Power Supply Noise Analysis Methodology for Deep-Submicron VLSI Chip Design

Howard H. Chen and David D. Ling

IBM Research Division
Thomas J. Watson Research Center
Yorktown Heights, NY 10598, U.S.A.

Abstract

This paper describes a new design methodology to analyze the on-chip power supply noise for high-performance microprocessors. Based on an integrated package-level and chip-level power bus model, and a simulated switching circuit model for each functional block, this methodology offers the most complete and accurate analysis of Vdd distribution for the entire chip. The analysis results not only provide designers with the inductive ΔI noise and the resistive IR drop data at the same time, but also allow designers to easily identify the hot spots on the chip and ΔV across the chip. Global and local optimization such as buffer sizing, power bus sizing, and on-chip decoupling capacitor placement can then be conducted to maximize the circuit performance and minimize the noise.

1 Introduction

One of the most important issues in today's deep submicron design is signal integrity. To preserve signal integrity, every circuit must have a built-in noise margin to allow for signal degradation. The signal degradation comes from various sources, such as the coupled noise from adjacent signals, the reflection noise from impedance discontinuities, and the power supply noise due to switching currents [1]. Excessive noise not only will introduce additional signal delay, but also may cause false switching of logic gates. This paper will focus on the analysis of power supply noise which includes the IR drop and ΔI noise, and discuss the use of on-chip decoupling capacitors to keep power supply within specification, provide signal integrity, and reduce EMI radiated noise.

In traditional VLSI design, the resistive IR drop occurs mostly on the chip, and the inductive ΔI noise only occurs on the package. They are often analyzed separately and designed with their respective noise budgets. However, as we move into deep submicron design with much smaller feature size, faster switching speed, and higher circuit density, the inductive component of wire impedance jwL becomes comparable to R, and the on-chip power bus inductance can no longer be ignored. The ΔI noise, also referred to as simultaneous switching noise or ground bounce, is caused by changes in current (ΔI) through various parasitic inductors. The simultaneous switching of I/O drivers and internal circuits can increase the voltage drop on the power supply by an amount of $\Delta V = \sum L\Delta I/\Delta t$, where L is the effective wire inductance of power busses, ΔI is the current change during transition, and Δt is the rise or fall time. Since the maximum ΔI noise occurs during switching when the current change ΔI is maximum, and the maximum IR drop occurs when the current I is at its peak, the worst-case ΔI noise and worst-case IR drop do not occur at the same time. It is therefore too pessimistic to calculate the maximum power supply noise by analyzing the ΔI noise and IR drop separately, and then adding the two worst cases together. An integrated package-level and chip-level power bus model with switching and timing information is needed to accurately analyze the Vdd variation over time.

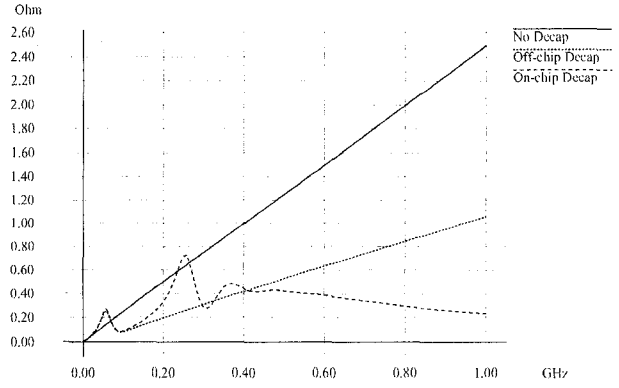

Figure 1: Package impedance frequency response

The excessive power supply voltage drop ΔV in deep submicron design also necessitates the use of on-chip decoupling capacitors, in addition to off-chip decoupling capacitors, to alleviate the noise problem. Fig. 1 shows the frequency response of impedance on a single-chip module package. If no decoupling capacitance is provided on the package, the impedance (jwL) will increase linearly with the frequency (w). The addition of off-chip decoupling capacitance on the package reduces the total impedance, but the magnitude of

"Permission to make digital/hard copy of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication and its date appear, and notice is given that copying is by permission of ACM, Inc. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee."

