DESIGN AND ANALYSIS OF POWER DISTRIBUTION NETWORKS IN VLSI CIRCUITS

by

Sanjay Pant

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering)
in The University of Michigan
2008

Doctoral Committee:

Professor David Blaauw, Chair
Professor John P. Hayes
Associate Professor Igor L. Markov
Associate Professor Dennis M. Sylvester
Eli Chiprout, Intel

© Sanjay Pant
All Rights Reserved
2008

To my Parents

ACKNOWLEDGEMENTS

Looking back at my stay in Ann Arbor, I see a number of people who have helped me in this endeavor. I would like to take this opportunity to express my sincere gratitude to all of them.

First of all I would like to thank my advisor Professor David Blaauw. I have always considered myself to be very fortunate to have had him as my advisor. I have learnt a lot from him. His enthusiasm and energy has been the source of my motivation and inspiration all throughout my journey through graduate school. This work would not have been possible without his guidance and encouragement.

I am thankful to Professor Hayes, Professor Markov, Professor Sylvester and Dr. Chiprout for serving on my committee and for their valuable feedback. I would also like to acknowledge all my co-authors who have helped me grow professionally.

I thank Edward Chusid, Bertha Wachsman, Amanda Brown, Denise DuPrie and Beth Stalnaker for the help provided in managing a lot of different things. Special thanks to Amanda for the cookies during paper-submission and tapeout deadlines.

Finally, I would like to thank my family and friends. My family has been a constant source of support and encouragement, and all credit goes to them for whatever I accomplish. Thanks to Visvesh, Ashish, Shidhartha, Sarvesh, Kavi, Ravi, Manav, Carlos, Rajeev and other friends who made my stay in Ann Arbor, some of the best years of my life.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF FIGURES vii

LIST OF TABLES xii

CHAPTER I

INTRODUCTION 1

1.2 Power-grid modeling and voltage-drop analysis 6

1.3 Impact of supply noise on circuit performance 11

1.4 Decoupling-capacitance modeling and allocation 12

1.5 Ringing in power supply networks 17

1.6 Active supply-voltage regulation and supply-noise measurement 21

1.7 Contributions of this work and organization 23

CHAPTER II

VECTORLESS ANALYSIS OF SUPPLY-NOISE INDUCED DELAY VARIATION 28

2.1 Introduction 28

2.2 Delay model for supply fluctuations 31

2.3 Voltage-drop sensitivity computation 38

2.4 Block-current constraints 42

2.5 Overall path-based delay-maximization formulation 46

2.6 Block-based circuit-delay model 49

2.7 Experimental results 55

2.8 Conclusions 63

CHAPTER III

POWER-SUPPLY-DROP ANALYSIS 64

3.1 Introduction 64

3.2 Constraint-based early supply-drop analysis 65

3.3 Statistical supply-drop analysis 68

3.4 Experimental results 79

3.5 Conclusions 85

CHAPTER IV

TIMING-AWARE DECOUPLING-CAPACITANCE ALLOCATION IN POWER SUPPLY NETWORKS 86

4.1 Introduction 86

4.2 Proposed global-optimization approach 89

4.3 Greedy path-based approach 100

4.4 Experimental results 101

4.5 Conclusions 104

CHAPTER V

INDUCTANCE, LOCALITY AND RESONANCE IN POWER SUPPLY NETWORKS 105

5.1 Introduction 105

5.2 Full-wave on-die inductive effects 107

5.3 Mid-size model and capacitive effects 112

5.4 Complete package-die model 117

5.5 Conclusions and CAD implications 121

CHAPTER VI

AN ANALOG ACTIVE DECAP CIRCUIT FOR INDUCTIVE-SUPPLY-NOISE SUPPRESSION 123

6.1 Introduction 123

6.2 Proposed active decap circuit 125
6.3 Opamp design 133
6.4 Simulation results 136
6.5 Conclusions... 143

CHAPTER VII
DIGITAL CIRCUIT-TECHNIQUES FOR ACTIVE INDUCTIVE-SUPPLY-NOISE SUPPRESSION 144
7.1 Introduction... 144
7.2 Charge-injection-based active decoupling circuit............................ 146
7.3 High-voltage charge-pump-based active decoupling circuit 165
7.4 High-voltage shunt-supply-based active decoupling circuit............ 170
7.5 Digital on-chip oscilloscope for supply-noise measurement............ 174
7.6 Conclusions... 177

CHAPTER VIII
CONCLUSIONS AND FUTURE DIRECTIONS 179
8.1 Contributions 179
8.2 Future directions 183

BIBLIOGRAPHY.. 185

LIST OF FIGURES

Figure 1.1: On-die IR voltage map of the PowerPC® microprocessor (source: Motorola) 2

Figure 1.2: Power delivery network of a high-performance microprocessor (source:Intel) 4

Figure 1.3: Device-capacitance modeling in power-grid analysis [68] 15

Figure 1.4: Simplified model of a power supply network 18

Figure 1.5: Transient and frequency response showing first and second droops 19

Figure 1.6: Current excitation at resonance frequency (top) and voltage response (bottom) (source: Motorola) 19

Figure 2.1: A driver-receiver pair in a non-ideal supply network 32

Figure 2.2: A path in a power supply network with worst-case voltage shifts causing the maximum path delay 33

Figure 2.3: A driver-receiver pair in a non-ideal supply network (a), propagation delay (b) and output transition time (c) 34

Figure 2.4: Variation of rise/fall propagation delays of a gate with respect to Vddg, Vssg, Vddin and Vssin 36

Figure 2.5: Library-characterization error for delay and transition time for more than 35,000 sample points 37

Figure 2.6: (a) Power grid as a linear system and (b) Time-varying block-current discretization into time-steps 40

Figure 2.7: Correlation between Multiplier and ALU block-currents (a) and correlation between IF-stage in cycle t and ID-stage in cycle t+1 (b) 45

Figure 2.8: Path-based delay-maximization accounting for transition-time variations 48

Figure 2.9: An example of a combinational circuit 49

Figure 2.10: Illustration of the removal of the max function... 51

Figure 2.11: Overall block-based circuit-delay-maximization flow 53

Figure 2.12: An illustration of delay-based circuit pruning to reduce the problem size .. 55

Figure 2.13: Comparison of block-based NLP with random runs for c880 59

Figure 2.14: Step response at a node due to different blocks (resonance frequency = 50MHz).. 60

Figure 2.15: Time-varying total chip-current obtained from the block-based delay maximization 62

Figure 3.1: Flow diagram of proposed statistical approach... 68

Figure 3.2: Variation of a block-current with time 72

Figure 3.3: PDF (a) and auto-correlation (b) of a block-current...................................... 73

Figure 3.4: Probability distribution of the overall worst-case drop 81

Figure 4.1: A combinational circuit in a power distribution network 89

Figure 4.2: Illustration of gradient computation using the modified adjoint-sensitivity method.. 97

Figure 4.3: Overall global-optimization flow 99

Figure 4.4: Optimization flow of path-based greedy algorithm 100

Figure 4.5: Decap area vs circuit-delay trade-off curve for c432 102

Figure 5.1: Illustration of the PEEC model of a wire segment...................................... 108

Figure 5.2: A 3D PEEC simulation of 500mx500m grid (left) compared to a 2D-modeling approach (right) 110

Figure 5.3: Voltage map generated due to high-frequency current transients 111

Figure 5.4: A 2mm X 2mm section of on-die grid attached to the middle of package shadow........ 113

Figure 5.5: Supply drop frequency response showing resonance frequencies for Case I and

Case II illustrated in Table 5.4 114

Figure 5.6: A 2mm X 2mm section of the grid with lumped (a) and distributed decoupling capacitors (b) 114

Figure 5.7: Voltage response of a 2mm X 2mm with 2x- and 4x-reduced grids for identical current excitation 115

Figure 5.8: Frequency response at all the M2 supply nodes illustrating locality as function of excitation frequency 117

Figure 5.9: The non-uniform block-based on-die decap distribution in the microprocessor....... 118

Figure 5.10: Transient current response of non-uniform decaps (left) and the magnified plot (right) 118

Figure 5.11: Locality of high-frequency currents provided by decaps 119

Figure 5.12: Distribution of mid-frequency decap currents which are strongly correlated to decap distribution map 120

Figure 5.13: A visualization of the flow of currents on die 120

Figure 6.1: Model of a power delivery network 125

Figure 6.2: Proposed active decoupling capacitance circuit 128

Figure 6.3: Need for external active supply for operational amplifiers 129

Figure 6.4: Comparison of frequency response with active and passive decaps 131

Figure 6.5: Transient current profile (a) and voltage response (b) using active and passive decaps 132

Figure 6.6: Operational amplifier schematic 134

Figure 6.7: Layout of the opamp with drive strength of 5pF 137

Figure 6.8: Gain of the opamp 137

Figure 6.9: Transient response of the opamp 138

Figure 6.10: Comparison of frequency response of power grid with different decap sizes 139

Figure 6.11: Comparison of transient response of power grid with different decap sizes.. 141

Figure 6.12: Worst-case voltage-drop variation with decap allocation 142

Figure 6.13: Impact of C4 distribution on worst-case voltage drop 143

Figure 7.1: A simplified power distribution model and the proposed regulation technique......... 147

Figure 7.2: Simulated unregulated and regulated supply waveforms and safety bounds148

Figure 7.3: Sampling clock generator 150

Figure 7.4: Vdd/Vss level shifter and clocked-comparator schematics.......................... 150

Figure 7.5: Comparator banks and generation of normal, undershoot and overshoot signals.. 151

Figure 7.6: Synthetic load-current generator 152

Figure 7.7: V-I converter-based drop-detector circuit 153

Figure 7.8: Die micrograph and implementation details of the test chip........................ 154

Figure 7.9: Area breakdown in unregulated and regulated test-cases 155

Figure 7.10: Measurement setup... 156

Figure 7.11: Measured unregulated and regulated supply noise waveforms for ramp load and during resonance 157

Figure 7.12: Measured unregulated and regulated supply waveforms as a function of trigger signal.. 158

Figure 7.13: Measured unregulated and regulated peak-to-peak supply noise for varying peak load-currents 159

Figure 7.14: Measured worst supply drop as a function of the active supply voltage.... 160

Figure 7.15: Comparison of measured frequency responses with and without active supply regulation 161

Figure 7.16: Measured worst regulated supply drop as a function of calibration voltages...... 162

Figure 7.17: Statistical analysis across 38 chips showing the affect of a global calibration vs. individual die-tuning 163

Figure 7.18: Measured supply waveforms with on-chip V-I detector and active on-chip probing 164

Figure 7.19: Proposed high-voltage charge-pump-based regulation circuit technique .. 166

Figure 7.20: Die micrograph and implementation details of the test chip 167

Figure 7.21: Measured unregulated and regulated supply noise waveforms for ramp load and during resonance 169

Figure 7.22: Schematic of the proposed high-voltage shunt-supply-based regulation technique.. 170

Figure 7.23: Die micrograph and implementation details of the test chip 171

Figure 7.24: Measured unregulated and regulated supply noise waveforms for ramp load and during resonance 173

Figure 7.25: Schematic of the proposed on-chip oscilloscope for supply-noise measurement.... 174

Figure 7.26: Unregulated (left) and regulated (right) resonance waveforms measured using the on-chip oscilloscope 176

Figure 7.27: Measured supply waveform using the on-chip oscilloscope (left) and the V-I converter (right) 177

LIST OF TABLES

Table 1.1: Summary of the key contributions........ 23

Table 2.1: Experimental results for DC current constraints 56

Table 2.2: Random-run comparison with NLP........ 59

Table 2.3: Experimental results for AC (time-varying) current constraints........ 61

Table 2.4: Run-time reduction due to circuit pruning 63

Table 3.1: Comparison of worst-case voltage drops........ 80

Table 3.2: Mean, standard deviation and 95% interval of the voltage drops........ 82

Table 3.3: Comparison with HSPICE........ 83

Table 3.4: Effect of correlations on accuracy and run-time........ 84

Table 4.1: Experimental results showing delay reduction for a given decap budget 103

Table 4.2: Experimental results showing reduction in decap area for specified timing constraint........ 104

Table 5.1: Run-time and peak memory-usage of the 2mm X 2mm model before and after multi-grid-based sparsification........ 116

CHAPTER I

INTRODUCTION

1.1 Overview and motivation

Power distribution networks deliver the power and the ground voltages from pad locations to all devices in a design. Shrinking device dimensions, faster switching frequency and increasing power consumption in deep submicron technologies cause large switching currents to flow in the power and ground networks. Rapidly switching currents cause spatial and temporal fluctuations in the supply voltage which may cause functional failures in a design, degrade circuit performance and create reliability concerns. A robust power distribution network is essential to ensure reliable operation of circuits on a chip. Power-supply integrity verification is, therefore, a critical concern in high-performance designs.

Due to the resistance of the interconnects constituting the network, there occurs a voltage drop across the network, commonly referred to as the IR drop. IR drop is predominantly caused by the parasitic resistance of metal wires constituting the on-chip power distribution network. The package supplies currents to the pads of the power grid either by means of package leads in wire-bond chips or through C4 (controlled collapsed chip connection) bump-array [26][78] in flip-chip technology. Although the resistance of package is quite small, the inductance of package leads is significant which causes a voltage drop at the pad locations due to time-varying currents drawn by devices on the die. This voltage

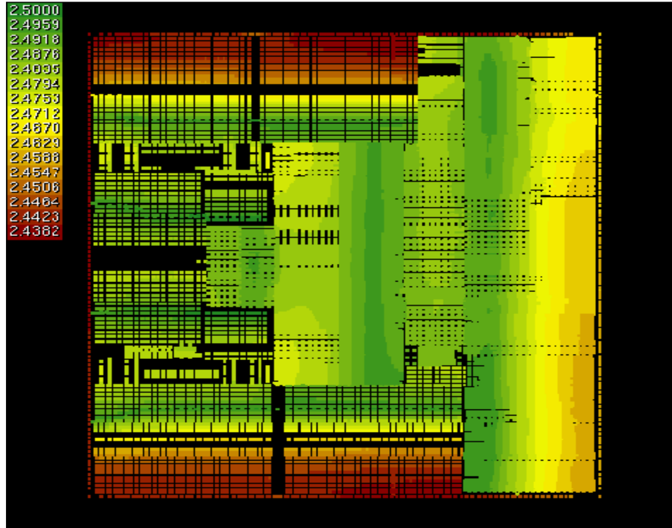


Figure 1.1. On-die IR voltage map of the PowerPC® microprocessor (source: Motorola)

drop is referred to as the d*i*/d*t* drop or Ld*i*/d*t* drop. Therefore, the voltage seen at the devices is the supply voltage minus the IR drop and the Ld*i*/d*t* drop. Figure 1.1 shows the voltage map of a high-performance microprocessor [27] illustrating the spatial variation in supply drop at the on-die devices.

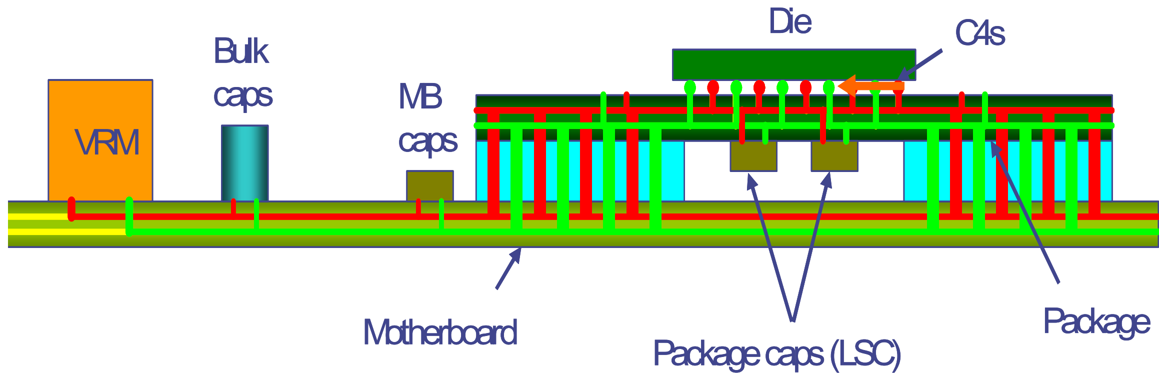
Excessive voltage drops in the power grid reduce switching speeds [16][66][79] and noise margins of circuits, and inject noise which may lead to functional failures. High average-current densities can also lead to undesirable wear-out of metal wires due to electromigration(EM) [7]. Therefore, the challenge in the design of a power distribution network is in achieving excellent voltage regulation at the consumption points notwithstanding the wide fluctuations in power-demand across the chip, and to build such a network using minimum area of the metal layers. These issues are prominent in high-performance chips such as microprocessors, since large amounts of power, in the order of hundreds of Watts, have to be distributed through a hierarchy of metal layers. A robust

power distribution network is vital in meeting performance guarantees and in ensuring the reliable operation of a chip.

Capacitance between power and ground distribution networks, referred to as the decoupling capacitance or *decap*, acts as local charge storage and is helpful in mitigating the voltage drop at supply points. Parasitic capacitance between metal wires of supply lines, device capacitance of the non-switching devices, and capacitance between N-well and substrate, occur as implicit decoupling capacitance in a power distribution network. Unfortunately, this implicit decoupling capacitance is not enough to constrain the voltage drop within safe bounds and designers have to often add intentional explicit decoupling-capacitance structures on the die at strategic locations [15]. These explicitly added decoupling capacitances are not free and increase the area and the leakage-power consumption of the chip. Parasitic interconnect resistance, decoupling capacitance and package/interconnect inductance form a complex RLC network which has its own resonance frequency [47][59]. If this resonance frequency lies close to the operating frequency of the design, large voltage drops can develop in the grid.

Figure 1.2 shows the power distribution network of a typical high-performance microprocessor in detail. The voltage regulator module (VRM), located on the motherboard (MB), provides the nominal supply voltage to the die through wire-bond pads or C4 bumps. Decoupling capacitances are inserted at various levels (on-die, package, motherboard and VRM) in order to suppress the Ldi/dt noise in the supply.

The crux of the problem in designing a power grid lies in the fact that there are many unknowns until the very end of the design cycle. Nevertheless, decisions about the structure, size and layout of the power grid have to be made at very early stages when a large



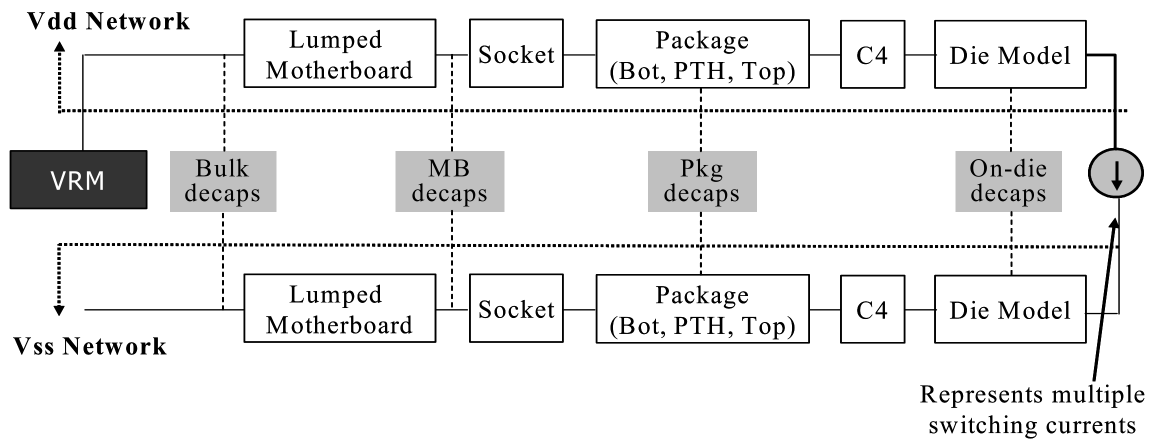


Figure 1.2. Power delivery network of a high-performance microprocessor (source:Intel)

part of the chip design has not even begun. Unfortunately, most commercial tools focus on post-layout verification of the power grid when the entire chip design is complete and detailed information about the parasitics of the power and ground lines and the currents drawn by the transistors are known. Power grid problems revealed at this stage are usually very difficult and/or expensive to fix. Also, due to the growth in power consumption and switching speeds of modern high-performance microprocessors, the di/dt effects are becoming a growing concern in these designs. Clock-gating or power-gating, which is a preferred scheme for power management in high-performance designs, can cause rapid surges in current demands of macro-blocks and increase di/dt effects. Designers rely on the on-chip parasitic capacitances and intentionally added decoupling capacitors to coun-

ter the L*di*/*dt* variations in the supply voltage. It is necessary to accurately model the inductance and capacitance of the package and chip in detail and then analyze the power grid to avoid any underestimation/overestimation in the amount of added decap. Also, it is necessary to maintain the efficiency of the analysis after including these detailed models.

A critical issue in the analysis of power grids is the large size of the network (typically millions of nodes in a state-of-the-art microprocessor). Simulating all the non-linear devices in the chip together with the power grid is computationally infeasible. To make the model-size manageable, the simulation is done in two steps. First, the non-linear devices are simulated assuming ideal supply voltages and the currents drawn by the devices are measured. Next, during power grid simulation, the non-linear devices are replaced by the measured time-varying current sources. Note that the non-linear simulation to generate the time-varying currents is performed assuming a nominal supply and does not account for any drops in the supply voltage. However, since the voltage drops are typically less than 10% of the nominal supply voltage, the error incurred by ignoring the interaction between the device currents and the supply voltage is small.

The circuit-currents are not independent because of signal correlations between different logic blocks. This has traditionally been addressed by deriving the inputs for individual blocks of the chip from the results of logic simulation using a common set of chip-wide input vectors. An important issue in power-grid analysis is to determine what these input vectors should be. For IR-drop analysis, patterns that produce maximum instantaneous currents are required, whereas for the verification of electromigration issues, patterns producing large sustained (average) currents are of interest. For L*di*/*dt*-drop analysis, input-vector search is even more challenging because the worst L*di*/*dt* drop at a particular point

in time is dependant on the previous history of currents. Maximizing Ldi/dt drop thus requires simulation of a long pattern of input vectors spanning multiple clock cycles.

In this chapter, we will present in detail the traditional methods for addressing the above-mentioned power-ground signal integrity issues, and will discuss their strengths and shortcomings. This chapter is organized as follows. Section 1.1 presents some popular power grid modeling and analysis techniques. Section 1.2 provides an overview of the impact of supply noise on circuit performance. Section 1.3 explains some commonly used decap modeling and optimization methodologies while Section 1.4 discusses the transient response and resonance in power grids. In Section 1.5, we present some prior-proposed circuit techniques for measurement and suppression of resonance and Ldi/dt noise in supply networks. The research contributions of our work are summarized in Section 1.6.

1.2 Power-grid modeling and voltage-drop analysis

Modeling of power distribution networks depends on the accuracy and run-time efficiency tradeoff in the required analysis. Although a purely resistive model of the power distribution network may be sufficient for IR drop computation, a more detailed model including the network capacitances, device parasitic capacitances, explicit decoupling capacitors and the inductance of the package is necessary when analyzing the network with time-varying currents. Due to the increasing switching speeds, the Ldi/dt drop is becoming a significant part of the total voltage drop [69]. Therefore, it has become crucial to model the effect of package inductance. Additionally, the impact and the significance of on-chip inductance on the di/dt effects need to be studied, which requires detailed 3-D inductance models [35][65] of the on-die power supply network. Finally, it is imperative

to accurately model the amount of implicit non-switching decaps and explicitly added decaps. An analysis of the inductance effects without accurately modeling the decoupling capacitance will provide pessimistic results since the high fluctuation of currents at the pads will cause very large voltage swings.

Methods for power-grid analysis can be broadly classified into *input-vector-dependent* methods and *vectorless* methods. The input-vector-dependent methods employ search techniques to find a set of input patterns which cause the worst drop in the grid. A number of methods have been proposed in literature [20][40][46] which use genetic algorithms or other search techniques to find vectors or a pattern of vectors that maximize the total current drawn from the supply network. Input-vector-dependent approaches are computationally intensive and are limited to circuit-blocks rather than full-chip analysis. Furthermore, these approaches are inherently optimistic, underestimating the voltage drop and thus allowing some of the supply noise problems go unnoticed. The vectorless approaches [45][43][61], on the other hand, aim to compute an upper bound on the worst-case drop in an efficient manner. These approaches have the advantage of being fast and conservative, but are sometimes too pessimistic and may therefore lead to over-design. Also, these approaches are limited to static power-supply analysis since they can only account for IR drop.

Most of the above discussed approaches model the power supply network as a linear system and rely on a fast and accurate solution of a linear system of equations. The next subsection describes a few linear system solution techniques which are commonly employed in supply-drop analysis.

1.2.1 Linear-system solution techniques

With the increasing number of devices on a chip, the size of power network has grown so large as to make the power-supply integrity verification very challenging. Several methods have been proposed in the literature to reduce the run-time and memory requirements of power grid simulation. Power grid simulation involves solving the following system of differential equations:

$$Gx(t) + C\dot{x}(t) = b(t) \qquad \text{(EQ 1.1)}$$

where G is the conductance matrix; C is the matrix resulting from capacitive and inductive elements; $x(t)$ is the vector of time-varying voltages at the nodes, currents through the inductors and currents through the ideal voltage sources; and $b(t)$ is the vector of the time-varying current sources and the ideal voltage sources.

A method known as Modified Nodal Analysis (MNA) [38][60] transforms the above system of differential equations into the following linear algebraic system which can be solved very efficiently in time-domain:

$$\left(G + \frac{C}{h}\right)x(t) = b(t) + \frac{C}{h}x(t-h) \qquad \text{(EQ 1.2)}$$

MNA uses the Backward Euler (BE) technique with a small fixed time-step, h. The BE reduction with a fixed time-step is advantageous for transient simulation since the left hand side (LHS) matrix $\left(G + \frac{C}{h}\right)$, referred to as the coefficient matrix, does not change during simulation, allowing for preprocessing or factoring of the matrix for a one-time cost and reusing the factors efficiently to solve the system at successive time points.

Several *direct* [60] and *iterative* [33] approaches are available to solve the linear system of equations as in EQ 1.2. Direct techniques rely on factoring the LHS matrix once into a

product of lower and upper triangular matrices (LU factors) and then using them repeatedly in a simple backward and forward substitution procedure [60] to solve the system at every time-step. Iterative methods, on the other hand, rely on efficient convergence techniques to steer the iterations from an initial guess to the final solution.

The size and structure of the conductance matrix of the power grid is important in determining the type of linear-solution technique (direct or iterative) that should be used. Although, the power grid consists of millions of nodes, the conductance matrix is very sparse (typically, fewer than 5 entries per row/column). Sparsity favors the use of iterative methods, but convergence is slowed down by ill-conditioning and can be accelerated to some extent by preconditioning methods [17]. Iterative methods do not suffer from size limitations so long as the (sparse) matrix and some iteration vectors can fit into the memory. Although the conductance matrix itself is sparse, its LU factors are extremely dense. The number of non-zero entries in the LU factors is of the order $O(N^2)$, where N is the number of rows/columns in the coefficient matrix. The single biggest problem with direct methods is the need for large amounts of memory to store the LU factors of the coefficient matrix. However, if only fixed time steps are used for transient analysis, then the initial factorization can be reused with subsequent current vectors, thus amortizing the large decomposition time. Iterative methods do not have this feature of reusability and the linear system needs to be solved iteratively from scratch at every new time step. Iterative methods are best suited for IR-drop analysis which requires simulation of only one time step, or for solving large systems using limited memory resources.

When the vector $x(t)$ in EQ 1.2 consists only of node voltages (power grid network of RC elements and current sources), the coefficient matrix, $\left(G + \frac{C}{h}\right)$ can be shown to be symmetric and positive-definite [32]. The symmetric-positive-definiteness of the coefficient matrix, which is also very sparse, is especially attractive since the system can now be solved very efficiently using specialized linear system solution techniques, such as Cholesky factorization (direct method) and conjugate gradient (iterative method) techniques. The Modified Nodal Analysis circuit formulation is no longer guaranteed to be positive-definite when inductance is included in the power grid model. However, using a simple nodal formulation or mesh current formulation [24], the RLC model of the power distribution network can also be converted into a symmetric-positive-definite system and above techniques can be used effectively. To further speedup the simulation by exploiting the hierarchy in the supply distribution network, a hierarchical macromodeling-based approach was presented in [87], where the power-grid analysis is performed based on a divide-and-conquer approach by splitting the supply network into a global grid and multiple local grids.

Several other approaches have also been proposed to compute the supply drop. The work in [44] proposes a multigrid-like method [10] for the simulation of large power grids. This method, which is particularly attractive for regular meshes, reduces the size of the network by solving several coarser meshes and interpolating the results to the original fine mesh However, this method has been shown to be error-prone while handling *via* resistances and after a certain level of grid-reduction of the original network. The approach in [43] formulates the IR-voltage-drop maximization problem as a linear optimi-

zation formulation with constraints on block-currents while [61] formulates the IR-drop computation as an integer linear programming (ILP) problem to estimate the worst case switching activity based on working modes of macro-blocks. Recently, a statistical approach [62] based on random walks [29] has been proposed which exploits the spatially localized nature of supply drop in power grids to achieve speed-up in simulation. This approach is very efficient for IR-drop computation but the convergence may be slow in transient L*di*/*dt*-drop computation. A part of this dissertation is dedicated to finding efficient vectorless approaches which can account for both IR and L*di*/*dt* drops. The approaches should be useful in the early design phase, when there is limited information about currents of the logic blocks, as well as with detailed simulation models during sign-off.

1.3 Impact of supply noise on circuit performance

The voltage fluctuations in a supply network can inject noise in a circuit, leading to functional failures of the design. Extensive work has therefore been focused on modeling and efficient analysis of the worse-case voltage drop in a supply network as discussed in the previous section. Also, with decreasing supply voltages, the gate delay is becoming increasingly sensitive to supply-voltage variation as the difference between the supply voltage and the threshold voltage is consistently getting reduced. With ever-diminishing clock-cycle times, accurate analysis of the supply voltage impact on circuit performance has therefore also become a critical issue.

Power-supply variation can impact the circuit delay in two ways: First, a reduced supply voltage lessens the gate drive strength, thereby increasing the gate delay. Second, a differ-

ence in the supply voltage between a driver and receiver pair creates an offset in the voltage with which the driver/receiver gates reference the signal transition. This has the effect of creating either a positive or negative time shift in the signal transition perceived at the receiver gate. This dual nature of the impact of supply voltage on circuit delay was observed in [16], and complicates the generation of simulation vectors that maximize the delay along a particular circuit path. Increasing the voltage drop at a particular location may worsen the delay of one gate while improving the delay of another. Therefore, determining the path with worst-delay under these often conflicting goals is a complicated task.

Traditionally, the impact of supply noise on delay has been accounted for by reducing the operating voltage of all library cells by the worst-case supply-voltage drop during library characterization. The library cells are characterized at the worst-case supply variation (which is assumed to be around 10% - 15%) and static timing analysis is performed at this worst supply voltage. However, this approach may be extremely pessimistic since it assumes the occurrence of worst-case voltage drop at all the gates in the design at the same point of time. Also, the traditional approach ignores any supply-voltage shifts between adjacent gates. In this dissertation, we will present approaches for accurate estimation of supply noise induced circuit-delay variation.

1.4 Decoupling-capacitance modeling and allocation

Decoupling capacitance acts as local charge storage and is helpful in attenuating fast transients in supply drop. In this section, we discuss some of the traditional methods for modeling and optimization of on-die decoupling capacitance. The following three sources of capacitances affect the supply voltages in a power grid:

(1) parasitic wire capacitances between power/ground wires, substrate, or signal nets,

(2) parasitic capacitance of transistors, and

(3) explicitly placed decoupling capacitors.

Parasitic wire capacitances can be extracted using either approximate formulae, which use the width and spacing between wires, or by using a commercial parasitic-extraction tool. A difficult issue is the analysis of the wire capacitances coupled to signal nets. The effect of these capacitors on the voltage in the power grid depends on the state of the signal net. For example, coupling from a power network to a signal net that is high, simply couples the power network to itself, with little or no effect on the voltage drop. Unfortunately, due to prohibitively large number of signal nets in a design, it is infeasible to model the signal nets and power grid simultaneously [13]. Therefore, a statistical approach for modeling of the coupling capacitance between the power grids and signal nets has been proposed in [13].

The switching activity of signals is determined by calculating the average number of signal nets that switch in a clock cycle. If the low-to-high and high-to-low switching probabilities of a signal are equal, their effects cancel each other out and therefore, switching nets can be ignored in the analysis. Of the remaining (non-switching) nets, half may be considered to be in stable high state and half in stable low state. Each coupling capacitance is replaced by an effective capacitance to ground, in series with a resistor. The effective value of the coupling capacitor [68] is

$$C_{eff} = \frac{1}{2}C_{coupling}(1 - P_{active}) \qquad \text{(EQ 1.3)}$$

where P_{active} is the average switching activity.

Device capacitances attenuate the voltage drop in the power grid and have a larger effect on the voltage in the power grid as they are much larger than the wire capacitances. The effect of device capacitances also depends on the state of the signal. This is explained with the help of an inverter chain in Figure 1.3, where we compute the implicit decoupling capacitance of inverters INV2 and INV3 based on their input and output states. Each transistor has 5 device capacitances, C_{sb}(*source to bulk*), C_{db}(*drain to bulk*), C_{gs}(*gate to source*), C_{gs}(*gate to drain*) and C_{gb}(*gate to bulk*). The C_{sb} can be ignored since the source and bulk for both the pmos and nmos are always at the same potential[1]. Figure 1.3(a) shows the remaining 4 device capacitances for an inverter. In Figure 1.3(b), the capacitances are arranged across three inverters to make the analysis more convenient. Capacitances of switching devices contribute to the current drawn from the grid, which is already modeled by the time varying current source in the power-grid analysis. Therefore, in power-grid analysis, we need to consider the device capacitances of only those gates that do not switch.

We first look at the case where net N in Figure 1.3(a) is in a low state. The device capacitances shown in Figure 1.3(a) can be modeled by the equivalent RC circuit shown in Figure 1.3(b). The resistance R_p corresponds to the effective pull-up resistance of inverter 1, the resistance R_n corresponds to the effective pull-down resistance of inverter 2, and the resistance R_{wp} and R_{wn} correspond to the P and N well resistances, respectively. Since net N is low, capacitances C_{dbn}, C_{gsn}, and C_{gbn} are discharged and do not contribute to the

1. This assumption is not valid in gates with series stacks of transistors.

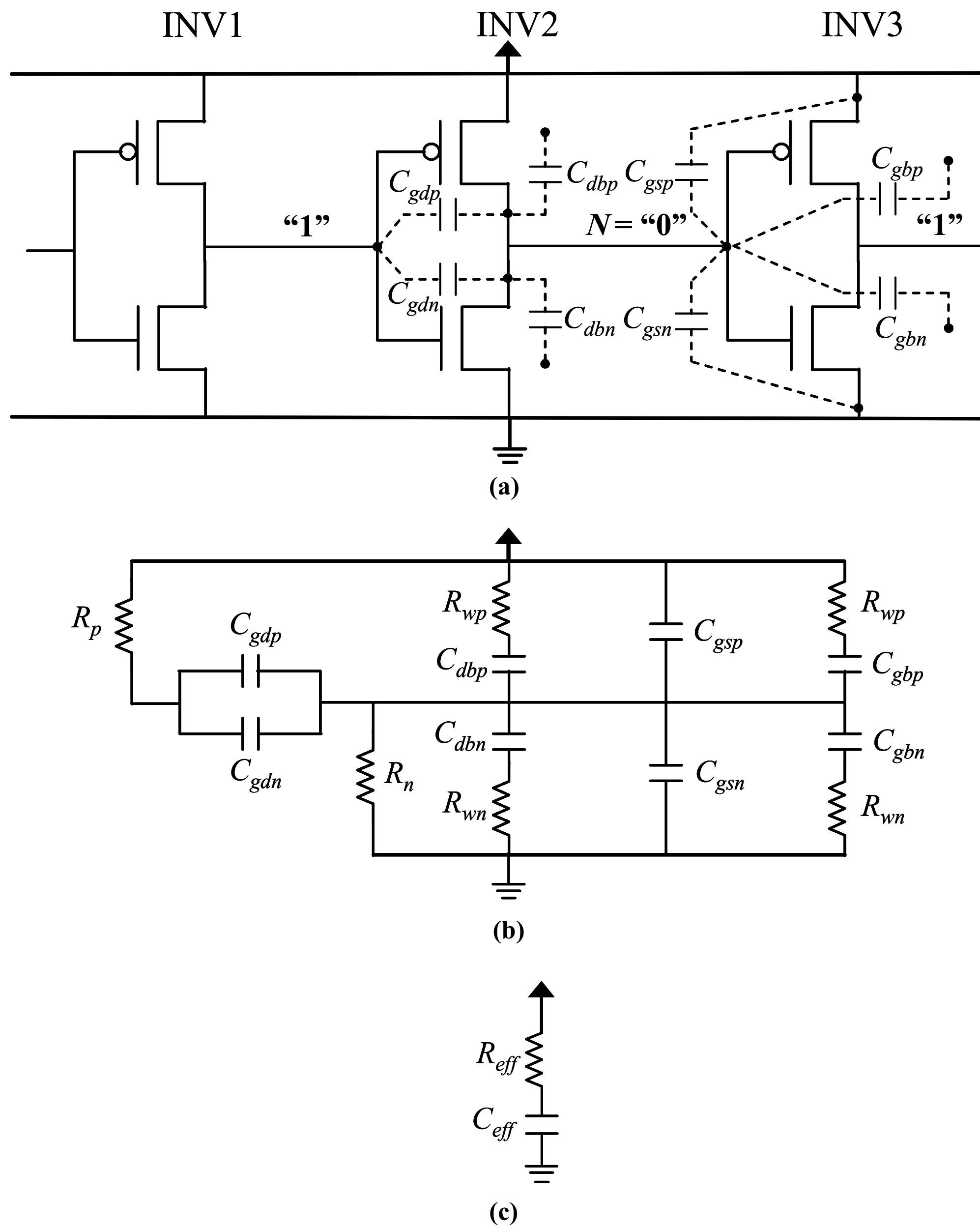


Figure 1.3. Device-capacitance modeling in power-grid analysis [68]

decoupling between the power and ground grid. Furthermore, since R_{wp} is a relatively high resistance and since C_{dbp} and C_{gbp} are small, they can be ignored without a significant loss in the accuracy. An analogous analysis can be made for the decoupling capacitances when the state of signal N is high. This state is assumed to have equal probability of

being high or low when the gate is not switching, although a different ratio of high-to-low signal states could easily be incorporated in the analysis. An approximate model of the device capacitances for power-grid analysis is shown in Figure 1.3(c). The effective decoupling capacitance in this simplified model is the sum of the effective high and low decoupling capacitances weighted by the probability of the gate being in either state:

$$C_{eff} = (1 - P_{active})(C_{gdp} + C_{gdn}) + \frac{1}{2}((1 - P_{active})(C_{gsp} + C_{gsn})) \qquad \text{(EQ 1.4)}$$

Similarly, the effective resistance is approximated by the sum of the high and low resistance of the gates:

$$R_{eff} = R_p + R_n \qquad \text{(EQ 1.5)}$$

The intrinsic N-well capacitance is also modeled as a series RC whose time constant and capacitance per unit well area are characterized using a process simulator. The intrinsic as well as the explicit decoupling capacitances are distributed either according to the layout or, when a layout is not available (as during the early design stage), uniformly across the power rails.

The intrinsic decoupling capacitance is usually not enough to confine the voltage drop within safe bounds and designers have to add specific (explicit) decoupling-capacitance structures on the chip. Several methods [72][73][88] have been proposed for optimal explicit decoupling-capacitance allocation. These methods formulate the decoupling-capacitance allocation as an optimization problem with the objective of decoupling-capacitance area minimization and constraints on the worst voltage drop. However, in high-performance designs, circuit performance is a more pressing concern and the above approaches, although optimal for supply-noise reduction, may not be optimal for best circuit performance. For instance, in a logic block, only the delay of gates on the critical and

near-critical paths are of concern and the gates having larger timing slacks can afford relatively higher voltage drop. In this dissertation, we will present an approach for decoupling-capacitance minimization with the objective of optimizing the circuit performance.

1.5 Ringing in power supply networks

Intrinsic and extrinsic on-die decoupling capacitances interact with package inductance and may cause power supply resonance, where even small changes in the load-current can cause excessive supply-voltage fluctuations in the power distribution network. Also, the decoupling capacitances present at the other levels of the power distribution network (Figure 1.2) interact with parasitic inductances to form a multiple LC system with multiple resonance frequencies. In this section, we explain the transient response and various resonance frequencies of a power supply network. Figure 1.4 shows a simplified model with parasitics and decoupling capacitances at various levels of the power supply network. In this simplified model, all the on-chip decoupling capacitances have been lumped into a single capacitor C_{die} and all the switching devices have been modeled by a single time-varying current, I_{die}. The L*di*/d*t* drop occurring in the supply network can be classified as zeroth droop, first droop, second droop and third droop. The third droop occurs at the node between the voltage regulator module (VRM) and the motherboard and at the node between the motherboard and the microprocessor socket. The third droop is controlled by the motherboard inductance (L_{MB}) and motherboard capacitors (C_{MB}). The second droop, which takes place at the node between the microprocessor's socket and the microprocessor's package, is controlled by the partial package inductance leading from the motherboard upto the package capacitors (L_{pkg1}) and package capacitors (C_{pkg}). The first droop

occurs at the node between the microprocessor's package and the chip itself. The first droop is controlled by partial package inductance leading from the package capacitors to I/O pads (L_{pkg2}) and on-chip decoupling capacitance (C_{die}). L_{pkg2} also consists of the inductance of wire-bonds or C4 bumps. Lastly, a zeroth droop can occur on the die if there is significant amount of on-die inductance in the power grid. Figure 1.5 shows the first and the second droop in the transient response for a step load-current and the ac frequency response of the supply voltage. The third drop is very small in magnitude as well as frequency than the first and the second droops. The expression for the first droop (supply drop and ground bounce combined) due to a step load-current of peak amplitude I_{max} is given as follows:

$$\Delta V_{die}(t) \cong 2I_{max}R + I_{max}\sqrt{\frac{2L_{pkg2}}{C_{die}}} \cdot e^{-\frac{R}{2L_{pkg2}}t} Sin(\omega_r - \theta) \qquad \text{(EQ 1.6)}$$

where, $R = R_{pkg} + R_{MB} + R_{die}$; ω_r is the resonance frequency of the first droop given by:

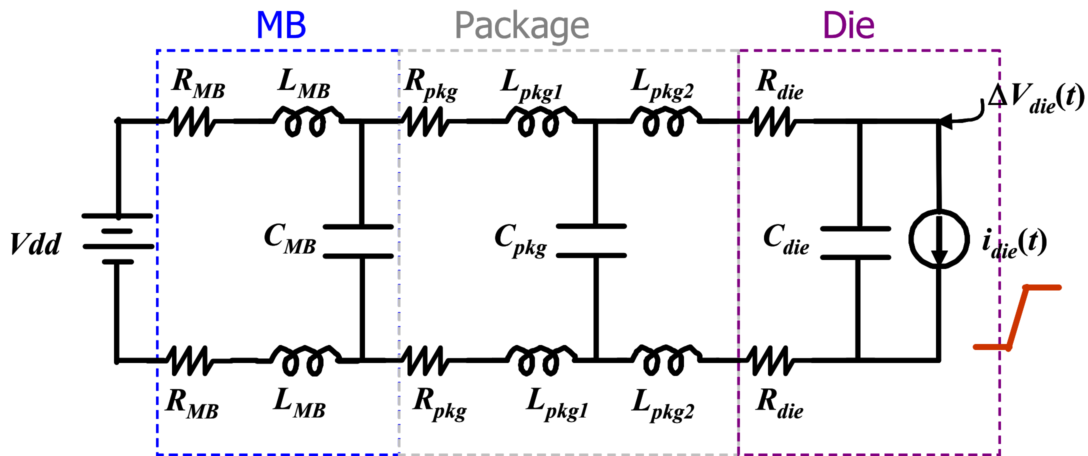


Figure 1.4. Simplified model of a power supply network

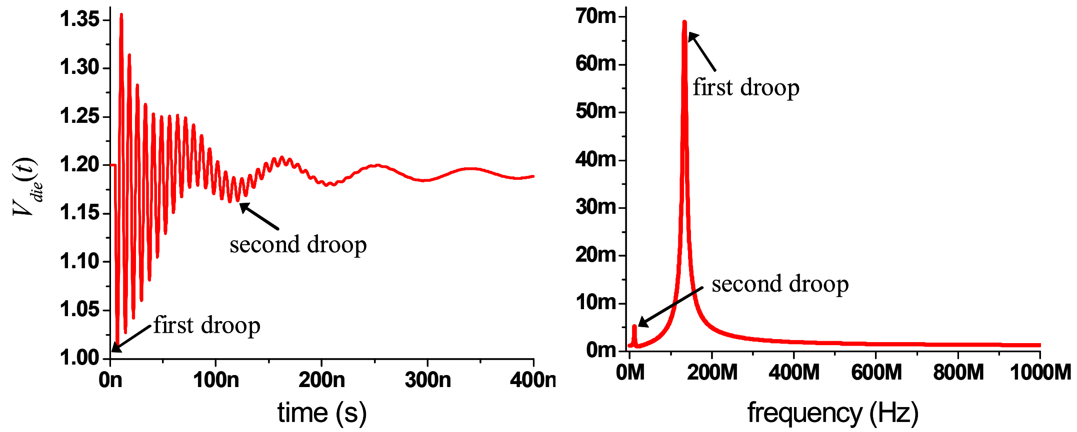


Figure 1.5. Transient and frequency response showing first and second droops

$$\omega_r = \frac{1}{\sqrt{L_{pkg2} C_{die}}} \quad \text{(EQ 1.7)}$$

and θ is the phase difference given by,

$$\tan\theta = \frac{\sqrt{R^2 C_{die}^2 - 4L_{pkg2} C_{die}}}{RC_{die} - 2L_{pkg2}/R} \quad \text{(EQ 1.8)}$$

Any current excitation which is centered at the resonance frequency, ω_r, for a sufficient amount of time can create excessive voltage swings in the power grid as shown in Figure

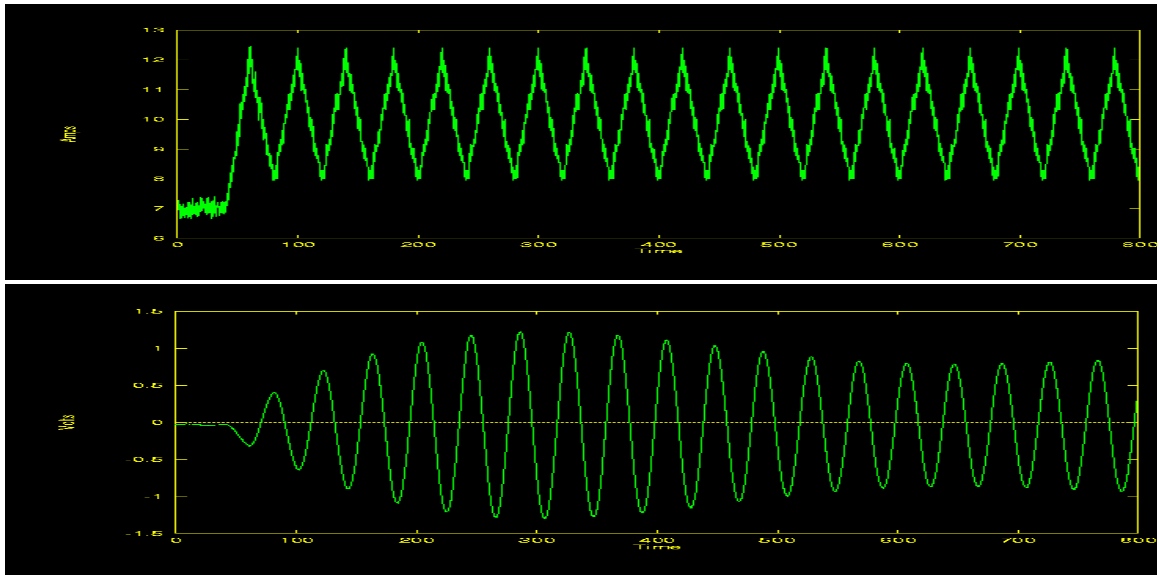


Figure 1.6. Current excitation at resonance frequency (top) and voltage response (bottom) (source: Motorola)

1.6. Due to the presence of decoupling capacitance at multiple locations (die, package and motherboard), there are multiple resonance frequencies in the supply network: the L_{MB}-C_{MB} combination causes a resonance at a few hundred kHz; the L_{pkg1}-C_{pkg} combination causes a resonance at a few MHz; and the L_{pkg2}-C_{die}causes a resonance of the order of a few hundred MHz. The goal of a power grid designer is to minimize package inductance (L_{pkg2}), allocate on-die decaps (C_{die}) and sufficient C4s in the pads, so as to minimize the first droop and to create the largest possible gap between the highest resonance frequency and the operating frequency of the design. Care should also be taken to ensure that the operating frequency does not directly coincide with the higher harmonics of the resonance frequencies [11]. The advent of C4 flip-chip technology, which have much smaller inductance than that of bond-wires, and advances in package technology have led to a significant decrease in package inductance. The impact of package inductance on the first droop can be further reduced by placing the package decaps (C_{pkg}) as close to the die as possible. Intelligent allocation and placement of explicit on-die decaps is also helpful in reducing the first droop as shown in [47][59][73][88]. Explicit decaps are, therefore, often added near regions with severe supply drops.

