at 100 MHz

4.2.2.6 Effect of Placement

A large amount of research effort has been expended on determining the optimum location of decaps [96]-[103] since the added resistance of the interconnect between the noise source and the decap reduces the decap effectiveness by increasing its overall impedance. In the 65nm technology, the intermediate metals (Metals 2 to 7), typically used for routing, have a resistance of approximately 1.4 Ω/μm for traces with a minimum width. Thus a decap placed 20 μm away from a noise source and routed using a trace of minimum width in a Metal between 2 and 7, will have an additional resistance of ~28 Ω in series with the decap. The effect of this added resistance is quantified in Figure 4.25 for selected structures, where, again, $R_{distance}$ represents the resistance of the trace used to route the decap to the noisy supply. As can be seen, a 20 μm trace can undesirably increase the impedance of the decap by approximately 1.5 times.

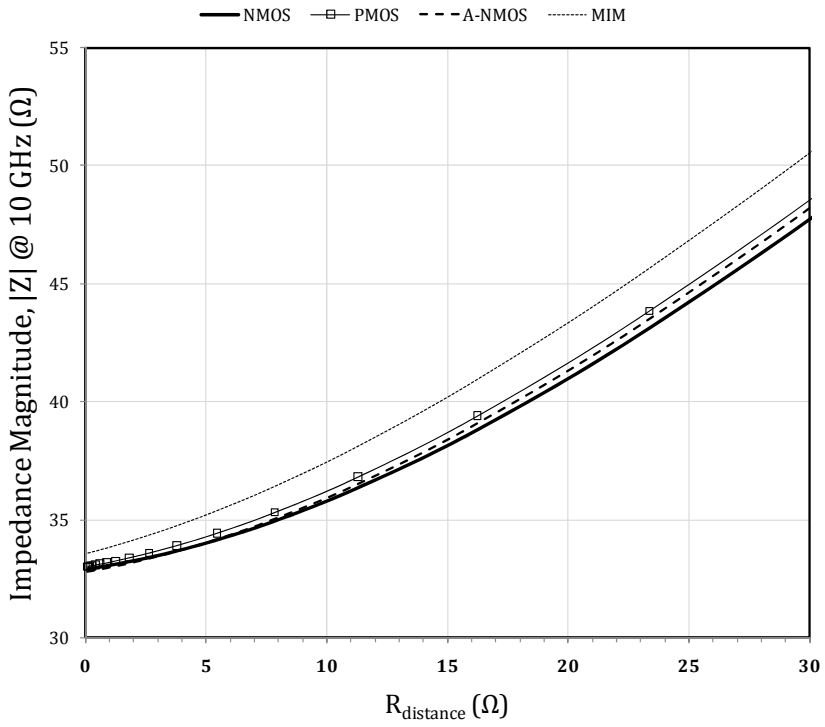


Figure 4.25 Effect of placement on effective capacitance for NMOS decaps

4.2.3 Layout Considerations

As was described earlier, the specific layout topology can have a significant impact on the area occupied by a particular decap. In the case of MOS decaps, as was seen earlier, the capacitance is primarily a result of the vertical oxide capacitance between the gate and silicon substrate. There is also a fringe component between the sidewalls and the substrate and a coupling component between the sidewalls and the substrate contacts. When implementing MOS decaps there is freedom in the number and size of the unit fingers used to implement the gate of the devices. Using multiple fingers to implement a particular decap compared to an equivalent single finger implementation, for example, can have two main effects. Firstly, the multi-finger decap will have a larger fringe component than its equivalent single finger equivalent due to the additional sidewall area. Secondly, the overall area of the multi-finger structure will be larger than its single finger equivalent due to the spacing requirements between the fingers in the multi-finger structure. The magnitude of these effects will of course vary depending on the specific layout configuration chosen. Four different NMOS layout configurations were simulated and these two effects quantified and summarized in Table 4-VII. The width of each gate finger was held constant and only the unit lengh and corresponding number of fingers was varied, thus keeping the overall gate area constant for all four structures. As can be seen, the decap area does not vary significantly for designs with 12 or less fingers and a relatively small (<2%) increase in capacitance is attributed to additional fringe capacitance where the number of fingers is increased in the dimension range considered.

TABLE 4-VII EFFECTIVE CAPACITANCE OF VARIOUS NMOS DECAP CONFIGURATIONS

UNIT L (µm)	NUMBER OF FINGERS	LAYOUT CONFIGURATION	UNIT W (µm)	AREA (um^2)	LAYOUT EXTRACTED C_{eff} @ 100 MHZ (fF)
1	36	6x6 array	0.83	~55	501
3	12	2x6 array	0.83	~49	508
6	6	1x6 array	0.83	~49	510
12	3	1x3 array	0.83	~49	511

Another important consideration, in determining the most optimal layout configuration of a MOS decap is the frequency response of the capacitance since the voltage along the channel length, *L*, of a MOS capacitor has been shown to vary with frequency [84]. The overall effective capacitance per unit area can be graphically observed for the three general types of MOS decaps (PMOS, NMOS and A_NMOS) in Figure 4.26. Here the length of each finger and thus total number of fingers is varied for each type of decap as indicated. As can be seen, the accumulation mode NMOS decap is the most sensitive to *L*, and the NMOS decap the least sensitive. The results further show that all of the NMOS, PMOS and accumulation mode NMOS decaps with an *L* of 3 μm provide the best capacitance per unit area over most frequencies considered. PMOS devices with an *L* of 12 μm and greater, and accumulation mode NMOS devices with an *L* of 6 μm and greater, should be avoided when high frequency noise components are present as a significant degradation in effective capacitance occurs with increasing frequency.

MIM decaps are similarly expected to exhibit variations in area based on their layout configuration although a more constant frequency response, however these decaps are not specifically considered in this analysis since their area efficiency was found to be approximately an order of magnitude smaller than that of NMOS decaps and are thus not recommended for the implementation of decaps. In the case of metal decaps, since a significant portion of their capacitance comes from the lateral capacitance between metal traces, adhering to minimum design rules for width and spacing is recommended to maximize the area efficiency of these structures.

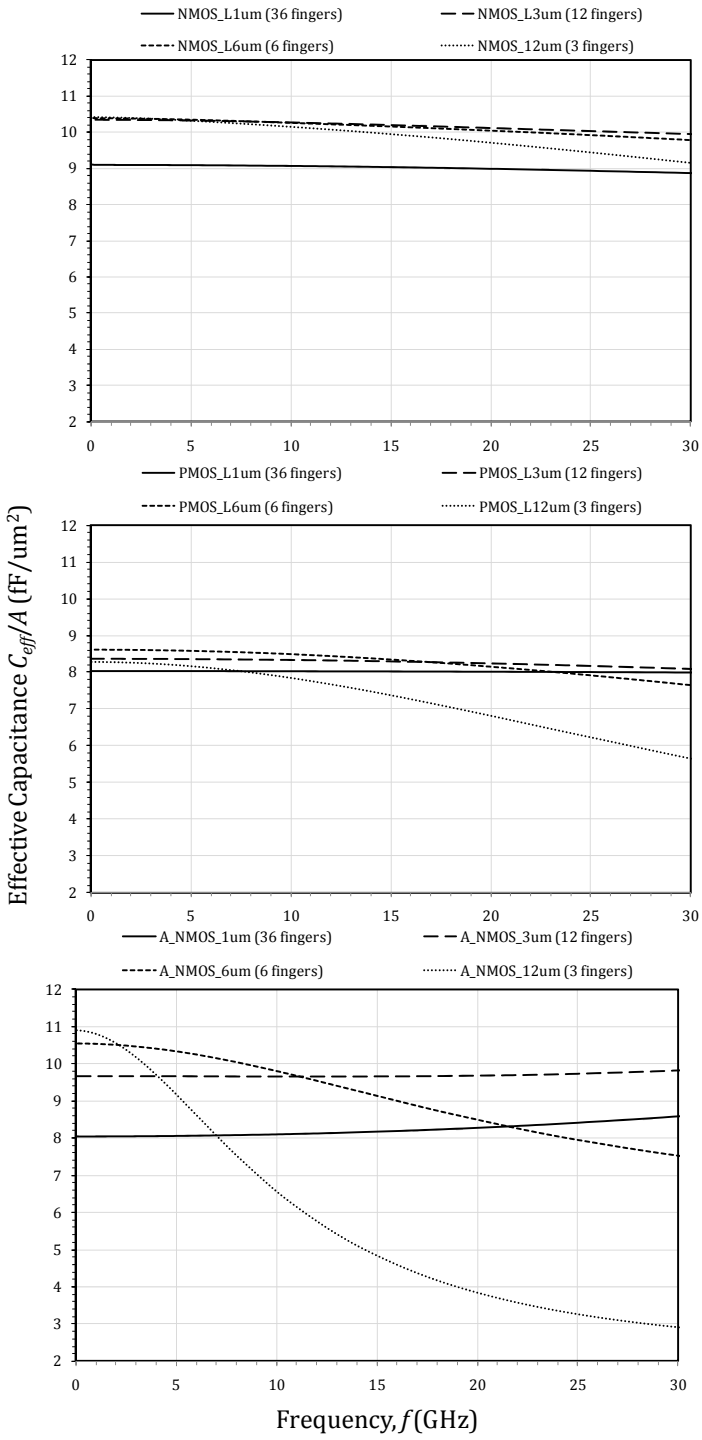


Figure 4.26 Effective capacitance per unit area for various lengths of NMOS, PMOS, and accumulation mode NMOS decaps

4.2.4 Multi-Layer Metal Decaps

Multi-layer MOM capacitors can be formed using interdigitated structures in multiple metal layers with vias used to connect the traces of each corresponding terminal. There is freedom in the orientation of the structure in each layer relative to other layers, for example, the structures can be parallel to each other and stacked as illustrated in Figure 4.27 (a), parallel to each other with the terminals alternating in the vertical direction as illustrated in Figure 4.27 (b) or perpendicular to each other as in woven or *rotative metal capacitors*, or RTMOM capacitors, as illustrated in Figure 4.27 (c). An alternate multi-layer metal decap can also be formed using sheets of metal layers with alternating terminals creating a series of vertical capacitors as illustrated in Figure 4.27 (d).

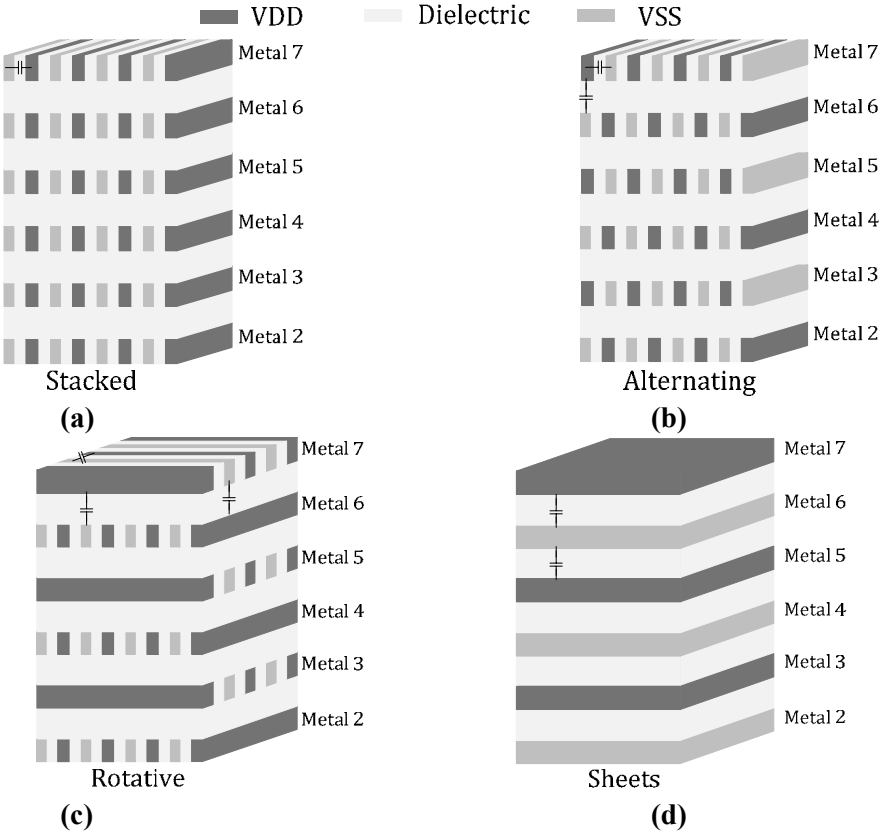


Figure 4.27 Cross-sectional view (not to scale) of (a) a stacked multi-layer, (b) an alternating multi-layer, and (c) a rotative MOM decap, and (d) a multi-layer decap using metal sheets

One question that arises when considering various metal decaps, is whether multi-layer interdigitated metal (MOM) decaps (Figure 4.27 (a), (b) and (c)) are more area efficient than a series of vertical decaps formed using sheets of metal layers (Figure 4.27 (d)). Which decap is more area efficient depends on the dielectric properties and achievable feature dimensions and spacings in the particular process technology used. Figure 4.28 (a) and (b) illustrate an alternating two-layer MOM decap and a two-layer metal decap using metal sheets, respectively, designed in a 65 nm CMOS technology. The MOM decap uses minimum dimensions for reliability and both structures are designed to have the same area.

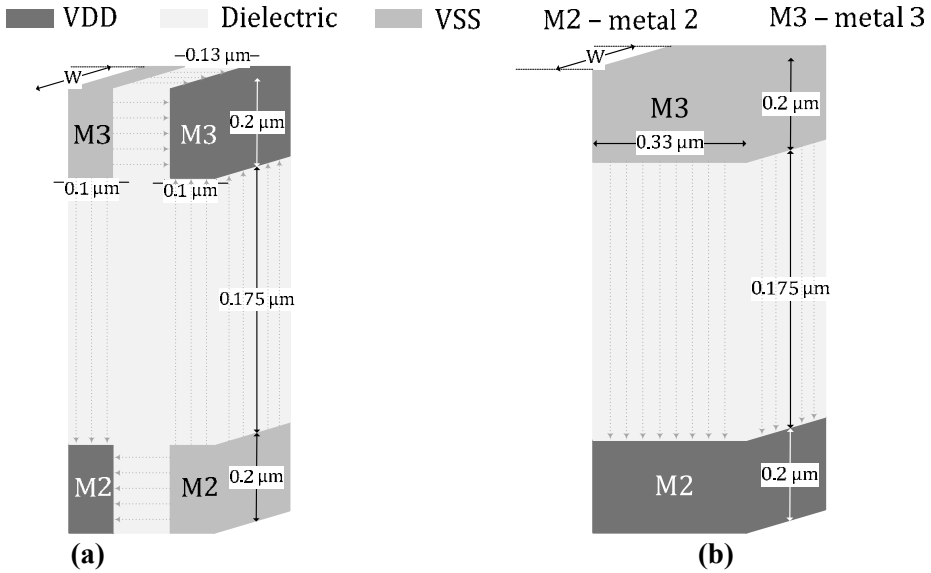


Figure 4.28 (a) Two-layer MOM decap, and (b) two-layer vertical metal decap.

The capacitances of the MOM decap, C_a, and the decap using metal sheets, C_b, per unit width, W, are given in Table 4-VIII, where $C_{a\_vertical}$ is the vertical capacitance and $C_{a\_lateral}$ the lateral capacitance of the MOM decap. As can be seen, the total capacitance of the metal decap using metal sheets is approximately 47% less than that of the MOM decap. The multilayer interdigitated structure is thus more desirable than a simple non-interdigitated metal decap using metal sheets in terms of capacitance per unit area. This advantage in capacitance is primarily a result of the minimum spacing dimension between metal features

being significantly smaller than the thickness of the dielectric between metal layers, which is typical of most processes. In addition, while it can be expected that the metal thickness will decrease with technology scaling, the minimum space requirements are correspondingly expected to decrease thus compensating for the effect of the decrease in metal thickness. The superiority of the interdigitated structure is thus expected to be evident over various processes.

TABLE 4-VIII COMPARISON OF A MOM DECAP AND A VERTICAL METAL DECAP

DECAP	TOTAL CAPACITANCE PER UNIT W	
MOM	$\frac{C_a}{W} = 2\frac{C_{a\_vertical}}{W} + 2\frac{C_{a\_lateral}}{W}$	0.117 fF/μm
Vertical Metal	$\frac{C_b}{W}$	0.062 fF/μm

4.2.4.1 Effective Capacitance (C_{eff})

Figure 4.29 shows the effective capacitance versus frequency for the multilayer metal structures of Figure 4.27. Metals 2 to 7 are used in these structures since Metal 1 is typically reserved for low-level routing, and Metals 8 and 9 have design rules that result in a minimal increase in capacitance of the overall structures and are typically reserved for top level routing. All the structures are designed to have the same area and the graph thus provides an indication of the relative area efficiencies of the structures (shown on secondary axis). The rotative structure, followed very closely by the alternating structure, provides the largest effective capacitance in the given area. The rotative structure also has the simpler layout of the two and it can be noted that the specific layout can cause a variation in the overall effective capacitance. The structures simulated in Figure 4.29 use minimum dimensions for reliability in a 65 nm technology. It can further be seen that the stacked structure is less area efficient than the rotative and alternating structures and the multi-layer decap using metal sheets has the poorest area efficiency compared to the other multi-layer metal decaps as was theoretically demonstrated earlier.

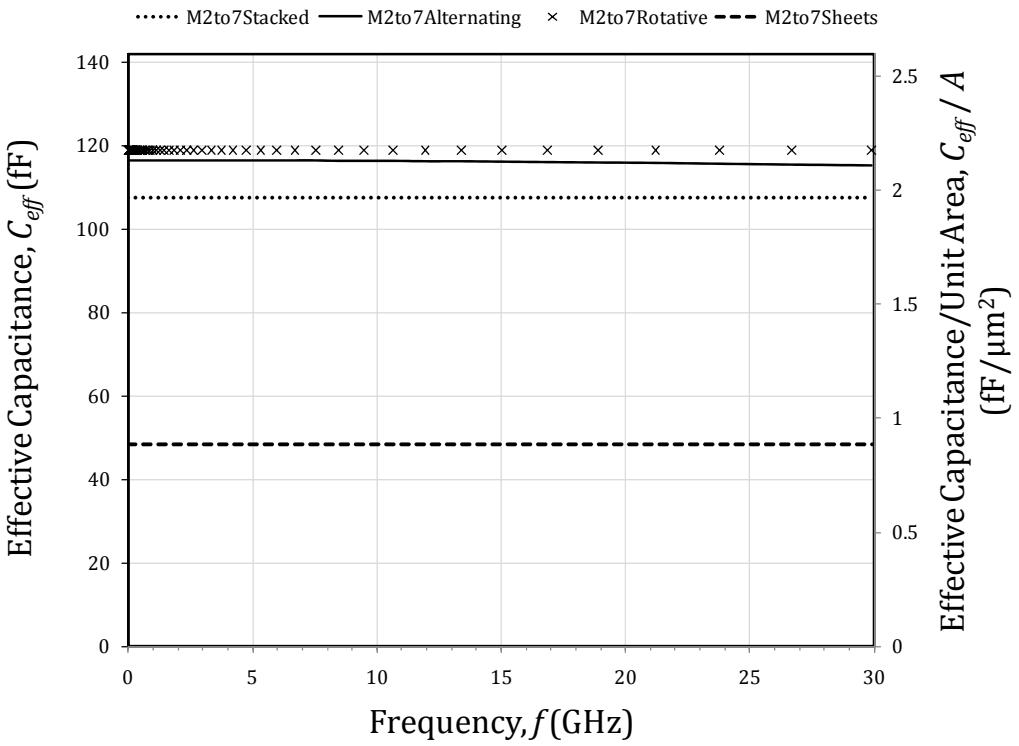


Figure 4.29 Effective capacitance and effective capacitance per unit area versus frequency for various multi-layer metal decaps

4.2.4.2 Equivalent Series Resistance (ESR)

Figure 4.30 shows the corresponding *ESR* for the structures simulated in Figure 4.29. The alternating structure exhibits the largest *ESR* due to its routing complexity. The stacked and rotative structures are comparatively simpler to realize in layout and thus exhibit correspondingly smaller *ESR*s. The decap structure realized with metal sheets has the smallest *ESR* due to the fact that the metal sheets have a smaller resistance than the traces used in the interdigitated structures.

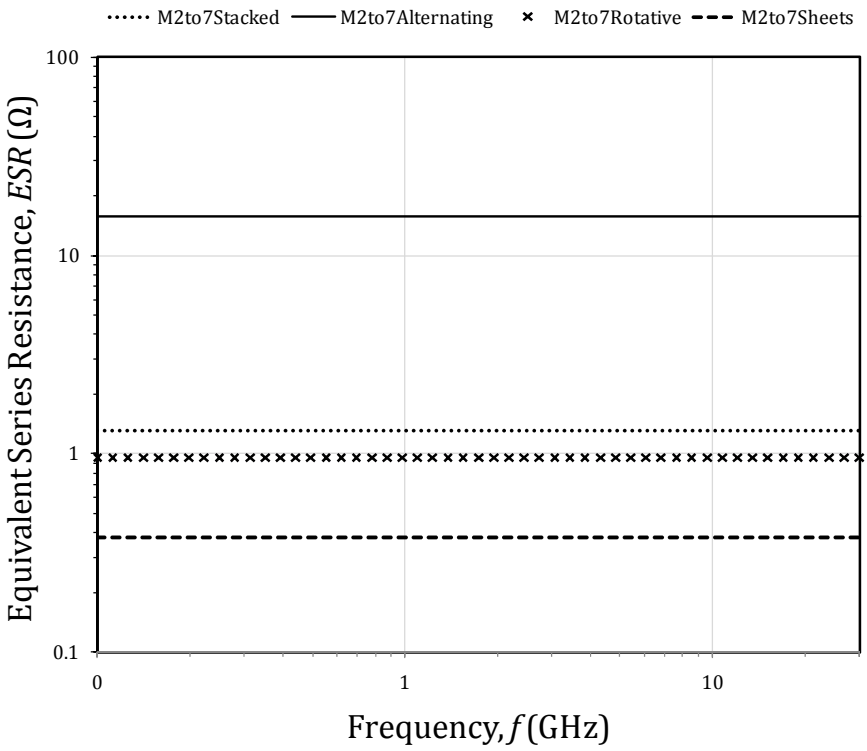


Figure 4.30 Equivalent series resistance for various multi-layer metal decaps

4.2.4.3 Impedance

The impedance magnitude with associated normalized impedance magnitude (on secondary axis) is plotted in Figure 4.31 for the various multi-layer metal decaps. Again, following from the effective capacitance trends, the rotative decap followed closely by the alternating decap provide the lowest impedance structures subsequently followed by the stacked structure, with the metal sheet structure significantly lagging.

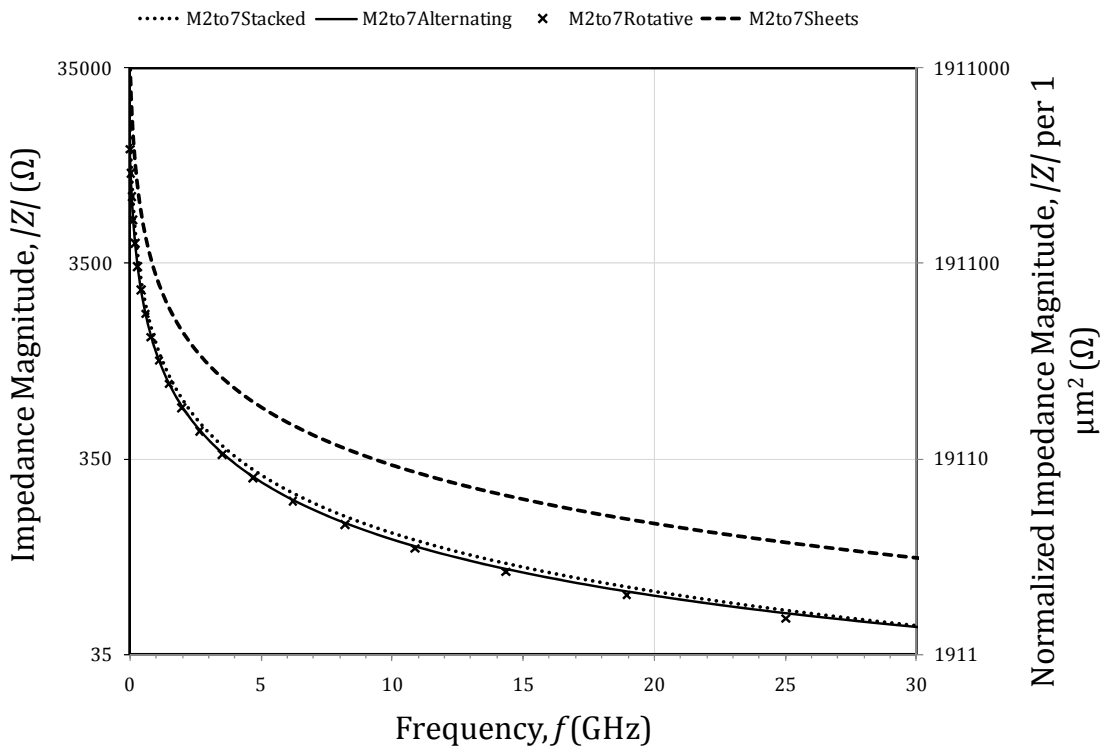


Figure 4.31 Impedance magnitude and normalized impedance magnitude for various multi-layer metal decaps

4.2.4.4 Placement of Multi-layer Metal Decaps

The multi-layer metal structures discussed thus far have the advantage of satisfying metal fill requirements of chips while providing decoupling capacitance between the supply and ground nodes. The interdigitated structures can be placed in each metal layer in any part of the chip, i.e. in chip 'whitespace', as well as on top of existing circuitry, where the corresponding metal layers are not being used for routing, thus allowing additional decap to be obtained at no additional cost in area. In other words, these multi-layer structures are desirable as metal fill patterns since they serve to simultaneously provide metal fill as well as decoupling capacitance.

4.2.5 Hybrid Decaps

In this section, the combination of more than one decap to form a decap hybrid structure is investigated. While it was shown earlier that MOS based decaps provide the most area efficient designs, these structures utilize only a few process layers, namely, the gate oxide, poly and typically a single metal layer for routing. This leaves multiple metallization layers available for providing additional capacitance within the same area. In a hybrid structure, a multi-layer metal decap is physically placed on top of, and electrically connected in parallel with, a MOS-based decap to increase its capacitance per unit area.

While metal is typically placed on top of various parts of the chip including decaps using an automated algorithm in a metal fill procedure to meet metal density requirements, this metal is typically not designed to provide capacitance between the supply and ground terminals. It is preferentially left floating however all the fill can also be tied to ground where sufficient computer resources are not available to deal with the floating metal. It is generally not connected to the supply as there can be floating supplies in a design due to power gating. Using the hybrid structure, the available metal layers can be used more efficiently in increasing overall decap capacitance while simultaneously fulfilling metal fill requirements.

Of course the metal-based decaps will not be as efficient in terms of capacitance in areas of the chip where some of the metals are required for routing. When decaps area added to a design in the design flow, they are first added to the ‘whitespace’ in a chip, i.e. areas occupied by metal traces only in interconnect limited designs. In these areas all the metal layers may not be available for the hybrid structures. However, further decap is then added to the chip which utilizes additional area. In these regions, the metal decaps can easily be integrated with the MOS decaps.

Reliability is another concern that should be considered. Since the metal decaps can cover relatively large areas of a chip, the likelihood of shorts between metals increases. The analysis conducted thus does

not use minimum spacing between metal traces but rather an increased spacing recommended for reliable fabrication for minimizing shorts.

Clearly, a number of MOS and metal decap combinations are possible. In this section, a combination of NMOS+RTMOM decaps, and NMOS+RTMOM+MIM decaps are investigated. In the NMOS+RTMOM decap, Metals 2 to 7 are used for the RTMOM. In the NMOS_RTMOM +MIM, only Metals 2 to 6 are used since the MIM decap is fabricated between Metals 7 and 8 and the design rules preclude the use of Metal 7 under a MIM capacitor.

Figure 4.32 illustrates the effective capacitance versus frequency for the hybrid decaps, where the area of each individual decap comprised in the hybrid structures is kept the same as that of the 500 fF (@100 MHz) NMOS decap considered previously. The data for a simple NMOS decap is also included for comparison. As shown in the figure, an increase in effective capacitance of ~25% is obtained with both hybrid structures compared to a simple NMOS decap. It can also be seen that omitting one metal layer (Metal 7) in the NMOS+RTMOM+MIM decap reduces any gains in capacitance from the MIM decap and no noticeable improvement is seen with this hybrid structure compared to the NMOS+RTMOM hybrid. Thus, including a MIM decap as part of the hybrid structure is not recommended. The corresponding *ESR,* and absolute and normalized impedance magnitudes are given in Figure 4.33 and Figure 4.34, respectively. A decrease in *ESR* is seen with both hybrid structures compared to the NMOS decap since the additional metal layers that form the MOM decap serve to reduce the overall series resistance. Again, following from the effective capacitance plot, the impedance magnitude of both hybrid decap structures is lower (by up to 28% in the frequency range considered) than that of the standard NMOS decap. Lastly, the dielectric current of the NMOS+RTMOM hybrid structure was observed to be negligibly higher than the NMOS decap alone. Therefore, the boost in effective capacitance is obtained at essentially no cost in leakage power.

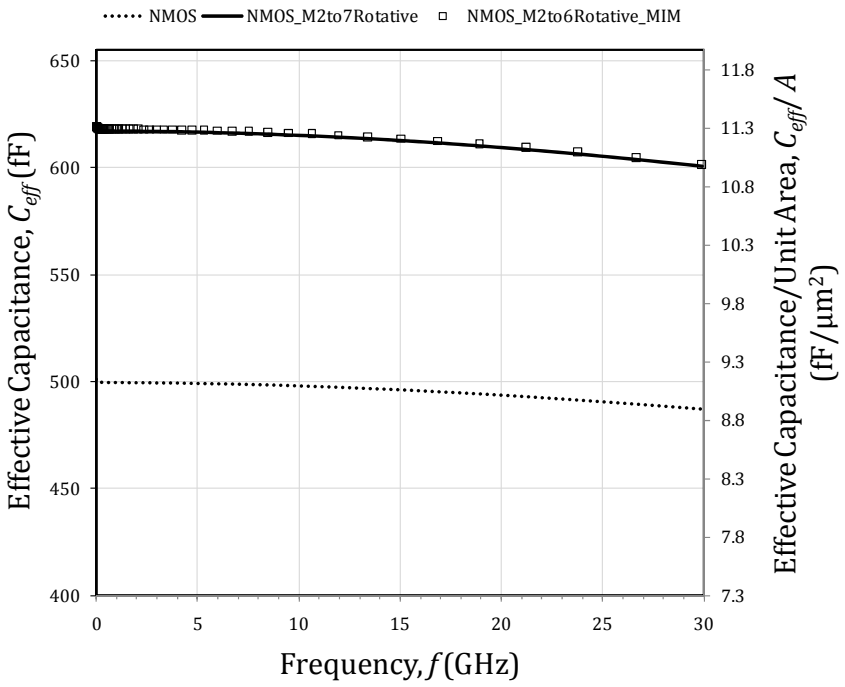


Figure 4.32 Effective capacitance and corresponding effective capacitance per unit area versus frequency for hybrid decaps and an NMOS decap

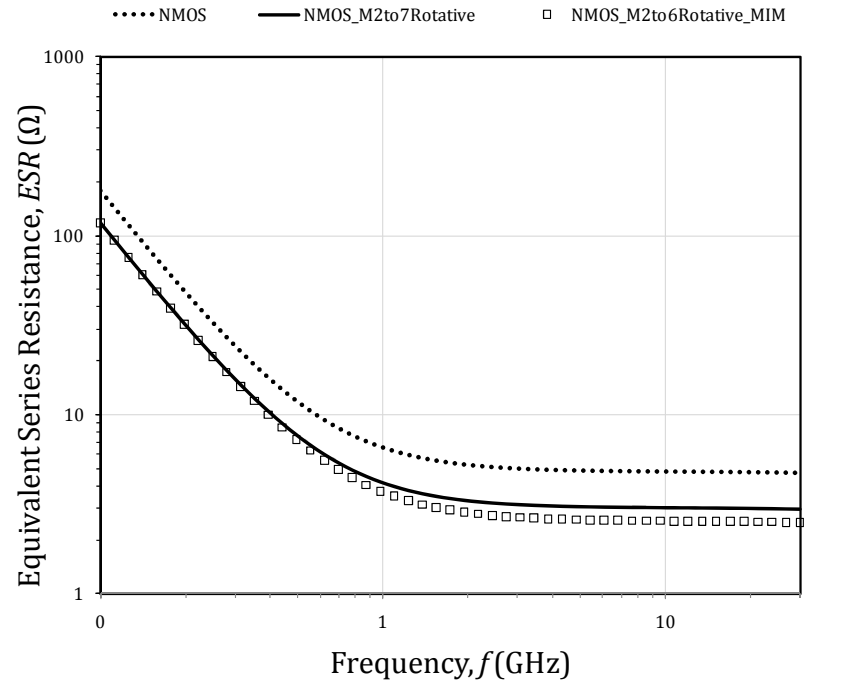


Figure 4.33 Equivalent series resistance for hybrid decaps and an NMOS decap

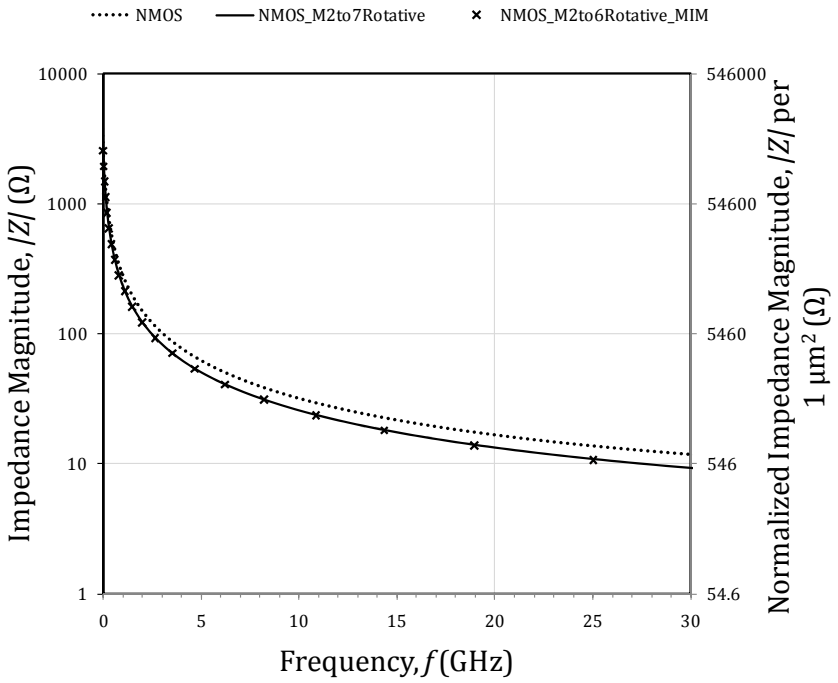


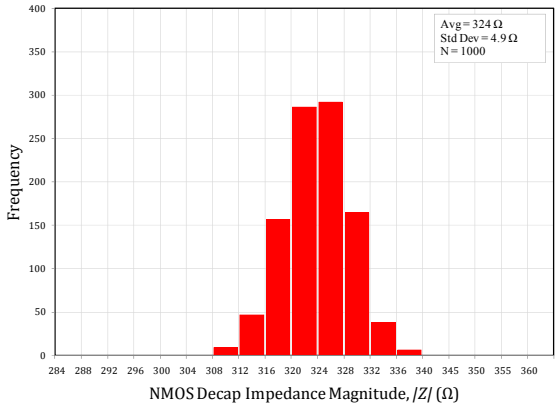
Figure 4.34 Impedance and corresponding normalized impedance magnitude for hybrid and an NMOS decap

4.2.6 Effect of Process Variations and Temperature

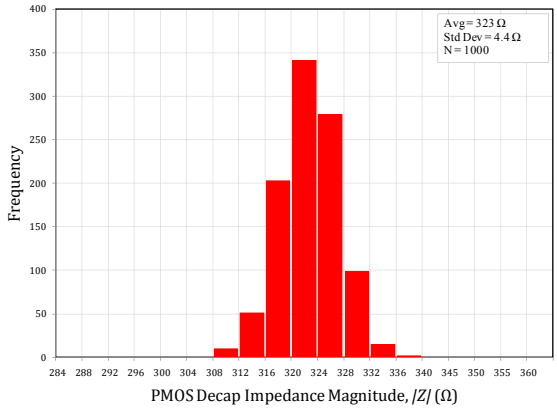
The overall impedance of the various decap structures will also deviate from their typical values with variations in processing parameters. Monte Carlo simulations were run on the NMOS, PMOS, RTMOM and NMOS+RTMOM hybrid decaps and the effect on the impedance magnitude plotted as illustrated in Figure 4.35. As can be seen, a sigma of ~1.5%, or a three-sigma variation of ~4.5%, can be expected for MOS based decaps. With respect to the multilayer metal decaps, variations in a number of dimensions can affect the overall capacitance of the structure, namely variation in the metal thickness, metal width and inter-metal dielectric thickness. The variations in these dimensions are given in Table 4-IX. Using the structure in Figure 4.28 (a) and the expected variations in the various dimensions of the structure, a three-

sigma variation of ~14% is expected in the overall effective capacitance of an alternating multilayer metal decap structure.