DAC97, Anaheim, California
(c)1997 ACM 0-89791-920-3/97/06..$3.50

the impedance still increases with the frequency. Finally, with additional on-chip decoupling capacitance, the impedance will taper off at high frequency. Therefore, for low-frequency ΔV problems, it may be adequate to use only the off-chip decoupling capacitors. However, for high-frequency ΔV problems, the on-chip decoupling capacitor is more effective due to its proximity to the switching activities. For today's 300 MHz CMOS RISC Microprocessor design [2], as much as 160 nF on-chip decoupling capacitance is used to control the power-supply noise. Therefore it is imperative to accurately estimate and optimize the use of on-chip decoupling capacitors for high-performance design.

In the following sections, we will describe how to model the power bus structure and switching activities to analyze the power supply noise. The resistive (IR) and inductive (LdI/dt) voltage drops are considered at both the chip level and package level, so that we can correctly identify the hot spots on the chip, and ΔV across the chip. To reduce the IR drop at the local hot spots, design guidelines are provided to resize the I/O buffers and power busses in the corresponding area. If excessive switching noise is present at the local hot spots, an iterative improvement procedure is proposed to estimate the on-chip decoupling capacitance needed to keep Vdd within specification. The additional decoupling capacitors will then be placed inside or adjacent to the macros to minimize the simultaneous switching noise.

To demonstrate the various applications of the chip-level noise analysis, we have included three case studies in this paper. The first benchmark compares the flip-chip C4 and peripheral I/O technologies, and their respective on-chip power supply voltage drops. The second case illustrates the effectiveness of on-chip decoupling capacitors by comparing the Vdd distribution before and after the placement of decoupling capacitors. The third experiment allows designer to predict the power supply noise of an SOI chip, based on an existing bulk CMOS design.

2 Power Bus Model

A complete power supply distribution model must include the package-level power distribution network, the on-chip power bus model, and the equivalent circuits to represent the various on-chip switching activities for each functional block (Fig. 2). Among the three major components in the model, the package-level power bus model is dominated by

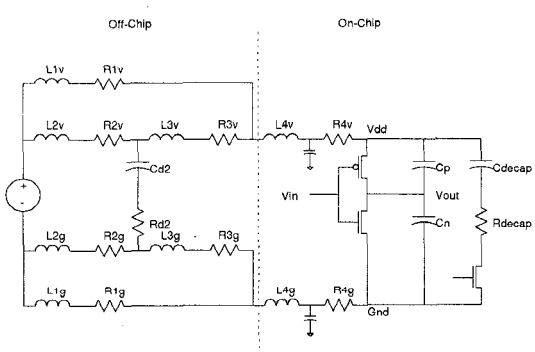

Figure 2: Power supply distribution model

the inductance, the on-chip power bus model is dominated by the resistance, and the switching circuit model determines the switching current.

Fig. 3 illustrates a simplified package-level power bus model for a single chip site. The power and ground distribution networks on the thin film and ceramic mesh planes are represented by their equivalent inductance model. The off-chip decoupling capacitors, multilayer ceramic vias, C4 connections [3] to the chip, and I/O pins to the board interface, are also included in the model.