However, these explicit decaps result in area overhead and hence directly increase the cost of the chip. In addition, explicit decaps increase the leakage power consumption of the chip due to their gate-leakage current. With technology scaling, gate leakage has become a significant percentage of the overall leakage which places a significant limitation on the maximum amount of decap that can be introduced [9]. Several circuit techniques have, therefore, been proposed to actively regulate the on-die supply voltage. The next section provides an overview of some of the previously proposed active circuits.

1.6 Active supply-voltage regulation and supply-noise measurement

Active supply voltage regulation techniques employ circuits to enhance the amount of charge transfer to-and-from the power supply network during a supply-voltage fluctuation. The objective of these approaches is to reduce the supply drop for the same amount of explicit decoupling capacitance or to minimize decoupling-capacitance area for the same worst-case supply drop.

Active guard-ring circuits, based on the use of active decaps, were proposed in [51] and [82] to suppress substrate noise. In [77], a Miller-coupling based capacitance enhancement technique was proposed to reduce crosstalk between digital and analog regions on a die. A switched-capacitor based circuit technique with two decap banks switched between series and parallel configurations was proposed in [4]. The same switched capacitor was used for suppressing resonance in [36]. In [84], a band-pass filter is used to detect supply noise resonance and an artificial shunt load, connected between Vdd and Vss, is periodically switched on and off with 180° phase shift to dampen the resonance. In [56], a shunt high voltage supply is connected to the regular power grid when power-gated logic blocks wake up from the sleep state. A linear regulator with a specific goal to source or sink large transients of current was proposed in [83]. Recently, several adaptive frequency-management techniques [30][76] have also been proposed to compensate for supply transients. These techniques employ supply-drop monitors at various locations on the die and the frequency of operation is altered to compensate for fluctuations in the power supply. In [37], the use of controlled incremental frequency changes to alleviate inductive noise in dynamically voltage scaled microprocessors was explored. In addition to these, several micro-

architectural control techniques such as selective issue, pipeline throttling and selective wake-up of clock-gated modules have been proposed in [34] and [41]. A major portion of this dissertation is devoted to exploring power efficient active circuit techniques for suppression of Ldi/dt drop and supply resonance.

Several circuits have been proposed to monitor the on-die supply noise. Circuits proposed in [31][42][54][55][74] constitute a sample-hold circuit to sample the supply voltage and a V-I converter circuit. The V-I converter circuit consists of a high-conductance transistor to convert the supply voltage samples into current. The current is then amplified using a current mirror and is transmitted out of the chip using a transmission line. Current-based sensing is particularly attractive due to its robustness to coupling noise. However, the use of an analog V-I converter with high gain, followed by current amplifiers makes this technique power inefficient. An A-D converter has been proposed to convert the analog samples of supply noise into a digital code [89] which is then transmitted off-chip. This approach requires a lot of area, making it less effective for fine-grained supply noise measurement. [53] presents an analog circuit that reports whether power supply or ground voltage at the location of comparators within a microprocessor core crosses a pre-defined threshold voltage in every clock cycle. [2] and [39] use repetitive time-shifted sampling of on-chip supply noise. In this dissertation, we propose a low-power, all-digital on-chip oscilloscope for measuring supply noise and compare its performance with that of the V-I converter based supply drop monitor. The proposed on-chip oscilloscope can also be used for measuring coupling noise in critical signals and clock jitter.

1.7 Contributions of this work and organization

This dissertation focuses on developing algorithms and circuit techniques to analyze and suppress the supply noise related signal integrity issues. While we do not purport to present complete solutions to these problems, we aim to provide state-of-the-art CAD tools and active circuit techniques to estimate and combat supply noise in modern high-performance VLSI systems. Table 1.1 summarizes the key problems, the prior-proposed solutions and the solutions proposed in this dissertation. The first half of this dissertation presents CAD approaches for modeling, analysis and optimization of power supply net-

	Problem Statement	Prior Works	Proposed Solutions	Chapter
1	Analysis of supply noise induced circuit delay variation	[1][16][66]	Path-based and block-based approaches for accurate analysis of delay variation under IR and L*di*/d*t* drops	Chapter II
2	Early-mode power-grid analysis	[19][43][44][61][87]	A constraint-based and a statistical approach for power-grid analysis in early design phase	Chapter III
3	Timing-aware passive decoupling-capacitance allocation	[12][48][73][88][86]	An exact optimization approach and a heuristic-based approach for timing-aware decap allocation	Chapter IV
4	Detailed study of the impact of inductance, resonance and locality in power distribution network of an industrial microprocessor	[15][27][52]	A detailed full-die dynamic model of a 90nm Intel Pentium®-class microprocessor design	Chapter V
5	Active circuit techniques for the suppression of inductive-supply-noise	[4][36][37][51][56][76][77][82][83]	Analog and fully-digital circuit techniques for inductive-supply-noise suppression	Chapters VI,VII
6	On-die supply-noise measurement	[2][31][39][42][53][54][55][74][89]	A power-efficient, fully-digital on-chip oscilloscope for on-die probing of high-speed signals	Chapter VII

Table 1.1. Summary of the key contributions

works. In the later half of the dissertation, we present circuit techniques for measurement and suppression of supply noise. The dissertation is organized as follows:

In Chapter II, we propose a path-based and a block-based analysis approach for computing the maximum circuit delay under power-supply fluctuations. The analyses are based on the use of superposition, both temporally and spatially across different circuit blocks. The approaches are vectorless and take both IR drop as well as Ldi/dt drop into account. The path-based approach computes the maximum possible delay of a given critical path in the presence of supply variations, while the block-based approach does not require apriori knowledge of the critical paths in a circuit and can be, therefore, effectively incorporated into an existing static timing analysis framework. The delay maximization problem is formulated as a non-linear optimization problem with constraints on currents of macros or circuit-blocks in the design. We show how correlations between currents of different circuit-blocks can be incorporated in the formulations using linear constraints. The proposed methods were validated on ISCAS85 benchmark circuits and an industrial power-supply grid, and demonstrate accurate worst-case circuit-delay computation.

In Chapter III, we propose two new approaches for analyzing the power-supply drop. The first approach conservatively computes the worst-case supply drop, early in the design flow when detailed information of the design is not available. The second approach computes the statistical parameters of supply-voltage fluctuations with variability in block currents. The proposed statistical analysis can be used to determine the portions of the grid that are most likely to fail. The analyses consider both IR drop and Ldi/dt drop in a power supply network and can take into account spatial and temporal correlations among block-currents. We show that the run-time is linear with the length of the current waveforms

allowing for extensive vectors, up to millions of cycles, to be analyzed. We implemented the approaches on a number of grids, including a grid from an industrial microprocessor to demonstrate their accuracy and efficiency.

In Chapter IV, we propose an approach for timing-aware decoupling-capacitance allocation which uses timing slacks to drive the optimization. Non-critical gates with larger timing slacks can tolerate a relatively higher supply-voltage drop as compared to the gates on the critical paths. The decoupling-capacitance allocation is formulated as a non-linear optimization problem using Lagrangian relaxation and a modified adjoint sensitivity method is used to obtain the sensitivities of objective function to decap sizes. A fast path-based heuristic is also implemented and compared with the global optimization formulation. The approaches have been implemented and tested on ISCAS85 benchmark circuits and grids of different sizes. Compared to uniformly allocated decaps, the proposed approach utilizes 35.5% less total decap to meet the same delay target. For the same total decap budget, the proposed approach is shown to improve the circuit delay by 10.1% on an average.

In Chapter V, we describe the first detailed full-die dynamic model of a 90nm Intel Pentium®-class microprocessor design, including package and non-uniform decap distribution. This model is justified from the ground up using a full-wave model and then increasingly larger but less detailed models with only the irrelevant elements removed. Using these models, we show that there is insignificant impact of on-die inductance in such a design, and that the package is critical to understanding the resonant properties of the grid. We also show that transient effects are sensitive to non-uniform decap distribu-

tion and that locality is a function of the excitation frequency and of the package-die resonance frequency.

In Chapter VI, we present an analog active decap circuit that significantly increases the effectiveness of decap in suppressing power supply fluctuations. The proposed circuit senses the supply drop and drives an amplified and inverted voltage fluctuation on the decap. The active decoupling circuit is powered by a separate power supply and we study the optimal allocation of the total C4s/pads between this second power supply and the regular supply, as well as the optimal allocation of the total decoupling capacitance between actively switched and traditional static decap. Finally, we demonstrate that the overhead of the proposed method is small compared to the area of the decaps. Simulations in a 0.13μm CMOS process demonstrate that the maximum supply drop is reduced by 45% compared to the use of only traditional decap, corresponding to an increase in the effective decap of approximately 8X.

In Chapter VII, we present three digital circuit techniques for inductive supply-noise suppression. The presented techniques effectively suppress supply noise caused by rapid current transients or due to resonance. The charge-injection-based active decoupling technique uses a nominal active supply and an active decap bank to inject extra charge into the power grid in case of an undershoot. This technique provides as effective decap of 10.5X (for a 10% supply regulation tolerance), does not require any high-voltage supplies and obviates the need for thick oxide (thick-ox) devices. Furthermore, the active decap acts as passive when the supply voltage is within the pre-specified safety bounds. The high-voltage charge pump based active circuit uses a high-voltage charge pump to dump extra charge into the power grid during excessive undershoots. The high-voltage shunt supply

based active circuit connects the regular nominal-supply power grid directly to an external high-voltage supply whenever an undershoot is detected, thus damping the transient response of the supply network. We also present a fully-digital on-chip oscilloscope which is more power efficient than a conventional supply drop monitor. All of the proposed circuits were implemented in a test-chip, fabricated in a 0.13μm, triple-well CMOS process. Measurement results demonstrate that the three active supply-noise suppression techniques suppress the inductive supply fluctuation by 57%, 33% and 43%, respectively for a step load-current. During resonance, the supply fluctuation is suppressed by 75%, 43% and 45%, respectively by the three proposed circuit techniques. The performance of the proposed active circuit techniques was validated with the V-I converter-based drop detector circuit, the proposed on-chip oscilloscope and by direct on-chip probing of Vdd and Vss metal lines.

In Chapter VIII, we summarize our contributions and discuss the future work.

CHAPTER II

VECTORLESS ANALYSIS OF SUPPLY-NOISE INDUCED DELAY VARIATION

2.1 Introduction

Power supply networks are essential in providing the devices on a die with a reliable and constant operating voltage. Due to the resistance and inductance of the on-chip and package supply networks, the supply voltage delivered to various devices on a die is non-ideal and exhibits both spatial and temporal fluctuations. In today's high-end designs, it is not uncommon for the supply network to conduct hundreds of Amperes of total current [75]. As semiconductor technology is scaled down and the supply voltage is reduced, the total current that must be supplied by the power network is expected to increase even further, making it more difficult to meet stringent supply-integrity constraints. In particular, the Ldi/dt voltage drop is expected to become more prominent as it increases with both increasing current demand and higher clock frequency [69]. Furthermore, IR drop and Ldi/dt drop interact in a non-trivial manner and total maximum drop is not always the sum of the two individual maximum voltage drops. With decreasing supply voltages, the gate delay is becoming increasingly sensitive to supply-voltage variation as the difference between the supply voltage and the threshold voltage is consistently reduced. With ever

diminishing clock-cycle times, accurate analysis of the impact of supply voltage on circuit performance has, therefore, become a critical issue.

Power supply variation can impact the circuit delay in two ways: First, a reduced supply voltage lessens the gate drive strength, thereby increasing the gate delay. Second, a difference in the supply voltage between a driver/receiver pair creates an offset in the voltage with which the driver/receiver gates reference the signal transition. This has the effect of creating either a positive or negative time shift in perceived signal transition at the receiver gate. This dual nature of the impact of supply voltage on circuit delay was observed in [16], and complicates the generation of simulation vectors that maximize the delay along a particular circuit path. Increasing the voltage drop at a particular location may worsen the delay of one gate while improving the delay of another. Therefore, a vector must be determined that results in an optimal combination of these often conflicting goals of introducing both reduced drive strengths and supply voltage shifts such that overall circuit delay is maximized.

Traditionally, the impact of supply noise on delay has been accounted for by reducing the operating voltage of all library cells by the worst-case supply-voltage drop during library characterization. This assumes that the expected worst-case voltage drop occurs at all places in the design. This yields a very conservative analysis since, in practice, the worst drop is localized to a small region at any one point in time. On the other hand, this approach ignores the impact of voltage shifts between driver/receiver pairs, thereby possibly underestimating the worst-case delay in certain situations.

In this chapter, we present two approaches for analyzing the impact of supply variations on circuit delay. The proposed approaches are vector-less, allowing for efficient analysis,

and address both IR drop and Ldi/dt drop effects. The first approach is a path-based approach which computes the maximum possible delay of a path in the circuit in presence of supply variation. The second approach is a block-based method which does not require apriori knowledge and enumeration of the critical paths in a circuit. The proposed approaches use a quadratic gate-delay model. We formulate the task of determining the worst-case impact of supply noise on delay using a constrained non-linear optimization problem where the currents of the different logic-blocks are the optimization variables. Linear constraints on block-currents are then formulated for the total power consumption of the chip and for individual block-currents. Constraints between currents of different blocks or a single block in consecutive clock cycles can be incorporated to account for any spatial or temporal correlations that exist between circuit blocks. The proposed approaches have the advantage that accurate constraints can be extracted from extensive gate level simulation data that is readily available during the design process, thereby significantly improving the accuracy of the analysis while avoiding the need for lengthy and time-consuming power grid simulation. The proposed methods were validated on ISCAS85 benchmark circuits, including a power grid from an industrial processor design, and demonstrate a significant reduction in pessimism during worst-case circuit-delay computation.

The remainder of this chapter is arranged as follows. Section 2.2 describes the gate-delay model for delay variations with respect to supply-voltage fluctuations. Section 2.3 presents the voltage-drop sensitivity computation while Section 2.4 presents the constraints on block-currents and a method to incorporate current correlations. Section 2.5 and Section 2.6 present the path-based and the block-based formulations respectively for

maximizing the impact of supply-voltage fluctuations on delay. Section 2.7 presents the experimental validation of the proposed approaches and conclusions are presented in Section 2.8.

2.2 Delay model for supply fluctuations

In this section, we model the impact of supply-voltage variations on delay. Since the supply-voltage variations in a power grid are typically very slow compared to the transition time of a switching gate [11], we can make the simplifying assumption that the supply voltages are constant during a switching transition. From the perspective of the circuit delay, we are therefore concerned with the impact of fixed voltage offsets from the nominal *Vdd* and *Vss* voltages on the delay of a circuit. Note however that *dynamic* IR drop and Ldi/dt drop effects will be the cause of these voltage offsets.

A voltage drop at a power-supply point can impact the delay of a gate through one of the following two mechanisms:

1. A decrease in the *Vdd* voltage or an increase in the *Vss* voltage at the gate under consideration decreases the *locally* observed supply voltage of the gate and will reduce its drive strength and hence increase its delay. The worst-case voltage drop is typically localized to a small region in the chip. Hence, only a few gates in a path will typically be operated with a worst-case drive strength. Gates with higher local supply voltage therefore compensate for the increased delay of gates with reduced local supply voltage in the path, and a global analysis of the impact of supply voltage on the path delay is therefore required.

2. A relative shift in the *Vdd* or *Vss* voltages between the driver and receiver gates of a signal net can introduce a voltage offset that will impact the delay of a gate. This is illustrated in Figure 2.1 where the *Vss* voltage of the receiver gate is increased relative to the *Vss* voltage of the driver gate. Since the input signal has a rising transition, the NMOS transistor of the receiver gate senses the input voltage relative to the local *Vss* voltage level. The shown voltage shift therefore results in an effective (negative) noise voltage at the receiver gate input that increases the delay of the receiver gate. Note that a shift in the supply voltage impacts the rising and falling transitions of a gate in opposite ways, meaning that an increase in the *Vss* voltage from driver to receiver results in an increased delay for a rising input transition while an increase in the *Vdd* voltage improves the delay for a falling input transition.

The relative shift in supply voltage between the driver and receiver gates is likely to be larger if the gates are farther apart. Therefore, nets that transmit signals across the chip will have a higher likelihood of shifts in supply voltage between their driver and receiver pair and hence are more susceptible to power grid noise. The relative magnitude of the

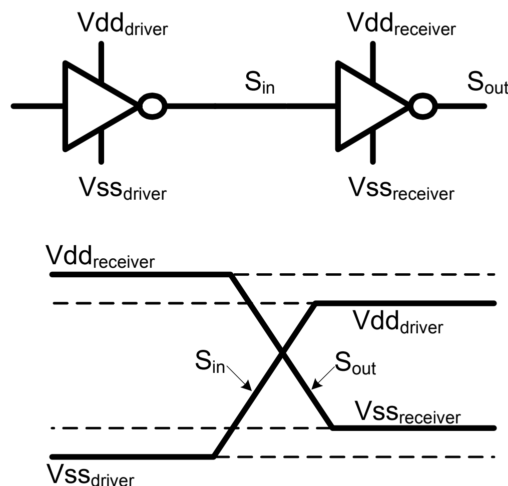


Figure 2.1. A driver-receiver pair in a non-ideal supply network

above two mechanisms depends on the input slope and output loading of a gate. The sensitivity of gate delay to driver-strength reduction will increase with output loading, while the sensitivity to voltage shifts will increase with slower input signal-transition times.

In order to maximize the delay of a path, it is necessary to induce voltage drops in the supply network such that the delay of each gate is increased through both mechanisms: reduction of driver strength and voltage shifts between successive gates in the path. A possible voltage assignment that maximizes the voltage shift between consecutive gates in a circuit path is shown in Figure 2.2. However, this assignment does not reduce the drive strength of each gate by the maximum possible amount. Furthermore, this assignment may not be feasible for grids where Vdd-drop and Vss-bounce are tightly correlated with each other. Maximizing the delay through reduced drive strength and through voltage shifts therefore, requires conflicting voltage assignments that cannot be realized simultaneously.

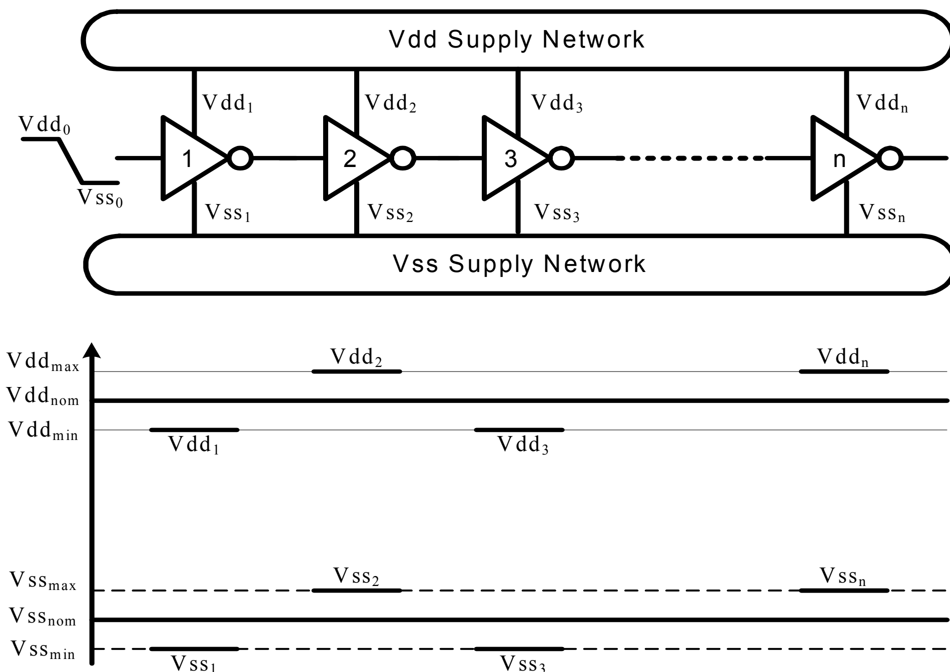


Figure 2.2. A path in a power supply network with worst-case voltage shifts causing the maximum path delay

A worst-case *realizable* voltage assignment that maximizes the overall path delay will depend on the specific conditions of the gates and their sensitivities to the different voltage-drop phenomena.

We now present our model for the dependence of the delay of a single gate on the voltage drops at that gate and at its preceding gate. We consider the delay of a gate *G*, shown in Figure 2.3(a), with local supply voltages Vdd_g and Vss_g and supply voltages Vdd_{in}, Vss_{in} at the preceding driver gate. As shown in Figure 2.3(b), the propagation delay τ between the input and output transitions of a gate is measured at 1/2 the *nominal* supply voltage point to ensure a common reference between successive gates. The delay of the receiver gate depends on the Vdd_g and Vss_g voltages at the receiver gate itself, the voltages Vdd_{in}, Vss_{in} at the preceding driver gate, the input transition time and the output load. To provide a

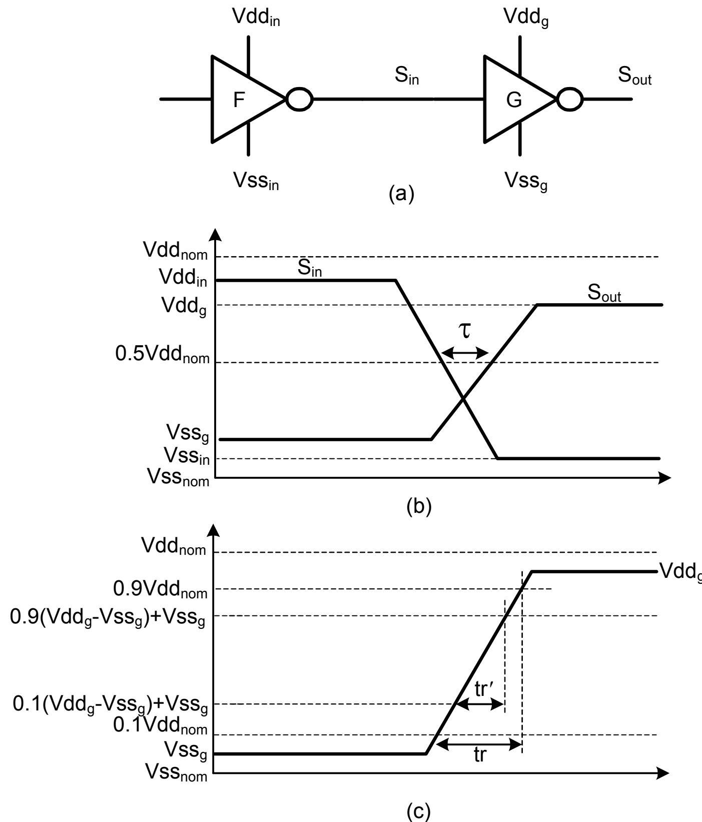


Figure 2.3. A driver-receiver pair in a non-ideal supply network (a), propagation delay (b) and output transition time (c)

common reference for transition time, we again define the transition time *tr* of a signal as the time between the 10% to 90% crossing of *nominal* supply voltage for an equivalent full-swing transition, as shown in Figure 2.3(c). Given the signal transition at the output of gate *G*, and given the *local* transition time tr', measured between 10% to 90% of the local supplies $V_{ss,g}$ to $V_{dd,g}$, the equivalent full-swing transition time *tr* is computed as follows:

$$tr = tr' \cdot \frac{Vdd_{nominal}}{Vdd_g - Vss_g} \qquad \text{(EQ 2.1)}$$

For a given input transition time and output slope, the delay and transition time at the output of gate *G* are expressed as follows:

$$\tau = f(Vdd_g, Vss_g, Vdd_{in}, Vss_{in}) \qquad \text{(EQ 2.2)}$$

$$tr_{out} = g(Vdd_g, Vss_g, Vdd_{in}, Vss_{in}) \qquad \text{(EQ 2.3)}$$

where, τ and *tr* are the propagation delay and output transition time of gate *G*.

In general, *f* and *g* are non-linear functions of their variables. However, the voltage drop in a power grid network is restricted and is typically within the range of $\pm 10\%$ of the *nominal* supply voltage. It was observed in [1] that within this range, the delay of a gate is close to a second degree polynomial. Figure 2.4 shows the rise and fall delays of an inverter in 0.13μ technology as Vdd_g, Vss_g, Vdd_{in} and Vss_{in} vary by $\pm 20\%$. The delay curves in Figure 2.4 show that functions, *f* and *g* can be accurately modeled as second degree polynomials for reasonable supply voltage variations. We therefore express the delay, τ, and output transition time, *tr* of a gate as follows:

$$\tau = \tau_0 + \sum_{\Omega} k_i \Delta Vdd_g^{\alpha_i} \Delta Vss_g^{\beta_i} \Delta Vdd_{in}^{\delta_i} \Delta Vss_{in}^{\gamma_i} \qquad \text{(EQ 2.4)}$$

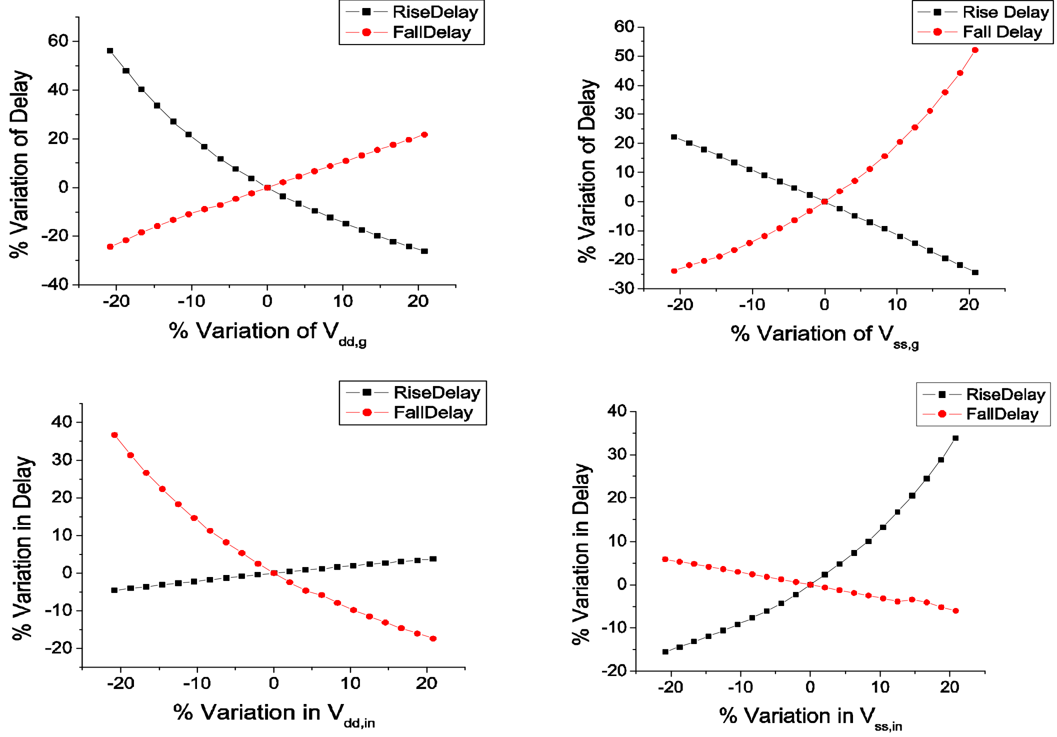


Figure 2.4. Variation of rise/fall propagation delays of a gate with respect to Vdd_g, Vss_g, Vdd_{in} and Vss_{in}

$$tr = tr_o + \sum_{\Omega} l_i \Delta Vdd_g^{\alpha_i} \Delta Vss_g^{\beta_i} \Delta Vdd_{in}^{\delta_i} \Delta Vss_{in}^{\gamma_i} \qquad \text{(EQ 2.5)}$$

$$\alpha_i + \beta_i + \delta_i + \gamma_i \leq 2,\ \alpha_i, \beta_i, \delta_i, \gamma_i \in \{0, 1, 2\},\ k_i, l_i \in \Re$$

where ΔVdd_g, ΔVss_g, ΔVdd_{in}, and ΔVss_{in} are the deviations of the four supply voltages from their nominal values, τ_0 is the propagation delay with nominal supply, tr_0 is the output transition time with nominal supply and Ω is the sample space of all possible values of α_i, β_i, δ_i and γ_i..

Traditional standard-cell libraries are composed of two dimensional tables with table entries representing delay and transition times for different load and transition-time combinations. We modify the library to incorporate the coefficients k_i and l_i in place of a fixed delay entry. For a given set of transition time, load and supply voltages at a gate and its

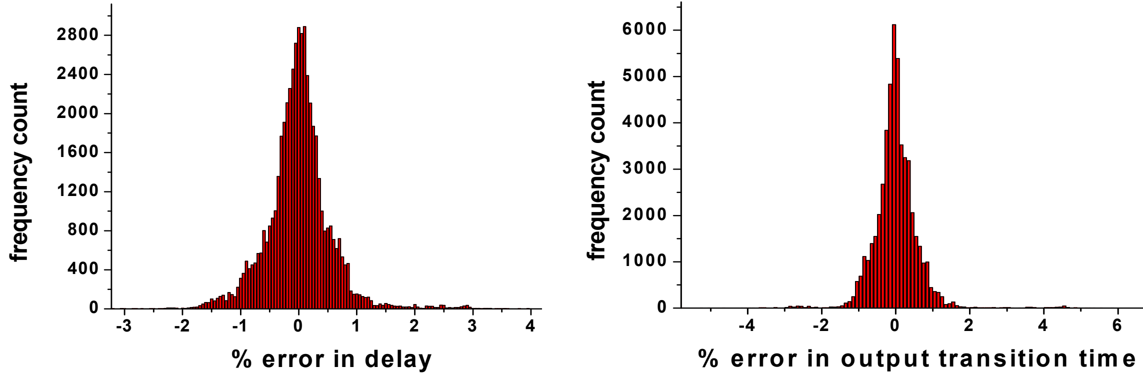


Figure 2.5. Library-characterization error for delay and transition time for more than 35,000 sample points

driver, the delay is computed appropriately using EQ 2.4 and EQ 2.5. Each set of load/transition time combination corresponds to a fixed set of coefficients k_i and l_i which are obtained by performing multiple linear-regression analyses where each gate is simulated over a range of supply-voltage variations and rise/fall transition changes. Figure 2.5 shows the error distribution of delay modeling for 0.13μ standard cells used in our experiments (inverters, nands, nors, xors, xnors, buffers of four different drive strengths) for different sets of input-slope, output-load and supply-drop combinations. The total size of the samples used in error validation exceeded 35,000. The model has the maximum error of 4.33% and 6.92% for delay and output transition time, respectively. The errors in propagation delay and transition have the standard deviation of 0.56% and 0.65%, respectively.

The input transition time at a gate *G* is a function of the supply voltages at the gates prior to gate *G*. In other words, a drop in supply voltage at a gate may impact its output transition time, and thus, the delay of all the *downstream* gates in the path. For now, we assume that the transition time of the input to a gate is independent of supply drop and equal to the transition time with nominal supply voltage. How to take into account the changes in transition times in the analysis will be discussed later in Section 2.5.

2.2.1 Path-based circuit-delay model

We now consider the variation of the delay, $\Delta\tau_{Path}$ of a circuit path due to supply-voltage variations at different supply connections along a path, as shown in Figure 2.2. Using EQ 2.4 and EQ 2.5, the change in the delay of a path consisting of N gates, numbered in topological order, is given by:

$$\tau_{path} = \tau_{path0} + \sum_{n=1}^{N} \sum_{\Omega} k_{n,i} \Delta Vdd_n^{\alpha_i} \Delta Vss_n^{\beta_i} \Delta Vdd_{n-1}^{\delta_i} \Delta Vss_{n-1}^{\gamma_i} \qquad \text{(EQ 2.6)}$$

$\alpha_i + \beta_i + \delta_i + \gamma_i \le 2$, $\alpha_i, \beta_i, \delta_i, \gamma_i \in \{0, 1, 2\}$ and $k_{n,i}, l_{ni} \in \Re$

$\forall 1 \le n \le N \wedge n - 1 \in input(n)$

where τ_{path0} is the path delay with nominal supply at all the gates constituting the path; $k_{n,i}$, $l_{n,i}$ are the delay coefficients of gate n; ΔVdd_n, ΔVss_n are the supply voltages at gate n.

For simplicity of our discussion, we assume an ideal transition between 0V and nominal Vdd at the primary input of the path, and hence, $\Delta Vdd_0 = \Delta Vss_0 = 0$. However, the analysis can be easily extended to account for non-ideal primary-input signal transitions also.

EQ 2.6 models the change in the delay of a path as a linear function of supply voltages at the individual gate connections. In the next section, we express these supply voltages as a linear function of block-currents and formulate the problem of maximizing the delay of a path as a quadratic optimization problem.

2.3 Voltage-drop sensitivity computation

The power supply network of a chip consists of the ideal supply voltage sources, power and ground wires modeled as a linear RLC network, time-varying current sources repre-

senting switching transistors and decoupling capacitances. A design can consist of millions of transistors forming a sea of gates. Simulation of a power supply network at the transistor level is not feasible even for moderate sized designs. In block-based power grid simulation, this sea of gates is grouped into large circuit-blocks and the minimum and maximum currents of each circuit-block are obtained using Powermill or Verilog simulations or estimated on the basis of a previously fabricated part. During early-mode power-grid analysis, detailed information about the distribution of total current within a circuit-block may not be available. Hence, we make the simplifying assumption that the total current in a circuit block is evenly divided among its power-supply points. This has the advantage that the voltage sensitivities can be computed with respect to the total current of a circuit-block, instead of with respect to each individual current-source point in a circuit-block. This therefore greatly reduces the number of optimization variables in our formulation and improves its efficiency.

When selecting circuit-blocks, it is therefore important that each block be sufficiently small to ensure that the spatial distribution of the currents within a circuit-block does not significantly impact the voltage response. For high-performance processors, with tight and uniform supply grids over multiple layers of metal, the spatial distribution of the total block-current is typically not significant for moderate size blocks. However, if the block size is deemed to be too large, blocks can be partitioned to obtain finer current granularity without modifying the analysis method. Also, the proposed approach can be extended for non-uniform current distributions if necessary. It is also desirable that circuit-blocks are selected such that their currents are independent, reducing the need to incorporate constraints between the currents of different blocks in the delay-maximization formulation.

The model of the power supply network consists of RLC elements, ideal time-varying current sources and ideal voltage sources. We are interested in the sensitivity of voltage variations at the supply points of logic blocks to the block-currents. Thus, the simplified linear model of the power supply network can be considered as a linear system with time-varying block-currents as inputs and voltage variations at the supply connections as outputs (Figure 2.6(a)). This system has the impulse-response function given by a matrix $H(t)$, whose element at row b and column n denotes the impulse response at node n due to current block b. Since the system is linear, the voltage response at node n, $V_{nb}(t)$ due to any current waveform of block b, i_b is given by convolution as follows:

$$V_{nb}(t) = \int_0^{\infty} i_b(t-\tau) \cdot h_{nb}(\tau) d\tau \qquad \text{(EQ 2.7)}$$

where, $h_{nb}(\tau)$ is the impulse response at node n due to the excitation at block b.

If the total number of blocks in the design is B, then the voltage response at node n due to all the blocks acting together, $V_n(t)$ is the superposition of individual responses as shown below.

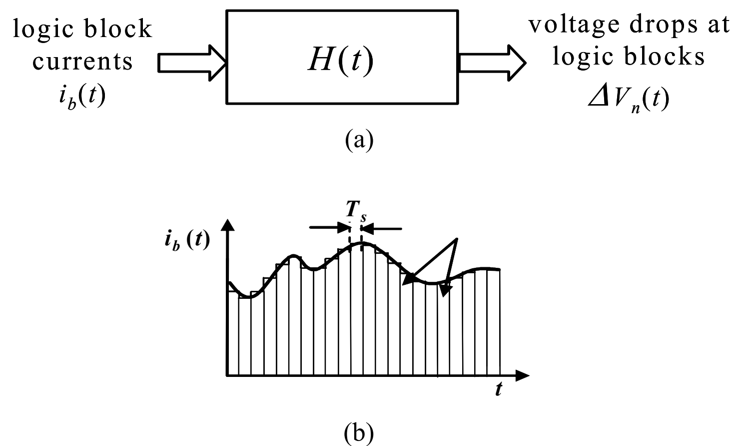


Figure 2.6. (a) Power grid as a linear system and (b) Time-varying block-current discretization into time-steps

$$V_n(t) = \sum_{b=0}^{B-1} \left(\int_0^{\infty} i_b(t-\tau) \cdot h_{nb}(\tau) d\tau \right) \quad \text{(EQ 2.8)}$$

For static or DC-only power-grid analysis, the impulse response $h_{nb}(t)$ is a DC value. There are many ways of computing the impulse response $h_{nb}(t)$ at a node n due to a block current i_b. For static analysis, the sensitivity can be computed using the inverse of conductance matrix as follows:

$$GV = I \Rightarrow V = G^{-1}V \quad \text{(EQ 2.9)}$$

$$h_{nb} = [G^{-1}]_{nb} \quad \text{(EQ 2.10)}$$

where, G is the conductance matrix of the grid and consists of nodes of both the power and the ground grid; I is the vector representing currents between the power and the ground grid and ideal voltage sources; V is the voltage drop at the power and the ground nodes.

However, since we are also interested in dynamic power-grid analysis where block-currents can vary with time, we first obtain the step response at node n by applying a unit-step current at all the supply nodes of block b and simulating the grid. In this work, we distribute each block's current evenly among all its supply nodes. However, if the block-current distribution among its supply nodes is known beforehand, unit-step response for the block can be appropriately distributed among its supply points. The unit-step response at all the nodes is discretized with a sufficiently small time-step T_s and then numerically differentiated to obtain the impulse response for the block. A key characteristic is that for typical grids, the unit-step response dampens out quickly (within K time steps) and the grid needs to be simulated only for a small period of time. Given a sufficiently fine grain discretization and sufficient simulation length of the unit-step response, arbitrary accuracy can be obtained. Any time-varying current waveform can be discretized by the time step T_s as

shown in Figure 2.6(b). We can, therefore, compute the response of any node in the network due to an arbitrary current source $i_b(t)$ using a single simulation of a unit current pulse and combining the scaled and shifted versions of this response.

After representing the voltage drops as a linear function of block-currents using EQ 2.8, we construct the constraints on block-currents. The block-currents as the optimization variables and the generation of constraints are discussed in the next section.

2.4 Block-current constraints

For dynamic IR or L*di*/d*t* analysis, the voltage drop in a time-step depends on the currents drawn in K previous cycles, where K is the number of time-steps after which the impulse response for all the block-currents have attained a steady-state value. Any switching activity which has happened more than K time-steps prior to the time-step of interest has no effect on the voltage drop and hence, has no effect on the change in delay during the time-step of interest. In all, $B \cdot K$ optimization variables need to be introduced where K is the number of variables for each block-current.

The block-currents, which are the optimization variables, are constrained by three sets of constraints. The first set of constraints states that, in each time-step, the current of a block must have a value between its maximum and minimum possible value. In this work, we assume that the minimum and maximum currents of logic-blocks are known by means of simulation or based on the estimate based on a prior manufactured part. The constraint on current of each block b in every time-step k is bounded as follows:

$$i_{b,\,min} \le i_b[k] \le i_{b,\,max} \forall k \in \{0, 1, 2 \ldots K-1\} \qquad \text{(EQ 2.11)}$$

where, $i_b[k]$ is the value of the current of block b in time-step k; $i_{b,min}$ and $i_{b,max}$ are lower and upper bounds on the current of block b; and K is the number of time-steps in the impulse response of the linear power-grid system.

We further constrain the sum of all block-currents to not exceed the peak current consumption of the chip. The upper-bound on the total chip-current is likely to be much smaller than the sum of the individual maximum of block-currents because all the logic-blocks are not likely to switch with maximum capacity at the same point in time. Hence, placing this constraint on the total chip-power consumption significantly aids in reducing the pessimism of the analysis. The second set of constraints forces an upper-bound on the total current of the chip in every time-step. This states that, while individual blocks may vary dramatically from cycle-to-cycle, the total power of the chip typically has a well known maximum current consumption. The constraint on peak power consumption of the chip is expressed as follows:

$$0 \leq \sum_{b=0}^{B-1} i_b[k] \leq I_{peak} \forall k \in \{0, 1, 2 \ldots K-1\} \qquad \text{(EQ 2.12)}$$

where, $i_b[k]$ is the current of block b in time step k; I_{peak} is the peak current consumption of the chip and B is the total number of logic-blocks in the design. The upper-bound on the total current can be computed using either chip-level Verilog simulation or by scaling the maximum power of a similar design in an older technology.

EQ 2.11 and EQ 2.12 express simple constraints on the current of individual blocks or the total current of the processor as a whole. However, in most processor designs, correlations between the currents of different blocks or between currents of a block in consecutive clock cycles will also arise. For instance, positive correlation between the current of

two pipeline stages can arise when data is passed from one pipeline stage to the next, or negative correlation may exist between the currents of two circuit blocks that operate mutually exclusively. A third set of constraints is enforced on block-currents to capture any such correlation between logic blocks.

We therefore propose to incorporate linear constraints in the formulation to express such correlations. It should be noted that the delay-maximization formulation is conservative, meaning that it will overestimate the change in delay due to supply-voltage fluctuations. This is the result of the optimization formulation, which automatically maximizes the delay change within the bounds of the provided constraints. Incorporating additional constraints in the analysis is therefore an effective method to reduce the conservatism of the analysis.

Any linear constraint can be represented in the proposed formulation and a number of different approaches for automatically generating such constraints can be used. In this chapter, we propose the use of gate-level power-simulator, such as a Verilog-based simulator, to extract correlation constraints. By simulating a large set of chip-level simulation vectors, the correlation between the current of different blocks in one clock cycle or between currents of blocks in different clock cycles can be observed and can be represented using linear constraints.

In Figure 2.7(a), we show an example of the correlation between the currents of a Multiplier and an ALU block in an Alpha processor. The X-axis of the scatter plot corresponds to the current of the Multiplier block and the Y-axis corresponds to the current of the ALU. The entire processor design was simulated, and the current of the ALU and Multiplier

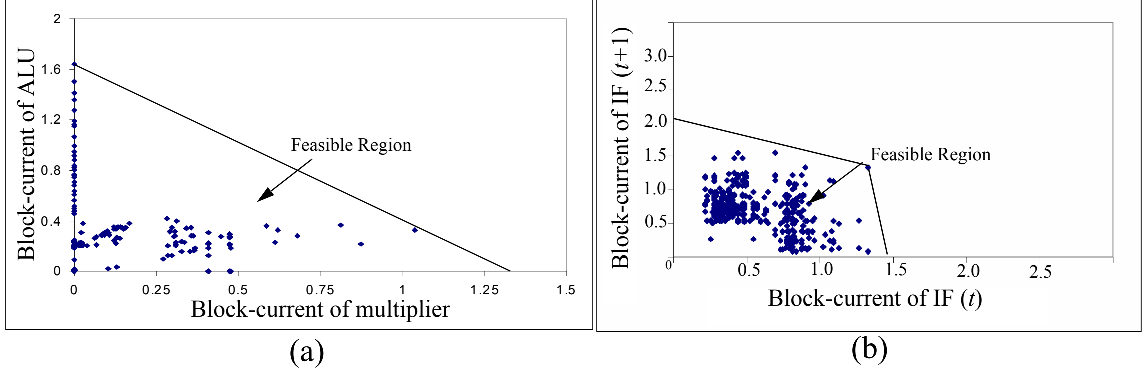


Figure 2.7. Correlation between Multiplier and ALU block-currents (a) and correlation between IF-stage in cycle t and ID-stage in cycle t+1 (b)

blocks were computed using pre-characterized power data in the cell library. Each point in the scatter plot represents a simulated clock cycle. Since the Alpha processor is a single-issue machine and was designed with clock gating for reduced power consumption, the multiplier and the ALU block cannot be active in the same clock cycle. This negative correlation is evident from the L-shaped scatter points in Figure 2.7(a). To express this correlation in the delay maximization formulation, we generate the linear constraint as shown by the solid line in Figure 2.7(a) and expressed it with the following inequality:

$$i_{mult}[k] + 1.36 i_{ALU}[k] \le 1.7 \qquad \text{(EQ 2.13)}$$

where, $i_{mult}[k]$ and $i_{ALU}[k]$ are the currents of multiplier unit and rest of ALU respectively in time step k. It is clear that the constraint in EQ 2.13 will reduce predicted delay increase of the analysis by preventing the Multiplier and the ALU from simultaneously exhibiting their maximum current values.

An example of a correlation between currents in different clock cycles is shown in Figure 2.7(b), where the current of the instruction-fetch stage in cycle t is plotted against the current of the instruction-decode stage in cycle t+1. Since data is passed from the instruction-fetch stage to the instruction-decode stage, a correlation can arise, as is visible from

the scatter plot in Figure 2.7(b). In this case, the correlation is captured using two constraints and expressed as follows:

$$1.7 i_{IF}[k] + i_{ID}[k+1] \le 3.5 \quad \text{(EQ 2.14)}$$

$$9.6 i_{IF}[k] + I_{ID}[k+1] \le 14.4 \quad \text{(EQ 2.15)}$$

where, $i_{IF}[k]$ and $i_{ID}[k]$ are the currents of instruction-fetch unit and instruction-decode unit, respectively, in time step k.

Although in this chapter, we manually extract constraints from the correlation data, it is clear that such constraints could be easily generated automatically by observing the current trace of blocks and finding a polyhedron that encompasses all generated current points. The use of gate-level power simulation has the advantage that very extensive suites of test vectors are readily available and block-current data can be obtained from them with minimum overhead during the design process. Also, gate-level simulation is typically performed for many millions of clock cycles. The proposed approach allows realistic constraints to be extracted, based on extensive simulation data. The approach obviates the need to simulate the power grid with long patterns of input vectors. In the next section, we finalize the overall path-based delay-maximization formulation.

2.5 Overall path-based delay-maximization formulation

This formulation requires the enumeration of critical paths in the circuit and application of the formulation on a path-by-path basis. The overall path-based delay-maximization problem is stated as follows:

$$\text{maximize} \quad \sum_{n=1}^{N} \sum_{i} k_{n,i} \Delta Vdd_n^{\alpha_i} \Delta Vss_n^{\beta_i} \Delta Vdd_{n-1}^{\delta_i} \Delta Vss_{n-1}^{\gamma_i}$$

$$\forall 1 \le n \le N \wedge n-1 \in input(n) \qquad \text{(EQ 2.16)}$$

$$s.t. \quad \Delta Vdd_n = \sum_{b=0}^{B-1} \sum_{k=0}^{K-1} i_b[K-k-1] \cdot h_{nb}^{Vdd}[k] \ \forall 1 \le n \le N \qquad \text{(EQ 2.17)}$$

$$\Delta Vss_n = \sum_{b=0}^{B-1} \sum_{k=0}^{K-1} i_b[K-k-1] \cdot h_{nb}^{Vss}[k] \ \forall 1 \le n \le N \qquad \text{(EQ 2.18)}$$

$$i_{b,min} \le i_b[k] \le i_{b,max} \quad \forall k \in \{0, 1 \ldots K-1\} \qquad \text{(EQ 2.19)}$$

$$0 \le \sum_{b=0}^{B-1} i_b[k] \le I_{peak} \quad \forall k \in \{0, 1 \ldots K-1\} \qquad \text{(EQ 2.20)}$$

and other block-current correlation constraints

EQ 2.16 maximizes the sum of the delay of gates in a path, numbered in a sequential manner from 1 to *N*. ΔVdd_n, ΔVss_n are the power-supply drop and ground bounce respectively in the supplies of gate *n*. EQ 2.17 and EQ 2.18 express the power-supply drop and ground bounce in the supplies of a gate *n* as a function of currents, i_b, of each block *b*. $h_{nb}^{Vdd}[k]$ and $h_{nb}^{Vdd}[k]$ are the impulse responses at time-step *k* in the power and ground grid respectively at node *n* due to a current excitation at block *b*. EQ 2.19 and EQ 2.20 are the constraints on the block-currents in each time-step *k*. Additional correlation constraints on block-currents can be incorporated in the formulation as explained in Section 2.4.

The input transition time at any gate *n* in a circuit is a function of the supply voltages at its preceding gates. It is possible that the supply voltages at the preceding gates affect input transition time at gate *n*. The delay and output transition time of the gate are depen-

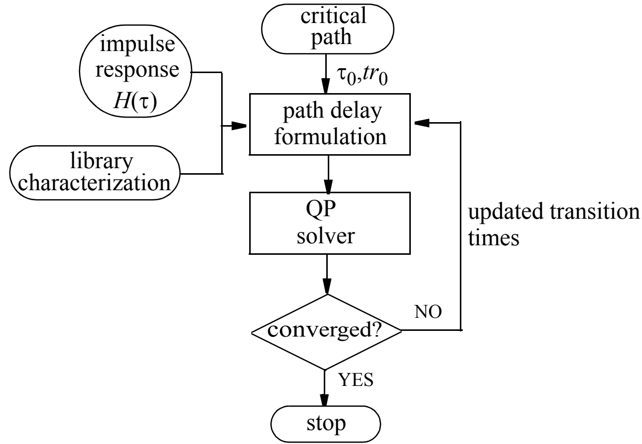


Figure 2.8. Path-based delay-maximization accounting for transition-time variations

dent on its input transition time and should, therefore, reflect the changes in input transition time. In other words, the delay coefficients for delay and output transition time for a gate *j* in EQ 2.4 and EQ 2.5 are a function of input transition time, which can vary depending on the supply at the prior gates. In our experiments, we observed that representing the delay coefficients as a function of input transition time leads to significant additional complexity of the non-linear solver used for delay-maximization formulations. Instead of adding this extra complexity, we propose a simple iterative approach (Figure 2.8) for delay correction due to changes in transition times. Given a critical path, the path delay is, first, formulated using the impulse responses of the grid and gate-delay characterization as a function of supply voltages. Next, a quadratic-programming solver (QP solver) is used to obtain the set of currents resulting in largest path delay. The new transition time of all the signals along the path computed based on these set of currents and the path delay are reformulated with the updated transition times. The new formulation is again solved and the iterations continue until the change in maximum delay from the previous iteration is less than a pre-specified value, ε.

