On-Chip Decoupling Capacitor Optimization for Noise and Leakage Reduction

Howard H. Chen, J. Scott Neely, Michael F. Wang, and Gricell Co
IBM Corporation
1101 Kitchawan Road, Yorktown Heights, New York, U.S.A.

Abstract

The on-chip decoupling capacitors are widely used in today's high-performance microprocessor design to mitigate the power supply noise problem. The continued reduction of oxide thickness in advanced nanotechnology, however, also significantly increases the tunneling current and leakage power of thin-oxide capacitors. This paper describes the modeling and simulation of a complete chip and package power supply distribution network, and the optimization of the placement of thin-oxide and thick-oxide capacitors to reduce the tunneling current, leakage power, and burn-in cost, while limiting the power supply noise within noise margin.

1. Introduction

During the past four decades, the semiconductor industry has followed the Moore's Law by doubling the performance and functionality per chip every technology node [1]. However, with the channel length of MOSFET expected to reach 9 nm by 2016 (Table 1), the continued scaling of CMOS devices seems to have approached its physical limit.

Table 1. Semiconductor technology roadmap

Year	Gate length (nm)	Oxide thickness (nm)	Gate leakage (uA/um)	Supply Voltage (V)
2001	65	1.3	0.01	1.2
2002	53	1.2	0.03	1.1
2003	45	1.1	0.07	1.0
2004	37	0.9	0.10	1.0
2005	32	0.8	0.30	0.9
2006	28	0.7	0.70	0.9
2007	25	0.6	1.00	0.7
2010	18	0.5	3.00	0.6
2013	13	0.4	7.00	0.5
2016	9	0.4	10.00	0.4

The advent of nanotechnology not only aggravates the power supply noise problem with lower supply voltage and smaller noise margin (10% of Vdd), but also significantly increases gate leakage due to the thinner gate dielectric that must be scaled with the gate length to fully realize performance gains [2].

The power supply noise (ΔV) is caused by the impedance (Z) of the power supply network and the current (I) that flows through the power supply lines. Figure 1 shows an example of the current spikes in the power supply network and the corresponding power supply voltage fluctuation during steady state from one of our microprocessor circuits. In order to accurately simulate the power supply noise, we need to take into account not only the IR drop due to wire resistance (R) in the power supply network, but also the $L\Delta I/\Delta t$ noise that results from the inductance (L) of the power supply network and the switching activities of the circuit ($\Delta I/\Delta t$). Figure 2 illustrates the dramatic effect of $L\Delta I/\Delta t$ noise on the transient power supply voltage from our simulation, when the circuits are in transition from idle power to maximum average power. It also shows why a traditional static IR drop analysis that does not consider the dynamic effect of switching activities cannot be used to estimate the total power supply noise that is best represented by $\Delta V=IR + L\Delta I/\Delta t$.

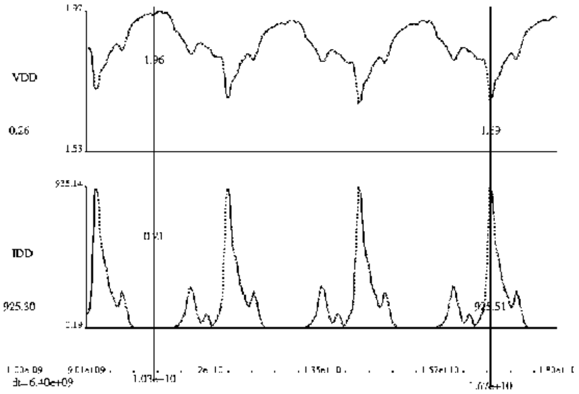

Figure 1. Steady-state Vdd and current waveform

The power supply voltage drop may result in false logic switching if the noise exceeds the threshold voltage during steady state. It will also affect the timing closure if the noise introduces additional delay during transient state. Since the device current is proportional to $(Vdd-Vt)^K$, where Vt is the threshold voltage and k is a super-linear parameter between 1 and 2, a 10% Vdd noise may have a 15% impact on circuit performance if k is equal to

0-7695-2009-X/03 $17.00 © 2003 **IEEE**

1.5. As the power supply voltage continues to scale down in future technologies, the power supply noise will have an increasingly significant impact on device current and circuit performance (figure 3).