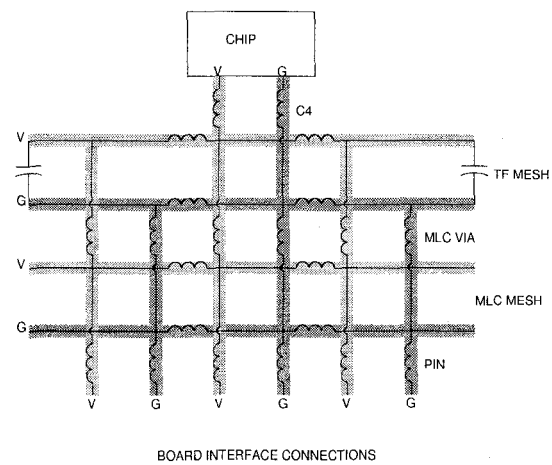

Figure 3: Simplified package-level power bus model

To analyze the on-chip power supply voltage drop, we need to model the resistance, capacitance, and inductance of each power bus segment. The nominal resistance at $25°C$, $R_{25} = R_s/width$, is determined by each layer's sheet resistance R_s and the width of the power bus. At a operating temperature of $85°C$, the resistance increases to $R_{85} = R_{25} \times (1+T_C \times (85-25)) \times (1+10\%)$, where T_C is the temperature coefficient, and an additional 10% is added to account for the electromigration induced resistance increase over the lifetime of the device. The total capacitance for the power bus consists of three components: the area capacitance, the fringe capacitance, and the line-to-line capacitance. The area capacitance is the parallel plate capacitance to the wiring planes above and below. The fringe capacitance is the capacitance from the left and right edges of the wire to the wiring planes above and below, based on semi-cylindrical approximation. The line-to-line capacitance is the coupling capacitance between adjacent wires on the same wiring plane. The modeling of wire inductance, however, is more complicated and cannot be represented by simple formula such as $LC = \epsilon_r/c^2$. For a periodic Vdd/Gnd structure, we use the $PROPCALC$ program [4] to calculate the propagation characteristics, with the assumption that the mesh plane on the multichip module constitutes the ground plane. Mutual inductance is also considered if the power busses are in close proximity to each other. Although the inductance on the package dominates the ΔI noise, on-chip power bus inductance generally cannot be ignored for wires wider than $5um$.

After we calculate the resistance, capacitance, and inductance for each power bus segment, an equivalent RLC power-bus

network can be generated. In order to reduce the complexity for full-chip analysis, a hierarchical approach is used to build the on-chip power bus model. At the chip level, a global routing grid is generated to subdivide the chip into global routing cells. All the switching activities within one global routing cell are lumped together, and adjacent cells are connected by the global power busses. At the macro level, where local hot spots are located, a finer grid will be generated to model the detailed power bus structure. Since the power supply voltage in one region can be affected by the switching activities in the neighboring regions, the finer detailed power bus model should always be connected to the adjacent global power bus model to ensure accurate analysis results.

3 Switching Circuit Model

To model the switching activities for each functional block, we build an equivalent circuit which consists of time-varying resistors (R_1, R_2, R_3), loading capacitors (C_1, C_2, C_3) and decoupling capacitors (C_{d1}, C_{d2}) (Fig. 4). The loading capacitance for the equivalent circuit is calculated by $C_L = P/V^2 f$, where P is the estimated power for the corresponding area, V is the power supply voltage, and f is the clock frequency. When the circuit is turned on, the time-varying resistance will be set to R_{on}, where $R_{on}C_L$ = switching time constant. When the circuit is switched off, the time-varying resistance will be set to R_{off}. Since not all circuits will switch at the same time, the circuit represented by the loading capacitance C_L can be further partitioned into subcircuits represented by $C_1, C_2, C_3, ...$, where $\sum C_i = C_L$, to simulate the distributed switching activities. The timing and delay pattern of each subcircuit can be controlled separately by switching on and off $R_1, R_2, R_3, ...$ at different times.

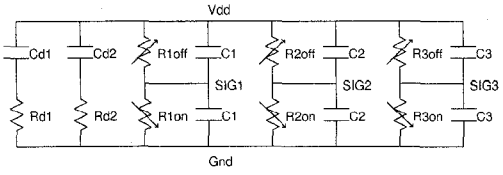

Figure 4: Equivalent switching circuit

If the simulation results of functional blocks are available, we can replace the nonlinear devices and capacitive loads in the switching circuit model with the piecewise linear current sources, which mimic the waveforms of the actual circuits. A triangular or trapezoidal current waveform, which is a simpler form of the piecewise linear current model, can also be derived by calculating the total average current I_{ave} and peak current I_{peak} for each macro in the procedure listed below.