Next, we present the block-based *circuit*-delay maximization problem which can work in conjunction with static timing analysis and does not require enumeration of critical paths.

2.6 Block-based circuit-delay model

The worst-*circuit*-delay computation problem is expressed in terms of arrival times at the gates. Computing the worst-case delay of a circuit involves determining the worst-case critical path under supply variations in the design and maximizing its delay. The arrival time at the output of a gate itself involves a *max* function. This *max* function renders the optimization problem non-convex and does not allow it to be solved directly using a convex-optimization-based non-linear-programming solver (NLP solver). We therefore show how the *max* operations can be eliminated by the introduction of slack variables with additional constraints, allowing the use of standard NLP optimization methods.

Consider a combinational circuit with *s* primary inputs (PIs), *t* primary outputs (POs) and *n* gates or wire segments. The power and ground supplies at a gate *i* are denoted as Vdd_i and Vss_i respectively. Two fictitious nodes *src* and *sink* are added to the circuit. All

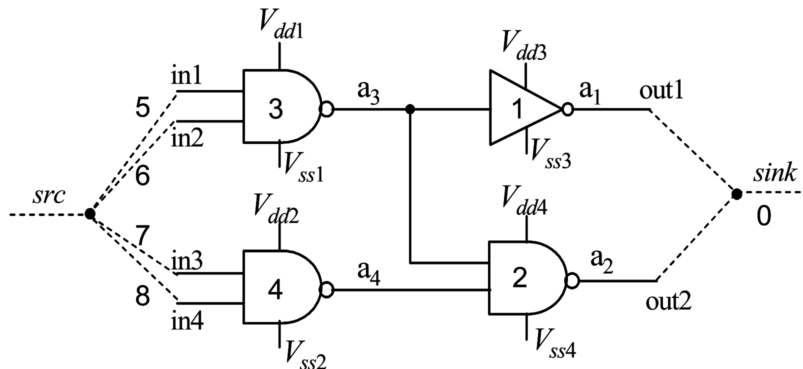


Figure 2.9. An example of a combinational circuit

the primary inputs are connected to the *src* node and all the primary outputs are joined together to form the *sink* node. The *sink* node connecting all the POs is labeled as node 0 and all other gates are numbered in the reverse topological order [14]. The arrival time at the output of a gate i is denoted by a_i. For $0 \le i \le n$, let *input*(i) be the set of indices of gates driving the inputs of gate i. For $1 \le i \le n + t$, let *output*(i) be the set of indices of gates in the fanout of gate i. For example in Figure 2.9, s=4, n=4, t=2, *input*(0)={1,2} and *output*(3)={1,2}.

For brevity, we do not differentiate between rising and falling transitions in the discussion although they have been taken into account in our implementation. Let τ_{ij} represent the delay of gate j from one of its inputs i and b_{kj} denote the delay coefficients for gate j. Then the delay-maximization problem, given the constraints on supply voltages can be stated as follows:

$$maximize\ a_0 \qquad \text{(EQ 2.21)}$$

$$s.t. \qquad a_j = max\{a_i + \tau_{ij}\} \quad 0 \le j \le n \wedge \forall i \in input(j) \qquad \text{(EQ 2.22)}$$

$$a_j = 0 \qquad n + 1 \le j \le n + t \qquad \text{(EQ 2.23)}$$

$$\tau_{ij} = \tau_0 + \sum_{\Omega} b_{kj} \Delta Vdd_i^{\alpha_i} \Delta Vss_i^{\beta_i} \Delta Vdd_j^{\delta_i} \Delta Vss_j^{\gamma_i} \quad 0 \le j \le n \wedge \forall i \in input(j) \qquad \text{(EQ 2.24)}$$

$$\tau_{i0} = 0 \qquad \forall i \in input(0) \qquad \text{(EQ 2.25)}$$

and constraints on voltage drops ΔVdd_i, ΔVss_i *at gate i*

The objective function is the worst-case arrival time at the output *sink* node. EQ 2.22, EQ 2.23 set the arrival time at the output of each gate and the *sink* node. EQ 2.24 expresses gate-delay as a function of the voltage drops at the gate and its driver. The constraints on voltage drops at each gate as a function of currents are incorporated in a set of additional constraints which will be discussed later in Section 2.5.

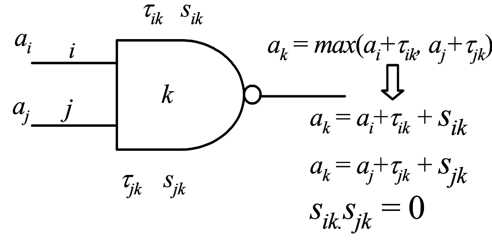


Figure 2.10. Illustration of the removal of the *max* function

Eliminating the *max* function

The delay-maximization problem described above involves the *max* function which cannot be used per se for optimization purposes. A common technique is to replace the *max* operation where inequalities are used in place of the *max* operator [14]. Although this works for problems such as gate sizing where circuit delay is minimized, it cannot to used in our proposed formulation for maximizing the delay because it makes the objective function unbounded. We, therefore, propose a different method of elimination of the *max* function by introduction of slack variables along all the possible input-output pairs in a gate as shown in Figure 2.10. For instance, in an n input gate, one slack variable is introduced for each possible path from inputs to output. This results in n slack variables for an n input gate. The complexity of the modified formulation depends on the number of multiple-input gates in the design. Using these slack variables, the arrival time at the output of an inverting gate j can be stated as follows:

$$a_j = a_i + \tau_{ij} + s_{ij} \qquad 0 \leq j \leq n \wedge \forall i \in input(j) \qquad \text{(EQ 2.26)}$$

At least one of the slacks, s_{ij} has to be zero since at least one of the possible paths from inputs to the output will be critical during circuit operation. Therefore, an additional constraint is added for each *max* function which ensures that at least one of the slack variables s_{ij} is 0:

$$\prod_{i \in input(j)} s_{ij} = 0 \qquad s_{ij} \geq 0,\ 0 \leq j \leq n \qquad \text{(EQ 2.27)}$$

This additional constraint places an upper bound on arrival times a_j and the modified maximization problem can be efficiently solved using industrial non-linear solvers. The next subsection presents the overall block-based delay-maximization formulation

2.6.1 Overall block-based circuit-delay maximization

The overall NLP formulation for the block-based circuit-delay maximization with constraints on block currents is expressed below:

$$maximize\ a_0 \qquad \text{(EQ 2.28)}$$

$$s.t. \qquad a_j = max\{a_i + \tau_{ij}\} \quad 0 \leq j \leq n \wedge \forall i \in input(j) \qquad \text{(EQ 2.29)}$$

$$a_j = 0 \qquad n+1 \leq j \leq n+t \qquad \text{(EQ 2.30)}$$

$$\tau_{ij} = \tau_0 + \sum_{\Omega} b_{kj} \Delta Vdd_i^{\alpha_i} \Delta Vss_i^{\beta_i} \Delta Vdd_j^{\delta_i} \Delta Vss_j^{\gamma_i} \quad 0 \leq j \leq n \wedge \forall i \in input(j) \qquad \text{(EQ 2.31)}$$

$$\Delta Vdd_n = \sum_{b=0}^{B-1} \sum_{k=0}^{K-1} i_b[K-k-1] \cdot h_{nb}^{Vdd}[k] \quad 1 \leq n \leq N \qquad \text{(EQ 2.32)}$$

$$\Delta Vss_n = \sum_{b=0}^{B-1} \sum_{k=0}^{K-1} i_b[K-k-1] \cdot h_{nb}^{Vss}[k] \quad 1 \leq n \leq N \qquad \text{(EQ 2.33)}$$

$$i_{b,min} \leq i_b[k] \leq i_{b,max} \quad \forall k \in \{0, 1 \ldots K-1\} \qquad \text{(EQ 2.34)}$$

$$0 \leq \sum_{b=0}^{B-1} i_b[k] \leq I_{peak} \quad \forall k \in \{0, 1 \ldots K-1\} \qquad \text{(EQ 2.35)}$$

$$\prod_{i \in input(j)} s_{ij} = 0 \qquad s_{ij} \geq 0,\ 0 \leq j \leq n \qquad \text{(EQ 2.36)}$$

and block current correlation constraints

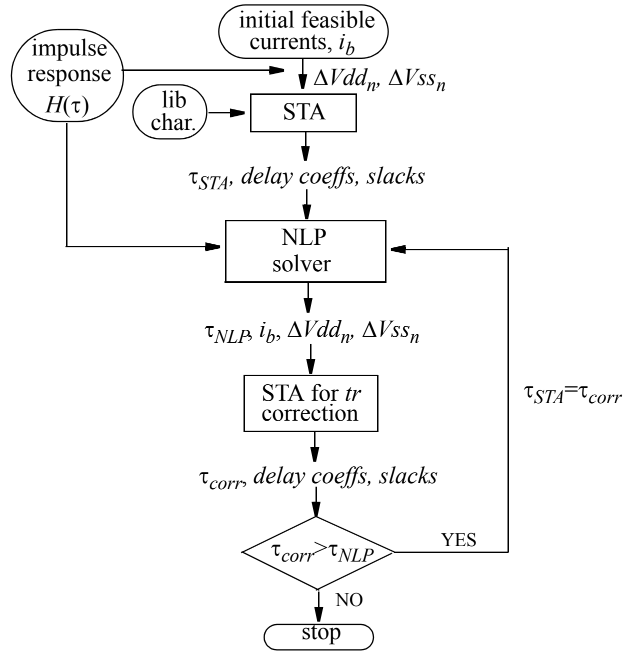


Figure 2.11. Overall block-based circuit-delay-maximization flow

EQ 2.29 and EQ 2.30 state the arrival times at the outputs of the gates and the inputs respectively. EQ 2.31 expresses gate delay as a function of the supply voltages at the gate and its driver. EQ 2.32 and EQ 2.33 express the voltage drop and ground bounce respectively at gate *n*. ($h_{nb}^{Vdd}[k]$ and $h_{nb}^{Vdd}[k]$ are the impulse responses at time-step k in the power and ground grid respectively at node n due to a current excitation at block b. EQ 2.34 and EQ 2.35 represent the constraints on currents of block b in time-step k.

In order to account for the variations in transition times of the signals, as explained in Section 2.5, we propose an iterative solution outlined in Figure 2.11. As a first step, an initial block current distribution i_b which satisfies the current constraints is identified. The constraints can be satisfied either by setting all block currents in all time steps to zero or by distributing the peak current appropriately among the blocks. These block currents are

then used to obtain the voltage drops, ΔV_n at the supply points of the gates in the given circuit. The static timing analyzer, which has been modified to incorporate supply variations, computes delay coefficients and slacks s_{ij} for each gate j on the basis of input transition times and output loading. The resulting worst-case delay from the STA, based on the initial current distribution, is denoted by τ_{STA}. The initial guess for current distribution, along with the delay coefficients and slacks obtained from STA, are used to formulate the delay-maximization problem as an NLP problem. NLP computes the block-current waveforms so as to maximize the worst-case delay of the circuit which is represented as τ_{NLP}. This worst-case NLP delay, τ_{NLP}, is corrected to account for any changes in the transition times by re-running STA with the new voltage-drop values computed by NLP. The delay coefficients computed during the corrected STA are used to re-formulate the NLP for the next iteration and the process is repeated until the corrected STA output, τ_{corr}, is either equal to or less than the pre-NLP delay, τ_{sta}, in each iteration.

The run-time of the NLP solver depends largely on the number of *max* functions for multi-input gates in the circuit. The next subsection describes a technique to reduce the number of inputs of a multi-input gate, thus reducing the run-time of the non-linear solver.

2.6.2 Delay-Based Circuit Pruning

Finally, we propose a delay-based circuit-pruning technique to, if possible, eliminate the inputs of multi-input gates in the circuit. Consider a single-output, n-input gate G with inputs numbered as $0,1...n$. The arrival time at the output of this gate is the maximum of the n arrival times through n possible paths from inputs to the output. As a first step, the arrival times at the inputs of the gate are computed using an ideal voltage supply and the

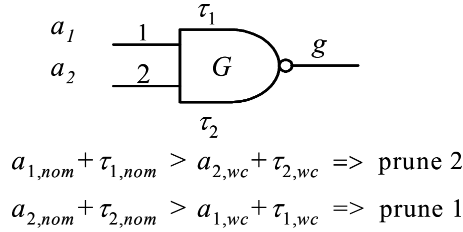


Figure 2.12. An illustration of delay-based circuit pruning to reduce the problem size

supply with the worst-case voltage margin. These are denoted as $a_{i,nominal}$. If output arrival time from any input under worst case voltage margin is found to be greater than the output arrival time from all other inputs under ideal voltage supply, the input is removed from the gate and the gate is treated as a *n*-1 input gate. Figure 2.12 shows an example when input 2 of a 2-input nand gate is removed by pruning.

The next section shows the experimental results obtained on ISCAS85 benchmarks for both static IR-only and dynamic IR+Ldi/dt power-grid analysis.

2.7 Experimental results

The proposed approach for determining the worst-case voltage drop of a given circuit was implemented and tested on ISCAS85 benchmark circuits synthesized in 0.13μ technology. A modified static timing analyzer was implemented in C++ and all experiments were run on a 1GHz SUN machine with 4GB of memory. The MINOS [70] non-linear solver, which uses the augmented Lagrangian method, was used for optimization and AMPL [3] was used as the programming interface. The power grid for the circuits was constructed in metal layers M2-M8 using pitches and widths of an industrial microprocessor design. However, if the block size is deemed to be too large, blocks can be partitioned

to obtain finer current granularity without modifying the analysis method. A partial element equivalent circuit (PEEC)-based extraction tool was used to extract the RLC parameters of the grid based on the pitches and widths of an industrial microprocessor. Package inductance and resistance of 1nH and 1mΩ were attached in series with all the C4 bumps. In all experiments, on-die decap was distributed uniformly among the lowest metal layers across the whole area of the die. The power supply network was assumed to consist of 10 logic blocks and the minimum and maximum currents of the blocks were generated according to their area estimates. The current of each block was assumed to be distributed evenly among all its supply nodes. The gates for the circuits were connected to the supply nodes in the lowermost metal layer. In order to compute voltage sensitivities to block currents, the power grid was simulated in HSPICE with a fast current ramp (approximating a step waveform) applied at the supply nodes constituting each block and the response is numerically differentiated to obtain the impulse responses at all the nodes in the grid.

ckt	num. of gates	nom. delay (ns)	traditional approach delay				path based approach delay			block based approach delay			HSPICE	
			10% margin		avg. currents									
			delay (ns)	% incr.	delay (ns)	% incr.	path delay (ns)	% incr.	run time	ckt delay (ns)	% incr.	run time	delay (ns)	% err
c17	7	0.109	0.154	41.3%	0.122	11.9%	0.131	20.2%	0.22s	0.131	20.2%	0.24s	0.129	1.81%
c432	212	1.224	1.728	41.2%	1.393	13.8%	1.471	20.2%	2.31s	1.471	20.2%	2.70s	1.479	0.54%
c499	553	0.824	1.156	40.3%	0.947	14.9%	0.964	16.9%	5.54s	0.999	21.2%	25.34s	1.011	1.20%
c880	568	1.550	2.190	41.3%	1.771	14.3%	1.854	19.6%	5.32s	1.857	19.8%	9.39s	1.916	3.08%
c1355	654	1.393	1.945	39.6%	1.586	13.9%	1.632	17.1%	6.87s	1.671	19.9%	19.46s	1.696	1.45%
c1908	543	1.912	2.680	40.2%	2.183	14.2%	2.296	20.1%	5.35s	2.296	20.1%	10.25s	2.338	1.80%
c2670	1043	1.512	2.125	40.5%	1.736	14.8%	1.815	20.0%	10.24s	1.848	22.2%	30.46s	1.957	5.55%
c3540	1492	2.063	2.907	40.9%	2.353	14.1%	2.430	17.8%	9.36s	2.462	19.3%	1m57s	2.407	2.29%
c5315	2002	1.966	2.798	42.3%	2.238	13.8%	2.338	18.9%	13.03s	2.379	21.0%	3m12s	2.492	4.54%
c6288	3595	5.175	7.424	43.5%	6.016	16.3%	6.417	24.0%	21.61s	6.591	27.4%	6m19s	6.749	2.33%
c7552	2360	2.957	4.349	47.1%	3.327	12.5%	3.582	21.1%	18.27s	3.696	24.5%	7m21s	3.806	2.89%
AVG				**41.6%**		**14.0%**		**19.7%**			**21.5%**			**2.5%**

Table 2.1. Experimental results for DC current constraints

Table 2.1 shows the results for the delay obtained using static power-grid analysis (*IR-drop* only) where only DC current constraints are applied. Each block has a minimum and maximum constraint on its current. There is a constraint on the peak current consumption of the chip. Columns 1 and 2 show the circuits and the number of gates in their netlists. Column 3 shows the nominal delay of each circuit obtained from STA. Columns 4 and 5 show circuit delay with minimum supply margin (10%) for both *Vdd* and *Vss* which is the general method currently used in traditional STA tools to estimate supply noise induced delay increase. The 10% margin is the worst case observed voltage drop in our experiments when all block currents are allowed to switch within their bounds. Circuit delay increases on average by 41.6% for this worst-case voltage margin. Columns 6 and 7 show the case when the peak current is uniformly distributed among all the blocks, which causes an increase in circuit delay by 14% on an average. Columns 8 and 9 show the *critical path delay* (only the most critical path considered) and the percentage delay increment of the circuit respectively. On average, the critical path delay is increased by 19.7%. This shows that applying 10% voltage drop to all the gates in the circuit is very pessimistic because worst-case voltage drop is typically localized to a small region on the die and most of the die area may observe considerably better supplies. On the other hand, applying average block currents underestimates the worst-case delay because it does not allow the blocks with a higher sensitivity to circuit delay to switch with higher currents and may lead to an under-estimation of the impact of supply noise on delay. It should be noted however, that the over-estimation depends on the placement of the gates in the path on the chip, giving a worse over-estimation of the delay increase for paths that are distributed over a significant area of the die. Columns 11 and 12 show the worst case *circuit delay* obtained from the

block-based formulation. On an average, the worst-case delay of the circuit increases, using the block-based formulation, by 21.52%. Column 14 presents the worst case delay of the circuit obtained from HSPICE when the circuit is simulated with voltage supply waveforms obtained from block-based circuit delay maximization. The approach has an average error of 2.5% (column 15) and a maximum error of 5.6% for the tested circuits. Columns 10 and 13 show the run-times for the path-based and block-based formulations respectively for each test circuit. The runtime includes the time for parsing the netlist, levelizing the circuit and the iterative algorithms shown in Figure 2.8 and Figure 2.11. The block-based circuit delay maximization, although more expensive in terms of runtime, is more accurate in predicting the worst-case delay since it targets all the paths in the circuit simultaneously. A path which is critical under nominal voltage supply may no longer remain critical under supply variations if it is located in the region with lower supply drop. On the other hand, even relatively non-critical paths can become critical if they are located in the worst voltage drop region. Since apriori knowledge of which path will be the most critical under worst-case supply fluctuations is difficult, the path-based formulation underestimates the circuit delay in certain test-cases.

To test the quality of the NLP solutions obtained from block-based formulation, 50,000 random runs were performed on all the test circuits. Currents for all the logic blocks were generated randomly so as to strictly meet the peak current consumption constraint:

$$\sum_{b=0}^{B-1} i_b = I_{peak}$$

Table 2.2 compares the worst-case delay obtained from block-based NLP with the delay obtained from random simulations. The difference between the delays has a maximum

ckt	delay (ns) random runs	delay (ns) NLP	%diff
c17	0.131	0.131	0.00%
c432	1.467	1.471	0.27%
c499	0.998	0.999	0.10%
c880	1.854	1.857	0.16%
c1355	1.667	1.671	0.24%
c1908	2.288	2.296	0.35%
c2670	1.832	1.848	0.87%
c3540	2.453	2.462	0.37%
c5315	2.379	2.379	0.00%
c6288	6.572	6.591	0.29%
c7552	3.693	3.696	0.08%
AVG			**0.25%**

Table 2.2. Random-run comparison with NLP

error of 0.87%. Additionally, the delay estimated by the block-based formulation was always greater than the delay obtained from random simulations. This indicates that the block-based NLP result is likely to be near the global maximum for circuit delay. Figure 2.13 plots the delay distribution obtained from 50,000 random runs of circuit c880. Every random run strictly met the current constraints and hence lies in the feasible region of the block-based NLP. The arrows in Figure 2.13 show the nominal delay, delay obtained from

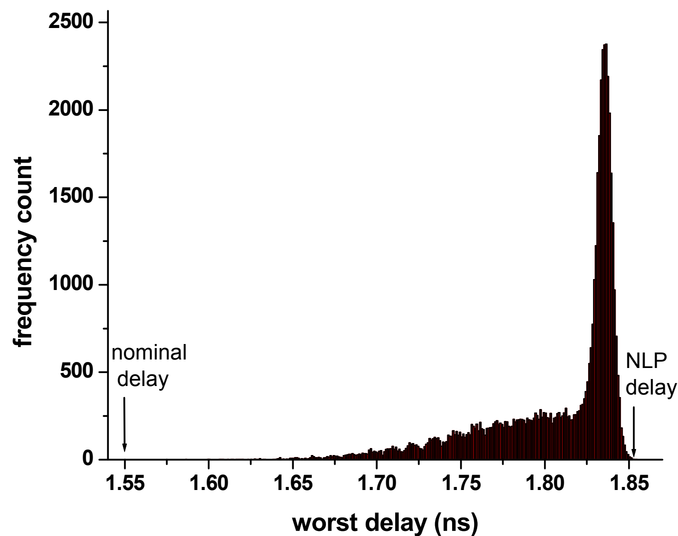


Figure 2.13. Comparison of block-based NLP with random runs for c880

average uniform block currents and the worst-case delay obtained from optimization. Table 2.2 shows that the worst-case block-based NLP delay for every circuit is always greater than the worst case delay obtained from random runs in all the circuits, showing the effectiveness of the approach. The plot also shows that there is a significant probability of obtaining a delay increase that is very close to the value obtained by the block-based NLP analysis.

The proposed approach was also tested for the dynamic case when the currents drawn by the logic blocks vary with time so as to maximize the worst-case circuit delay. In the dynamic scenario, the delay of a circuit in a given cycle depends not only on the values of currents in the current cycle, but also on the history of currents in the previous *K* cycles/ time-steps where *K* is the total number of cycles/time-steps in the impulse response of the linear power grid system. Figure 2.14 shows the step response at a node due to four different block currents. Assuming the operating frequency to be 1GHz, the step response is observed to reach the steady DC state after 50 cycles. In our experiments, the time step was taken to be one clock cycle and the delay of a circuit in a cycle depends on the activity of block currents in prior *K*=50 cycles. Table 2.3 shows the worst circuit delays, run-time

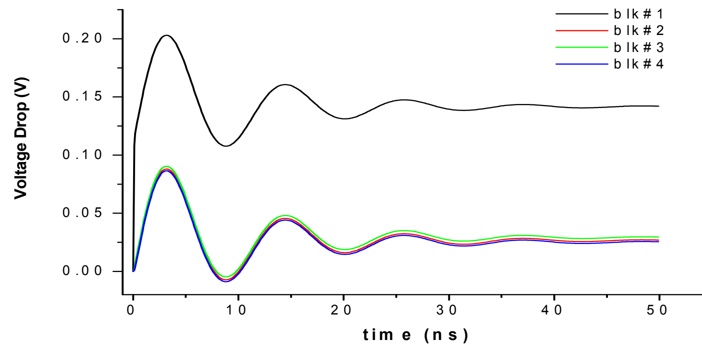


Figure 2.14. Step response at a node due to different blocks (resonance frequency = 50MHz)

ckt	nom. delay (ns)	current ramped from zero to avg.		path based approach delay			block based approach delay		
		delay(ns)	%incr.	delay(ns)	%incr.	run time	delay(ns)	%incr.	run time
c17	0.109	0.129	18.35%	0.145	33.03%	0.31s	0.145	33.03%	0.32s
c432	1.224	1.441	17.73%	1.616	32.03%	3.65s	1.616	32.03%	7.54s
c499	0.824	0.973	18.08%	1.004	21.84%	20.01s	1.089	32.16%	44.21s
c880	1.550	1.823	17.61%	2.009	29.61%	35.27s	2.034	31.22%	41.05s
c1355	1.393	1.639	17.66%	1.747	25.41%	31.66s	1.826	31.08%	70.21s
c1908	1.912	2.268	18.62%	2.524	32.00%	23.50s	2.524	32.00%	40.95s
c2670	1.512	1.783	17.92%	1.972	30.42%	37.57s	2.010	32.94%	2m11s
c3540	2.063	2.422	17.40%	2.683	30.05%	51.70s	2.755	33.54%	5m08s
c5315	1.966	2.333	18.67%	2.593	31.89%	52.48s	2.627	33.62%	11m05s
c6288	5.175	6.225	20.29%	6.878	32.91%	2m11s	7.089	36.99%	33m51s
c7552	2.957	3.497	18.26%	3.804	28.64%	50.19s	3.932	32.97%	26m54s
AVG			**18.24%**		**29.80%**			**32.87%**	

Table 2.3. Experimental results for AC (time-varying) current constraints

and memory usage for the different benchmarks. Column 2 shows the circuit delay under nominal supply. Columns 3 and 4 show the increase in circuit delay and percentage delay increment when block currents are switched from zero to their respective average values within a clock cycle. The delay of the circuits is computed in the clock cycle of the occurrence of the worst-case voltage drop. Columns 5 and 6 list the critical path delay and percentage increase in delay obtained from the path-based optimization. Worst-case circuit delay obtained from the block-based delay maximization is shown in columns 8 and 9. On average, the delay of the circuits was observed to increase by 32.87%, demonstrating the need to incorporate the Ldi/dt drop while estimating the increase in circuit delay due to supply fluctuations. Columns 7 and 9 present the runtimes for the path-based and block-based formulations respectively, demonstrating again that the block-based approach, although more expensive, is more accurate in predicting the worst circuit delay. Since it is difficult to ascertain apriori which path will be the most critical under worst-case supply fluctuations, the path-based formulation underestimates the circuit delay in certain cases.

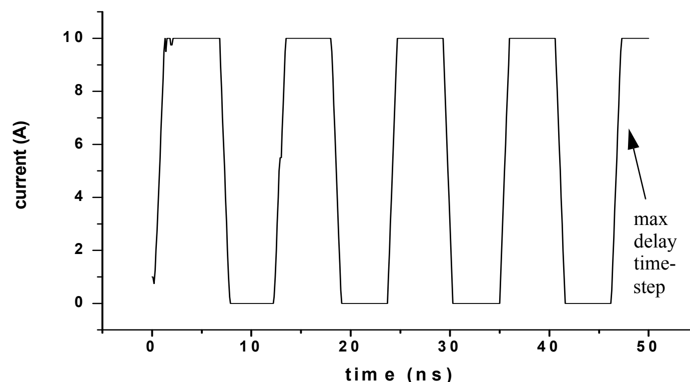


Figure 2.15. Time-varying total chip-current obtained from the block-based delay maximization

Figure 2.15 shows the total time-varying chip current obtained after the block-based NLP maximizes circuit delay at t=50ns. It is interesting to observe that the total current of the chip oscillates between the minimum possible current to the maximum possible current with a frequency of 50MHz, which matches exactly to the resonance frequency of the step response as shown in Figure 2.14. Hence, the NLP has naturally captured the innate behavior of the system. The total current gets distributed among the individual blocks based on the sensitivities of the supply nodes of the circuit and the dependence of circuit delay on the supplies at different nodes in the power grid.

Table 2.4 shows reduction in run-time of the block-based formulation due to circuit pruning. A worst-case voltage margin of 10% was assumed for pruning. Column 2 and 3 show the number of nets removed and percentage reduction in the pruned circuits. The run-time of the implementation (including pruning time) of the proposed approach on the pruned circuits is shown in column 5. Column 6 shows the reduction in runtime for the test circuits after pruning, which is 47% of the intial runtime on average. The computed worst-case delay was identical for the original and the pruned circuit.

ckt	# gates initial	# inputs pruned	runtime initial	runtime pruned	runtime reduction
c17	7	2	0.32s	0.30s	6.25%
c432	212	123	7.54s	4.65s	38.33%
c499	553	234	44.21s	20.66s	53.27%
c880	568	287	41.05s	18.46s	55.03%
c1355	654	300	70.21s	30.08s	57.16%
c1908	543	260	40.95s	21.50s	47.50%
c2670	1043	453	2m11s	49.09s	62.53%
c3540	1492	723	5m08s	2m00s	61.04%
c5315	2002	1123	11m05s	4m52s	56.09%
c6288	3595	1048	33m51s	23m06s	31.76%
c7552	2360	1085	26m54s	13m44s	48.95%

Table 2.4. Run-time reduction due to circuit pruning

2.8 Conclusions

In this chapter, we proposed a path-based and a block-based analysis approach for computing the maximum circuit delay under power supply fluctuations. The analyses are based on the use of superposition, both spatially across different circuit blocks, and temporally in time. The approaches are vectorless and take both IR drop as well as L*di*/d*t* drop into account. The path-based approach computes the maximum possible delay of a given critical path in the presence of supply variations, while the block-based approach does not require apriori knowledge of the critical paths in a circuit. The delay maximization problem is formulated as non-linear optimization problem with constraints on currents of macros or circuit blocks in the design. We show how correlations between currents of different circuit blocks can be incorporated in the formulations using linear constraints. The proposed methods were validated on ISCAS85 benchmark circuits and an industrial power supply grid, and demonstrate a significant reduction in pessimism during worst-case circuit delay computation.

CHAPTER III

POWER-SUPPLY-DROP ANALYSIS

3.1 Introduction

In this chapter, we propose approaches for computing the supply-voltage drop in a power distribution network. We propose an approach for supply-drop analysis based on constraints on block-currents as described in Chapter II. This is particularly useful in early-mode power-grid analysis when detailed information about various block-currents is unknown. The worst-case supply-drop computation is formulated as a linear-optimization problem with constraints on the currents drawn by different logic blocks. The analysis considers both IR drop and Ldi/dt drop in a power supply network and can take into account both spatial and temporal correlations in block-currents.

We also propose an approach for the computation of statistical parameters of supply drops (IR + Ldi/dt) at all the nodes in a power grid. The variability is defined over the input-vector space where different vectors cause different currents to be drawn from the power supply network. We model the currents drawn by major blocks in the design as stochastic processes and extract their statistical information which includes correlations between different blocks both in space and in time. We present an approach to propagate this information through the linear model of a power grid and show how we can obtain the distribution of voltage drops at any node in the grid.

The statistical characteristics of supply variations can be useful in a number of ways. First, they enable the designer to identify the regions in the grid which are more likely to fail and should be given higher priority when the grid is corrected. The probability distributions of the voltage drops can also be used to obtain the distribution of the delay of gates in the critical paths of a circuit, which can be used in statistical timing analysis to compute the probability distribution of the circuit delay. In our analysis, we found that the occurrence of the worst-case drop is an extremely rare event. This demonstrates that the traditional worst-case analysis, where each gate delay is characterized with the worst-case drop, can be very conservative and illustrates the need for a statistical power-grid analysis approaches. A particular advantage of the proposed approach is that the statistical information is obtained directly from block-currents and allows for very large sets of input vectors to be incorporated in the analysis, which is not feasible in traditional power-grid simulation. We implemented the proposed approach on a number of grids, including a power grid extracted from an industrial processor design. We compared the results against SPICE simulation and demonstrate the efficiency and accuracy of the proposed method.

The remainder of the chapter is arranged as follows. Section 3.2 and Section 3.3 describe the proposed constraint-based worst-case supply-drop analysis and the proposed statistical supply-drop analysis, respectively. Section 3.4 presents the experimental validation of the proposed approaches and conclusions are summarized in Section 3.5.

3.2 Constraint-based early supply-drop analysis

We propose a worst-case supply-drop computation approach early in the design cycle, when only limited information about the details of the design is available. The proposed

approach allows for the determination of a quick initial estimate of the worst-case drop in early phase of a design cycle. The uncertainty in currents consumed by the logic blocks is modeled as constraints on the block-currents. As more detailed information about the placement of decaps, logic blocks and their current consumption becomes available in later periods of the design cycle, the constraints are refined and used to computed the worst-case supply drop more accurately.

In Chapter II, we observed that a power grid can be expressed as a linear system and the voltage response at node *n* due to any current waveform of block *b* is given as follows:

$$V_n(t) = \int_0^{\infty} i_b(t-\tau) \cdot h_{b,n}(\tau) \cdot d\tau \qquad \text{(EQ 3.1)}$$

where, $h_{b,n}(\tau)$ is the impulse response at node *n* due to the excitation at block *b*, $i_b(t)$ is the current waveform of block *b* and $V_n(t)$ is the voltage response at node *n*. If the total number of blocks in the design is *B,* then the voltage response at node *n* due to all the current blocks acting together is the superposition of individual responses as shown below:

$$V_n(t) = \sum_{b=0}^{B-1} \left(\int_0^{\infty} i_b(t-\tau) \cdot h_{b,n}(\tau) d\tau\right) \qquad \text{(EQ 3.2)}$$

which can be discretized as:

$$\Delta Vdd_n = \sum_{b=0}^{B-1} \sum_{k=0}^{K-1} i_b[K-k-1] \cdot h_{bn}[k] \quad \forall 1 \le n \le N \qquad \text{(EQ 3.3)}$$

where, *K* is the number of time-steps before the impulse response reaches a steady state and *N* is the total number of nodes in the power grid. ΔVdd_n is the drop in the supply voltage at a node *n* which depends on the history of block-currents in the past *K* time steps.

We can obtain $h_{b,n}(\tau)$ at a node n by simulating the grid with a unit step current at block b and then numerically differentiating the resulting unit step response. The worst-case supply-drop computation is formulated as a linear-optimization problem with constraints on block-currents. The types of block-current constraints and their generation have been described in detail previously in Section 2.4 of Chapter II. The supply drop maximization problem at a node n in the power grid is formulated as follows:

$$maximize \qquad \Delta Vdd_n \tag{EQ 3.4}$$

$$s.t. \quad \Delta Vdd_n = \sum_{b=0}^{B-1} \sum_{k=0}^{K-1} i_b[K-k-1] \cdot h_{bn}[k] \quad \forall 1 \le n \le N \tag{EQ 3.5}$$

$$i_{b,\,min} \le i_b[k] \le i_{b,\,max} \quad \forall k \in \{0, 1 \ldots K-1\} \tag{EQ 3.6}$$

$$0 \le \sum_{b=0}^{B-1} i_b[k] \le I_{peak} \quad \forall k \in \{0, 1 \ldots K-1\} \tag{EQ 3.7}$$

where, N is the total number of nodes in the power grid; $i_b[k]$ is the current of block b in time-step k; B is the total number of logic blocks in the design; K is the number of time-steps in the impulse response of the power grid; $h_{bn}[k]$ is the impulse voltage response at node n, due to current of block b, in time step k; $i_{b,min}$ and $i_{b,max}$ are the minimum and maximum current, respectively, of block b; and I_{peak} is the peak current consumption of the chip. The above described linear optimization problem can be solved very efficiently using commercial linear programming (LP) solvers to compute the worst-case drop at each node in the power grid.

In the next section, we describe a statistical approach to power-grid analysis.

3.3 Statistical supply-drop analysis

In this section, we present the proposed method for computing the statistical parameters of supply voltage variations at any node in the power supply network. Figure 3.1 shows the general flow of the analysis. A chip design can consist of millions of transistors forming a sea of gates. As a first step, this sea of gates is grouped into large blocks such that there is minimum correlation in the currents drawn by these blocks. Each block is simulated using a suitable power simulator, such as PowerMill to obtain the currents drawn by the blocks over time. Next, the mean, auto-correlation function, and cross-correlation functions are computed for each block-current. Since the complexity of computing these correlation functions is linear with the vector length, very large vector sequences can be accommodated, consisting of millions of cycles or more. The power supply network is modeled as a linear system with block-currents as stochastic processes characterized by the extracted statistical information. We then compute the impulse response for every node due to each block-current by simulating the power grid in SPICE or a fast linear solver. The impulse response attains its steady state quickly and the grid needs to be simu-

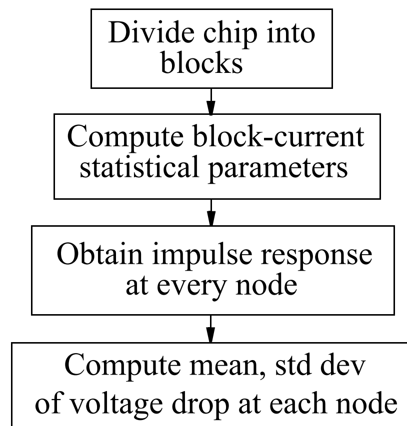


Figure 3.1. Flow diagram of proposed statistical approach

lated only for a short period of time. These impulse responses, along with the statistical parameters of block-currents are then used to generate statistical parameters for the voltage drop. Initially, the block-currents are assumed to be independent and later both spatial and temporal correlations in block-currents are incorporated in the statistical voltage-drop analysis. Finally, we show that the voltage drops closely approximate normal distributions, allowing arbitrary confidence points on the voltage distribution to be obtained. We now discuss each of the analysis steps in more detail.

3.3.1 Block-currents and voltages as random processes

During the operation of a chip, the blocks consume currents that vary in time. The actual waveforms of these currents have rather complicated shapes. In order to completely analyze the behavior of a power supply network, it is required to simulate the power grid for all possible current waveform patterns, which is clearly not feasible. Usually, power supply networks are simulated using current waveforms for only a very limited period of time which does not guarantee that the voltage variation for all possible situations is analyzed. In this Section, we discuss how currents are modeled as random processes characterized by their statistical characteristics. This approach allows for very long waveforms to be analyzed, covering a large input vector space.

We represent each block-current as a random process $i_b(t)$. This implies that the value of a block-current at each time point is considered a random variable and the current over time is a random function of time. Note that the current values at different points in time are not independent. Their values depend on their history which means that we must account for possible correlations between current values over different time points. This

random variable is described by a probability density function $p(i_b)$, which in the general case varies over time. Complete description of a random process requires specification of joint probability density function, $p(i_b(t_1),i_b(t_2),...)$ for all block-currents at any time which very difficult to obtain and analyze. Our goal therefore is to compute the statistical characteristics of the voltage drop only, i.e. the mean and variance (standard deviation), which greatly simplifies the overall computation. Hence, we only need to determine the mean, auto-correlation function and cross-correlation functions of the block-currents in order to compute the mean and standard deviation of the voltage response at any node in the power grid.

The mean and variance of a random process $X(t)$ are the expectation and the variance respectively of the random variable obtained by observing the process at some time t.

$$\mu_X(t) = E(X(t)) \tag{EQ 3.8}$$

$$\sigma_X^2(t) = E((X(t) - \mu_X)^2) \tag{EQ 3.9}$$

The auto-correlation function (or simply the correlation function) of a random process $X(t)$ is defined as the expectation of the product of two random variables obtained by observing the random process at times t_1 and t_2:

$$R_X(t_1, t_2) = E(X(t_1) \cdot X(t_2)) \tag{EQ 3.10}$$

The cross-correlation function of two random processes $X(t)$ and $Y(t)$ is the expectance of the product of two random variables obtained by observing *two* processes at times t_1 and t_2 respectively:

$$R_X(t_1, t_2) = E(X(t_1) \cdot Y(t_2)) \tag{EQ 3.11}$$

We assume that each block-current is a stationary random process. This implies that the statistical characteristics of block-currents do not depend on any time shift or equivalently, the probability density functions of block-currents do not depend on any time shift. This assumption is reasonable for currents and voltages because the only non-repeating part of current waveforms is chip initialization which is negligibly short compared to the total time of operation of a chip. For stationary processes all the moments do not depend on time and the auto-correlation function depends only on time difference $\tau = t_2 - t_1$. Hence, the mean, auto-correlation and the cross-correlation functions are given as follows:

$$\mu_X = \mu_X(t) = E(X(t)) \quad \text{(EQ 3.12)}$$

$$R_X(\tau) = E(X(t) \cdot X(t+\tau)) \quad \text{(EQ 3.13)}$$

$$R_{XY}(\tau) = E(X(t) \cdot Y(t+\tau)) \quad \text{(EQ 3.14)}$$

Another reasonable assumption which significantly simplifies the analysis is to assume the block-currents to be ergodic processes. Ergodicity implies that averaging a random process over the *sample space* for each particular time (ensemble average) gives the same result as averaging it over the *time* of one implementation of the random process (time average). This assumption is justified by the fact that if a chip operates long enough it definitely exposes all possible operation modes and all its states. Therefore, investigating any single waveform long enough is sufficient for predicting the statistical behavior of a chip.

For stationary ergodic processes with an observation window $-T \le t \le T$, the mean and the auto-correlation functions can be expressed as stated below:

$$\mu_X = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} X(t) d\tau \quad \text{(EQ 3.15)}$$

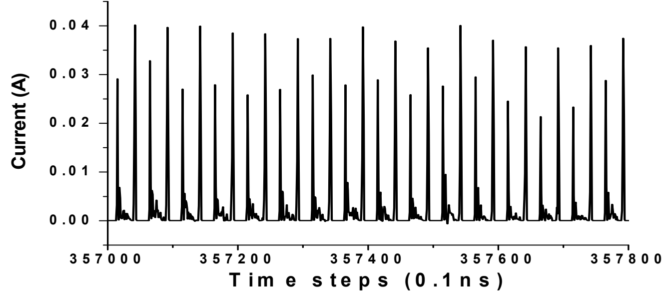


Figure 3.2. Variation of a block-current with time

$$R_X(\tau) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} X(t) \cdot X(t+\tau) dt \qquad \text{(EQ 3.16)}$$

and the cross-correlation of two random processes $X(t)$ and $Y(t)$ is given as:

$$R_{XY}(\tau) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} X(t) \cdot Y(t+\tau) dt \qquad \text{(EQ 3.17)}$$

The limits in these integrals indicate that these expressions are valid for a large value of T, i.e. the observation window for the random process $X(t)$ should be large. For our approach, this implies that the time interval of the block-current waveforms should be sufficiently large.

Now we discretize EQ3.15 through EQ3.17 for practical computation of the statistical parameters of block-currents. Figure 3.2 shows the variation in current for a short period of time for a block in an industrial microprocessor. The spikes correspond to periods of high activity in a clock cycle. As a first step, this waveform is divided into small time steps of width Δt such that there is no significant variation in block-current within a time step. Then, we compute the mean of the block-currents as follows:

$$\overline{I_b} = E(I_b) = \frac{1}{K} \sum_{k=0}^{K-1} I_{b_k} \qquad \text{(EQ 3.18)}$$

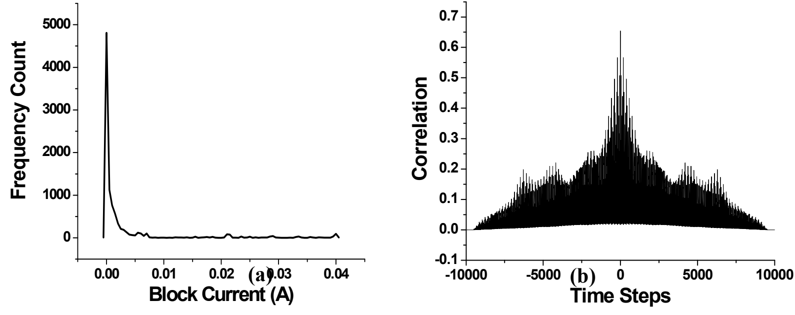


Figure 3.3. PDF (a) and auto-correlation (b) of a block-current

where K is the total number of time steps for which the waveform is observed. Similarly, the values of auto-correlation and cross-correlation functions in time step n are computed as follows:

$$R_{I_b}(n) = \frac{1}{K-n+1}\sum_{k=0}^{K-n} I_{b_k} \cdot I_{b_{k+n}} \qquad \text{(EQ 3.19)}$$

$$R_{I_i, I_j}(n) = \frac{1}{K-n+1}\sum_{k=0}^{K-n} I_{i_k} \cdot I_{j_{k+n}} \qquad \forall -K \le n \le K \qquad \text{(EQ 3.20)}$$

where, I_{b_k} is the current of block b in time step k. Figure 3.3 shows the probability distribution and auto-correlation functions for the time varying block-current shown in Figure 3.2. To compute the correlation functions for all possible values of n ($-K \le n \le K$) results in a complexity $O(K^2)$. However, we later show that we can restrict n to values less than the response time of the power grid without loss in accuracy, which is significantly smaller than K for large vector sequences and hence, the complexity is linear with K.

In the next subsection, we describe the computation of mean and standard deviation of voltage supply fluctuations, using the statistical parameters of block-currents as derived in this section.

3.3.2 Linear system with stochastic excitations

An important property of a linear system is the fact that if we know the impulse response function and the expectance and auto- correlation functions of a signal at the input of linear system, we can compute the expectance and auto-correlation function at the output of the system. The output of a linear system, $Y(t)$ can be expressed as the convolution of impulse response, $h(t)$ with its input random process $X(t)$ as given in EQ3.1. Using stationarity and ergodicity assumptions, we can average the expression over time to get the expectance of the signal at the output of the linear system.

$$\bar{Y} = E(Y(t)) = E\left(\int_0^{\infty} X(t-\tau)\cdot h(\tau)\cdot d\tau\right) = \int_0^{\infty} E(X(t-\tau))\cdot h(\tau)d\tau \qquad \text{(EQ 3.21)}$$
$$= \bar{X}\cdot\int_0^{\infty} h(\tau)d\tau$$

This implies that the mean of the random process at the output of a linear time-invariant system in response to random process $X(t)$ is equal to the mean of $X(t)$ multiplied by the dc response of the system. This property is useful in power-grid analysis because we can obtain the exact value of mean of the supply voltage variations without any complex computation if we know the mean of the input block-currents and the dc response of the network. We simulate the power grid in order to obtain the unit step response of the system and observe the response till it dies out and attains a steady state value. This steady state value is the dc response of the system and can be used in computing the mean of the supply voltage fluctuations.

Similarly, we can compute the second moment of the random signal at the output of the linear system using the auto-correlation function of the input signal as follows:

$$\overline{Y^2} = E(Y^2) = E\left(\int_0^{\infty} X(t-\tau)\cdot h(\tau)d\tau \cdot \int_0^{\infty} X(t-\tau)\cdot h(\tau)d\tau\right) \quad \text{(EQ 3.22)}$$

$$= E\left(\int_0^{\infty} X(t-\tau)\cdot h(\tau)d\tau \cdot \int_0^{\infty} X(t-\tau)\cdot h(\tau)d\tau\right) = \int_0^{\infty}\int_0^{\infty} R_X(\tau_1-\tau_2)\cdot h(\tau_1)\cdot h(\tau_2)d\tau_1 d\tau_2$$

Once we determine the second moment and the mean of the output process, we can obtain its variance using the following well known expression:

$$\sigma_Y^2 = \overline{Y^2} - (\bar{Y})^2 \quad \text{(EQ 3.23)}$$

EQ3.21 through EQ3.23 can be used to compute the mean and variance of the voltage drop at any node *n*. Let the number of blocks in the circuit be *B*. The voltage response at node *n* due to all the current blocks acting together is the superposition of individual responses and is given by EQ3.2. Thus the mean of the voltage response at node *n* is given by,

$$\overline{V_n} = \sum_{b=0}^{B-1} E(I_b)\int_0^{\infty} h_{b,n}(\tau)\cdot d\tau \quad \text{(EQ 3.24)}$$

For practical implementation, we can discretize this expression as follows:

$$\overline{V_n} = \sum_{b=0}^{B-1} \overline{I_b} \sum_{i=0}^{M-1} h_{b,n}[i], \quad \text{(EQ 3.25)}$$

where M is the total number of time steps after which the impulse response $h_{b,n}$ remains zero.

Now, we discuss the computation of variance of the voltage drop for the cases when the block-currents are independent and when they are correlated. We also discuss the run time complexity in both cases.

Block-currents considered independent

If the block-currents are assumed to be independent, then the second moment (variance) of the voltage response due to all the blocks acting together depends only on the auto-correlation functions of the individual block-currents and is equal to the sum of the second moments (variances) of voltage responses due to individual block-currents.

$$\overline{V_n^2} = \sum_{b=0}^{B-1} \int_0^{\infty} \int_0^{\infty} R_{I_b}(\tau_1 - \tau_2) \cdot h_{b,n}(\tau_1) \cdot h_{b,n}(\tau_2) d\tau_1 d\tau_2 \qquad \text{(EQ 3.26)}$$

We again discretize the above expression as follows:

$$\overline{V_n^2} = \sum_{b=0}^{B-1} \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} R_{I_b}[i-j] \cdot h_{b,n}[i] \cdot h_{b,n}[j] \qquad \text{(EQ 3.27)}$$

where M is the number of time steps until the voltage response to the impulse current becomes zero or the step response reaches its steady state, K is the number of time steps used in computing the auto-correlation function of $I_b(t)$ (the number of time steps over which the block-currents are observed) and B is the total number of blocks. It is important to note that the values of M and B are much less than K. In EQ3.27, we need to compute the auto-correlation function only for time steps n before which the impulse response has died down and is zero i.e. $1 \leq n \leq M$. Thus any temporal correlation in current waveforms, shifted in time beyond the time period taken by the impulse response to reach its steady state can be ignored. Hence, the complexity of computing auto-correlation functions for B

blocks (EQ3.19) is $O(KMB)$ which is linear in the length of a block-current waveform K and, the complexity of EQ3.27 is $O(M^2B)$ which is independent of block-current waveform length. Thus, it is possible to run millions of vectors and use the statistical information of corresponding current waveforms to obtain the statistical parameters of voltage drop.