Figure 4.36 illustrates the effect of temperature on the various decap structures. As can be seen, the effective capacitance is only a very weak function of temperature and insignificant variations in the capacitance occur with change in temperature.



(a)



(b)

Figure 4.35 Monte Carlo simulation results for (a) NMOS, and (b) PMOS decaps

TABLE 4-IX VARIATIONS IN METAL AND INTER-METAL DIELECTRIC DIMENSIONS

DIMENSION	VARIATION
Metal Width	0.5%
Metal Thickness	15%
Inter-Metal Dielectric Thickness	10%

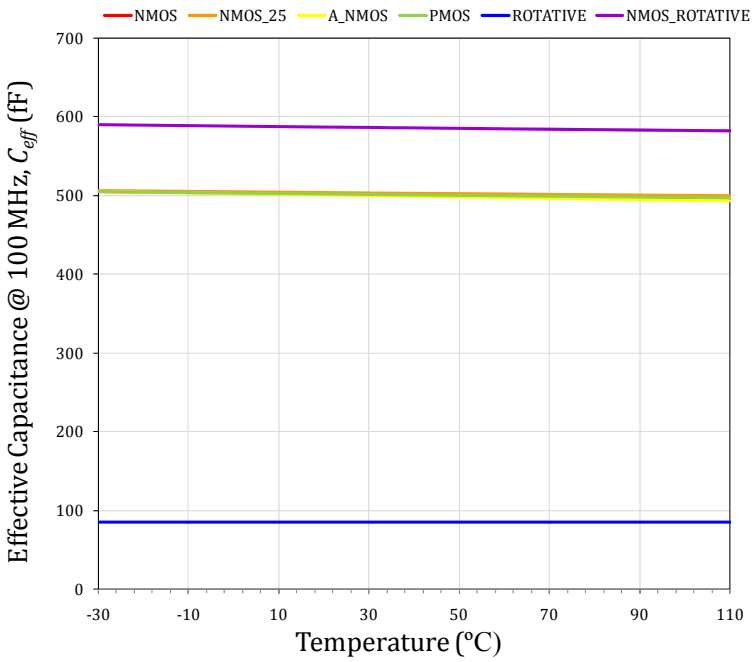


Figure 4.36 Effect of temperature on effective capacitance of various decaps

4.2.7 Selection of Decaps

The following guidelines can be used to determine which decap structure is most appropriate based on given design constraints for a 65nm CMOS technology.

1) Where *area* is the primary constraint in a design:
    a. NMOS+RTMOM multi-layer metal hybrid decaps should be placed in areas where additional non-whitespace decap is required. In regions where metal traces are present, if practical, the hybrid structure should be modified to utilize the available metal layers. NMOS_LVT decaps can replace NMOS decaps where available due to their slightly improved admittance characteristics at higher frequencies.
    b. Within standard cells, CMOS+RTMOM hybrid structures should be used and the RTMOM structure modified to exclude unavailable layers if necessary.

c. In all other areas where active circuitry is present, RTMOM structures should be placed to provide decap while simultaneously satisfying metal fill requirements. The RTMOM structures can be formed using all available metal layers from Metals 2 to 7.

2) Where both *area* and dielectric *leakage* power are of concern:
   a. A_NMOS+RTMOM hybrid or PMOS+RTMOM hybrid decap should be used in place of the NMOS+RTMOM decap in the previous scenario. PMOS and accumulation mode NMOS decaps provide a good trade-off between area efficiency and leakage (~70% less leakage for ~10% more area compared to the NMOS decap), with the accumulation mode NMOS performing slightly better in terms of area efficiency at frequencies >10 GHz and the PMOS having slightly lower leakage.
3) Where *leakage* power is the primary constraint in a design and must be minimized at cost of area:
   a. Thick oxide NMOS+RTMOM hybrid decaps should be used in place of NMOS+RTMOM decaps in scenario 1). Thick oxide NMOS decaps have negligible leakage current associated with them (at the cost of ~60% in area).
4) Gating decaps can further minimize leakage current in scenario 2) and 3) by isolating the decaps when not required. The magnitude of the overall savings in power, is based on the specific circuit to which the decaps are attached. Gating can be considered as an additional means of saving leakage power.
5) For noise components in the range of 100 MHz to 30 GHz, an L of ~3 μm should be used for MOS-based decaps.
6) The alternating interdigitated multi-layer metal decap performs similarly to the RTMOM decap in terms of area efficiency with respect to effective capacitance and impedance, and can replace the RTMOM decap. The alternating structure does however have a greater layout complexity.

The decaps recommended for various design constraints are summarized in Table 4-X.

TABLE 4-X SELECTION OF DECAPS BASED ON VARIOUS DESIGN CONSTRAINTS

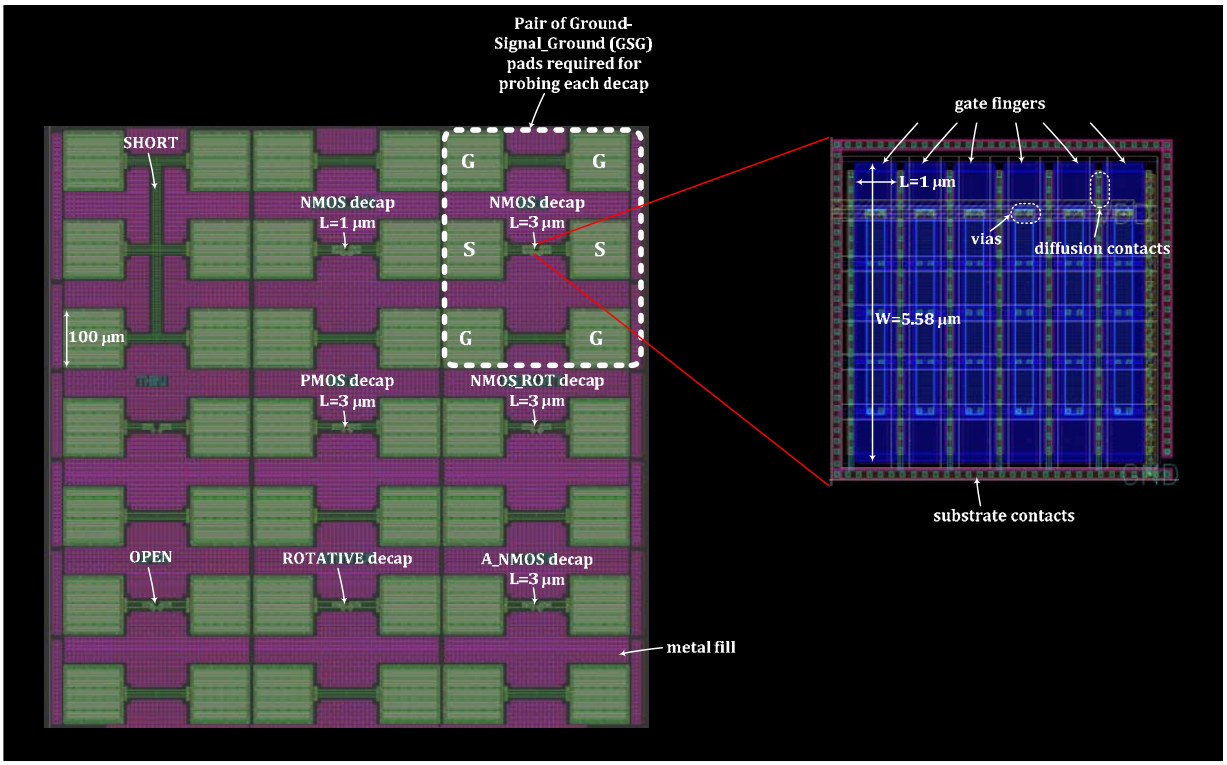
CONSTRAINT	LOCATION	DECAP
Area	Decap Space	NMOS+RTMOM*
	Standard Cell	CMOS+RTMOM*/**
	Device Whitespace	NMOS
Power‡	Decap Space	Thick NMOS+RTMOM*
	Standard Cell	Thick CMOS+RTMOM*/**
	Device Whitespace	Thick NMOS
Area and Power‡	Decap Space	PMOS***/A_NMOS****+RTMOM*
	Standard Cell	CMOS+RTMOM*/**
	Device Whitespace	PMOS*** /A_NMOS****

* Alternating multilayer metal decaps provide similar capacitance.
** RTMOM can be modified where there is dense metal routing present.
*** Where power is more important than area.
**** Where area is more important than power.
‡ Gating can also be used to reduce power consumption.

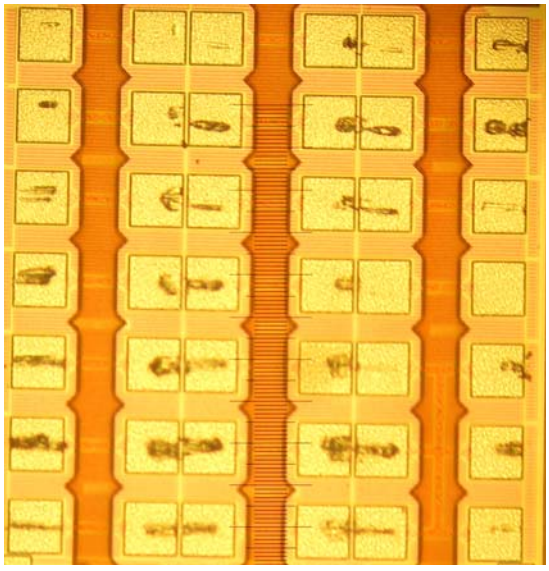
4.2.8 On-chip Measurement of Decap Parameters

Characteristics of selected decap structures were further obtained directly from a fabricated silicon chip in 90 nm CMOS technology. Figure 4.37 (a) shows the layout of the test chip and Figure 4.37 (b) the corresponding chip micrograph, with the characteristics of each structure identified in Table 4-XI.

The on-chip decap parameters are obtained using two-port measurements acquired directly by on-chip probing. Picoprobe® High Frequency Ground-Signal-Ground (GSG) Probes and a Cascade Microtech Manual RF Probe Station were used for probing and a Vector Network Analyzer (VNA) was used to obtain the associated scattering-parameters, or S-parameters. Figure 4.38 shows the various pieces of equipment used for testing.



(a)



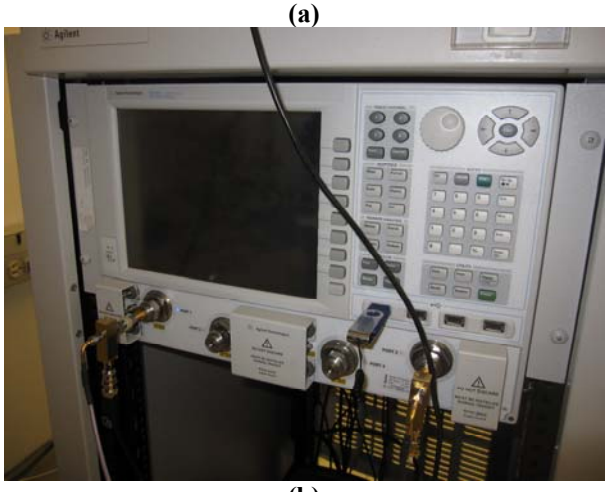
(b)

Figure 4.37 (a) Chip layout showing various decaps, de-embedding structures and close-up of a 90 nm CMOS NMOS decap, and (b) a chip micrograph

TABLE 4-XI DESIGN PARAMETERS FOR VARIOUS DECAP CONFIGURATIONS FABRICATED

DECAP	LAYOUT CONFIGURATION	UNIT L (µm)	UNIT W (µm)	LAYOUT EXTRACTED C_{eff} @ 100 MHz (fF)
NMOS	2x1 array	3	5.63	501.2
NMOS_ROTATIVE	2x1 array Interdigitated	3 6.07	5.63 0.16	587
A_NMOS	2x1 array	3	5.59	500.3
NMOS_1µM	6x1 array	1	5.58	499.5
PMOS	2x1 array	3	6.05	501.5
ROTATIVE	Interdigitated	6.07	0.16	84.84
SHORT	Short Circuit Ports	-	-	-
OPEN	Open Circuit Ports	-	-	-

(a)



(b)

Figure 4.38 (a) Probe station, and (b) network analyzer, used to test chip

4.2.8.1 S-Parameters

S-parameters are related to signal power and are generally used for high frequency measurements since direct current and voltage measurements are not straightforward at high frequencies. The VNA is capable of generating and measuring well defined travelling waves, and in a two-port measurement, four sets of travelling waves are present as illustrated in Figure 4.39. The travelling waves a_1, and a_2 represent the incident waves at port 1 and 2, respectively, and b_1, and b_2, the reflected waves at port 1 and 2, respectively, as illustrated in Figure 4.39.

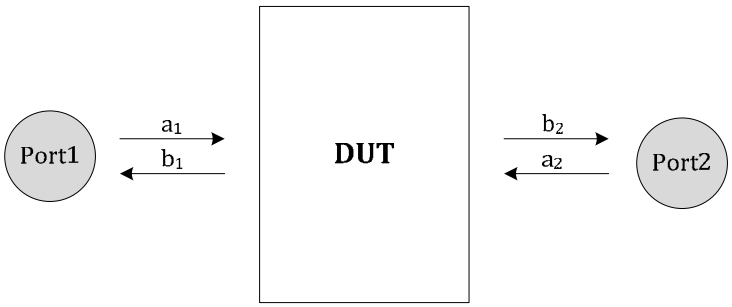


Figure 4.39 Two-port S-parameter measurement setup

Four sets of S-parameters are obtained from a two-port measurement and are defined as [104]

$$S_{11} = \left.\frac{b_1}{a_1}\right|_{a_2=0},\ S_{12} = \left.\frac{b_1}{a_2}\right|_{a_1=0},\ S_{21} = \left.\frac{b_2}{a_1}\right|_{a_2=0},\ S_{22} = \left.\frac{b_2}{a_2}\right|_{a_2=0} \tag{4-24}$$

These parameters completely describe the DUT behavior and can be converted into other types of parameters, such as Z (impedance)-parameters and Y(admittance)-parameters.

4.2.8.2 Calibration and De-Embedding

When performing a high frequency on-chip measurement, the S-parameters measured by the VNA include not only the characteristics of the device under test (DUT), but also a number of unwanted parasitic characteristics that become important at high frequencies. These characteristics include effects of the probes themselves, and cables and connectors between the probes and VNA, as well as the effects of

the on-chip pads required for probing and the related metal traces connecting the pads to the DUT. A two-step procedure is thus required to eliminate these extraneous effects. The first procedure involves using a standard calibration to eliminate the effect of the probes and all the related upstream components and wiring. The tools for this calibration are provided with any commercially available VNA. The SHORT-OPEN-LOAD-THROUGH (SOLT) equipment calibration technique is used here which is the most commonly used technique [105]. In this procedure, calibration standards comprising of SHORT, OPEN, LOAD and THROUGH standards as illustrated in Figure 4.40, are used. Software within the VNA is then used to remove, or de-embed, the characteristics of the probes and upstream components and wiring from the displayed results.

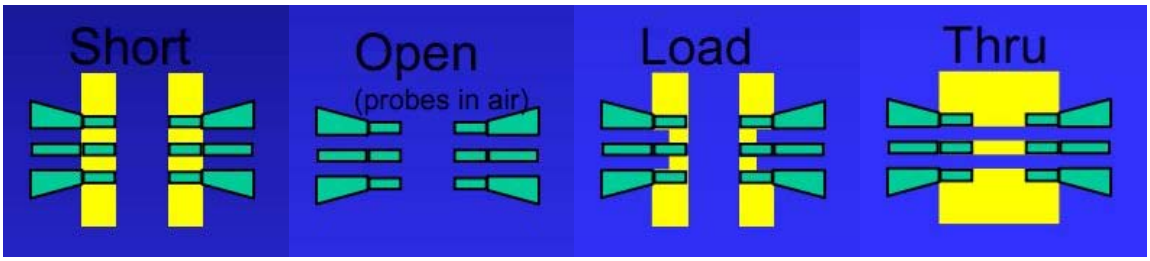


Figure 4.40 Calibration standards for SOLT calibration [105]

Secondly, the effects of the on-chip pads and metal traces must also be de-embedded from the measured data since the calibration removes the effects of the components only up to the probe tips. The de-embedding procedure essentially mathematically subtracts the extraneous on-chip networks from the measured results. In order to perform the on-chip de-embedding process, certain on-chip de-embedding structures are required depending on the particular de-embedding procedure used [106]-[111]. The de-embedding procedure here uses an OPEN-SHORT method which is an industry standard technique [106]. This technique requires the OPEN and SHORT test structures identified in Figure 4.37. The short structure short circuits the two measurement ports together and the open structure is the measurement pads and routing with the DUT omitted. The electrical model in Figure 4.41 represents the various

components in a single decap measurement structure. Here Y_{P1}, Y_{P2} and Y_{P3} represent the pad and wiring Y-parameters, and Z_{L1}, Z_{L2} and Z_{L3} represent the pad and wiring series impedances. The corresponding models for the open and short structures are given in Figure 4.42 (a) and (b), respectively. Again, the objective of de-embedding is to remove the effects of the unwanted parasitics thus providing only the information of the DUT.

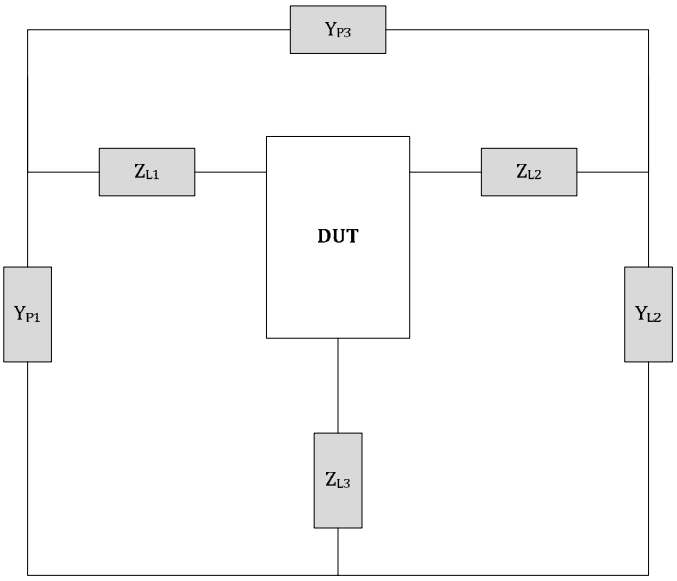


Figure 4.41 Two-port on-chip test structure model

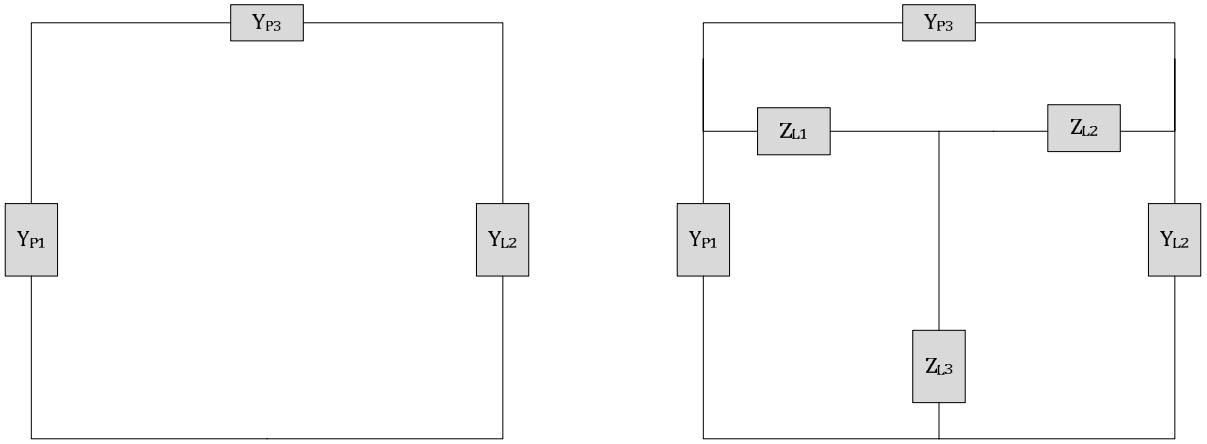


Figure 4.42 Electrical model for (a) open, and (b) short structures

The following procedure [106] is then used to obtain the Y-parameters of the DUT alone:

1. Subtract the admittance of the open structure Y_{OPEN} from the overall measured admittance Y_{ALL} and convert to impedance of structure with open structure parasitics removed $Z_{DUT+SHORT}$

$$Z_{DUT+SHORT} = \frac{1}{Y_{ALL} - Y_{OPEN}} \tag{4-25}$$

2. Subtract Y_{OPEN} from the admittance of the short structure Y_{SHORT} and convert to impedance of short structure only Z_{SHORT}

$$Z_{SHORT} = \frac{1}{Y_{SHORT} - Y_{OPEN}} \tag{4-26}$$

3. Subtract Z_{SHORT} from $Z_{DUT+SHORT}$ to get impedance of DUT structure alone

$$Z_{DUT} = Z_{DUT+SHORT} - Z_{SHORT} \tag{4-27}$$

In the work done, the S-parameters were imported into the Advanced Design System (ADS) software program by Agilent Technologies where the mathematical manipulations were conducted.

4.2.8.3 Measurement Procedure and Results

Based on the capability of the equipment used, the frequency was swept from 100 MHz to 27 GHz for each measurement. The required DC voltage on the gate of the MOS decaps was applied using an external DC supply with the signal coupled in using a T connector. While the parasitics of the DC supply path is not expected to affect the measurements, the calibration was performed with all the wiring in place.

The amplitude of the AC signal is also an important measurement set-up parameter since this should be kept relatively small during the measurements. Typical voltage amplitudes used in the small-signal characterization of MOS devices are approximately 100 mV or lower, which corresponds to a power setting of -15 dBm for the system used (50 Ω characteristics impedance) [112]. The noise level of the system used is approximately -40 dBm. The measurements were thus conducted at this power setting. The

measurements were also conducted with the microscope lamps turned off in order to prevent any light-induced generation of carriers in the semiconductor.

In order to determine the effective decap capacitance and *ESR* from the de-embedded S-parameters, the equivalent one-port S_{11} parameter is first obtained from [104]

$$\left(S_{11}\right)_{1port} = \left(S_{11} - S_{12} \cdot \frac{S_{21}}{1+S_{22}} \right)_{2port} \tag{4-28}$$

The effective capacitance is then obtained from

$$C_{eff} = \frac{1}{\left| \mathrm{Im}\left(\frac{1}{Y_{11}} \right)_{1port} \right| \cdot 2\pi f} \tag{4-29}$$

and the *ESR* from

$$ESR = \left| \mathrm{Re}\left(\frac{1}{Y_{11}} \right)_{1port} \right| \tag{4-30}$$

Figure 4.43 shows the simulated versus measured effective capacitance versus frequency for NMOS, rotative, and NMOS+ROT hybrid decaps in the 90 nm technology. As can be seen there is excellent agreement between the measured and simulated results for the rotative multilayer metal decap. Furthermore, the expected ~20% increase in capacitance from the standard NMOS to the NMOS+ROT hybrid is verified in the measured results.

There is however, a significant discrepancy in the frequency response of the NMOS decap. The effective capacitance falls off with frequency at a much greater rate than predicted by the simulator. The design kit used for the simulations incorporate the BSIM4 and BSIM3v3 models. It is well known that the core models do not implement the MOSFET poly-silicon gate resistance [113]-[115]. The source and drain resistance are also not explicitly modeled in the basic models [116]. While omitting the gate and

source/drain resistances do not cause a problem at relatively low frequencies, it leads to significant discrepancies in high frequency simulations. RF CMOS models therefore generally include an additional resistor that represents the gate resistance in the models [117]-[119]. A difference in the measured and simulated frequency response can therefore be expected for the decaps analyzed. While RF models are available in the design kit used, their models are limited to very small ranges of device dimensions. The core models were thus used to simulate the behavior of the decaps in this work. The actual drop-off in frequency response is and interesting observation, however, and it should be noted that the decaps are expected to be less effective than anticipated for higher frequency supply noise components. Although, considering the stagnation of clock frequencies with the move to multi-core computer architectures, the frequency of the dominant noise component is expected to remain in the low GHz range.

Figure 4.44 illustrates the effective capacitance for the NMOS decaps with gate lengths of 3μm and 1μm. As expected, the device with L=3μm, has a poorer frequency response due to the larger channel resistance. Again, a significant discrepancy is seen between the simulated and measured frequency response. Given this behavior, and even though the simulations indicate that using L=3μm is more area efficient, it is recommended that smaller lengths be used for decaps due to the large difference in frequency response observed on-chip. Figure 4.45 illustrates similar effective capacitance curves for an accumulation mode NMOS decap and a PMOS decap. The frequency response of the PMOS is seen to fall off at a very large rate as expected due to the larger channel resistance in PMOS devices.

Figure 4.47 and Figure 4.48 illustrate the ESR as a function of frequency. As can be seen, while the order of the ESR values for the various decaps is almost the same in both the simulated and measured cases, the relative magnitude of the measured values are higher as expected from the preceding discussion.

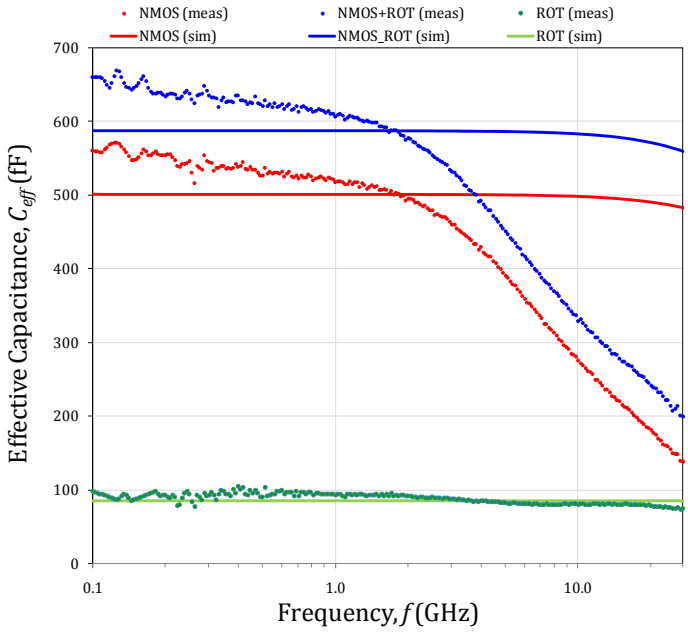


Figure 4.43 Simulated versus measured effective capacitance for NMOS, rotative and NMOS+ROT hybrid decaps

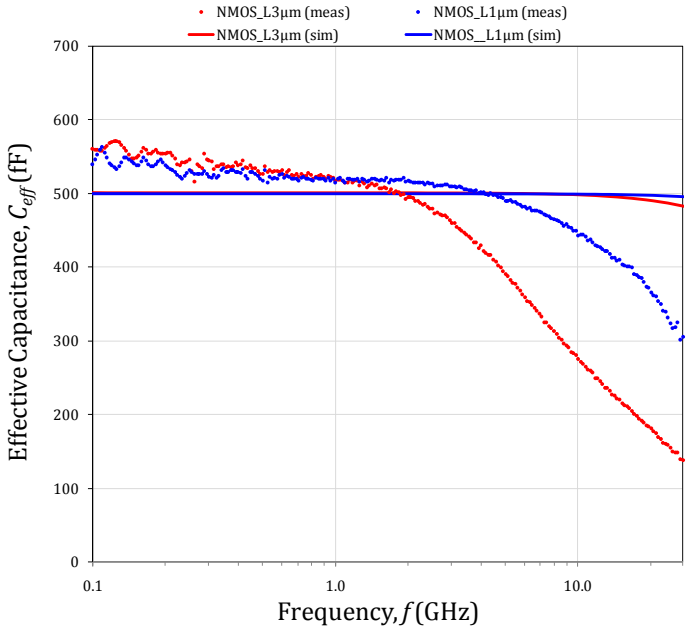


Figure 4.44 Simulated versus measured effective capacitance for NMOS with L=3µm, and NMOS with L=1 µm

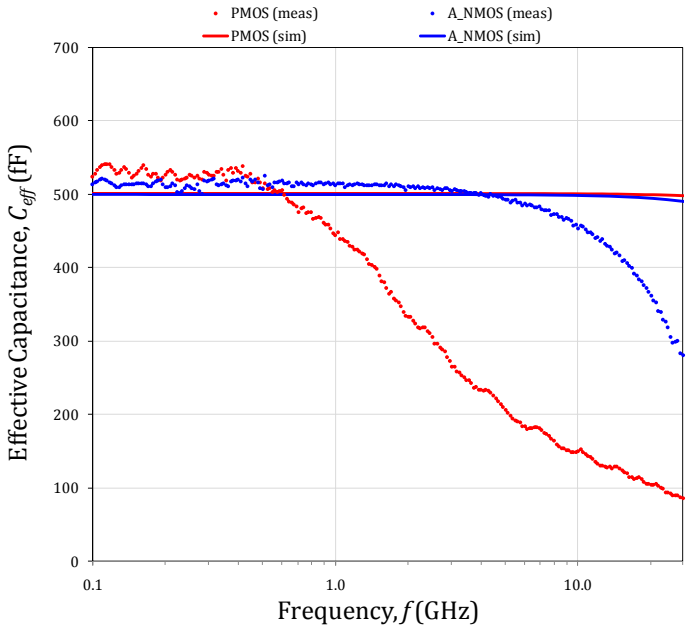


Figure 4.45 Simulated versus measured effective capacitance for PMOS decaps and accumulation mode NMOS decaps

In order to correlate the behavior seen in the measured device to the omitted resistances in the model, the gate resistance and source/drain resistances were added to the basic model and the effective capacitance re-simulated for the case of the NMOS decap. The gate resistance was calculated to be approximately 10 ohms and the source and drain metal resistances were each calculated to be approximately 5 ohms. Increasing the gate resistance was found to have a negligible effect on the effective capacitance values, however, as illustrated in Figure 4.46, the resistance at the source and drain were found to produce the behavior observed in the measured data. The calculated source and drain resistance of 5 ohms included only the metal traces at the diffusions, however, from the simulated data, additional resistance at these terminals is seen to be present. These resistances are likely due to contact/via resistances not accounted for in the model.

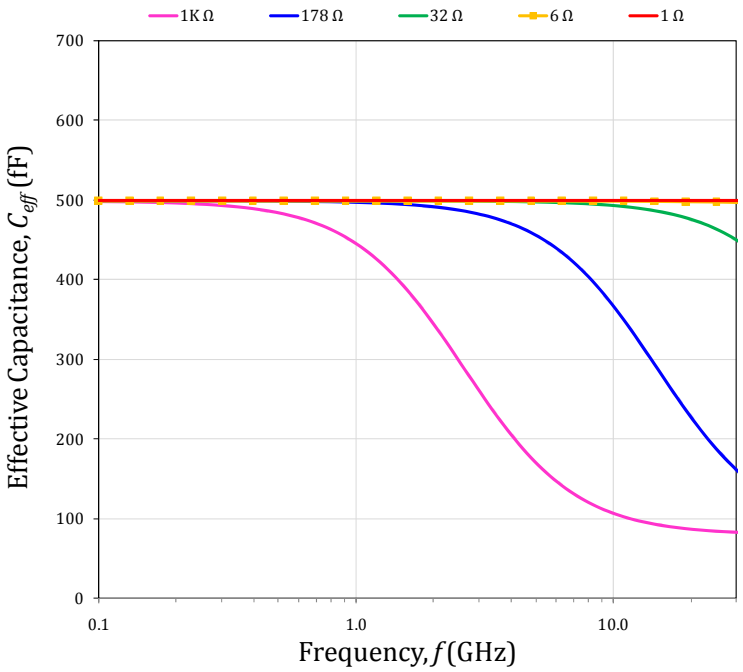


Figure 4.46 Effective capacitance versus frequency for varying amounts of resistance added at each of the drain and source terminals

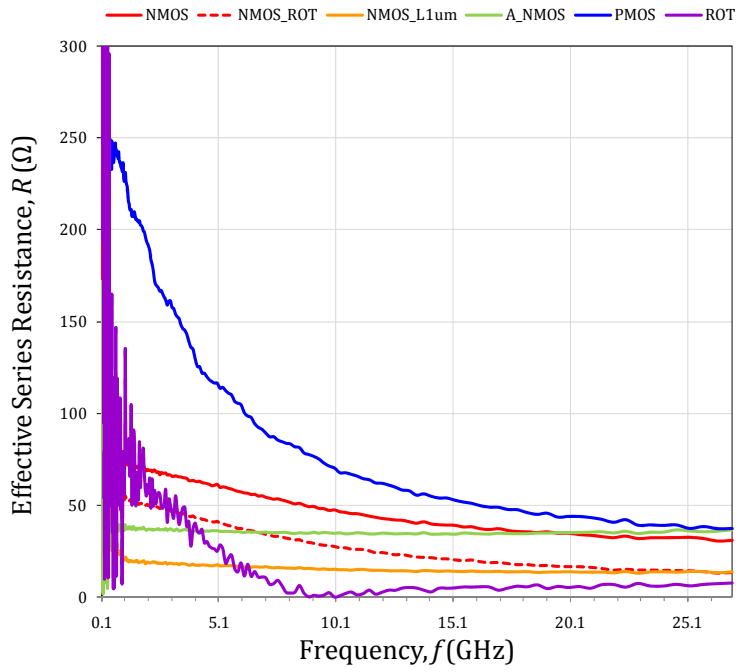


Figure 4.47 Measured equivalent series resistance for various decaps

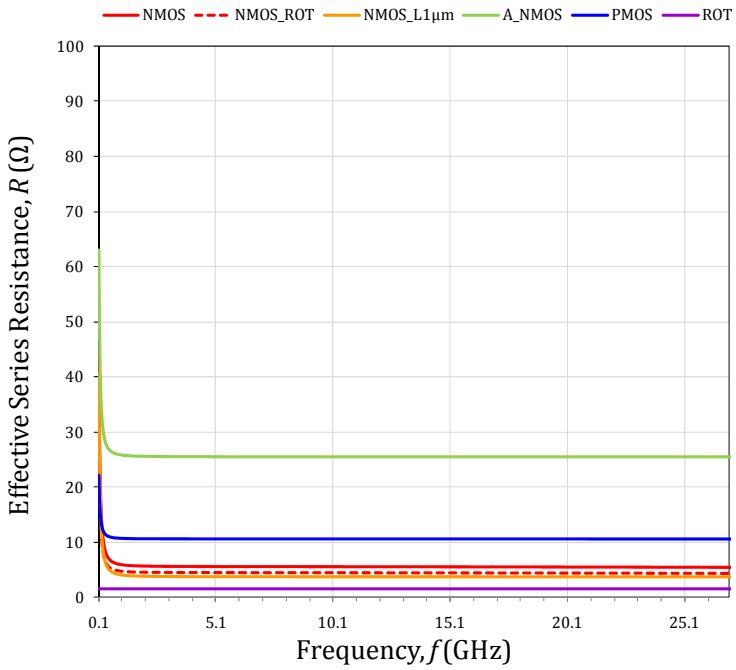


Figure 4.48 Simulated equivalent series resistance for various decaps

The gate leakage of the various fabricated structures was also measured with a full-swing bias (1 V) across each decap structure and a comparison of the measured versus simulated values is provided in Figure 4.49. As can be seen, while the magnitude of the measured versus simulated leakage currents vary slightly, the relative lower leakage current of the PMOS decap is confirmed by the measured results. The difference in leakage between NMOS and PMOS devices is due to the fact that the main tunneling component in a PMOS device in inversion is hole tunneling from the valence band, as opposed to electron tunneling from the conduction band, as opposed to electron tunneling from the conduction band in NMOS devices, which results in PMOS gate currents being smaller than NMOS devices [120].