1. Simulate circuits without loading to obtain internal I_{ave} and I_{peak}.
2. Calculate the total output capacitance C_{out} from all output nets.
3. $I_{ave}(total) = I_{ave}(internal) + C_{out} * Vdd * f$, where Vdd is the power supply voltage and f is the frequency. The switching factor SF is ignored for the purpose of calculating I_{ave} when the circuits are switching.
4. $I_{peak}(total) = I_{peak}(internal) * n$, where n is an empirical ratio between the peak current with loading and the peak current without loading.
5. Calculate the total power. $P = \frac{1}{2}Vdd(I_{ave}(internal) + C_{out} * Vdd * f * SF)$.

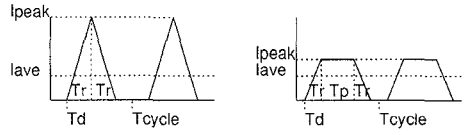

Figure 5: Triangular and trapezoidal current waveforms

Fig. 5 shows the simple triangular and trapezoidal current waveforms which are derived from the average current I_{ave}, peak current I_{peak}, cycle time T_{cycle}, and delay time T_d, assuming that the rise time and fall time are approximately equal. If I_{ave} is less than or equal to one half of I_{peak}, a triangular current waveform, characterized by

$$\frac{1}{2}I_{peak}(T_r + T_r) = I_{ave}T_{cycle},$$

can be generated, where T_r is the rise time. If the average current I_{ave} is greater than $\frac{1}{2}I_{peak}$, the circuit can be represented by a trapezoidal current waveform with

$$I_{peak}(T_r + T_p) = I_{ave}T_{cycle},$$

where T_r is the rise time and T_p is the time period when the current stays at peak. If the pulse width T_{pw}, which is equal to $(T_r + T_p + T_r)$, is known, T_r can be derived easily from $I_{peak}(T_{pw} - T_r) = I_{ave}T_{cycle}$. Otherwise any waveform which satisfies the constraint of $T_{cycle}(I_{ave}/I_{peak}) \leq T_{pw} \leq T_{cycle}$, can be generated.

After the equivalent circuit for each functional block is generated, it will then be assigned to the global routing cell(s) where the functional block is located, and connected to the corresponding points on the power bus.

4 Decoupling Capacitor Model

The model for on-chip decoupling capacitors consists of 3 major components: the n-well capacitor C_{nw}, the circuit capacitor C_{ckt}, and the thin-oxide capacitors C_{ox}. The n-well capacitor C_{nw} is the reverse-biased pn junction capacitor between the n-well and p-substrate (Fig. 6). The time constant for C_{nw} is process dependent, but usually can be characterized between 250ps and 500ps. The circuit capacitor C_{ckt} is derived from the built-in capacitance between Vdd and Gnd in non-switching circuits (Fig. 7). The total capacitance $\sum(C_p + C_n)$ from non-switching circuits is estimated to be $(P/(V^2 f)) * (1 - SF)/SF$, where P is the power of the circuit, V is the supply voltage, f is the frequency, and SF is the switching factor. The time constant for C_{ckt} is determined by the switching speed of the device, and typically ranges from 50ps to 250ps. The thin-oxide capacitor C_{ox} uses the