Block-currents considered to be correlated

If the block-currents are correlated, the second moment and the variance of the voltage response due to all the blocks acting together depends not only on the auto-correlation functions, but also on the cross-correlations between different block-currents. In general, when each block-current is correlated both spatially and temporally to other block-currents, the second moment of the voltage response at any node n is given by,

$$\overline{V_n^2} = \int_0^{\infty}\int_0^{\infty}\sum_{j=0}^{B-1}\sum_{k=0}^{B-1} R_{I_j, I_k}(\tau_1 - \tau_2) \cdot h_{j,n}(\tau_1) \cdot h_{k,n}(\tau_2)(d\tau_1)d\tau_2 \qquad \text{(EQ 3.28)}$$

or,

$$\overline{V_n^2} = \sum_{i=0}^{M-1}\sum_{j=0}^{M-1}\sum_{k=0}^{B-1}\sum_{l=0}^{B-1} R_{I_j, I_k}[i-j] \cdot h_{j,n}[i] \cdot h_{k,n}[j] \qquad \text{(EQ 3.29)}$$

where R_{I_j, I_k} is the cross-correlation function between block-currents $i_j(t)$ and $i_k(t)$. In general, not all the blocks are strongly correlated with each other and cross-correlation of only those blocks which have significant correlations can be considered with limited impact on the accuracy. If each block b in the design is correlated with C other blocks, then only BC correlations need to be considered instead of a total possible B^2, where $C \ll B$. Again, although the complexity of computing the cross-correlation function between two waveforms of length K is quadratic in K, we only need values of $R_{I_j, I_k}[n]$ such that $n<M$. Thus, the complexity of computing all the auto-correlations for all the blocks and cross-correla-

tions for the possible BC combinations is $O(KMBC)$ while the complexity of computing the second moment from the correlation functions is $O(M^2BC)$. The overall complexity is therefore still linear with K.

Till now, we considered both spatial and temporal correlations in block-currents. Typically in a design, blocks will have much larger correlation in space as compared to the correlation in time. If the block-current correlations in time are ignored and each block is assumed to be spatially correlated to C other blocks, then as a special case, EQ3.29 reduces to:

$$\overline{V_n^2} = \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} \sum_{k=0}^{B-1} \sum_{l=0}^{B-1} R_{I_j, I_k}[0] \cdot h_{j,n}[i] \cdot h_{k,n}[j] \qquad \text{(EQ 3.30)}$$

The run time of the analysis improves in this case because now the complexity of computing cross-correlation coefficient for the block-currents is $O(KBC)$ instead of a previous $O(KMBC)$ for computing the auto-correlation and cross-correlation functions.

3.3.3 Voltage-drop probability distribution

The currents of a block vary in accordance to the input vectors applied to the circuit and generally show a large variation. The shape of the probability distribution of these block-currents is not at all fixed and may vary depending on the size of the blocks, the functionality of the block and various other factors. Thus we cannot definitely state anything about the shape of the probability distribution of the current blocks.

However, the overall voltage drop is the sum of the voltage drops due to all the current sources acting together. The blocks have been formed such that there is minimum correlation between the current of the blocks. Thus, assuming that the number of the blocks is rel-

atively large and most of the blocks are independent from each other, the central limit central limit theorem can be applied which states that if X_1, X_2...X_n are independent random variables, then the random variable formed by summing the variables $Y=X_1+X_2+...+X_n$ has a distribution which approaches a normal distribution for large values of n. Thus, if the design is divided into large number of blocks with most of the blocks being independent, the overall drop at any node can be approximated as having a Gaussian distribution function.

This means that while the proposed approach computes only the mean and variance of the voltage drop at a particular node, the voltage distribution can be approximated as a normal distribution. We show in Section 3.4 that the normal distribution closely fits the exact distribution. Using a Gaussian approximation of the voltage distribution, we can compute any desired confidence interval of the voltage drop, such as the 95% or 99% voltage confidence points.

3.4 Experimental results

In this section, we present the experimental results for the proposed supply-drop analysis techniques.

3.4.1 Constraint-based early supply-drop analysis

The proposed constraint-based approach for determining the worst-case voltage drop was implemented and tested on a number of grids of different sizes for both flip-chip and wire bond package models. Grid-1 through Grid-8 are grids of different sizes in 9 layers of metal, generated using pitches and widths of an industrial microprocessor design. Grid-9

is the grid of an industrial processor, extracted using a commercial extraction tool. For each chip, the design was partitioned into a number of blocks. The maximum and minimum current of each block, and the total maximum power of the chip were then obtained through area estimates. Table 3.1 shows the results for worst-case voltage drop computation, using the approach described in Section 3.2. We compare the obtained results with two traditional approaches for voltage-drop analysis. In the first approach (*Peak Curr*) shown in Table 3.1, all blocks are assigned their maximum switching current, so as to draw peak simultaneously. In the second approach (A*vg. Curr*), we assign an average current to each block. The last column shows the voltage drop obtained from the constrained maximization approach, where blocks with low sensitivity will be switching with lower currents while blocks with higher sensitivity will switch with higher currents. The current drawn by each block will change in every clock cycle so as to maximize the voltage drop at a given node due to both IR drop and L*di*/d*t* drop. Table 3.1 shows that the peak current approach overestimates the worst-case voltage drop by a maximum of 64% and by 37%

Table 3.1. Comparison of worst-case voltage drops

Grid	Grid Type	# of nodes	# of Blocks		Worst voltage drops	
				Peak Curr/ Block (mV)	Average Curr/ Block (mV)	Constr. Max (mV)
Grid-1	WB	1051	10	258.2	96.8	170.8
Grid-2	WB	1051	16	295.3	105.5	193.3
Grid-3	FC	1691	16	121.9	43.5	109.0
Grid-4	WB	1691	20	195.2	90.1	166.8
Grid-5	FC	2438	20	172.2	57.4	147.7
Grid-6	WB	2438	25	232.8	76.7	141.9
Grid-7	FC	3818	25	149.1	43.9	112.9
Grid-8	WB	3818	30	247.2	81.9	178.3
Grid-9	FC	157180	30	190.3	69.2	134.7

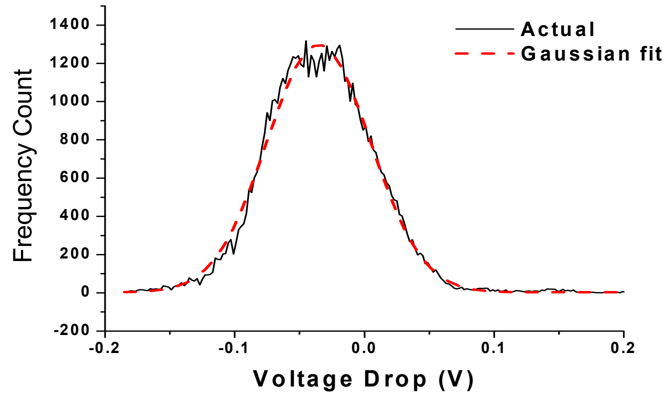


Figure 3.4. Probability distribution of the overall worst-case drop

on average over all test cases. On the other hand, the average current approach underestimates the worst-case drop by as much as 61% and by 51% on average.

3.4.2 Statistical supply-drop analysis

The proposed approach for determining the mean and standard deviation of voltage drop was implemented and tested on a number of grids of different sizes for both flip-chip and wire bond package models. In this case, the block-currents were known and generated by simulating an industrial microprocessor for thousands of cycles. The statistical parameters of the voltage drop were computed both assuming the block-currents to be independent as well as by taking block-current correlations (spatial and temporal) into account. We discussed in Section 3.3.3 that the probability distribution function of the voltage response was expected to be close to Gaussian if the number of blocks in the chip is large and most of the blocks are uncorrelated. We found that this assumption held for all the grids that we tested. Figure 3.4 shows the exact probability distribution function (PDF) of the voltage drop at a node in a grid and its Gaussian fit, which forms a close approximation. The exact

Table 3.2. Mean, standard deviation and 95% interval of the voltage drops

Grid	Grid type	# of nodes	# of blocks	Mean (mV)	Std dev (mV)	95% conf int. (mV)	Maxdrop (mV)
Grid1	WB	3772	12	50.3	66.1	179.9,-79.3	327.6
Grid2	FC	3772	12	39.6	49.4	136.4,-57.2	281.3
Grid3	WB	7712	20	56.9	67.3	188.8,-75.0	297.1
Grid4	FC	7712	20	48.1	51.5	149.0,-52.8	169.1
Grid5	WB	17037	30	39.0	47.6	132.3,-54.3	231.7
Grid6	FC	17037	30	33.5	36.5	105.0,-38.0	186.0
Grid7	WB	32897	40	103.4	111.2	321.4,-114.6	376.9
Grid8	FC	32897	40	86.6	94.5	271.8,-98.6	292.0
Grid9	FC	157180	40	88.2	113.6	310.9,-134.5	-

PDF was obtained by simulating the grid with SPICE for the entire length of the current vectors and statistically analyzing the resulting voltage drops.

Table 3.2 shows the mean and standard deviation computed using the proposed approach, assuming block-current independence. The 95% confident points are also shown, obtained using a Gaussian fit of the voltage drop PDF. The maximum voltage drop, as observed during SPICE simulation of the complete set of vectors is also shown. Grid-1 through Grid-8 are different size grids in 9 layers of metal, generated using pitches and widths of an industrial microprocessor design with a PEEC based extraction tool. The metal lines in the grid are modeled as an RLC network, which consists of the self inductance, capacitance and resistance of the wires. Mutual inductance of on-chip interconnects was ignored for simplicity. Grid-9 is the grid of an industrial processor, extracted using a commercial extraction tool and consists of over 1 million elements. I/O pads were modeled using an industrial package model. It is interesting to observe that the maximum voltage drop is between 2.17 and 4.89 times the standard deviation higher than the mean. This demonstrates that the occurrence of the worst-case drop is extremely rare.

Table 3.3. Comparison with HSPICE

Grid	Proposed Approach		HSPICE		% Error in Std Dev	Run Time
	Mean (mV)	Std Dev (mV)	Mean (mV)	Std Dev (mV)		
Grid-1	50.3	66.1	50.3	78.2	15.47%	48s
Grid-2	39.6	49.4	39.6	57.8	14.53%	48s
Grid-3	56.9	67.3	56.9	79.0	14.81%	1m40s
Grid-4	48.1	51.5	48.1	60.1	14.31%	1m29s
Grid-5	39.0	47.6	39.0	54.6	12.82%	2m37s
Grid-6	33.5	36.5	33.5	40.7	10.32%	2m41s
Grid-7	103.4	111.2	103.4	123.8	10.18%	3m14s
Grid-8	86.6	94.5	86.6	103.8	8.96%	3m25s
Grid-9	88.2	113.6	-	-	-	3m31s

Table 3.3 compares the mean and standard deviation of the voltage drops computed with the proposed approach against those obtained from SPICE simulation. Grid-9 could not be simulated in SPICE for the whole current waveforms because of its size. The mean of the voltage drop does not depend on correlations between block-currents and is only affected by the error due to the discretization of current waveforms. Hence, the mean computed with the proposed method was found to be within 1% of those obtained with SPICE simulation. There is some error in the standard deviation of the voltage drops since the block-currents are assumed to be independent. As a general trend, the error reduced with increasing number of nodes in the grid, which is a favorable characteristic since most industrial grids are very large. This results from the fact that with increasing size of the grid, the number of C4 bumps/wire bond pads in the grids increase, which attenuates the effect of current correlations on the voltage drop. Also, wire bond chips had a greater error due to correlation as compared to chips with flip-chip pads. The last column shows the run time

Table 3.4. Effect of correlations on accuracy and run-time

Degree of Correlation	No. of Cross Correlations	% Error in Std Dev	Run Time
None	0	10.32%	2m41s
2	60	8.03%	14m53s
4	120	6.64%	24m48s
6	180	5.89%	36m36s
8	240	5.40%	51m03s
10	300	4.42%	1h9m24s
15	450	2.84%	1h37m41s

for finding the mean and standard deviation at a single node. This run time includes the time to compute the mean, the auto-correlation functions of the currents and computation of the standard deviation. Note that the required time for computing the voltage statistics for additional nodes in the circuit would be substantially less since the auto-correlation and the cross-correlation functions of block-currents need to be computed only once.

In Table 3.4, we show the effect of incorporating spatial and temporal correlations between the block-currents on the run time and accuracy. We assume blocks that are close together to have a large correlation among them and assume distant blocks to be independent. Column 1 indicates the degree of correlation between adjacent blocks. A degree of correlation n implies that each block-current is correlated with its n neighbors. Column 2 gives the total number of cross-correlations taken into account. We compute the cross-correlation functions of a particular block with all its neighboring blocks and choose n blocks with the largest cross-correlation coefficient. We observe that the accuracy of the computation can be improved with reasonable additional run-time.

3.5 Conclusions

In this chapter, we presented two new approaches for analyzing the power supply drop. The first approach conservatively computes the worst-case supply drop, early in the design flow when detailed information of the design is not available. The second approach computes the statistical parameters of supply voltage fluctuations with variability in block-currents. The analyses consider both IR drop and Ldi/dt drop in a power supply network and can take into account both spatial and temporal correlations in block-currents. The analyses were implemented and tested on a number of grids, including the power grid of an industrial processor.

CHAPTER IV

TIMING-AWARE DECOUPLING-CAPACITANCE ALLOCATION IN POWER SUPPLY NETWORKS

4.1 Introduction

With technology scaling and decreasing supply voltage, the gate-delay is becoming increasingly sensitive to supply-voltage variation as explained in Chapter II. Therefore, it is extremely important to consider the effect of power-supply noise on circuit performance and improve the robustness of the power delivery network from the aspect of circuit timing. Capacitance between the power and ground distribution networks provides local charge storage and is helpful in mitigating the voltage drop in the presence of rapid switching current transients. Parasitic capacitance between metal lines of the power distribution grid, device capacitance of the non-switching transistors and N-Well substrate capacitance occur naturally in a power distribution network and act as implicit decoupling capacitance. Unfortunately, the amount of the naturally occurring decoupling capacitance is not sufficient to meet stringent power-supply-integrity constraints and designers have to often add substantial amount of explicit decoupling capacitance on the die at various strategic locations.

Gate capacitance of *n* or *p* type devices is normally used for the explicit decap. These explicitly added decaps not only result in area overhead, but also increase the leakage-

power consumption of the chip due to their gate-leakage current. With technology scaling, gate leakage has become a significant percentage of the overall leakage and has been cited as a significant limitation on the maximum amount of decap that can be introduced [9]. Hence, the goal of the designers is to meet the desired performance and signal-integrity constraints with the least possible total amount of explicitly added decaps.

A number of methods have been proposed to allocate decap in order to confine the voltage drops in the power grid within a pre-specified bound. The decap-allocation problem was formulated as a non-linear optimization problem in [48][73][88] with constraints on the worst-case voltage drop. An adjoint-sensitivity-based method [28] is used in [48] and [73] to obtain the sensitivities of decaps to the power-supply-noise metric. Decap-allocation methods tend to be computationally intensive because of expensive transient power-grid and adjoint-grid simulations. The method in [48] proposes a partitioning-based approach to reduce the power-grid simulation run-time during decap allocation.

The above mentioned approaches aim at reducing the voltage drop at all the supply nodes below a certain threshold for a given total decap budget. However, in high-performance designs, circuit performance is a more pressing concern and the above approaches, although optimal for supply-noise reduction, may not be optimal for optimizing circuit performance. For instance, in a logic block, only the delay of gates on the critical and near-critical paths are of concern and the gates having larger timing slacks can afford relatively higher voltage drop. To address this issue, two recent approaches [12][86] have been proposed for timing-aware reduction of power-supply noise. The approach in [12] requires enumeration of all critical paths and then formulates the problem as a non-linear optimization problem. This approach is computationally intensive, requiring many

adjoint-circuit simulations (equal to the number of gates in the enumerated paths in each iteration) during each optimization iteration. The approach in [86] uses a *prediction* and *correction* based algorithm for power noise reduction. In the *prediction* step, the amount of decap at various locations is predicted based on switching frequency and placement of standard cells. The *correction* step involves gate sizing to improve timing after placement. However, this approach is heuristic-based and may lead to over-design in certain scenarios.

In this chapter, we propose an approach that is based on static timing analysis of a circuit and does not require the enumeration of circuit paths. The approach naturally incorporates timing slacks at the gates in the decap allocation algorithm. The objective of the proposed approach is to improve the circuit performance, given a total decap budget. Rather than confining the voltage drop at all the nodes in a power grid within a pre-specified bound, the proposed approach automatically reduces the voltage drop near the timing critical regions while non-critical gates may have relatively higher voltage drops. Arrival time constraints are handled using Lagrangian relaxation and the relaxed subproblem is solved efficiently using the modified adjoint sensitivity method. We show how the sensitivities of the node voltages with respect to the decap sizes can be computed with a single adjoint simulation, significantly improving the runtime of the algorithm. We also propose a fast and accurate heuristic to allocate decaps for improving circuit performance. We implemented the proposed approaches and tested them on benchmark circuits. We demonstrate that, on average, using the global approach improves the timing delay by 10.1% or the amount of required decap by 35.5% compared to uniform decap allocation. The results

also show that the heuristic approach results in slightly lesser delay/ decap reduction with significantly smaller runtimes.

The remainder of the chapter is organized as follows. Section 4.2 presents the decap minimization using Lagrangian relaxation. Section 4.3 presents the proposed heuristic approach. Experimental results and conclusions are presented in Section 4.4 and Section 4.5, respectively.

4.2 Proposed global-optimization approach

A typical power supply network model of a chip consists of the ideal supply voltage sources, power and ground wires modeled as a linear RLC network, time varying current sources representing switching transistors and decoupling capacitances [71]. Each logic block in the design is then simulated at nominal supply voltage to obtain the current drawn by the blocks over time. Each block of non-linear devices is then replaced by its current distributed among its power grid supply points. For illustration, Figure 4.1 shows a combinational logic circuit in a power grid consisting of two metal layers. For clarity, only the

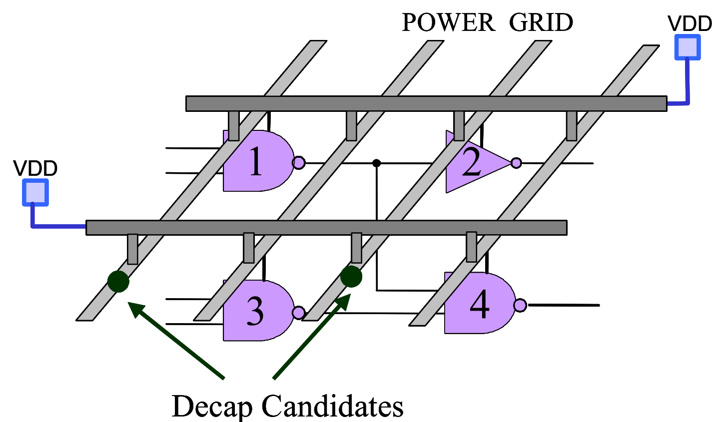


Figure 4.1. A combinational circuit in a power distribution network

Vdd distribution grid is shown. The ideal voltage sources, time-varying currents and decaps are not shown in the figure.

The objective of our optimization problem is to allocate decaps such that the circuit delay in each clock cycle is less than a pre-specified period, *T*. Conversely, the optimization can be formulated to minimize the decap allocation while meeting a specified delay constraint. The optimization variables are the decap sizes C_i attached to the power grid node *i*. We assume that the white space available for decap insertion is known. Furthermore, the proposed approach is a post-placement method and the placement information of gates is assumed to be available. Since the voltage variations in a power grid are typically very slow compared to the transition time of a switching gate [11], we make the simplifying assumption that the supply voltages are constant during the switching transition of a gate. The next subsection describes the problem formulation for global timing-aware decap allocation.

4.2.1 Gate-delay model

Consider a combinational circuit with *s* primary inputs (PIs), *t* primary outputs (POs) and *n* gates. The power and ground supplies at a gate *i* are denoted as Vdd_i and Vss_i respectively. Two fictitious nodes *src* and *sink* are added to the circuit. All the primary inputs are connected to the *src* node and all the primary outputs are joined together to form the *sink* node. The *sink* node connecting all the POs is labeled as node 0 and all other gates are numbered in the reverse topological order. Let the arrival time at the output of a gate *i* be denoted by a_i. Let *input*(*i*) be the set of indices of gates driving the inputs of gate *i* and let *output*(*i*) be the set of indices of gates in the fanout of gate *i*.

The delay of a gate i from one of its inputs j, D_{ji}, and output transition time, tr_{ji}, are represented as a linear function of the drops in supply voltages at the gate i and the driver gate driving the node j. This linear approximations are accurate for power supply variations within a range of ±10 %.

$$D_{ji} = D_{ji}^0 + k_{ji}\Delta Vdd_i + l_{ji}\Delta Vss_i + m_{ji}\Delta Vdd_j + n_{ji}\Delta Vss_j \quad \text{(EQ 4.1)}$$

$$tr_{ji} = tr_{ji}^0 + p_{ji}\Delta Vdd_i + q_{ji}\Delta Vss_i + r_{ji}\Delta Vdd_j + s_{ji}\Delta Vss_j \quad \text{(EQ 4.2)}$$

$\forall j \in input(i)$

where, D_{ji}^0 is the delay of gate i from input j under ideal supply voltages; tr_{ji}^0 is the transition time at the output of gate i under ideal supply voltages; ΔVdd_i and ΔVss_i are power voltage drop and ground bounce respectively at gate i; k, l, m, n, p, q, r and s are constants obtained by simulating the gate delay over a range of supply voltages and performing multi-variable linear regression.

Traditional standard-cell libraries are composed of two dimensional tables with table entries representing delay and transition times for different load-transition time combinations. The cell-library is re-characterized for different input slope-output load combinations to also incorporate the constants k through s along with the nominal delay and transition time entries, D_{ji}^0 and tr_{ji}^0. For a given set of transition time, load and supply voltages at a gate and its driver, the delay is computed appropriately using EQ 4.1 and EQ 4.2.

In the next section, we state the primal optimization problem (PP) for decap allocation under timing constraints, followed by the Lagrangian relaxation formulation. The discussions are focussed on the decap minimization problem with constraints on circuit delay.

The problem of circuit timing improvement with constraints on total decap can be solved in a similar manner.

4.2.2 Primal problem

The variables in the primal problem are the decap sizes C_is and arrival times at the output of gates, a_is. The objective of the primal problem (*PP*) is to minimize the total decap area:

$$minimize \quad \sum_{i=0}^{N-1} C_i \qquad \text{(EQ 4.3)}$$

where, N is the total number of decap candidate locations. The maximum decap sizes are bounded based on the white space available at each candidate decap-location:

$$Cmin_i \le C_i \le Cmax_i \qquad \text{(EQ 4.4)}$$

The constraints on arrival times, a_is, are stated as follows (For simplicity in explanation, but without loss of generality, we do not differentiate between rise and fall transitions):

$$a_j \le T \qquad \forall j \in input(0)$$

$$a_j + D_{ji} \le a_i \qquad \forall j \in input(i) \wedge (1 \le i \le n) \qquad \text{(EQ 4.5)}$$

where, T is the pre-specified delay requirement of the given circuit.

The delay of a gate i from one of its inputs j, D_{ji}, are expressed as a linear function of supply voltages as shown in EQ 4.1. The supply voltages, on the other hand, are a function of decap sizes, and are given by the following modified nodal analysis (MNA) relation explained in Chapter I:

$$\left[G + \frac{C}{h}\right]x[n] = i[n] + \frac{C}{h}x[n-1] \qquad \text{(EQ 4.6)}$$

where, x is the vector of unknowns: node voltages, inductor currents and currents from voltage sources; i is the vector of current and voltage sources; G and C are the conductance and capacitance matrices of the power grid; n is the simulation time; and h is the time-step for simulation.

The primal problem is difficult to solve in the current form because of the large number of unknowns and constraints in the problem. Also, it requires a prohibitively large number of circuit simulations. In the next subsection, we describe the use of Lagrangian relaxation to remove the constraints and present an efficient formulation that reduces the number of required simulations.

4.2.3 Lagrangian relaxation problem

Lagrangian relaxation is a standard technique to eliminate difficult constraints in the primal problem [6]. A non-negative Lagrange multiplier, λ_{ji}, is associated with each input-output pair (j,i) for gate $i,$ and the corresponding constraint is incorporated into the objective function. For a given set of Lagrange multipliers, the problem in Section 4.2.2 can be expressed as the following Lagrangian relaxation problem:

$$\text{minimize} \sum_{i=0}^{N-1} C_i + \sum_{j \in input(0)} \lambda_{j0}(a_j - T) + \sum_{i=1}^{n} \sum_{j \in input(i)} \lambda_{ji}(a_j + D_{ji} - a_i) \qquad \text{(EQ 4.7)}$$

subject to:

$$Cmin_i \le C_i \le Cmax_i$$

$$D_{ji} = D_{ji}^0 + k_{ji}\Delta Vdd_i + l_{ji}\Delta Vss_i + m_{ji}\Delta Vdd_j + n_{ji}\Delta Vss_j \qquad \forall j \in input(i)$$

$$\left[G + \frac{C}{h}\right]x[n] = i[n] + \frac{C}{h}x[n-1]$$

For a given set of Lagrange multipliers λ, the above problem has two sets of variables: arrival times, *a* and decap sizes *C*. Kuhn-Tucker conditions [6] state that if (a^*, C^*) is at the optimal solution to the above problem, then the derivative of the objective function with respect to all the variables must be zero:

$$\left.\frac{\partial}{\partial a_i} obj\right|_{a = a^*,\, C = C^*} = 0$$

and hence,
$$\sum_{j \in input(i)} \lambda_{ji} = \sum_{k \in output(i)} \lambda_{ik} \qquad \forall i \qquad \text{(EQ 4.8)}$$

This condition states that at the optimal solution, the sum of Lagrange multipliers from the inputs of a node *i* is equal to the sum of Lagrange multipliers emanating from node *i* to all its fanout gates. Let Ω_λ be the set of Lagrange multipliers satisfying the above condition in EQ 4.8. Thus, if we search the Lagrange multipliers only in the set Ω_λ, we can eliminate arrival time variables a_i from the objective function. The simplified objective function is expressed below:

obj:
$$\sum_{i=0}^{N-1} C_i - \sum_{j \in input(0)} \lambda_{j0} T + \sum_{i=1}^{n} \sum_{j \in input(i)} \lambda_{ji} D_{ji} \qquad \text{(EQ 4.9)}$$

Substituting the expression for D_{ji} from EQ 4.1 into EQ 4.9, we get the objective function as follows:

$$\sum_{i=0}^{N-1} C_i - \sum_{j \in input(0)} \lambda_{j0} T + \sum_{i=1}^{n} \sum_{j \in input(i)} \lambda_{ji} D_{ji}^0 + \sum_{i=1}^{n} \alpha_i \Delta V dd_i + \sum_{i=1}^{n} \beta_i \Delta V ss_i \qquad \text{(EQ 4.10)}$$

where,
$$\alpha_i = \sum_{j \in input(i)} \lambda_{ji} k_{ji} + \sum_{k \in output(i)} \lambda_{ik} m_{ik} \qquad \text{(EQ 4.11)}$$

and
$$\beta_i = \sum_{j \in input(i)} \lambda_{ji} l_{ji} + \sum_{k \in output(i)} \lambda_{ik} n_{ik} \qquad \text{(EQ 4.12)}$$

Thus, given the optimal value of λ, we can solve the simplified Lagrangian relaxation problem to arrive at the optimal solution. With the simplified objective function in EQ 4.10, removal of arrival time variables and arrival time constraints, the Lagrangian problem is much easier to solve as compared to the Primal Problem. The next subsection describes the solution of the Lagrangian problem for a given set of Lagrange multipliers.

4.2.4 Solving the Lagrangian relaxation

To solve the Lagrangian subproblem described in the previous subsection, we need to compute the sensitivities of the objective function expressed in EQ 4.10 to decaps C_i. Adjoint sensitivity [21] is the preferred method when sensitivity of one measurement response to multiple parameters is required. On the other hand, the direct sensitivity method [60] is efficient in computing the sensitivity of multiple measurement responses to a single parameter. In our case, as shown in objective function in EQ 4.10, the sensitivity of worst supply voltage values at each gate, to changes in all the decap sizes is required. Thus, using either the adjoint or the direct method in its standard form will require multiple power grid simulations. For instance, using the direct sensitivity method would require N transient power simulations, where N is the number of decaps. On the other hand, the adjoint sensitivity method would require one original power grid simulation and n adjoint network simulations, where n is the number of gates in the combinational circuit. We therefore use a modified adjoint sensitivity method which uses the principle of superposition and computes the sensitivities of multiple measurement variables to multiple circuit parameters in only one adjoint simulation. Since, the objective function in EQ 4.10 is a linear function of voltage drops at all the gate locations, the principle of superposition

holds and the modified adjoint sensitivity method can be used to solve the Lagrangian relaxation problem.

Sensitivity computation

We first describe the sensitivity computation of worst voltage drop, ΔVdd_i, at a gate i. The power grid is simulated with a given current profile and the derivative of the voltage waveform across all the capacitors, $\dot{V}_i(t)$, is stored. The time point of the occurrence of the worst voltage drop at the supply node of gate i is observed. Let τ_i denote the time of occurrence of the worst drop at gate i. Then, the adjoint power grid network is constructed with all the voltage sources shorted to ground and all the current sources removed from the network (refer to Figure 4.2). The adjoint network is excited backwards in time with unit delta, $\delta(t-\tau_i)$ current waveform applied at supply node of gate i. The voltage waveform at all the supply nodes of all the decaps, denoted by $\Psi_i(t)$ is observed. Lastly, the convolutions between $\dot{V}_i(t)$ and $\Psi_i(t)$ provide the sensitivity of worst case voltage drop at the gate i to all the decap sizes. Since there are n gates in the design, this method will require n different adjoint grid simulations, one for each gate, to compute the sensitivity of the voltage drops at all the gates.

However, the objective function of the Lagrangian relaxation subproblem in EQ 4.10 consists of a linear combination of voltage drops ΔVdd_i and ΔVss_i, weighted by constants α_i and β_i. As mentioned above, the principle of superposition is therefore used and all the current excitations are applied simultaneously to the adjoint circuit. Let $[\tau]$ be an nx1 vector (n is the total number of gates), representing the time of occurrence of worst drop at all

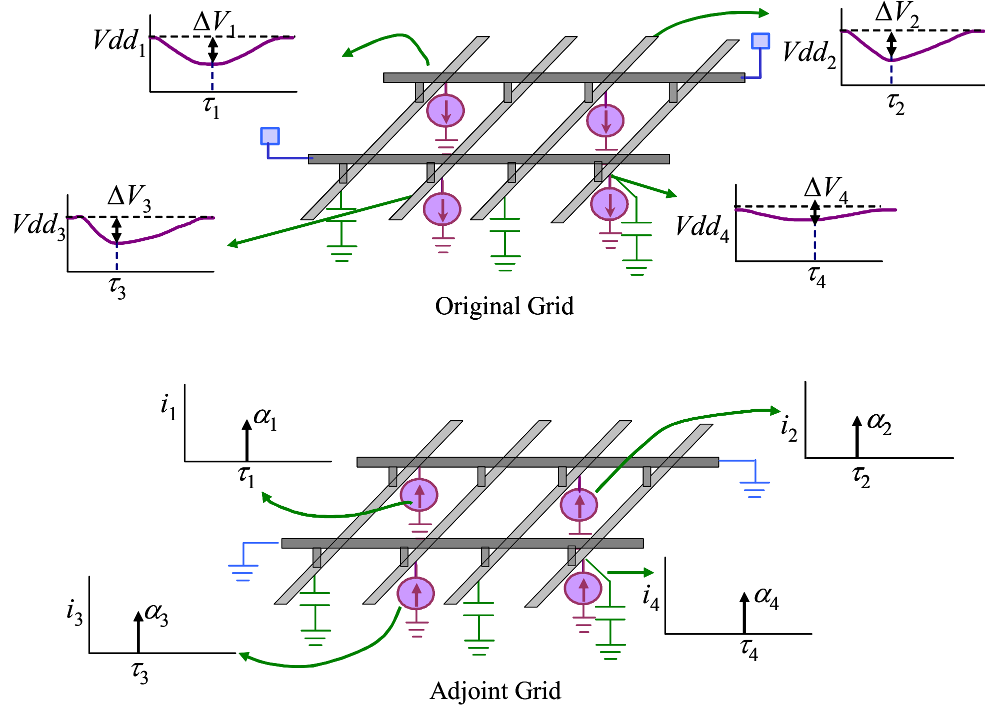


Figure 4.2. Illustration of gradient computation using the modified adjoint-sensitivity method

the gates. In the adjoint network, for every gate *i*, a current source, represented by a scaled delta function, $\alpha_i\delta(t-\tau_i)$ for power grid ($\beta_i\delta(t-\tau_i)$ for ground grid) is applied and voltages at the decaps $\Psi(t)$ are observed. Then, the convolutions between the derivative of voltages across the decaps in the original grid and voltages across the decaps in the adjoint grid provide the sensitivities. Thus, the sensitivity of EQ 4.10 to C_i is given as follows:

$$\frac{\partial}{\partial C_i}obj = 1 + \int_0^T \Psi_i(T-t) \cdot \dot{V}_i(t)dt \qquad \text{(EQ 4.13)}$$

where, T is the duration of original grid simulation; i is the power grid node where decap C_i is connected; $\Psi_i(t)$ is the voltage waveform at node i in the adjoint network and $V_i(t)$ is the voltage waveform at node i in the original network.

The gradients of the objective function to decap values obtained in this manner can be used to solve the Lagrangian relaxation problem using the conjugate gradient method [6].

4.2.5 Finding the optimal λ

The previous subsection described the modified adjoint method to find the optimal solution given a set of Lagrangian multipliers λ. In this subsection, we present a method for obtaining the optimal set of λs based on timing slacks available in the circuit. The Lagrange multiplier λ_{ji} intuitively represents the criticality of the delay of gate i from its input j. If the circuit consists of only one prominent critical path under supply variations, then at the optimal solution point, only the Lagrange multipliers along the critical path will remain non-zero, while all other λs associated with the non-critical paths will be zero. Thus, we propose to update the set of Lagrange multipliers based on the timing slack available at each node.

At the start of the algorithm, we assume that every path is critical and as an initial guess, all the Lagrange multipliers are non-zero and satisfy condition EQ 4.8. Then the Lagrangian relaxation subproblem is solved optimally as described in Section 4.2.4 based on the given set of Lagrange multipliers. In the next iteration, timing slack is computed at each node in the circuit based on arrival times (AT) and required arrival times (RAT). Each Lagrange multiplier λ_{ji} is decreased in proportion to the slack available at the output of gate i as follows:

$$\lambda_{ji}^{k+1} = \lambda_{ji}^{k} + \rho_k \cdot s_i^k \qquad \forall(1 \le i \le n), j \in input(i) \tag{5}$$

where, λ_{ji}^k and s_i^k is the value of Lagrange multiplier in iteration k; s_i^k is the slack at the output of gate j in iteration k; ρ_k is the step-size in iteration k.

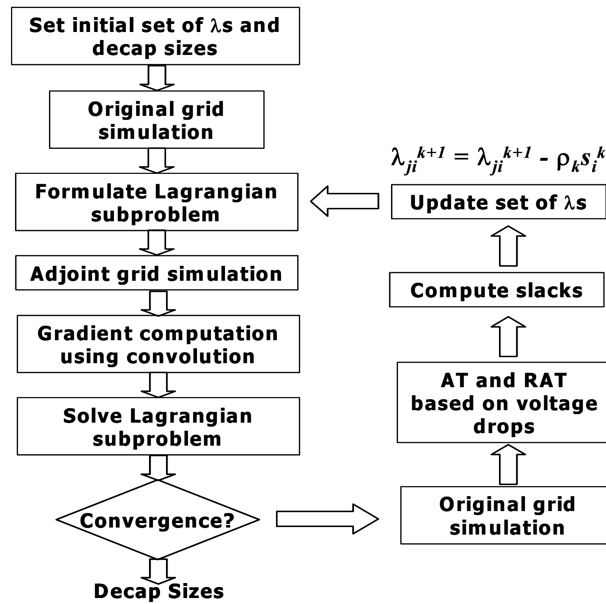


Figure 4.3. Overall global-optimization flow

4.2.6 Overall optimization flow

The overall optimization flow for the algorithm is presented in Figure 4.3. We start with all Lagrange multipliers as non-zero such that the condition in EQ 4.8 is satisfied. Then, the Lagrangian relaxation problem is formulated and values of voltage drop coefficients α_i and β_i are obtained. The voltage drop coefficients are used to provide current excitations to the adjoint network. Convolution is performed between the original and adjoint network waveforms to obtain the sensitivities of the objective in the Lagrangian subproblem. Using these sensitivities, the Lagrangian subproblem is solved optimally for the given set of Lagrange multipliers. Lagrange multipliers are then updated based on the slack available at gates in the combinational circuit and the procedure is repeated until the convergence of Lagrange multipliers.

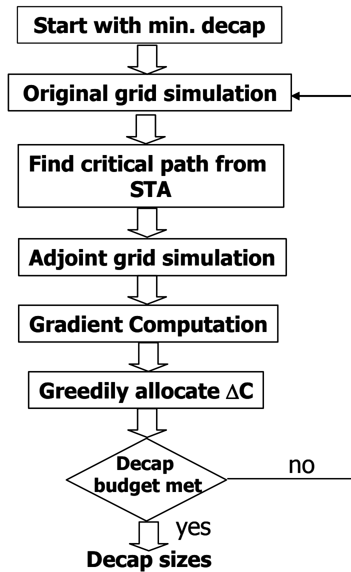


Figure 4.4. Optimization flow of path-based greedy algorithm

4.3 Greedy path-based approach

The Lagrangian subproblem described in the prior subsections involves updating the Lagrange multipliers in each major iteration and solving the Lagrange relaxation subproblem within each such iteration. In this subsection, we present a path based greedy heuristic to reduce the number of iterations. This approach is fast and has been found to give a favorable trade-off in terms of run time and optimization quality compared to the optimal results.

The optimization flow for the greedy approach is illustrated in Figure 4.4. At the start of the optimization, all the decaps are assigned a small value. Power grid is simulated with these decap values to obtain the worst voltage drops at all the gates in the circuit. Static timing analysis is performed on the circuit with the voltage drops to obtain the path with maximum delay. In the Lagrangian relaxation based algorithm, this amounts to setting all the Lagrange multipliers to 1 in the critical path and 0 in the non-critical paths. Path delay

is formulated as a linear function of voltage drops using EQ 4.1. Gradient of the path delay to all the decaps is then computed using the results from original and adjoint grid simulations. Based on the gradients, a small amount of decap ΔC is allocated in the best search direction. The new allocated capacitance values are stamped in the MNA coefficient matrix (G+C/h) and original grid is re-simulated to obtain the new worst voltage drops at the gates. STA is again performed to obtain the critical path in the circuit which may have changed due to decap insertion in the previous iteration. The process is repeated until the decap budget has been exhausted. In every iteration, only a small amount of decap is available to the optimizer for allocation.

4.4 Experimental results

The proposed Lagrangian relaxation-based and heuristic decap allocation algorithms were implemented in C++ and tested on ISCAS85 benchmark circuits with power grids of different sizes. For the experiments, two power grids, Grid1 and Grid2, consisting of 4 metal layers were constructed using pitches and widths of an industrial microprocessor in 0.13μ technology. Grid1 is a M2-M5 power grid of 700μx700μ die-area consisting of 10,804 nodes, 17,468 elements, 12 *Vdd* C4s and 12 *Vss* C4s. Grid2 is a M2-M5 1.2mm X 1.2mm grid with 17,530 nodes, 29,746 elements, 28 *Vdd* C4s and 28 *Vss* C4s. To model package impedance, an inductance of 1nH was connected in series with a resistance of 0.1mΩ at each C4 location in the grid. The benchmark circuits were synthesized using 0.13μ standard-cell library and placed using Cadence Silicon Ensemble. Some of the white space available was chosen to be the candidate for decap allocation and the decap in

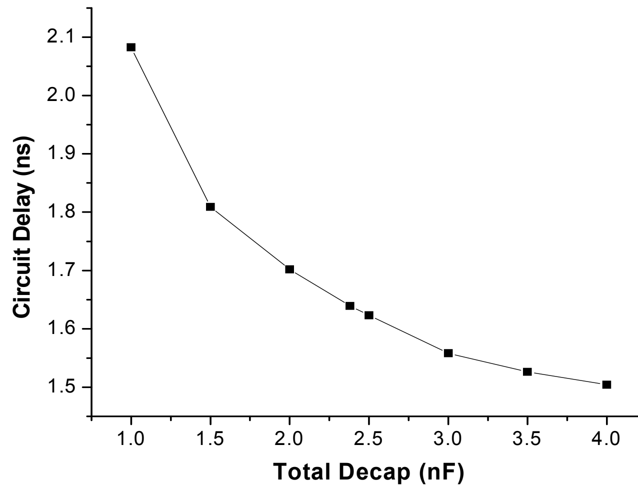


Figure 4.5. Decap area vs circuit-delay trade-off curve for c432

each candidate location was modeled as a capacitor connected to the power grid nodes over the region. For the gates in the modeled power grid region, the peak currents were approximated using a triangular waveform of 1ns duration. The peak currents were scaled to cause an appreciable (~15%) voltage drop in the grid. Preconditioned conjugate gradient method was applied for optimization, using the LANCELOT non-linear solver [22].

Figure 4.5 shows the delay-decap trade-off curve for circuit c432 placed in test-grid Grid1. The dotted line represents the nominal delay of the circuit under ideal supply voltages. The figure shows the effectiveness of the proposed timing-aware approach in improving the circuit delay. Table 4.1 presents the improvement in circuit delay for a fixed decap budget. Columns 1 and 2 show the circuit name and its power grid. Columns 3 and 4 show the number of gates in the circuit and number of candidate decap locations in the layout. The total decap budget is shown in column 5. Columns 6, 7, 8 and 9 show the circuit delay under ideal voltage supplies, delay with uniform decap distribution, delay obtained after Lagrangian optimization and delay after application of the heuristic, respec-

grid	ckt	# gates	# decaps	decap budget	circuit delay				% delay redn	runtimes	
					nom.	uniform	global optim.	greedy optim.		global optim.	greedy optim.
grid1	c432	212	476	2.38n	1.498ns	1.798ns	1.621ns	1.640ns	9.84%	11m15s	1m15s
grid1	c499	553	595	2.98n	1.233ns	1.480ns	1.308ns	1.394ns	11.62%	9m41s	1m57s
grid1	c1355	654	793	3.97n	1.839ns	2.207ns	1.878ns	1.913ns	14.90%	11m43s	2m58s
grid1	c1908	543	579	2.89n	2.088ns	2.506ns	2.251ns	2.256ns	10.17%	20m24s	25s
grid1	c2670	1043	1190	3.57n	1.622ns	1.946ns	1.754ns	1.764ns	9.86%	52m33s	8m41s
grid2	c3540	1492	1559	7.79n	2.301ns	2.761ns	2.498ns	2.564ns	9.52%	109m59s	23m49s
grid2	c5315	2002	2217	6.65n	2.080ns	2.769ns	2.409ns	2.416ns	9.97%	221m24s	61m18s
grid2	c6288	3595	3712	8.15n	5.186ns	6.223ns	-	5.820ns	6.48%	>4hrs	188m36s
grid2	c7552	2360	2571	7.18n	2.975ns	3.571ns	-	3.262ns	8.65%	>4hrs	63m03s

Table 4.1. Experimental results showing delay reduction for a given decap budget

tively. It should be noted that the uniform decap allocation, Lagrangian optimization and greedy heuristic utilize the same total decap budget. The greedy heuristic-based approach gave near-optimal results for all the circuits. Column 10 shows the percentage improvement in circuit delay by Lagrangian optimization over uniform decap allocation. On average, the circuit delay was observed to improve by 10.11% for the given decap budget. The last two columns show the runtimes of the Lagrangian-based optimization and the greedy heuristic. The Lagrangian-based optimization, although optimal, had considerably larger runtimes than the greedy algorithm. The Lagrangian-based optimization could not converge within 4 hours of runtime for two circuits. Runtimes can be improved by using a better non-linear solver and by accelerating the transient power grid simulations using moment-based solvers [18].

Table 4.2 presents the comparison of total decap budget required to meet a pre-specified timing budget using the heuristic method. Column 3 states the circuit delay with ideal supply. The constraint on circuit delay under power supply fluctuations is shown in column 4. Columns 5 and 6 show the amount of total decap required to meet the given timing con-

grid	ckt	nom. delay	delay constr.	decap allocated		% decap redn
				uniform.	greedy optim.	
grid1	c432	1.498ns	1.640ns	3.55nF	2.38nF	32.98%
grid1	c499	1.233ns	1.394ns	3.49nF	2.98nF	17.60%
grid1	c1355	1.839ns	1.913ns	6.65nF	3.97nF	40.33%
grid1	c1908	2.088ns	2.256ns	6.15nF	2.89nF	52.92%
grid1	c2670	1.622ns	1.764ns	6.96nF	3.57nF	95.80%
grid2	c3540	2.301ns	2.564ns	10.04nF	7.80nF	22.37%
grid2	c5315	2.080ns	2.416ns	12.20nF	6.65nF	45.56%
grid2	c6288	5.186ns	5.820ns	9.74nF	8.15nF	16.31%
grid2	c7552	2.975ns	3.262ns	13.26nF	7.18nF	45.85%

Table 4.2. Experimental results showing reduction in decap area for specified timing constraint

straint under uniform distribution and allocated using the proposed heuristic approach. Column 6 shows the percentage reduction in total decap area compared to uniform decap distribution, which is 35.51% on an average.

4.5 Conclusions

In this chapter, we proposed an approach for timing aware decoupling capacitance allocation which utilizes the timing slacks available at the gates in a design. The decoupling capacitance allocation is formulated as a non-linear optimization problem and Lagrange relaxation in conjunction with the modified adjoint method is used for optimization. We also presented a fast and near-optimal greedy heuristic for timing-aware decap allocation. The approach has been implemented and tested on ISCAS85 benchmark circuits and power grids of different sizes. Compared to uniformly allocated decaps, the proposed approach utilizes 35.51% less total decap to meet the same delay target. For the same total decap budget, the proposed approach is shown to improve the circuit delay by 10.11%.

CHAPTER V

INDUCTANCE, LOCALITY AND RESONANCE IN POWER SUPPLY NETWORKS

5.1 Introduction

In a power distribution network (PDN), rapid transient currents flow through the transistors onto the power grid, charging an discharging various capacitances, then flow onto the package through the C4 bumps, and eventually make their way to the voltage regulator module (VRM). This flow of currents causes spatial and temporal voltage variation in the PDN, degrading circuit performance and reliability. PDN modeling and verification is challenging due to the presence of decaps, on-die inductance, various resonance effects, and simply the enormous size of the PDN. The following issues must be resolved for an accurate supply-drop analysis:

1) How significant is the impact of on-die inductance?

2) How localized are the currents as they flow outward from a device?

3) Does the decap charge respond locally or globally to supply drops?

4) Do resonance effects occur, and if so, how?

The answers to these questions are critical to addressing the type of models and CAD algorithms required to deal with the PDN verification and chip-package codesign. For

example, if supply-drop effects are localized, then it is possible to considerably simplify the analysis by verifying several partitions of the PDN in parallel, as proposed in [19]. Researchers have described investigations of some of these effects on large-scale industrial designs [15][27][52]. Previous work, however, has not comprehensively spanned the entire range of modeling parameters, from detailed PDN modeling to full-die simulation, including a package model and non-uniform decoupling capacitor (decap) distribution. To the best of our knowledge, ours is the first comprehensive simulation study of an entire industrial processor, covering in detail, these modeling and analysis issues.

In this chapter, we concentrate our study on the core region; we do not cover the I/O region. We electrically model a full-core die in the highest level of detail possible within computational-power constraints, and then justify the model from bottom up. This entails beginning with a full-wave model for a small section of the die area and progressing in steps to a full-die and package co-simulation model containing all the essential elements required to attain the desired accuracy level. Simulations of the package-die model at every step highlight the critical and non-critical elements constituting the model. We ignore the non-critical elements, which do not significantly affect the simulation accuracy, to incorporate a larger die area at the next abstraction level. This enables the analysis of a larger region of the die for the same simulation time.

Using these models, we demonstrate the following for an Intel Pentium® class microprocessor, designed in a 90nm technology: First, popular 2D inductive models, often used to model on-die inductance [85], overestimate the impact of on-die inductance on supply noise. High-frequency (> 5GHz) current transients, which can excite on-die inductive effects, are comparatively smaller, highly localized (with a radius of a few microns) to the

switching device and transient in nature (decaying quickly). The on-die power grid behaves otherwise as an RC network. Therefore, we can ignore the on-die inductance for important frequencies and scales, considerably simplifying modeling and analysis. Second, the package has a significant impact on the accuracy of on-die power-grid analysis. This necessitates including an accurate package model for a CAD approach, targeting transient PDN analysis and optimization. Third, decoupling capacitors act both globally (full-chip) and locally, depending on the frequency of excitation currents. They act globally at the main resonance frequency because of their interaction with package inductance (low frequency of about 50 MHz to 200 MHz). But the impact of decaps becomes increasing more localized at frequencies higher than this resonance frequency. This observation is important in CAD for placement, sizing and optimization of decoupling capacitors. Fourth, localized (about a 1,000 micron radius), mid-frequency (about 1 GHz to 2 GHz) resonance effects are possible due to the resistive isolation of pockets of capacitors interacting with localized C4 and package inductors, and these pockets collectively act as several mini-dies.

We begin with an explanation of our smallest but most detailed model, progress to a 2mm X 2mm model and finally, describe our full-chip microprocessor model. We conclude with a coherent explanation of the various highlighted effects and their interaction and the implications on CAD modeling and optimization.