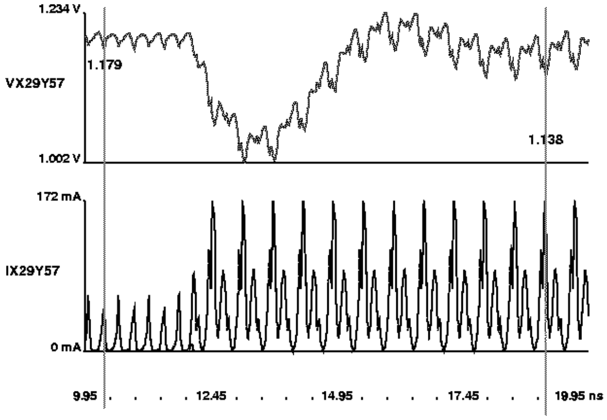

Figure 2. Transient Vdd and current waveform

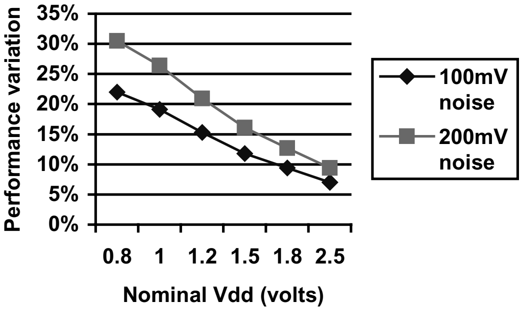

Figure 3. Performance impact from Vdd noise

To reduce the power supply fluctuation, decoupling capacitors [3] are often used to support the large current transients generated by the simultaneous switching of on-chip circuits and off-chip drivers. By charging up during the steady state, the decoupling capacitors can assume the role of power supply and provide the current needed during switching.

In a simplified circuit model, the electric charge before switching can be represented by C_d•Vdd, where C_d is the decoupling capacitance and Vdd is the nominal power supply voltage. The electric charge after switching can be represented by $(C_d + C_s)$•(Vdd + ΔV), where C_s is the switching capacitance, and ΔV is the power supply voltage fluctuation. From the conservation of charge, where C_d•Vdd = $(C_d + C_s)$•(Vdd + ΔV), we can easily derive the upper bound on transient power supply voltage fluctuation ΔV = -Vdd•C_s / $(C_d + C_s)$. To limit ΔV within 10% of Vdd and prevent the decoupling capacitors from being significantly discharged during circuit switching, a conservative value of 5 to 9 times the switching capacitance is often used as the guideline for decoupling capacitance.

The proper amount of decoupling capacitance should also be carefully selected, so as not to generate a resonant frequency $f = 1/(2\pi\sqrt{LC})$ near the operating frequency, which will significantly increase the impedance and power supply noise. Ironically, the parasitic resistance that causes IR drop and latch-up problems can help to resolve the resonance problem by introducing a damping effect and reducing the resonance impedance Z=L/(RC) [4].

Depending on the locations of the decoupling capacitors, on-chip decoupling capacitors are effective in reducing the high-frequency noise, while off-chip decoupling capacitors are effective in reducing the low-frequency noise. The on-chip decoupling capacitors include the intrinsic decoupling capacitors such as the n-well capacitors for bulk CMOS devices, the non-switching circuit capacitors, and the wiring capacitors between Vdd and Gnd. Additional decoupling capacitors such as the gate-oxide capacitors (figure 4) and trench capacitors can also be added to minimize the power supply noise.

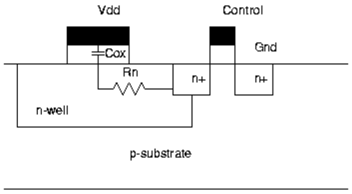

Figure 4. Gate-oxide decoupling capacitor

The gate-oxide capacitor has a capacitance per unit area of $C_{ox} = \varepsilon_{ox}/t_{ox}$, where ε_{ox} and t_{ox} are the dielectric constant and the thickness of the oxide respectively. In order for the gate to retain more control over the channel than the drain, the gate oxide thickness must be scaled proportionally to the channel length. Thin oxide provides more capacitance per unit area than thick oxide. However, the thinner the oxide, the higher the electric field across the gate and the higher the leakage current that may lead to an oxide breakdown. Table I shows that an oxide thickness of 9Å provides 44% more capacitance than an oxide thickness of 13Å, but its gate leakage will increase by an order of magnitude. Furthermore, an oxide thickness of less than 10Å has only a few layers of atoms and is subject to quantum-mechanical tunneling that exponentially increases the gate leakage current.