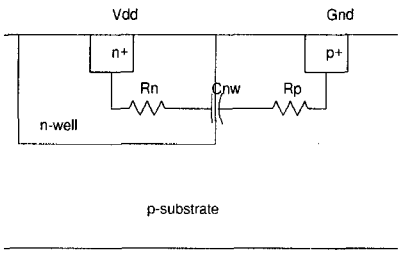

Figure 6: N-well junction decoupling capacitor

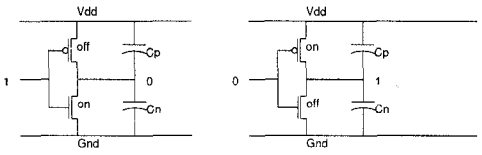

Figure 7: Non-switching circuit decoupling capacitor

thin-oxide layer between n-well and polysilicon gate (Fig. 8) to provide the additional decoupling capacitance needed to alleviate the switching noise problem. The thin-oxide capacitors are usually added near the drivers, high-power macros, or any available empty space on the chip. Depending on the size and shape of C_{ox}, its RC time constant can range from 100 ps to 300 ps.

5 Switching Noise Analysis

According to the switching patterns and placement of functional units, we can generate the equivalent circuits and attach them to the corresponding points on the power bus. The on-chip power busses in turn are connected to the power bus structure on the package for circuit simulation and noise analysis. A common mistake which has often been overlooked is to perform a chip-level noise analysis without the use of a package-level model. Since a chip is always mounted on a module or board, it is impractical to assume constant power supply voltage at the chip I/O's. Therefore, a complete chip-level noise analysis must include a package-level model which considers the effect of package inductance, to account for voltage drops on both the package level and chip level.

Fig. 9 illustrates the current waveforms of 6 circuits that are

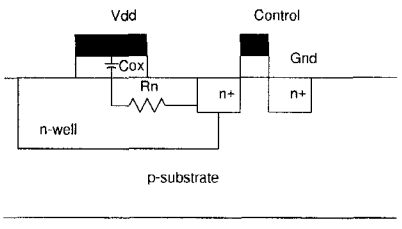

Figure 8: Thin-oxide decoupling capacitor

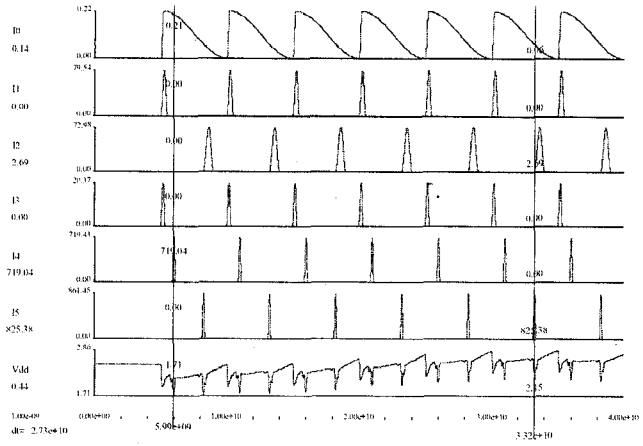

Figure 9: Vdd variation due to delayed switching

attached to the same power bus at one local hot spot. Signals can be switched simultaneously or with their respective delay patterns. The corresponding Vdd waveform of the noisy power supply is shown at the bottom of Fig. 9.

Another major concern during switching noise analysis is the voltage differential between various locations on the chip. We are concerned with not only the steady-state noise of the hot spots, but also the transient noise when circuits switch from one power level to another. To examine the differential noise between different units on the chip, we partition the chip site into 9 (3x3) regions. Assuming a power supply voltage of 2.5V for the $0.25um$ CMOS technology [5], and the circuits are switched from 20% idle power to 100% full power, we measure the transient voltage and the steady-state voltage in each region (Table 1). If the flip-chip C4 technology is used to provide the on-chip power supply, the minimum steady-state Vdd in the center region will be 2.37V. If the C4's are replaced by the wire-bond peripheral I/O's, the minimum steady-state Vdd in the center region will drop to 2.00V. Therefore, beyond the I/O density advantage, the use of C4's for high performance design provides significant leverage on noise reduction, especially as the chip size and power increase.