5.2 Full-wave on-die inductive effects

The PDN is usually modeled as a linear network consisting of RLC elements, nominal voltage sources, and independent time-varying currents representing the switching cur-

rents of on-die devices. Computational constraints make it infeasible to model and analyze the entire PDN in complete detail. Hence, we began with a fully detailed but smaller (500 micron x 500 micron) section of the die area, consisting of metal layers M4 through M7. We used this model to observe high-frequency inductive effects and their locality. For this purpose, we began with a full-wave modeling method known as partial-element equivalent circuit (PEEC) [64], which has been used extensively in package-level analysis. Using the grid dimensions for each layer, we broke up the grid description into detailed via-to-via metal segments, including vias for all layers. The PEEC method, as shown in Figure 5.1, models every metal segment with its self-resistance, self-inductance, and capacitance to ground, as well as its capacitive and inductive coupling to every other metal segment. This results in a dense, full-wave electromagnetic model that is highly accurate but extremely CPU and memory intensive.

The PEEC capacitors, which model the dielectric and metal charge interaction effects, serve only to dampen the inductive ringing, and hence, can be removed in order to highlight any inductive effects. The PEEC model consisted of 67,150 electrical nodes, 84,470 R and L elements, and 12 million mutual inductors. In addition, we assumed that total intrinsic and extrinsic decoupling capacitance was a low value of 10 pF for that area. We

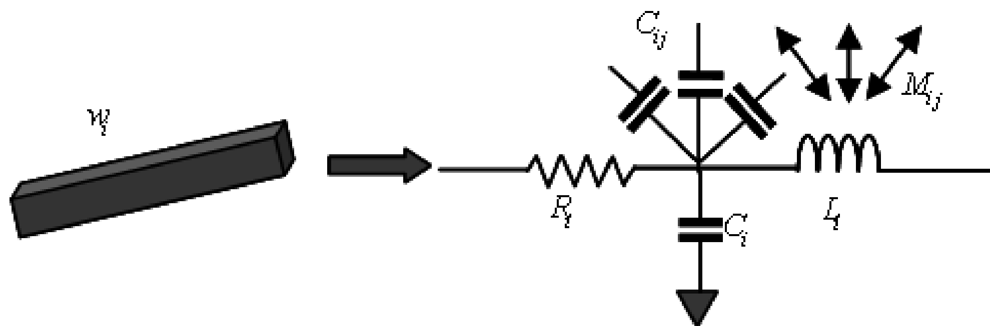


Figure 5.1. Illustration of the PEEC model of a wire segment

attached the total decoupling capacitance in a uniformly distributed manner to the lowest metal layer as 5,512 individual capacitors. This low value of decoupling capacitance is useful for highlighting potential inductive effects. The simulated area contained 13 C4 bumps, each modeled as a series RL element of 0.01 ohm and 0.325 nH, representing both the C4 and package input impedance. This is, in fact, a low inductance value per C4 for the package input. But, again, we used this low value to highlight any on-die inductive effects. We attached a source of variable rise times (10ps to 100 ps) to the lowest metal layer. The model also made the following simplifications:

- It avoids discretizing the wires for skin effects
- It does not model signals in the neighborhood of the grid (which have the effect of absorbing inductive noise)
- It injects current directly into M4 rather than model M1-M3

All of the above assumptions effectively increase inductive effects and are a worst case scenario in order to detect them.

5.2.1 PEEC simulation results

The 3D PEEC simulation results were compared to a 2D model which modeled every layer separately in 2D, discretized the per-unit-length values to via-to-via segments, and then stitched the 2D layers together using resistive vias. The simulation results were compared for an R model (resistive only PEEC grid, with de-caps attached to M4), an RL model (R model with self inductance) and an RLM model (RL model with mutual inductors), all of which had R-L C4 models attached to M7. The simulation results for PEEC 3D are given in Figure 5.2, left. It is clear that the PEEC model, in spite of all the assump-

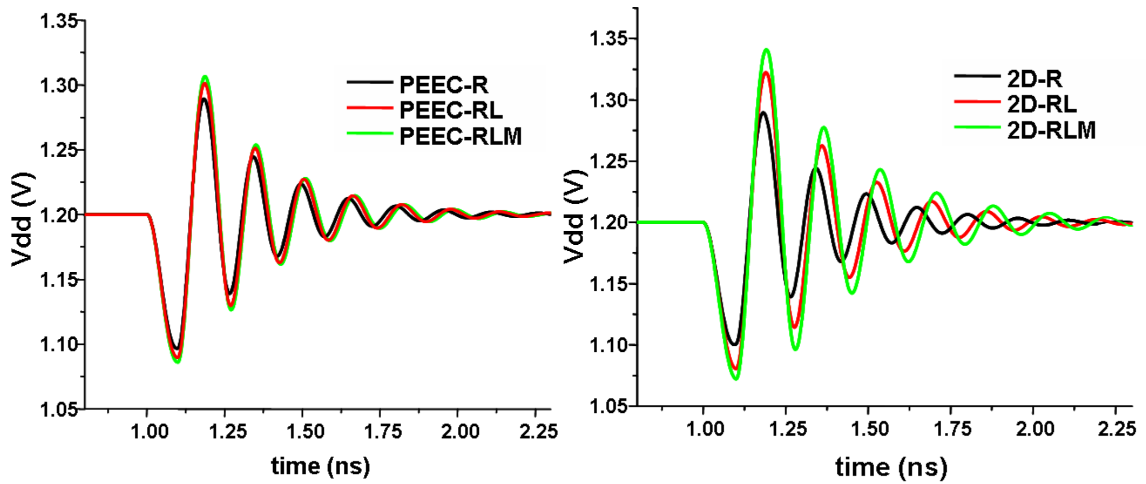


Figure 5.2. A 3D PEEC simulation of 500mx500m grid (left) compared to a 2D-modeling approach (right)

tions intended to highlight inductive effects, shows little impact of inductance. However, we note that the 2D model (Figure 5.2, right), shows significant inductive differences. Unfortunately, the 2D models have been extensively used to model inductance on power grids, yielding subsequent tenuous conclusions based on those models. The same study above was performed for various uniform or non-uniform sources, different rise times (down to 10ps) and for increased C4 L values (conforming to actual values) or device capacitance values (conforming to actual de-cap densities) and the results were found to be similar.

The only difference observed in the full-wave RLC model was for very fast transients of 10ps, which caused an initial high frequency localized (a few micron radius) transient "blip" which quickly degenerated into a wave fully described by an RC grid model. Given that we overestimated the inductance, even this small localized inductive effect should be smaller than what we observed, if all of the details of the localized model were in place.

Figure 5.3 shows the voltage map generated when a large driver, switching at 10ps, draws high frequency current from the power grid.

5.2.2 High-frequency supply noise

It is important to note that the PEEC model describes all potential inductive interactions for the full dimensions of the model. However, remote potential interactions do not determine the return path and the high frequency currents (>5GHz) tend to remain extremely localized (Figure 5.3). If the model were to be extended to a larger area, the locality of the high frequency would not change. This result may be explained mainly due to these reasons:

- There are frequent power rail vias in a microprocessor PDN, providing frequent return pathways.
- The grid is loaded with wire resistance, device and wire capacitance. These help to dampen the transient response.

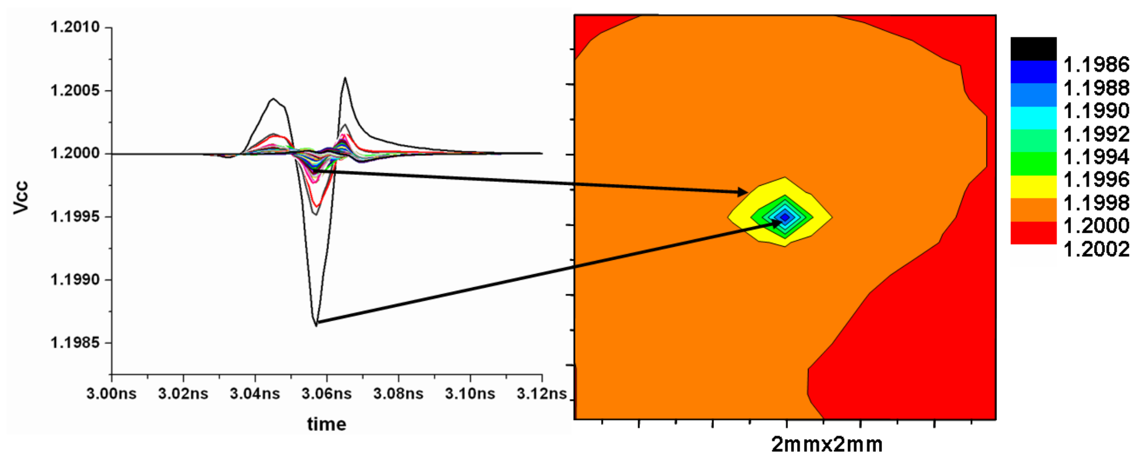


Figure 5.3. Voltage map generated due to high-frequency current transients

Although the high frequency response is highly localized, the midlow frequency currents (1-2 GHz) dissipate outward from the source to affect several gates. Therefore, when there are multiple switching sources of current, the high frequency transients of each gate will only have local impact around the gate while the mid-low frequency transients will have additive impact at every neighboring gate, overwhelming each gate's localized high frequency effect in amplitude. This is another significant reason for not requiring on-die inductance in a power grid model. The model necessary for understanding the full-die requires only a resistive grid with device capacitance and a C4/package model.

5.3 Mid-size model and capacitive effects

Given the conclusions of the previous study, we eliminated on-die inductance in our larger models, used an R-only model for the grid with device caps, and extended our detailed via-to-via metal segment model to 2mm X 2mm, consisting of metal layers M2 to M7. This allowed us to determine a larger area of interaction and to understand the properties of this larger grid in order to build a full-chip model. We attached this die model to an RLC package model which modeled the package from the die shadow (the package area where the die is placed) to the VRM, and which was discretized to 9x9 in the die shadow area as shown in Figure 5.4. Our segment of the grid was only big enough to cover a 2x2 section of that area. In the middle of our grid at M2 we placed a single frequency domain current source in order to observe its impact on the surrounding droop. What to do with the other discrete die shadow pins was an open question and we tried two cases: (I) attach all the die capacitance under the 2mm X 2mm die model or (II) distribute it evenly in the

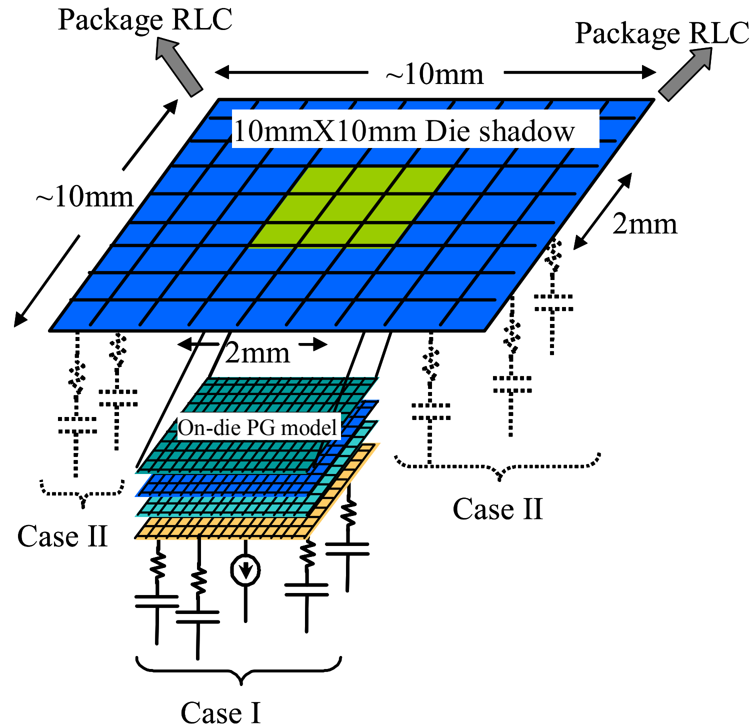


Figure 5.4. A 2mm X 2mm section of on-die grid attached to the middle of package shadow

9x9 die shadow, with 2x2 of the sections placed under the attached die model, and the rest directly attached to the package/die interface pins (Figure 5.4). From the package perspective, a resonance frequency of ~200MHz was expected, which was verified through package simulations and also validated in silicon. This was seen for case (I) as showin in Figure 5.5(left). However, when the die capacitance was distributed outside of the 2mm X 2mm section, another spurious frequency was observed, that of 60 MHz (Figure 5.5(right)). We deduced that this spurious resonance was due to the high impedance path from the caps outside of the 2mm X 2mm section to those inside the section, since all remote de-cap currents had to travel through the package without an on-die connection. This showed that a full-die grid resistance model is essential for a correct global die behavior.

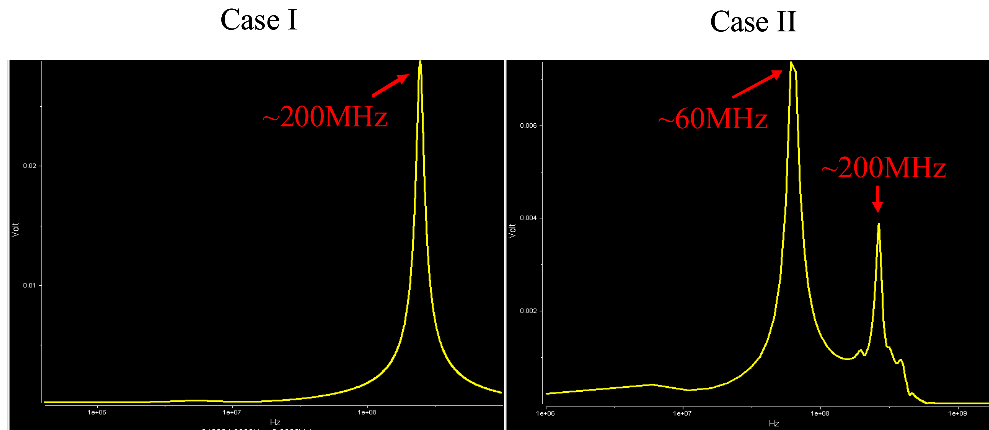


Figure 5.5. Supply drop frequency response showing resonance frequencies for Case I and Case II illustrated in Table 5.4

We illustrate this principle more clearly in the following simple example. We simulated the same 2mm X 2mm grid section but with four individual capacitors placed in the middle of the four 1mm X 1mm quadrants, with values of 1nF, 2nF, 0.5nF and 4nF (Figure 5.6, top). We attached RL models attached to the C4 pins with values equal to the input impedance of the rest of the package. When we probed the frequency response of the volt-

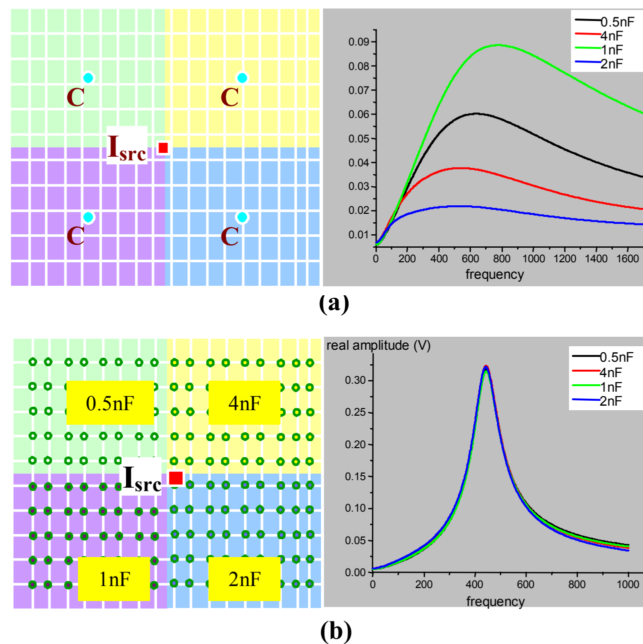


Figure 5.6. A 2mm X 2mm section of the grid with lumped (a) and distributed decoupling capacitors (b)

age over the caps, using one current source in the middle of the grid, we observed 4 distinct resonant frequencies. However, when we spread the cap values randomly around the 4 quadrants (while maintaining the same total amount), we observed only a single resonance frequency (Figure 5.6, bottom). This implied that resistive isolation between capacitive regions, together with the limited number of C4 inductors above the regions, caused them to act as distinct mid-frequency resonant circuits (four mini die, in some sense). This implied a principle of locality which is explained in the next section. The fact that one could have isolated pockets of mid-frequency (greater than the die-package resonance) was an important new effect that was exposed by this analysis.

5.3.1 Model reduction

In order to progress to a full-die model that fit into memory, it was important to reduce the resistive grid from the level of detail contained in the 2mm X 2mm model. Using the same level of accuracy was not feasible for a full die. However, we needed to determine how much to reduce the grid and still maintain the accuracy of the detailed effects we wanted to observe, especially with respect to resonance. We applied the prior-proposed multi-grid method for this purpose [44]. We used this method to reduce our 2mm X 2mm

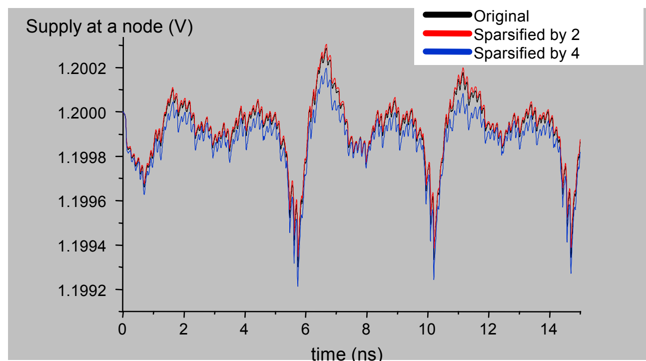


Figure 5.7. Voltage response of a 2mm X 2mm with 2x- and 4x-reduced grids for identical current excitation

2mm X 2mm model	# nodes	# elements	Runtime (s)	Peak mem usage
original	877259	1249250	1453.75	2.2G
2X reduction	222475	323874	797.83	650M
4X reduction	57861	85110	140.46	300M

Table 5.1. Run-time and peak memory-usage of the 2mm X 2mm model before and after multi-grid-based sparsification

model by a factor of 2 and then by 4 in order to determine the accuracy of the resultant models. When we placed unit transient currents on the original grids and the reduced grids we obtained the results in Figure 5.7. A 4X reduction allowed entire grid to fit in memory and not suffer much accuracy loss. Table 5.1 lists the run-time and peak memory usage of the original and sparsified models.

5.3.2 Locality in power grids

Flip-chip power grids in DC have been shown to have property of locality, in which the voltage droop from a single current source stays in the proximity of that source due to the C4 sources [19]. However, it was not clear what this locality principle meant for a package-die PDN model in the time and frequency domain. For this purpose, a 4X reduced M2-M7 resistive grid was attached to an RL C4/package model and with a uniform cap distribution on M2. A single frequency source was placed in the middle. All the voltage nodes on M2 were probed in the frequency domain and simulated across DC to mid-frequencies (~1-2 GHz). The results are shown in the information-rich graph in Figure 5.8. Each curve represents the frequency response of one node, on M2, to a single source in the middle of the grid. On the DC (left) side, there is clear locality because there is a decreasing response of nodes as one moves away from the source (downward movement on the

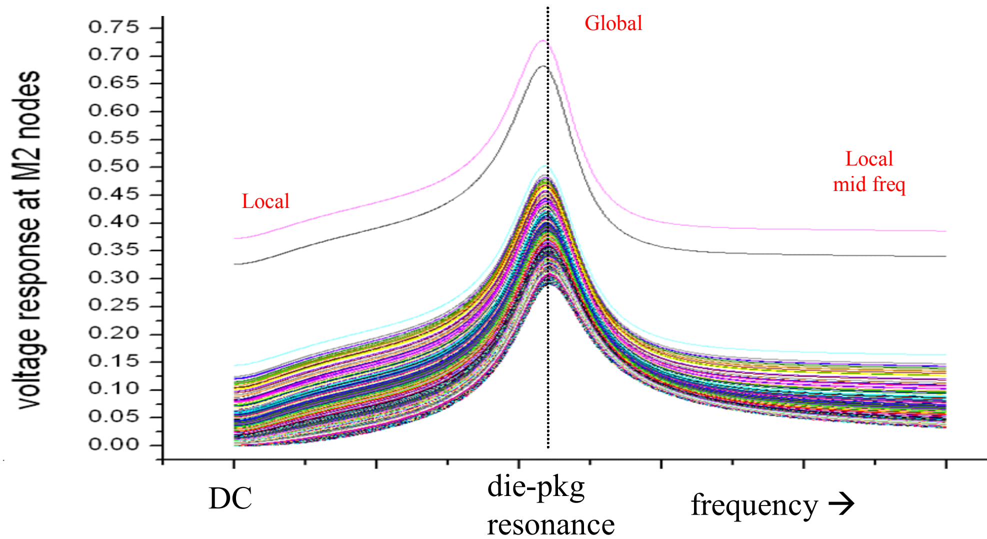


Figure 5.8. Frequency response at all the M2 supply nodes illustrating locality as function of excitation frequency

X-axis) until there is a zero response. On the mid-frequency right side, there is a quasi-locality as the response gets smaller with distance but never goes to zero, indicating that some diminishing capacitor currents are always supplied at a distance. At the main low frequency package/die resonance in the middle, all locality effect is lost. This indicates that at the main resonant frequency, both the die and package are acting as one and charge is flowing everywhere on the die. However, at other frequencies, the caps and de-caps tend to act in a local manner.

5.4 Complete package-die model

The most realistic full-die model was constructed using the 4x reduced M2-M7 full-die grid, the package model and a realistic nonuniform decap distribution. We reduced the package model to a series impedance connected at each C4 location. Further, we took the actual non-uniform full-die decap distribution as shown in Figure 5.9) and placed it on the

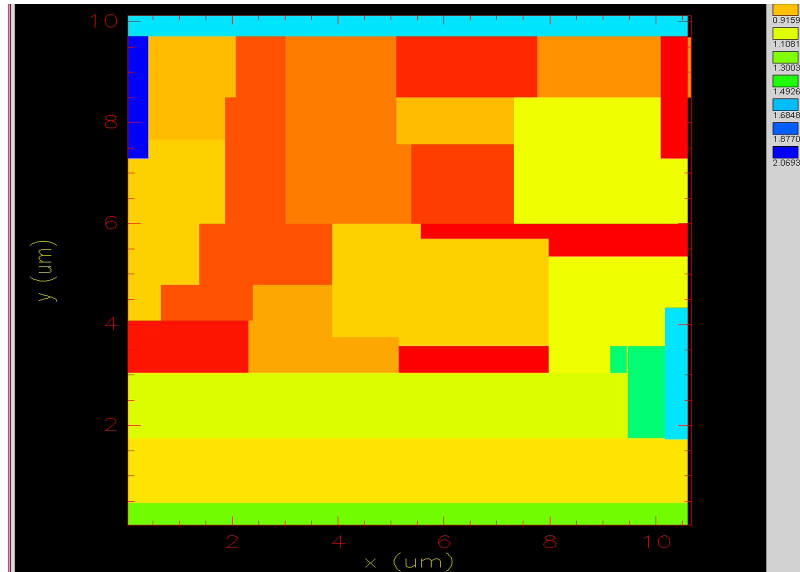


Figure 5.9. The non-uniform block-based on-die decap distribution in the microprocessor

M2 metal nodes. With a single 10ps source placed in the middle of a central unit, we observed the time domain current waveforms of all of the nonuniformly distributed caps on M2 (Figure 5.10, left). Note that as time progresses, all the cap currents sync up with the global resonant frequency described by the die-package resonance. As understood from Figure 5.8, this is the stage where locality is lost and all caps are charge-sharing. However, note that in the beginning, the response to the fast transient consists of multiple frequencies much higher than the global resonant one.

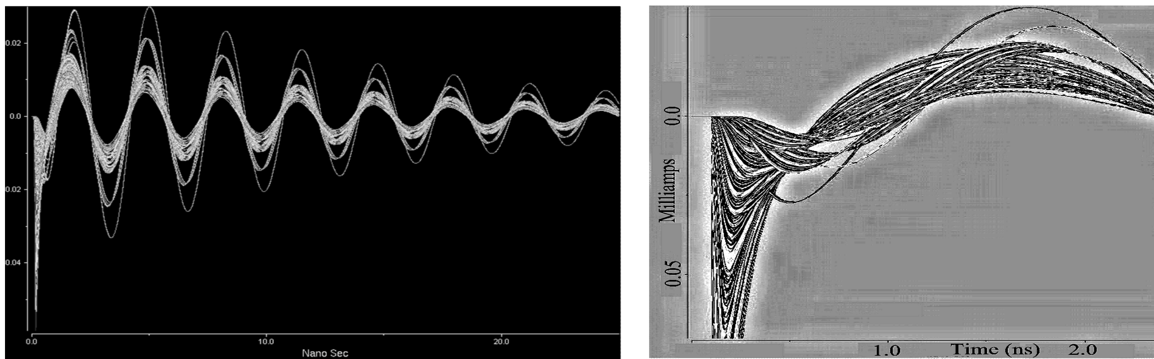


Figure 5.10. Transient current response of non-uniform decaps (left) and the magnified plot (right)

The highlighted waveforms of the first dip in Figure 5.10 (left) are expanded in the graph of Figure 5.10 (right). Here we see that the frequencies are multiple with some being higher than the main resonant frequency. This demonstrated that there were midfrequency effects due to non-uniform cap distribution and resistive grid isolation. In order to understand the locality of these midfrequency transients as compared to the global resonance, we plotted the amplitude of the currents at two specific time points: at the bottom of the first dip in Figure 5.10 (left), where the mid-frequency effects were visible (Figure 5.11), and one at the bottom of the second dip where the responses have almost converged to a global resonance (Figure 5.12). We observe clearly in the 3D plot in Figure 5.11 that the mid-frequency effects are "local" to a radius of approximately 1mm. However the current magnitude plot of Figure 5.12 shows that by the second dip there is almost global convergence and the cap currents reflect an almost perfect correlation with the full-die cap distribution in Figure 5.9. Thus mid-frequency effects may be resonant at less than full-die, but low frequency die-package effects are global. Please note, that this kind of RC locality is very different and of a wider area as compared to the high frequency locality

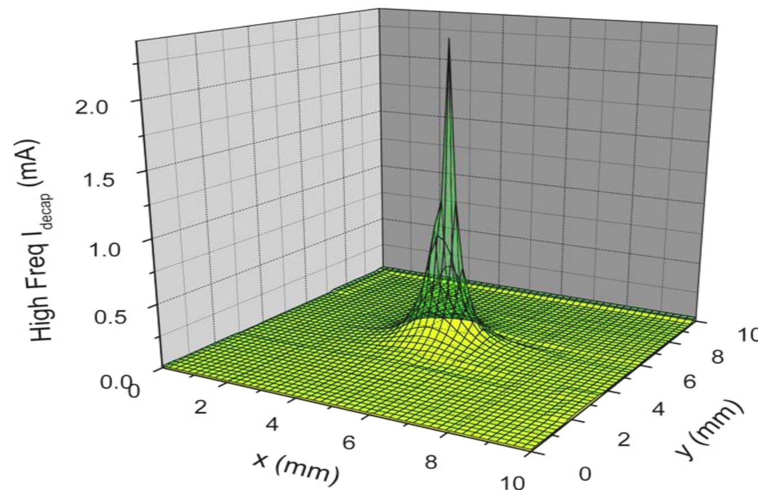


Figure 5.11. Locality of high-frequency currents provided by decaps

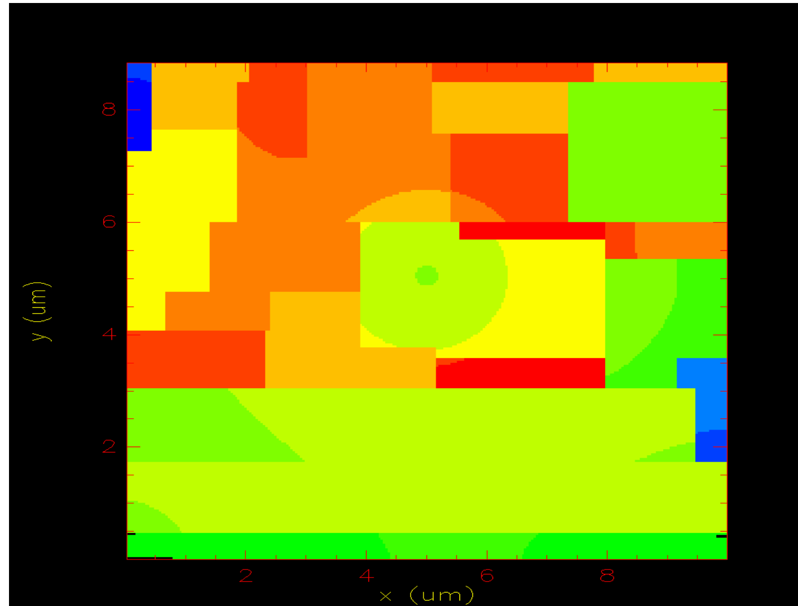


Figure 5.12. Distribution of mid-frequency decap currents which are strongly correlated to decap distribution map

spoken of earlier in discussing inductive effects where the locality was observed to be much smaller.

A visualized explanation of all of the above effects is shown in Figure 5.13. When a single gate switches (middle), it pulls in power delivery current from various sources. At high frequencies, the package with large parasitics is effectively isolated from the die. If the frequency is high enough in a small local area, it will excite on-die inductance but this

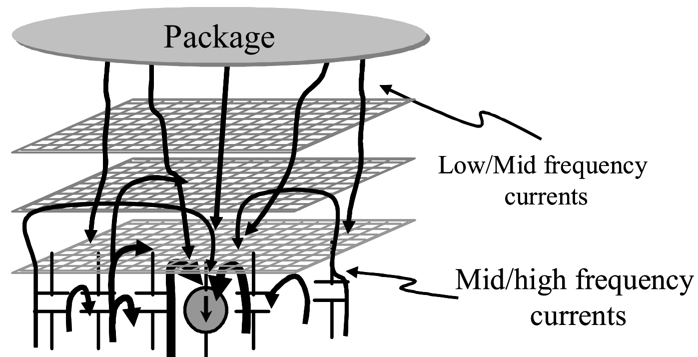


Figure 5.13. A visualization of the flow of currents on die

effect will be highly transient and within a few microns radius before the RC background absorbs the high frequency energy. The high frequency currents are immediately satisfied by caps nearby, either explicit de-caps, active device caps or wire cap. The farther away the cap, the less current will be supplied but the supply radius grows larger as the frequency lowers.

At some mid frequency (greater than the global die-package frequency) the package comes into effect and mid-frequency currents are supplied through the C4's. However, even at these frequencies, they continue to also be supplied by the caps. If the gate is surrounded by pockets of cap that are resistively isolated (partially, or completely) from other caps further out, the local caps will resonate only with the local C4 bump/package inductance above them, causing a mid-frequency resonance of a radius of a few hundred or more microns. This effectively is a small version of the total die at resonance. When the frequency is low enough (main die-package resonance), all of the caps and all of the C4's will interact to produce a global resonant frequency that is full-die in nature.

5.5 Conclusions and CAD implications

In this study, we demonstrated that in an industrial microprocessor design, the package inductance overwhelms the effect of on-die inductance even in 90nm technology and the impact of on-chip inductance remains insignificant and highly localized except at frequencies > 5Hz. It is also shown that the 2-D or lumped inductance models, which are commonly used, significantly overestimate the impact of on-die grid inductance compared to 3D models. The dominance of package inductance over on-die inductive effects necessitates accurate modeling of the package and cosimulation of package and on-die power dis-

tribution networks. The impact of decap distribution on the first and second transient voltage droops was shown in Section 5.4. A new mid-frequency localized effect was found to be dependent on the nearby nonuniform decap distribution, while the main resonant behavior depends on the total global decap. It was noted that resistively isolated pockets of decaps can lead to multiple resonances. Thus, accurately modeling the decap distribution is extremely important to analyze the transient behavior of power grid accurately. Due to the distributed nature of C4s in flip-chip packaging, voltage drop induced due to a current excitation may be limited to the vicinity of the current source. Recently, several works [19][62] have been proposed to exploit this locality in power grids to accelerate voltage-drop analysis. However, we demonstrated that transient locality is a strong function of the excitation frequency. We showed that although the voltage drop exhibits locality for the DC and high frequency excitations, it is global at frequencies around the resonance frequency caused by package inductance and on-die decaps. Moreover, the area of locality is dependent on the frequency of excitations as illustrated in Section 5.3.2. Thus, although locality can be used to simplify and accelerate the static power-grid analysis as proposed in [19], its usage for transient power-grid analysis and optimization may lead to erroneous results, unless integrated with these effects.

CHAPTER VI

AN ANALOG ACTIVE DECAP CIRCUIT FOR INDUCTIVE-SUPPLY-NOISE SUPPRESSION

6.1 Introduction

Increasing power consumption and clock frequency have significantly exacerbated the L*di*/d*t* drop, such that it is now considered the dominant portion of the overall supply drop in high performance designs [11][69]. Also, methods which are effective in IR-drop mitigation, such as increasing on-chip power grid metallization are less effective in reducing L*di*/d*t* drop which is primarily caused by package inductance. This has given rise to an urgent need for the suppression of inductive noise in power distribution networks in the presence of large transient switching currents.

With technology scaling, gate leakage has become a significant percentage of the overall leakage which places a significant limitation on the maximum amount of thin oxide decap that can be added on the die. Recently, on-die metal-insulator-metal (MIM) decoupling capacitors have been suggested in [67] to reduce the area overhead and leakage of on-die decaps. MIM decaps require extra processing steps and have low capacitance to area efficiency as compared to thin oxide decaps. We refer to these explicit decaps as *passive* decaps.

Active decaps [4][37][51][77][82] employ a circuit to increase the amount of charge transfer to-and- from the power supply network during a supply voltage fluctuation. The objective of these approaches is to obtain a smaller supply drop for the same amount of explicit decoupling capacitance. In this chapter, we propose a novel active decap circuit to suppress L*di/dt* noise in power supply networks. The proposed active decap circuit is powered from a separate supply network (referred to as the *active* supply) and the existing power supply pads/C4s are split between the regular supply network and the active supply network. The proposed decap circuit senses and amplifies the voltage drop in the power and ground lines of the regular supply network. This amplified and inverted voltage drop is then placed on one of the terminals of the active decap, allowing it to pump significantly more charge into the regular power grid (or absorb more charge from the ground grid) as compared to passive decaps. We study the trade-offs involved in the optimal allocation of the total decap between active and passive decap. Similarly, we study the trade-offs in the allocation of the total number of pads/C4's allocated between the regular supply network and the active supply network.

Finally, we show that the area overhead of the amplifier is small compared to the area consumed by the decaps themselves. Also, since the active supply network is dedicated only to the active decap circuits, it does not require an extensive power distribution network and hence does not lead to a significant supply grid overhead. We perform extensive simulations of the proposed technique and show that it can improve the frequency response of the power distribution network dramatically. The transient voltage drop was shown to reduce by 45% compared to the use of only passive decaps, corresponding to an increase in the effective decap of approximately 8X.

The rest of the chapter is organized as follows. Section 6.2 explains the proposed circuit in detail. Practical considerations in implementing of the proposed circuit are presented in Section 6.3. Section 6.4 presents the experimental validation of the circuit and conclusions are drawn in Section 6.5.

6.2 Proposed active decap circuit

This section presents the proposed active decap circuit and compares its performance to passive decap. We first present the analysis of a power delivery model with passive decaps in Section 6.2.1. The proposed active circuit is then presented in Section 6.2.2 and compared with passive decaps in Section 6.2.3.

6.2.1 Power delivery network with passive decap

Figure 6.1 shows a commonly used model of a power distribution network with on-die passive decaps represented by C_d. The lumped inductors L_{Vdd} and L_{Vss} model the inductance attributed to the power supply network in the motherboard, socket, package and on-die inductance, whereas resistors R_{Vdd} and R_{Vss} represent the resistive parasitics in the

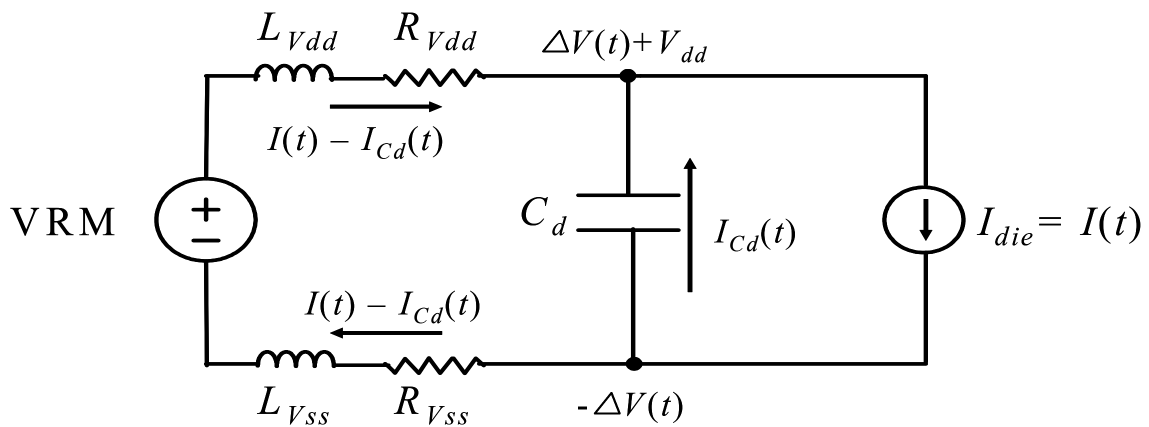


Figure 6.1. Model of a power delivery network

supply network. I_{die} is the current drawn from the supply network by the on-die devices. In addition to the on-die decaps, discrete decaps are also placed at various locations (VRM decap, motherboard decap and package decap) in a power distribution network. The resonance frequencies due to these additional decaps are considerably lower [11] (~kHz-5MHz) than the resonance frequency due to package inductance and on-die decaps (~100MHz). Extra board and package decaps can be added with relative ease but such is not the case with on-die decaps due to the leakage and area constraints. Hence, in our analysis, we concentrate only on the effectiveness of on-die decaps.

Let $\Delta V(t)$ be the transient voltage drop (ground bounce) seen by the devices on-die due to time-varying current $I(t)$. The decap C_d observes a net change in potential of $2\Delta V(t)$ across its two terminals and provides a charge of $2C_d\Delta V(t)$ to both the power and ground lines. The current, $I_{Cd}(t)$, provided by the decaps to the on-die devices is therefore given by:

$$I_{Cd}(t) = 2C_d \cdot \frac{\mathrm{d}}{\mathrm{d}t}\Delta V(t) \quad \text{(EQ 6.1)}$$

The DC part of on-die current $I(t)$, along with the rest of transient current ($I_{pkg}(t)$) is supplied by the VRM through the inductors L_{Vdd} and L_{Vss} is:

$$\Delta V(t) = I_{pkg}(t) \cdot R + L\frac{\mathrm{d}}{\mathrm{d}t}I_{pkg}(t) \quad \text{(EQ 6.2)}$$

The current provided by the VRM, $I_{pkg}(t)$, causes the drop in the power supply $\Delta V(t)$ as follows:

$$\Delta V(t) = I_{pkg}(t) \cdot R + L\frac{\mathrm{d}}{\mathrm{d}t}I_{pkg}(t) \quad \text{(EQ 6.3)}$$

It can be observed that the voltage drop decreases as the amount of on-die decaps is increased. The resonance frequency, ω_r, due to the interaction of package inductance with on-die decaps is given by:

$$\omega_r = \sqrt{L_{pkg} C_d} \qquad \text{(EQ 6.4)}$$

where L_{pkg} is the net inductance of the supply network (L_{Vdd}+L_{Vss} in Figure 6.1). As the amount of on-die decaps increases, the resonance frequency as well as the peak Ld*i*/d*t* noise is lowered. The aim is, therefore, to maximize the total amount of decap C_d within the area and leakage constraints, thus minimizing both the IR as well as the Ld*i*/d*t* drop in the network. In the next subsections, we present the proposed active decap circuit and compare it with passive decap.

6.2.2 Active decap circuit

The proposed active decap circuit consists of two operational amplifiers (*opamps*) which amplify and invert the voltage drop across the power and ground lines of the supply network. The total available passive decap is divided evenly among the power and ground networks. The two opamps drive one terminal of each decap as shown in Figure 6.2. The opamps operate with a separate, so called *active*, supply and hence, the total number of C4s available for power delivery in the chip is divided between the active and the regular supply. For simplicity in our discussion in this section, we assume that the active supply is ideal and does not affect the gain or the current drive of the opamps. However, later in Section 6.4.3, we will study what percentage of the total available C4s should be dedicated to the active supply versus the regular power grid. We find that only a small percentage of

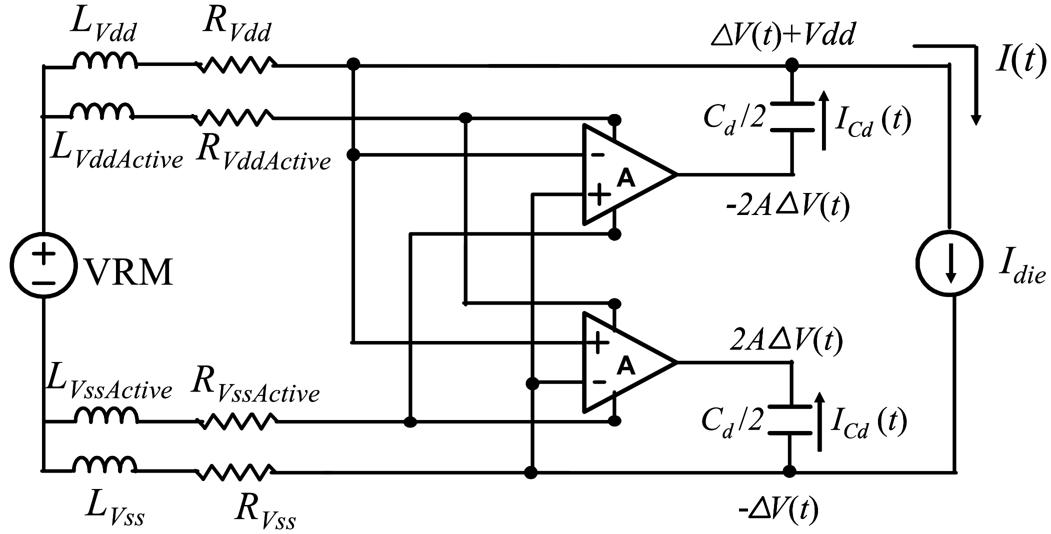


Figure 6.2. Proposed active decoupling capacitance circuit

the total C4s need to be allocated to the active supply and that the non-ideality of the active supply has only minor impact on the effectiveness of the method.

We now compute the theoretical effectiveness of the proposed method. The transient voltage drop (ground bounce), $\Delta V(t)$, is amplified to a potential of $-2A\Delta V(t)$ and $2A\Delta V(t)$, respectively, by the two amplifiers driving the *Vdd* and *Vss* decaps. As a result, each of the two decaps observes a difference in potential of $(2A+1)\Delta V(t)$ across its terminals. Thus, the current $I_{Cd}(t)$, provided by the decaps to the devices is given by:

$$I_{Cd}(t) = (2A+1)\frac{C_d}{2} \cdot \frac{\mathrm{d}}{\mathrm{d}t}\Delta V(t) \qquad \text{(EQ 6.5)}$$

As compared to EQ 6.1, the current provided by the active decaps is therefore considerably higher if the gain A of the opamp is sufficiently high. In general, for the same on-die current, $I(t)$, and same amount of decap, C_d, the active decap will lead to a lesser voltage drop, $\Delta V(t)$, provided the following condition holds:

$$(2A+1)\frac{C_d}{2}\cdot\frac{\mathrm{d}}{\mathrm{d}t}\Delta V(t) \geq 2C_d\cdot\frac{\mathrm{d}}{\mathrm{d}t}\Delta V(t) \quad \text{(EQ 6.6)}$$

or, $A \geq 1.5$

If the gain of the opamp is greater than 1.5, a better power supply regulation can be achieved or the amount of decap can be reduced for the same supply regulation. The *effective* decoupling capacitance achieved using this active circuit is:

$$C_{eff} = (2A+1)\frac{C_d}{2} \quad \text{(EQ 6.7)}$$

where, C_d is the amount of decap present in the chip driven by the opamps and A is the gain of opamps. In a practical implementation, the gain A of the opamp is a function of frequency and the gain decreases as frequency of operation is increased as will be discussed later in Section 6.3.

It should be noted that the opamps in the active circuit must operate with an external active power supply and cannot be connected to the regular on-die power supply. Figure 6.3 explains why the active circuit cannot operate if its supply rails are connected to the same power supplies which provide current to the other on-die devices. Let $I(t)$ be the cur-

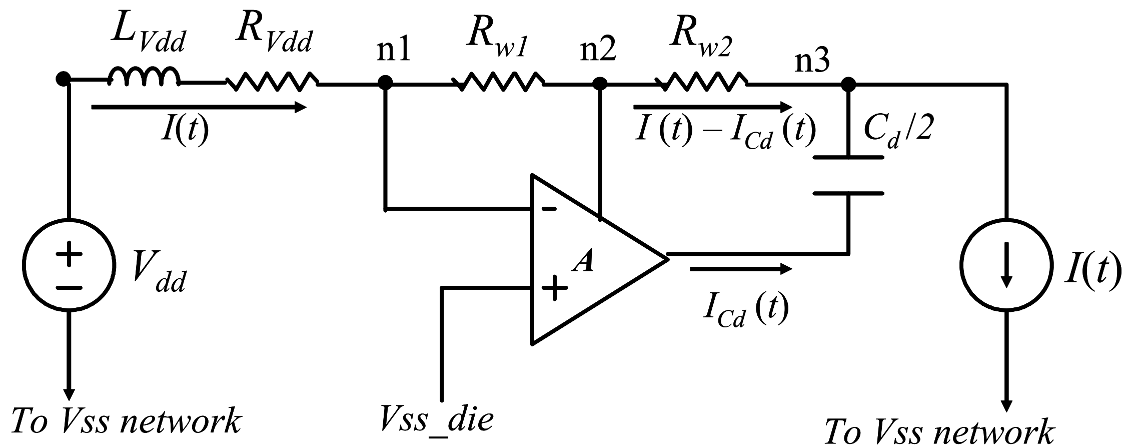


Figure 6.3. Need for external active supply for operational amplifiers

rent consumed by the chip and $I_{Cd}(t)$ be the current provided by decap C_d. Then, the current flowing in the resistance R_{w2} is $I(t)$-$I_{Cd}(t)$. On the other hand, the current flowing in the package remains I(t) as shown in the figure. Thus, the only improvement in supply drop is in the IR-drop due to a reduced current now flowing through the resistor R_{w2}. The package inductance L_{pkg} has full on-chip current flowing across it, resulting in a worse inductive effect as compared to the use of passive decaps.

While separate active supply rails are needed in the proposed method, we show in Section 6.4.3 that only a small portion of the total C4s need to allocated to the active supply. Furthermore, since the active supply is only connected to the active decap circuits, a full supply grid design is not needed for the active supply. Instead, C4s in the immediate vicinity of the active decap circuits can be directly routed to the circuits, thereby minimizing the routing overhead incurred by the additional supply.

6.2.3 Comparison of active and passive decaps

This subsection explains and compares the response of the proposed active decap circuit with that of design with only passive decap, assuming ideal opamps. The practical design constraints and opamp design are presented in Section 6.3.

Figure 6.4 compares the frequency response of the power supply network with passive decaps and active decaps. The HSPICE simulations were performed in 130nm triple-well technology with nominal supply of 1.2V. The capacitive and inductive parasitics were based on an industrial power supply grid for a high performance processor design. The simulated circuit represents a 0.7mm X 0.7mm section of the die since it is not possible to

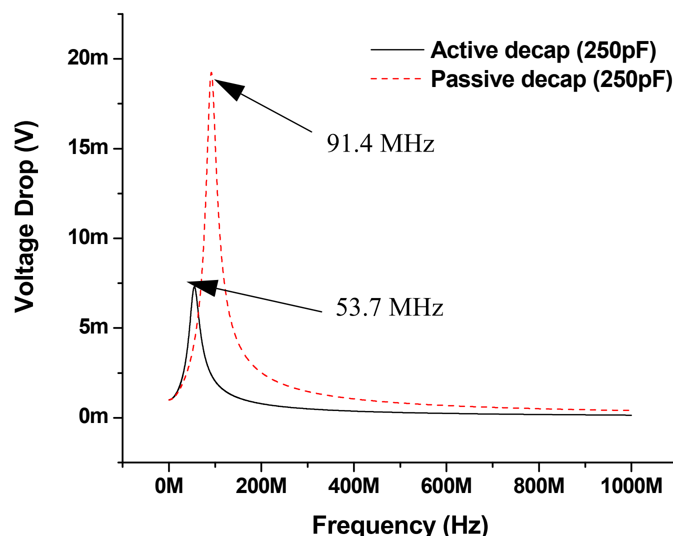


Figure 6.4. Comparison of frequency response with active and passive decaps

simulate a detailed model of the entire supply network. However, since the active decap circuit can be replicated across the die, the results for the whole die are similar.

The total amount of decap available, C_d, (passive and active combined), was 250pF. The gain A of the ideal opamps was set to be 5 which results in an effective capacitance of $3C_d$ using EQ 6.7. As a result, the resonance frequency was reduced by a factor of $\sqrt{3}$. The resonance frequency of the grid due to package inductance with passive on-die decaps was observed to be 91.4MHz. The use of active decaps reduced the resonance frequency to 53.7MHz. Moreover, use of active decaps also led to a significant reduction in peak noise at the resonant frequency as shown in Figure 6.4.

