c01.qxd 12/16/2003 11:21 AM Page 1

# 1

# INTRODUCTION

As power supply voltage continues to drop with the VLSI technology scaling associated with significantly increasing device numbers in a die, power network design becomes a very challenging task for a chip with millions of transistors. The common task in VLSI power network design is to provide enough power lines across the chip to reduce the voltage drops from the power pads to the center of the chip. The voltage drops are mainly caused by the resistance or inductance of the power network metal lines.

The power network can be modeled as a low-pass filter with RL segments in series, attached with capacitors at each end. The current sources of the switching gates and the intentional decoupling capacitors are also inserted in the model. The IR drop is proportional to the average current consumed by the circuit in the chip. The $L \cdot di/dt$ drop is proportional to the time-domain change of the current, due to the switching of logic gates in the chip operations.

This chapter is organized into seven sections. Section 1.1 discusses the general trend of power supply noise with the process technology scaling. Section 1.2 shows the modeling methodology for on-chip power networks. Section 1.3 discusses the switching current modeling methodology for the power distribution network, which is critical for the accuracy of power grid analysis. Once we obtain the models, the power network can be characterized as a linear network with $R$, $L$, $C$, and current sources, in order to solve the voltage distributions across the power network.

*Power Distribution Network Design for VLSI,* by Qing K. Zhu
ISBN 0-471-65720-4 © 2004 John Wiley & Sons, Inc.

c01.qxd 12/16/2003 11:21 AM Page 2



Section 1.4 discusses a special topic in power network design: the decoupling capacitor optimization to allocate enough decoupling capacitors between $V_{dd}$ and $V_{ss}$ nets, but not over-allocating so as to result in enlargment of the die area. Section 1.5 discusses the on-chip inductance effects on power network modeling. We show the metal configurations used in the power line design in order to minimize the inductance delay. In general, many thin-width $V_{dd}$ and $V_{ss}$ lines interleaved with each other in the power distribution network are preferred in order to minimize the area of the return current loop or on-chip inductance.

Section 1.6 discusses process technology scaling impacts for the future power network design. We discuss the technology scaling impacts in two scenarios. Section 1.7 provides the summary to this chapter.

## 1.1 POWER SUPPLY NOISE

Noise problems in microprocessor power distribution networks have been discussed in the literature [1, 2, 3, 4, 5, 6]. The supply voltage is continually dropping in microprocessor design to reduce the power consumption and matche the reduced gate oxide thickness in the scaled IC process technology generations. Figure 1-1(a) shows the supply voltage drop trend in new technologies; and Figure 1-1(b) shows the gate oxide thickness reduction during the process scaling.

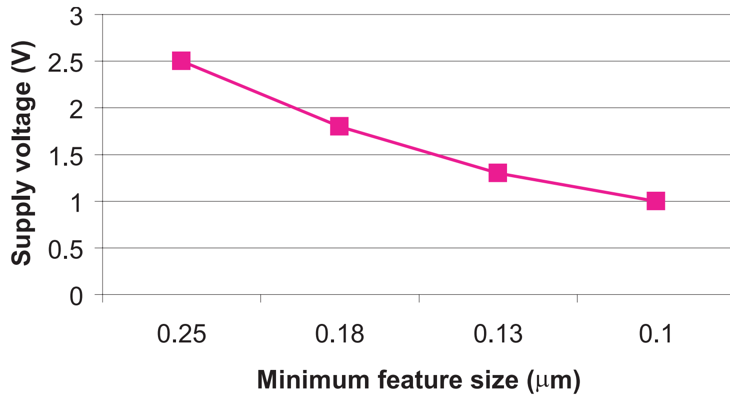
The on-chip decoupling capacitor is constructed by using the dummy transistors connected to $V_{cc}$ with the gate, and $V_{ss}$ with the drain and source. A conventional method for on-chip decoupling capacitance allocation is based on a percentage (i.e., 10%) area in each layout window (e.g., 100 × 100 μm) allocated for the decoupling capacitance.

The decoupling capacitors are inserted near the large-size buffers, such as clock buffers or phase-locked loops. The conventional method, based on the layout area percentage, is not optimal, either being overestimated for a large layout area or underestimated for meeting the power noise requirements.

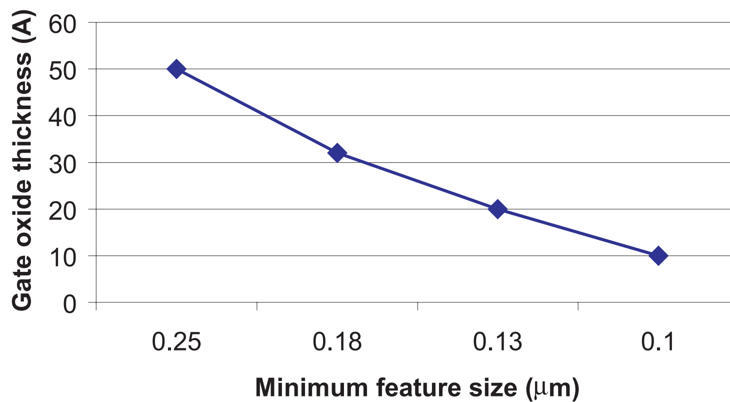
The power distribution design techniques used for DEC Alpha chips, such as the C4 package and on-chip power planes, can be found in [1]. The decoupling capacitance optimization technique, based on the layout floor plan graph and path-finding algorithm,

c01.qxd 12/16/2003 11:21 AM Page 3






(a)




(b)

**Figure 1-1.** Power supply (a) and gate oxide scaling (b) trends.

can be found in [2]. The power network modeling and analysis techniques for PowerPC microprocessors can be found in [3]. A power network modeling and simulation CAD tool is described in [4].

The reliability problems (i.e., electromigration) and CAD tool for the power network are discussed in [5]. The basics of VLSI power distribution can be found in [6]. The description of a high-performance power network scaling model and decoupling capacitance optimization method is proposed in [7]. A criterion to include the inductance in on-chip interconnect modeling was dis-

c01.qxd 12/16/2003 11:21 AM Page 4



cussed in [8]. The VLSI design basic to the power network design, such as metal sizing equations, can be found in [9]. Interconnect scaling issues in the deep-submicron process can be found in [10].

## 1.2 POWER NETWORK MODELING

The layout and C4 package of a high-performance microprocessor power network is illustrated in Figure 1-2. It is a five metal process, and M5 and M4 (the top two metal layers in this process) are used for the full-chip power distribution, although signal lines can still be routed between the spaces between the power lines in these top metal layers. Note that the local power networks are not shown in Figure 1-2; they will be routed on lower metal layers to deliver the power to the circuits.