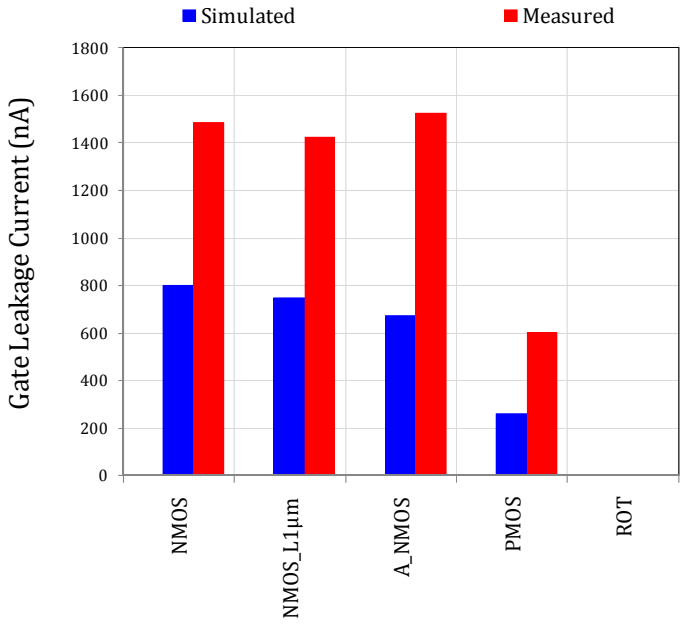


Figure 4.49 Simulated and measured gate leakage currents for various fabricated decaps

4.2.9 Decap Scaling Trends

In the previous section it was shown that both MOS-based structures and metal-based structures are desirable depending on various design constraints. The 65nm technology specifically was considered. This section looks at how MOS and metal-based decaps are affected by scaling of their various dimensions. The NMOS decaps is selected as a representative MOS-based structure and (post-layout) simulated across commercially available 180 nm, 90 nm, 65 nm and 45 nm CMOS processes. As was previously seen, the alternating (Figure 4.27 (b)) and rotative (Figure 4.27 (c)) multilayer interdigitated metal decaps both produce the most area efficient designs. The alternating structure is considered here for simplicity and also (post-layout) simulated across these technologies. Data from the 2010 ITRS [7] is further used to predict the capacitance of the decaps in newer and future technologies. CMOS production is currently well under way in the 28 nm technology node, and predictive models available down to the 16

nm node. While the ITRS no longer endorses the term ‘technology node’ it predicts device behavior over approximately the next decade down to transistor gate lengths of 8 nm.

4.2.9.1 MOS-Based Decaps

Figure 4.50 (a) and (b) illustrate the change in gate oxide thickness across selected existing technologies [128], and newer and future technologies as predicted by the ITRS [7], respectively. The gate oxide thickness shown is the *electrical oxide thickness* which is normalized to a relative dielectric permittivity of 3.9, the relative permittivity of silicon dioxide. A transition to a high-k gate dielectric is made in the 45 nm technology, thus the physical thickness of the dielectric is larger than previous technologies for the purpose of reducing leakage current, however, the high permittivity of the material results in an electrical oxide thickness that is comparable to its predecessor technologies.

As can be seen, there is a progressive decrease in electrical oxide thickness with technology, except for a slight initial increase where a change in device type is implemented. Thus the capacitance per unit decap *active area* is, in general, expected to increase with scaling as predicted in Chapter 3. Active area refers to the poly over gate oxide area only and excludes any additional contacts and routing that may be needed. These additional components add area to the overall decap structure. As will be shown in a subsequent section, the latter can have a significant impact on the total decap area and thus the effective area efficiency of the decap. As technology progresses, however this effect is expected to diminish as will be seen. An overall increase in the capacitance of MOS-based decaps can therefore be expected as technology continues to scale.

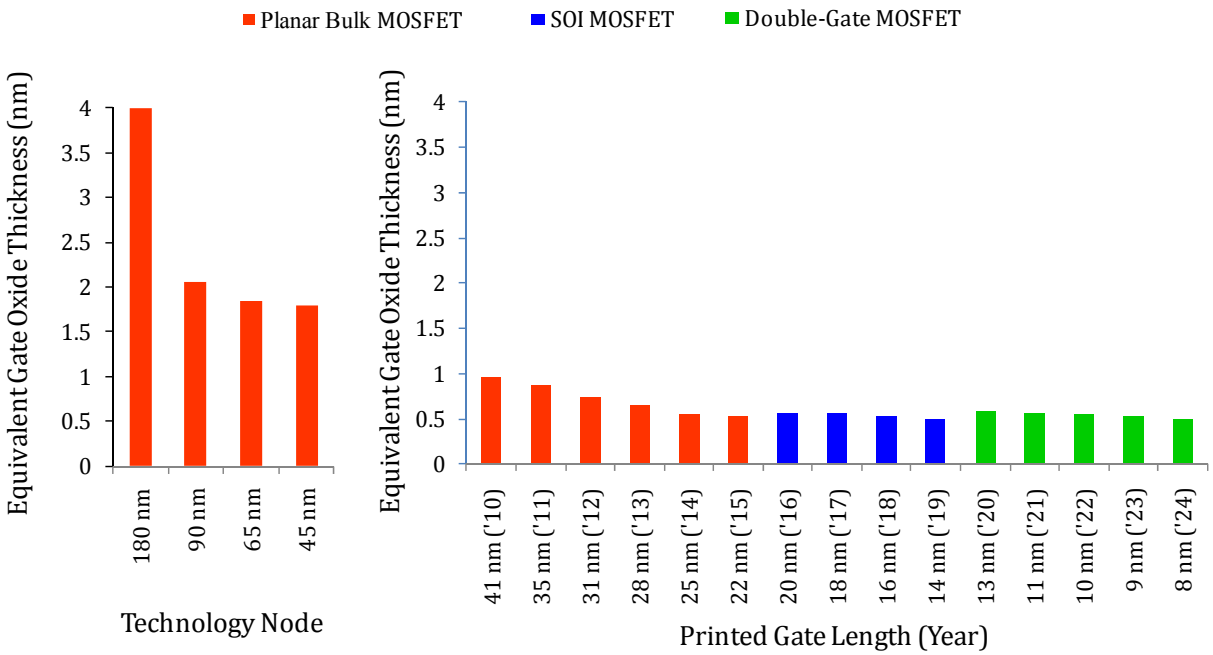


Figure 4.50 Electrical gate oxide thickness for (a) selected existing CMOS technologies, and (b) newer and future (data from [7]) CMOS technologies

As was seen earlier, another important parameter in considering decaps is the gate oxide leakage current and the thin gate oxide of MOS-based decaps cause these structures to suffer from relatively large leakage currents compared to other decap structures. Figure 4.51 (a) and (b) show the leakage current for MOS structures in selected existing, and newer and future technologies [7], respectively. A significant increase (note log scale) in leakage is seen in the 90 nm to 65 nm transition, a well known problem of the latter technology. An even more significant drop, however, is seen in the transition from the 65 nm to 45 nm technology due to the use of a high-k gate dielectric at the 45 nm technology node that enables the gate oxide thickness to be increased in order to reduce the transistor leakage currents. This sizeable reduction in leakage enables thin gate oxide MOS decaps to be used in post-65 nm technologies where thick oxide decaps would have been recommended in the past in power critical designs. As can be seen, the use of high-k gate dielectrics is expected to keep the leakage levels below the 65 nm levels over

approximately the next decade, however the undesirable increase with technology will make these decaps less and less attractive with scaling.

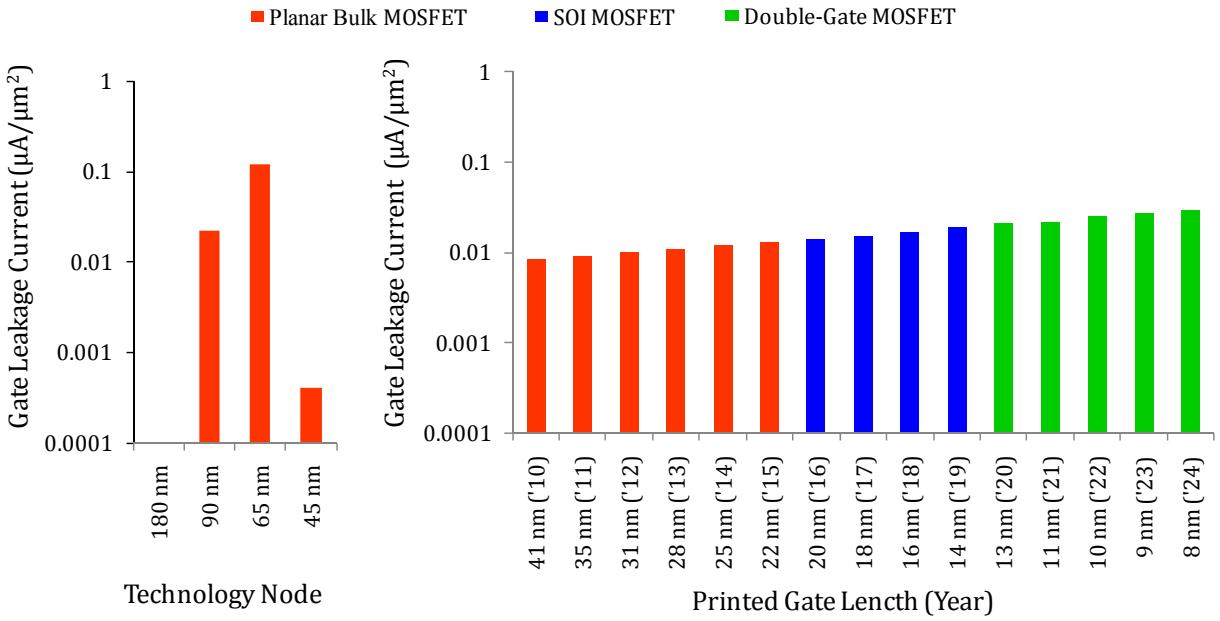


Figure 4.51 MOS leakage currents in (a) existing, and (b) newer and future (data from [7]) technologies

4.2.9.2 Multi-Layer Interdigitated Metal Decaps

With progressively smaller transistor gate lengths with evolving technologies, the back-end process similarly undergoes a corresponding scaling of dimensions which has an effect on the overall capacitance of these metal-based decaps. The total capacitance of an alternating multi-layer interdigitated metal decap, C_{metal}, per unit length, L, of each interdigitated trace, is given by

$$\frac{C_{metal}}{A} = (n-1)\cdot m \cdot \frac{C_{lateral}}{L} + (m-1)\cdot n \cdot \frac{C_{vetical}}{L} + \frac{C_{fringe}}{L} \tag{4-31}$$

Here n is the total number of metal traces in each metal layer, m is the total number of metal layers used in the decap structure, $C_{lateral}$ is the lateral capacitance between metal traces with different potentials,

$C_{vertical}$ is the vertical capacitance between two metal layers at different potentials and C_{fringe} any fringe capacitances between the various metal traces that may be present. The capacitances $C_{lateral}$ and $C_{vertical}$ are illustrated in the Metal 2-Metal 3 structure shown in Figure 4.52.

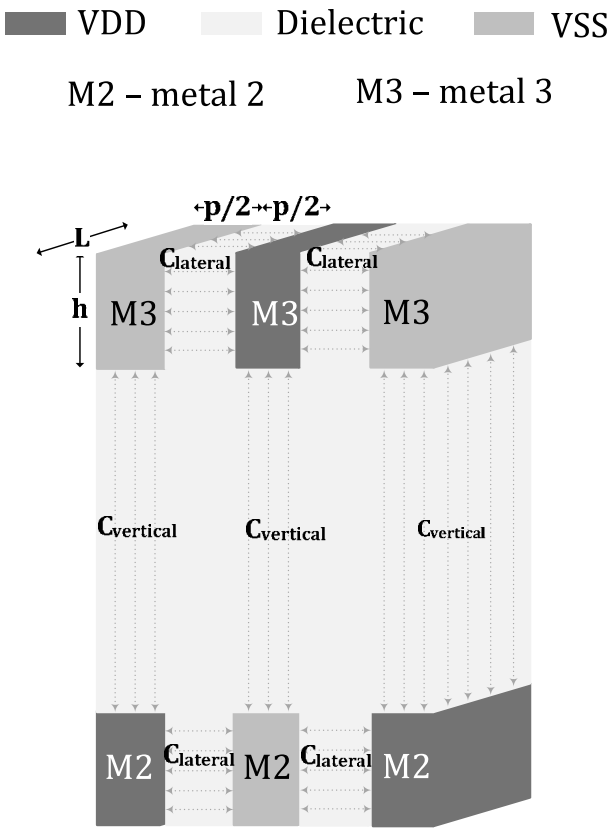


Figure 4.52 Inter-metal capacitances in multi-layer interdigitated metal decaps

The parallel plate capacitance of the *lateral* capacitance between two adjacent metal traces in terms of capacitance per unit length, L, is

$$\frac{C_{lateral}}{L} = \frac{\varepsilon \cdot h}{p/2} \tag{4-32}$$

where h is the thickness and p the pitch of the metal traces. As shown in Figure 4.52, the minimum metal width is typically equal to the half-pitch of the metal. The aspect ratio, AR, of the metal trace is thus given by

$$AR = \frac{h}{p/2} \tag{4-33}$$

and equation (9) can be re-written as

$$\frac{C_{lateral}}{L} = \varepsilon \cdot AR \tag{4-34}$$

Therefore, where the aspect ratio stays constant, any increase in lateral capacitance due to a decrease in pitch is negated by a corresponding decrease in metal thickness.

Figure 4.53 (a) to (d) show the trends in intermediate metal trace parameters for 180 nm, 90 nm, 65 nm and 45 nm technologies with parameters normalized to preserve confidentiality. Figure 4.54 (a) to (d) show the trends for similar parameters in newer and future technologies [7]. As can be seen in Figure 4.53 (a) and Figure 4.54 (a), the aspect ratio stays approximately steady with the normalized ratio remaining between 1 and 2 across the technologies shown.

The *vertical* capacitance per unit area shown in Figure 4.53 (b) for the selected existing technologies, can be seen to initially increase then decrease over the technologies shown, with significant differences seen in this parameter from technology to technology. It can be noted, however, that the physical dimensions result in the magnitude of $C_{vertical}$ per unit L_{metal} being approximately an order of magnitude smaller than the magnitude of $C_{lateral}$ per unit L_{metal}, where L_{metal} is the length of the metal line. The effect of the variation in $C_{vertical}$ is thus significantly muted in the overall capacitance of the metal decap. While this component of the capacitance was not available in the ITRS, Figure 4.54 (b) shows the total trace capacitance as two vertical, two lateral, and all fringe capacitances summed together. This trend desirably increases the overall capacitance of the metal decap structure.

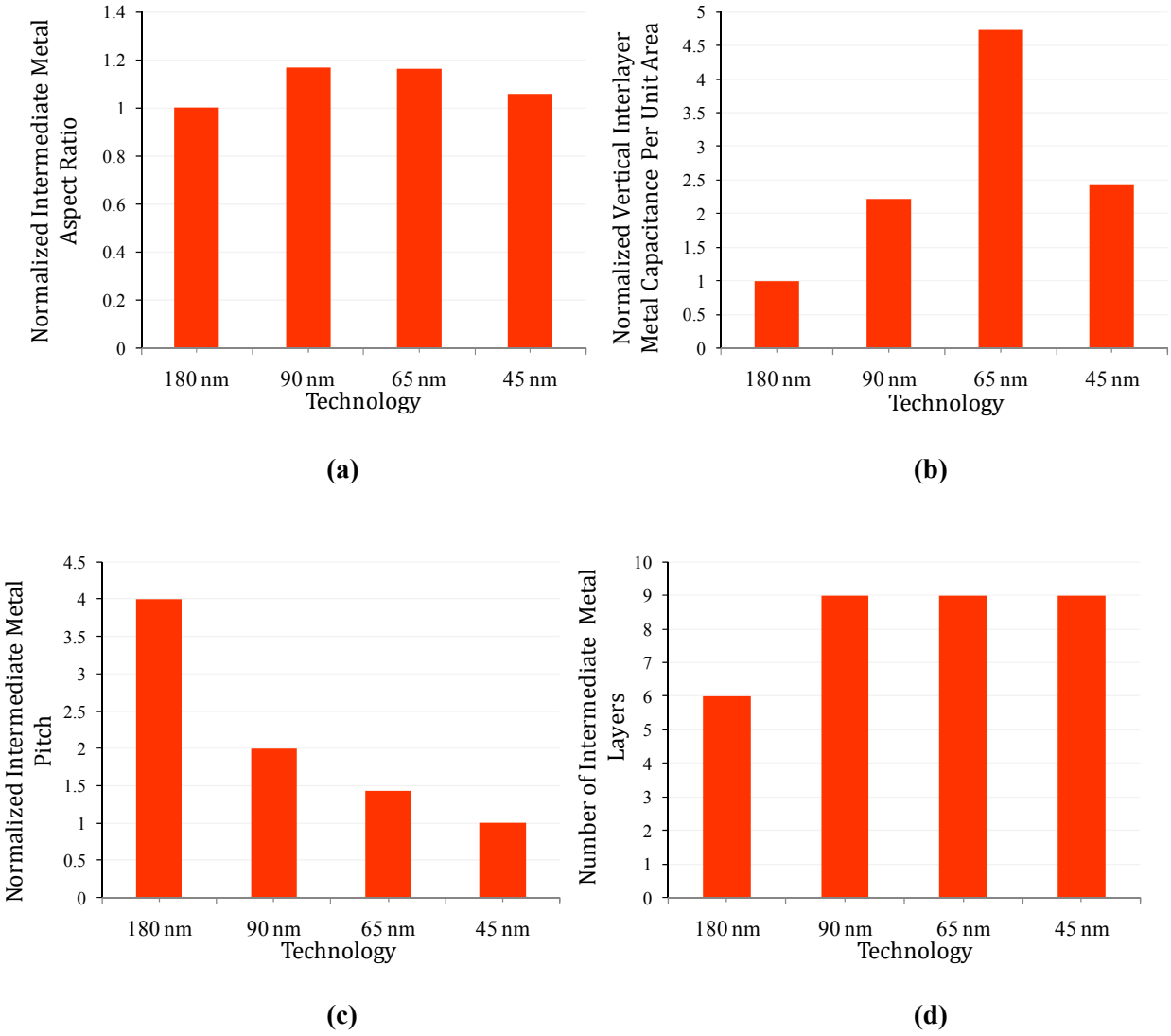


Figure 4.53 Various process parameters for intermediate metal layers in selected existing technologies

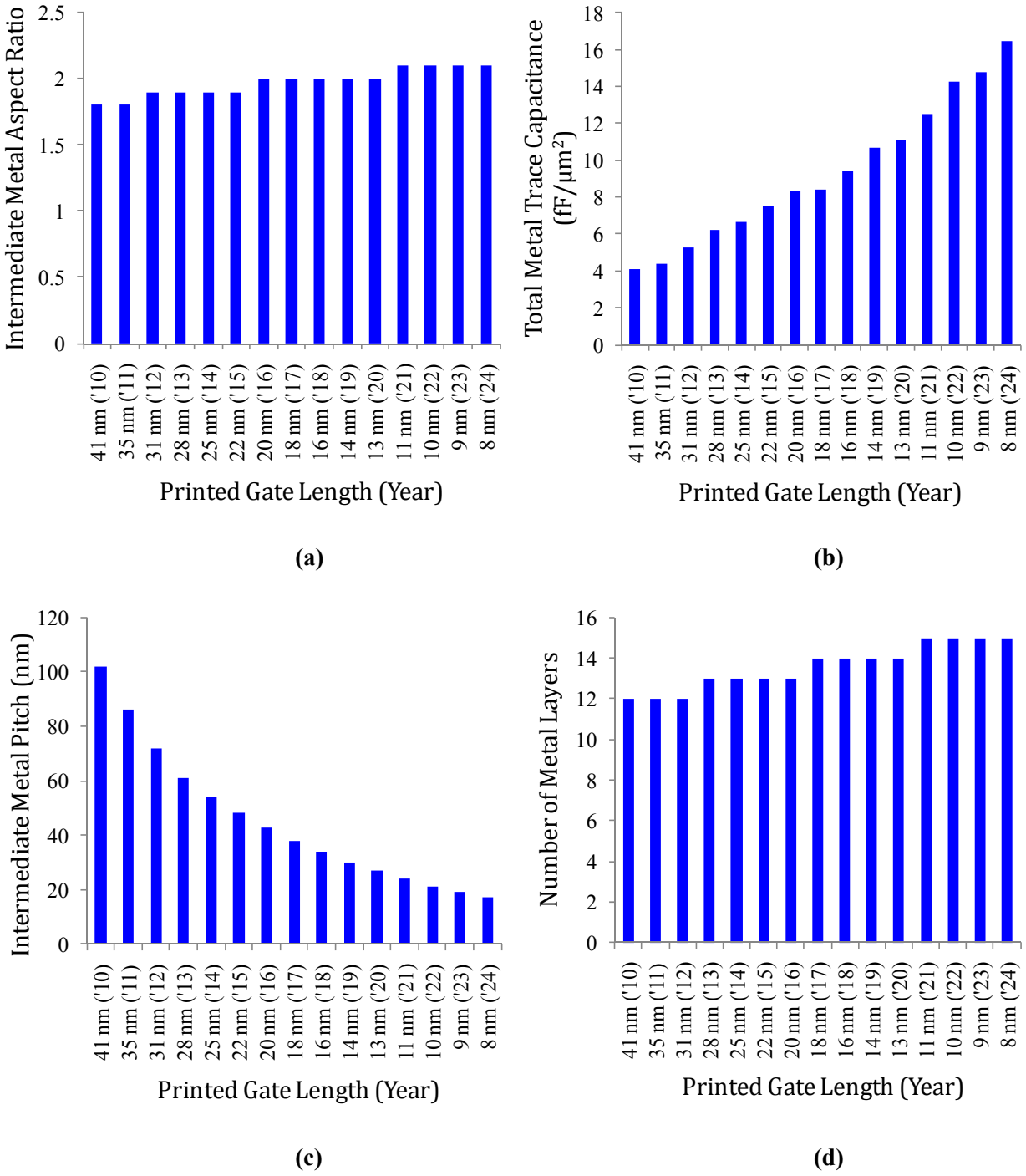


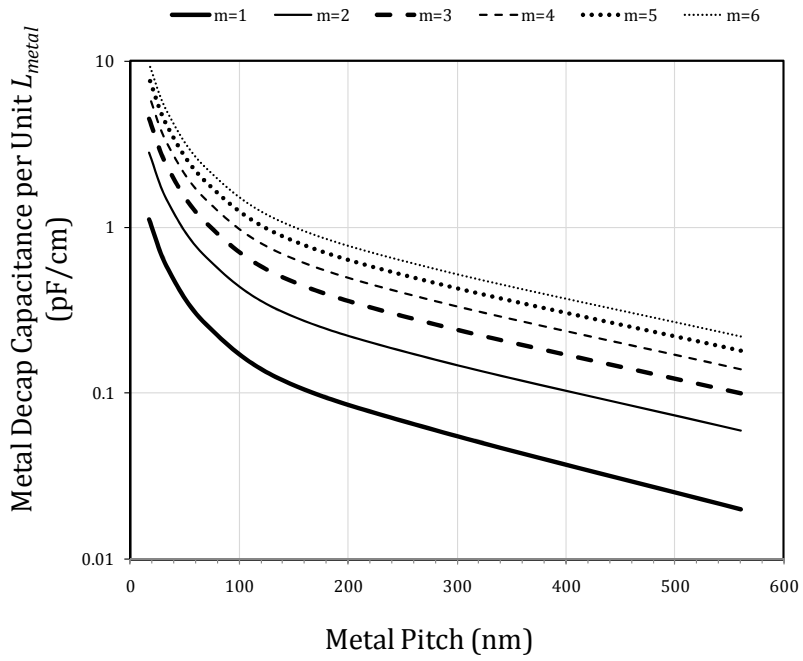
Figure 4.54 Scaling trends in intermediate metal interconnect parameters (data from [7])

The number of metal traces, n, is determined by the metal pitch. The smaller the pitch, the larger is the number of traces achievable in a given area. The number of metal traces is related to the pitch as

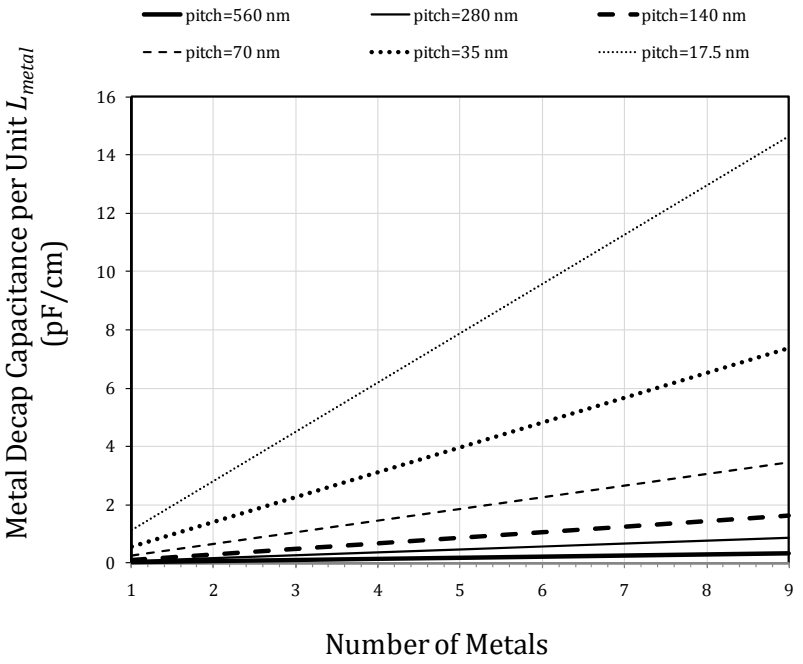
$$n = Quotient\left(\frac{W + p/2}{p}\right) \tag{4-35}$$

where, W is the width of the metal decap as shown in Figure 4.52. Since the pitch decreases with progressing technology as shown in Figure 4.53 (c) and Figure 4.54 (c), an increase in n and thus the overall capacitance per unit area of metal decaps can be expected with progressing technology. The total number of metal layers, m, further increases the overall capacitance according to (4-31). As shown in Figure 4.53 (d) and Figure 4.54 (d), the number of metals increases periodically over the technologies shown.

Neglecting C_{fringe}, and assuming constant typical values of $C_{lateral}$ and $C_{vertical}$, the total overall capacitance, C_{metal}, per unit L_{metal} is plotted for decaps with varying metal pitch, and number of metals, in Figure 4.55 (a) and (b), respectively. All decaps assume a width of 1 μm and use (4-35) to determine the total number of traces per layer in each decap. As shown in Figure 4.55 (a), the capacitance of the metal decap increases very rapidly as a power function of the metal pitch. As shown in Figure 4.55 (b), the overall metal decap capacitance is a linear function of the number of metal layers and further increases in metal decap capacitance. Metal-based decaps are therefore expected to become important sources of capacitance in future technologies.



(a)



(b)

Figure 4.55 Multi-layer interdigitated metal decap capacitance per unit *L* as a function of metal pitch and number of metals

4.2.9.3 NMOS Decaps

The layout simulated frequency response of C_{eff} for NMOS decaps in each of the 180 nm, 90 nm, 65 nm and 45 nm technologies is given in Figure 4.56. The effective capacitance decreases slightly with frequency. As described in [84], the effective capacitance of MOS based capacitors degrades with frequency due to the finite channel resistance of the devices.

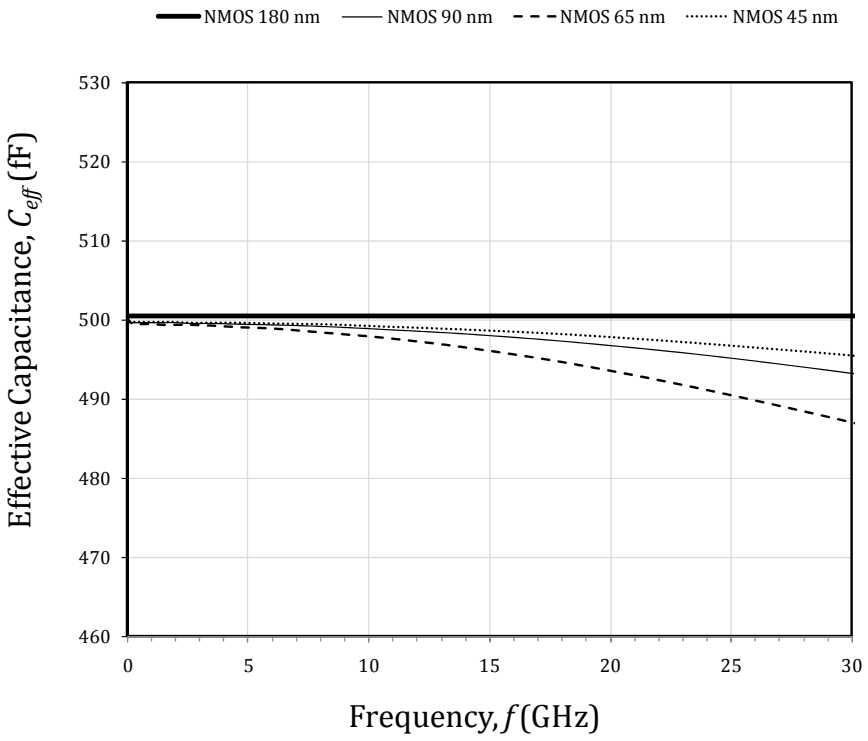


Figure 4.56 Frequency response of effective capacitance for NMOS decaps in selected existing technologies

The C_{eff} per unit total decap area (measured at 100 MHz) and the *ESR* are given in Figure 4.57 for the NMOS decaps in each of the technologies shown. Each NMOS decap was designed to provide a total C_{eff} of 500 fF (measured at 100 MHz) and laid out as a 6x6 array with transistor lengths of 1 μm and widths adjusted to obtain the target capacitance.

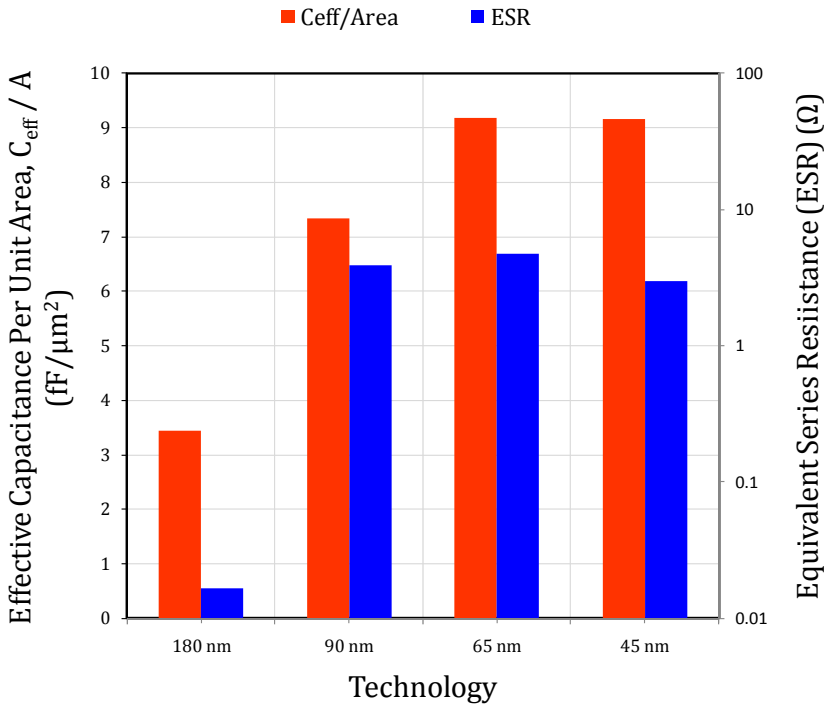


Figure 4.57 Effective capacitance per unit area and ESR of NMOS decaps in selected existing technologies.

Figure 4.58 illustrates the total decap area and the total active area of each 500 fF NMOS decap in each technology. The total area includes minimum spacing requirements between transistor fingers, contacts, and any routing that may be required. With progressing technologies, the design rules for these features scale correspondingly. Therefore, the reduction in decap area is both a function of an increase in capacitance of the MOS structure as well as a reduction in spacing, and metal and via dimensions. As can be seen in Figure 4.58, while there is a small change in active area, there is a larger change in the overall area of the decaps with technology. Also identified in the figure, is the area scaling factor, S_A, from technology to technology for comparison. As can be seen, the decap area does not scale at the same rate as other feature areas on a chip. Unless otherwise specified, 'area' refers to total decap area throughout this paper. As suggested earlier, the area attributed to routing, vias and spacing rules is seen to become

less important with scaling and the effective capacitance per unit area stays approximately constant in the 65 nm to 45 nm transition closely following the trend of equivalent gate oxide thickness.

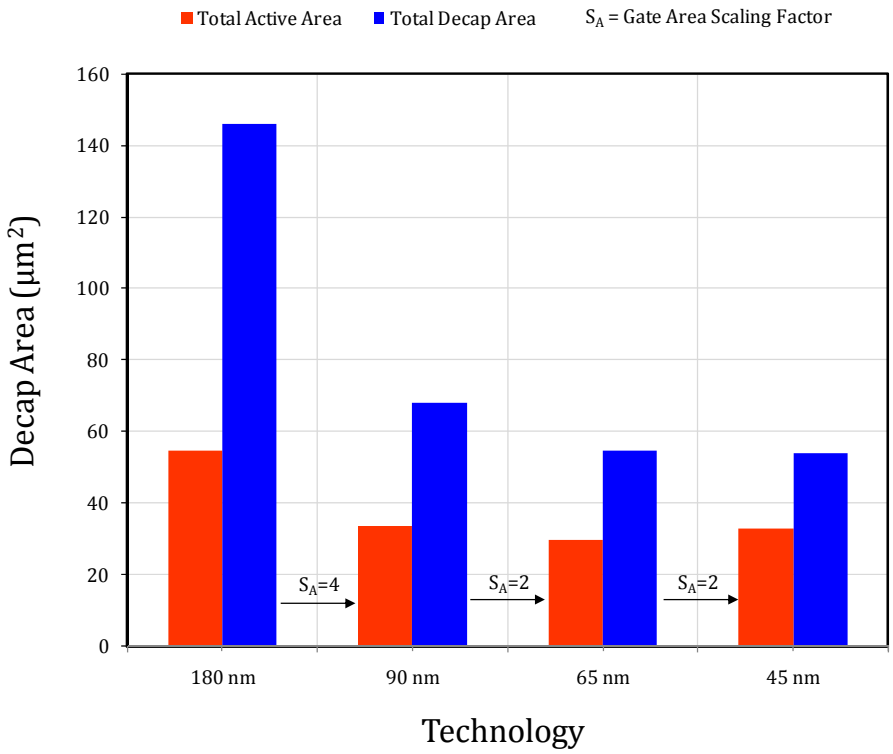


Figure 4.58 Area of NMOS decaps in various process technologies

The overall impedance of each decap directly determines the extent of supply noise suppression that it provides. The impedance magnitude normalized to an area of 1 μm^2 is further plotted in Figure 4.59 for each of the decap configurations considered. Numerical values of the normalized impedance magnitude are further provided for convenience in Table 4-XII at approximately 100 MHz, 1 GHz, 10 GHz and 30 GHz. As can be seen, the reduction in impedance becomes less significant with progressing technology. Although the C_{eff} is approximately constant in the 65 nm to 45 nm technologies, the slight decrease in impedance seen is due to the slight reduction in *ESR* over this technology transition.