In order to extend the battery life of portable devices and other low-power application, it is absolutely necessary to control the gate leakage current, as well as the sub-threshold and junction leakage current. Device performance can only be maximized after the low leakage current requirement is met. Alternate gate dielectric material with a higher dielectric constant (high κ) may also be needed to control the increasing gate leakage current and satisfy the requirements of low power logic. For a gate dielectric of thickness T_d and relative dielectric constant κ, the equivalent oxide thickness in Table 1 is equal to T_d / (κ/3.9), where 3.9 is the relative dielectric constant of silicon dioxide.

2. Power supply distribution model

To address the deficiency of a static IR drop analysis and prevent any potential chip failure due to the collapse of power rails, we have developed a complete power supply noise model that includes both the package model and the on-chip power bus model to simultaneously simulate the inductive LΔI/Δt noise and the resistive IR drop (figure 5).

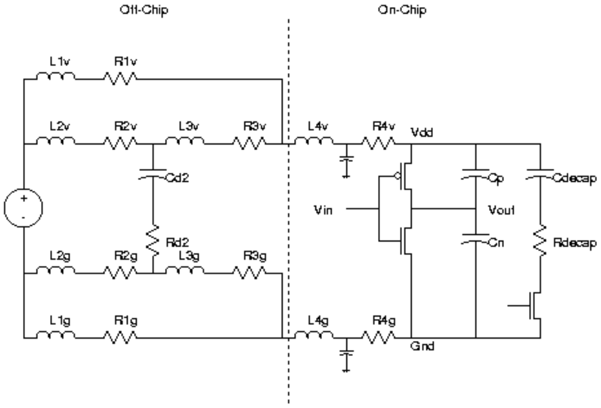

Figure 5. Power supply distribution model

In order to reduce the complexity of a full-chip power supply noise analysis, a hierarchical approach is used to build the chip and package power distribution model. At the package level, a coarse-grid birthday-cake model [5] is generated to represent the equivalent inductance between adjacent regions on a single-chip or multi-chip module package. At the chip level, a fine-grid model with C4 pitch is used to represent the multilayer RLC power bus network.

More importantly, in order to ensure the accuracy of a full-chip power supply noise analysis, we employ a state-of-the-art switching-circuit model [6] that truly captures the dynamic effect of transient current. Based on the circuit simulation results of our common power analysis methodology [7], we model the switching activities of each functional unit with a piecewise linear current source that mimics the switching pattern and current signature of the real circuits. For example, if the circuits operate with a certain power level such as hold power, maximum average power, clock-gated hold power, or clock-gated maximum average power, within a given cycle, then the waveform that best represents the current switching condition in one of the several possible states will be selected. As the circuits switch from one state to another state, the waveform will change accordingly from cycle to cycle to facilitate a vector-based dynamic power supply noise analysis.

Figure 6 illustrates one example that produces the worst-case scenario when all the circuits are switching between hold power and maximum average power at resonant frequency. To prevent excessive Vdd noise due to circuit resonance, we have to make certain in our power-aware design methodology that the clock and power supply are not gated at or near the resonant frequency. We also need to optimize the placement of decoupling capacitors to mitigate the power supply noise.

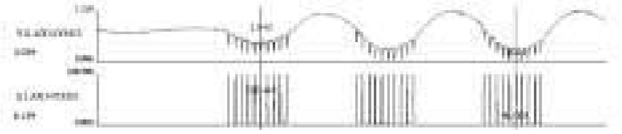

Figure 6. Resonant power supply voltage

It is worth noting the important role that timing plays in the noise analysis, as the noise doubles if two identical drivers switch at the same time, and the noise can be reduced by half if the same two drivers switch at different times. Therefore, in areas where hundreds of drivers are located, it is critical to properly model the switching factor and signal delay of each circuit, to minimize the compounding effect of noises that may be erroneously superimposed.

Finally, we need to model the intrinsic decoupling capacitors and additional decoupling capacitors with their respective time constants. The potential candidates for intrinsic decoupling capacitors include the device and junction capacitors that are connected between Vdd and Gnd. For a simple inverter buffer, about ½ of the gate capacitance, ½ of the diffusion capacitance, and ¾ of the gate-to-diffusion capacitance contribute to the intrinsic decoupling capacitance. Since only the circuits that are not switching can provide decoupling capacitance, the non-switching device capacitance is calculated by subtracting the switching capacitance from the total device capacitance.

The switching capacitance from clock circuits which charge and discharge at the frequency of f cycles per second can be calculated from $C=P/V^2f$, where P is the power dissipation and V is the power supply voltage. The switching capacitance for logic circuits which usually charges and discharges in alternating cycles can be calculated from $C=2P/V^2f$ [8].