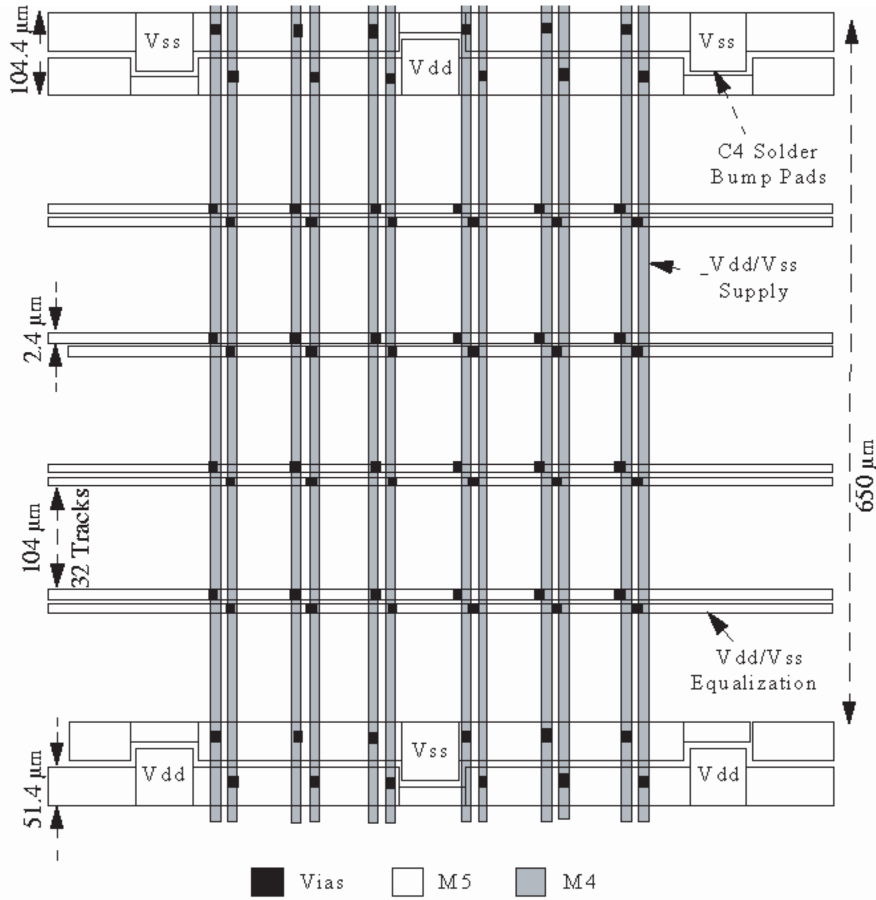
The on-chip power lines are modeled in *RLC* segments, as illustrated in Figure 1-3. $R_{vcc}$ and $L_{vcc}$ are the unit-length resistance and unit-length inductance (self and mutual) of the power line, multiplied by the line length between two nodes in the power grid.

$R_d$ and $C_d$ are the resistance and capacitance in the series, used to model the decoupling capacitor that is implemented by the dummy transistors. $I_s$ is the switching current of devices and it is time varying. $R_s$ and $C_s$ represent the turn-on resistance and the capacitance load of the devices connected at the power grid nodes (AC, BD, etc.).

The model in Figure 1-3 contains only the linear elements such as $R$, $L$, $C$, and current sources. It suggests to us that a linear circuit simulator can be used to speed up the large-size microprocessor power network analysis based on the proposed model. The key parameters of decoupling capacitors (dummy transistors) are $C_{\text{decap}}$ and $R_{\text{decap}}$, as shown in Figure 1-4.

The charges in $C_{\text{decap}}$ are used to help the supply voltage stability in $C_{\text{sw}}$ (switching gates) before the charges eventually come from the supply voltage source via the long current loop from the package.

To improve the efficiency of the decoupling capacitors, the $R_{\text{decap}}$ needs to be sufficiently small. When $V_{cc}$ is applied to the gate, as shown in Figure 1-5, the inversion channel is created between the D and S with the $R_{\text{ds-on}}$ resistance. The $R_{\text{ds-on}}$ resistance is the 1/slope of the $I/V$ curves of the resistor at $V_{ds}$ = 0V. The



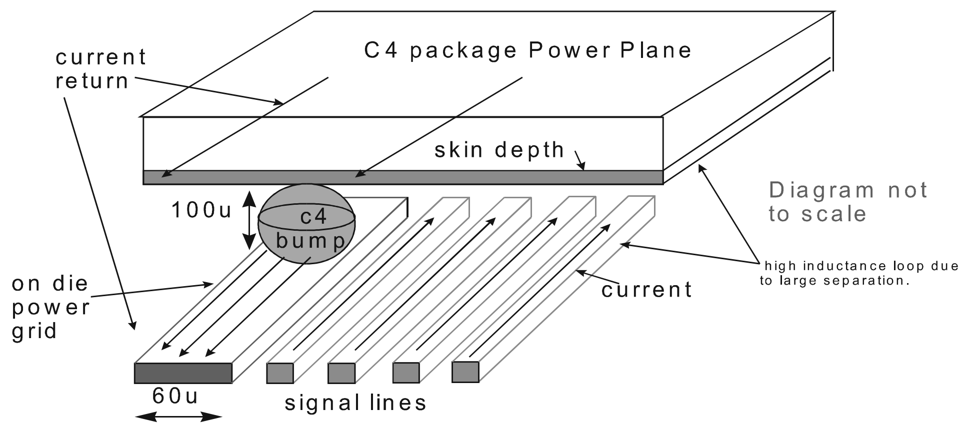




**Figure 1-2.** Power distribution for high-performance microprocessors.

c01.qxd 12/16/2003 11:21 AM Page 6



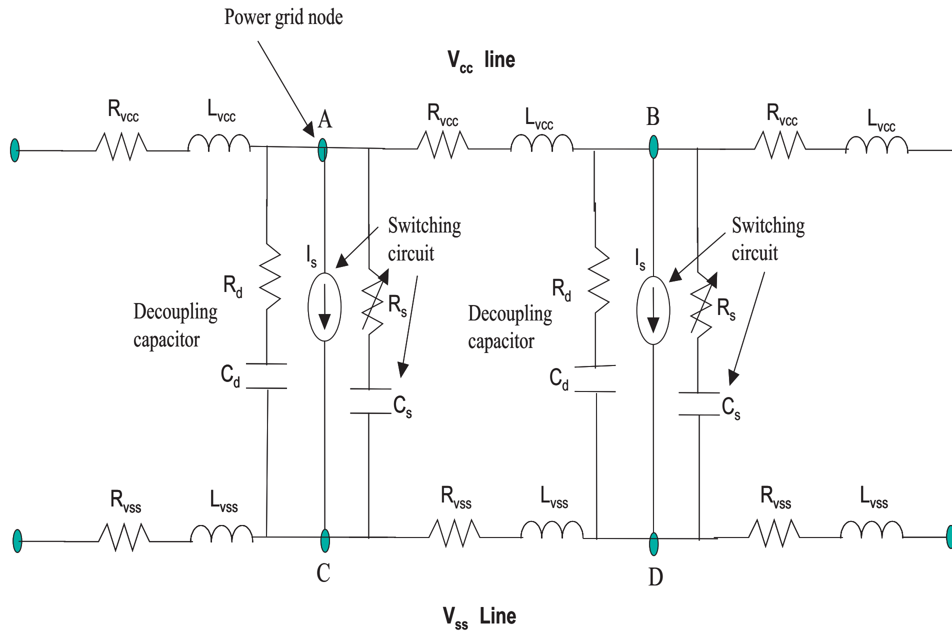



**Figure 1-3.** On-chip power grid *RLC* modeling.

$R_{\text{ds-on}}$ and $C_{\text{gate}}$ form a distributed *RC* network. $C_{\text{gate}}$ is in series with two $R_{\text{ds-on}}/2$ resistors connected in parallel, resulting in $R_{\text{ds-on}}/4$ added in series with $C_{\text{gate}}$, as shown in Figure 1-5.