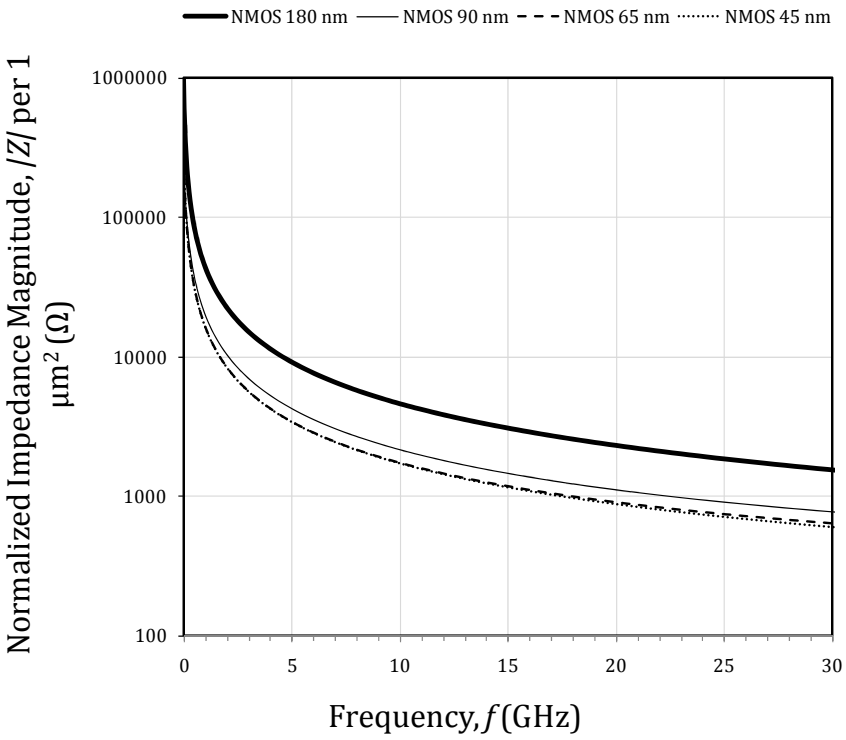


Figure 4.59 Impedance magnitude of NMOScaps in various existing CMOS technologies

TABLE 4-XII NUMERICAL VALUES OF NORMALIZED IMPEDANCE MAGNITUDE FOR NMOS DECAPS IN SELECTED EXISTING TECHNOLOGIES

TECHNOLOGY	NORMALIZED |Z|			
	100 MHz (kΩ)	1 GHz (kΩ)	10 GHz (kΩ)	30 GHz (kΩ)
180 nm	463	50.9	4.23	1.40
90 nm	217	23.8	2.00	0.74
65 nm	174	19.1	1.61	0.61
45 nm	172	18.8	1.58	0.55

Figure 4.60 shows the dielectric leakage currents per unit area measured at DC for the decaps in each technology considered. While these values are determined using the total decap area the trend observed in Figure 4.51 (a) is closely followed.

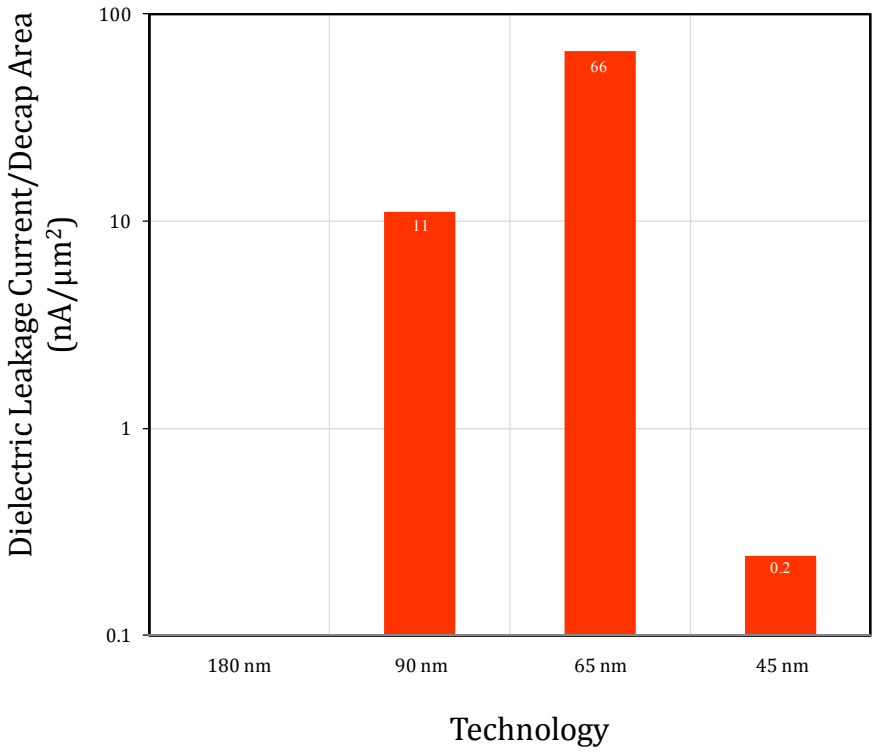


Figure 4.60 Dielectric leakage current per unit area for NMOS decaps in selected existing technologies

4.2.9.4 Multi-Layer Interdigitated Metal Decaps

The rotative multi-layer interdigitated metal decap was similarly simulated in each of the 180 nm, 90 nm, 65 nm and 45 nm technologies using the reliability recommended design rules. Only the 'intermediate' metal layers were used in these structures since, again, Metal 1 is used for local transistor routing and the higher, thicker metals are typically reserved for long range routing. The 45 nm technology available had a limited number of metal layers available and the capacitance values were extrapolated for a typical process. Figure 4.61 illustrates the C_{eff} per unit area and *ESR* for the metal decaps in each technology considered. As can be seen, the effective capacitance per unit area desirably increases with technology scaling as expected from the theoretical analysis presented in section 4.2.9.2. It can be noted

that the *ESR* value for the 45 nm technology provide an upper bound since the reported value is for a reduced number of metal layers available (5 available as opposed to the typical 7 intermediate metals).

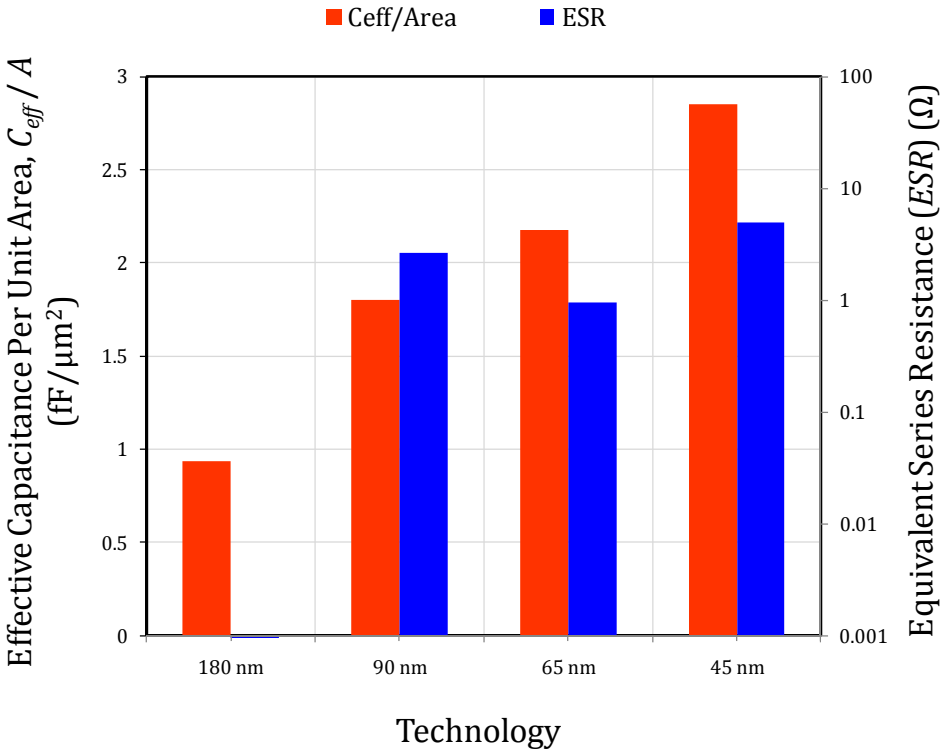


Figure 4.61 Effective capacitance per unit area for rotative multi-layer interdigitated metal decaps in selected existing technologies

4.2.9.5 Hybrid Decaps

As mentioned, it is desirable to physically stack multi-layer interdigitated metal decaps on top of MOS-based decaps in order to enhance the amount of capacitance in a given area. The total capacitance obtained is the approximate sum of each individual decap since the two decaps are essentially electrically placed in parallel with each other.

Figure 4.62 illustrates the C_{eff} per unit area of each hybrid decap simulated in the given technologies. The NMOS decap values are identified for comparison. As was observed earlier, an increase in both the

effective capacitance of the NMOS decap and metal decap is seen. The percentage enhancement in capacitance, however, varies depending on the rate of increase of each capacitance compared to the previous technology.

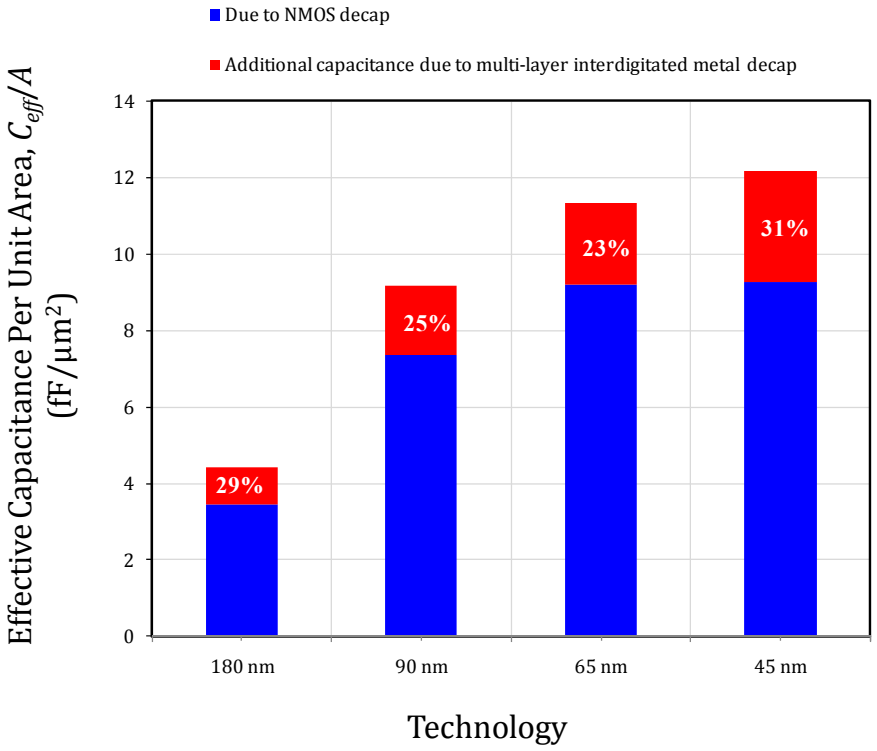


Figure 4.62 Effective capacitance per unit area for hybrid decaps in selected existing technologies with contributions from individual components identified

The effective capacitance for MOS decaps and multi-layer interdigitated metal decaps is calculated based on ITRS data [7] and shown in Figure 4.63. The values for the MOS decaps are determined using the active area and the difference between this area and the total decap area is assumed to be negligible. The values for the metal decaps are obtained using (4-31).

Figure 4.64 further shows the relative contributions of each decap in the overall hybrid structure. As in the case of the simulated decaps, the relative contribution of the metal decap increases only where the slope of the metal decap line in Figure 4.63 is steeper than that of the MOS decap line.

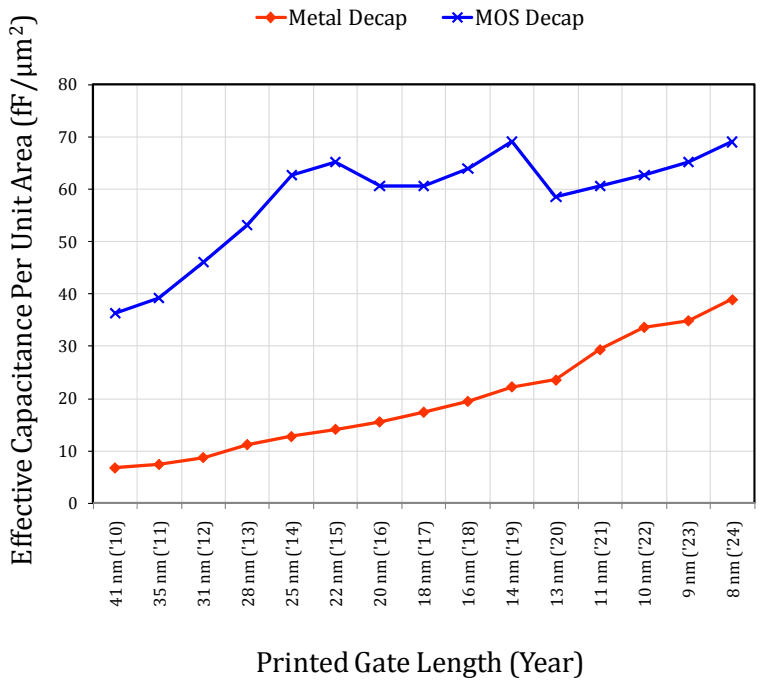


Figure 4.63 Effective capacitance per unit area of MOS decaps and metal decaps

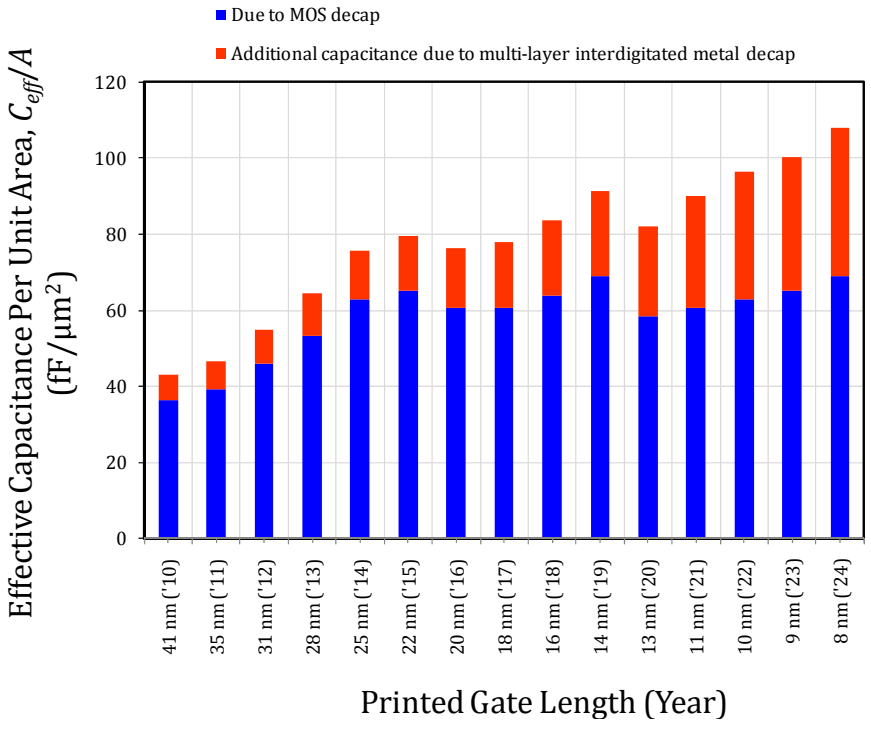


Figure 4.64 Relative contributions of metal and MOS decaps to overall hybrid decap structures

The relative enhancements in capacitance with the hybrid structures are extracted and plotted in Figure 4.65 where almost a 40% increase in relative contribution is expected over approximately the next decade. An approximately 55% capacitance enhancement is predicted with a MOS-metal hybrid decap compared to a MOS decap alone by the 8 nm gate length technology. Metal decaps are thus expected to become increasingly important with technology scaling and are important structures in enhancing the capacitance per unit area of traditionally used MOS decaps.

Figure 4.63 shows the simulated *ESR* values of the hybrid decaps compared to the NMOS decaps in each technology shown. Placing the NMOS decap and metal decap in parallel has the effect of reducing the overall *ESR* in the hybrid structure as these theoretical resistors in each structure are essentially placed in parallel with each other. Thus, the increase in capacitance and reduction in *ESR* both have desirable effects on the overall impedance of the hybrid decap. Table 4-XIII provides selected impedance values of the simulated hybrid decaps in each technology shown. An improvement in the values relative to the NMOS only values can be observed.

TABLE 4-XIII NUMERICAL VALUES OF NORMALIZED IMPEDANCE MAGNITUDE FOR HYBRID DECAPS IN SELECTED EXISTING TECHNOLOGIES

TECHNOLOGY	NORMALIZED |Z|			
	100 MHz (kΩ)	1 GHz (kΩ)	10 GHz (kΩ)	30 GHz (kΩ)
180 nm	360	39.5	3.28	1.09
90 nm	174	19.1	1.61	0.58
65 nm	141	15.4	1.30	0.47
45 nm	128	14.0	1.17	0.41

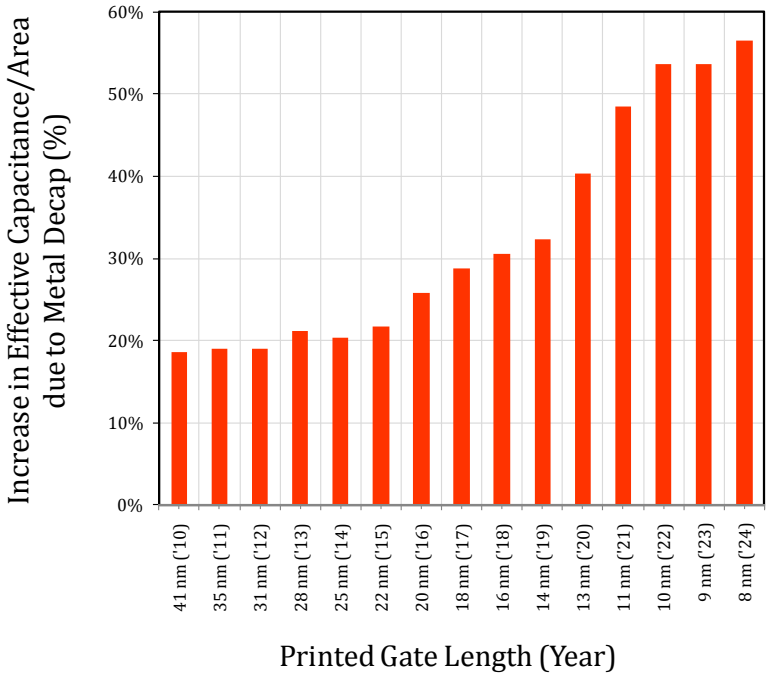


Figure 4.65 Percentage increase in effective capacitance per unit area due to placing a metal decap in parallel with a MOS decap in a hyrid decap structure

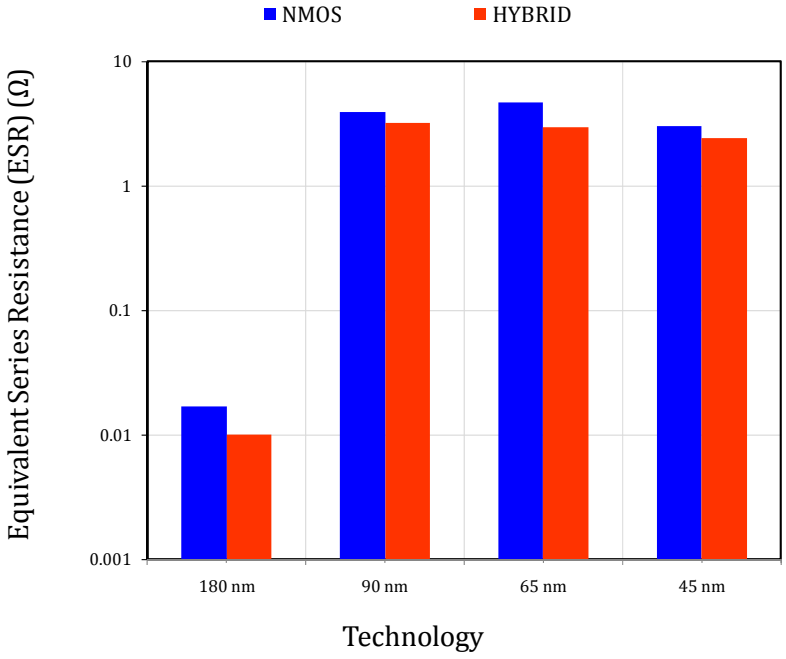


Figure 4.66 Effective series resistance of hybrid decaps in selected existing technologies with data for NMOS decaps included for comparison

4.2.10 Conclusions

Various decap structures are characterized in a commercial 65 nm CMOS process, and a set of guidelines provided to enable the most optimal decap design selection for given chip constraints. Standard decap structures including MOS-based structures and metal-based structures are studied. MIM decaps are also studied as they are a common process option. Post-layout simulations show that NMOS decaps are the most area efficient decap type with respect to overall decap capacitance and impedance, while PMOS and accumulation mode NMOS decaps provide a good compromise between area and leakage currents. The low threshold voltage NMOS decap further provides slightly improved admittance characteristics at high (> ~20 GHz) frequencies. It is further shown that multi-layer metal decap structures in combination with MOS-based structures provide a hybrid structure that is an attractive alternative to traditionally used structures. For example, an NMOS+RTMOM hybrid decap has ~25% greater capacitance than an NMOS decap. The multi-layer metal decap structures themselves can also be implemented in areas of the chip with existing circuitry in replacement of existing metal-fill patterns. Optimal sizing for MOS-based decaps is also investigated. It is shown that a gate length of ~ 3 μm is the most optimal choice in the given technology studied.

Chip measurements confirm the improvement in capacitance predicted for a 90 nm technology. As expected, the core BSIM models insufficiently model the high frequency behavior of MOS devices. Based on the observed results, care should therefore be taken to minimize MOS decap gate lengths where possible within practical area constaints.

The effect of technology scaling on MOS-based decaps and multilayer metal decaps was also investigated over a wide range of technologies (180 nm to 8nm). It was shown that the multilayer metal decap structures will become increasingly important with technology scaling and correspondingly so will

the hybrid structure. The increase in capacitance provided by the hybrid structure compared to a traditional NMOS decap is expected to rise to ~55% by the 8 nm gate length technology.

Chapter 5
A Further Look at Supply Noise Management

"...more than 20% of the total die area has been occupied by [passive] decaps leading to a significant waste of active die area"

…J. Gu, R. Harjani and C. H. Kim (University of Minnesota)

In this chapter a method of managing supply noise is proposed based on shifting currents within the clock cycle. The method is demonstrated by implementation of the technique in an ALU. An analysis highlighting the limitations of the switched capacitor active noise suppression technique is further provided.

5.1 Noise Management within a Clock Cycle

Power supply noise, as was seen earlier, is managed through a combination of techniques. One of these techniques, briefly introduced previously, is by means of current spreading. Essentially, in this technique, the instantaneous current demand is spread out over a finite period of time in order to the reduce the peak current experienced at any given moment. With this method, the supply noise generation itself is suppressed, rather than countering noise that is already generated as in the case of decaps, for example.

The supply noise management technique proposed is applicable to a level of circuit abstraction where there are parallel evaluations with varying latencies executing in parallel. An arithmetic and logic unit (ALU) which forms part of a datapath in a microprocessor [121][122], is a typical example of such a circuit block and is used to demonstrate the proposed technique. In the technique, the switching of non-critical path circuits is shifted to a later time during the same clock cycle by delaying the input vectors to these circuits.

5.1.1 Prior Work

Designers, in the past, have been cognizant of the benefits of distributing switching currents. For example, in [66], variable delays are introduced in high-speed memory I/O interfaces to reduce cross-talk induced jitter. Cross-talk is a significant problem in memories due to their finite area requirements and staggering of the I/O transitions have shown a reduction in this cross-talk. In addition, in [67], a power gating structure that employs a technique where one set of sleep transistors is turned on one half of the resonant oscillation frequency later than another set of sleep transistors in order to suppress the supply noise fluctuation during mode transition. Furthermore, in [68], a staggering approach is used to perform asynchronous set/reset operations in order to reduce supply fluctuations and any related unwanted interactions between independent stages during a partial set/reset, for example. Other staggering techniques have also been presented in the literature [123]-[125], however none of these techniques propose current shifting at the level of abstraction proposed here. In [124], clock scheduling is used to control the input to various combinational logic blocks to minimize simultaneous switching, however this optimizes the switching at the clock level by skewing the time at which the clock arrives at a given set of flip-flops. In our work, we introduce a delay element after the data has been latched by the flip-flops and within the clock evaluation cycle and our proposed method can thus be used in combination with clock optimization for further reductions in supply noise.

5.1.2 Noise Management in an ALU using Current Shifting

5.1.2.1 ALU Design

An ALU implemented in a 65 nm CMOS technology is used to demonstrate the current shifting technique presented in this work. The 64-bit ALU design is a typical design [121] and is illustrated in Figure 5.1. It comprises of an adder block, a logic unit consisting of AND, OR and XOR gates, and a 5-bit shifter. The ALU is designed using static logic since this logic style is becoming more prevalent with progressing technologies due to the need for robust designs and the move to multi-core computer architectures allowing the use of lower clock frequencies. The adder block was designed using the Kogge-Stone architecture illustrated in Figure 5.2. For simplicity, the instruction bits were hard coded in. As can be seen, the addition and logic/shift operations are performed in parallel and the required output selected using a multiplexer. While it may seem wasteful to evaluate all operations simultaneously when only a single output is selected by the multiplexor, ALUs generally employ this design since activating/deactivating the different blocks is itself relatively resource intensive. Especially in the case where static logic is used, for which the activity factors are relatively low (10-20%) [126].

In order to emulate a representative switching pattern on a typical chip, the test input bit stream was selected such as to provide a repeating series of low power input vectors followed by a maximum power input vector while maintaining the average data activity close to the characteristic levels for static logic. A high frequency decoupling capacitor (decap) model was also placed in parallel with the circuit to bring the supply noise within ~20% of the nominal level. An additional high capacitance model was further placed in parallel to shift the resonance frequency to the range typically seen on complete chips (low hundreds of MHz) [17][39]. The package parasitics were modeled using an ideal inductance and resistance along the supply path.

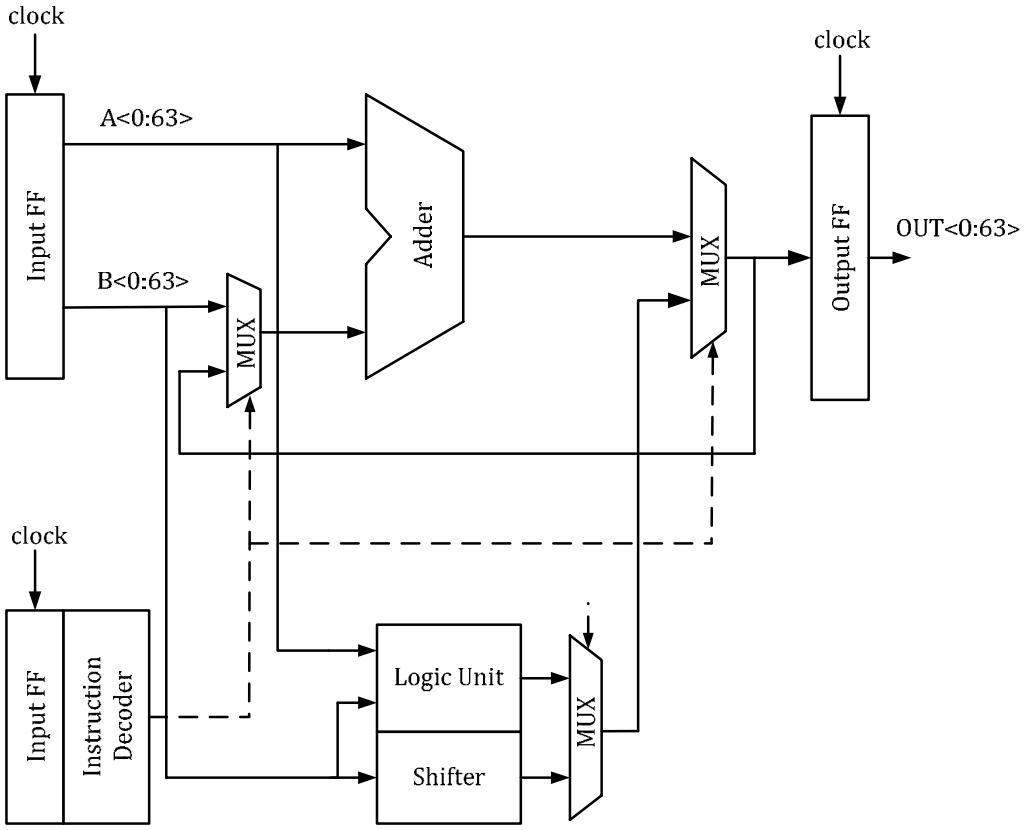


Figure 5.1 64-Bit ALU block diagram

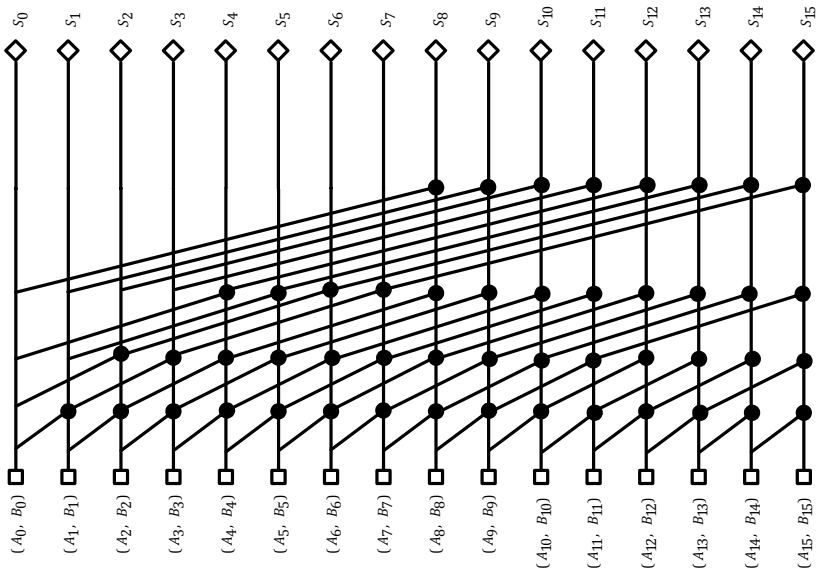


Figure 5.2 64-bit Kogge-Stone adder architecture

Figure 5.3 illustrates the ideal current draw pattern (i.e. at a constant 1 V supply voltage) and the corresponding power supply voltage waveform and clock pattern. The data is latched into the flip-flops at a rate of 2 GHz. The peaks in current identify the instances of high power input vectors. Again, this sequence of low power input vectors followed by a high power input vector is used to emulate a typical input bit stream. As is characteristic for a single-clock edge-triggered clocking style, the largest current peaks within the clock cycle occur shortly following the clock edges [125]. As expected and can be observed, the peaks in current demand correspond to the largest dips in the power supply voltage level. Resonance of relatively high frequency can also be observed between one high power input vector and the next. Input patterns such as the one used where the inputs alternate between high and low power, can also stimulate low frequency resonance peaks, so a relatively long simulation was performed to ensure any low frequency resonance was sufficiently damped. A relatively small variation in the minimum supply voltage from one high input pattern to another was seen. Nevertheless, the second high power input pattern was found to exhibit the largest dip and the supply voltage dip, ΔV_{min}, at this edge was measured in the evaluation of the proposed method.

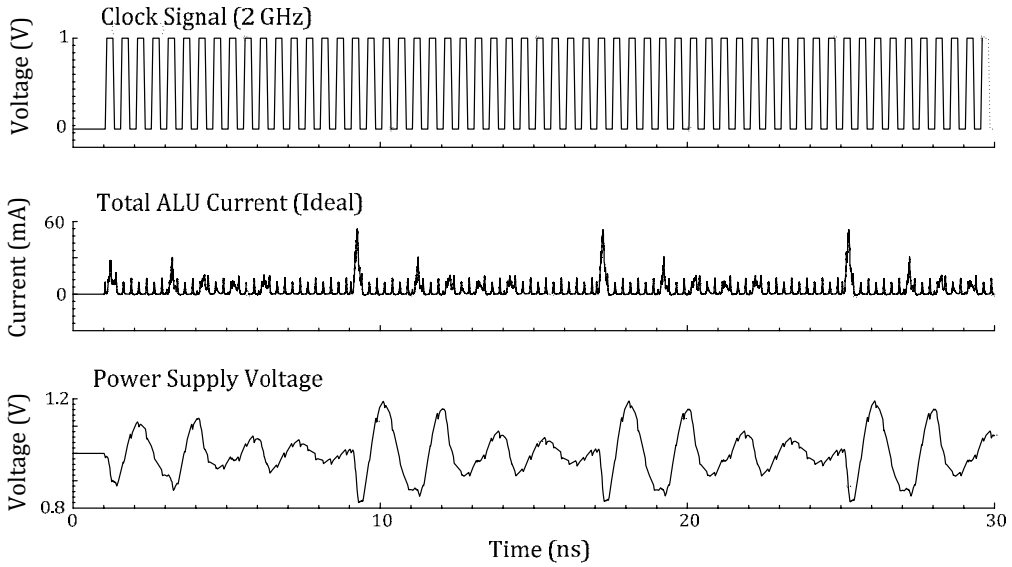


Figure 5.3 64-bit ALU current consumption pattern and supply voltage waveform

5.1.2.2 Current Redistribution Within the Clock Cycle

Since the logic unit and shifter blocks have a smaller delay path compared to the adder block, evaluation of the logic/shifter blocks can be delayed to later within the given clock cycle. The current demand of these blocks can therefore be shifted away from the clock edge thus reducing the overall current demands that immediately follow the clock edge. Again, the proposed method is intended for use in addition to known clock skewing methods of supply glitch minimization [125], since the delay is introduced after the clock edge and within the clock period without the need of additional data storage elements.

One known method that can be applied to move the current consumption of the logic/shifter block away from the clock edge is to use an approach similar to data gating [127]. Here an enable signal would be used to delay the evaluation of the non-critical blocks. However, what we propose is the use of a delay element directly in the path of the input data. This provides a simpler and more elegant solution since this implements a hard shift in the evaluation of the non-critical blocks and a dedicated enable signal is not necessary. Gating requires a delay element from the existing clock edge in addition to an enable transistor, whereas placing the delay element directly in the data path eliminates the need for the enable transistor (as well as the circuitry needed to generate the enable signal). Furthermore, where the delay is inserted directly in the data input path, the delay element switches, and thus consumes power, only when there is a change in data. In the case of gating based on the clock, the delay element would switch and consume power every clock cycle. In the proposed method, the input data to the logic/shifter blocks is thus delayed by inserting a delay element (buffer) between each input data bit and the logic/shifter blocks as illustrated in Figure 5.4, such that the adder and logic/shifter blocks no longer begin evaluating simultaneously.

Figure 5.5 illustrates the ideal relative current consumption patterns for the adder block, the logic/shifter blocks and the flip-flops/other circuitry. As can be seen and as expected, the current peak for the logic/shifter blocks approximately coincides with the current peak of the adder. The peak in current demand of the flip-flops occurs when the input data pattern is latched in which is slightly before the adder/log/shifter peak current demands occur. The adder is seen to exhibit the highest peak in current consumption. Adding a delay in the input data path of the logic/shifter unit, results in the shifted logic/shifter current curve which is observed to be moved by approximately 100 ps. Here, a relatively weak 2-stage buffer is used to realize the delay. After shifting, the peak in logic/shifter current now occurs at a time when all other current demands are relatively low. Judging from the area under the two logic/shifter current curves, a slight increase in power consumption is expected. Shown in Figure 5.6, are the corresponding supply voltage waveforms for both the standard ALU and the ALU with the data delay element included for current shifting. As can be seen delaying the evaluation of the logic/shifter blocks suppresses the dip seen in the power supply voltage.