Table 1: C4 and wire-bond Vdd comparison

I/O	Vdd (min-max)	Chip center	Chip corner
C4	Transient	2.19V - 2.69V	2.21V - 2.69V
C4	Steady state	2.37V - 2.62V	2.37V - 2.62V
WB	Transient	1.96V - 2.44V	2.16V - 2.50V
WB	Steady state	2.00V - 2.20V	2.24V - 2.46V

After identifying the hot spots on the chip and their corresponding switching noise, we can resize the power bus or detune the circuits to minimize the noise. If it is still not possible to contain the amount of noise, then we need to consider the use of on-chip decoupling capacitors. The following section describes a decap optimization procedure to minimize the sizes and optimize the locations of on-chip decoupling capacitors, subject to the floor-planning constraints.

6 Decap Optimization Procedure

Since the power supply voltage directly affects the driving capability and signal delay of VLSI circuits, most designs now require ΔV to be contained within 10% of Vdd. To achieve this goal, decoupling capacitors are added to minimize the switching noise. For high-performance circuits with a cycle time of 5 ns or less, it is estimated that as much as 10% of the chip area may be needed to serve this purpose [6]. Therefore, it is important to estimate and allocate the area needed for on-chip decoupling capacitors during the early floor-planning stage.

The floor planning of flexible decoupling capacitors is restricted by the topological and ordering constraints of the preplaced functional blocks. Given the relative placement of a set of functional blocks B, we can generate two directed acyclic graphs (G_H, G_V), where G_H is the horizontal constraint graph and G_V is the vertical constraint graph. For each block $b_i \in B$, there is a corresponding node in both G_H and G_V. The chip boundaries are represented by the $LEFT$ and $RIGHT$ nodes in G_H, and the TOP and $BOTTOM$ nodes in G_V. If (b_i, b_j) is an edge in G_H, then b_i is to be placed to the left of b_j. If (b_i, b_j) is an edge in G_V, then b_i is to be placed below b_j. The weight x_i of node b_i in G_H represents the x-dimension (width) of the block, while the weight x_{ij} of edge (b_i, b_j) in G_H represents the horizontal spacing between adjacent blocks b_i and b_j. Similarly, the weight y_i of node b_i in G_V represents the y-dimension (height) of the block, while the weight y_{ij} of edge (b_i, b_j) in G_V represents the vertical spacing between adjacent blocks b_i and b_j. As decoupling capacitors fill the empty spaces where $x_{ij} > 0$ or $y_{ij} > 0$, additional nodes b_k and edges (b_i, b_k) (b_k, b_j) may be introduced dynamically to represent the pseudo blocks of decoupling capacitors. If $L(G_H)$ is the length (total weight) of the path from $LEFT$ to $RIGHT$ in G_H and $L(G_V)$ is the length of the path from $BOTTOM$ to TOP in G_V, then $L(G_H) \times L(G_V)$ is the total chip area which must be fixed or minimized.

The optimization of on-chip decoupling capacitors involves an iteration process between circuit simulation and floor planning. Given the specifications and location of each functional block, the circuit simulator will analyze the switching noise on the power bus, identify the hot spots, and determine the amount of decoupling capacitance C_n needed for each global cell n. The floor planner then translates the amount of decoupling capacitance into physical area A_n, generates pseudo blocks b_k in each region n, and determines their locations and dimensions (x_k, y_k) such that $\sum(x_k \times y_k) \geq A_n$, and $L(G_H) \times L(G_V)$ in the updated constraint graph is minimized. The added decoupling capacitors will be modeled and simulated with the new floor plan during the next iteration until ΔV is contained. Since it may be possible to allocate a certain portion of the decoupling capacitor inside the macro, the area needed for decoupling capacitors can be reduced to $A_n{}'$ after the physical layout of each macro is complete.