The simulation of the power network depends on the accuracy and turnaround time of the power grid modeling. In most cases, only the resistance and capacitance of the power lines are needed,

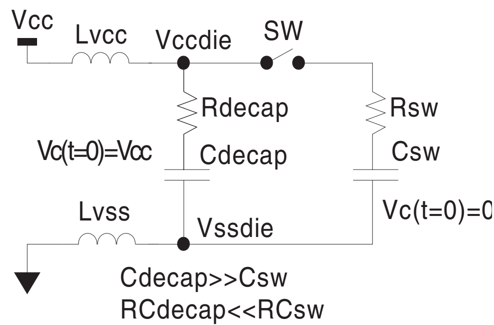



**Figure 1-4.** Switching model of decoupling capacitor.

c01.qxd 12/16/2003 11:21 AM Page 7



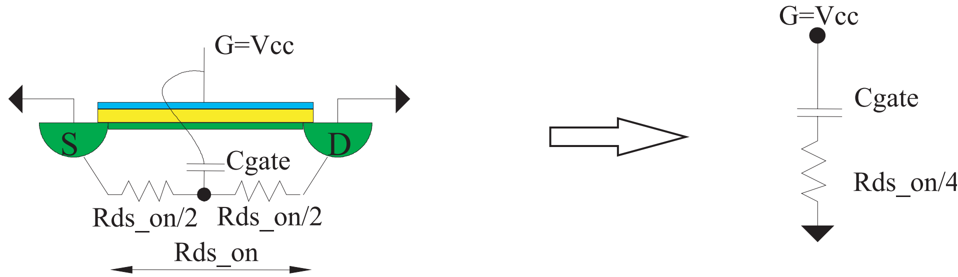



**Figure 1-5.** Decoupling capacitor modeling.

excluding the metal inductances for the on-chip power network. Many CAD tools are available for the purpose of extracting the interconnect *RC* for power grids, as summarized in Table 1-1.

The on-chip inductance for the power grid can be ignored by using special design rules, shortening the return loop of $V_{dd}$ and $V_{ss}$ by using several interleaved $V_{ss}$ and $V_{dd}$ lines, as shown in Figure 1-2, for example, to implement the power grid.

During the *RC* modeling process, each metal segment can be represented in two forms as follows: (1) the lumped capacitive parasitic, or (2) the distributed *RC* parasitic, as shown in Figure 1-6(a). The lumped capacitive parasitic represents the total wire capacitance from each driver circuit in the signal net. The distributed *RC* parasitic includes the resistance (*R*) of the metal line in the modeling.

Power grid modeling usually uses the *RC* model, since the metal line resistance of the power grid is significant at the full-chip level. A long metal line can be broken into multiple *RC* segments, as shown in Figure 1-6(b).

**Table 1-1.** Well-known *RC* extraction CAD tools

| Tool | Manufacturer |
|---|---|
| Fire & Ice | Cadence Design Systems |
| Star-RCXT | Synopsys |
| xCalibre | Mentor Graphics |
| HyperExtract | Cadence Design Systems |
| Arcadia | Synopsys |
| Columbus | Sequence Design |
| Nautilus | Cadence Design Systems |
| QuickCap | Random Logic |

c01.qxd 12/16/2003 11:21 AM Page 8



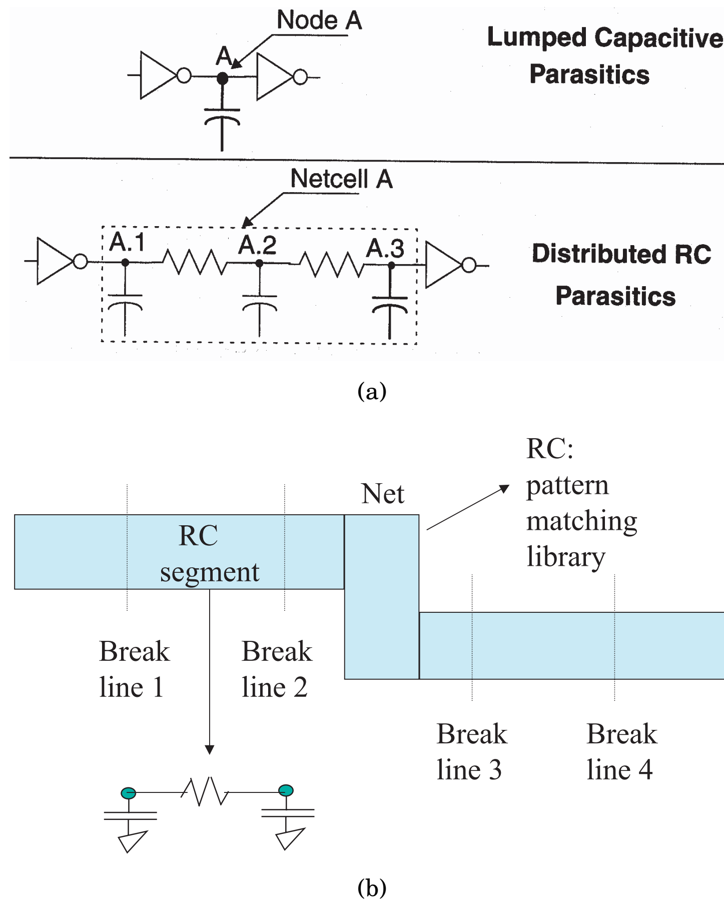



**Figure 1-6.** Lumped and distributed *RC* models.

Each *RC* segment is modeled with a series resistor, together with two capacitors at two ends of the resistor. The metal segment capacitance is evenly divided by two capacitors. This is usually called the Pai–*RC* model since it looks like a pi ($\pi$) symbol, as shown in Figure 1-6(b).

The extracted *RC* data from the layout are saved in a standard parasitic format (SPF) file. It includes a list of nets and detailed *RC* values. The *R* and *C* elements with the node names are specified either as schematic-based labels or layout-based labels, depending on the options used in the *RC* netlisting stage.

The schematic node names are preferred in the SPF, since this SPF can be back-annotated to the prelayout schematic netlist [33,

c01.qxd 12/16/2003 11:21 AM Page 9



34]. In addition, the SPF can include the device section that models the extracted devices from the physical layout.

In general, the capacitance can be formed between any polygons in the layout, although the closer ones have more significant capacitances, and thus have more impact on the total capacitance of the net. Figure 1-7 shows the possible capacitances between the gates and metal lines in the physical layouts.

The capacitance to the substrate is dominant over other coupling capacitances in the old one or two metals technology. But the situation changes in the latest submicron technology with seven to eight metal layers, since the top-level metals are far away from the substrate, and the total capacitance of these top-level metals is more impacted by the coupling capacitances between adjacent lines in the same layer or adjacent layers of the layout.

In addition, the spacing between metal lines is continually scaled, so the coupling capacitance between neighboring metal lines becomes more and more important. The calculation of the resistance or capacitance can be done through the direct solution of the well-known Maxwell's EM equations or Green's functions [17].