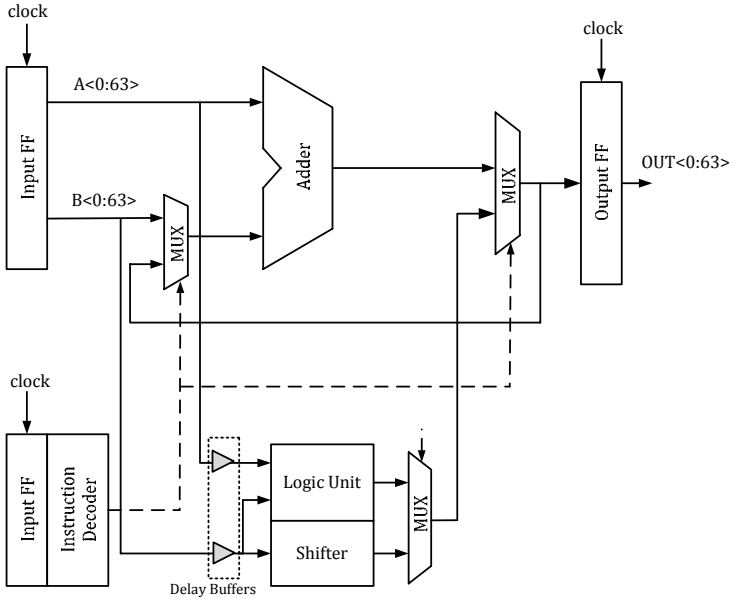


Figure 5.4 ALU with delay buffer

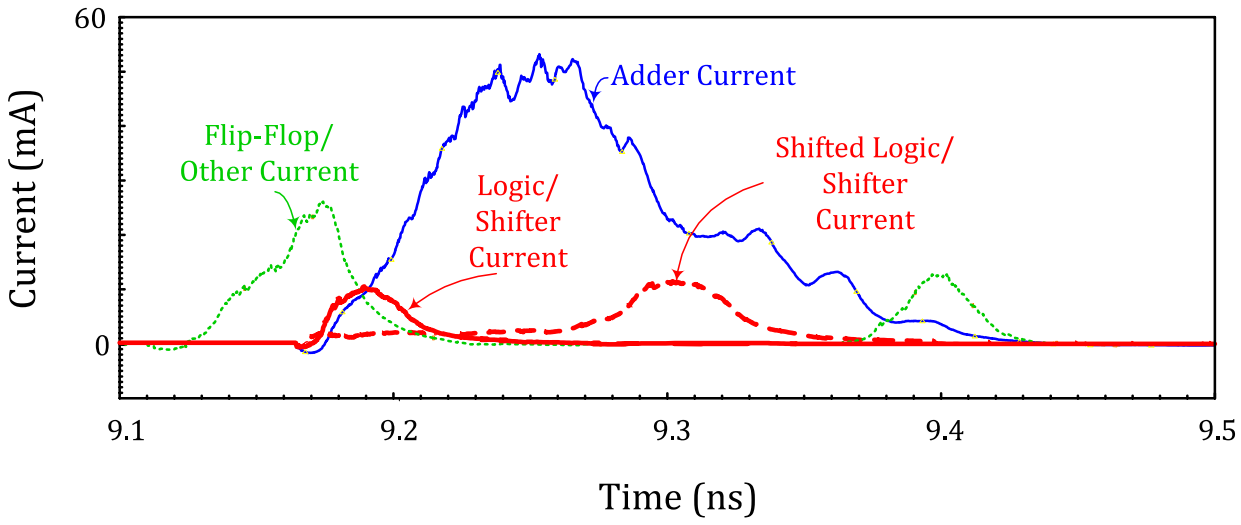


Figure 5.5 Relative current consumption of various blocks within the ALU with and without current shifting

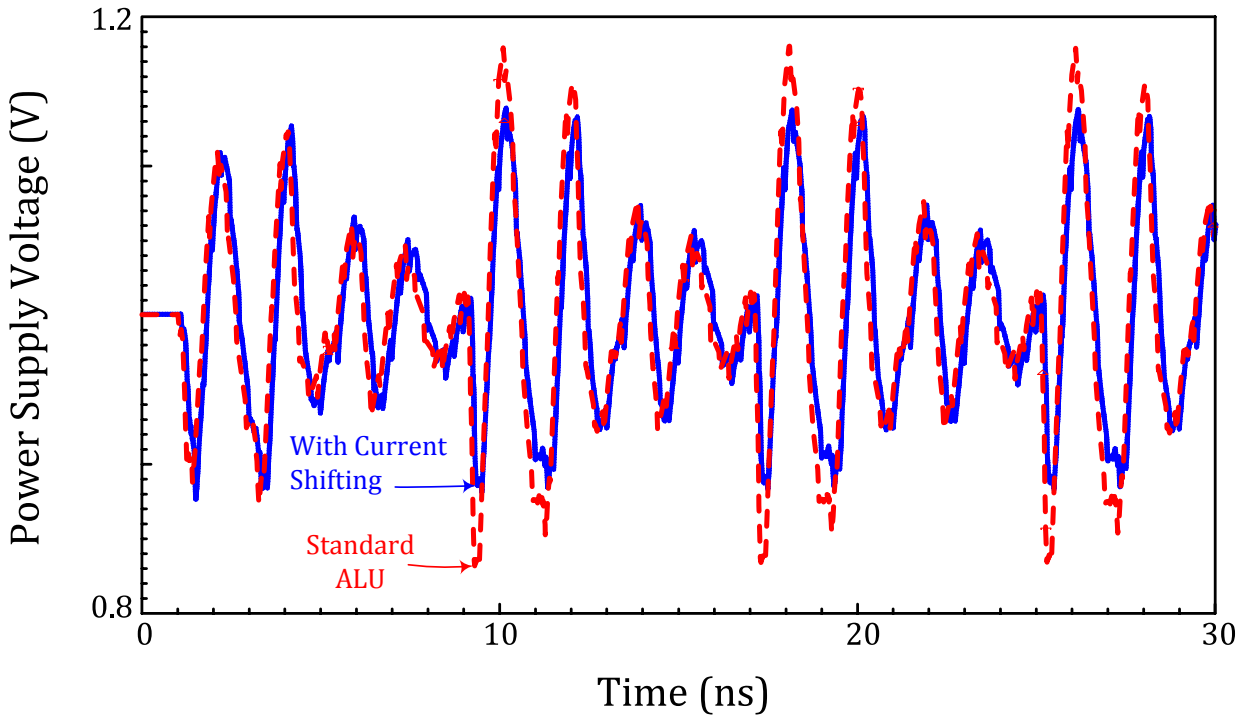


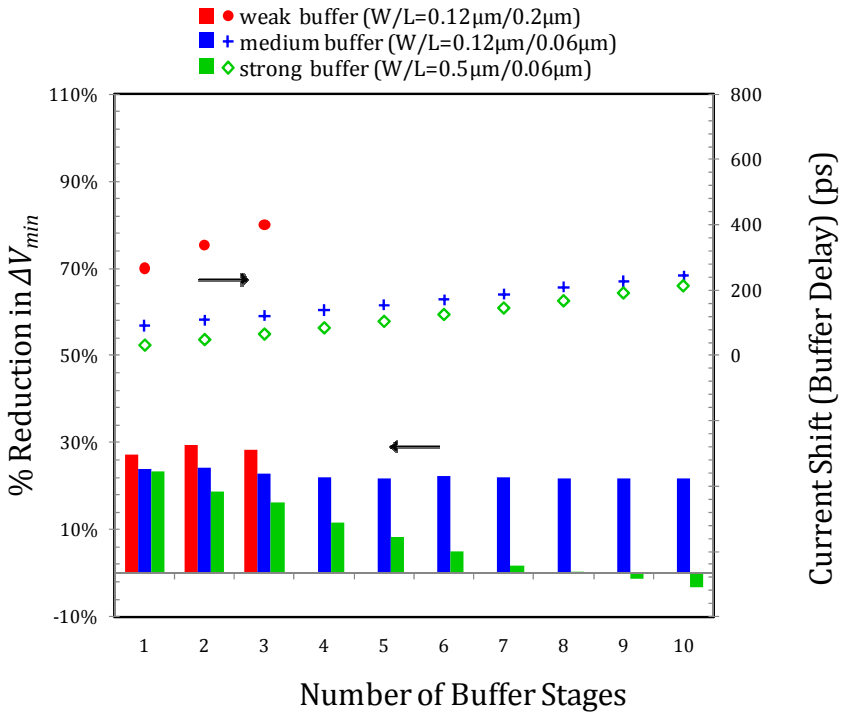
Figure 5.6 Supply voltage waveforms for ALU with and without current shifting

5.1.2.3 Delay Element Optimization

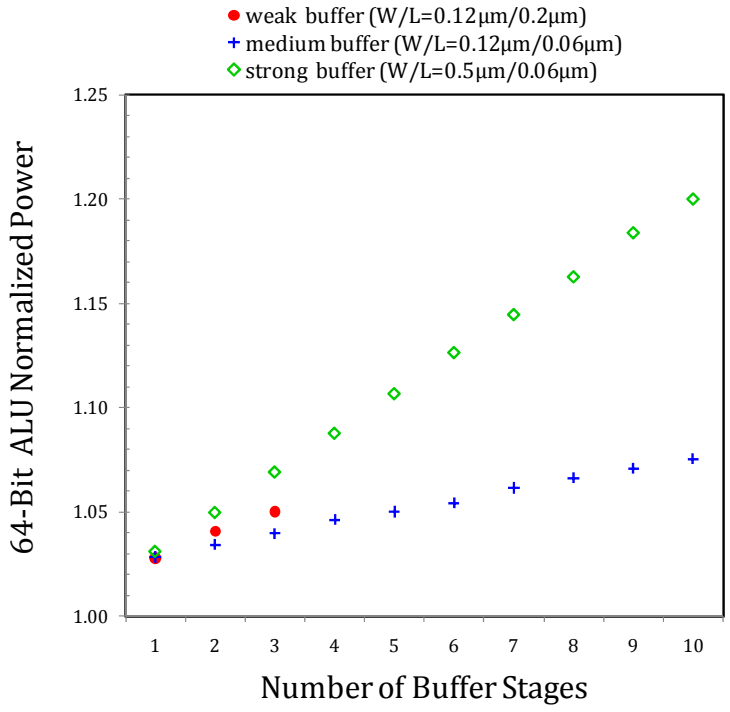
Ideally, the current demands of the logic/shifter block should be shifted as far away from the adder current peak within the clock period as possible. However, how and the duration that the logic/shifter block current peak is pushed has related power consumption and area implications. Very weak buffers can provide larger delays in logic/shifter evaluation, however, can have large short circuit (crowbar) currents. Strong buffers, on the other hand, minimize the short circuit power however provide smaller delays necessitating thus necessitating additional stages for a given shift in current. Therefore, there are clearly trade-offs between the duration of the logic/shifter block current shift, power consumption,, area overhead and supply noise suppression.

Several buffers were designed with a varying number of stages, and transistor gate widths (W) and lengths (L). the last inverting stage was held at a constant dimension in order to provide the logic/shifter blocks with a consistent input signal rise time. Figure 5.7 (a) illustrates the supply voltage dip (ΔV_{min}) and duration of current shift (buffer delay) for the various buffers used. The corresponding normalized ALU power consumption and area overhead are also given in Figure 5.7 (b) and (c), respectively. The related adder latency and average supply voltage levels for each design were observed to remain relatively constant in all the various designs.

As can be seen from the plots in Figure 5.7, the weaker buffers provide larger delays at some cost to area and power. The level of noise suppression provided by the strong buffer designs decreases rapidly with an increase in the number of stages due to the associated increase in current draw of these buffers. Weak buffers (W/L = 0.12 μm/0.2 μm) were found to be the most optimum, with a weak 1-stage buffer providing an improvement in supply droop of 27% at a minimal cost in area (~3%) and power consumption (~3%).


weak buffer (W/L=0.12µm/0.2µm)
medium buffer (W/L=0.12µm/0.06µm)
strong buffer (W/L=0.5µm/0.06µm)
% Reduction in ΔV_{min}
Current Shift (Buffer Delay) (ps)
110%
90%
70%
50%
30%
10%
-10%
800
600
400
200
0
1
2
3
4
5
6
7
8
9
10
Number of Buffer Stages

(a)


weak buffer (W/L=0.12µm/0.2µm)
medium buffer (W/L=0.12µm/0.06µm)
strong buffer (W/L=0.5µm/0.06µm)
64-Bit ALU Normalized Power
1.25
1.20
1.15
1.10
1.05
1.00
1
2
3
4
5
6
7
8
9
10
Number of Buffer Stages

(b)

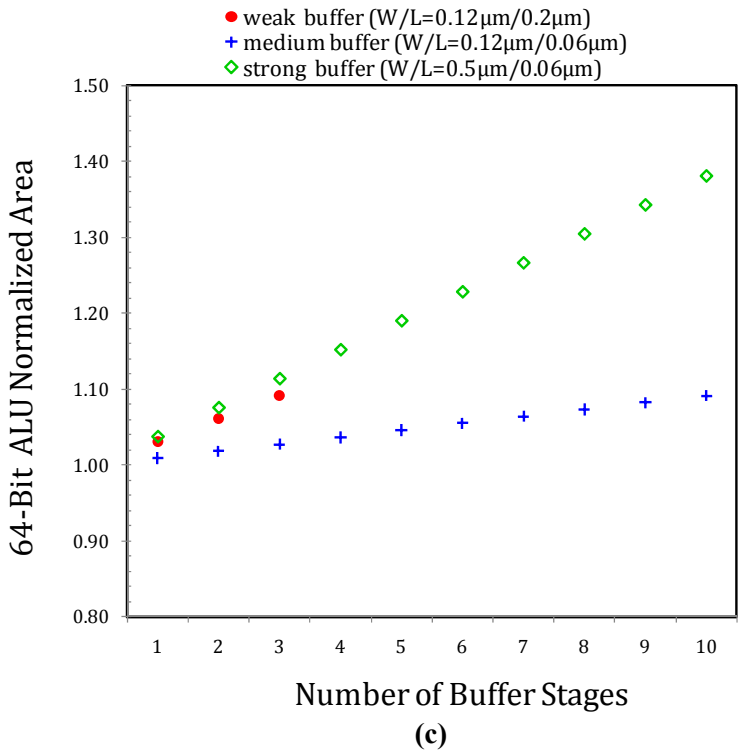


Figure 5.7 (a) Percentage reduction in ΔV_{min} and buffer delay, (b) normalized power consumption, (c) and normalized area overhead

Process variations are not expected to have an important effect on the effectiveness of the proposed circuit. While the variations can result in the adder and logic/shifter current peaks being offset from each other, there is sufficient design space to accommodate such variations. It can be observed from Figure 5.7 (a) that a single-stage buffer provides comparable benefits to a two-stage buffer and a slight overdesign can mitigate any variations due process variability. As can be seen in Figure 5.7 (b) and (c), the costs in area and power of a reasonable over-design are minimal.

5.1.3 Implications of Supply Waveform Reshaping

5.1.3.1 Area

When considering the area overhead, while minimal, leads to a worthy question: what if the same area required by the buffers were used simply for additional decap at the supply node for noise suppression in place of the current shifting design proposed? Figure 5.8 shows the minimum supply voltage for the standard ALU with varying amount of decoupling capacitance. Also shown is the minimum supply voltage for the ALU with current shifting and with a decoupling capacitance of 15 pF. As can be seen, approximately 3X more decoupling capacitance would be required to achieve the same level of voltage droop with the standard ALU compared to the current shifted ALU design. The area required by the additional decoupling capacitance in significantly larger than the area overhead of the current shifting scheme (~0.3% of additional decap area. The additional deacap would also come coupled with an undesirable increase in leakage current. It can be noted that this increase of 3.3X in apparent decoupling capacitance is applicable to the case presented. The specific relative advantage of the proposed current shifting technique is a function of the amount of decoupling capacitance used in addition to the current shifting technique, as well as the desired level of noise suppression required.

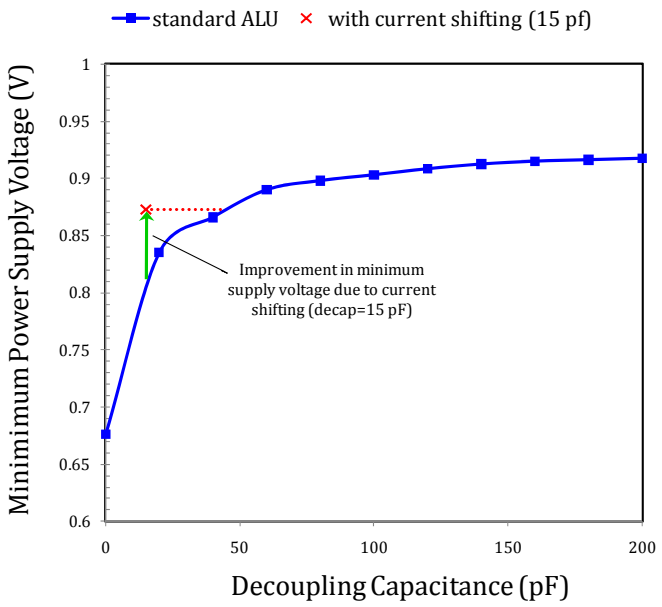


Figure 5.8 Minimum supply voltage for standard ALU with varying amounts of decap, and for current shifted ALU at 15 pF

5.1.3.2 Effect on Chip Nominal Voltage

As demonstrated, the proposed scheme of current shifting exhibits a significant improvement in ΔV_{min}. This has an important implication in terms of the stability of storage cells (SRAMs, flip-flops, latches, etc.) on a chip, which typically place a lower limit on the supply voltage (V_{min}) and thus a corresponding limit on the nominal supply voltage of the entire chip. In other words, in cases where the nominal voltage can be reduced for arithmetic blocks, for example, in low power circuits, the V_{min} imposed by the storage cells puts a limit on the how low the nominal voltage can be reduced. Therefore, if the ΔV_{min} of the supply voltage can be reduced, then the nominal voltage of circuit blocks that can continue to function at a lower voltage can be reduced such that the minimum supply voltage remains within the V_{min} constraint of the storage cells. Figure 5.9 illustrates the supply waveform for the standard ALU design and that of the current shifted design at a reduced nominal voltage (0.933V). The circuit on the lowered nominal supply

voltage results in a power reduction of ~18% (calculated including leakage power) without affecting the stability of any neighboring storage cells since both circuits display the same minimum voltage level.

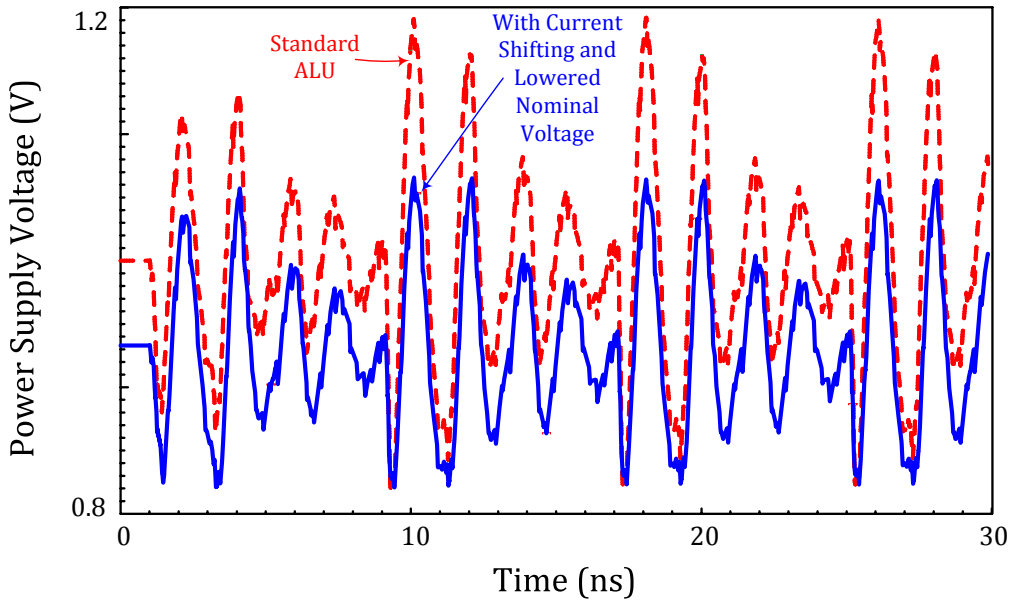


Figure 5.9 Power supply voltage waveforms for the standard ALU design, and the current-shifted design at a reduced nominal supply voltage of 0.933 V (18% power savings)

5.1.3.3 Clock Jitter

The suppression of noise in the supply voltage also affects the jitter of timing signals. The reduced variation in the voltage desirably decreases the jitter in these signals. The effect of the supply reshaping on the jitter of the clock signal was thus investigated. The eye diagrams for the clock signals with, and without, the proposed current shifting scheme are illustrated in Figure 5.10 (a) and (b), respectively. A reduction in jitter of ~29% is observed with the implementation of the proposed scheme.

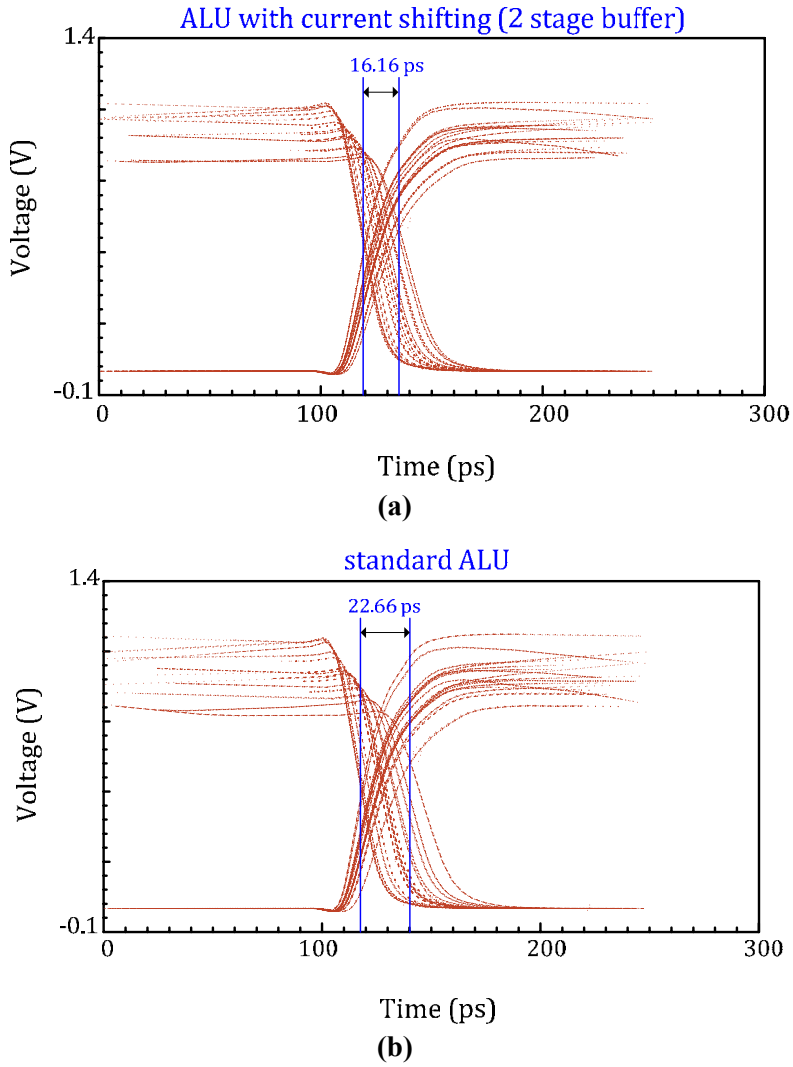


Figure 5.10 Clock signal jitter plots (a) with, and (b) without, current shifting

5.1.4 Effect of Current Shifting in Low Power (Reduced Voltage) Circuits

The relative effect of delaying the evaluation of non-critical circuit blocks (logic/shifter in this case) was further investigated at various nominal supply voltage levels. Figure 5.11 shows the percentage reduction in supply voltage dip (ΔV_{min}) for ALUs with the proposed current shifting scheme at nominal

supply voltages of 0.7 V, 0.8 V, 0.9 V and 1 V. As can be observed, the relative reduction in ΔV_{min} remains significant even at low nominal supply voltages.

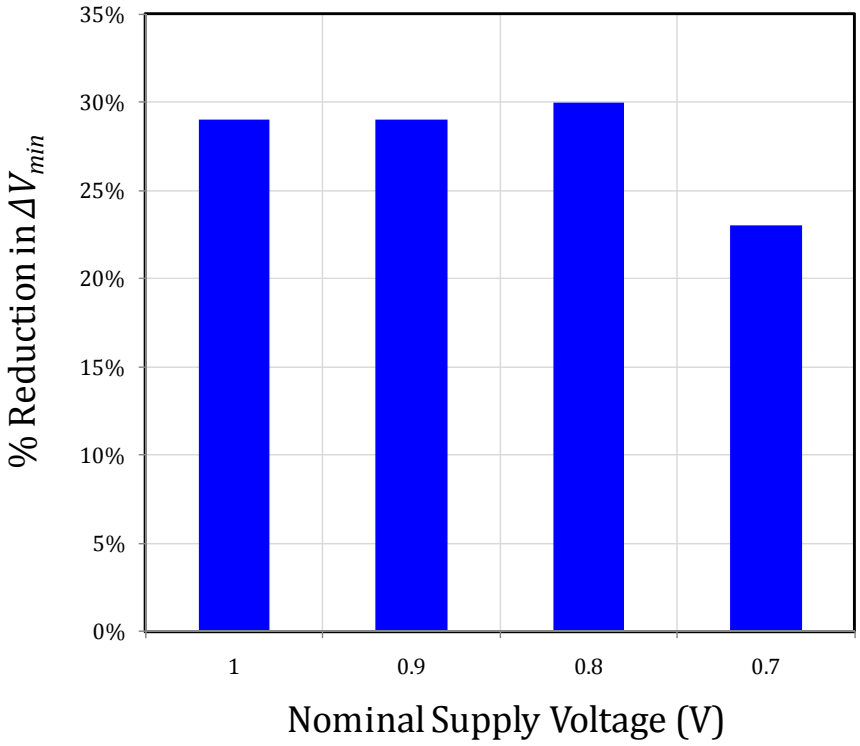


Figure 5.11 Percentage reduction in ΔV_{min} for varying nominal supply voltages

5.1.5 Conclusions

This work proposes a power supply noise suppression technique whereby the evaluation of non-critical circuitry is delayed within a clock cycle in order to shift its current consumption away from the clock edge to a later time within the cycle when switching currents are low. Using a 65 nm CMOS ALU circuit test bench, this technique shows ~27% improvement in the worst case supply voltage droop at a relatively

small cost to area (~3%) and power (~3%). The proposed technique is essentially shown to boost the decoupling capacitance effect by 3X for a given ALU design. Furthermore, the technique allows the nominal supply voltage of a chip to be reduced where desirable since the improved noise performance allows minimum voltage constraints on a chip to be met at a lower nominal supply voltage. An improvement in clock jitter is also observed with the proposed scheme. Lastly, the proposed scheme is shown to remain significantly effective at reduced voltage levels of up to at least 0.7 V.

5.2 Analysis of Switched Capacitor Based Noise Suppression

The switched capacitor technique is the most common active mitigation technique investigated in the literature. Compared to the other mitigation techniques presented in the previous section, the switched capacitor technique has the advantage that it can be used at high frequencies. It is therefore an attractive alternative to the conventional decaps and was thus selected for further analysis. In addition, after having been shown as an effective supply noise mitigation technique approximately nine years ago [70], much of the subsequent literature has primarily focused on the various triggering techniques for the switched capacitor circuit. As such, there is a lack of sufficient analysis available with respect to the limits and caveats of the switched capacitor circuit itself independent of the triggering mechanism. This section attempts to bridge this gap in information regarding this mitigation technique for the CMOS 65 nm technology node.

The model in Figure 5.13 is used to represent the noise generation circuit for the analysis. Typical values for L, R and R_{Cnon} are used, with the remainder of the parameters extracted from a 64-bit Kogge-Stone adder circuit. The various component values used are listed in Table 5-I. The current pattern was modeled as a sinusoidal waveform for simplicity.

TABLE 5-I SUPPLY NOISE MODEL PARAMETERS

COMPONENT	VALUE	COMPONENT	VALUE
$\boldsymbol{L}$	4 nH	R_{non}	5 MOhm
$\boldsymbol{R}$	200 mOhm	C_{non}	700 pF
$\boldsymbol{I_{ac}}$	2.5 GHz (positive axes) sine wave, 8.72 mA peak	R_{Cnon}	500 Ohms

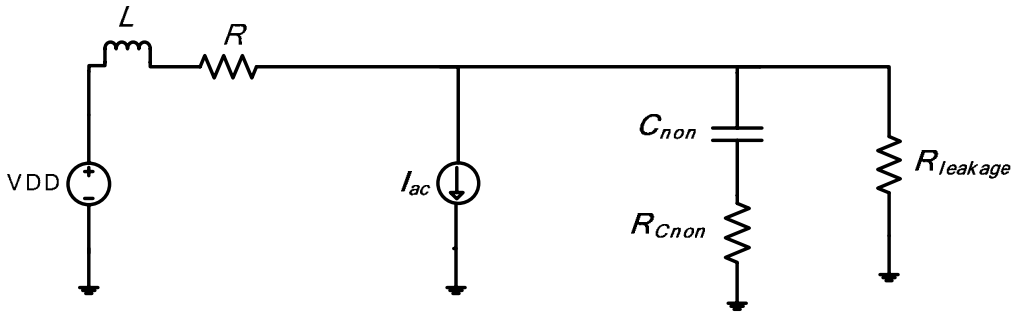
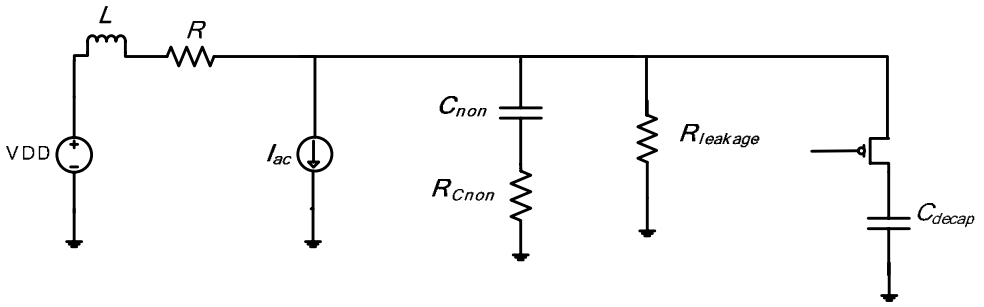
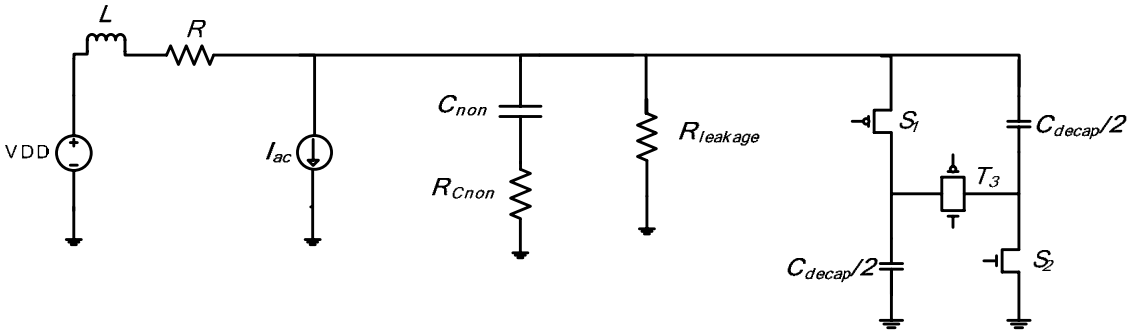


Figure 5.12 Supply noise generation model

In each analysis conducted, the adder-based model in Figure 5.12 was simulated to determine the supply noise level with no suppression applied for reference. The adder-based model with a passive decoupling capacitor, and the adder-based model with the switched capacitor circuit were then simulated and compared. The latter two models are shown in Figure 5.13 (a) and (b). A decap isolation switch is also included as is often done in practice. It should be noted that for the switched capacitor circuit, the switching is triggered at the start of the dip in voltage in attempt to maximize noise mitigation, and that in practice, depending on where the voltage that triggers the switching is detected, this may not be the case. Ideal clock signals are used to decouple the effect of the triggering circuitry and determine the ideal maximum achievable level of mitigation. Real NMOS and PMOS switches have also been used, and are sized to minimize resistance. NMOS capacitors are used to implement the decaps. Since the noise waveforms are sinusoidal in shape, only the minimum and maximum voltages have been reported.



(a)



(b)

Figure 5.13 Supply noise simulation models (a) with a passive decap, and (b) with a switched capacitor decap circuit

5.2.1 Theoretical Analysis

As described earlier, the switched capacitor circuit ideally provides a voltage doubling effect. Assuming the capacitance of each capacitor is equal, the total charge accumulated at each capacitor, ΔQ, is

$$\Delta Q = \frac{C_{decap}}{2} V_1 \tag{5-1}$$

where, V_1 is the voltage of the supply before switching. When a droop is detected in the supply, a triggering circuit causes the capacitors to switch from a parallel configuration to a series configuration resulting in the equivalent capacitance between the supply rails, C_{eq}, to become

$$C_{eq} = \frac{C_{decap}}{4} \tag{5-2}$$

and the supply voltage after switching, V_2, to ideally become

$$V_2 = \frac{C_{decap}/2}{C_{decap}/4} V_1 = 2V_1 \tag{5-3}$$

Considering the ability of the mitigation method to supply charge to the actively switching circuitry, the passive decap structure can provide charge, ΔQ_{decap}

$$\Delta Q_{decap} = C_{decap} \cdot \Delta V \tag{5-4}$$

where, ΔV is the dip in supply voltage. The switched capacitor implementation, on the other hand, is able to provide charge, ΔQ_{sw}

$$\Delta Q_{sw} = \frac{C_{decap}}{4}\left(VDD + \Delta V\right) \tag{5-5}$$

Therefore, for the switched capacitor circuit to provide more charge than the simple decap circuit

$$\Delta V < \frac{VDD}{3} \tag{5-6}$$

which is typically the case in most practical circuits.

Considering the effect of the switched capacitor implementation on the impedance between the supply and ground, the passive decap impedance, Z_{decap}, is

$$Z_{decap} = \frac{1}{j\omega C_{decap}} + R_{decap} + R_{switch} \tag{5-7}$$

where, R_{switch} and R_{decap} are the resistance of the switch in series with the capacitor, and any resistance associated with the capacitor, respectively. The impedance of the switched capacitor based scheme in the series configuration, Z_{sw_series}, is

$$Z_{sw\_series} = \frac{4}{j\omega C_{decap}} + 2R_{decap} + R_{switch} \tag{5-8}$$

Therefore, the impedance of the switched capacitor based scheme in the series configuration is larger than that of the passive decap implementation, and serves to counter, to some extent, the gains in noise mitigation made from the voltage doubling effect.

Furthermore, as seen in the model of Figure 5.13 (b), there is an additional capacitance, C_{non}, present between the supply rails. This causes the supply voltage, V_2, to become

$$V_2 = \left(\frac{C_{non} + C_{decap}/2}{C_{non} + C_{decap}/4} \right) V_1 \tag{5-9}$$

Therefore, if there is a large amount of additional capacitance between the rails, it can be seen that the boost in voltage can become quite small.

The maximum improvement can alternately be shown to be represented in terms of a gain, G [6]

$$G = \frac{n + k - 1}{k \cdot n^2} \tag{5-10}$$

where k is the voltage regulation tolerance and kV_1 the permissible drop in supply voltage, and n the number of parallel capacitors. Based on this equation, there exists a value of k such that a boost in voltage cannot be theoretically achieved [6].

The above equations all consider ideal cases, and in practice, due to switch resistances and leakage currents these boosts in voltage are not physically achievable. The following analysis using simulation of supply noise models aims to provide a more practical look at the behavior of the switched capacitor circuit versus the conventional passive decap in the 65 nm technology.

5.2.2 Effect of Capacitor Size

The purpose of this analysis is to investigate how the advantage of each mitigation method varies with the amount of decoupling capacitance added. Figure 5.14 shows the effect of increasing decap on supply noise suppression.