To illustrate the effects of on-chip decoupling capacitors, we analyzed the Vdd distribution of a 15W chip, where C4's are used to provide the on-chip power supply, the relative positions of functional blocks are fixed, and 5% of the total chip area is available for on-chip decoupling capacitors. Based on

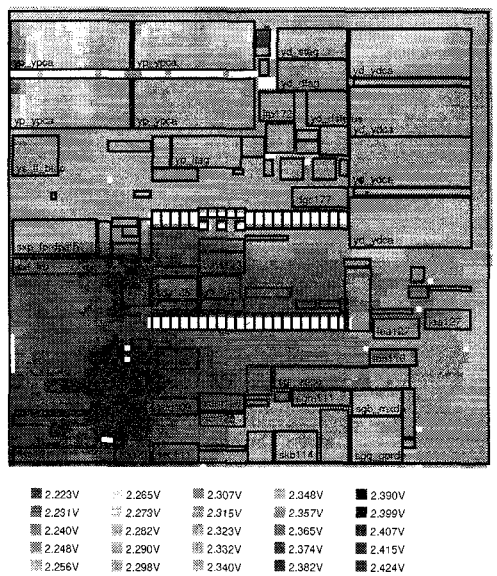

Figure 10: Vdd distribution without additional on-chip decoupling capacitors

the same voltage-scale color map, the minimum steady-state Vdd distribution with only built-in on-chip decoupling capacitors $(C_{nw} + C_{ckt})$ is shown in Fig. 10, and the minimum steady-state Vdd distribution with additional on-chip decoupling capacitors $(C_{nw} + C_{ckt} + C_{ox})$ is shown in Fig. 11. The color change from Fig. 10 to Fig. 11 clearly illustrates how the additional on-chip decoupling capacitors help to reduce the power supply noise in their respective regions. By optimizing the size and location of decoupling capacitors, we are able to limit the power supply voltage drop within 10% of Vdd. The circuit simulation for a 20-cycle full-chip power supply noise analysis takes 140 minutes on an IBM R/S 6000 workstation with a memory requirement of 692M using the optimized L/U factorization solution method. The same analysis will require 272 CPU minutes, but only 76M memory, if row-wise Gaussian elimination is used as the solution method.

7 Noise Estimation for SOI Circuits

Deep submicron CMOS silicon-on-insulator (SOI) circuits offer performance advantages over the traditional bulk CMOS circuits due to the reduction of parasitic capacitance, enhanced channel mobility, and higher circuit density. However the reduced parasitic capacitance and faster switching speed also lead to severe switching noise problems for CMOS SOI circuits. Furthermore, the degradation of power supply voltages may cause circuit instability and device *latch-up*, due to SOI's floating substrate structure [7, 8]. The power supply noise problem is even more critical for SOI devices with low threshold voltages, when the power supply voltage is scaled down from 2.5V to the projected 0.9V in 2010. By tying a transistor's gate to its body [9], we can lower the threshold voltage when the device is turned on, thereby increasing the switching speed, and raise the threshold voltage when the device is turned off, thus reducing the leakage current. The lower threshold voltage makes the device more

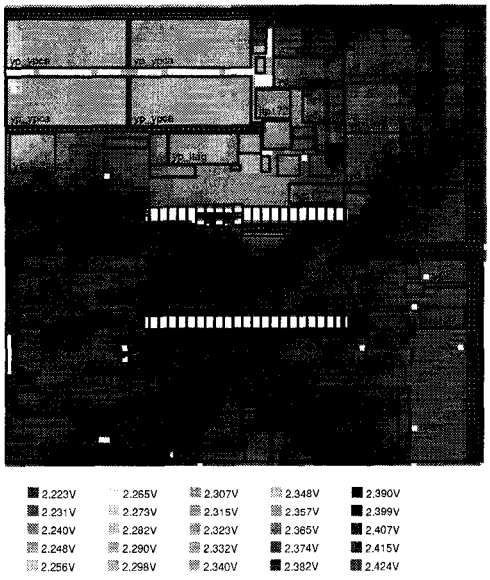

Figure 11: Vdd distribution with additional on-chip decoupling capacitors

noise sensitive, and it is therefore imperative to analyze and identify the potential noise problems for high-performance SOI circuit design.