A complex geometrical layout can require an extremely long computational time using the direct EM field solution. Therefore,

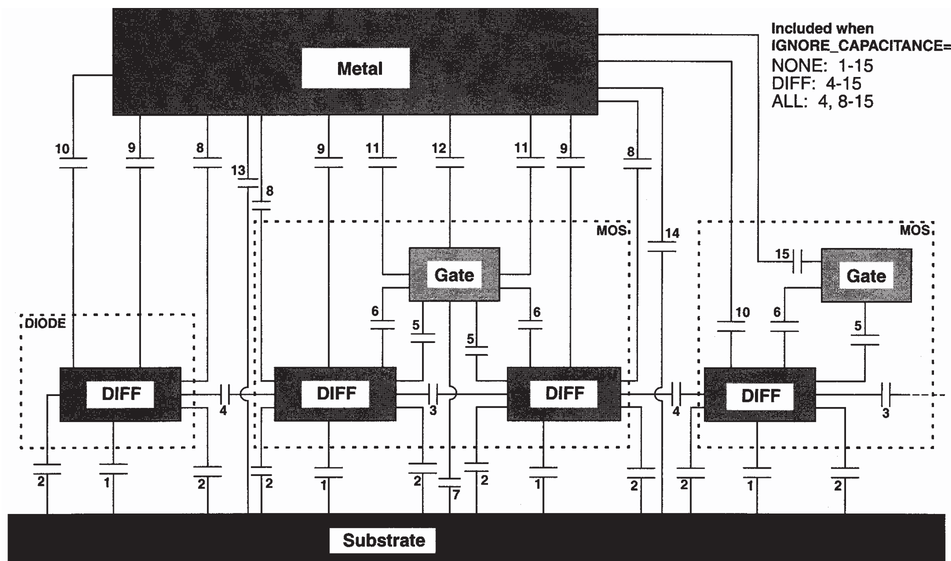



**Figure 1-7.** Coupling capacitances between conductors in a VLSI layout [33].

c01.qxd 12/16/2003 11:21 AM Page 10



equations or capacitance models are usually adopted in the capacitance calculation for a large-scale layout.

Once the capacitance equations have been established, they are used in the RC extraction, which is fast enough to handle a large-scale layout. The RC extraction works on the physical database together with the specified RC equations.

Let us review the basic resistance equation:

$$R = sl/w \text{ (ohm)} \tag{1-1}$$

In Equation (1-1), $s$ is the sheet resistance in the unit of ohm/square, $l$ is the length of the line in μm, and $w$ is the width of the line in μm.

Table 1-2 shows the sheet resistance data in a 0.18 μm technology. Metal four and metal five have significantly lower resistances, making them suitable for long metal routes. The polysilicon and metal one layers have high resistance, making them suitable for short metal connects.

The contacts or vias between metal layers, as shown in Figure 1-8, are usually modeled as resistors. Each contact or via has a fixed resistance based on design rules. The contact represents the metal hole between metal one to the diffusion or poly layer, whereas the via represents the metal hole between metal one and metal two. Contacts or vias will introduce many *RC* segments and significantly increase the *RC* parasitic file size and simulation time.

The unit-length capacitance models are based on the results in [41] as follows.

a. *Overlap capacitance:* the bottom/top surface of one line to the bottom and top surfaces of another line in two layers. Two lines are overlapped in the vertical direction. The overlap capacitance is modeled as $Ca = \varepsilon_0\varepsilon_r \cdot A/d_{l1l2}$, where $A$ is the overlap area of line $l_1$ and $l_2$, $\varepsilon_0$ is the permittivity of free

**Table 1-2.** Metal sheet resistances in 0.18 μm technology

| Layer | Polysilicon | Metal 1 | Metal 2 | Metal 3 | Metal 4 | Metal 5 |
|---|---|---|---|---|---|---|
| Sheet Resistance (Ω square) | 5.5 | 0.1 | 0.05 | 0.05 | 0.01 | 0.01 |

c01.qxd 12/16/2003 11:21 AM Page 11



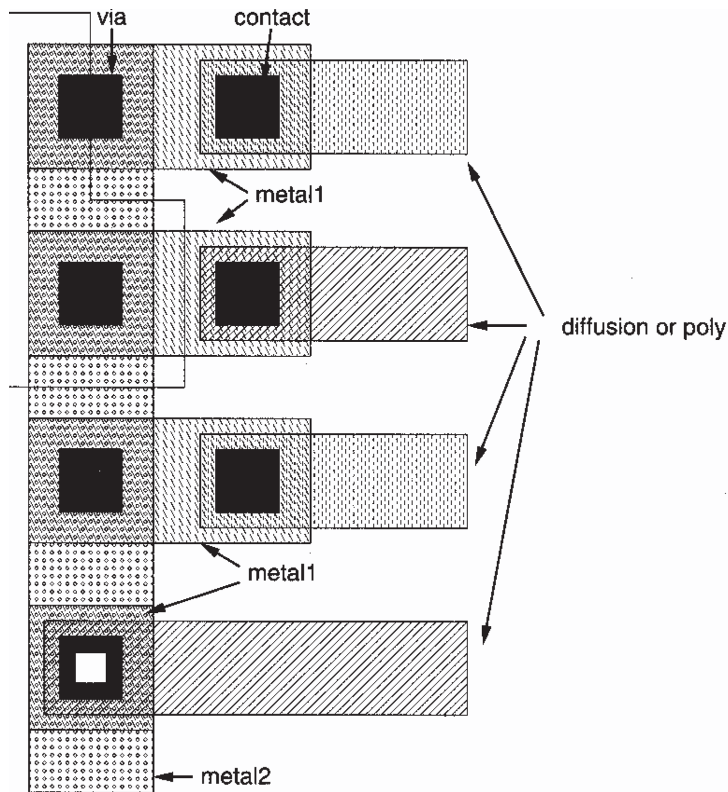



**Figure 1-8.** Contacts and vias [9].

space ($8.854 \cdot 10^{-14}$ F/cm$^2$), $\varepsilon_r$ is the relative permittivity between $l_1$ and $l_2$, and $d_{l1l2}$ is the vertical spacing between two lines.

b. *Fringe capacitance:* the side surface of one line to the bottom or top surface of another line in two layers. Two lines may or may not be overlapped in the vertical direction. The fringe capacitance is modeled as $C_{fr} = C_{fr0} \cdot l \cdot (e^{-x1/x0} - e^{-x2/x0})$. $x_1$ is the distance from $l_1$ (side edge) to $l_2$ (near-end edge), and $x_2$ is the distance to $l_2$ (far-end edge). $l$ is the length of $l_1$ (side edge). $C_{fr0}$ and $x_0$ are model coefficients that are characterized based on different vertical profiles. In a special case, two side edges may coincide in $l_1$ and $l_2$ ($x_1$ = 0 and $x_2$ = width of $l_2$) and the model becomes $C_{fr} = C_{fr0} \cdot l \cdot (1 - e^{-x2/x0})$.

c. *Lateral capacitance:* the side surface of one line to the side surface of the adjacent line in the same layer. The lateral ca-

c01.qxd 12/16/2003 11:21 AM Page 12



pacitance is modeled as $C_{lt} = F_{l1l2}(d) \cdot l$, and $F_{l1l2}(d) = C_0 + C_1/d + C_2/d^2 + C_3/d^3 + C_4/d^4$. $l$ is the parallel length of two neighboring lines or conductors, $F_{l1l2}(d)$ is the lateral capacitance per unit length, and $d$ is the spacing between two lines. $C_0$, $C_1$, $C_2$, $C_3$, and $C_4$ are coefficients that are characterized for the given process technology.