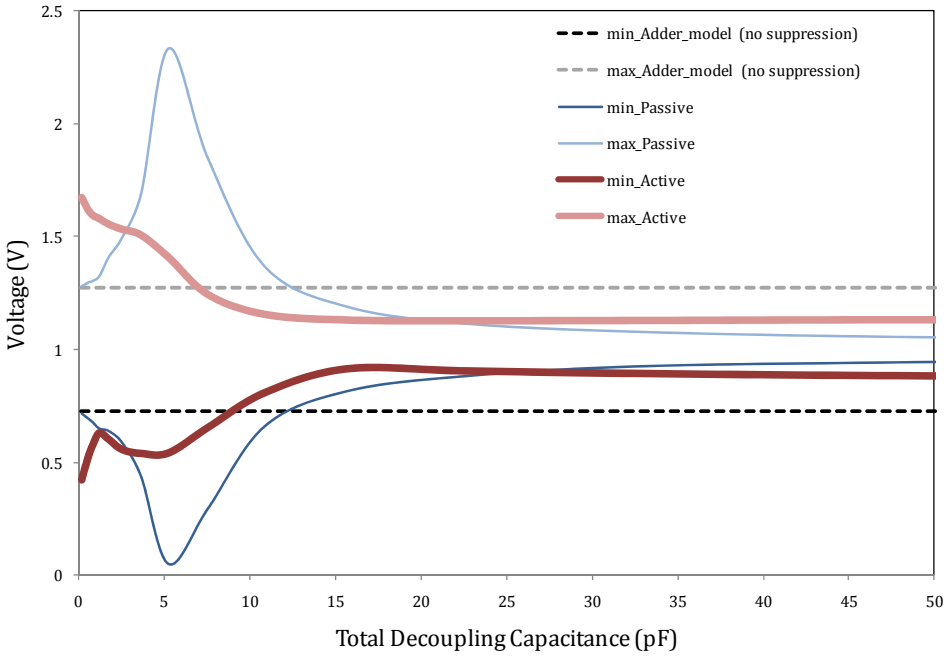


Figure 5.14 Effect of amount of decap on noise suppression

As can be seen there is a point between 20 and 25 pF where the decap circuit becomes preferable over the switched capacitor circuit. Upon inspection of the supply waveforms, it was observed that where the switched capacitor circuit mitigation levels off, the mitigation is limited by the timing of the triggering signals, therefore, some improvement in the mitigation may be seen when the circuit is coupled to a practical triggering circuit depending on the speed of detection and triggering. It can also observed that that increasing the decap beyond ~20 pF provides diminishing returns, which is a well known characteristic [40].

It is further visible that below a certain voltage, ~11 pF for the passive decap and ~7 pF for the switched capacitor circuit, the addition of the mitigation circuit actually exacerbates the noise problem. The reason for this is clear from inspection of the noise waveform where the voltage is seen to oscillate. This is a result of the resonance in the RLC network formed. In these instances it is, again, important to ensure sufficient resistance in the network to suppress the oscillations [85].

5.2.3 Effect of Placement

The distance between the mitigation circuitry and the noise source is known to be a critical parameter in supply noise mitigation. This is modeled by including a resistor between the supply and mitigation circuit, R_{dist}, to represent distance. A total decap value of 7 pF is used for this analysis. The effect of increasing the resistance between the supply voltage and mitigation circuit for each mitigation circuit is shown in Figure 5.15.

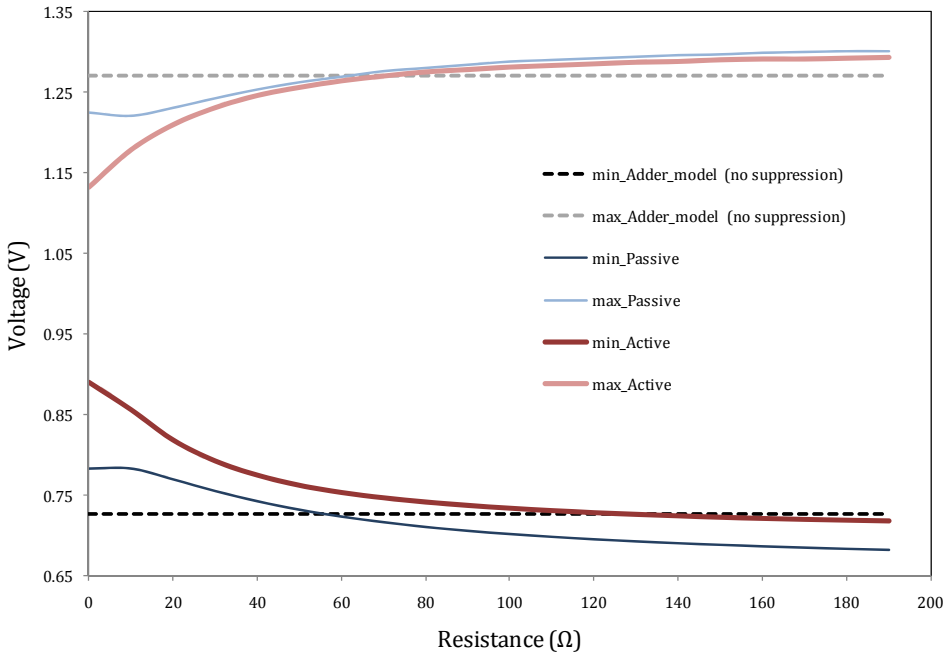


Figure 5.15 Effect of resistance between a supply and noise suppression circuitry

As expected, the mitigation efficiency is reduced with increasing resistance, or distance, between the supply and mitigation circuit. Furthermore, after ~60 Ω, the mitigation circuit has no effect on the noise in the case of the circuits modeled.

5.2.4 Effect of Timing and Triggering Circuitry

As described above, the time at which the switched capacitor circuit is switched affects the extent of mitigation observed. In practical circuits, a finite amount of time is required between when the supply voltage level is sensed and when the switches are triggered. Triggering the switching when the supply voltage is on an upward swing can increase the peak noise, and triggering the switches late in a downward cycle can miss the optimum mitigation opportunity. Figure 5.16 illustrates the effect of triggering the switches at various times across the 400 ps noise period.

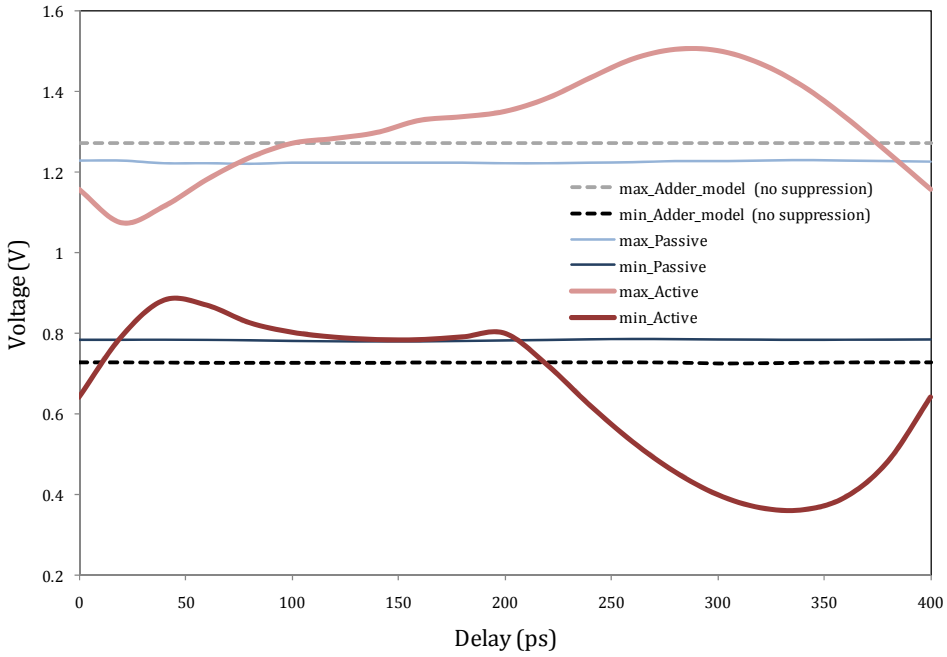


Figure 5.16 Effect of switch timing on noise suppression

5.2.5 5.5. Effect of Circuit Activity

With regards to the noise of the adder model with no suppression and the adder model with passive decap, it can be expected that increasing the circuit activity, or current draw, will simply increase the noise levels with the decap circuit attenuating the noise amplitude accordingly as shown in Figure 5.17.

With the switched capacitor circuit, however, it can be seen that the voltage levels play an important role. When the activity level is below approximately 60% to 80%, the switched capacitor can cause a spike in voltage that deteriorates the noise levels of the supply. And only after this region does it follow a similar pattern to the decap mitigation circuit.

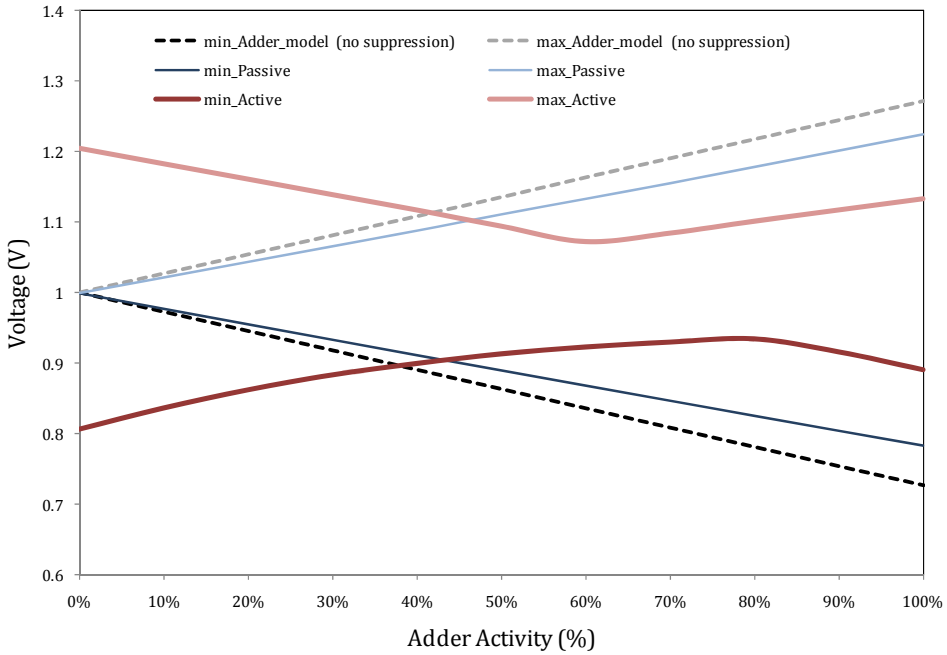


Figure 5.17 Effect of circuit activity on supply noise suppression

5.2.6 Conclusions

The analysis conducted highlights the limitations of the switched capacitor noise suppression technique found in the literature. It is theoretically shown that the larger the chip capacitance present due to non-switching circuitry, the less effective the switched capacitor technique. It is also shown that whether the switched capacitor technique is superior to a passive decap depends on the particular amount of decap implemented. Furthermore, the switched capacitor circuit is shown to be more sensitive to placement than the passive decap in the range R_{dist} < ~20Ω. As expected timing is also critical in determining whether the

switched capacitor circuit will outperform the passive decap or vice versa. Circuit activity is lastly shown to be an important consideration when designing the switched capacitor suppression circuit since at low activity levels, the technique can result in over compensation leading to worsening of the noise levels.

Chapter 6

Supply Noise Detection

"Discovery consists of seeing what everybody has seen and thinking what nobody else has thought."

…Jonathan Swift

This chapter deals with the area of supply noise detection. The ability to detect supply variations has important implications, especially during the testing phase of the design flow. A new method of detecting the minimum supply voltage level is proposed here. The method can also be used to reduce the complexity of existing detection schemes.

6.1 Detection Techniques and Challenges

A major challenge with on-chip noise detection techniques is that the noise being measured not only affects the probed circuitry but affects the operation of the measurement circuitry itself. This often necessitates the use of external supply and/or ground signals to be routed to the measurement cells [129]-[139], which leads to increased pin counts and added routing complexity. One method of dealing with this problem is to filter the on-chip supplies using a simple first order *RC*-filter [140][141], an example of

which is illustrated in Figure 6.1. Here, the resistance has been implemented as a transistor which provides the added benefit of acting as a power switch to the entire measurement cell. On-chip filtering is an effective way of providing a less noisy on-chip supply to the measurement circuitry, however depending on the supply noise immunity and current requirements of the measurement circuitry, the size of the measurement cell can become significantly large due to the large capacitors that may be required.

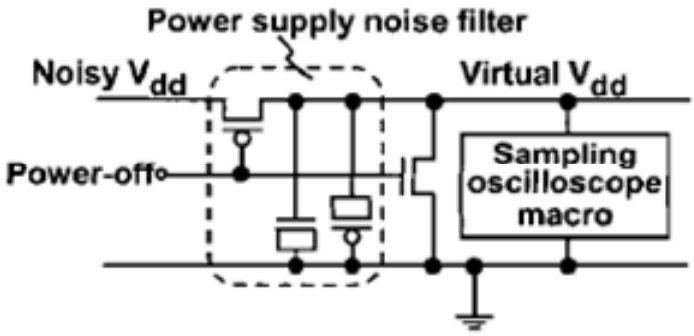


Figure 6.1 On-chip supply filter [140]

Another challenge in the design of measurement circuitry is that these circuits are designed in the same technology as the probed circuitry and thus cannot generally operate faster than the probed circuitry. Therefore, techniques that involve taking multiple samples of the noisy signal within a clock period in order to regenerate the waveform are difficult to employ. As an alternative, some of the early measurement methods [129][133][140][142][143] used a *sub-sampling* technique used in sampling oscilloscopes [144]. In this method, a periodic probed signal is assumed and a single sample is taken in each period, with each successive sample slightly delayed compared to the previous sample. In this way, the various sampled points can be accumulated to determine the measured waveform. This technique, although very useful, has the clear disadvantage that the probed signal must be periodic or that the probed signal must be regenerated multiple times. This precludes this method from being used in real-time applications.

As is evident from Appendix B, majority of the current noise measurement techniques utilize a *voltage comparator-based* approach [130][131][135][138][140][141][143][145][146]-[151] an example of which is illustrated in Figure 6.2. In these methods, a supply/ground line or any noisy signal is sampled and compared to a reference voltage to infer its value. This typically requires multiple comparisons to be made in order to accurately determine the voltage level [131], with the successive approximation technique often used [141][143][150]. This technique is used in analog-to-digital converters (ADC's) and efficiently determines which voltage references should be used for the comparisons. Once again, methods that require such repeated comparisons require a periodic noise source to be present and are not suitable for real-time measurement. Furthermore, an important disadvantage of using voltage comparators is that a number of reference voltages must be generated on-chip or provided externally. Digital-to-analog converters (DACs) have been used to generate these references [141] which can result in large measurement cells, increased power consumption, added routing complexity and the need for digital inputs. These references are also susceptible to noise and have been routed as currents instead of voltages until they are in close proximity to the comparator circuits to reduce their susceptibility to noise [135].

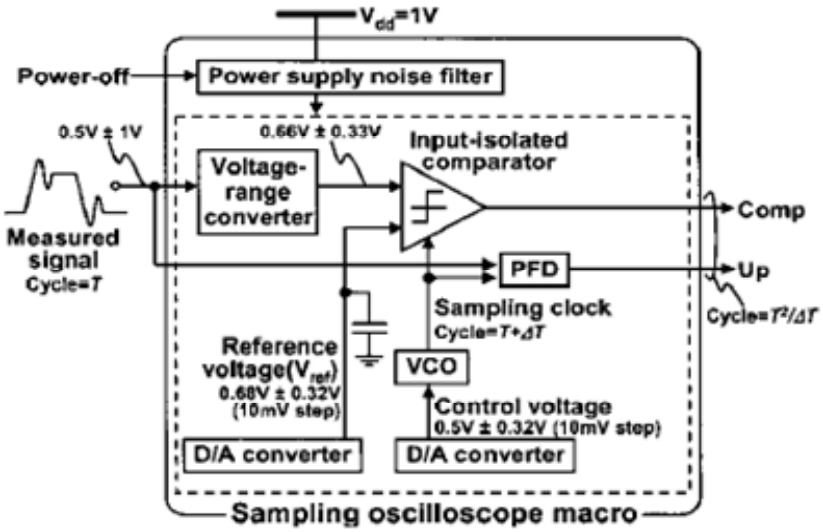


Figure 6.2 Voltage comparator based supply noise detection technique [140]

In attempt to overcome the problem of requiring a periodic noise source, alternate techniques have been proposed where the voltage levels are evaluated at predetermined times during the clock cycle [138] or at user triggered times [139]. Similarly, there are techniques that detect and report significant events that have occurred, where these significant events are defined by certain predetermined criteria [132][134][135][152][138]. For example, an overshoot detector circuit [132] detects every instance that the supply exceeds a certain threshold and outputs the number of times this event occurred. In [138], the circuitry provides programmability with respect to polarity, duration and magnitude to define the significant event. In [138] a peak detection circuit is, however, first needed to narrow down the window during which the noise event occurs and a flash ADC is then used to measure the noise only in the given window thus reducing the number of comparators required. These event detection methods allow real-time or almost real-time noise event detection. They are still, however, relatively large and complex.

Another technique [138][151] is to simply use a flash ADC. Here multiple comparators are present each with a different reference level applied to it. This method can quickly lead to a very large cell size and power consumption due to the large number of comparator stages and reference voltages that need to be generated depending on the required resolution. Figure 6.3 illustrates the noise sensing circuitry of [138], noting that the peak detection circuitry serves to narrow down the measurement window thus reducing the number of comparator stages needed.

Techniques based on the delay and/or errors generated in digital circuits within the measurement cell have also been proposed [134][139][154]. For example, errors in the output of two NAND gates can be used to identify fluctuations in the power supply line using the circuit illustrated in Figure 6.4. If the supply dips below a particular level, the outputs of the error detection circuitry can both be either high or low instead of being complementary as expected under a clean supply line. Another method based on the delay of a chain of inverters [154], monitors the delay across the chain and correlates this to supply

voltage. A similar method [139] connects the inverters in a ring oscillator configuration and correlates the frequency of oscillations to the supply voltage. These techniques have the disadvantage of requiring significant post-processing, calibration and in some cases separate clean supplies for the detection circuitry. The latter technique further has the disadvantage of being limited to low bandwidths since the ring oscillator frequencies are susceptible to the same filtering problems as direct on-chip methods [136][137].

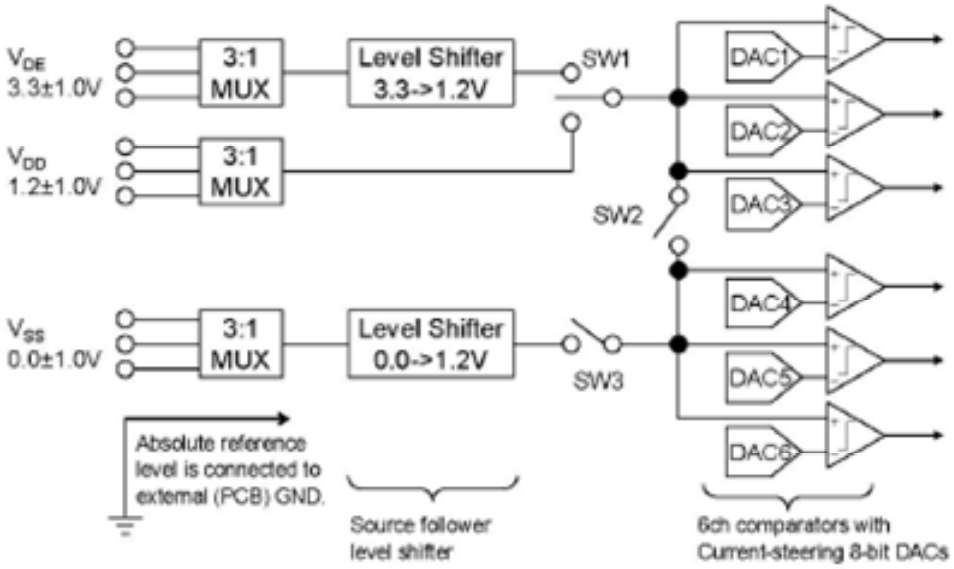


Figure 6.3 Flash ADC-like supply noise detection technique [138]

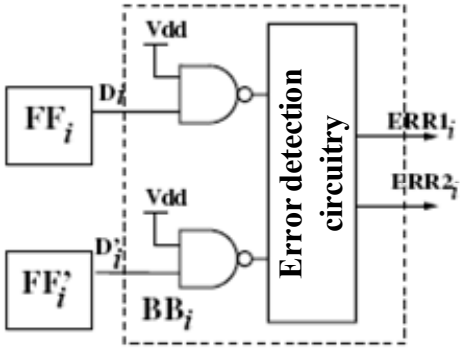


Figure 6.4 NAND gate-based supply noise detection scheme [134]

A further method of measuring supply and/or ground noise includes *dithering* low resolution measurements to increase resolution [146][155], which has the disadvantage of significant post processing required. Yet another method [130][145] involves monitoring the variation in the metastability point of comparators with respect to variations in substrate noise. This method suffers from the need for external noiseless signals and the requirement for the noise to be periodic.

When overshoots occur on the power supply, the voltage levels can exceed the input range of the sensing circuitry. This is dealt with by stepping down the supply voltage prior to measurement using a voltage divider as illustrated in Figure 6.5. Lastly, calibration remains a significant challenge for many of the measurement techniques since they rely on accurate reference voltages and accurate timing signals to be available. A means of calibration is thus often added to the circuitry to ensure accuracy of the measurement techniques [142][143].

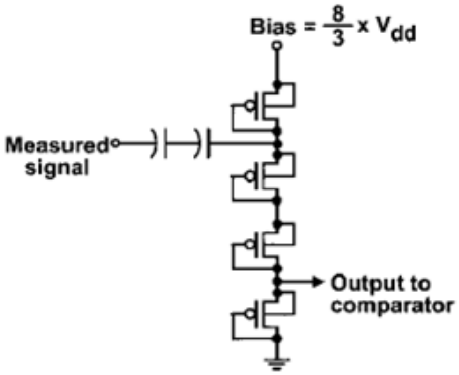


Figure 6.5 Voltage level shifter [140]

In the following sections, two alternate supply noise detection circuits are proposed. The first detects the minimum point of the supply voltage and can similarly be applied to peak detection. AS will be shown this technique has a significantly smaller footprint compared to previous designs and correspondingly requires low power and is further amenable to high-speed detection. The second circuit proposed is a digital detection scheme based on half-latches and desirably provides a digital output signal

making it less susceptible to noise. It further eliminates the need to generate multiple reference levels as in the case of flash ADC-based techniques.

6.2 A Further Look: A Supply Noise Minimum Point Detection Circuit

6.2.1 Concept of Detection Circuit

The basic circuit concept of the proposed minimum supply voltage detection scheme is illustrated in Figure 6.6. Initially, capacitor, C, is reset to a known high reference voltage by switch, S_1. Switch, S_1, is then turned off and switch, S_2, activated to allow capacitor, C, to discharge through diode, D, when the voltage at node X falls to a particular level below the turn-on voltage of the diode. The voltage on the capacitor will settle to a value that is dependent on both the length of the detection window as well as the RC-constant of the discharge path. Since diode, D, only allows current to flow away from capacitor, C, the voltage of capacitor, C, settles to a minimum level based on the voltage at node X before the next cycle when switch S_1 resets the capacitor voltage. For the capacitor to discharge the voltage at node X must be below the turn-on voltage of the diode, therefore a voltage shifter block is used to step down the noisy supply voltage such that relatively small fluctuations in supply voltage can be detected with the proposed scheme. The voltage shifter can be implemented using a simple voltage divider circuit or a source follower circuit. The proposed circuit can thus be used to detect various minimum levels of supply noise by correlating the different minimum supply voltages to the minimum voltage on capacitor, C, within a given period. A straightforward calibration can be used to relate the measured voltage levels to the actual on-chip minimum supply voltage level. To be used as a peak detector, the direction of diode can simply be modified to reverse the direction of current flow.

The signal at node *Detector_out* needs to be converted to a form that can straightforwardly be carried off-chip. If the detection window is small, the analog output can simply be read off-chip as the circuit essentially behaves as a sample-and-hold circuit that holds on to the minimum voltage level. If the detection window is relatively small, however, reading the signal directly can be challenging. In this case, a number of the known supply voltage detection techniques described earlier can be used to read the voltage stored on the capacitor. An alternate method proposed in Figure 6.7 may also be used. Here, a number of capacitors C_1 to C_n, and related switches S_{1_1} to S_{1_n} and S_{2_1} to S_{2_n}, are consecutively used to sample the minimum voltage level, the number of capacitors selected will depend on the sampling speed and the speed at which the output signal can be captured. For example, If the detection frequency is 500 MHz, implementing five capacitors allows each capacitor to be read out at a rate of 100 MHz, or implementing ten capacitors reduces the output rate to 50 MHz. This would result in possibly a large number of outputs and an analog multiplexer would be required to efficiently manage the output signals.

6.2.1.1 Advantages

While an additional measurement block may be required, the advantage of this method lies in the fact that the minimum supply voltage is detected and a proportional signal moved to a predictable point in time. This eliminates the uncertainty in where the minimum voltage should be measured and the need to sample the voltage multiple times in a given period in order to locate the minimum voltage. This makes the proposed method both conducive to real time detection as well as reduces the complexity and size of the overall detection scheme. Furthermore, compared to the detection schemes in [152][153] where an opamp/comparator is used for voltage comparison, the proposed detection scheme utilizes only a single device (diode) with an elegant circuit implementation as will be shown. The area and speed of the detector are thus significantly reduced compared to the existing implementations. Lastly, the size of capacitor, C, can be made relatively small thus making the power requirements of V_{ref} correspondingly

small. This makes the circuit amenable to the use of a locally generated voltage reference as opposed to an external dedicated quiet supply.

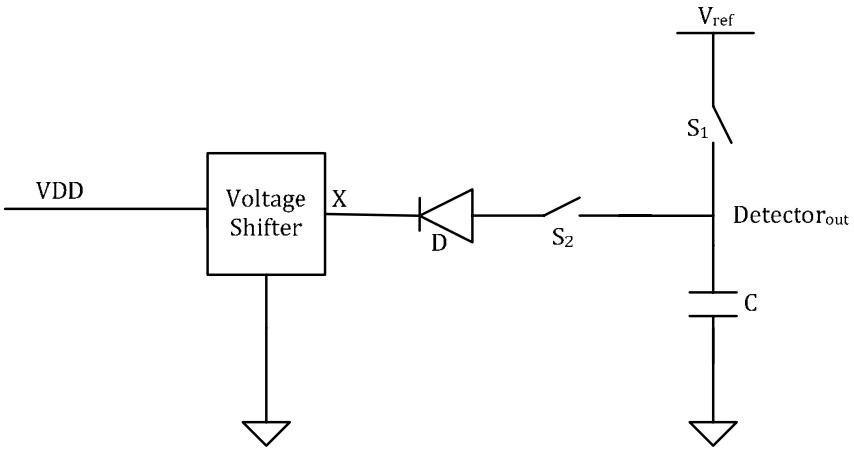


Figure 6.6 Basic circuit concept of supply noise detector circuit

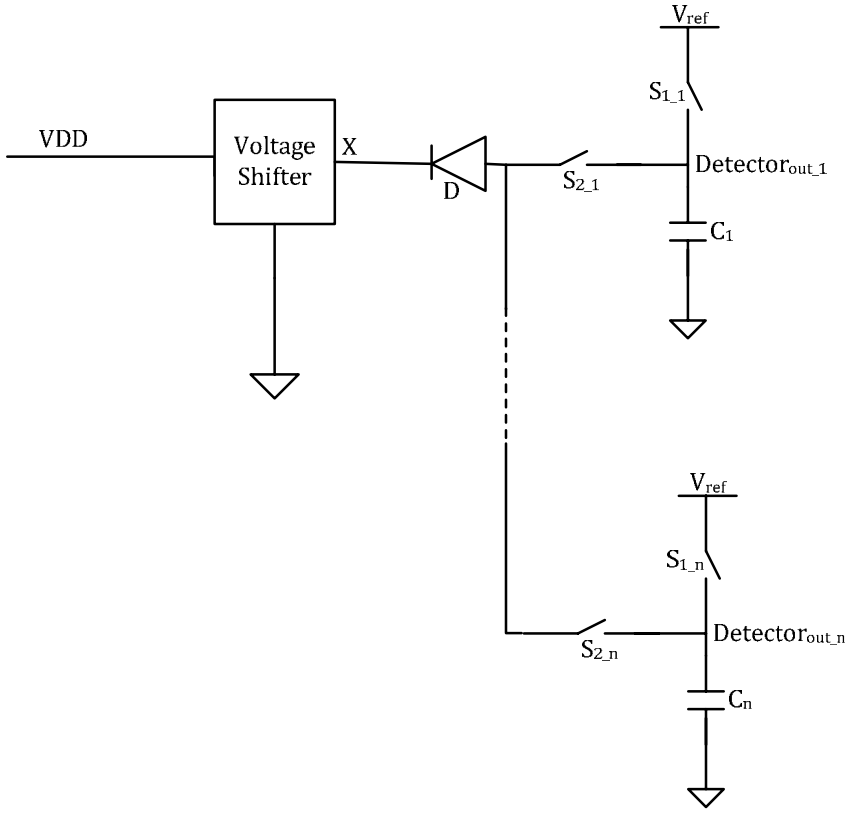


Figure 6.7 Detector circuit for high speed detection

6.2.2 Circuit Implementation

Figure 6.8 illustrates the circuit implementation of the proposed supply noise detection scheme of Figure 6.6. Diode, *D*, is implemented using a diode-connected transistor in order to realize a small turn-on voltage of the device. Switch, S_2, is implemented using a transmission gate. A dummy transmission gate is also implemented at node $Detector_{out}$ such that the switching of the reset signals in an opposite direction to the active transmission gate signals results in cancelling of clock feedthrough effects at the output of the detector. Switch, S_1, is further implemented using a simple PMOS device and sized appropriately to minimize clock feedthrough to the node $Detector_{out}$. Capacitor, *C*, is implemented using an NMOS capacitor to minimize its area. The capacitance value of the NMOS cap is relatively small with a value 14 fF. The stepping down of the supply voltage is obtained using a source follower as shown and sized to minimize the power consumption and maximize the input range. An inverter is further required to generate the complementary reset signal, where the reset signal can easily be generated from the clock of the noise generating circuitry itself. For simplicity, an externally supplied voltage reference is used here, however, as described previously a local voltage reference circuit can be used to implement this voltage.

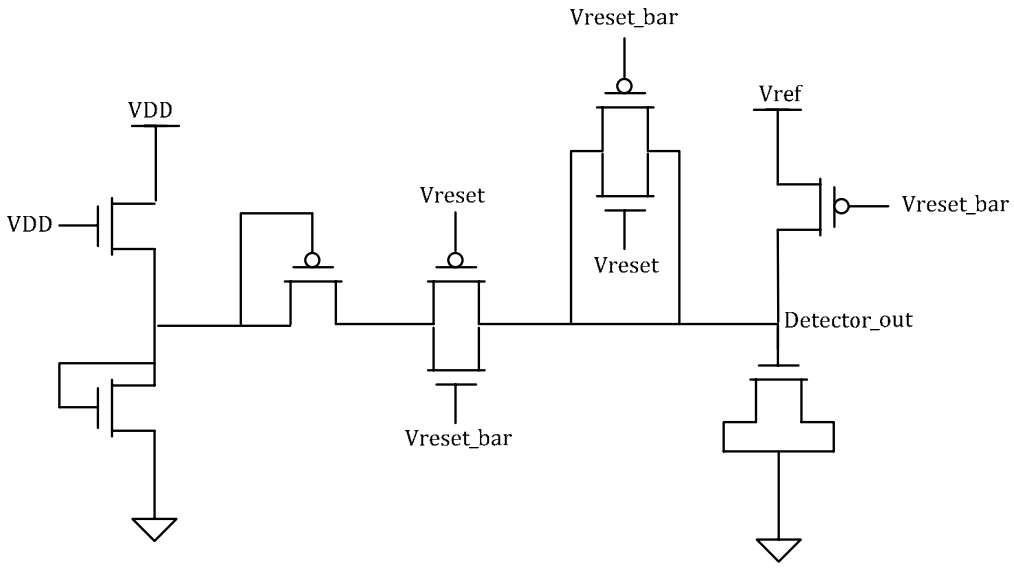


Figure 6.8 Circuit implementation of supply noise detection circuit

6.2.3 Simulation Results

In order to demonstrate the performance of the proposed detection scheme, it was implemented with a 64-bit static ALU block and the supply noise generated by the switching activity of the ALU block measured. Figure 6.9 illustrates the overall testbench used. The ALU switching generated supply noise with a frequency of ~1 GHz. In order to determine the frequency limits of the proposed detection scheme, a sine wave was also used in place of the ALU block to enable generation of supply noise with a wide range of amplitudes and frequencies.

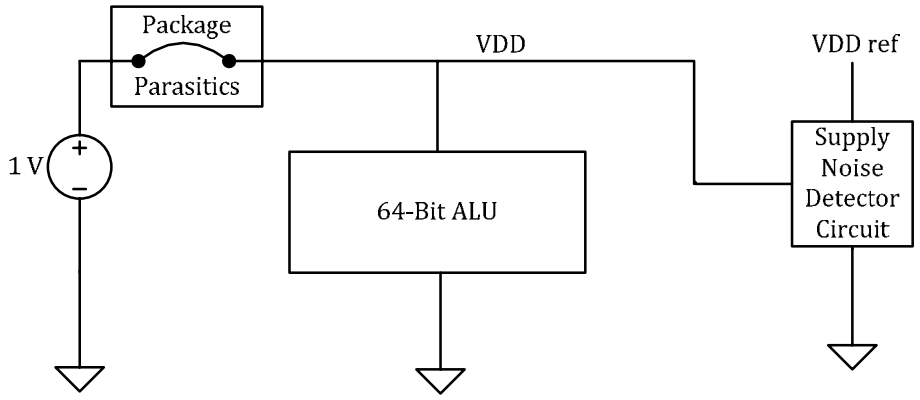
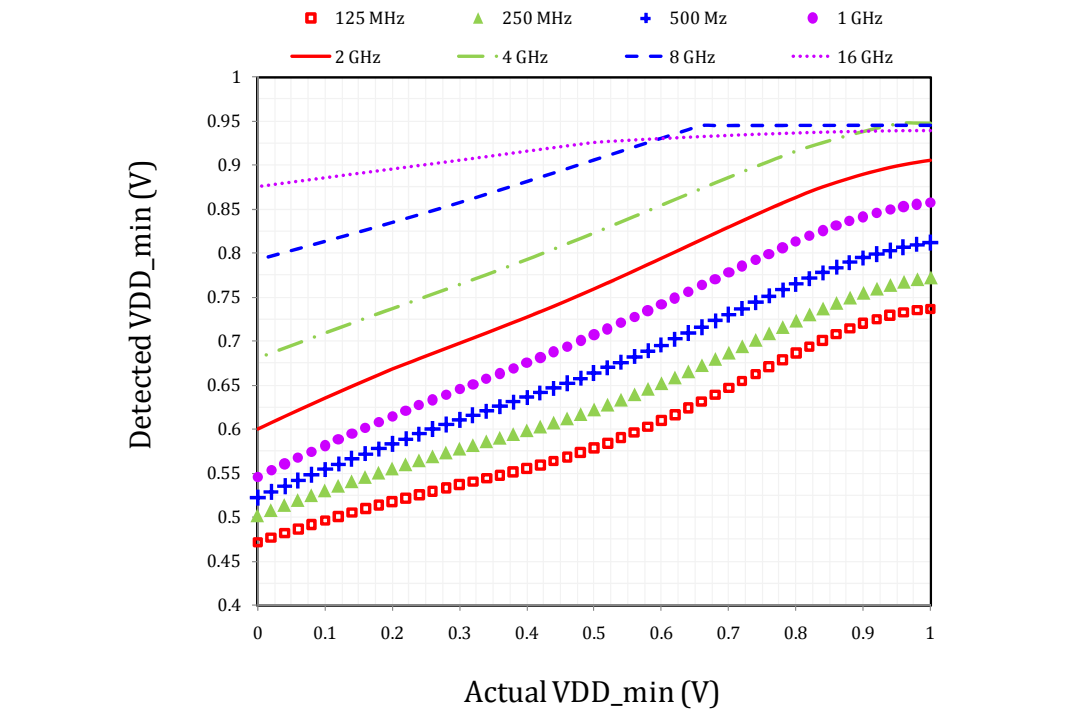
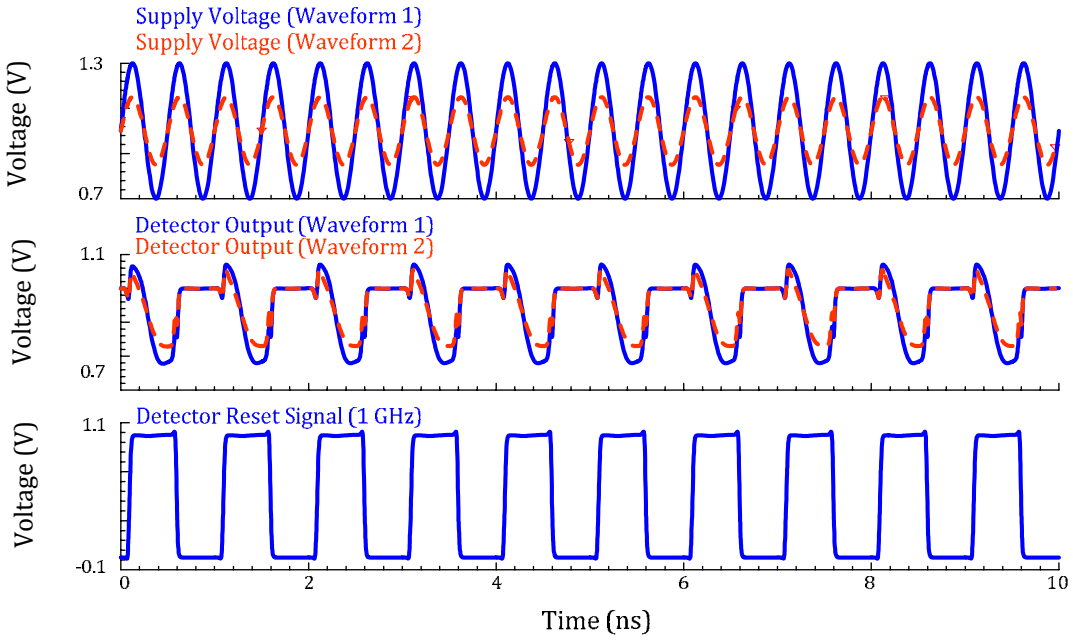


Figure 6.9 Circuit testbench for noise generation

Figure 6.10 (a) shows the correlation between the actual minimum supply voltage and the minimum voltage from the detection circuit for various detection frequencies. The noise frequency is set to twice that of the measurement frequency to ensure the minimum voltage is captured within the detection window. Again, a sine wave is used as the noise source in this case to enable a wide range of amplitudes and frequencies to be obtained for characterization of the circuit. Figure 6.10 (b) shows a sample pair of supply and detector output curves where a change in detector output is visibly correlated to a change in minimum supply voltage.