Figure 6.5b shows the transient response of the power grid with active and passive decaps to the current profile shown in Figure 6.5a. The current pattern was modeled on a high performance microprocessor switching from a state of low activity operations to a state of high activity operations [11] in 5 clock cycles and back. The processor is first in a

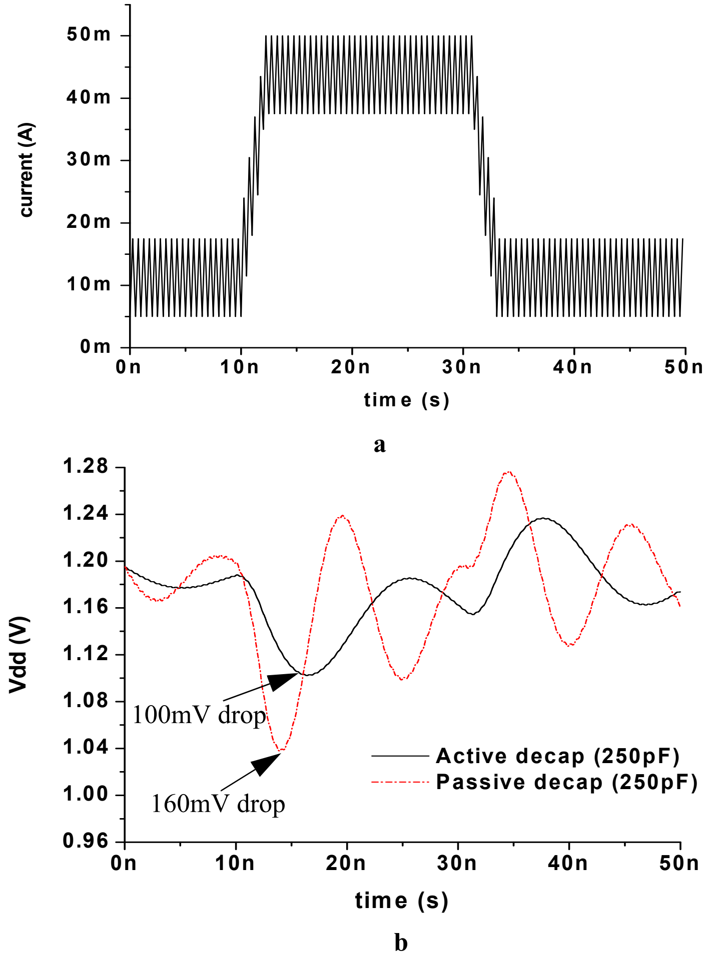


Figure 6.5. Transient current profile (a) and voltage response (b) using active and passive decaps

state of low power instructions consuming 10% to 30% of its peak power. The state of the processor is then ramped from this low power mode to a high power mode in 5 clock cycles, resulting in current levels oscillating between 75% to 100%. Finally, the processor is ramped down to the low power mode again. The operating clock frequency was assumed to be 4GHz. With an opamp gain of 5, the worst-case supply drop was reduced from 160mV to 100mV, resulting in a percentage improvement of 37.5% as shown in Figure 6.5b. It is clear that higher reduction in supply drop can be obtained by increasing the gain of the opamp.

The results in Figure 6.4 and Figure 6.5 demonstrate the effectiveness of the proposed circuit with ideal opamps, in the presence of large current transients. The next section presents the opamp design and other practical considerations in the silicon implementation of the proposed approach.

6.3 Opamp design

The previous section described the proposed circuit with ideal opamps having fixed gain, infinite input impedance and zero output impedance. In reality, the gain of an operational amplifier is a function of operating frequency and drops as frequency is increased. Hence the net effective capacitance given in EQ 6.7 decreases with increasing frequency. In order to reduce the impedance of a power supply network at the resonance frequency, the gain of the amplifier should be as high as possible. On the other hand, the bandwidth of the opamp should be high enough to suppress high frequency inductive noise as well. These are contradicting requirements since the gain-bandwidth product is usually near-constant in a given technology. In all, the opamp used in the active circuit should have the following characteristics:

1. High gain at the resonance frequency for suppressing the dominant voltage fluctuations and resonance effects.
2. High bandwidth for mitigating high frequency inductive noise.
3. High current drive in order to drive large decaps.
4. Small amplification delay for fast reaction to high frequency current transients.
5. Robustness of opamp to its own active power supply variations.

The opamp used in the proposed active circuit consists of three stages. The inputs to the opamp have common-mode voltages of *Vdd* and *Vss*. This entails the use of level shifters as the first stage of the opamp, so as to bring the inputs to their common-mode bias levels for the gain stage. The gain stage forms the second stage of the opamp. For improving the output impedance and current drive, a high drive strength output stage is used after the gain stage.

Figure 6.6 shows the circuit schematic of the opamp, which amplifies the difference between *Vdd* and *Vss* supply drops and applies it to a terminal of the decap connected to *Vdd* grid. The bias generation circuitry is not shown for clarity. As mentioned in Section 6.2.3, the opamp needs to operate using an active supply different from the main power supply of the other devices on die. All the bias supplies and currents are generated from the active supply to the opamp. Transistors M1, M2 and bias currents I_{b1}, I_{b2} form the

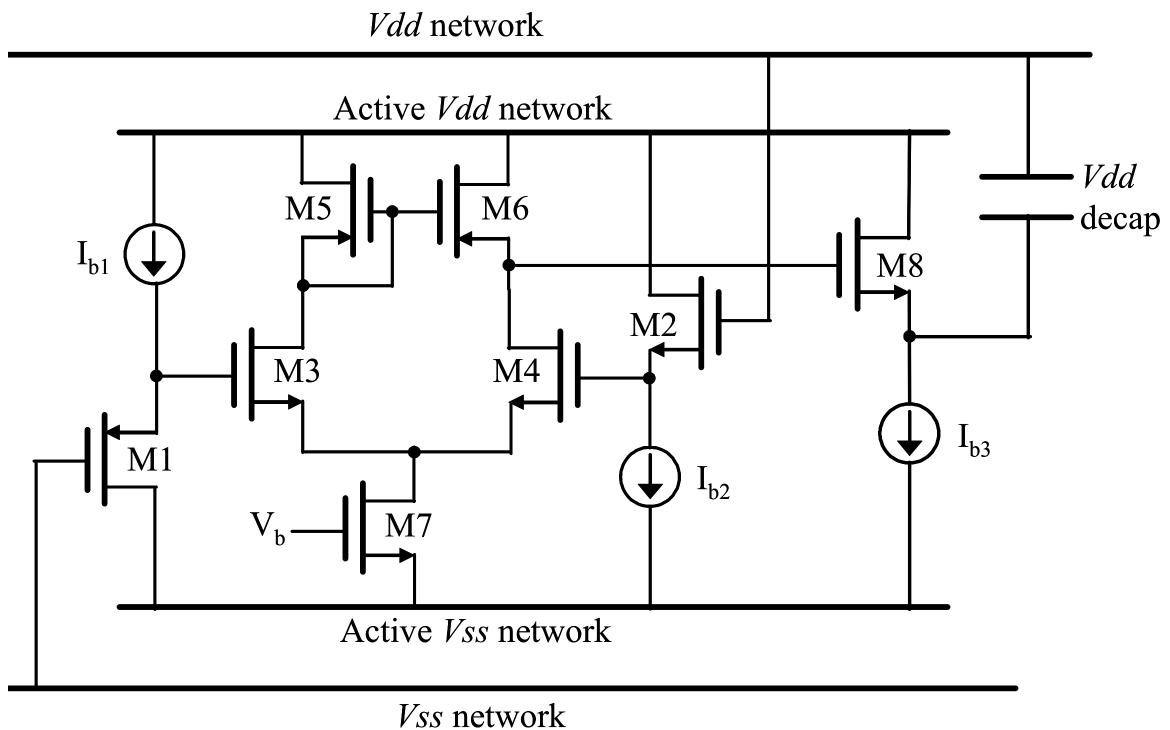


Figure 6.6. Operational amplifier schematic

input level shifters which convert the input common bias supplies of regular *Vss* and *Vdd* to the input common-mode voltage of the gain stage, which was set at 500mV. The gain stage, which succeeds the level shifters is formed using a single-ended differential amplifier pair formed by transistors M3 and M4. Transistors M5 and M6 are the active current mirror loads to the diffamp and M7 provides its biasing current. The diffamp is followed by a source follower (M8 and I_{b3}) as the output stage, which reduces the output impedance and provides high current drive capability to the amplifier.

The gain of the above opamp, A has a dominant pole at frequency ω_0 and can be approximated as follows [63]:

$$A = \frac{A_0}{\left(1 + j\frac{\omega}{\omega_0}\right)} \quad \text{(EQ 6.8)}$$

where A_0 is the gain of the opamp at ω=0.

Let us consider a power grid network consisting only of active decaps, C_d and no passive decaps. Then, the effective on-die capacitance in the circuit C_{eff} is given by:

$$C_{eff} = \left(2\frac{A_0}{\left(1 + j\frac{\omega}{\omega_0}\right)} + 1\right)\frac{C_d}{2} \quad \text{(EQ 6.9)}$$

assuming that the total decap C_d is evenly distributed for the *Vdd* and *Vss* lines. The impedance of the power supply network in presence of active decaps is expressed as follows:

$$Z(s) = \frac{(R + sL)\left(1 + \frac{s}{s_0}\right)}{\frac{s^3 LC_d}{s_0} + \left(LC_d(1 + 2A_0) + \frac{RC_d}{s_0}\right)s^2 + \left(\frac{RC_d}{2} + A_0RC_d + \frac{1}{s_0}\right)s + 1} \quad \text{(EQ 6.10)}$$

where L is the package inductance of the grid, C_d is the total active decap present in the grid and $s_0=j\omega_0$. Simulation results showing the impedance of the supply network obtained for different ratios of active and passive decaps are presented in Section 6.4.2.

EQ 6.9 shows that the effective decap starts reducing after ω_0. As a result, the package inductance becomes a dominant part of the overall power supply network's impedance, thereby resulting in a linear increase in impedance with increasing frequency as shown in EQ 6.10. To improve the high frequency behavior, a small amount of passive decaps is therefore used in conjunction with the active decaps to suppress high frequency noise. This is explored further in Section 6.4.1.

6.4 Simulation results

The proposed active decap circuit was designed in 130nm triple-well technology with a nominal supply of 1.2V and tested with HSPICE simulations. The next subsection presents the layout and simulations of the opamp used in the proposed active decap circuit.

6.4.1 Opamp simulations

Figure 6.7 shows the layout of the opamp with a drive capability of 5pF. Several of these opamps are used to drive banks of decaps placed at different locations on the die. The area of the opamp was 14.1μmx11.8μm. In the technology used for experiments, this area amounts to approximately 0.39pF of capacitance. The area overhead of the opamp, when compared to the area needed for decap insertion is therefore small at approximately 8%. The area overhead of the opamp design is addressed by reducing the total decap by this

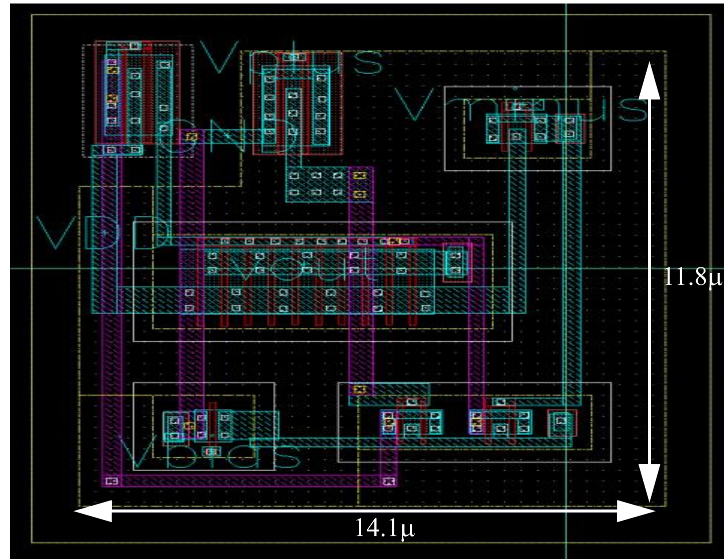


Figure 6.7. Layout of the opamp with drive strength of 5pF

amount, resulting in a slightly lower total available decap. However, this reduction in available decap is easily compensated by the much larger effective capacitance that results from the active decap circuit.

Figure 6.8 shows the simulated gain of the opamp as a function of frequency while Figure 6.9 shows the transient response of the opamps obtained when used to drive a decap of

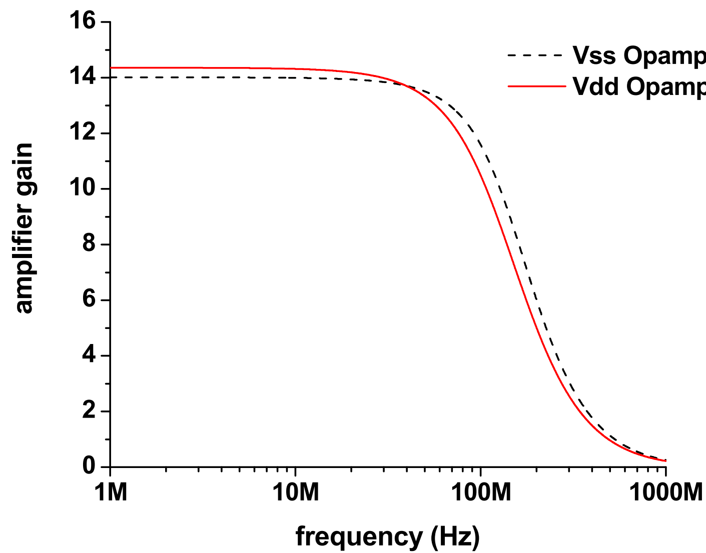


Figure 6.8. Gain of the opamp

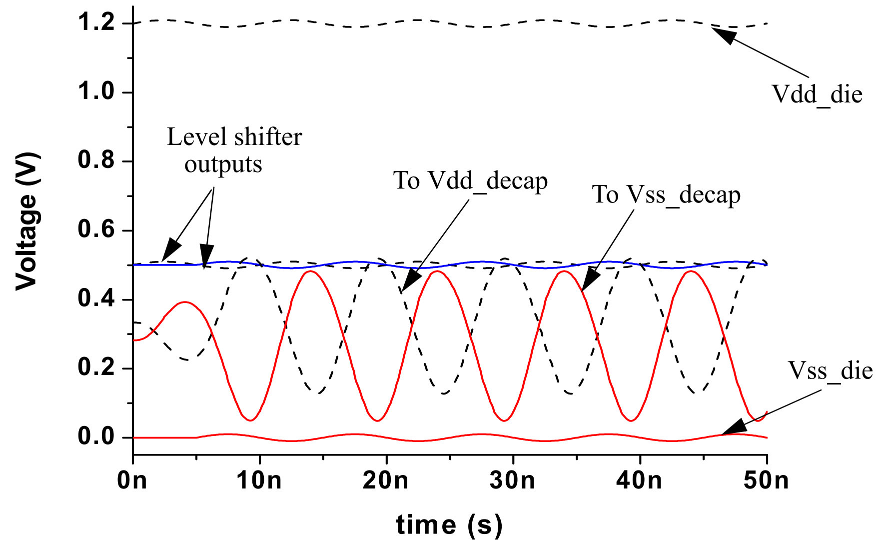
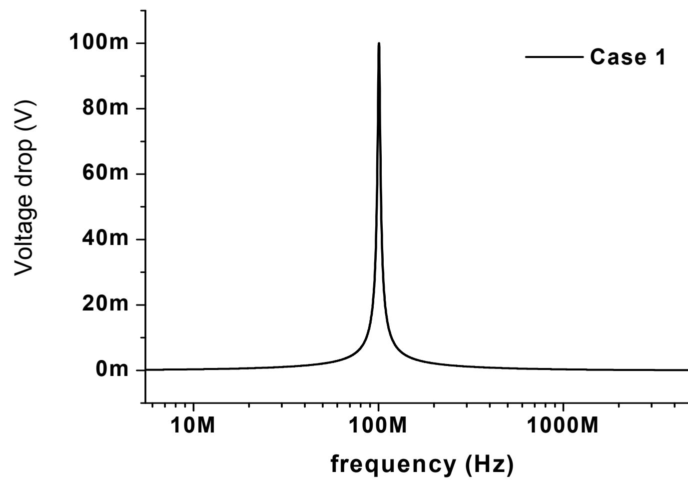


Figure 6.9. Transient response of the opamp

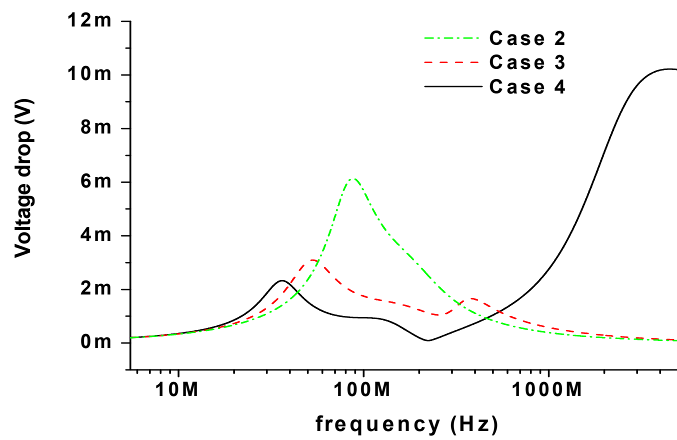
10pF (5pF each of *Vdd* and *Vss* decaps). The gains of *Vdd decap* and *Vss decap* were 14 and 14.2 respectively. The bandwidths of the two opamps were 197MHz and 190MHz respectively. The two signals at the extreme top and bottom of the Figure 6.9 are the supply fluctuations in *Vdd* and *Vss* grid respectively. Using input level shifters, the common-mode voltages of these supply fluctuations (1.2V and 0V respectively) are converted to a DC voltage of 500mV, the bias voltage of the gain stage. In the figure, the "level shifter outputs" signals represent the signals input to the gain stage of the opamp. Vss_decap and Vdd_decap signals are the output of the output stage of the opamps used to drive the decaps connected to *Vdd* and *Vss* grid respectively.

6.4.2 Comparison of active and passive decaps

In this subsection, we compare the frequency and transient response of the power supply drop with different ratios of active and passive decaps present in the supply network. Figure 6.10 compares the frequency response of the voltage drop for 4 different cases:



(a)



(b)

Figure 6.10. Comparison of frequency response of power grid with different decap sizes

Case 1: 250pF passive decap

Case 2: 180.5pF passive + 50pF active decap

Case 3: 80.5pF passive + 200pF active decap

Case 4: 230.5pF active decap.

Note that in all but the first, all passive case, the total available decap is reduced slightly to account for the area overhead of the active decap circuit.

As shown in Figure 6.10, the use of active decaps in Case 2, 3 and 4 results in a reduction in the worst case supply voltage drop observed at resonance. Case 1, with only passive decaps present in the grid, has a single resonance frequency at 110MHz. Case 4, which only has active decaps and consequently the maximum amount of effective decap, has the maximum suppression of the inductive noise at resonance frequency and also the maximum lowering of the resonance frequency. In particular, the resonance frequency is lowered to 35.7MHz from 110MHz, indicating an effective decap of 2nF, an 8x increase in the amount of decap. However, the effective capacitance starts to decrease rapidly with frequency resulting in an increase in supply drop at high frequencies, although the peak voltage drop still remains at a much lower values as compared to the one in Case 1. Cases 2 and 3, which use a combination of active and passive decaps, show noise suppression at all the frequencies. These cases also show a very slight second resonance at a frequency higher than the one in Case 1 due to the interaction between package inductance and passive decap present in the grid. As the amount of passive decap in the grid is reduced (active decap is increased), the first resonance peak is suppressed and the second extra resonance peak increases.

Figure 6.11 compares the transient response of the voltage drop for different ratio of on-die active and passive decaps present in the power grid. Figure 6.11a shows the current profile used, which resembles the ramp-up of a microprocessor for low-power instructions to high-power instructions in 15 cycles. The clock frequency of the design was assumed to be 4GHz in the experiment. Figure 6.11b shows the transient response for three configurations: 250pF passive, 50pF active + 180.5pF passive and 100pF active + 130.5pF passive

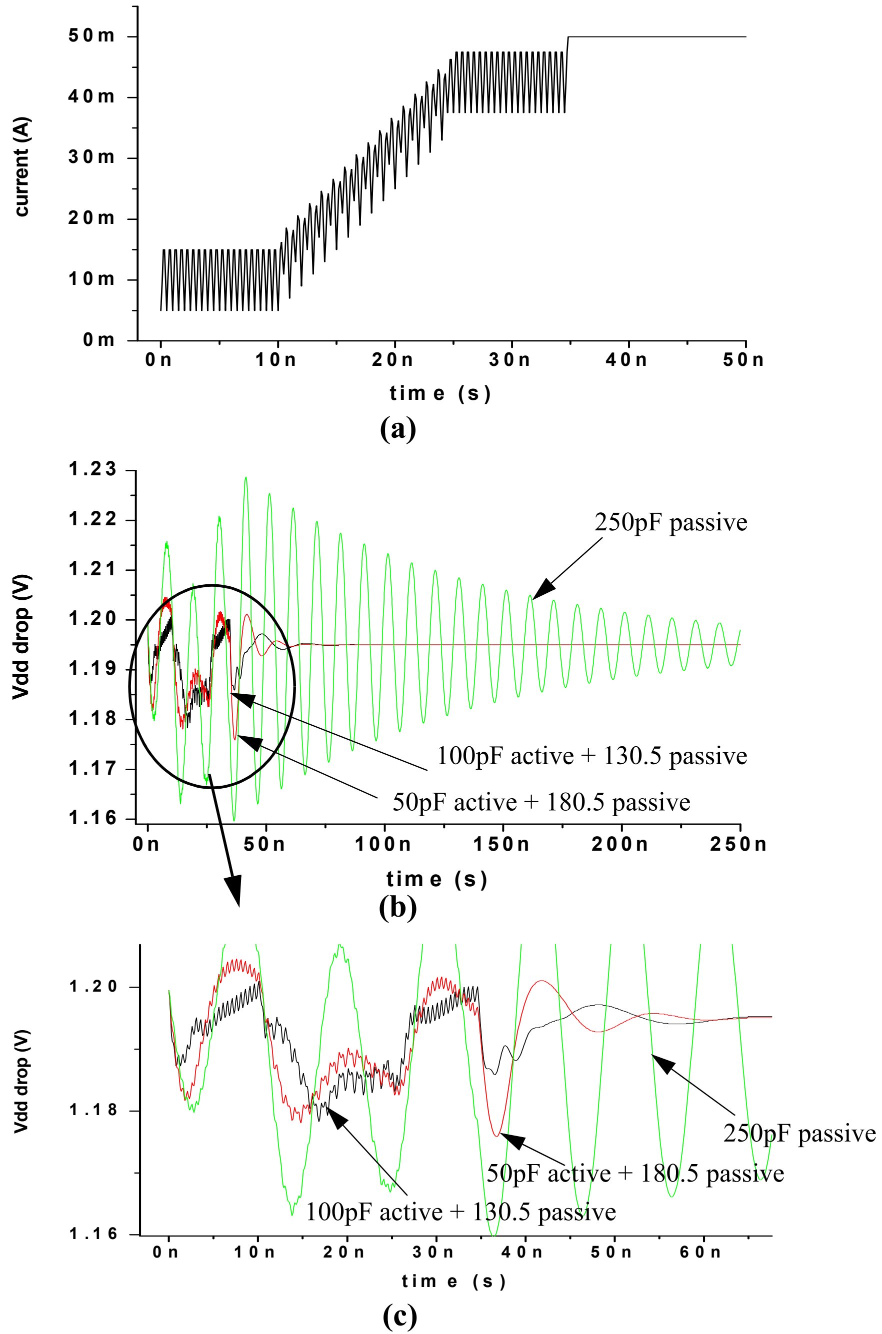


Figure 6.11. Comparison of transient response of power grid with different decap sizes

decaps. The worst-case supply drop is reduced from 40mV (in 250pF passive case) to 22mV (in 100pF active + 130.5pF passive case) resulting in a 45% reduction in worst voltage drop. In addition, the low frequency ringing is much less with the usage of active

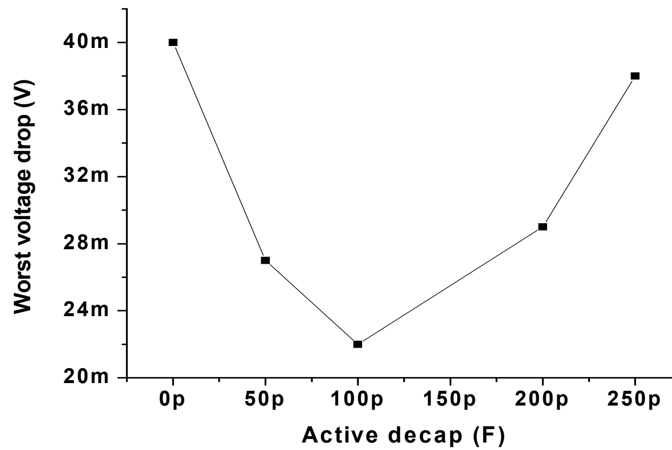


Figure 6.12. Worst-case voltage-drop variation with decap allocation

decaps. Figure 6.12 shows the worst voltage drop as a function of the total 250pF decap distributed between active and passive decaps in different ratios.

6.4.3 Impact of C4 distribution

In this subsection, we study the impact of the distribution of C4s among the active and regular supply networks. Figure 6.13 plots the worst-case voltage drop as a function of the percentage of total C4s allocated to the active supply. As the number of C4s allocated to the active supply is increased beyond a limit, the overall package inductance of the regular supply increases, thus worsening the Ldt/dt noise in the regular supply. In addition, the IR drop in the regular supply also increases with fewer pads allocated to the regular supply. This results in an increase in the voltage drop in the regular grid. On the other hand, if only a few pads are allocated to the active supply, the fluctuations in the supply provided to the opamps increase, thus lowering their respective gains and drive strengths. Figure 6.13 shows that the worst-voltage drop is at its minimum when 6% of the total C4s (3 C4s) are allocated to the active supply and the remaining 47 C4s are allocated to the regular supply.

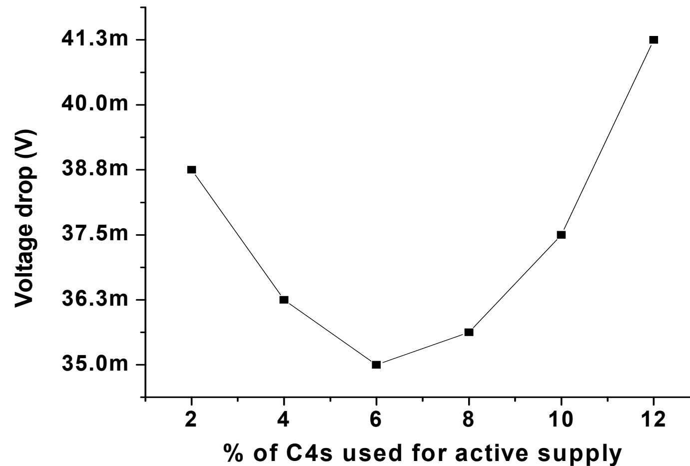


Figure 6.13. Impact of C4 distribution on worst-case voltage drop

6.5 Conclusions

An active decap circuit was proposed to suppress the inductive noise in power distribution networks. The proposed circuit senses the supply drop and drives an amplified and inverted voltage fluctuation on the decap. The active decoupling circuit is powered by a separate power supply and we studied the optimal allocation of the total C4s between this second power supply and the regular supply as well as the optimal allocation of the total decoupling capacitance between actively switched and traditional static decap. Using the proposed method, the maximum supply drop is reduced by 45% compared to the use of only traditional decap, corresponding to an increase in the effective decap of approximately 8X.

CHAPTER VII

DIGITAL CIRCUIT-TECHNIQUES FOR ACTIVE INDUCTIVE-SUPPLY-NOISE SUPPRESSION

7.1 Introduction

Aggressive scaling and increasing clock frequency have exacerbated inductive (L*di*/*dt*) supply noise, impacting the robustness of power delivery networks. L*di*/*dt* drop is further aggravated by commonly used power reduction techniques such as power/clock-gating and frequency-stepping in DVS systems. On-die passive decap, which has traditionally been used for suppressing L*di*/*dt*, has become expensive due to its area and leakage power overhead.

In chapter VI, we proposed an analog circuit technique to suppress excessive L*di*/*dt* supply noise. The proposed technique utilizes an operational amplifier (opamp) to amplify the supply noise and transfer a larger amount of charge from the decoupling capacitances. The efficacy of the proposed circuit relies strongly on the gain and bandwidth characteristics of the opamp. In general, a higher gain and a larger bandwidth over the entire frequency spectrum is imperative for the effective operation of the circuit. However, the gain-bandwidth product is constant for a particular technology node and the gain drops after the cut-off frequency, f_0, of an amplifier. This gain reduction at frequencies larger than f_0 results in degraded performance and a reduction in effective decap of the proposed circuit. Also,

the output stage of the analog circuit consumes a large amount of quiescent current, adversely affecting the power consumption of the design. Lastly, supply voltage scaling with each technology generation makes it difficult to design high-gain, single stage amplifiers. Increasing the stages in an amplifier increases the amplification delay which may adversely affects the stability of the proposed circuit at high frequencies. All the aforementioned reasons led us to investigate digital circuit techniques for supply noise regulation.

Several partial or all-digital circuits have recently been proposed in order to actively regulate the supply against sudden surges in load-current. A switched capacitor based supply voltage regulator was proposed in [4]. The authors proposed switching decap banks between series and parallel configuration depending on the supply drop amplitude. In [84], a band-pass filter is used to detect supply noise resonance. An artificial shunt load is connected between Vdd and Vss to dampen the power delivery system during resonance. In [56], a shunt high voltage supply is connected to the regular power grid when power-gated logic blocks wake up from the sleep state.

The above techniques deliver only limited charge [4], are suited only for resonance damping [84] or do not have integrated undershoot/overshoot detectors [56]. Recently, several adaptive frequency management techniques [30][76] have also been proposed to compensate for supply transients. These techniques employ supply drop monitors at various locations on the die. The frequency of operation is altered to accommodate for the fluctuations in the power supply. This adaptive frequency management results in improved average-case performance since the worst-case supply drops may not be very

frequent. However, the worst-case performance is unaffected since the worst-case supply drop remains the same.

In this chapter, we propose three all-digital active circuit techniques to detect and suppress excessive supply voltage undershoots and overshoots caused by large current transients or by excitation of supply resonance. We propose all-digital supply drop undershoot and overshoot detectors which are used to monitor excessive supply fluctuations in the three proposed circuit techniques. We also propose an all-digital supply noise measurement circuit capable of sampling the noise at 20Gsamples/sec. All the circuits were integrated in a prototype fabricated in a 0.13µm, triple-well CMOS technology.

The remainder of the chapter is organized as follows. Section 2 describes the charge injection based active decoupling circuit and the undershoot/overshoot detectors. Section 3 and 4 describe the high-voltage charge pump based active circuit and high-voltage shunt supply based active circuit respectively. The proposed noise measurement circuit is presented in section 4. The conclusions are summarized in section 6.

7.2 Charge-injection-based active decoupling circuit

The proposed circuit uses an active decap bank, C_a, and a nominal-voltage active supply, Vdd_a, to inject extra charge into the power grid during an undershoot emergency. The use of a nominal-voltage Vdd_a obviates the need for any high-voltage supplies and enables use of decaps and switching transistors with nominal oxide thickness. Furthermore, the proposed method has the advantage that C_a behaves as a passive decap when the supply voltage is within safety margins.

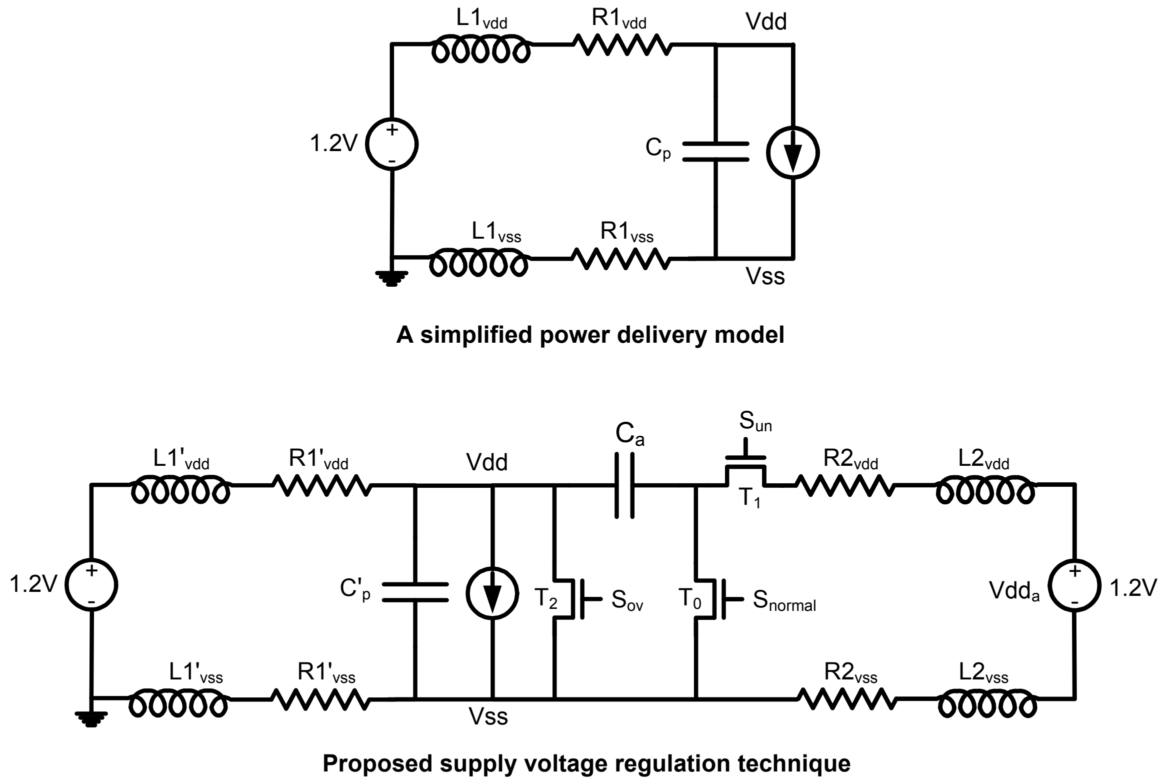


Figure 7.1. A simplified power distribution model and the proposed regulation technique

7.2.1 Proposed regulation technique

Figure 7.1(a) shows a simplified model of a power delivery network to illustrate the concept. The impedance of the power distribution networks is modeled as a series combination of lumped elements, $L1_{vdd}$ and $R1_{vdd}$. Similarly, $L1_{vss}$ and $R1_{vss}$ represent the parasitic inductance and resistance of the ground network. The total amount of non-switching as well as explicitly added on-die decap is represented by the lumped capacitance C_p. A time-varying current source models the switching current of all the logic blocks in the chip. Figure 7.1(b) shows the schematic of the proposed charge injection based active decoupling circuit. A small fraction of the total pads available for Vdd are allocated to Vdd_a such that the total number of supply pads is kept constant.

$$N1_{vdd} \cdot L1_{vdd} = N1_{vdd} \cdot L1'_{vdd} + N2_{vdd} \cdot L2_{vdd} \qquad \text{(EQ 7.1)}$$

where, $N1_{vdd}$ and $N2_{vdd}$ are the number of pads for Vdd and Vdd_a respectively; $L1_{vdd}$ and $L2_{vdd}$ are the inductances of Vdd and Vdd_a respectively.

Similarly, the total area of the passive decap, C_p, is reduced to incorporate C_a and the area overhead of the active circuit, ΔA. The control signals, S_{normal}, S_{un} and S_{ov} are generated by the undershoot/overshoot detectors to be described in Section 7.2.2.

$$C'_p = C_p + C_a + \Delta A \qquad \text{(EQ 7.2)}$$

The solid curve in Figure 7.2 shows the simulated transient response of the unregulated supply distribution model in Figure 7.1(a). The shaded regions in the graph indicate regions where the supply voltage exceeds the designer-specified margins, V_H and V_L. C_a is connected between Vdd and Vss and acts as a normal passive decap when the supply voltage is within safe bounds, V_H and V_L. When a supply drop below V_L is detected, transistor T_0 in Figure 7.1(b) is turned off and T_1 ramps up the negative terminal of C_a from 0 to

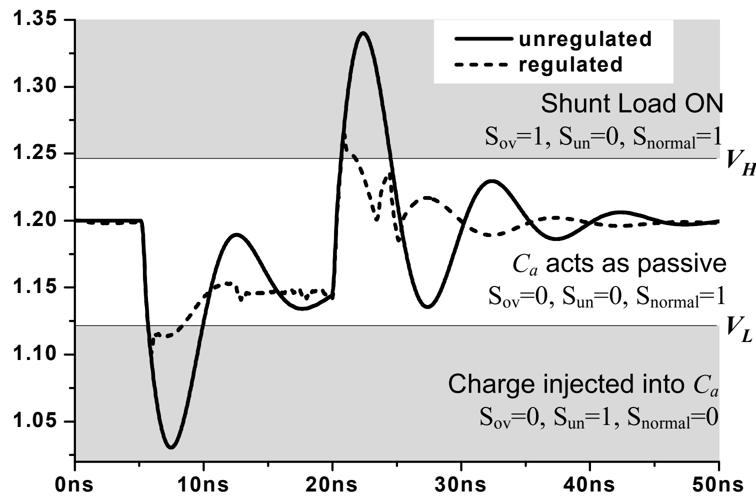


Figure 7.2. Simulated unregulated and regulated supply waveforms and safety bounds

Vdd_a, injecting a charge C_aVdd_a into the regular power grid. To prevent excessive overshoots, a shunt load T_2 is turned on while C_a is simultaneously recharged. The dashed curve in Figure 7.2 shows the simulated regulated supply waveform, confined within the safety bounds. For a supply voltage regulation tolerance of kVdd, the decap amplification factor, G, is (0.5+1/k) in the proposed circuit. A typical voltage regulation tolerance of k=10% yields a decap amplification factor of 10.5X which is significantly greater than the one in the circuit proposed in [4]. The next sub-section describes the generation of control signals S_{normal}, S_{un} and S_{ov}.

7.2.2 Supply-noise undershoot and overshoot detection

One of the key concerns in overshoot/undershoot regulation is detection speed. Analog detection techniques are either slow or consume a large amount of quiescent current. Therefore, we propose all-digital detectors with a simulated worst-case response time of 330ps. A ring oscillator based sampling clock generator (Figure 7.3) is used to generate a 6-phase (ϕ1-ϕ6) 3.3GHz clock. Supply waveforms are sampled at the rising edge of each phase of the clock, resulting in an effective sampling rate of 20GS/sec. This sampling rate is sufficient to capture ~100 samples before maximum first supply droop.

Two sets of clocked comparator banks [81] (one each for undershoot and overshoot detection) are used to sample the supply noise at a high frequency (Figure 7.4). A level-shifter first translates Vdd and Vss noise to a common-mode reference, Vref, of 600mV. The RC time constant of the level shifter is 10ns. Our prototype implementation uses an external reference supply for Vref. Any noise in Vref affects the common-mode voltage of

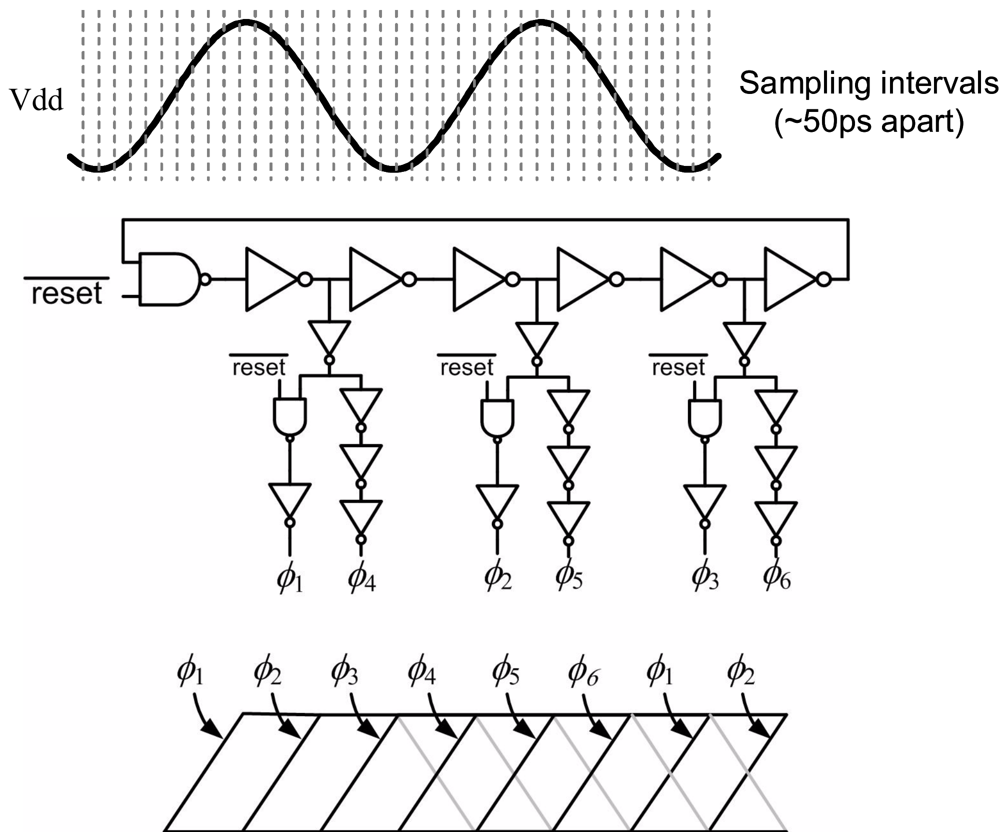


Figure 7.3. Sampling clock generator

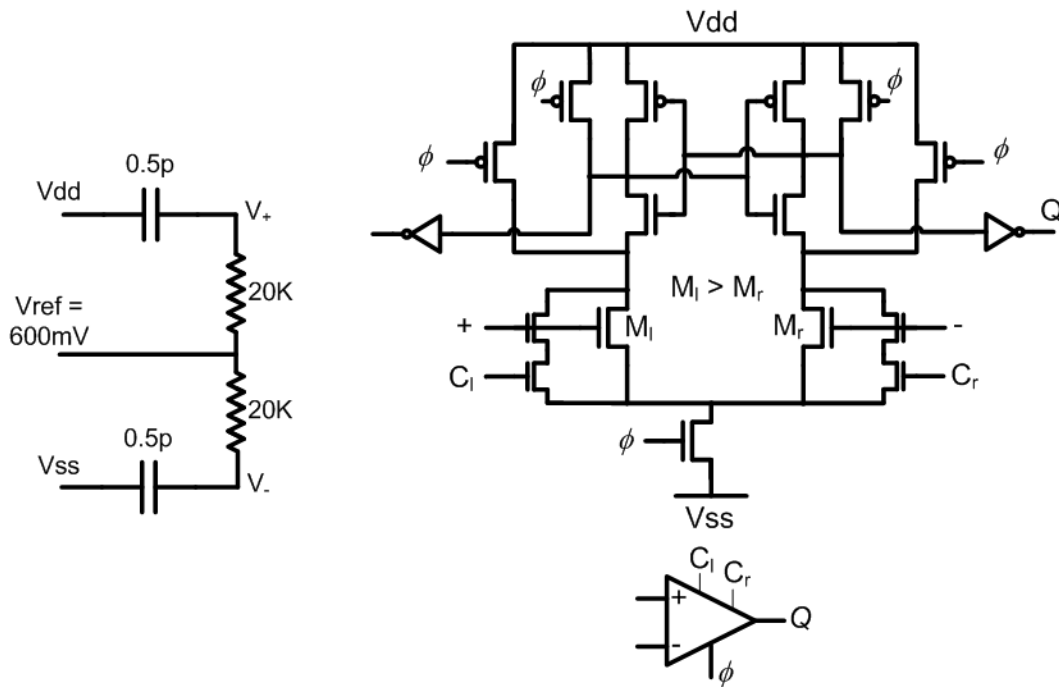


Figure 7.4. Vdd/Vss level shifter and clocked-comparator schematics

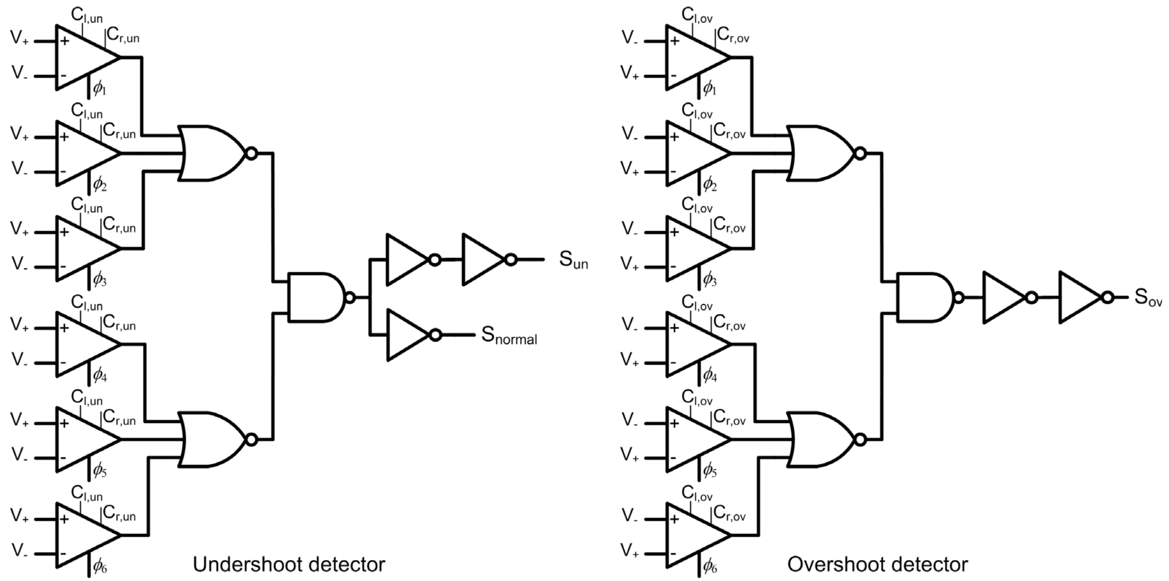


Figure 7.5. Comparator banks and generation of normal, undershoot and overshoot signals

V_+ and V_- only, making the differential sensing immune to the noise. Therefore, Vref can be easily generated on-chip using resistor dividers. The translated waveforms, V_+ and V_- are differentially sensed by 2 banks of 6 clocked comparators. The output of each comparator must resolve within 150ps of the rising edge of its clock. During simulations, the worst case delay of the comparator was found to be 80ps for an input differential of 1mV. Transistors, M_l and M_r in each comparator are skewed ($M_l > M_r$) to create a switching threshold between V_+ and V_-. This switching threshold automatically creates the safety margins, V_H and V_L without the need of any external reference supplies. Each comparator consists of calibration voltages $C_{l,un}$, $C_{r,un}$ ($C_{l,ov}$, $C_{r,ov}$), which provide post-silicon tuning of V_H (V_L), if required. Calibration voltages are shared among all the comparators in a bank. Outputs from comparator banks are ORed together and buffered to generate S_{normal}, S_{un} and S_{ov} control signals (Figure 7.5). The latency of control signal generation was

found to be 330ps in simulations. The frequency of first supply drop, which is set by the resonance frequency of package inductance and on-die decap, is typically in the range of 50MHz-250MHz [11]. Thus, the undershoot/overshoot detection delay of 330ps is small compared to the period of the first droop and enables the proposed circuit to react quickly to any sudden drops or surges in supply voltage.