## 1.3 MODELING OF SWITCHING CURRENTS

The high current consumption in some regions of the die produces "hot spots." In these hot spots, significant current transition occurs and the power network voltage fluctuation will be high. Accurate transition current modeling and power network simulation are necessary to calculate the noise and temperature distributions across the entire chip power network.

Figure 1-9(a) shows the current waveforms of multiple nearby drivers with three combinations of the transition patterns for these drivers. The simulation results are obtained when all drivers are charging (case: ALL UP), all on discharging (case: ALL DN), and half are charging and half discharging (case: UP_DN). In Figure 1-9(a), the X-coordinate is the time (ns) and the Y-coordinate is the voltage (V).

The waveforms illustrate the need to include the driver transition patterns (UP/DOWN) to model the transition currents. In our simulation, a 295.2 μm long bus with 130 signals is simulated in the minimum M5 width and pitch. Figure 1-9(b) shows the circuit schematic to be simulated. Figure 1-9(c) shows the entire power grid modeling for the simulation. Figure 1-9(d) shows the structure of bus lines and $V_{cc}/V_{ss}$ lines on the M5 layer included in the simulation.

In general, the total current consumption $I(t)$ of the CMOS circuit shown in Figure 1-10 consists of three components: $I_d$, $I_{sc}$, and $I_l$. $I_d$ is the charge or discharge current to the output load:

$$I_d = C_{\text{load}} V_{cc} f \tag{1-2}$$

In Equation (1-2), $C_{\text{load}}$ is the total output load of the driver, including the gate load and interconnect load; $V_{cc}$ is the supply voltage; and $f$ is the switching activity of $C_{\text{load}}$. Although the charge and discharge dynamic current $I_d$ is a predominant component of

c01.qxd 12/16/2003 11:21 AM Page 13



the total current consumption, other two current components ($I_{sc}$, $I_l$) are still significant in the submicron CMOS process.

The short-circuit current $I_{sc}$ is due to the fact that pMOS and nMOS transistors are both in the transition region of the inverter. The leakage current $I_l$ is due to the reverse-biased diode's leakage between the diffusion region and the substrate or well. Although the sum of the short-circuit and leakage currents accounts for less than 15% of the total current consumption of the microprocessor chip, the percentage will go up in future CMOS processes.

Figure 1-10(b) shows the current waveforms based on the estimated current components; the waveform is assumed to be a tri-

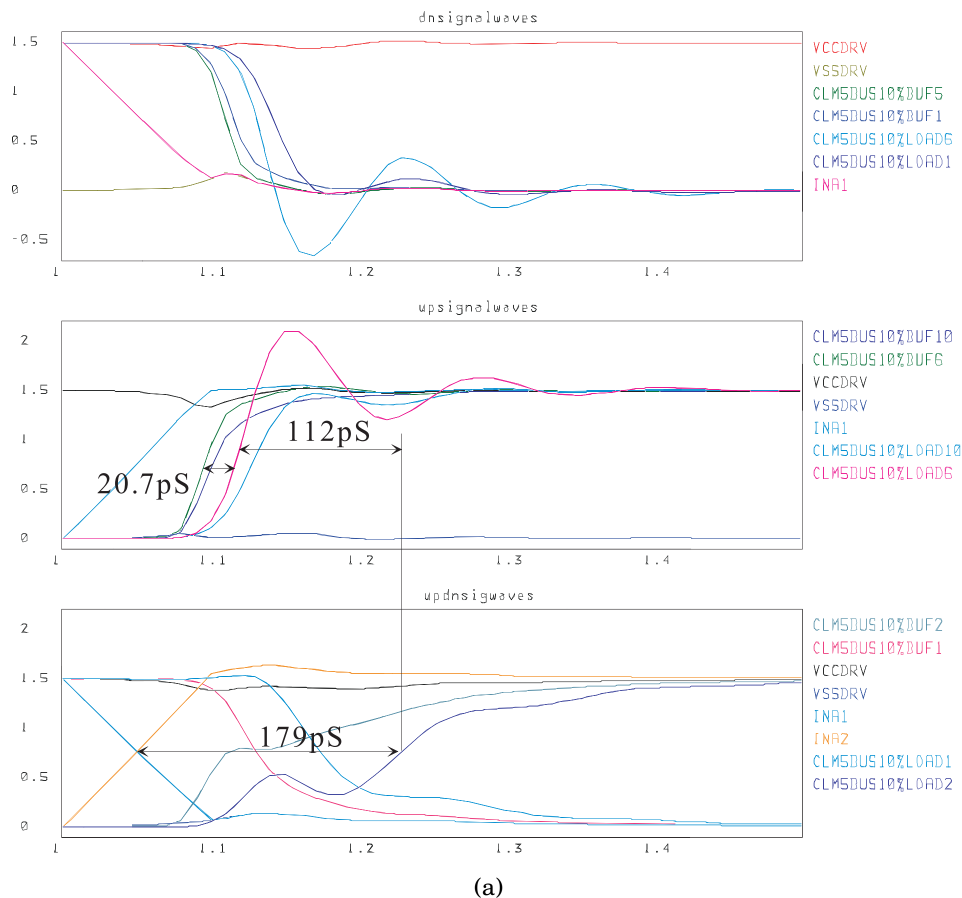
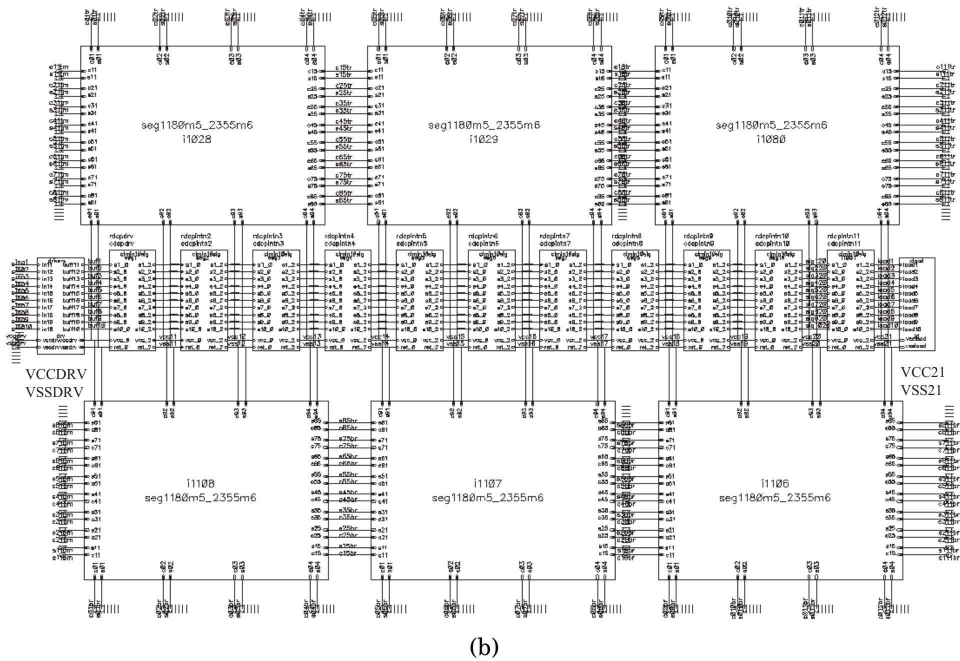