(a)



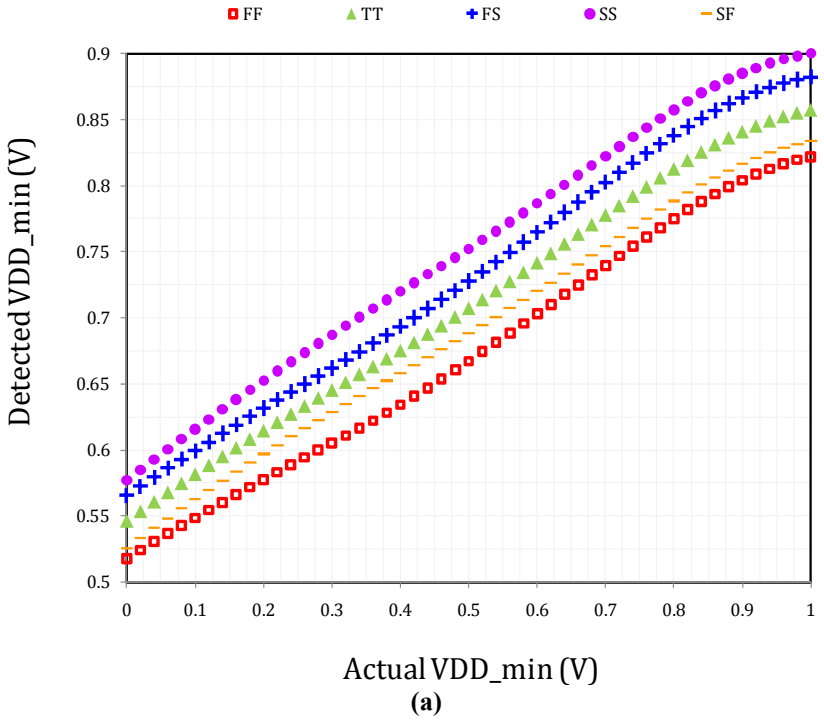
(b)

Figure 6.10 (a) Correlation between actual minimum supply voltage and minimum detector output level, and (b) a sample pair of supply and detector output curves

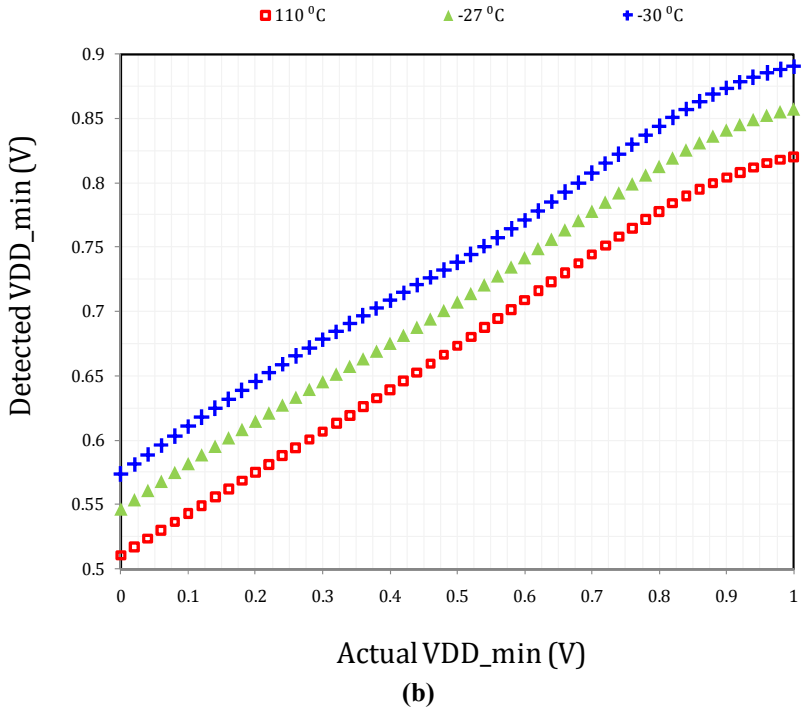
As can be seen in Figure 6.10 (a), an approximately linear or piecewise linear correlation exists between the actual minimum supply voltage and the minimum detector output voltage for a wide range of supply fluctuations. Up to a detection frequency of between 2-4 GHz the circuit output continues to exhibit an approximately linear correlation to the actual minimum supply voltage level. The overall resolution of the detection scheme will be dependent upon the technique utilized to move the signal off-chip. Assuming a resolution of 5 mV of the output technique, the proposed scheme provides a measurement resolution of ~15 mV at a detection frequency of 2 GHz.

6.2.4 Effect of Process Variations and Temperature

There will be an effect of process variation and temperature on the relationship between the detected minimum voltage and the actual minimum voltage. Figure 6.11 (a) illustrates the extent of the process variation effects at the various corners, and Figure 6.11 (b) illustrates the effect of temperature. As can be seen there is no significant degradation in resolution caused by neither process not temperature variations.



(a)



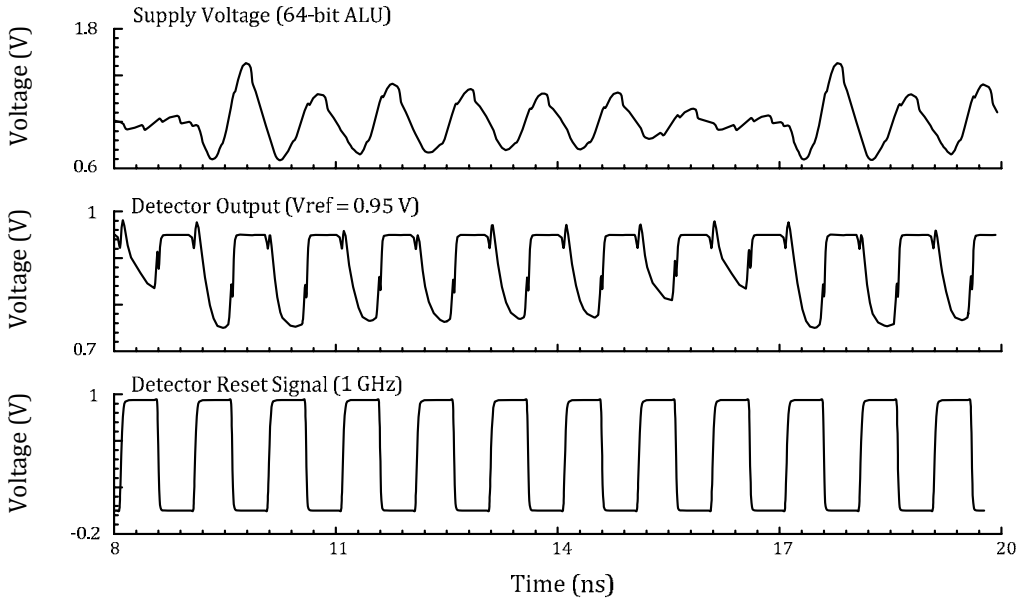
(b)

Figure 6.11 Effect of (a) process and (b) temperature variation on the calibration between the detected and actual minimum supply voltage levels

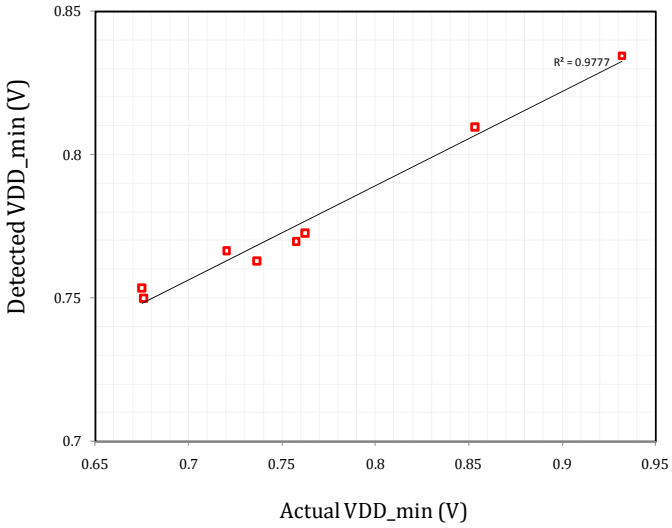
6.2.5 Noise Detection on ALU Block

The proposed noise detection scheme was used to measure the minimum supply voltage levels during the actively switching 64-bit ALU block implemented in a 65 nm CMOS technology. A maximum noise input pattern was used to generate the supply fluctuations where the storage-flips generate most of the noise seen on the supply voltage. A reference voltage V_{ref} of 0.95 V was used for the detection circuit leaving some voltage leeway for a reference voltage generation circuit. A detection frequency of 1 GHz was selected to enable capture of each droop in supply voltage level. Figure 6.12 (a) illustrates a sample supply voltage waveform with its corresponding detector output waveform. As can be seen, there is a correlation between the minimum supply voltage level within the detection window and the minimum

detector output level at the end of each detection period. This correlation is quantified in Figure 6.12(b) which shows a similar resolution to that in Figure 6.10 (a). TABLE I shows the relative cost in area and power of the detection circuit at a measurement frequency of 1 GHz. As shown, the cost in both these resources is only ~2% compared to the 64-bit ALU block itself.



(a)



(b)

Figure 6.12 (a) 64-bit ALU supply waveform with corresponding detector output, and (b) correlation between the actual supply voltage minimum and detector output minimum voltage

TABLE 6-I RELATIVE AREA AND POWER CONSUMPTION OF DETECTION CIRCUIT AT A DETECTION FREQUENCY OF 1 GHZ

	64-BIT ALU BLOCK	SUPPLY NOISE DETECTION CIRCUIT
Area (WxL)	403 μm^2	10.6 μm^2 (2.6%)
Power Consumption	3.0876 mW	0.0631 mW (2%)

6.2.6 Conclusions

The minimum supply voltage levels in a chip are of significant importance due the minimum voltage constraints of on-chip storage elements such as memories. The nature of the supply grid and data patterns however make it difficult to determine the minimum supply levels in real-time operation, making the real-time detection of noise a desirable capability during chip operation and testing. This work proposes a simple, small and real-time power supply noise detection circuit for the detection of minimum/maximum supply voltage levels. It is also amenable for use with existing detection techniques for high-frequency detection with its advantage lying in that it generates a signal at a known point in time that is correlated to the minimum voltage, thus eliminating large overheads in scanning for the minimum voltage level. The circuit is further conducive to local reference voltage generation mitigating the need for resource intensive external supplies. The detection scheme is shown to be functional for a wide range of supply fluctuations up to at least a 4 GHz detection frequency in a 65 nm technology. The detection scheme is also used to measure the minimum supply voltage for a 64-bit ALU with a cost in each of area and power of only ~2%.

Chapter 7

Conclusions

"We shall not cease from exploration. And the end of all our exploring will be to arrive where we started and know the place for the first time."

… T. S. Eliot

7.1 Conclusions and Contributions to the Field

The work carried out over the duration of this PhD degree and presented in this thesis has led to a number of conclusions and contributions. Each major conclusion is briefly presented below:

1) **Power supply noise will get progressively worse with scaling but the rate will not be as fast as previously predicted:** The initial work on scaling trends shows that existing power supply noise trend predictions in the field are inadequate, especially given modern scaling trends and design practices, and scaling trend predictions are made. The predicted move to an SOI process will, however, result in a sizeable spike in decap requirements since the intrinsic nwell capacitance will no longer be present.
2) **Guidelines for which decap to use for given design constraints and metal decaps:** The detailed analysis on various decap structures enables optimal selection of decap for given constraints in area and/or power consumption. It was shown that multilayer metal decaps are a significant source of decoupling capacitance when used in combination with the traditional MOS-based structures. The metal structures further simultaneously fulfill chip metal-fill requirements. These decaps will

become increasingly important as CMOS technology progresses and is an attractive means of countering the increase in supply noise predicted with scaling. The increase in capacitance provided by the metal decaps when combined with traditional structures is confirmed by chip measurements.

3) **Supply noise generation can be reduced by current shifting within a clock cycle:** Previous techniques of current shifting to minimize supply noise generation have focused on shifting the clock edge. The work done here shows that an additional suppression in noise generation can be obtained by current shifting within a clock cycle and this is demonstrated for a 64-bit ALU.
4) **Switched capacitor based supply noise suppression has its limitations:** It is shown that the switched-capacitor based supply noise technique described in the literature is advantageous within a bound range of design parameters. Specifically, the amount of capacitance added, its placement, and switching circuit size and activity were shown to be important considerations. The timing of the switched-capacitor circuit was also shown to be important.
5) **Proposed a fast, small and power efficient supply noise minimum point detector:** Given that the minimum supply noise level is of great significance on a chip due to its effect on the stability of storage elements and timing signal jitter, detection of this minimum voltage level is very useful. The proposed method of detection has the advantage of being small and fast, two largely desirable qualities of noise detection circuits.

7.2 Contributions to Literature in the Field

The following publications are either accepted, submitted or in preparation for submission to the current literature in the field:

1) **T. Charania,** P. Chuang, A. Opal, and M. Sachdev, "Analysis of power supply noise mitigation circuits," *Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering*, pp. 1250-55, 2011.

2) **T. Charania**, A. Opal, and M. Sachdev, "Design and Analysis of On-Chip Decoupling Capacitors," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* (in press).

3) **T. Charania,** P. Chuang, A. Opal and M. Sachdev, "Suppression of on-chip power supply noise generated by a 64-bit static logic ALU block," *IEEE Conference on Very Large Scale Integration,* (submitted).

4) **T. Charania,** A. Opal and M. Sachdev, "Effect of technology scaling on on-chip decoupling capacitor structures," *International System-on-Chip Conference,* (submitted).

5) **T. Charania**, A. Opal and M. Sachdev, "Modern trends in on-chip power supply noise," (in preparation for submission to *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*).

6) **T. Charania,** P. Chuang, A. Opal and M. Sachdev, "A fast and compact real-time power supply minimum voltage detection circuit," (in preparation for submission to *Transactions on Circuits and Systems II*).

7) **T. Charania,** A. Opal and M. Sachdev, "Importance of multilayer interdigitated metal capacitors for use as decaps," (in preparation for submission to *Journal of Solid State Circuits*).

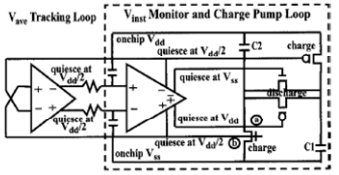
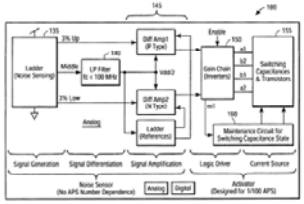
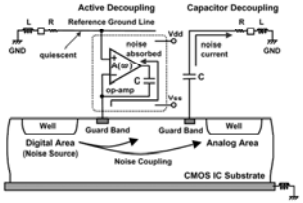
7.3 Future Work

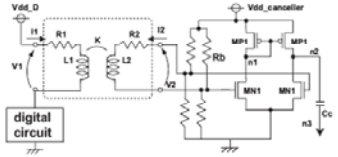
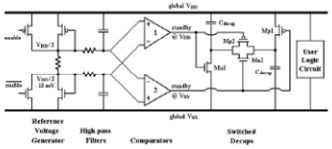
The work undertaken during the degree, has led to a number of areas of future research in the field of power supply noise. Each potential area is described briefly below:

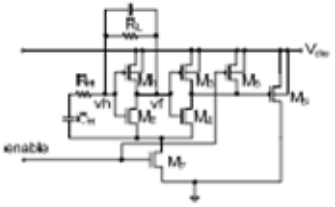
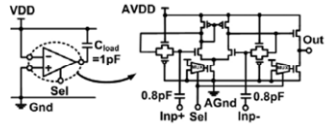
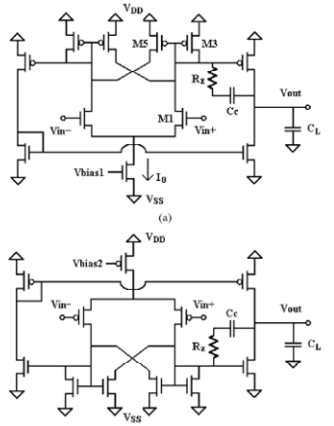
1) **Decap analysis in 3D EM simulator:** As was seen, the BSIM model used for simulation of the decaps does not sufficiently model MOS devices at high frequencies. A reproduction of the analysis conducted in a 3D EM simulator would allow the conclusions to be verified as well as provide clearer insight into the optimum unit length for MOS decaps in various processes.
2) **Implementation of supply minimum point detection circuit with additional output circuitry:** The advantage of the minimum point detection circuit presented lies in that it eliminates the overhead in scanning for the minimum voltage in a given period. Implementation of the proposed detection with a complementary detection scheme for high frequency supply noise detection would be useful in demonstrating this advantage.

Appendix A
Active Supply Noise Suppression Literature Survey

TABLE A.1 SUPPLY NOISE SUPPRESSION LITERATURE SURVEY

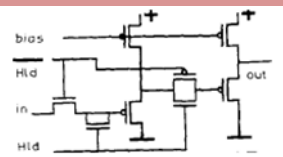
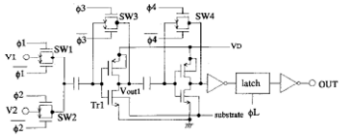
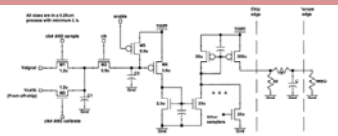
REF	YR	PUBLIC[N]	AFFIL[N]	*(NOISE TYPE)* DESCRIPTION	EXT SIGNALS	TECHN./ SUPPLY	UNIT PARAMETERS	CHALLENGES	ADVANTAGES	FIGURES
[70] --	2000	ISSCC US Patent # 6,509,785	Sun Microsys.	Active decaps/ switched capacitors (first publication) Supply ac coupled via HPF Autonomous operation Pseudo cascade inverter based comparator Placed under power grid to save space	None	0.15 µm/ 1.5 V	Chip cap 0.8 nF Passive decap 1.8 nF Active decap 1.35 nF Improvement (min) 4.6% Noise frequency 50 MHz	Slow	More efficient than passive decaps Space efficient	
[78]	2003	US Patent # 6,744,242	Fujitsu	Switched capacitors Diffamps + inverter chain to trigger Filtered supply for biasing				Slow		
[71]	2005	JSSC	STARC, Japan	Miller capacitance of opamp	Separate ground	0.13 µm/ 1 V	Feedback capacitance 30 pF Frequency 400 MHz (can go up to GHz) Improvement (peak) compared to passive 36%		Very efficient	

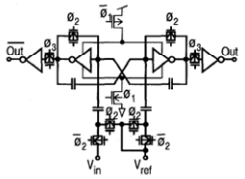
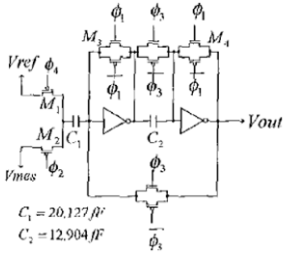
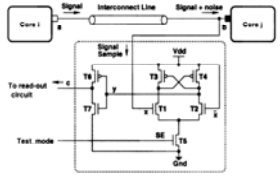
REF	YR	PUBLIC[N]	AFFIL[N]	*(NOISE TYPE)* DESCRIPTION	EXT SIGNALS	TECHN.	PARAMETERS	CHALLENGES	ADVANTAGES	FIGURES
[72]	2007	ASP-DAC	University of Tokyo	*(substrate noise)* Transformer senses changes in current Anti-phase amplified current generated by amplifier Current injected into substrate	Separate power supply	0.35 µm	Frequency 300 MHz (100-700 MHz possible) Improvement (peak) 9% compared to guard ring	Slow		
[79]	2007	VLSIC Digest of Technical Papers	University of Minnesota	*(resonant noise)* Switched capacitors Digital noise sensing: delay lines + phase comparator	None	0.13 µm/ 1.2 V	Frequency 100 MHz Improvement (min) 30% compared to passive decap	Slow		
[80]	2007	ISCAS	UBC	*(supply and ground noise)* Based on**[70]** (switched capacitor + HPF + comparator) Two-stage opamp (cascode + current mirror) comparator Voltage divider biases comparator		90 nm/ 1 V	Frequency ~166 MHz Improvement (min) 3% compared to passive decap	Slow		

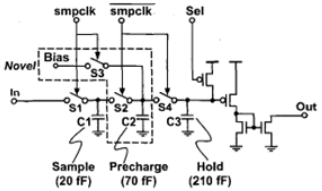
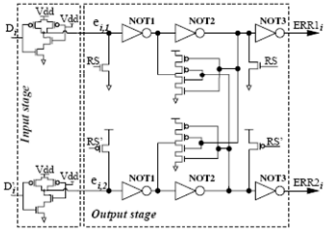
REF	YR	PUBLIC[N]	AFFIL[N]	*(NOISE TYPE)* DESCRIPTION	EXT SIGNALS	TECHN.	PARAMETERS	CHALLENGES	ADVANTAGES	FIGURES
[73]	2007	ISSCC	Intel	*(substrate resonance noise)* BPF + amplifier (inv) + comparator (inv) + current generator Current generator dumps current into substrate during noise event	None	90 nm	Frequency 140 MHz	Not applicable to supply noise High power 2.42 mA		
[54]	2009	TVLSI	University of Minnesota	Base d on **[71]** Improved opamp: no external biasing, larger output swing, smaller input caps		1.8 µm				
[81]	2009	JSSC	UBC	Same as **[80]** with improved comparator Latch-based comparator		90 nm/ 1V	Frequency 500 MHz (up to 2GHz possible) Improvement 2% (min) and 14% (avg) compared to passive decap with series resistance (i.e. no switching)			

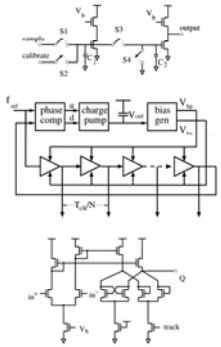
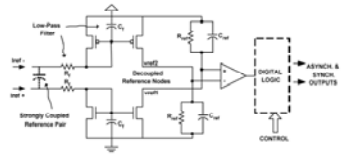
Appendix B
Supply Noise Detection Literature Survey

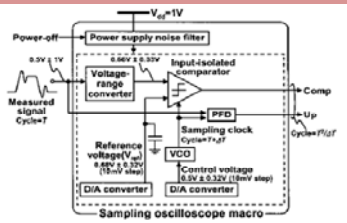
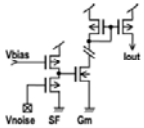
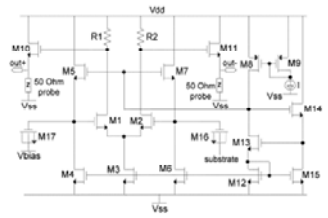
TABLE B.1 SUPPLY NOISE DETECTION LITERATURE SURVEY

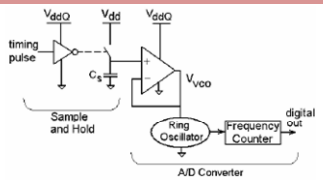
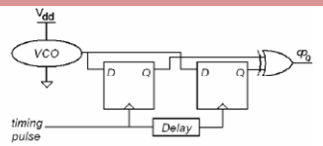
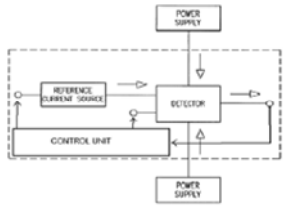
REF	YR	PUBLIC[N]	AFFIL[N]	*(NOISE TYPE)* DESCRIPTION	EXT SIGNALS	TECHN./ SUPPLY	PARAMETERS	CHALLENGES	ADVANTAGES	FIGURES
[129]	1993	Electronics Letters	Linkoping Institute of Technol. (P.Larsson)	*(any high frequency signal)* Sub-sampling - shifted sampling over periodic signal 3 Master-slave type S/H	Control signal for delay line	1 µm	S/H bandwidth 5 GHz	Signal must be periodic	High frequency signals can be measured on-chip	
[130]	1996	JSSC	Hitachi	*(substrate noise)* Clocked comparator; metastability point correlated to substrate noise Selectable high/low frequency	Separate power and ground Sampling clock	0.8 µm 3 V	Resolution 0.1 mV Clock freq 5 MHz Input range -20 mV to 10 mV	Not real time Analog I/O External sampling clock		
[142]	1998	VLSIC Digest of Technical Papers	Stanford	Based on Larsson 93 M/S S/H + 2 current mirrors to drive current off-chip Individual sampler calibration signals	External sampler clock Multiple calibration signals	0.25 µm	RC bandwidth 5 GHz	Signal must be periodic		

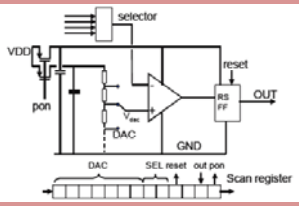
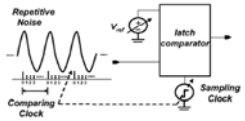
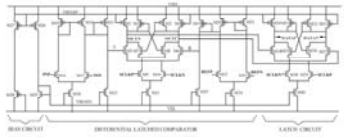
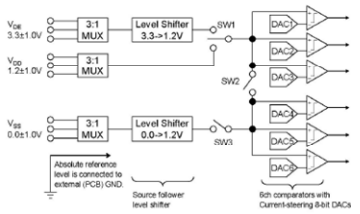
REF	YR	PUBLIC[N]	AFFIL[N]	*(NOISE TYPE)* DESCRIPTION	EXT SIGNALS	TECHN.	PARAMETERS	CHALLENGES	ADVANTAGES	FIGURES
[145]	1999	CICC	Hiroshima University	*(substrate noise)* Differential latch comparator Threshold measured to determine noise	Clock Reference signal Voltage input	0.4 µm	Time resolution 100 ps	Signal must be periodic		
[131]	2000	ICMTS	University of Tokyo	*(supply noise)* Voltage comparator Assuming successive approximation	External reference	0.6 µm	Supply 4.5 V Offset 100 mV Resolution 5 mV Time resolution 0.4 ns	Signal must be periodic Distributed single ended reference voltages		
[132]	2002	T-R	University of Texas at Dallas	Signal overshoots Sense amp to detect number of overshoots	If measured signal is supply, clean *VDD* is needed		8051 microprocessor ~ 1GHz	External signal required Does not give magnitude of overshoot		

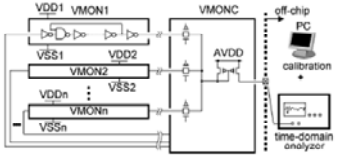
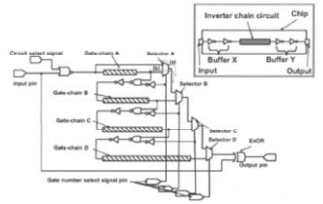
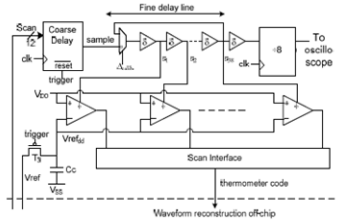
REF	YR	PUBLIC[N]	AFFIL[N]	*(NOISE TYPE)* DESCRIPTION	EXT SIGNALS	TECHN.	PARAMETERS	CHALLENGES	ADVANTAGES	FIGURES
[133]	2002	ISSCC	NEC	*(supply and substrate noise)* Based on Stanford 98 S/H + current mirrors + progressively shifted sampling Additional caps to charge share and increase input range *VDD*/*VSS* cleaned using decaps Delay line for clocks instead of PLL	Separate supply	0.13 µm	Supply 1.2 V S/H BW 6.4 GHz Clock freq 800 MHz Input range -0.3 V to *VDD*+0.3 V	Signal must be periodic		
[134]	2002	DATE	University of Bologna	Two FFs to two NAND gates to error detection cct Supply dip causes delay and unexpected output	*VDD* clean for error detection cctry? Clock/reset signals	0.35 µm	Supply 3.3 V	Clean *VDD* req'd? Slow due to large number of components?		

REF	YR	PUBLIC[N]	AFFIL[N]	*(NOISE TYPE)* DESCRIPTION	EXT SIGNALS	TECHN.	PARAMETERS	CHALLENGES	ADVANTAGES	FIGURES
[143]	2003	TVLSI	Columbia University	Sub-sampling Master-slave charge sharing S/H + Track and latch comparator based ADC + DLL based delay line for clocks Successive approximation	Off chip calibration	TSMC 0.25 µm	200 MHz	Signal must be periodic External calibration required	Clocks accurately generated on-chip with DLL	
[135]	2004	JSSC	Intel	*(supply and ground noise)* Voltage comparator Route reference currents then convert to voltage near comparator	Separate supply and ground External calibration	90 nm	Supply 1 V Clock freq ~ GHz Input range *VDD*/2	References still req'd Separate supply and ground lines req'd Noise threshold and duration must be preset Calibration No timing information	Real time	

REF	YR	PUBLIC[N]	AFFIL[N]	*(NOISE TYPE)* DESCRIPTION	EXT SIGNALS	TECHN.	PARAMETERS	CHALLENGES	ADVANTAGES	FIGURES
[140]	2004	VLSIC Digest of Technical Papers	NEC	Voltage range reducer + input isolated latch comparator + supply filter + phase-frequency-detector for timing Sub-sampling On-chip clock generation with DAC/VCO	Digital signals needed for 2 DACs	90 nm	Supply 1 V Clock freq 2.2 GHz Input range -0.5 *VDD* to 1.5*VDD* Filter cutoff 2 MHz	Signal must be periodic DACs need digital inputs (for Vref and VCO)	Timing info High-speed comparator O-chip filtering Large input range On-chip clock generation Power switch	
[136]	2005	JSSC	Kobe University	Small like standard cell Source follower + single transistor (Gm) + current output	Separate supply and ground	0.18 µm	Supply 1.8 and 2.5 V	Limited input range Separate supply	Can be very localized	 
[137]	2005	JSSC	Oregon State University	On-chip amplifiers	Separate supply and ground	0.35 µm	Bandwidth 700 MHz	Low bandwidth		

REF	YR	PUBLIC[N]	AFFIL[N]	*(NOISE TYPE)* DESCRIPTION	EXT SIGNALS	TECHN.	PARAMETERS	CHALLENGES	ADVANTAGES	FIGURES
[146] **[147]**	2004 2005	VLSIC JSSC	Stanford/ Rambus	Dual sampling S/H-based system with VCO based ADC Measures noise power spectral density (PSD) from autocorrelation of two variably spaced samples	Yes, higher than nominal	0.13 µm	~1 mV resolution; 100 mV noise, 1V supply 5 mV noise GHz bandwidth	Requires high-bandwidth, low leakage S/H circuits and analog buffers Difficult to scale S/H Post processing		
[155]	2005	ESSCIRC	Stanford/ Rambus	VCO-based ADC Short integration time Low-resolution measurement dithered to improve resolution	No	90 nm SOI	~1 mV resolution; 100 mV noise, 1V supply 5 mV noise GHz bandwidth	Large number of measurements required	No S/H	
[148]	2005	CICC	Intel	Differential current pair generates noise free currents Noise capacitively coupled to input of analog comparator – capacitors selectable for different frequencies	No, Dedicated *VSS*!! External calibration signals External thresholds	0.13 µm	Range +/- 105 mV Bandwidth 1 ns pulse width or 500 MHz	Needs calibration Dedicated ground line Control signals needed for detection thresholds		

REF	YR	PUBLIC[N]	AFFIL[N]	*(NOISE TYPE)* DESCRIPTION	EXT SIGNALS	TECHN.	PARAMETERS	CHALLENGES	ADVANTAGES	FIGURES
[141]	2006	ISSCC	Philips	*(supply noise)* Detects dips/peaks Unclocked comparator + successive approximation w/ DAC	DAC/SRlatch signals from scan chain	90 nm	Resolution 10 mV	Signal must be periodic	Standard cell compatible Fully digital On-chip *VDD* filter	
[149]	2007	VLSI-DAT	ITRI	Multiple samples per clock period Latch comparator	Vref	0.18 µms	10 MHz clock	Too slow	Signal does not need to be periodic	
[150]	2007	T-IM	Stanford/ National Taiwan University	Amplified signal + Latched comparator + successive approximation One phase differential clock	Externally programmable Vref	0.13 µm	Sensor bandwidth 1.6 GHz	Signal must be periodic External Vref	One phase clock Differential signaling does not require separate supply/ground	
[138]	2007	ISSCC	Fujitsu	Pseudo real-time (detection window determined before measurement) Sensor + Peak detection Sensor: 6 comparators+6 DACs+level shifters (akin to flash ADC)	External ground	90 nm	Sampling rate 480 MHz Resolution 10 mV Bandwidth 700 MHz	Detection window must be predetermined	Small number of comparators	

REF	YR	PUBLIC[N]	AFFIL[N]	*(NOISE TYPE)* DESCRIPTION	EXT SIGNALS	TECHN.	PARAMETERS	CHALLENGES	ADVANTAGES	FIGURES
[139]	2007	JSSC	Hitachi	Real time Ring oscillator + current mirror + off-ship DSP Small standard cell	Separate supply	90 nm	Time resolution 5 ns	Slow – bandwidth limited by transmission length Separate supply	Small Digital output Frequent calibration needed due to process variations	
[154]	2008	EPEP	Hitachi	Delay observed in inverter chain circuits Off-chip delay to voltage conversion	None	90 nm	Resolution 1 mV	Extensive post processing Low bandwidth		
[151]	2008	ESSCIRC	AMD/ University of Michigan	Multiple comparators (akin to flash ADC)	Vref and sampling trigger signal	0.13 µm	Time resolution 50 ps	User triggered sampling point External Vref	*VDD-VSS* detected	

References

[1] P. R. Gray, P. J. Hurst, S. H. Lewis and R. G. Meyer, *Analysis and Design of Analog Integrated Circuits*, John Wiley, 4th Edition, 2001.