3. Optimization of decoupling capacitors

Based on the floor plan of the chip, we can connect the switching circuit models to the corresponding points in the chip and package power-supply distribution model, and perform a full-chip power-supply noise analysis to estimate the steady-state and transient noise, and optimize the placement of decoupling capacitors. Both the thin-oxide and thick-oxide decoupling capacitors are available as options to reduce the noise and leakage. The thin-oxide decoupling capacitors are selectively placed in noisy hot spots due to their effectiveness in reducing the power supply noise. The thick-oxide decoupling capacitors, on the other hand, are placed in other areas that are less noisy to reduce the gate leakage.

The optimization of the placement of thin-oxide and thick-oxide decoupling capacitors involves an iterative process that is bounded by one initial simulation with 100% thin-oxide usage and another simulation with 100% thick-oxide usage for a given placement of decoupling capacitors. If the simulation result from 100% thin-oxide usage shows excessive noise that exceeds the noise margin, more thin-oxide decoupling capacitors must be added in the hot spots until the noise is contained. On the other hand, if the simulation result from 100% thick-oxide usage does not show any noise violations, unnecessary decoupling capacitors can be removed from the non-critical area to further reduce the leakage.

For a given placement of decoupling capacitors, if the simulation result from 100% thin-oxide usage shows noise containment, but the simulation result from 100% thick-oxide usage shows noise violation, ensuing optimization will continue until the proper distribution of thin-oxide usage and thick-oxide usage is determined. Depending on the manufacturing technology and performance target, the usage of thin-oxide and thick-oxide decoupling capacitors can be optimized to minimize the noise, subject to leakage constraints. It can also be optimized to minimize the leakage, subject to noise constraints.

In practical applications, it may not be necessary to simultaneously minimize the noise and leakage with sophisticated sensitivity analysis, because desirable results can often be achieved by following the simple procedure below.

1. Add decoupling capacitors to each functional unit, such that the total amount of decoupling capacitance is about 5 to 9 times the switching capacitance.
2. If the objective is to minimize leakage, subject to noise constraints, add thin-oxide decoupling capacitors to all available space on the chip. If the noise still exceeds the noise margin, add more decoupling capacitors at hot spots or change circuit layout as necessary. If the noise is contained, gradually replace thin-oxide capacitors with thick-oxide capacitors in noncritical area to reduce leakage, until the noise can no longer be contained.
3. If the objective is to minimize noise, subject to leakage constraints, add thick-oxide decoupling capacitors to all available space on the chip. Gradually replace thick-oxide capacitors with thin-oxide capacitors near hot spots to reduce noise, until the maximum leakage power limit is reached.

4. Benchmark analysis

To balance and optimize the use of thin oxide and thick oxide decoupling capacitors, we analyzed a benchmark microprocessor with seven different configurations, where the thin-oxide area ranges from 0 to 17 mm^2, and the thin-oxide capacitance ranges from 0 to 230 nF. The amount of thin-oxide and thick-oxide decoupling capacitance used in each configuration is shown in figure 7, with configuration A representing 100% thin oxide usage and configuration G representing 100% thick oxide usage.

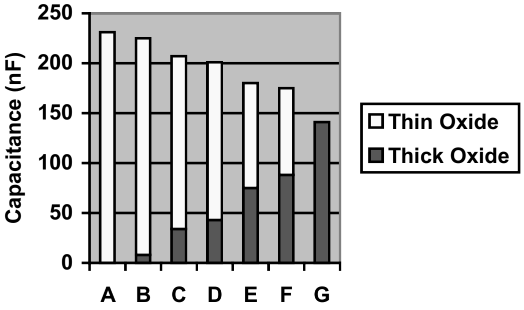

Figure 7. Thin-oxide decoupling capacitor usage

To simulate the general circuits with both linear and nonlinear elements, we use ACES [9] to measure the thermal power and the power supply current of each functional unit. Based on the current signatures in different states, circuit switching activities are modeled as a piecewise linear current source, and multi-cycle transient noise simulation can be performed by using the linear periodic function of time in PowerSPICE [10].