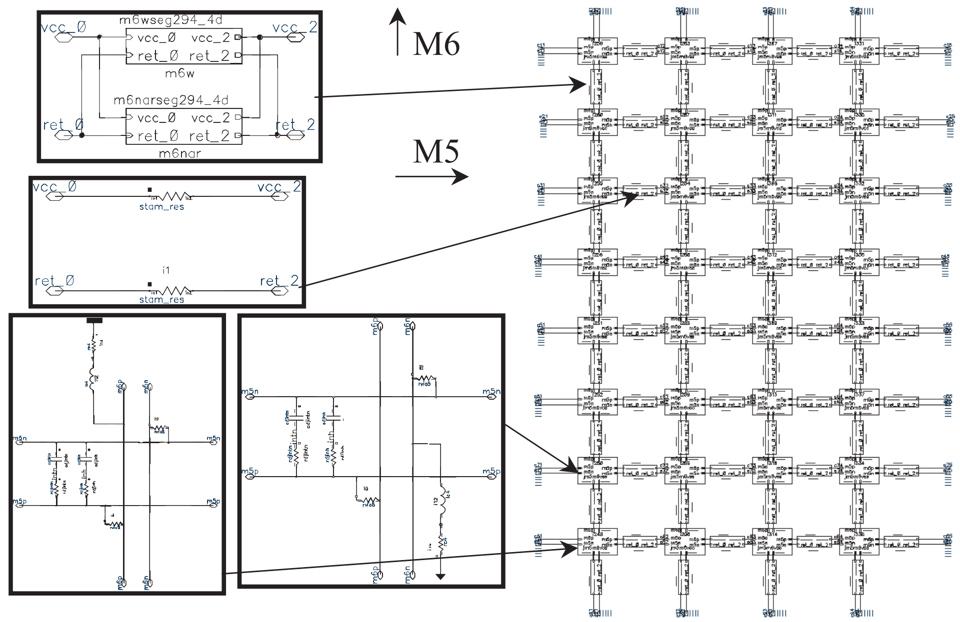

**Figure 1-9.** Switching noise simulation based on power grid modelling. (a) Simulation result. (*Figure continues on next page*)

c01.qxd 12/16/2003 11:21 AM Page 14



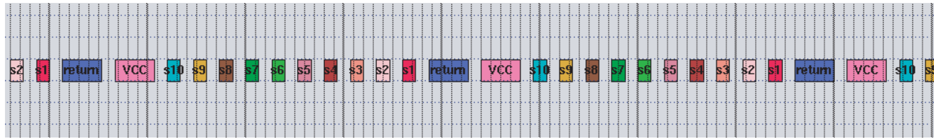



(b)



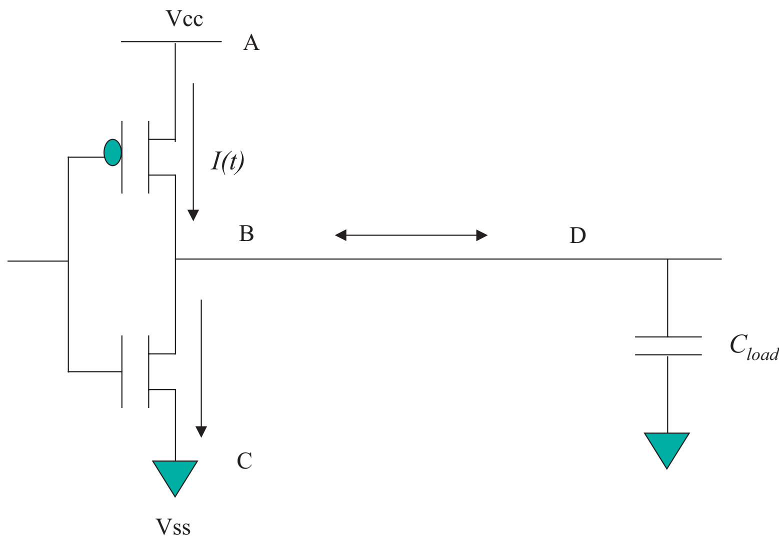

(c)

**Figure 1-9** *(continued).* (b) Simulated circuit. (c) M5 and M6 power grid modelling. *(Figure continues on next page)*

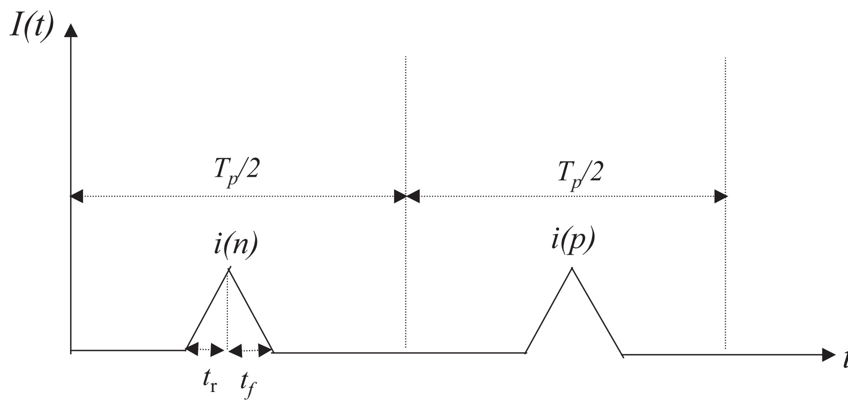



(d)

**Figure 1-9** *(continued).* (d) Bus lines layout structure.




(a)




(b)

**Figure 1-10.** Modeling of switching currents.

c01.qxd 12/16/2003 11:21 AM Page 16



angle. The current waveforms are back-annotated into the power network model, as shown in Figure 1-3. To improve the accuracy of the current waveforms, a current simulation tool such as Synopsys, Inc.'s PowerMill™ can be used, although the result largely depends on the (0, 1) patterns at the input ports.

## 1.4 ON-CHIP DECOUPLING CAPACITANCE

To prevent the supply level from collapsing when many gates switch simultaneously at the same clock transition, it is necessary to add decoupling capacitors at "hot spots" to reduce the peak voltage drops. These decoupling capacitors should be designed such that they do not occupy an excessively large area, which would decrease the yield.

It is important to realize that the on-chip decoupling capacitors reduce the $di/dt$ noise generated by the on-chip circuitry, but do not reduce the noise due to the simultaneous switching of off-chip drivers. Placing many low-inductance decoupling capacitors on the package and board to provide multiple low-inductance power/ground pins for output buffers should minimize the transient noise due to off-chip drivers.