7.2.3 Synthetic load-current generator and analog drop detector

This sub-section describes the artificial load-current generator used for injecting noise into the power grid and an analog supply drop measurement circuit. A configurable load-current generator with variable duty cycle and period was implemented using an array of variable-width transistors connected between Vdd and Vss as shown in Figure 7.6. The low and the high-periods (T_{high} and T_{low}) of the load-current are independently tunable from 500ps to 2μs. Two 10-bit counters, enabled by an externally fed *trigger* signal, set

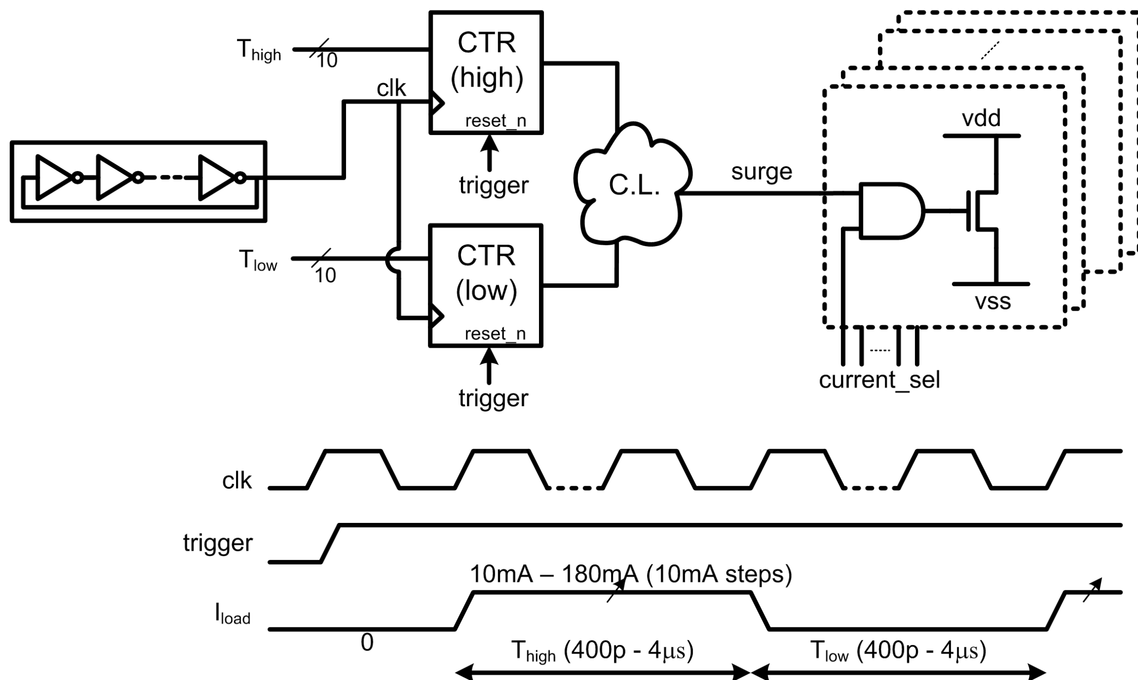


Figure 7.6. Synthetic load-current generator

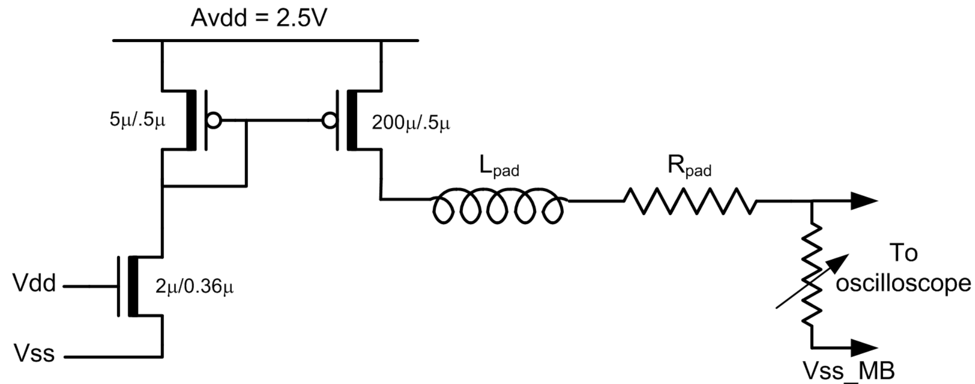
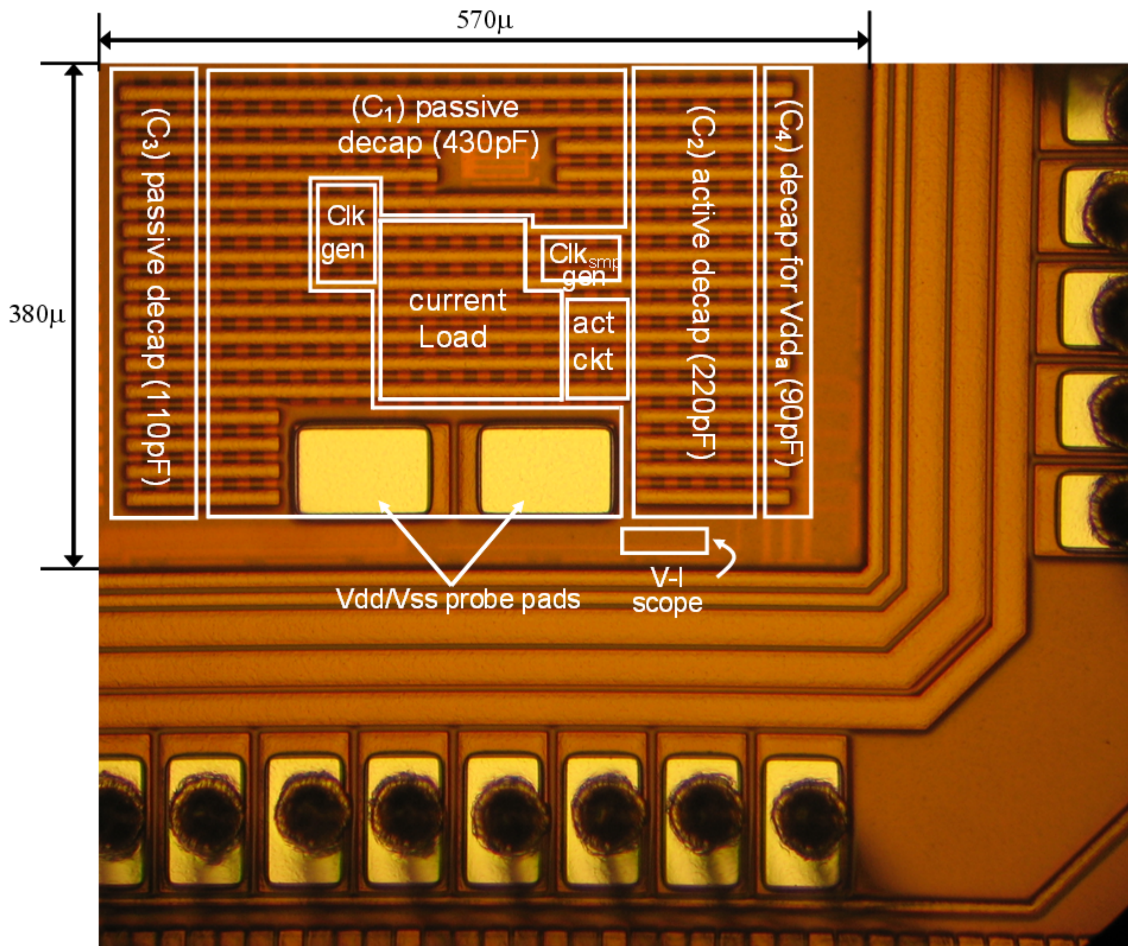


Figure 7.7. V-I converter-based drop-detector circuit

the high and the low period of the load-current waveform. The *trigger* signal invokes the current generation and starts the test-case. A 5-bit current-sel signal configures the peak magnitude of the load-current, which is configurable from 10mA to 180mA in steps of 10mA.

A V-I converter-based drop detector circuit [54] was implemented to measure the on-chip supply noise. The detector consists of a high conductance (Gm) transistor to convert the supply voltage variations into current. The current is amplified using a current mirror and is transmitted out of the chip using a transmission line. The transmission line is resistively terminated on the PCB. All the transistors in the circuit are designed to be larger than nominal oxide thickness. This enables capture of both undershoots and overshoots and also improves the gain of the circuit, which can now employ a higher than nominal supply voltage. Current-based sensing is particularly attractive due to its simplicity, robustness to coupling noise and bandwidth-limited nature of the I/O pads. The prototype was also designed to have Vdd and Vss probe-pads which were used to verify the on-die supply noise measurements.

The main drawback of the above V-I converter-based circuit is its power consumption, which can be of the order of mW. We propose a fully-digital on-chip oscilloscope for supply noise which consumes considerably lower power and has improved accuracy. The on-chip oscilloscope will be described later in Section 7.5. The next sub-section provides the details of the prototype implementation of the charge injection based technique.



Test-case	Pad Allocation				Decap Allocation			
	Vdd	Vss	Vdd_a	Total	Passive (C_p)	Active (C_a)	Vdd_a decap	Total
Unregulated	3	3	0	6	760pF ($C_1+C_2+C_3$)	0	-	760pF
Regulated	2	3	1	6	430pF (C_1)	220pF (C_2)	90pF (C_4)	740pF

Figure 7.8. Die micrograph and implementation details of the test chip

7.2.4 Prototype-implementation details

The proposed charge injection based technique was implemented in a test-chip, fabricated in a 0.13μm, 1.2V triple-well CMOS process. Figure 7.8 shows the die-micrograph and implementation details of the prototype. The test chip consists of unregulated and regulated test-cases, which were implemented for an iso-area, iso-pad comparison. The unregulated case utilized 3 Vdd pads, 3 Vss pads and 760pF of C_p. For the regulated case, 1 pad was re-allocated to Vdd_a resulting in 2 regular Vdd and 3 Vss pads. Simulations showed an optimal Vdd_a pad allocation of 6% of the total pads, which, however, could not be implemented in this pad-limited design. To be conservative, a small amount of decap was allocated for Vdd_a to prevent excessive ringing. The values of C'_p, C_a and decap for Vdd_a were 430pF, 220pF and 90pF, respectively. The active circuit area, which includes the sampling-clock generator, undershoot/overshoot detectors and switches T_0, T_1, T_2, was equivalent to the area of 10pF decap, which corresponds to 1.54% of the total decap area in the unregulated test case. Figure 7.9 shows the area breakdown of different decap

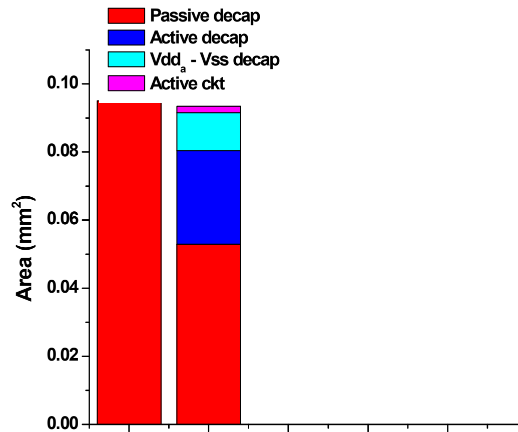


Figure 7.9. Area breakdown in unregulated and regulated test-cases

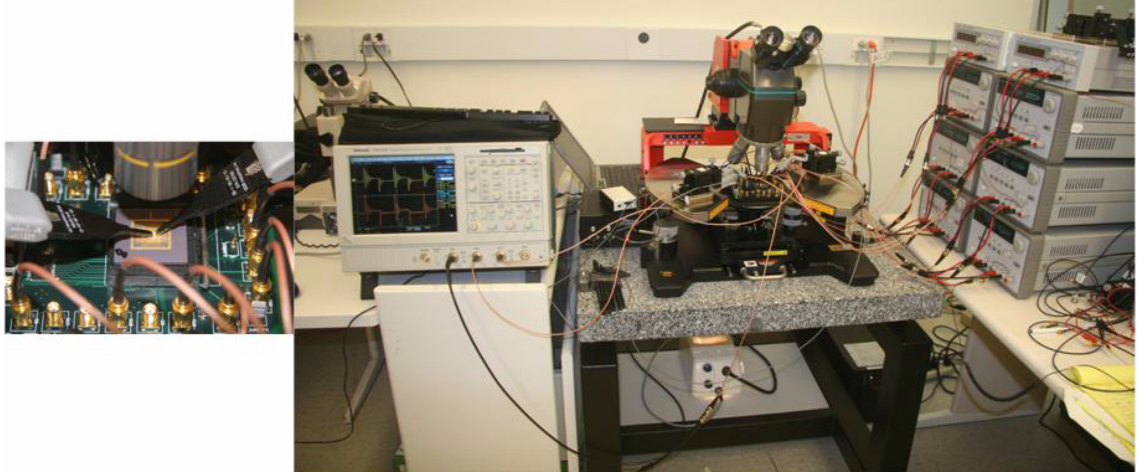


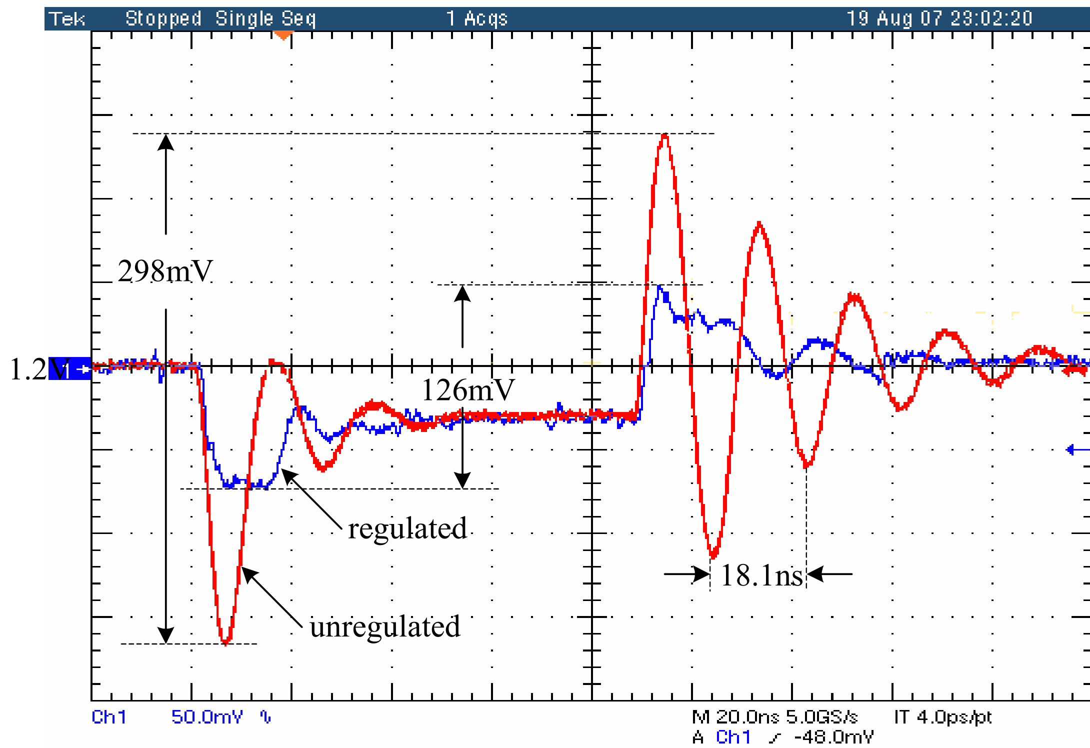
Figure 7.10. Measurement setup

banks and the active circuit in the unregulated and regulated test case implementations. The next sub-section presents the test chip measurement results.

7.2.5 Measurement results

Figure 7.10 shows the measurement setup of the test chip on a probstation. The supply noise was measured using the V-I converter based drop detector and the measurements were verified with direct differential on-chip probing. The results were also validated using an all-digital on-chip oscilloscope which will be presented later in Section 7.5.

Figure 7.11 shows a comparison of the measured on-die supply noise with and without supply regulation for an average die. In Figure 7.11(top), the excitation load-current is ramped up from 0 to 40mA and back to 0. This load-current is representative of the wake-up and turn-off of a power/clock-gated module [11]. The unregulated peak-to-peak supply drop was found to be 298mV. Active regulation reduces the peak-to-peak supply drop from 298mV to 126mV, which is an improvement of 57%. It is important to note that the steady-state IR-drop remains the same in both the unregulated and the regulated case. Fig-



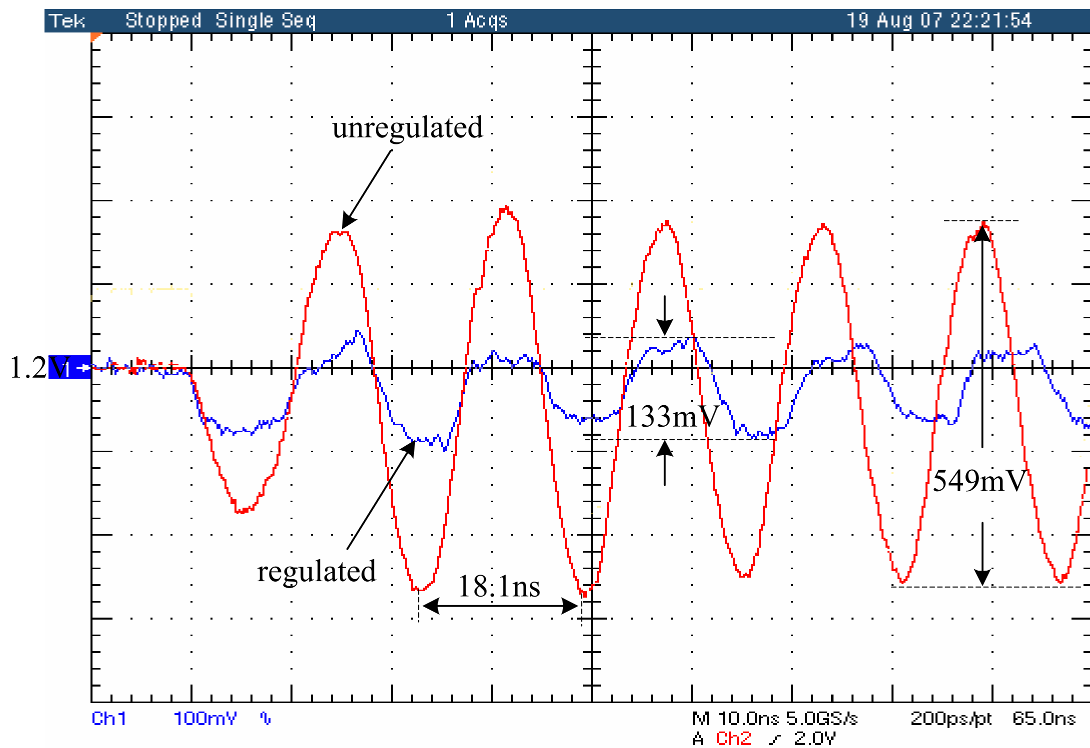
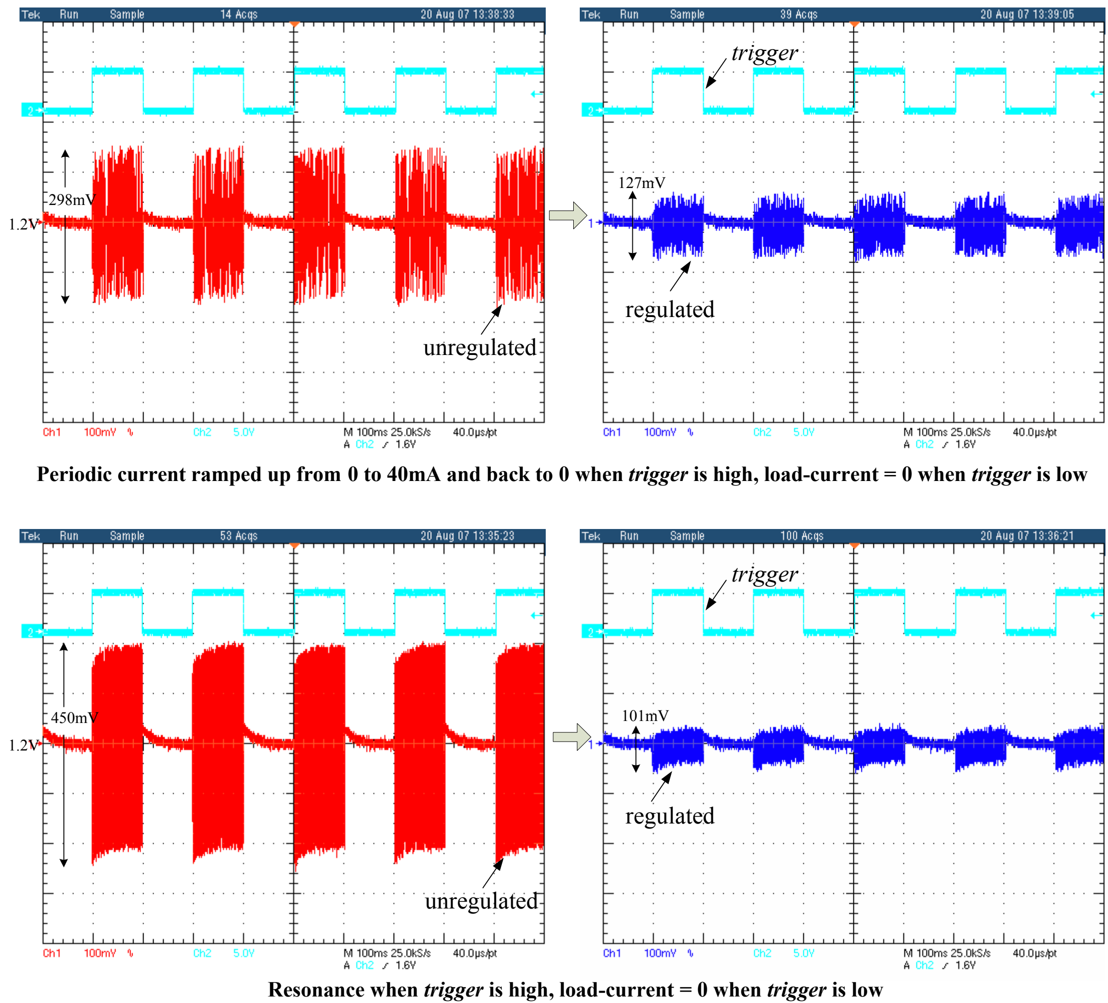


Figure 7.11. Measured unregulated and regulated supply noise waveforms for ramp load and during resonance

ure 7.11(bottom) shows the resonance build-up when a periodic load-current is applied at the resonance frequency. The resonance frequency (caused by package inductance and on-

die decap) of the power distribution network was found to be at 55.25MHz. The proposed circuit reduces the peak-to-peak supply drop from 549mV to 133mV, which is an improvement of 75%. Figure 7.12 shows the zoomed-out snapshot of the measured regulated and unregulated on-die supply noise. Also shown is the *trigger* signal, which invokes the test cases. A high *trigger* signal repetitively generates the desired load-current waveform and creates noise in the otherwise steady supply voltage. The results demonstrate effective regulation of both undershoot and overshoot supply noise during any rapid current transients as well as during resonance.



Periodic current ramped up from 0 to 40mA and back to 0 when *trigger* is high, load-current = 0 when *trigger* is low

Resonance when *trigger* is high, load-current = 0 when *trigger* is low

Figure 7.12. Measured unregulated and regulated supply waveforms as a function of *trigger* signal

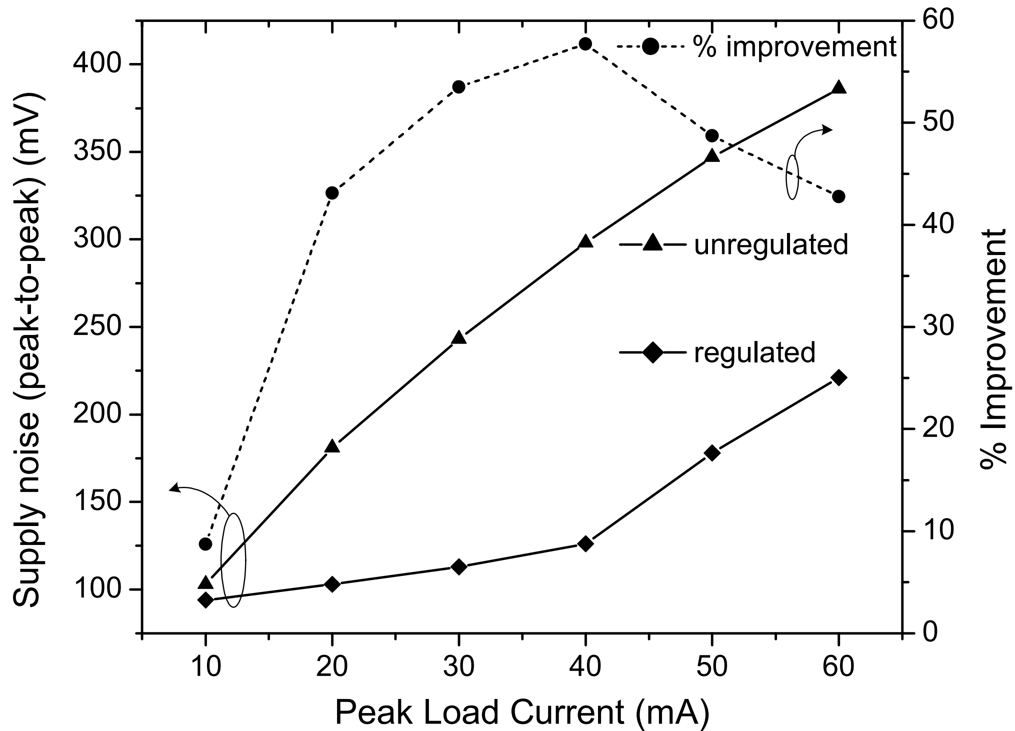


Figure 7.13. Measured unregulated and regulated peak-to-peak supply noise for varying peak load-currents

Measurements were performed for different load-currents of varying frequency and peak magnitude. Figure 7.13 plots the measured worst drop and improvement as a function of peak load-current (I_{max}) for one die. At very low load-currents, the worst supply drop or bounce is closer to the safety margins, V_H and V_L. As a result, the regulated supply drop is close to the regulated drop and the percentage improvement is lower. As the peak current is increased, the active circuit has to inject more charge to compensate for an increased supply drop, resulting in increased percentage improvement. As the load-current is increased beyond a certain extent, the charge injection gets limited by the size of the active decap and the switch though which it is driven (T_1 in Figure 7.1). Hence, there exists an optimal regulation point which is a function of the size of the active decap and

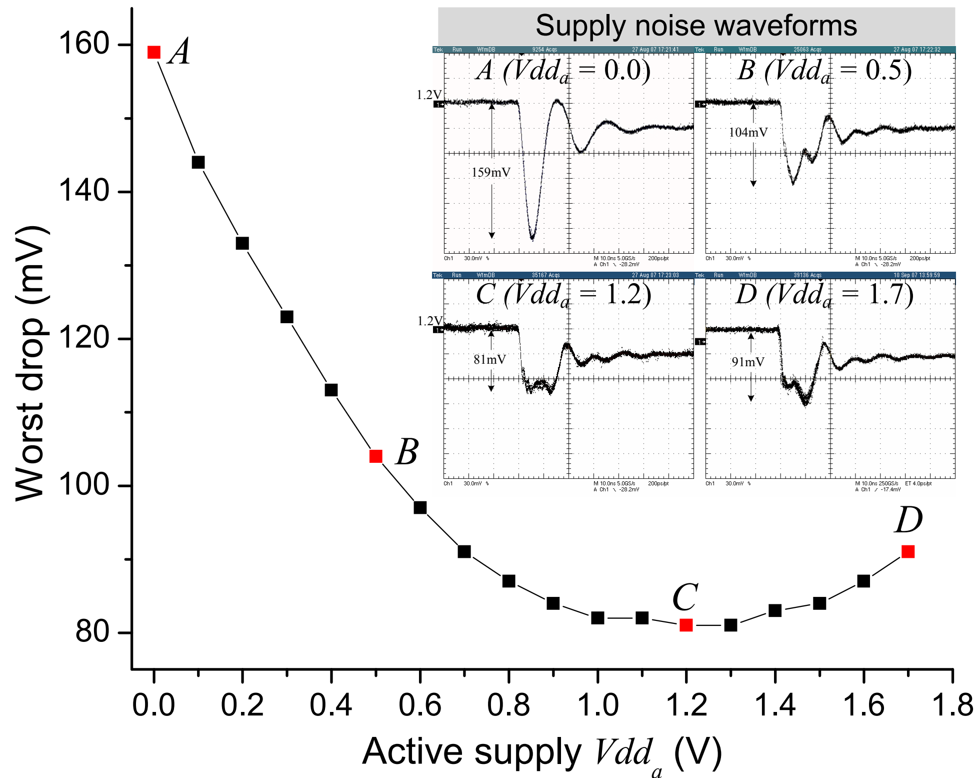


Figure 7.14. Measured worst supply drop as a function of the active supply

the switching transistor T_1. Thus, it is imperative to extract the worst-case current profile during design time and size the elements of the active circuit accordingly. In this implementation, the best improvement of 57.7% was measured for an I_{max} of 40mA.

Figure 7.14 shows the measured regulated worst drop as a function of Vdd_a when a step load-current of 40mA peak amplitude is applied. Also shown are the supply voltage snapshots for different values Vdd_a (Vdd_a = 0V, 0.5V, 1.2V and 1.7V). The amount of charge injected into the regular power grid increases with Vdd_a and the worst-case supply drop reduces as Vdd_a is increased to 1.2V. The regulated supply voltage exhibits a second dip immediately after the first one, which is attributed to the recharging of C_a once Vdd is above the undershoot threshold V_L. As explained earlier, the magnitude of the first dip is

dependent on the sizing of C_a and switching transistor T_1 in Figure 7.1. Also, the injected charge increases with Vdd_a, resulting in a reduction in the first dip. The recharging of C_a depends on the size of switching transistor T_0. An increase in the size of T_0 causes faster charging of C_a resulting in an increased second dip. Correct sizing of various elements in the active circuit is, therefore, essential for efficient supply regulation. As Vdd_a is increased above 1.2V, the second dip becomes more prominent, increasing the worst drop. An optimal regulation point is obtained when Vdd_a is at the nominal supply voltage of 1.2V. Since the supply regulation curve is flatter around the optimal operating point of 1.2V, the efficiency of the proposed circuit does not degrade significantly when the operating supply voltage is reduced for power reduction as in DVS systems.

Figure 7.15 shows the frequency dependence of the worst supply drop with and without active supply regulation. Also shown in the supply drop undershoot threshold, V_L which

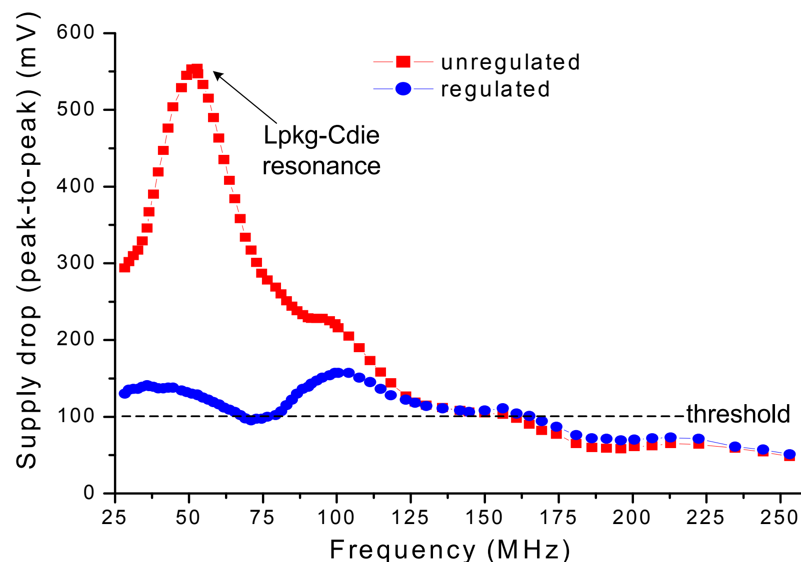


Figure 7.15. Comparison of measured frequency responses with and without active supply regulation

is set at 100mV. The regulated supply drop is less than the unregulated one whenever the drop exceeds the undershoot threshold. Also, it can be inferred that the impedance of the supply network improves considerably near its resonance frequency. Since the regulated test-case has fewer regular pads and lower amount of passive decap (Section 7.2.4), the regulated supply drop is slightly larger than that in the unregulated case when the supply drop is less than the undershoot threshold.

As shown in Figure 7.5, the clocked comparator banks are provided with calibration voltages ($C_{l,un}$, $C_{r,un}$, $C_{l,ov}$ and $C_{r,ov}$) for extra post-silicon tuning of the undershoot and overshoot thresholds, if required. Figure 7.16 shows the measured worst regulated supply drop as a function of calibration voltages. In this experiment, the calibration voltages $C_{l,un}$ and $C_{l,ov}$ are set to 0V ($C_{l,un} = C_{l,ov} = C_l = 0$) while $C_{r,un}$ and $C_{r,ov}$ are each equally varied

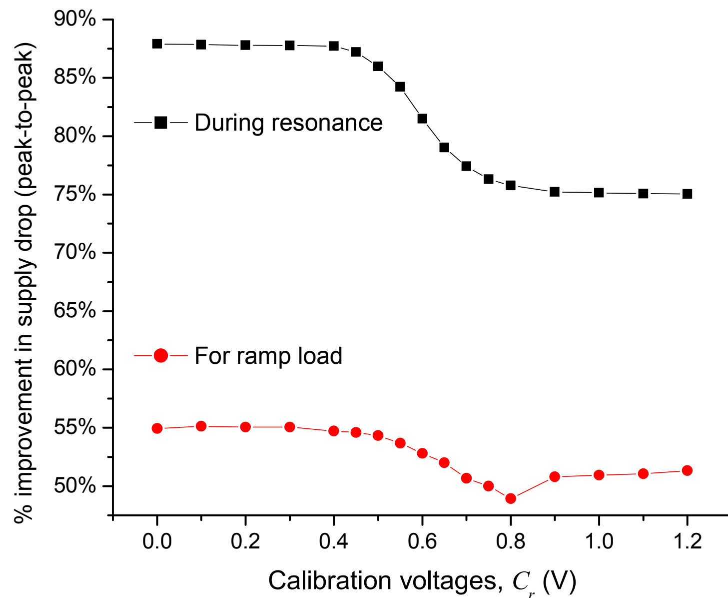


Figure 7.16. Measured worst regulated supply drop as a function of calibration voltages

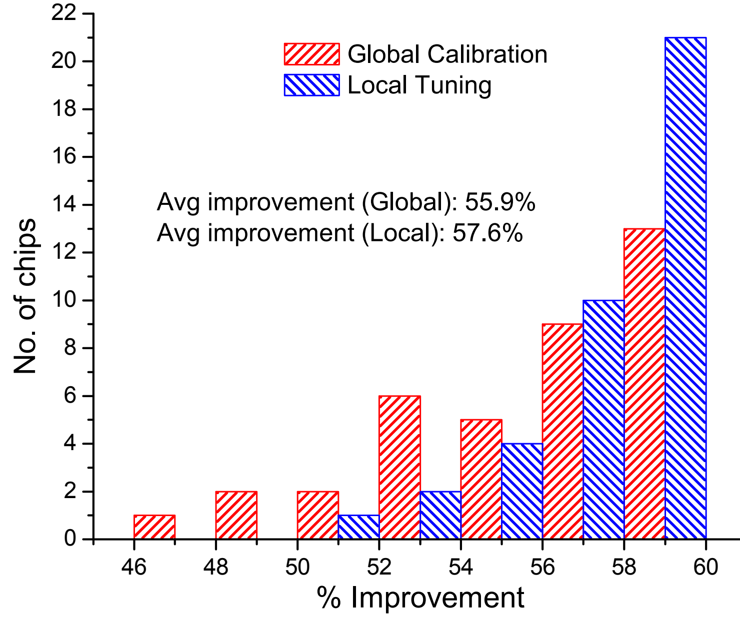


Figure 7.17. Statistical analysis across 38 chips showing the affect of a global calibration vs. individual die-tuning

from 0 to 1.2V ($C_{r,un} = C_{r,ov} = C_r$). An increase in C_r results in an increase in threshold V_H and V_L by the same amount. Since the active circuit starts to inject charge only when the supply drop exceeds the thresholds, the percentage improvement reduces with increasing C_r as shown in Figure 7.16.

A statistical analysis (Figure 7.17) was performed on 38 chips for step current-loads, to evaluate the effect of a single global setting of C_l and C_r for all chips as opposed to individual calibration tuning of each die. The minimum, maximum and average drop improvements for a global calibration setting were 47.1%, 59.6% and 55.9%, respectively. When each die is tuned for its best calibration, the minimum, maximum and average improvements increased to 51%, 59.7% and 57.6% respectively. Although individual die calibration shifts the low performance chips to the right, the improvements are only marginal and we concluded that the overhead of individual die calibration can be avoided.

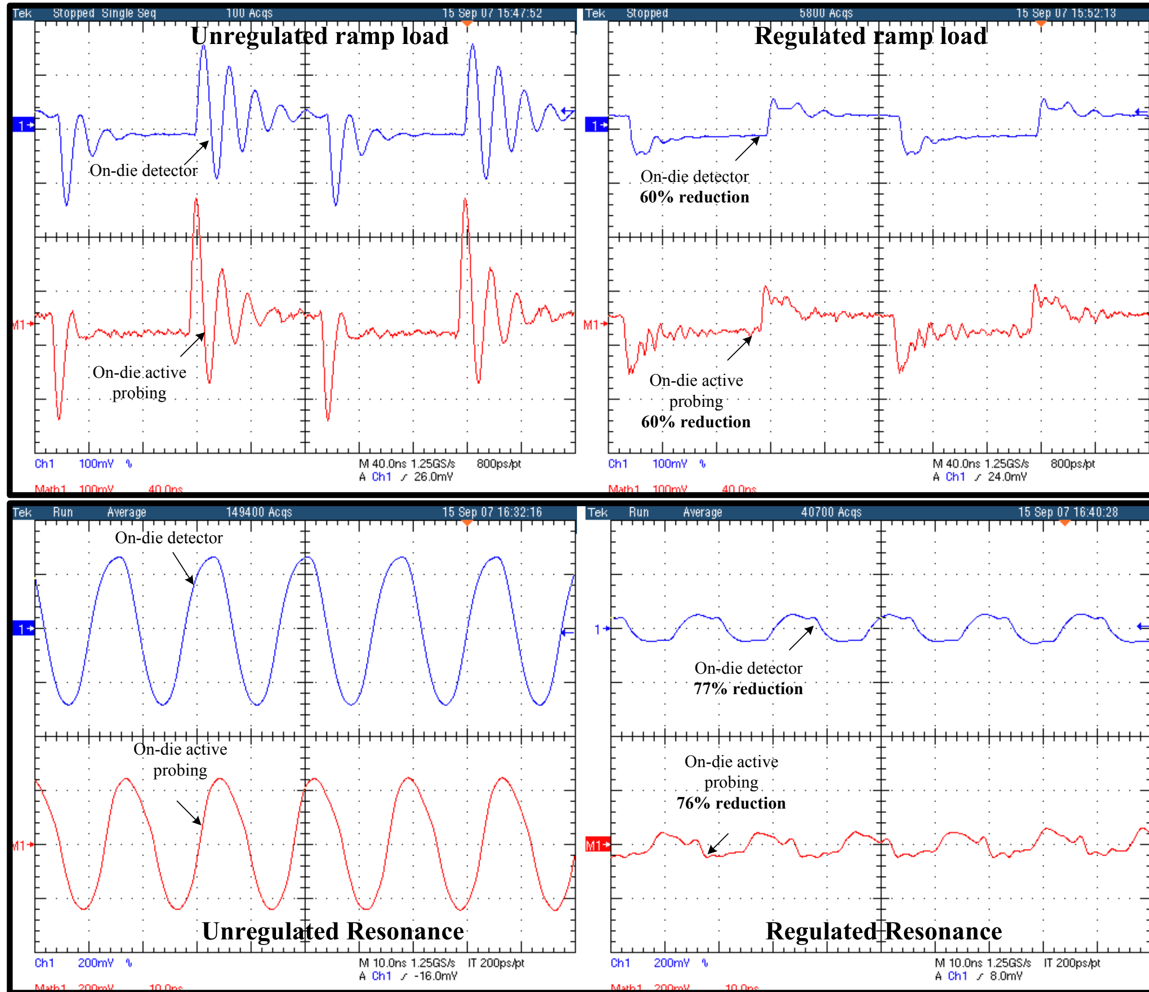


Figure 7.18. Measured supply waveforms with on-chip V-I detector and active on-chip probing

In Figure 7.18, a comparison of unregulated and regulated supply voltage waveforms is shown, measured with the on-chip V-I drop detector and through active on-chip probing of the Vdd and Vss probe pads. The top-layer Vdd and Vss metal layers are connected to the probe pads, which are differentially probed using active probes. The experiment was performed for several load-current configurations. Figure 7.18 shows that the supply voltage waveforms are very similar and the percentage improvements are almost identical. The V-I drop detector, although simplistic in design and fairly accurate, however, consumes about 25mW of power. The power overhead of the active circuit, which includes the sam-

pling clock generator, undershoot/overshoot detectors and switches T_0, T_1, T_2 was measured to be less than 1% of the peak power consumption of 48mW. The performance of the proposed circuit was also validated using an all-digital on-chip oscilloscope which consumes considerably lower power than the V-I drop detector. The details of the on-chip oscilloscope and measurement results will be presented later in Section 7.5. Next, we present the high-voltage charge pump based and high-voltage shunt supply based active supply regulation techniques.

7.3 High-voltage charge-pump-based active decoupling circuit

Figure 7.19 shows the schematic of the proposed high-voltage charge pump based supply regulation technique. As in Section 7.2.1, some of the pads used for the regular supply are allocated for a high-voltage supply, Vdd_H. Similarly, some of the passive decap area is allocated for the active decap, C_H, which forms the charge pump. When the supply voltage is between the pre-specified undershoot and overshoot thresholds, V_H and V_L, C_H is connected between Vdd_H and Vss. Whenever a supply drop greater than V_L is detected, the positive terminal of C_H is switched from Vdd_H to Vdd, dumping a charge of $(Vdd_H - Vdd)C_H$ into the Vdd power grid. For a Vdd_H = 2Vdd, the amount of charge dumped is the same as the charge injected in the technique proposed in Section 7.2. In case of excessive overshoots, an artificial load, connected between Vdd and Vss is turned on. Since the proposed circuit uses a high-voltage supply, the switching transistors and C_H need to be high-voltage tolerant to avoid any reliability issues. Therefore, T_0, T_1, T_2 and C_H are implemented with thick-ox transistors. This results in larger area overhead of T_0, T_1, T_2 and

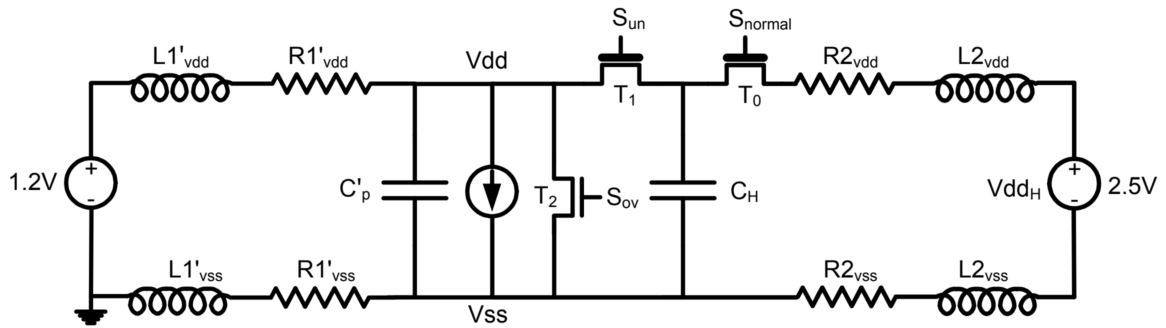


Figure 7.19. Proposed high-voltage charge-pump-based regulation circuit technique

lower area efficiency of decap C_H as compared to the circuit proposed in Section 7.2. However, this proposed circuit has C_H directly connected between Vdd and Vss in case of undershoots, creating a tight charging-discharging loop. This tight charging-discharging loop causes an identical amount of charge to be pumped into the Vdd grid and out of the Vss grid. This ensures that there is no DC voltage shift in the on-die Vss voltage, which may cause reliability concerns during data transfers between the chip and the motherboard.

The control signals, S_{normal}, S_{un} and S_{ov} are generated by the undershoot and overshoot detection circuits which are very similar to the ones shown in Section 7.2.2. The only difference is in the OR gates and the output buffers which use thick-ox transistors and powered by a high voltage supply of Vdd_H. The level-conversion of the clocked comparator outputs from nominal-voltage to high-voltage is performed at the OR gates which are implemented in DCVS logic. The level-conversion and usage of thick-ox transistors degrades undershoot and overshoot detection delay. A synthetic load-current generator

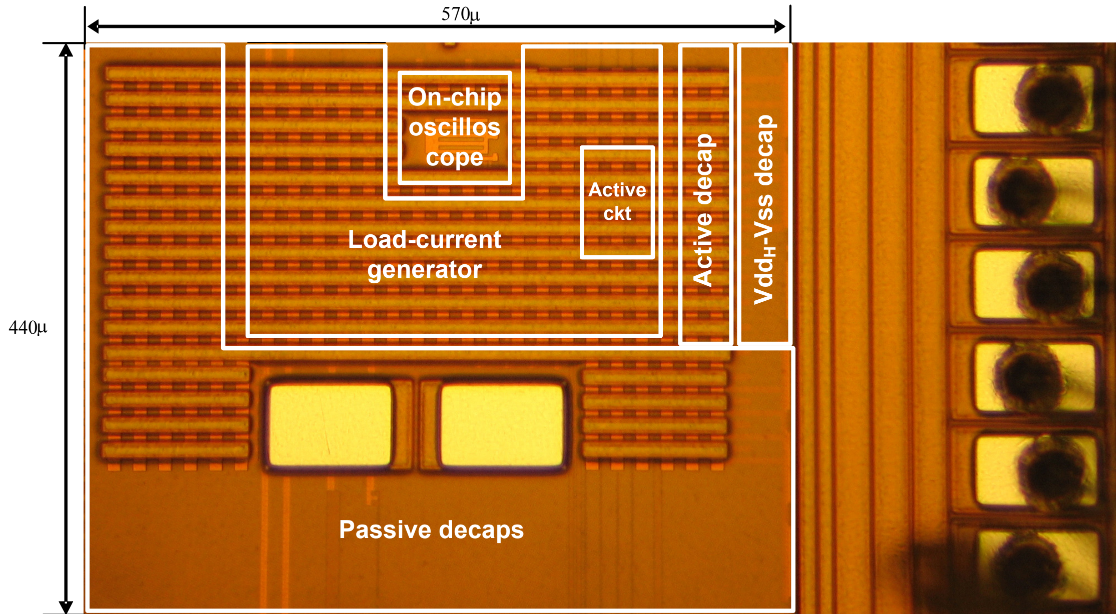


Figure 7.20. Die micrograph and implementation details of the test chip

and a V-I converter based drop detector (Section 7.2.3) is used for the creation and measurement of supply noise respectively.

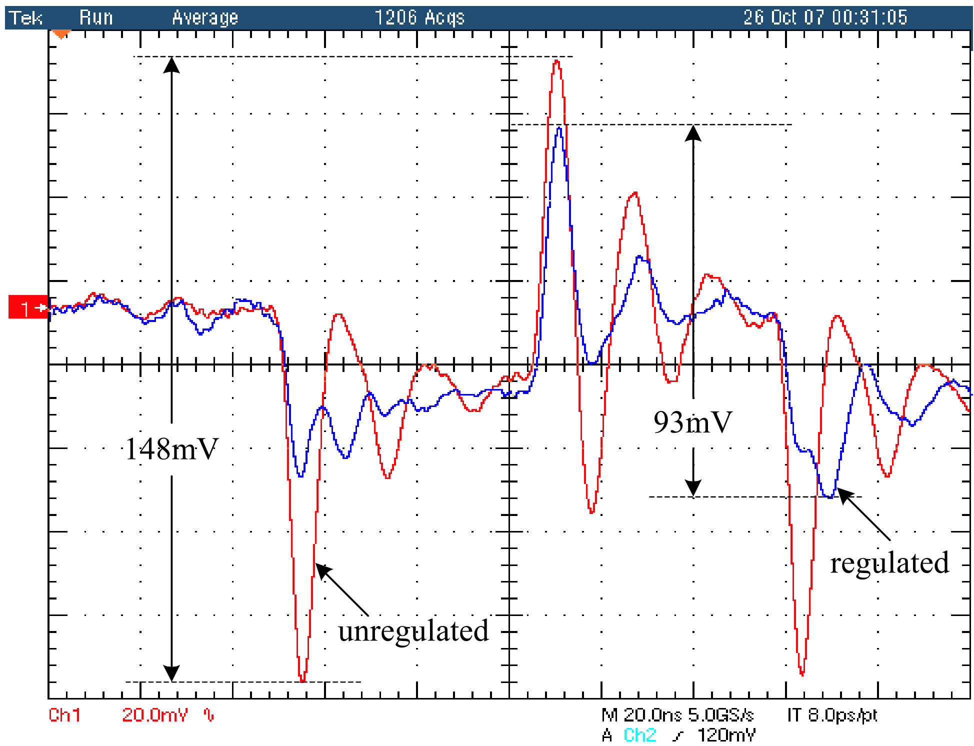
7.3.1 Prototype-implementation details

The proposed high-voltage charge pump based supply regulation technique was implemented in a test-chip, fabricated in a 0.13µm, 1.2V triple-well CMOS process. Figure 7.20 shows the die-micrograph and implementation details of the prototype. The test chip consists of unregulated and regulated test-cases, which were implemented for an iso-area, iso-pad comparison. The unregulated case utilized 3 Vdd pads, 3 Vss pads and 670pF of C_p. For the regulated case, 1 pad was re-allocated to Vdd_H resulting in 2 regular Vdd and 3 Vss pads. The values of C'_p and C_H in the regulated test case were 356pF and 113pF respectively. To prevent excessive ringing in Vdd_H, a decap of 52pF was added between Vdd_H and Vss. The next sub-section discusses the measurement results on the prototype.

7.3.2 Measurement results

The proposed high-voltage charge pump based active regulation technique was implemented in a prototype designed in a 0.13µm triple-well CMOS process. The prototype consists of the load-current generator and the V-I converter based supply drop detector described in Section 7.2.3. Figure 7.21(top) shows the measured supply waveforms with and without active regulation when the load-current is ramped up from 0A to 40mA and then back to 0. The worst-case peak-to-peak supply variation was measured to be 148mV in the unregulated case. Through active regulation, the overall peak-to-peak supply variation is reduced to 93mV showing an improvement of 37%. Figure 7.21(bottom) shows the measured supply waveforms during resonance, which is created by the periodically switching load-current. The proposed regulation technique reduces the peak-to-peak supply voltage fluctuation from 290mV to 164mV during resonance, resulting in an improvement of 43%. Note that the proposed circuit is able to suppress the supply noise undershoots much more effectively as compared to attenuating the overshoots.

In the next section, we present the proposed high-voltage shunt supply based supply voltage regulation technique.