[2] K. L. Shepard and V. Narayanan, "Noise in deep submicron digital design," in Proceedings of the 1996 IEEE/ACM International Conference on Computer-Aided Design, November 10,1996 - November 14, 1996, pp. 524-531.

[3] I. A. Ferzli and F. N. Najm, "Statistical estimation of circuit timing vulnerability due to leakage-induced power grid voltage drop," in 2004 International Conference on Integrated Circuit Design and Technology, ICICDT, may 17,2004 - may 20, 2004, pp. 17-24.

[4] H. H. Chen and J. S. Neely, "Interconnect and circuit modeling techniques for full-chip power supply noise analysis," IEEE Transactions on Components, Packaging, and Manufacturing Technology Part B: Advanced Packaging, vol. 21, pp. 209-215, 1998.

[5] K. Aygun, M. J. Hill, K. Eilert, K. Radhakrishnan, A. Levin, "Power Delivery for High-Performance Microprocessors," Intel Technology Journal, Electronic Package Technology Development, vol. 09(04), pp. 273-277, 2005.

[6] K. Arabi, R. Saleh and M. Xiongfei, "Power supply noise in SoCs: metrics, management, and measurement," *IEEE Design & Test of Computers,* vol. 24, pp. 236-44, 05. 2007.

[7] International Technology Roadmap for Semiconductors, 2010, http://www.itrs.net/Links/2010ITRS/Home2010.htm, accessed on May 10, 2012.

[8] R. H. Dennard, F. H. Gaensslen, Hwa-Nien Yu, V. L. Rideout, E. Bassous and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," IEEE J Solid State Circuits, vol. sc-9, pp. 256-68, 10. 1974.

[9] J. M. Rabaey, A. Chandrakasan and B. Nikolic, Digital Integrated Circuits: A Design Perspective, Prentice Hall, 2nd Edition, 2003.

[10] T. Lopez, R. Elfrich and E. Alarcon, Voltage Regulators for Next Generation Microprocessors, Springer, 2011.

[11] P. Larsson, "Power supply noise in future IC's: A crystal ball reading," in *Proceedings of the IEEE 1999 Custom Integrated Circuits Conference,* 1999, pp. 467-74.

[12] M. Nourani and A. R. Attarha, "Detecting signal-overshoots for reliability analysis in high-speed system-on-chips," *IEEE Trans. Reliab.,* vol. 51, pp. 494-504, 2002.

[13] A. Muhtaroglu, G. Taylor and T. Rahal-Arabi, "On-Die Droop Detector for Analog Sensing of Power Supply Noise," *IEEE J Solid State Circuits,* vol. 39, pp. 651-660, 2004.

[14] N. Weste and K. Eshragian, *Principles of CMOS VLSI Design*, Addison-Wesley, 1992.

[15] M. Saint-Laurent and M. Swaminathan, "Impact of power-supply noise on timing in high-frequency microprocessors," *IEEE Transactions on Advanced Packaging,* vol. 27, pp. 135-144, 2004.

[16] C. Tirumurti, S. Kundu, S. Sur-Kolay and Y. Chang, "A modeling approach for addressing power supply switching noise related failures of integrated circuits," in *Proceedings - Design, Automation and Test in Europe Conference and Exhibition, DATE 04, February 16,2004 - February 20,* 2004, pp. 1078-1083.

[17] S. Pant, D. Blaauw, V. Zolotov, S. Sundareswaran and R. Panda, "Vectorless analysis of supply noise induced delay variation," in *International Conference on Computer Aided Design,* 2003, pp. 184-91.

[18] T. B. Hook, M. Breitwisch, J. Brown, P. Cottrell, D. Hoyniak, C. Lam and R. Mann, "Noise margin and leakage in ultra-low leakage SRAM cell design," *IEEE Trans. Electron Devices,* vol. 49, pp. 1499-501, 08. 2002.

[19] M. Khellah, D. Khalil, D. Somasekhar, Y. Ismail, T. Karnik and V. De, "Effect of power supply noise on SRAM dynamic stability," in *2007 Symposium on VLSI Circuits, VLSIC, June 14,2007 - June 16,* 2007, pp. 76-77.

[20] J. A. Strak and H. Tenhunen, "Investigation of timing jitter in NAND and NOR gates induced by power-supply noise," in *13th IEEE International Conference on Electronics, Circuits and Systems,* 2007, pp. 1160-3.

[21] S. Zhao and K. Roy, "Estimation of switching noise on power supply lines in deep sub-micron CMOS circuits," in *The 13th International Conference on VLSI Design: Wireless and Digital Imaging in the Millennium, January 03,2000 - January 07,* 2000, pp. 168-173.

[22] M. G. Xakellis and F. N. Najm, "Statistical estimation of the switching activity in digital circuits," in *Proceedings of the 31st Design Automation Conference, June 06,1994 - June 10,* 1994, pp. 728-733.

[23] C. Wang and K. Roy, "Maximum power estimation for CMOS circuits using deterministic and statistical approaches," *IEEE Transactions on very Large Scale Integration (VLSI) Systems,* vol. 6, pp. 134-140, 1998.

[24] R. Burch, F. N. Najm, P. Yang and T. N. Trick, "A Monte Carlo approach for power estimation," *IEEE Transactions on very Large Scale Integration (VLSI) Systems,* vol. 1, pp. 63-71, 03. 1993.M. Butler, V. B. Jayaram, S. Kundu, N. V. Arvind, P. Sreeprakash and M. Hachinger, "A case study of IR-drop in structured at-speed testing," in *Proceedings International Test Conference 2003, September 30,2003 - October 02,* 2003, pp. 1098-1104.

[25] J. Wang, D. M. Walker, X. Lu, A. Majhi, B. Kruseman, G. Gronthoud, L. E. Villagra, d. W. van and S. Eichenberger, "Modeling power supply noise in delay testing," *IEEE Des. Test Comput.,* vol. 24, pp. 226-234, 2007.

[26] J. Saxena, K. M. Butler, V. B. Jayaram, S. Kundu, N. V. Arvind, P. Sreeprakash and M. Hachinger, "A case study of IR-drop in structured at-speed testing," in *Proceedings International Test Conference 2003, September 30,2003 - October 02,* 2003, pp. 1098-1104.

[27] K. R. Mistry, T. F. Fox, R. P. Preston, N. D. Arora, B. S. Doyle and D. E. Nelsen, "Circuit design guidelines for n-channel MOSFET hot carrier robustness," *IEEE Trans. Electron Devices,* vol. 40, pp. 1284-95, 07. 1993.

[28] B. Doyle, M. Bourcerie, J. -. Marchetaux and A. Boudou, "Interface state creation and charge trapping in the medium-to-high gate voltage range (Vd/2VgVd) during hot-carrier stressing of n-MOS transistors," *IEEE Trans. Electron Devices,* vol. 37, pp. 744-54, 03. 1990.

[29] M. Healy, F. Mohamood, H. S. Lee and S. K. Lim, "A unified methodology for power supply noise reduction in modern microarchitecture design," in *2008 Asia and South Pacific Design Automation Conference, ASP-DAC, March 21,2008 - March 24,* 2008, pp. 611-616.

[30] M. Benoit, S. Taylor, D. Overhauser and S. Rochel, "Power distribution in high-performance design," in *Proceedings of the 1998 International Symposium on Low Power Electronics and Design, August 10,1998 - August 12,* 1998, pp. [d]274-278.

[31] G. Bai, S. Bobba and I. N. Hajj, "Static timing analysis including power supply noise effect on propagation delay in VLSI circuits," in *38th Design Automation Conference, June 18,2001 - June 22,* 2001, pp. 295-300.

[32] M. Popovich, A. Mezhiba and E. G. Friedman, Power Distribution Networks with On-Chip Decoupling Capacitors, Springer, 2007.

[33] D. Blaauw, R. Panda and R. Chaudhry, "Design and Analysis of Power Distribution Networks," *Design of High-Performance Microprocessor Circuits,* Chandrakasan, Bowhill, and Fox, (Eds.), 2001.

[34] M. K. Gowan, L. L. Biro and D. B. Jackson, "Power considerations in the design of the alpha 21264 microprocessor," in *Proceedings of the 1998 35th Design Automation Conference,* 1998, pp. 726-731.

[35] P. E. Gronowski, W. J. Bowhill, R. P. Preston, M. K. Gowan, and Y. L. Allmon, "High-performance microprocessor design," *IEEE Journal of Solid State Circuits,* Vol. 33, No. 5, pp. 676-686, May 1998.

[36] A. Dharchoudhury, R. Panda, D. Blaauw, R. Vaidyanathan, B. Tutuianu and D. Bearden, "Design and analysis of power distribution networks in powerPC microprocessors," *Proceedings of the IEEE/ACM Design Automation Conference,* pp. 738-743, June 1998.

[37] D. Halliday, R. Resnick and J. Walker, Fundamentals of Physics, Wiley, 2007.

[38] L. D. Smith, R. E. Anderson, D. W. Forehand, T. J. Pelc and T. Roy, "Power distribution system design methodology and capacitor selection for modern CMOS technology," *IEEE Transactions on Advanced Packaging,* vol. 22, No. 3, august 1999.

[39] W. D. Becker, J. Eckhardt, R. Frech, G. Katopis, E. Klink, M. F. McAllister, T. G. McNamara, P. Muench, S. R. Richter and H. H. Smith, "Modeling, simulation and measurement of mid-frequency simultaneous switching noise in computer systems," *IEEE Transactions on Components, Packaging, and Manufacturing Technology, Part B,* vol. 21, pp. 286-297, May 1998.

[40] S. R. Nassif and O. Fakhouri, "Technology trends in power-grid-induced noise," in *2002 International Workshop on SystemLevel Interconnect Prediction, April 06,2002 - April 07,* 2002, pp. 55-59.

[41] C. W. Fok and D. L. Pulfrey, "Full-chip power-supply noise: The effect of on-chip power-rail inductance," in Workshop on Frontiers in Electronics, 2002, pp. 573-82.

[42] M. Popovich, E. G. Friedman, M. Sotman, A. Kolodny and R. M. Secareanu, "Maximum effective distance of on-chip decoupling capacitors in power distribution grids," in GLSVLSI'06 - 2006 ACM Great Lakes Symposium on VLSI, April 30,2006 - may 02, 2006, pp. 173-179.

[43] A. V. Mezhiba and E. G. Friedman, "Scaling trends of on-chip power distribution noise," *IEEE Transactions on very Large Scale Integration (VLSI) Systems,* vol. 12, pp. 386-394, 2004.

[44] R. R. Tummala, E. J. Rymaszewski, and A. G. Klopfenstein, Microelectronics

Packaging Handbook. Boca Raton, FL: Chapman & Hall, 1997.

[45] S. Gong, H. Hentzell, S. T. Persson, H. Hesselbom, B. Lofstedt and M. Hansen, "Packaging impact on switching noise in high-speed digital systems," *IEE Proceedings of Circuits, Devices and Systems,* vol. 145, pp. 446-452, 1998.

[46] W. S. Song and L. A. Glasser, "Power distribution techniques for VLSI circuits," *IEEE J Solid State Circuits,* vol. SC-21, pp. 150-6, 02. 1986.

[47] H. B. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI,* Addison-Wesley, 1990.

[48] L. Huang, D. Weiss and M. Rausch, "Effects of on-package decoupling on microprocessor power delivery design," *IEEE 5th Topical Meeting Elect. Perform. Electron. Packag.,* Napa, CA, pp. 86-89, October 1996.

[49] R. F. Sechler and G. F. Grohoski, "Design at the system level with VLSI CMOS," *IBM Journal of Research and Development,* vol. 39, pp. 5-22, 01. 1995.

[50] P. Larsson, "Parasitic resistance in an MOS transistor used as on-chip decoupling capacitance," *IEEE Journal of Solid State Circuits,* vol. 32, no. 4, April 1997.

[51] J. Gu, H. Eom, J. Keane and C. Kim, "Sleep transistor sizing and adaptive control for supply noise minimization considering resonance," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 17, no. 9, September 2009.

[52] J. Xu, P. Hazucha, M. Huang, P. Aseron, F. Paillet, G. Schrom, J. Tschanz, C. Zhao, V. De, T. Karnik and G. Taylor, "On-die supply-resonance suppression using band-limited active damping," in *2007 IEEE International Solid-State Circuits Conference,* 2007, pp. 286-603.

[53] B. Garben, G. A. Katopis and W. D. Becker, “Package and chip design optimization for mid-frequency power distribution decoupling,” *Electrical Performance of Electronic Packaging*, 2002.

[54] J. Gu, R. Harjani and C. H. Kim, "Design and implementation of active decoupling capacitor circuits for power supply regulation in digital ICs," *IEEE Transactions on very Large Scale Integration (VLSI) Systems,* vol. 17, pp. 292-301, 2009.

[55] “Intel Core i7 Processor Extreme Edition and Intel i7 Processor,” Datasheet, Volume 1, November 2008.

[56] D. J. Herrell and B. Baker, “Modeling of power distribution systems for high-performance microprocessors,” *IEEE Transactions of Advanced Packaging,* vol. 22, no. 3, August 1999.

[57] Y. Jiang, K. Cheng, “Estimation of maximum power supply noise for deep sub-micron designs,” *ISLPED,* Monterey, CA, 1998.

[58] Larsson, "di/dt noise in CMOS integrated circuits," *Analog Integr. Cir. Signal Proc.,* vol. 14, pp. 113-129, 1997.

[59] S. N. Lalgudi, M. Swaminathan and Y. Kretchmer, "On-chip power-grid simulation using latency insertion method," *IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications,* vol. 55, pp. 914-31, 04. 2008.

[60] J. Choi, M. Swaminathan, N. Do and R. Master, “Modeling of power supply noise in large chips with nonlinear circuits,” *Electrical Performance of Electronic Packaging,* pp. 257-260, 2002.

[61] G. A. Katopis, "Delta-I noise specification for a high-performance computing machine," Proc IEEE, vol. 73, pp. 1405-15, 1985.

[62] L. W. Schaper and D.I. Amey, “Improved electrical performance required for future MOS packaging,” *IEEE Transactions on Components, Hybrids, and Manufacturing Technology,* vol. 6, no. 3, pp. 283-289, September 1983.

[63] R. Senthinanathan, G. Tubbs and M. Schuelein, “Negative feedback influence in simultaneously switching CMOS outputs,” *IEEE Custon Integrated Circuits Conference,* pp. 5.4.1-5.4.5, 1988.

[64] L. A. Arledge and W. T. Lynch, “Scaling and performance implications for power supply and other signal routing constraints imposed by I/O pad limitations,” *Symposium on Ic/Package Design Integration,* pp. 45-50, 1998.

[65] K. C. Saraswat and F. Mohammadi, "Effect of scaling of interconnections on the time delay of VLSI circuits," *IEEE Trans. Electron Devices,* vol. ED-29, pp. 645-50, 04. 1982.

[66] K.Sham and R.Harjani, “I/O Staggering for Low-Power Jitter Reduction,” European Microwave Conference, 2008.

[67] C. J. Akl, R. A. Ayoubi and M. A. Bayoumi, “An Effective Staggered-Phase Damping Technique for Suppressing Power-Gating Resonance Noise during Mode Transition,” 10th Int’l Symposium on Quality Electronic Design, 2009.

[68] R. Rimolo-Donadioa, A. J. Acosta and W. Krautschneider, “Asynchronous Staggered Set/Reset Techniques for Low-Noise Applications*,” IEEE International Symposium on Circuits and Systems*, 2007.

[69] D. K. Su, M. J. Loinaz, S. Masui and B. A. Wooley, "Experimental results and modeling techniques for substrate noise in mixed-signal integrated circuits," *IEEE J Solid State Circuits,* vol. 28, pp. 420-30, 04. 1993.

[70] M. Ang, R. Salem and A. Taylor, "An on-chip voltage regulator using switched decoupling capacitors," in *2000 IEEE International Solid-State Circuits Conference 47th Annual ISSCC,* 2000, pp. 438-439.

[71] T. Tsukada, Y. Hashimoto, K. Sakata, H. Okada and K. Ishibashi, "An on-chip active decoupling circuit to suppress crosstalk in deep-submicron CMOS mixed-signal SoCs," *IEEE J Solid State Circuits,* vol. 40, pp. 67-79, 01. 2005.

[72] T. 1. Kazama, T. 1. Nakura, M. 1. Ikeda and K. 1. Asada, "Design of active substrate noise canceller using power supply di/dt detector," in *2007 Asia and South Pacific Design Automation Conference,* 2007, pp. 2.

[73] J. Xu, P. Hazucha, M. Huang, P. Aseron, F. Paillet, G. Schrom, J. Tschanz, C. Zhao, V. De, T. Karnik and G. Taylor, "On-die supply-resonance suppression using band-limited active damping," in *2007 IEEE International Solid-State Circuits Conference,* 2007, pp. 286-603.

[74] C. F. Webb, C. J. Anderson, L. Sigal, K. L. Shepard, J. S. Liptay, J. D. Warnock, B. Curran, B. W. Krumm, M. D. Mayo, P. J. Camporese, E. M. Schwarz, M. S. Farrell, P. J. Restle, Averill,R.M.,,III, T. J. Siegel, W. V. Huott, Y. H. Chan, B. Wile, P. G. Emma, D. K. Beece, Ching-Te Chuang and C. Price, "A 400 MHz S/390 microprocessor," in *Digest of Technical Papers,* 1997, pp. 168-9.

[75] B. Garben, R. Frech, J. Supper and M. F. McAllister, "Frequency dependencies of power noise," *IEEE Transactions on Advanced Packaging,* vol. 25, pp. 166-73, 05. 2002.

[76] S. Rusu and G. Singer, "First IA-64 microprocessor," *IEEE J Solid State Circuits,* vol. 35, pp. 1539-1544, 2000.

[77] R. Heald, K. Aingaran, C. Amir, M. Ang, M. Boland, P. Dixit, G. Gouldsberry, D. Greenley, J. Grinberg, J. Hart, T. Horel, Wen-Jay Hsu, J. Kaku, C. Kim, S. Kim, F. Klass, H. Kwan, G. Lauterbach, R. Lo, H. McIntyre, A. Mehta, D. Murata, S. Nguyen, Yet-Ping Pai, S. Patel, K. Shin, K. Tam, S. Vishwanthaiah, J. Wu, G. Yee and E. You, "A third-generation SPARC V9 64-b microprocessor," in *Digest of Technical Papers,* 2000, pp. 1526-38.

[78] C. Giacomotto, R.P. Masleid and A. Harada, “Four-State Switched Decoupling Capacitor System for Active Power Stabilizer,” *US Patent # 6,744,242*, 2004.

[79] J. Gu, H. Eom and C. H. Kim, "A switched decoupling capacitor circuit for on-chip supply resonance damping," in *2007 Symposia on VLSI Technology and Circuits,* 2007, pp. 31-2.

[80] X. Meng, K. Arabi and R. Saleh, "A novel active decoupling capacitor design in 90nm CMOS," in *2007 IEEE International Symposium on Circuits and Systems, ISCAS 2007,* 2007, pp. 657-660.

[81] X. Meng and R. Saleh, "An improved active decoupling capacitor for “hot-spot” supply noise reduction in ASIC designs," *IEEE J Solid State Circuits,* vol. 44, pp. 584-93, 02. 2009.

[82] J. Gu, H. Eom and C. H. Kim, "A switched decoupling capacitor circuit for on-chip supply resonance damping,"On-Chip Supply Noise Regulation Using a Low-Power Digital Switched Decoupling Capacitor Circuit,” *IEEE J Solid State Circuits,* vol. 44, pp. 1765-75, 06. 2009.

[83] X. Meng, R. Saleh and S. Wilton, “Charge-borrowing decap: a novel circuit for removal of local supply noise violations,” *IEEE Custom Integrated Circuits Conference,* pp. 25-28, 2009.

[84] J. Rius and M. Meijer, "A high-frequency nonquasi-static analytical model including gate leakage effects for on-chip decoupling capacitors," *IEEE Transactions on Advanced Packaging,* vol. 29, pp. 88-97, 2006.

[85] P. Larsson, "Resonance and damping in CMOS circuits with on-chip decoupling capacitance," *IEEE Trans. Circuits Syst. I Fundam. Theor. Appl.,* vol. 45, pp. 849-58, 08, 1998.

[86] X. Meng, R. Saleh and K. Arabi, "Layout of decoupling capacitors in IP blocks for 90-nm CMOS," *IEEE Transactions on very Large Scale Integration (VLSI) Systems,* vol. 16, pp. 1581-8, 11, 2008.

[87] S. Mukhopadhyay, C. Neau, R. T. Cakici, A. Agarwal, C. H. Kim and K. Roy, "Gate leakage reduction for scaled devices using transistor stacking," *IEEE Transactions on very Large Scale Integration (VLSI) Systems,* vol. 11, pp. 716-30, 08, 2003.

[88] S. Dimitrijev, Principles of Semiconductor Devics, Oxford University Press, 2nd Ed., 2012.

[89] A. Amerasekera and C. Duvvury, *ESD in Silicon Integrated Circuits,* 2nd ed. Hoboken, NJ: Wiley, 2002.

[90] O. Semenov, H. Sarbishaei and M. Sachdev, *ESD Protection Device and Circuit Design for Advanced CMOS Technologies,* Springer, 2008.

[91] Y. Chen, H. Li, K. Roy and C. Koh, "Gated decap: Gate leakage control of on-chip decoupling capacitors in scaled technologies," *IEEE Transactions on very Large Scale Integration (VLSI) Systems,* vol. 17, pp. 1749-1752, 2009.

[92] H. Samavati, A. Hajimiri, A. R. Shahani, G. N. Nasserbakht and T. H. Lee, "Fractal capacitors," *IEEE J. Solid State Circuits,* vol. 33, pp. 2035-2041, 1998.

[93] R. Aparicio and A. Hajmiri, “Capacity limits of matching properties of integrated capacpitors,” *IEEE Journal of Solid State Circuits,* vol. 37, no. 3, March 2002.

[94] T. H. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits,* 2nd Edition, Cambridge University Press, New York, New York, 2004.

[95] C. Pei, “A novel, low-cost deep trench decoupling capacitor for high-performance, low-power bulk CMOS applications,” *9th International Conference on Solid-State and Integrated-Circuit Technology,* 2008.

[96] K. Shimakazi and T. Okumura, "Minimum decap allocation technique based on simultaneous switching of nanoscale SoC," *Custom Integrated Circuits Conference,* 2009.

[97] "Decoupling capacitor sizing and placement," U S Patent # 6,898,769.

[98] M. Popovich, M. Sotman and R. Secareanu, "Maximum effective distance of on-chip decoupling capacitors in power distribution grids," *Great Lakes Symposium on VLSI,* 2006.

[99] H. Su, S. Sapatnekar and S. Nassif, "Optimal decoupling capacitor sizing and placement for standard-cell layout designs," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 22, no. 4, April 2003.

[100] M. D. Pant, P. Pant and D. S. Wills, "On-chip decoupling capacitor optimization using architectural level prediction," *IEEE Transactions of Very Large Scale Integration (VLSI) Systems,* vol. 10., no. 3, June 2002.

[101] J. Choi, S. Chun, N. Na, M. Swaminathan and L. Smith, "A methodology for the placement and optimization of decoupling capacitors for gigahertz systems," *Thirteenth International Conference on VLSI Design,* pp. 156-161, 2000.

[102] S. Zhao, K. Roy and C. Koh, "Decoupling capacitance and its application to power-supply noise-aware floorplanning," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 21, no. 1, January 2002.

[103] H.Chen and S. Schuster, "On-chip decoupling capacitor optimization for high-performance VLSI design," *International Symposium on VLSI Technology, Systems, and Applications,* pp. 99-103, 1995.

[104] D. Pozar, Microwave Engineering, Wiley, 2004.

[105] "Advanced RF Calibration Techniques," Presentation by Anthony Lord at Cascade Microtech, available at: http://ekv.epfl.ch/files/content/sites/ekv/files/mos-ak/wroclaw/MOS-AK_AL.pdf, accessed June 13, 2012.

[106] M. C. A. M. Koolen, J. A. M. Geelan and M. P. J. G. Versleijen, "An improved de-embedding technique for on-wafer high-frequency characterization," *IEEE Bipolar Circuits and Technology Meeting,* pp. 188-191, 1991.

[107] H. Cho and D. E. Burk, "A three-step method for the de-embedding of high-frequency s-parameter measurements," *IEEE Transactions on Electron Devices,* vol. 38, no. 6, June 1991.

[108] E. P. Vandamme, D. M. M. P. Schreurs and C. Dinther, “Improved three-step de-embedding method to accurately account for the influence of pad parasitics in silicon on-wafer RF test-structures,” *IEEE Transactions on electron Devices,* vol. 48, no. 4, pp. 737-742, April 2001.

[109] P. J. van Wijnen, H. R. Claessen, and E. A. Wolsheimer, “A new straightforward calibration and correction procedure for “on-wafer” high frequency S-parameter measurements (45 MHz–18 GHz),” in *Proc. IEEE Bipolar Circuits and Technology Meeting,* pp. 70-73, 1987.

[110] R. Torres-Torres, R Murphy-Arteaga and J. A. Reynoso-Hernandez, “Analytical model and parameter extraction to account for the pad parasitic in RF-CMOS,” *IEEE Transactions on Electron Devices,* vol. 52, no. 7, July 2005.

[111] I. M. Kang, S. –J. Jung, T. –H. Choi, J. –H. Jung, C. Chung, H. –S. Kim, H. Oh, H. W. Lee, G. Jo, Y. –K. Kim, H. –G. Kim and K. –M Choi, “Five-step (pad–pad short–pad open–short–open) de-embedding method and its verification,” *IEEE Electron Device Letters,* vol. 30, no. 4, pp. 398-400, April 2009.

[112] G. T. Sasse, “Reliability engineering in RF CMOS,” PhD thesis, university of Twente, The Netherlands, 2008.

[113] X. Jin, J. –J. Ou, C. –H. Chen, W. Liu, M. J. Deen, P. R. Gray and C. Hu, “An effective gate resistance model for CMOS RF and noise modeling,” *International Electron Device Meeting,* pp. 961-964, 1998.

[114] E. Torres-Rios, R. Torres-Torres, G. Valdovinos-Fierro and E. A. Gutierrez-D, “A method to determine the gate bias-dependent and gate bias-independent components of MOSFET series resistance from S-Parameters,” *IEEE Transactions on Electron Devices,* vol. 53, no. 3, pp. 571-573, March 2006.

[115] S. Lee and H. K. Yu, “A new extraction method for BSIM3v3 model parameters of RF silicon MOSFETs,” *International Conference on Microelectronic Test Structures,* vol. 12, pp. 95-98, March 1999.

[116] Y. Cheng, M. J. Deen and C. –H. Chen, “MOSFET modeling for RFIC design,” *IEEE Transactions on Electron Devices,* vol. 52, no. 7, pp. 1286-1303, July 2005.

[117] W. Liu, R. Gharpurey, M. C. Chang, U. Erdogan, R. Aggarwal and J. P. Mattia, , "RF MOSFET modeling accounting for distributed substrate and channel resistance with emphasis on the BSIM3v3

SPICE model," *IEEE International Electron Devices Meeting, Technical Digest,* pp. 309-312, December 1997.

[118] D. R. Pehlke, M. Schroter, A. Burstein, M. Matloubian and M. F. Chang, "High frequency application of MOS compact model and their development for scalable RF model libraries," *Proceedings of the IEEE Custon Integrated Circuits Conference,* pp. 219-222, May 1998.

[119] J.-J. Ou, X. Jin, I. Ma, H. Chenming and P. R. Gray, "CMOS RF modeling for GHz communication IC's," *Symposium on VLSI Technology, Digest of Technical Papers,* pp. 94-95, June 1998.

[120] R. S. Guindi and F. N. Najm , "Design techniques for gate-leakage reduction in CMOS circuits," *Proceedings of the 4th International Symposium on Quality Electronic Design,* pp. 61-65, 2003.

[121] M. M. Mano and C. R. Kine, Logic and Computer Design Fundamentals, 4th ed., Prentice Hall, 2007.

[122] S. B. Wijeratne, N. Siddaiah, S. K. Mathew, M. A. Anders, R. K. Krishnamurthy, J. Anderson, M. Ernest and M. Nardin, "A 9-GHz 65-nm Intel® Pentium 4 Processor Integer Execution Unit," *IEEE Journal of Solid-State Circuits*, vol.42, no.1, pp.26-37, 2007.

[123] M. Badaroglu, *et al.*, "Methodology and Experimental Verification for Substrate Noise Reduction in CMOS Mixed Signal ICs with Synchronous Digital Circuits", *IEEE Journal of Solid-State Circuits*, Vol. 37, No. 11, pp. 1383-1395, Nov. 2002.

[124] P. Parra, A. Acosta, R. Jiménez, M. Valencia. "Selective Clock-Gating for Low-Power Synchronous Counters", *Journal of Low Power Electronics* 1, pp. 11-19, 2005.

[125] P. Vuillod, L. Benini, A. Bogliolo, and G. D. Micheli, "Clockskew optimization for peak current reduction", *IEEE Int. Symposium Low Power Electronics and Design*, pp. 265-270, 1996.

[126] N. Weste and D. Harris, CMOS VLSI Design: A Circuits and Systems Perspective, 4th ed. Addison Wesley. 2010.

[127] A. Rushton, VHDL for Logic Synthesis, 3rd ed., Wiley, 2011.

[128] Arizona State Predictive Models, http://ptm.asu.edu/, accessed November 16, 2011.

[129] P. Larsson and C. 1. Svensson, "Measuring high-bandwidth signals in CMOS circuits," *Electron. Lett.,* vol. 29, pp. 1761-2, 1993.

[130] K. Makie-Fukuda, T. Anbo, T. Tsukada, T. Matsuura and M. Hotta, "Voltage-comparator-based measurement of equivalently sampled substrate noise waveforms in mixed-signal integrated circuits," *IEEE J Solid State Circuits,* vol. 31, pp. 726-731, 1996.

[131] H. Aoki, M. Ikeda and K. Asada, "On-chip voltage noise monitor for measuring voltage bounce in power supply lines using a digital tester," in *Proceedings of the 2000 International Conference on Microelectronic Test Structures,* 2000, pp. 112-17.

[132] M. Nourani and A. R. Attarha, "Detecting signal-overshoots for reliability analysis in high-speed system-on-chips," *IEEE Trans. Reliab.,* vol. 51, pp. 494-504, 2002.

[133] M. Takamiya, M. Mizuno and K. Nakamura, "An on-chip 100 GHz-sampling rate 8-channel sampling oscilloscope with embedded sampling clock generator," in *2002 IEEE International Solid-State Circuits Conference,* 2002, pp. 182-183.

[134] C. Metra, L. Schiano, M. Favalli and B. Ricco, "Self-checking scheme for the on-line testing of power supply noise," in *Proceedings 2002 Design, Automation and Test in Europe Conference and Exhibition,* 2002, pp. 832-6.

[135] A. Muhtaroglu, G. Taylor and T. Rahal-Arabi, "On-Die Droop Detector for Analog Sensing of Power Supply Noise," *IEEE J Solid State Circuits,* vol. 39, pp. 651-660, 2004.

[136] M. Nagata, T. Okumoto and K. Taki, "A built-in technique for probing power supply and ground noise distribution within large-scale digital integrated circuits," *IEEE J Solid State Circuits,* vol. 40, pp. 813-819, 2005.

[137] B. E. Owens, S. Adluri, P. Birrer, R. Shreeve, S. K. Arunachalam, K. Mayaram and T. S. Fiez, "Simulation and measurement of supply and substrate noise in mixed-signal ICs," *IEEE J Solid State Circuits,* vol. 40, pp. 382-91, 02. 2005.

[138] T. Sato, A. Inoue, T. Shiota, T. Inoue, Y. Kawabe, T. Hashimoto, T. Imamura, Y. Murasaka, M. Nagata and A. Iwata, "On-die supply-voltage noise sensor with real-time sampling mode for low-power processor applications," in *2007 IEEE International Solid-State Circuits Conference,* 2007, pp. 290-603.

[139] Y. Kanno, Y. Kondoh, T. Irita, K. Hirose, R. Mori, Y. Yasu, S. Komatsu and H. Mizuno, "In-situ measurement of supply-noise maps with millivolt accuracy and nanosecond-order time resolution," *IEEE J Solid State Circuits,* vol. 42, pp. 784-789, 2007.

[140] M. Takamiya and M. 1. Mizuno, "A sampling oscilloscope macro toward feedback physical design methodology," in *Digest of Technical Papers,* 2004, pp. 240-3.

[141] V. Petrescu, M. Pelgrom, H. Veendrick, P. Pavithran and J. Wieling, "A signal-integrity self-test concept for debugging nanometer CMOS ICs," in *Digest of Technical Papers,* 2006, pp. 10.

[142] R. Ho, B. Amrutur, K. Mai1, B. Wilburn, T. Mori and M. Horowitz, "Applications of on-chip samplers for test and measurement of integrated circuits," in *1998 Symposium on VLSI Circuits Digest of Technical Papers,* 1998, pp. 138-9.

[143] Y. Zheng and K. L. Shepard, "On-chip oscilloscopes for noninvasive time-domain measurement of waveforms in digital integrated circuits," *IEEE Transactions on very Large Scale Integration (VLSI) Systems,* vol. 11, pp. 336-344, 2003.

[144] J. Cerven and R. Kukuca, "PC sampling oscillograph for subnanosecond time region," *Electron. Lett.,* vol. 26, pp. 1950-1952, 1990.

[145] M. Nagata, Y. Kashima, D. Tamura, T. Morie and A. Iwata, "Measurements and analyses of substrate noise waveform in mixed signal IC environment," in *Proceedings of the IEEE 1999 Custom Integrated Circuits Conference,* 1999, pp. 575-8.

[146] E. Alon, V. Stojanovic and M. Horowitz, "Circuits and techniques for high-resolution measurement of on-chip power supply noise," in *2004 Symposium on VLSI Circuits, Digest of Technical Papers,* 2004, pp. 102-105.

[147] E. Alon, V. Stojanovic and M. A. Horowitz, "Circuits and techniques for high-resolution measurement of on-chip power supply noise," *IEEE J Solid State Circuits,* vol. 40, pp. 820-8, 04. 2005.

[148] C. Chansungsan, "Auto-referenced on-die power supply noise measurement circuit," in *Proceedings of the IEEE 2005 Custom Integrated Circuits Conference,* 2005, pp. 39-42.

[149] Hsien-Hung Wu, Chin-Hsin Fu, Yaw-Feng Wang, Pei-Wen Luo, Yen-Ming Chen, Liang-Chia Cheng and Cheng-Hsing Chien, "Characterization of supply and substrate noises in CMOS digital circuits," in *2007 International Symposium on VLSI Design, Automation, and Test Organization,* 2007, pp. 244-7.

[150] C. Iorga, Y. Lu and R. W. Dutton, "A built-in technique for measuring substrate and power-supply digital switching noise using PMOS-based differential sensors and a waveform sampler in system-

on-chip applications," *IEEE Transactions on Instrumentation and Measurement,* vol. 56, pp. 2330-2337, 2007.

[151] S. Pant and D. Blaauw, "Circuit techniques for suppression and measurement of on-chip inductive supply noise," in *ESSCIRC 2008 - 34th European Solid-State Circuits Conference,* 2008, pp. 134-7.

[152] Y. Zheng and K. L. Shepard, "On-chip oscilloscopes for noninvasive time-domain measurement of waveforms in digital integrated circuits," *IEEE Transactions on very Large Scale Integration (VLSI) Systems,* vol. 11, pp. 336-344, 2003.

[153] Hwang-Cherng Chow and Zhi-Hau Hor, "A high performance peak detector sample and hold circuit for detecting power supply noise," in *APCCAS 2008 - 2008 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS),* 2008, pp. 672-5.

[154] Y. Uematsu, H. Osaka, E. Suzuki, M. Yagyu and T. Saito, "Measurement techniques for on-chip power supply noise waveforms based on fluctuated sampling delays in inverter chain circuits," in *2008 IEEE 17th Conference on Electrical Performance of Electronic Packaging (EPEP),* 2008, pp. 69-72.

[155] V. A. Abramzon, E. Alon, B. Nezamfar and M. Horowitz, "Scalable circuits for supply noise measurement," in *31st European Solid-State Circuits Conference,* 2005, pp. 463-6.

[156] P. E. Allen and D. R. Holberg, *CMOS Analog Circuit Design,* Oxford University Press, 2nd Edition, 2002.