If decoupling capacitors are placed, an upper limit or bound of the transient voltage fluctuation can be determined by modeling the power lines behind the capacitor as an infinitely large inductor. Immediately after switching, based on the decoupling capacitor model, as shown in Figure 1-4, no current flows through this large inductor and a capacitance divider is established based on the charge conservation law:

$$C_{\text{decap}} V_{CC} = (V_{CC} + \Delta V)(C_{\text{decap}} + C_{sw})$$
$$\Delta V = -\frac{C_{sw}}{C_{\text{decap}} + C_{sw}} V_{CC} \tag{1-3}$$

Based on Equation (1-3), to ensure a small voltage fluctuation $\Delta V$, the $C_{\text{decap}}$ (decoupling capacitance) should be much larger than the $C_{\text{sw}}$ (switching capacitance). Accordingly, for a microprocessor chip with a 14 nF load, we need $10 \cdot 14$ nF = 140 nF to achieve a 10% $V_{dd}$ power noise threshold in the worst case. Equation (1-3) provides the calculation of an upper bound of the total on-chip decoupling capacitance to satisfy the voltage fluctuation $\Delta V$ bound.

c01.qxd 12/16/2003 11:21 AM Page 17



The objective of the decoupling capacitance optimization problem is to minimize the total amount of decoupling capacitance as needed. Meanwhile, all the nodes in the power network model are satisfied with the specified supply voltage noise thresholds. Formally, we can describe the objective and constraints as follows [83]:

$$\text{Min} \sum_{n_i} (C_d)_i \qquad \text{Subject to } V_1 \leq V(n_i) \leq V_2 \tag{1-4}$$

In Equation (1-4), $(C_d)_i$ is the decoupling capacitance and $V(n_i)$ the voltage at node $n_i$ of the power network model, as shown in Figure 1-3; $V_1$ and $V_2$ are the lower and upper thresholds required for feasible supply voltages. We define a *noisy node* in the power network model as one in which, at some time, the voltage exceeds the required $[V_1, V_2]$ thresholds, as shown in Figure 1-11.

The thresholds are at the upper bound and lower bound away from the nominal supply voltages to guarantee the correct circuit timing. For example, with a nominal voltage of 1.3 V and 10% away allowed, the upper and lower thresholds are $[V_1, V_2]$ = [1.17 V, 1.43 V].

The power network, with each node's transient voltages in the electrical model satisfying the given thresholds, is called a *feasible power network*. Adding the decoupling capacitors at noisy nodes will turn a power network into a feasible one. Figure 1-12(a) shows

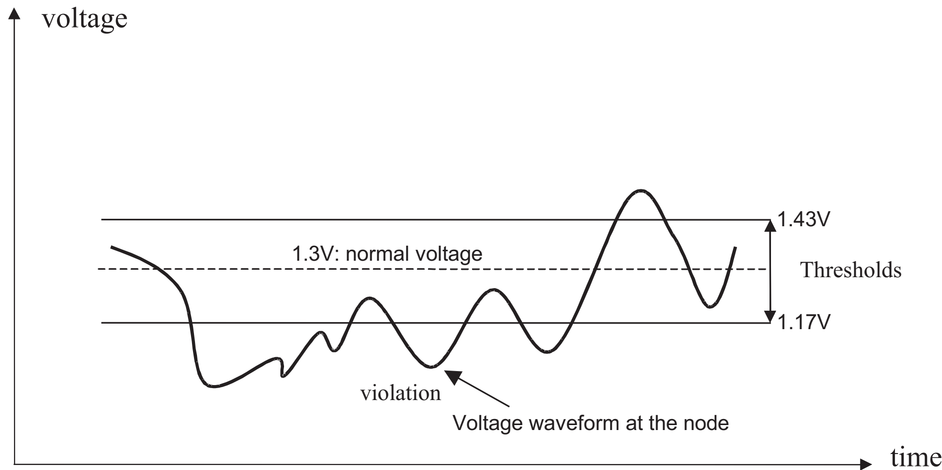
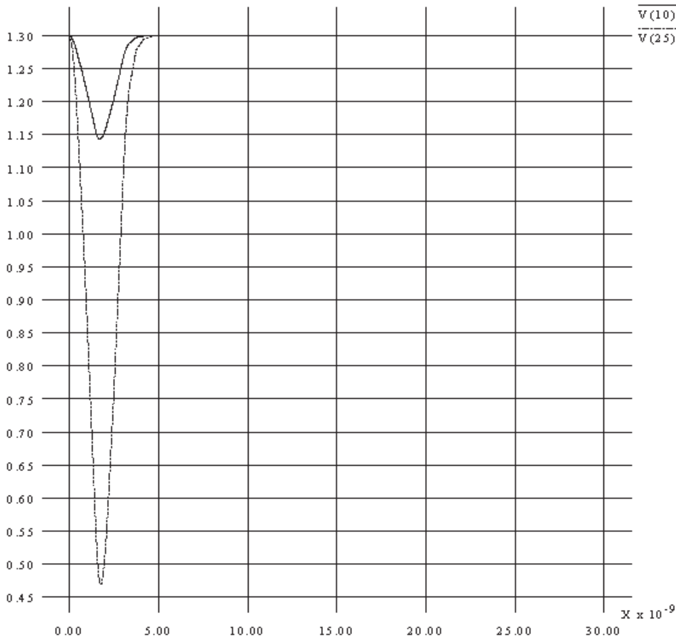



**Figure 1-11.** Supply voltage thresholds and noisy nodes definition [83].

c01.qxd 12/16/2003 11:21 AM Page 18






***Nominal voltage:***

Vcc = 1.3V

***Voltage thresholds:***

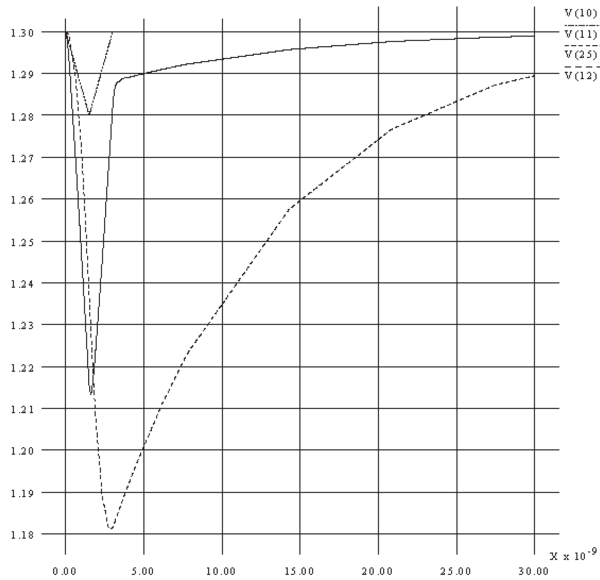
Vcc:[1.17V - 1.43V]

***Noisy nodes:***

Node 25 (min V = 0.47V)

Node 10 (min V = 1.15V)

(a)




***Nominal voltage:***

Vcc = 1.3V

***Voltage thresholds:***

Vcc:[1.17V - 1.43V]

***Noisy nodes:***

None

***Decoupling capacitors:***

Node 25

Node 10

(b)

**Figure 1-12.** Adding decoupling capacitors at noisy nodes [83]. (a) Nodes 10 and 25 are noisy. (b) Adding more capacitors on Nodes 10 and 25.

c01.qxd 12/16/2003 11:21 AM Page 19



one example with the simulated voltages of two nodes (Node 25 and Node 10) in the power network.

The minimum voltages (0.47 V and 1.15 V) of these nodes are less than the required lower threshold (1.17 V), and thus they are noisy nodes. The decoupling capacitor is added at each of these two noisy nodes and the voltages eventually satisfy the required thresholds, as shown in Figure 1-12(b).

Figure 1-13 shows the high-level decoupling capacitance optimization flow [83]. Procedure I adds the decoupling capacitors at the noisy nodes. Procedure II removes the unnecessary decoupling capacitance overallocated initially.