After extensive analysis of the power supply noise for each decoupling capacitor configuration (Table 2), configuration F is selected as the final configuration that limits the thin-oxide area to 6.7 mm^2 with a balanced use of thin-oxide and thick-oxide decoupling capacitors, while suppressing the worst-case transient power supply noise within the 200 mV noise margin. Our leading-edge technology enables us to analyze transient power supply noise under different scenarios such as power ramp-up and power ramp-down. It also allows us to evaluate the noise impact from various power-saving techniques such as clock gating and Vdd gating when the clocks or the power supplies to the logic blocks are cut off during the sleep mode. The gate leakage power is reduced by 61% after the area of thin-oxide decoupling capacitors decreases from 17.8 mm^2 to 6.7 mm^2. The optimal use of thin-oxide decoupling capacitors in the hot spots and thick-oxide decoupling capacitors in the cool spots contributes about 20% power saving under the maximum operating frequency of 1.8 GHz, and 30% power saving under the burn-in supply voltage of 1.8V and burn-in temperature of 140°C.

Table 2. Power supply noise comparison

Decoupling Capacitors	Thin-oxide Area (mm^2)	Transient Noise (mV)	Leakage Power (W)
A	17.8	191	26
B	16.7	191	25
C	13.3	193	20
D	12.2	193	18
E	8.1	197	12
F	6.7	198	10
G	0.0	232	0

5. Conclusion

A full-chip power supply noise analysis methodology has been developed to optimize the placement of on-chip decoupling capacitors for simultaneous noise and leakage reduction. In our current technology, the thin-oxide decoupling capacitors can provide 70% more capacitance per unit area than the thick-oxide decoupling capacitors, but the use of thin-oxide decoupling capacitors also generates more than 10 folds the leakage current than the thick-oxide decoupling capacitors. Without an optimal decoupling capacitor placement strategy in place, designers will have to adopt a conservative approach by populating the chip with the thick-oxide decoupling capacitors only to prevent a catastrophic increase of tunneling current, leakage power, and burn-in cost.

The noise and leakage reduction achieved in our benchmark analysis demonstrates the effectiveness of our decoupling capacitor optimization procedure, which selectively places the thin-oxide decoupling capacitors in critical hot spots only. As the leakage and noise problems continue to grow, due to the thinner oxide, smaller noise margin, and larger transient current, the control of power supply noise and the optimization of decoupling capacitor usage will become one of the most significant challenges for 90nm system-on-chip designs and beyond.

6. Acknowledgement

The authors would like to thank Erwin Cohen, Antje Mueller, Rolf Hilgendorf, Joachim Clabes, Robert Dennard, Mike Carlson, David Appenzeller, Daniel Ostapko, Thomas Bucelot, Leon Stok, and Eric Kronstadt for their contribution and support of this project.

7. References

[1] http://public.itrs.net

[2] Yuan Taur, "CMOS design near the limit of scaling," *IBM Journal of Research and Development,* Vol. 46, No. 2/3, March/May 2002, pp. 213-222.

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

[4] Patrik Larsson, "Resonance and damping in CMOS circuits with on-chip decoupling capacitance," *IEEE Transactions on Circuits and Systems – Part I: Fundamental Theory and Applications*, Vol. 45, No. 8, August 1998, pp. 849-858.

[5] Bradley McCredie and Wiren Dale Becker, "Modeling, Measurement, and Simulation of Simultaneous Switching Noise," *IEEE Transactions on Components, Packaging and Manufacturing Technology – Part B: Advanced Packaging*, Vol. 19, No. 3, August 1996, pp. 461-472.

[6] Howard H. Chen and J. Scott Neely, "Interconnect and Circuit Modeling Techniques for Full-Chip Power Supply Noise Analysis," *IEEE Transactions on Components, Packaging and Manufacturing Technology – Part B: Advanced Packaging*, Vol. 21, No. 3, August 1998, pp. 209-215.

[7] J. Scott Neely, Howard H. Chen, Steven G. Walker, James Venuto, Thomas J. Bucelot., "CPAM: A Common Power Analysis Methodology for High-Performance VLSI Design," *IEEE 9th Topical Meeting on Electrical Performance of Electronic Packaging*, October 2000, pp. 303-306.

[8] Gary K. Yeap, *Practical Lower Power Digital VLSI* Design, Kluwer Academic Publishers, Boston, 1998.

[9] Anirudh Devgan and Ronald A. Rohrer, "Adaptively controlled explicit simulation," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 13, No. 6, June 1994, pp. 746-762.

[10] William T. Weeks, Alberto Jose Jimenez, Gerald W. Mahoney, Deepak Mehta, Hassan Qassemzadeh, and Terence R. Scott, "Algorithms for ASTAP – a network-analysis program," *IEEE Transactions on Circuit Theory*, Vol. CT-20, No. 6, November 1973, pp. 628-634.