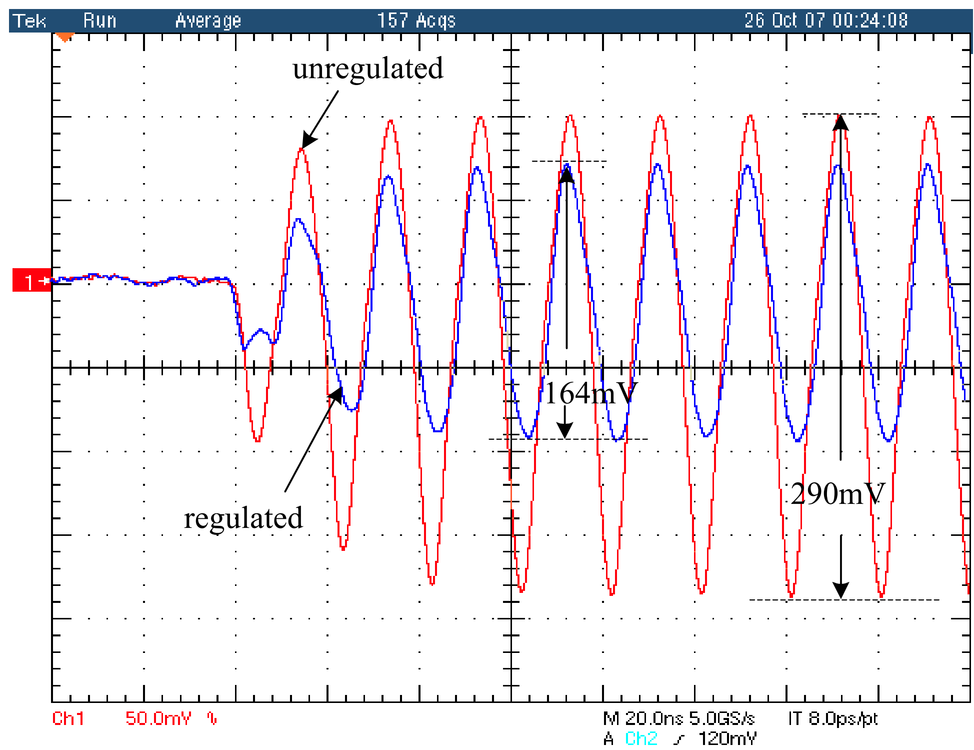


Figure 7.21. Measured unregulated and regulated supply noise waveforms for ramp load and during resonance

7.4 High-voltage shunt-supply-based active decoupling circuit

The efficacy of the regulation techniques proposed so far in Section 7.2 and Section 7.3 is limited by the amount of charge stored in the active decap banks. The amount of charge stored in the active decap must be sufficient to compensate for the supply drop created by the largest possible load-current. We, now propose a supply regulation technique which eliminates the need of any active charge storing elements. A similar technique has recently been proposed in [56] which uses a software-generated control signal to switch on the regulator. The regulator in [56] suppresses excessive supply undershoots, caused by sudden wake-up of power-gated logic blocks. We extend the technique proposed in [56] with integrated supply undershoot and overshoot detectors which eliminates the need of any software-generated control signals. Additionally, the proposed circuit automatically detects and suppresses resonance in the power supply network.

Figure 7.22 shows the schematic of the proposed high-voltage shunt supply based supply regulation technique. The regular nominal supply grid is connected in shunt to a high-voltage supply whenever an undershoot is detected. An artificial load is turned on to prevent excessive overshoots. The control signals, S_{un} and S_{ov} are generated using under-

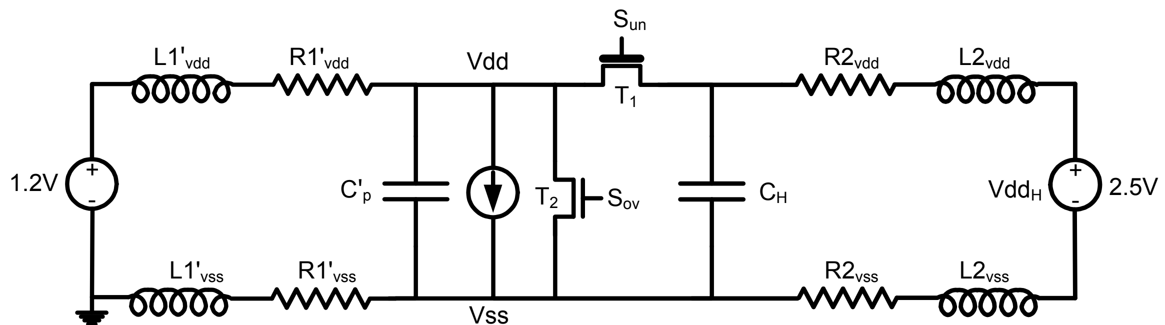


Figure 7.22. Schematic of the proposed high-voltage shunt-supply-based

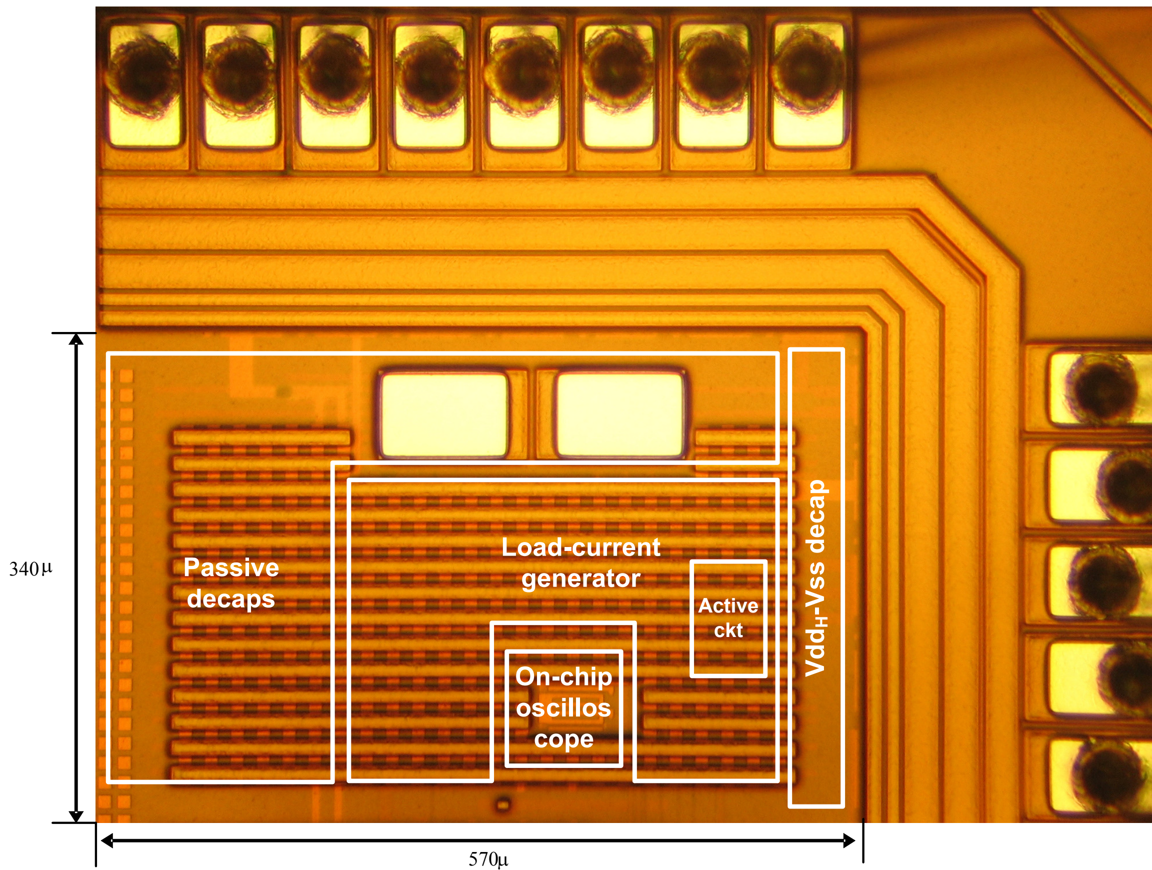


Figure 7.23. Die micrograph and implementation details of the test chip

shoot/overshoot detectors described in Section 7.3. T_1 and C_H use thick-ox devices to avoid reliability issues.

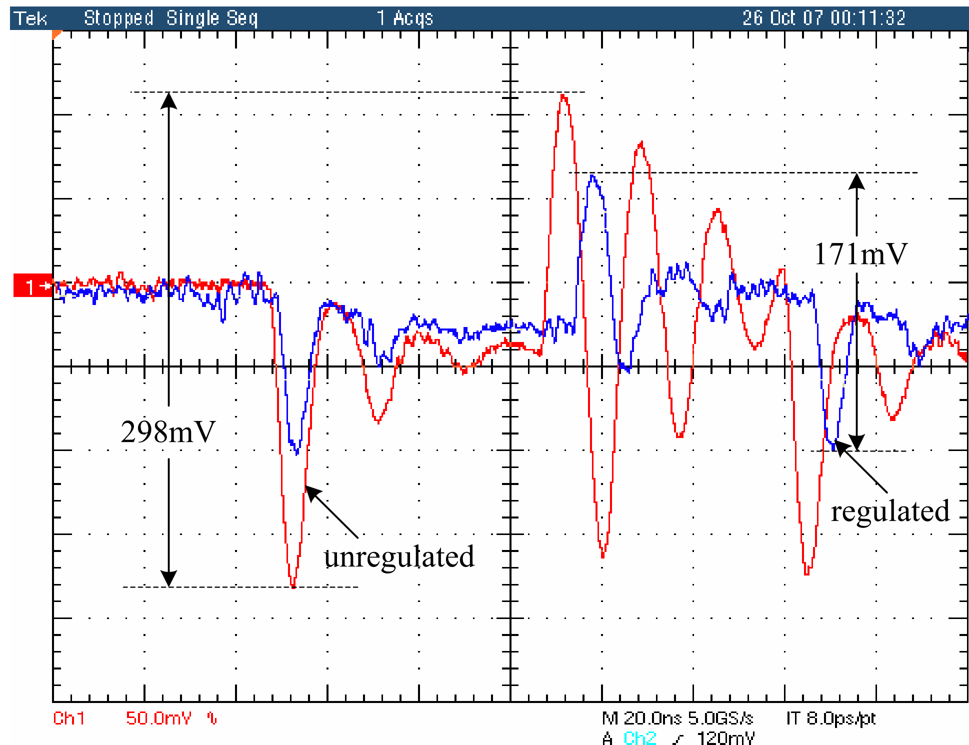
7.4.1 Prototype-implementation details

The proposed high-voltage shunt supply based regulation technique was implemented in a test-chip, fabricated in a 0.13µm, 1.2V triple-well CMOS process. Figure 7.23 shows the die-micrograph and implementation details of the prototype. The test chip consists of unregulated and regulated test-cases, which were implemented for an iso-area, iso-pad comparison. The unregulated case utilized 3 Vdd pads, 3 Vss pads and 570pF of C_p. For the regulated case, 1 pad was re-allocated to Vdd_H resulting in 2 regular Vdd and 3 Vss

pads. The values of C'_p and C_H in the regulated test case were 356pF and 113pF respectively. The next sub-section discusses the measurement results on the prototype.

7.4.2 Measurement results

The proposed high-voltage shunt supply based active regulation technique was implemented in a prototype designed in a 0.13µm triple-well CMOS process. The prototype consists of the load-current generator and the V-I converter based supply drop detector described in Section 7.2.3. Figure 7.24(top) shows the measured supply waveforms with and without active regulation when the load-current is ramped from 0A to 40mA and back. The worst-case peak-to-peak supply variation was measured to be 298mV in the unregulated case. Through active regulation, the peak-to-peak supply variation is reduced to 171mV showing an improvement of 43%. During resonance (Figure 7.24(bottom)), the peak-to-peak supply voltage fluctuation is reduced from 544mV to 299mV, resulting in an improvement of 45%.



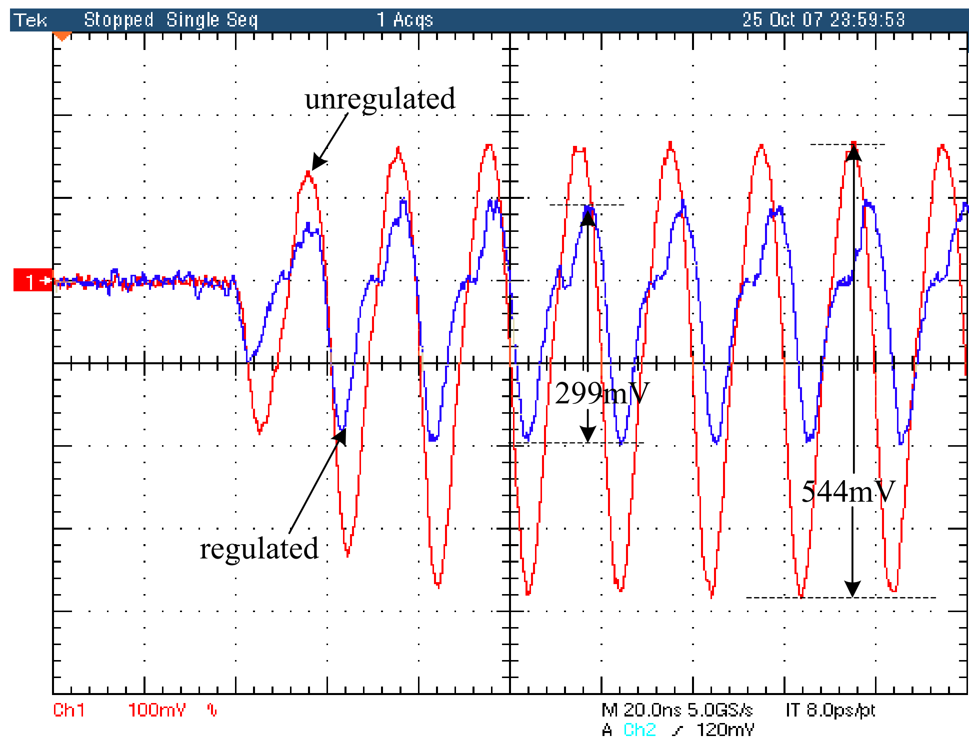


Figure 7.24. Measured unregulated and regulated supply noise waveforms for ramp load and during resonance

7.5 Digital on-chip oscilloscope for supply-noise measurement

The analog V-I converter based supply drop detector, described in Section 7.2.3 suffers from excessive power consumption (measurements on the test-chip show an average power consumption of 25mW). Also, non-linearity may adversely affect the accuracy of the detector in the presence of very large fluctuations in supply voltage. To address these concerns, we propose a power-efficient and accurate all-digital on-chip oscilloscope. Relying on the periodicity of the waveform to be measured, the proposed oscilloscope generates its time-shifted snapshots. Different time-shifted snapshots are then combined off-chip to reconstruct the original waveform. Unfortunately, the supply noise fluctuations are non-periodic in nature. Therefore, the device-under-test (DUT) executes the same test-case repeatedly, such that the overall supply noise waveform becomes periodic.

Figure 7.25 shows the schematic of the proposed on-chip oscilloscope. The oscilloscope consists of a coarse delay generator and fine delay line. The coarse delay generator con-

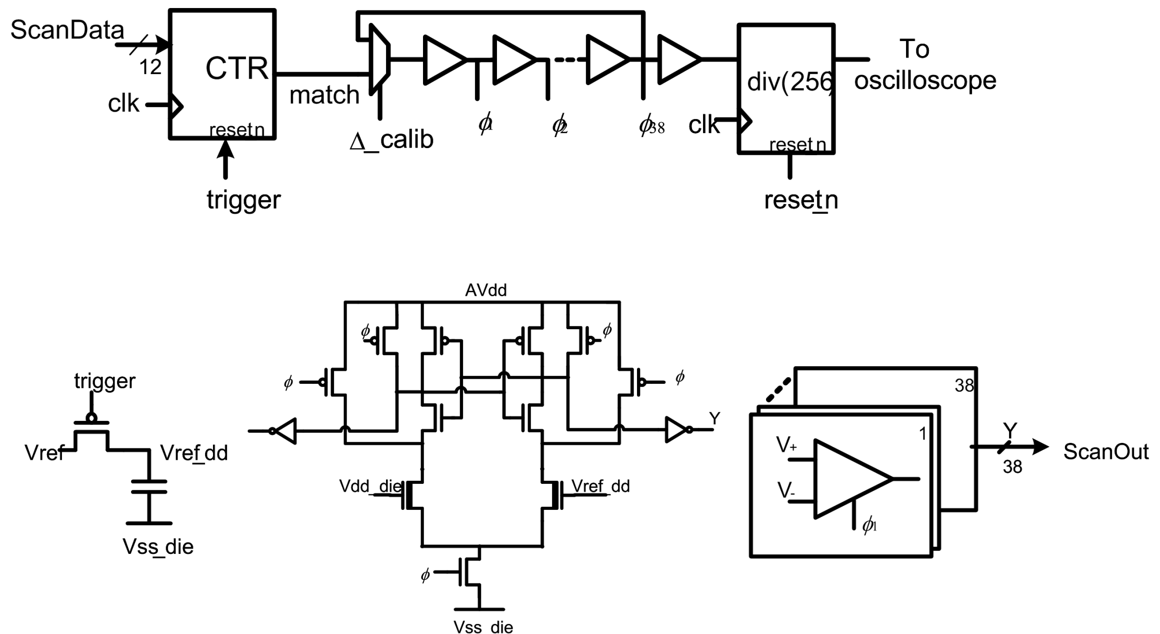


Figure 7.25. Schematic of the proposed on-chip oscilloscope for supply-noise measurement

sists of a counter running at the system clock (~1GHz) and asserts the *match* signal high whenever the counter meets a user-specified 12-bit value. The fine delay line is implemented using a chain of inverters connected in series, to generate time-shifted versions of the *match* signal. The minimum granularity that can be achieved in the fine delay line is equal to two inverter delays which was measured to be ~50ps. The proposed oscilloscopr is therefore capable of collecting samples at the rate of 20GSamples/sec.

A calibration control signal Δ_calib connects the fine delay line in feedback to form a ring oscillator in order to measure the approximate delay of each individual delay element. An external reference supply, Vref is provided as a reference for the comparison of the Vdd supply voltage (Vdd_die). Vref, which is supplied from off-chip, is translated to the common chip ground, Vss_die, using a low-pass RC network. The low-pass filter also helps in eliminating any high frequency noise in Vref. The capacitance in the RC network is realized using a MIM-cap (Metal-Insulator-Metal-cap) in order to minimize leakage. The measurement process is invoked by means of a control signal, *trigger*, which connects Vref to Vref_dd through its on-resistance. Our implementation employes 38 comparators clocked at time-shifted versions of the *match* signal (ϕ_1-ϕ_{38}). The output of the 38 clocked comparators is scanned-out as a thermometer code, indicating whether a particular sample is greater or less than the reference voltage. The same test-case is repeatedly executed on the DUT, with different reference voltages and coarse delay counter values. The complete supply noise waveform is constructed off-chip from the scanned-out thermometer code of each test-run.

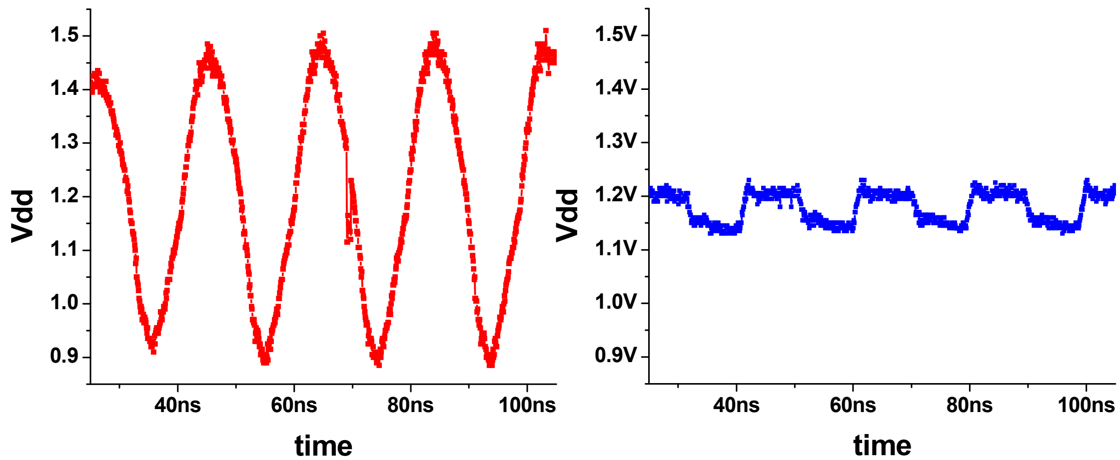


Figure 7.26. Unregulated (left) and regulated (right) resonance waveforms measured using the on-chip oscilloscope

7.5.1 Measurement Results

The proposed on-chip oscilloscope was integrated in the supply regulation prototypes of Section 7.2, Section 7.3 and Section 7.4, implemented in a 0.13µm triple-well CMOS process. The total layout area for the oscilloscope was 80µmx150µm. During measurements, the process-variation induced mismatch in the devices constituting each clocked comparator is calibrated out. Figure 7.26 shows the measured supply noise waveforms during resonance with and without the charge injection based supply regulation technique of Section 7.2. Mismatch in devices constituting the clocked comparator may result in different switching The measured waveforms conform well with the direct on-chip probe measurements shown in Figure 7.18.

Figure 7.27 shows the supply noise waveform measured using the digital on-chip oscilloscope and the V-I converter based drop detector described in Section 7.2.3. In the experiments, the results from the on-chip oscilloscope were found to conform well with

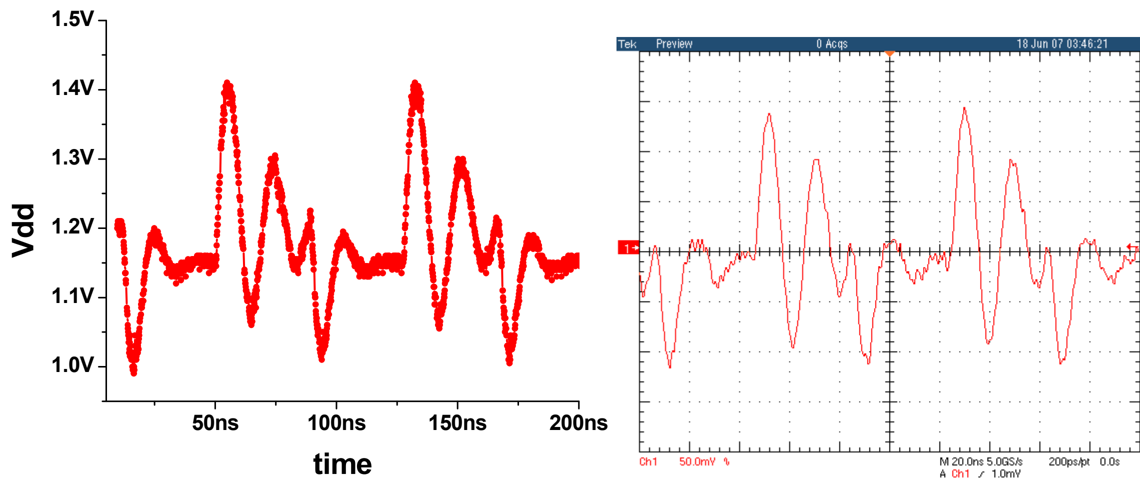


Figure 7.27. Measured supply waveform using the on-chip oscilloscope (left) and the V-I converter (right)

measurements from direct on-chip probing. The power overhead for the on-chip oscilloscope was <1mW as compared to ~25mW for the analog V-I converter based drop detector. The proposed on-chip oscilloscope may also be used an embedded block for other on-chip signal-integrity related measurements such as cross-talk noise.

7.6 Conclusions

In this chapter, we presented three digital circuit techniques for inductive supply noise suppression. The presented techniques effectively suppress supply noise caused by rapid current transients or due to resonance. The charge injection based active decoupling technique uses a nominal active supply and an active decap bank to inject extra charge into the power grid in case of an undershoot emergency. This technique does not require any high-voltage supplies and obviates the need of any thick-ox devices. Furthermore, the active decap acts as passive when the supply voltage is within the pre-specified safety bounds. The high-voltage charge pump based active circuit uses a high-voltage charge pump to

dump extra charge into the power grid during excessive undershoots. The high-voltage shunt supply based active circuit connects the regular nominal supply power grid directly to an external high-voltage supply whenever an undershoot is detected, thus damping the transient response of the supply network. We also presented a fully-digital on-chip oscilloscope which consumes much leower power than the prior proposed V-I converter based supply drop monitor. All the proposed circuits were implemented in a test-chip, fabricated in a 0.13μm, triple-well CMOS process. Measurement results demonstrate that the three active supply noise suppression techniques suppress the inductive supply fluctuation by 57%, 33% and 43%, respectively for a step load-current. During resonance, the supply fluctuation is suppressed by 75%, 43% and 45%, respectively by the three proposed circuit techniques. The performance of the proposed active circuit techniques was validated with the V-I converter-based drop detector circuit, the proposed on-chip oscilloscope and by direct on-chip probing of Vdd and Vss metal lines.

CHAPTER VIII

CONCLUSIONS AND FUTURE DIRECTIONS

Shrinking device dimensions, faster switching frequencies and increasing power consumption in deep submicron technologies cause large switching currents to flow in the power and ground networks. Rapid current transients in turn lead to supply noise which degrades performance and reliability. A robust power distribution network is essential to ensure reliable operation of circuits on a chip. Power supply integrity verification is, therefore, a critical concern in high-performance designs. This dissertation was aimed at proposing novel solutions to issues related to power grid integrity verification and exploring active circuits for supply noise suppression.

8.1 Contributions

The key contributions of this dissertation are summarized as follows:

- We proposed a path-based and a block-based analysis approach for computing the maximum circuit delay under power supply fluctuations. The analyses are based on the use of superposition, both temporally and spatially across different circuit blocks. The approaches are vectorless and take both IR as well as Ldi/dt drop into account. The path-based approach computes the maximum possible delay of a given critical path in the presence of supply variations, while the block-based approach does not require a-priori knowledge of the critical paths in a circuit and can be effectively incorporated in an existing static timing analysis framework. The delay

maximization problem is formulated as non-linear optimization problem with constraints on currents of macros or circuit blocks in the design. We showed how correlations between currents of different circuit blocks can be incorporated in the formulations using linear constraints. The proposed methods were validated on ISCAS85 benchmark circuits and an industrial power supply grid, and demonstrate a significant reduction in pessimism during worst-case circuit delay computation.

- We proposed two new approaches for analyzing the power supply drop. The first approach conservatively computes the worst-case supply drop, early in the design flow when detailed information of the design is not available. The second approach computes the statistical parameters of supply voltage fluctuations with variability in block currents. The proposed statistical analysis can be use to determine which portions of the grid are most likely to fail. The analyses consider both IR drop and L*di*/d*t* drop in a power supply network and can take into account both spatial and temporal correlations in block currents. We showed that the run time is linear with the length of the current waveforms allowing for extensive vectors, up to millions of cycles, to be analyzed. We implemented the approaches on a number of grids, including a grid from an industrial microprocessor to demonstrate their accuracy and efficiency.

- We proposed an approach for timing aware decoupling capacitance allocation which uses global time slacks to drive the optimization. Non-critical gates with larger timing slacks can tolerate a relatively higher supply voltage drop as compared to the gates on the critical paths. The decoupling capacitance allocation is formulated as a non-linear optimization problem using Lagrangian relaxation and the modified adjoint method is used to obtain the sensitivities of objective function to decap sizes. A fast path-based heuristic is also implemented and compared with the global optimization formulation. The approaches were implemented and tested

on ISCAS85 benchmark circuits and grids of different sizes. Compared to uniformly allocated decaps, the proposed approach utilizes 35.5% less total decap to meet the same delay target. For the same total decap budget, the proposed approach was shown to improve the circuit delay by 10.1% on an average.

- We described the first detailed full-die dynamic model of a 90nm Intel Pentium®-class microprocessor design, including package and non-uniform decap distribution. This model was justified from the ground up using a full-wave model and then increasingly larger but less detailed models with only the irrelevant elements removed. Using these models, we showed that there is little impact of on-die inductance in such a design, and that the package is critical to understanding resonant properties of the grid. We also showed that transient effects are sensitive to non-uniform de-cap distribution and that locality is a tight function of frequency and of the package-die resonance

- We presented an analog active decap circuit that significantly increase the effectiveness of decap in suppressing power supply fluctuations. The proposed circuit senses the supply drop and drives an amplified and inverted voltage fluctuation on the decap. The active decoupling circuit is powered by a separate power supply and we studied the optimal allocation of the total C4s/pads between this second power supply and the regular supply as well as the optimal allocation of the total decoupling capacitance between actively switched and traditional static decap. We demonstrated that the overhead of the proposed method is small compared to the area of the decaps. Simulations in a 0.13mm CMOS process demonstrate that the maximum supply drop is reduced by 45% compared to the use of only traditional decap, corresponding to an increase in the effective decap of approximately 8X.

- We presented three fully-digital circuit techniques for inductive supply noise suppression. The presented techniques effectively suppress supply noise caused by rapid current transients or due to resonance. The charge injection based active decoupling technique uses a nominal active supply and an active decap bank to inject extra charge into the power grid in case of an undershoot emergency. This technique provides as effective decap of 10.5X (for a 10% supply regulation tolerance) and does not require any high-voltage supplies and obviates the need of any thick-ox devices. Furthermore, the active decap acts as passive when the supply voltage is within the pre-specified safety bounds. The high-voltage charge pump based active circuit uses a high-voltage charge pump to dump extra charge into the power grid during excessive undershoots. The high-voltage shunt supply based active circuit connects the regular nominal supply power grid directly to an external high-voltage supply whenever an undershoot is detected, thus damping the transient response of the supply network. We also presented a fully-digital on-chip oscilloscope which is more power efficient compared to a conventional supply drop monitor. All the proposed circuits were implemented in a test-chip, fabricated in a 0.13µm, triple-well CMOS process. Measurement results demonstrate that the three active supply noise suppression techniques suppress the inductive supply fluctuation by 57%, 33% and 43%, respectively for a step load-current. During resonance, the supply fluctuation is suppressed by 75%, 43% and 45%, respectively by the three proposed circuit techniques. The performance of the proposed active circuit techniques was validated with the V-I converter-based drop detector circuit, the proposed on-chip oscilloscope and by direct on-chip probing of Vdd and Vss metal lines.

8.2 Future directions

In this section, we list some of the future challenges in power delivery design and analysis.

Multi-core power-delivery design and analysis

In the last decade, power consumption has increased from a few Watts to hundreds of Watts and frequency has increased from a few MHz into the GHz range in high performance microprocessors. Integrating multiple cores on the same die offers a significant area and power advantage over single-core designs, for the same performance. Therefore, there has been a paradigm shift in high performance microprocessor design from power-intensive, single-core designs towards power-efficient multi-core designs. Multi-core designs can further improve energy efficiency through fine-grained power management where different cores can operate at different supply voltages depending on the distribution of work-load.

Multi-core designs introduce new challenges to power distribution design. A multi-core design may consist of a globally unified power delivery system or a split system with separate power distribution network for each core. A split system is attractive for frequency management, where each core may run at its individual operating voltage. On the other hand, a unified system has the advantage of better connectivity between the cores. Thus, there is a further need for investigation into the issues related to multi-core power delivery design.

Supply drop analysis has also been complicated tremendously with the advent of multi-core design. The size of the supply network increases manifold, increasing the modeling complexity and simulation run-times. Thus, there is a need for efficient algorithms for efficient, large-scale supply noise analysis algorithms.

Package-Chip co-design

We discussed in Chapter V that package inductance has a significant impact on the Ld*i*/d*t* drop occurring in a supply network. However, with decreasing supply voltages and increasing power consumption, even the IR drops occurring in the power planes of packages may need to be factored into the chip design in the future. Package-chip co-design and co-simulation will lead to better control of Ld*i*/d*t* and IR drops which in turn improve the performance and reliability of the design. Designing the package to closely match the requirements of the chip also reduce wasted power and increase battery power. Hence, there is a need for CAD tools to support concurrent design of chip and package.

Power generation

This dissertation has only focussed on the power delivery related issues in high performance designs. However, reliable low-voltage power generation has become a significant challenge with the scaling of operating voltage. Power generation gets further complicated by power management techniques such as dynamic voltage scaling. Traditionally, power generation has been performed off-chip at the motherboard level. However, the response time of the voltage regulator is no longer fast enough to react to sudden surges in current demands. On-die decoupling capacitance and active regulation techniques can mitigate the rapid supply fluctuations to some extent. However, there is a need to explore power generation techniques which can be integrated on-die in order to minimize excessive supply drops and overshoots.

BIBLIOGRAPHY

[1] R. Ahmadi and F.N. Najm, “Timing analysis in presence of power supply and ground voltage variations,” in *Proceedings of International Conference on Computer-Aided Design*, 2003, pp. 176-183.

[2] E. Alon, V. Stojanovic and M.A. Horowitz, “Circuits and techniques for high-resolution measurement of on-chip power supply noise,” *IEEE Journal of Solid-State Circuits*, vol. 40, no. 4, pp. 820 - 828, April 2005.

[3] AMPL: A modeling language for mathematical programming. *www.ampl.com*.

[4] M. Ang, R. Salem and A. Taylor,“An on-chip voltage regulator using switched decoupling capacitors,” in *International Solid-State Circuits Conference Dig. Tech. Papers*, 2000, pp. 438-439.

[5] G. Bai, S. Bobba and I.N. Hajj, “RC power bus maximum voltage drop in digital VLSI circuits,” in *Proceedings on International Symposium on Quality Electronic Design*, 2001, pp. 205-210.

[6] M.S. Bazaraa, H.D. Sherali and C.M. Shetty, *Nonlinear Programming: Theory and Algorithms*, John Wiley & Sons, second edition 1993.

[7] J.R. Black, “Electromigration failure modes in aluminum metallization for semiconductor devices,” in *Proceedings of IEEE*, vol. 57, no.9, Sept. 1969, pp. 1587–1594.

[8] D. Blaauw, “Industrial perspectives on emerging CAD tools for low power processor design,” in *Proceedings of International Symposium on Low Power Electronics and Design*, keynote address,1998.

[9] S. Bobba, T. Thorp, K. Aingaran and D. Lu, “IC power distribution challenges,” in *Proceedings of International Conference on Computer Aided Design*, 2001, pp. 643-650.

[10] W.L. Briggs, *A Multigrid Tutorial*, PA: SIAM, 1987.

[11] A. Chandrakasan, W.J. Bowhill and F. Fox, *Design of high performance microprocessor circuits*. NY: IEEE Press, 2001.

[12] A. Chandy and T. Chen, "Performance driven decoupling capacitor allocation considering data and clock interactions," in *Proceedings of Design, Automation and Test in Europe,* 2005, pp. 984-985.

[13] R. Chaudhry, R. Panda, T. Edwards and D. Blaauw, "Design and analysis of power distribution networks with accurate RLC models," in *Proceedings of VLSI Design*, 2000, pp. 151-155.

[14] Chung-Ping Chen, C.C.N. Chu, and D.F. Wong, "Fast and Exact Simultaneous Gate and Wire Sizing by Lagrangian Relaxation", *Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 18, no. 7, pp. 1014-1025, July 1999.

[15] H. Chen and D. Ling, "Power Supply Noise Analysis Methodology for Deep-submicron VLSI Chip Design," in *Proceedings of Design Automation Conference*, 1997, pp. 638-643.

[16] L.H. Chen, M. Sadowska and F. Brewer, "Coping with buffer delay change due to power and ground noise," in *Proceedings of Design Automation Conference*, 2002, pp. 860-865.

[17] T.H. Chen and C.C. Chen, "Efficient large-scale power grid analysis based on preconditioned Krylov-subspace iterative methods", in *Proceedings of Design Automation Conference*, 2001, pp. 559-562.

[18] T.H. Chen, C. Luk, C.C.P. Chen, "INDUCTWISE: inductance-wise interconnect simulator and extractor," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 22, no. 7, pp. 884-894, July 2003.

[19] E. Chiprout, "Fast flip-chip power grid analysis via locality and grid shells," in *Proceedings of International Conference on Computer-Aided Design*, 2004, pp. 485-488.

[20] S. Chowdhry and J.S. Barkatullah, "Estimation of maximum currents in MOS IC logic circuits," *IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems*, vol. 9, pp. 642–654, June 1990.

[21] A.R. Conn, R.A. Haring, C. Visweswariah and C.W. Wu, "Circuit optimization via adjoint Lagrangians," in *Proceedings of International Conference on Computer Aided Design*, 1997, pp. 281-288.

[22] A.R. Conn, N.I.M. Gould and P.H. Toint, *LANCELOT: A Fortran Package for Large-Scale Nonlinear Optimization*, Springer Verlag, 1992.

[23] G.R. Cooper, C.D. McGillem, *Probabilistic Methods of Signal and System Analysis*, Oxford Series, 1998.

[24] L. Daniel, A. Sangiovanni-Vincentelli and J. White, “Techniques for including dielectrics when extracting passive low-order models of high speed interconnect,” in *Proceedings of International Conference on Computer Aided Design*, 2001, pp. 240-244.

[25] S. Das, D. Roberts, S. Lee, S. Pant, D. Blaauw, T. Austin, K. Flautner and T. Mudge, “A self-tuning DVS processor using delay-error detection and correction,” *Journal of Solid-State Circuits*, vol. 41, no. 4, pp. 792-804, April 2006.

[26] K. DeHaven and J. Dietz, “Controlled collapse chip connection (C4)-an enabling technology,” in *Proceedings of Electronic Components and Technology Conference*, pp. 1 - 6, 1994.

[27] A. Dharchoudhury, R. Panda, D. Blaauw, R. Vaidyanathan, B. Tutuianu, and D. Bearden, “Design and analysis of power distribution networks in PowerPC microprocessors,” in *Proceedings of Design Automation Conference*, 1998, pp. 738–743.

[28] S. Director and R. Rohrer, “The generalized adjoint network and network sensitivities,” *IEEE Transactions on Circuits and Systems*, vol. 16, no. 3, pp. 318-323, Aug. 1969.

[29] P.G. Doyle and J.L. Snell, *Random Walks and Electric Networks*, Washington, DC: Mathematical Association of America, 1984.

[30] T. Fischer, J. Desai, B. Doyle, S. Naffziger and B. Patella, “A 90-nm variable frequency clock system for a power-managed Itanium architecture processor,” *IEEE Journal of Solid-State Circuits*, vol. 41, no. 1, pp. 218-228, Jan. 2006.

[31] M. Fukazawa, T. Matsuno, T. Uemura, R. Akiyama, T. Kagemoto, H. Makino, H. Takata and M. Nagata, “Fine-Grained In-Circuit Continuous-Time Probing Technique of Dynamic Supply Variations in SoCs,” in *International Solid-State Circuits Conference Dig. Tech. Papers*, 2007, pp.288-289.

[32] K. Gala, D. Blaauw, V. Zolotov, P.M. Vaidya and A. Joshi, “Inductance model and analysis methodology for high-speed on-chip interconnect,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 10, no. 6, pp. 730 - 745, Dec. 2002.

[33] G. Golub and C. Van Loan, *Matrix Computations*, The John Hopkins University Press, 1989.

[34] E. Grochowski, D. Ayers and V. Tiwari, “Microarchitectural simulation and control of di/dt-induced power supply voltage variation,” in *Proceedings of International Symposium on High-Performance Computer Architecture*, 2002, pp. 7-16.

[35] F.W. Grover, *Inductance Calculations*. New York: Dover, 1954.

[36] J. Gu, H. Eom, and C.H. Kim, “A Switched Decoupling Capacitor Circuit for On-Chip Supply Resonance Damping,” in *Symposium on VLSI Circuits Dig. Tech. Papers*, 2007, pp. 126-127.

[37] E. Hailu, D. Boerstler, K. Miki, J. Qi, M. Wang and M. Riley,“A circuit for reducing large transient current effects on processor power grids,” in *International Solid-State Circuits Conference Dig. Tech. Papers*, 2006, pp. 548-549.

[38] C. Ho, A.E. Ruehli, and P. Brennan, “The modified nodal approach to network analysis,” *IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems*, vol. 22, pp. 504–509, June 1975.

[39] R. Ho, B. Amrutur, K. Mai, B. Wilburn, T. Mori and M. Horowitz, “Applications of on-chip samplers for test and measurement of integrated circuits,” in *Symposium on VLSI Circuits Dig. Tech. Papers*, 1998, pp. 138–139.

[40] Y.M. Jiang, T. Young and K. Cheng, “VIP - an input pattern generator for identifying critical voltage drop for deep submicron designs,” in *Proceedings of International Symposium on Low Power Electronic Design*, 1999, pp. 156-161.

[41] R. Joseph, D. Brooks and M. Martonosi, “Control techniques to eliminate voltage emergencies in high performance processors,” in *Proceedings of International Symposium on High-Performance Computer Architecture*, 2003, pp. 79-90.

[42] Y. Kanno, Y. Kondoh and T. Irita, “In-Situ Measurement of Supply-Noise Maps with Millivolt Accuracy and Nanosecond-Order Time Resolution,” in *Symposium on VLSI Circuits Dig. Tech. Papers*, 2006, pp. 78-79.

[43] D. Kouroussis and F.N. Najm, “A static pattern-independent approach for power grid voltage integrity verification,” in *Proceedings of Design Automation Conference*, 2003, pp. 99-104.

[44] J. Kozhaya, S.R. Nassif, and F.N. Najm, “A multigrid-like technique for power grid analysis,” *Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 21, no. 10, pp. 1148-1160, Oct. 2002.

[45] H. Kriplani, F. Najm and I. Hajj, “Pattern independent maximum current estimation in power and ground buses of CMOS VLSI circuits,” *Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 14, no. 8, pp. 998-1012, Aug. 1995.

[46] A. Krstic and K. Cheng, “Vector generation for maximum instantaneous current through supply lines for CMOS circuits,” in *Proceedings of Design Automation Conference*, 1997, pp. 383-388.

[47] P. Larsson, "Resonance and damping in CMOS circuits with on-chip decoupling capacitance," *Transactions on Circuits and Systems-I*: *Fundamental Theory and Applications*, vol. 45, no.8, pp 849-858, Aug. 1998.

[48] H. Li, Z. Qi, S.X.D. Tan, L. Wu, Y. Cai and X. Hong, "Partitioning-based approach to fast on-chip decap budgeting and optimization," in *Proceedings of Design Automation Conference,* 2005, pp.170-175.

[49] S. Lin and N. Chang, "Challenges in Power-Ground Integrity," in *Proceedings of International Conference on Computer-Aided Design*, 2001, pp. 651-654.

[50] T. Lin, M.W. Beaftie and L.T. Pileggi, "On the efficacy of simplified 2D on-chip inductance models", in *Proceedings of Design Automation Conference*, 2002, pp. 757-762.

[51] K. Makie-Fukuda and T. Tsukada,"On-chip active guard band filters to suppress substrate-coupling noise in analog and digital mixed-signal integrated circuits," in *Symposium on VLSI Circuits Dig. Tech. Papers*, 1999, pp. 57-60.

[52] A.V. Mezhiba and E.G. Friedman, "Impedance Characteristics of Power Distribution Grids in Nanoscale Integrated Circuits," *IEEE Transaction on Very Large Scale Integration (VLSI) Systems*, vol. 12, no. 11, pp. 1148-1155, Nov. 2004.

[53] A. Muhtaroglu, G. Taylor and T. Rahal-Arabi, "On-die droop detector for analog sensing of power supply noise," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 4, pp. 651–660, Apr. 2004.

[54] 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 Journal of Solid-State Circuits*, vol. 40, no. 4, pp. 813-819, April 2005.

[55] M. Nagata, J. Nagai, T. Morie and A. Iwata, "Measurements and Analyses of Substrate Noise Waveform in Mixed-Signal IC Environment," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Syst*ems, vol. 19, no. 6, pp. 671-678, June 2000.

[56] Y. Nakamura, M. Takamiya and T. Sakurai, "An on-chip noise canceller with high voltage supply lines for nanosecond-range power supply noise," in *Symposium on VLSI Circuits Dig. Tech. Papers*, 2007, pp. 292-293.

[57] S.R. Nassif and J.N. Kozhaya, "Fast power grid simulation," in *Proceedings of Design Automation Conference*, 2000, pp. 156-161.

[58] R. Panda, D. Blaauw, R. Chaudhry, V. Zolotov, B. Young and R. Ramaraju, "Model and analysis for combined package and on-chip power grid simulation," in *Proceedings of International Symposium on Low Power Electronic Design*, 2000, pp. 179-184.

[59] M.D. Pant, P. Pant, and D.S. Wills, “On-chip decoupling capacitor optimization using architecture level prediction,” *IEEE Transactions on Very Large Integration (VLSI) Systems*, vol. 10, pp. 319–326, June 2002.

[60] L.T. Pillage, R.A. Rohrer and C. Visweswariah, *Electronic Circuit and System Simulation Methods*, McGraw-Hill, 1995.

[61] H. Qian, S.R. Nassif and S.S. Sapatnekar, “Early-stage Power Grid Analysis. for Uncertain Working Modes”, *Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, no. 5, pp. 676-682, May 2005.

[62] H. Qian, S.R. Nassif and S.S. Sapatnekar, “Random walks in a supply network,” in *Proceedings of Design Automation Conference*, 2003, pp. 93-98.

[63] B. Razavi, *Design of Analog CMOS Integrated Circuits*, McGraw-Hill, 2002.

[64] A.E. Ruehli, “Equivalent Circuit Models for Three Dimensional Multi-conductor Systems,” *IEEE Transactions on Microwave Theory Tech.*, vol. MTT-22, pp. 216-221, Mar. 1974.

[65] A.E. Ruehli, “Inductance calculations in a complex integrated circuit environment,” *IBM Journal of Research and Development*, pp. 470–481, Sept. 1972.

[66] R. Saleh, S.Z. Hussain, S. Rochel, and D. Overhauser, “Clock skew verification in the presence of IR-drop in the power distribution network,” *IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems*, vol. 19, pp. 635–644, June 2000.

[67] H. Sanchez, B. Johnstone, D. Roberts, O. Mandhana, B. Melnick, M. Celik, M. Baker, J. Hayden, B. Min, J. Edgerton and B. White, “Increasing microprocessor speed by massive application of on-die high-K MIM decoupling capacitors,” in *International Solid-State Circuits Conference Dig. Tech. Papers*, 2006, pp. 538-539.

[68] L. Scheffer, L. Lavagno and G. Martin, *Electronic Design Automation for Integrated Circuits Handbook*, CRC Press, 2006.

[69] Semiconductor Industry Association. International Technology Roadmap for Semiconductors, 2004.

[70] Stanford Business Software Inc. *www.sbsi-sol-optimize.com.*

[71] G. Steele, D. Overhauser, S. Rochel and Z, Hussain, “Full-chip verification methods for DSM power distribution systems,” in *Proceedings of Design Automation Conference*, 1998, pp. 744-749.

[72] H. Su, K.H. Gala, and S.S. Sapatnekar, "Analysis and optimization of structured power/ground networks," *IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems*, vol. 22, pp. 1533–1544, Nov. 2003.

[73] H. Su, S.S. Sapatnekar and S.R. 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, pp. 428-436, April 2003.

[74] M. Takamiya, M. Mizuno and K. Nakamura, "An On-Chip 100 GHz-Sampling Rate 8-Channel Sampling Oscilloscope with Embedded Sampling Clock Generator," in *International Solid-State Circuits Conference Dig. Tech. Papers*, 2002, pp. 182-183.

[75] S. Taylor, "The challenge of designing global systems," in *Proceedings of Custom Integrated Circuits Conference*, 1999, pp. 429-435.

[76] J. Tschanz, N. S. Kim, S. Dighe, J. Howard, G. Ruhl, S. Vangal, S. Narendra, Y. Hoskote, H. Wilson, C. Lam, M. Shuman, C. Tokunaga, D. Somasekhar, S. Tang, D. Finan, T. Karnik, N. Borkar, N. Kurd and V. De, "Adaptive frequency and biasing techniques for tolerance to dynamic temperature-voltage variations and aging," in *International Solid-State Circuits Conference Dig. Tech. Papers*, 2000, pp. 438-439.

[77] 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 Journal of Solid-State Circuits*, vol. 40, no. 1, pp. 67-79, Jan 2005.

[78] R.R. Turnmula and E.J. Rymwzewski. *Microelectronics Packaging Handbook*, New York: Van Nostrand Reinhold, 1989, ch. 6, pp. 366-391.

[79] S.R. Vemuru, "Effects of simultaneous switching noise on the tapered buffer design" *IEEE Transactions on Very Large Integration (VLSI) Systems*, vol. 5, pp. 290–300, Sept. 1997.

[80] A. Waizman and C.Y. Chung, "Resonant-free Power Network Design Using Extended Adaptive Voltage Positioning (EAVP) Methodology," *IEEE Transactions on Advanced Packaging,* vol. 24, issue 3, pp. 236 – 244, Aug. 2001.

[81] N. Weste and D. Harris, *CMOS VLSI Design A Circuits and Systems Perpective*, Addison Wessley, 2004.

[82] W. Winkler and F. Herzel,"Active substrate noise suppression in mixed-signal circuits using on-chip driven guard rings," in *Proceedings of Custom Integrated Circuits Conference*, 2000, pp. 357-360.

[83] A.M. Wu and S.R. Sanders, "An active clamp circuit for voltage regulation module (VRM)applications" *IEEE Transactions on Power Electronics*, vol. 16, no. 5, pp. 623-634, Sep. 2001.

[84] 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 *International Solid-State Circuits Conference Dig. Tech. Papers*, 2007, pp. 268-269.

[85] M. Xu and L. He, "An Efficient Model for Frequency-based Onchip Inductance", in *Proceedings of Great Lakes Symposium on VLSI*, 2001, pp. 115-120.

[86] C.Y. Yeh and M.M. Sadowska, "Timing-aware power noise reduction in layout," in *Proceedings of International Conference on Computer Aided Design*, 2005, pp. 627-634.

[87] M. Zhao, R.V. Panda, S.S. Sapatnekar and D. Blaauw, "Hierarchical analysis of power distribution networks," *Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 21, no. 2, pp. 159-168, Feb. 2002.

[88] S. Zhao, K. Roy and C.K. Koh, "Decoupling capacitance allocation and its application to power-supply noise-aware floorplanning," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 21, no. 1, pp. 81-92, Jan. 2003.

[89] 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 Integration (VLSI) Systems*, vol. 11, no. 3, pp. 336-344, June, 2003.