To estimate the switching noise for an SOI chip, we replace the functional blocks of an existing bulk CMOS chip with the corresponding SOI circuits. The remapped SOI chip uses the same power bus structure and floor plan as the bulk CMOS chip, but different design parameters such as the reduced parasitic capacitance and faster switching speed. Table 2 shows some of the scaling factors that we use during simulation when the bulk CMOS chip is remapped to an SOI chip. As far as decoupling capacitance and switching noise are concerned, the SOI chip has less circuit capacitance C_{ckt}, no n-well capacitance C_{nw}, and higher peak current I_{peak} due to its faster switching speed.

Table 2: Normalized circuit parameters

Parameters	Bulk/CMOS	CMOS/SOI
I_{ave}	1.0	1.0
I_{peak}	1.0	1.3
C_{ox}	1.0	1.0
C_{ckt}	1.0	0.8
C_{nw}	1.0	0.0

For two 15W chips implemented in the bulk CMOS and SOI technologies respectively, we have found that the worst-case minimum steady-state Vdd drops to 1.00V for the SOI chip, compared to 1.54V for the bulk CMOS chip. The additional voltage drop can be attributed to SOI's faster switching speed and lack of intrinsic decoupling capacitance provided by the n-well and substrate structures in bulk CMOS. Since the excessive noise not only will introduce additional signal delay, but also may lead to device latch-up or false switching, the potential noise problem for SOI circuits is much more serious than the traditional bulk CMOS circuits.

8 Conclusions

We have developed a methodology to analyze the chip-level power supply switching noise, identify the hot spots where the most significant Vdd drops occur, and estimate the amount of on-chip decoupling capacitance needed to minimize the noise. By integrating the hierarchical on-chip power bus model and the macro-based switching circuit model with the package-level power distribution model, we are able to provide a complete and accurate switching noise analysis for high performance VLSI design. Three case studies are presented to illustrate the various applications of our power supply noise analysis methodology. As we continue to scale down the feature size and power supply voltage in deep submicron circuits, this methodology will play a vital role in preserving the reliability and achieving the performance targets of future VLSI design.

9 Acknowledgments

The authors would like to thank Stanley Schuster, Gerard Kopcsay, Barry Rubin, Steve Schmidt, Peter McCormick, Robert Hatch, and Dale Becker for their help in developing the early models and providing the test cases.

References

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

[2] W. Bowhill *et al.*, "A 300 MHz 64b quad-issue CMOS RISC microprocessor," in *Proc. International Solid-State Circuits Conference*, pp. 182–183, February 1995.

[3] L. Miller, "Controlled collapse reflow chip joining," *IBM Journal of Research and Development*, vol. 13, no. 3, pp. 239–250, 1969.

[4] B. Rubin, "An electromagnetic approach for modeling high-performance computer packages," *IBM Journal of Research and Development*, vol. 34, pp. 585–600, July 1990.

[5] B. Davari *et al.*, "A high performance 0.25um CMOS technology," in *Proc. International Electron Devices Meeting*, pp. 56–59, December 1988.

[6] D. Dobberpuhl *et al.*, "A 200-MHz 64-bit dual-issue CMOS microprocessor," *IEEE Journal of Solid-State Circuits*, pp. 1555–1567, November 1992.

[7] J. Seliskar *et al.*, "Voltage limitation of 0.5um CMOS on thin SOI," in *Proc. International Symposium on Silicon-on-Insulator Technology and Devices*, pp. 118–119, May 1992.

[8] A. Acovic *et al.*, "Hot carrier reliability of fully depleted accumulation mode SOI MOSFETs," in *Proc. IEEE International SOI Conference*, pp. 134–135, October 1992.

[9] F. Assaderaghi *et al.*, "A dynamic threshold voltage MOSFET (DTMOS) for ultra-low voltage operation," in *Proc. International Electron Devices Meeting*, pp. 809–812, December 1994.