We have done experiments on a power network model with about 100 *RLC* grids and decoupling capacitors. Current sources have been added at each node in the model for transistor transitions with the current waveforms, as shown in Figure 1-10(b). The

**Procedure I: Decoupling Capacitance Increment**

```
Simulate the power network model with RLC elements and current sources.
Identify the “noisy” nodes by comparing the voltage results with the specified thresholds.
While (there is “noisy” node){
    For (each “noisy” node){
        Add a step size of the decoupling capacitance.
    }
    Simulate the power network model with the updated decoupling capacitance.
    Identify “noisy” nodes by comparing simulation voltages with the required thresholds.
}
```

**Procedure II: Decoupling Capacitance Decrement**

```
For (each node){
    Mark the node as “deductible”;
}
While (there is still “deductible” node){
    Deduct a step size of decoupling capacitance from each “deductible” node;
    Simulate the power network model with the updated decoupling capacitance;
    Identify the “noisy” nodes by comparing simulation voltages with the required thresholds;
    For (each “noisy” node){
        Add a step size of the decoupling capacitance;
        Make the node as “nondeductible”;
    }
}
```

**Figure 1-13.** Decoupling capacitance optimization flow [83].

c01.qxd 12/16/2003 11:21 AM Page 20



cycle time is 3 ns or 330 MHz frequency in the experiments. Two voltage sources are added to model the C4 package power pads. The *RL* parasitic (200 Ω and 0.5 nH) of the package layer are included in the model. The nominal supply voltage is 1.3 V.

The power grid simulation is done using a fast linear circuit simulator [20]. The flow shown in Figure 1-13 is used to determine the locations and amounts of on-chip decoupling capacitors. Figure 1-14 shows the experimental results for a sensitivity study to decoupling capacitances. The decoupling capacitance is most sensitive to the changes in the noise margin and device transition currents.

This suggests to us that the model of the current consumption is the key to getting the accurate voltage drop and decoupling capacitance amounts. In addition, we want to reduce the on-chip decoupling capacitance size by improving the noise margin. This can be achieved by improving the power distribution on the package and the board. The changes of power line *RLC* values, as well as the absolute supply voltages with the same noise thresholds, do not show significant impact on the decoupling capacitance.

In the experiment, we assigned the initial *RLC* values at each node of the power network as follows: $R = 40\ \Omega$, $L = 0.005$ nH, $C = 0.3$ pF (without the decoupling capacitance at this initial assignment). The change of on-chip power line inductance does not lead to a lot of variation in decoupling capacitance, as shown in Figure 1-14(b); this is due to the very small $L/R$ delay (0.12 ps) compared to the *RC* delay (12 ps) in this example.

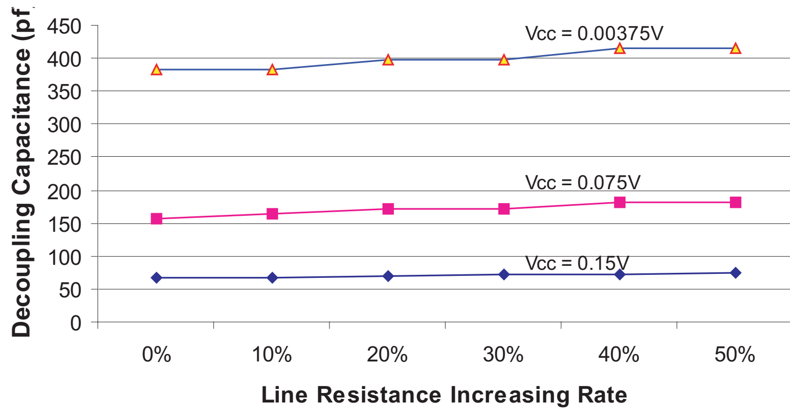
The decoupling capacitor can be improved by using either the PN junction or a MOS varactor device [43]. As shown in Figure 1-15(a), the PN junction is formed by diffusing p+ doping in an n-well. As shown in Figure 1-15(b), the MOS varactor is formed by placing an nMOS in an n-well. The n-well is added to form a channel between the source and drain. In addition, $V_{\text{tune}}$ and $V_{\text{gate}}$ voltages are controlled to vary the gate capacitance used for the decoupling capacitances between $V_{dd}$ and $V_{ss}$.

## 1.5 ON-CHIP INDUCTANCE

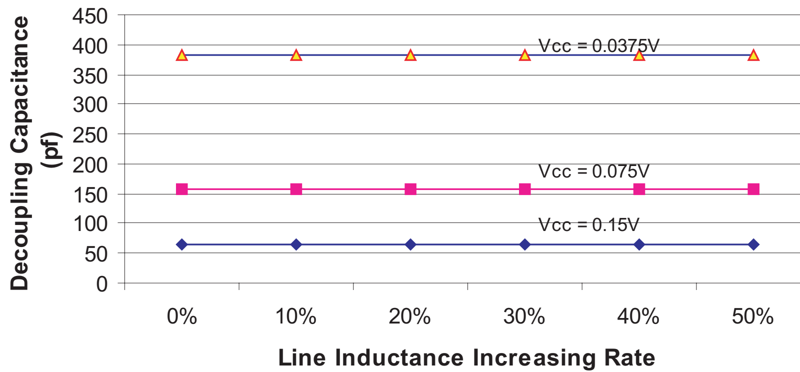
The inductive drop or noise ($L \cdot di/dt$) on the power lines becomes significant for high-speed microprocessor chips [14, 15], especially

c01.qxd 12/16/2003 11:21 AM Page 21

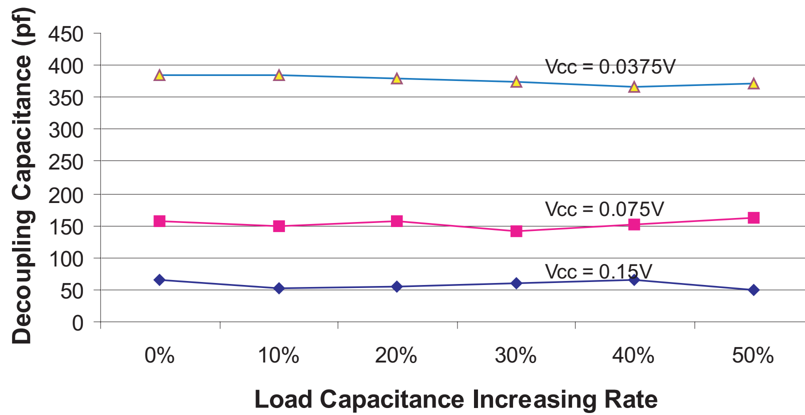





(a)

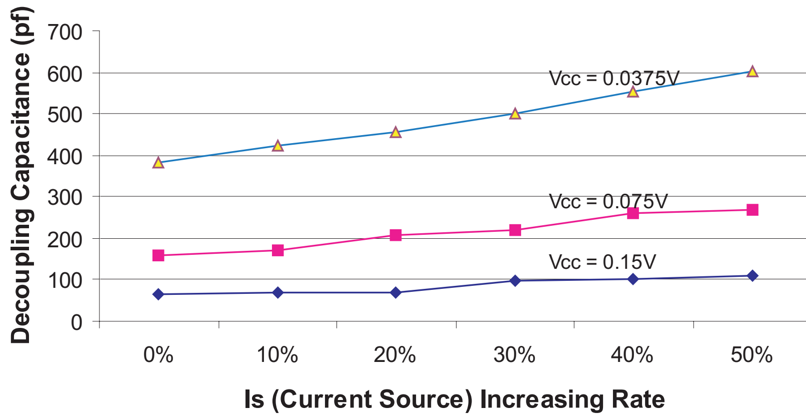



(b)

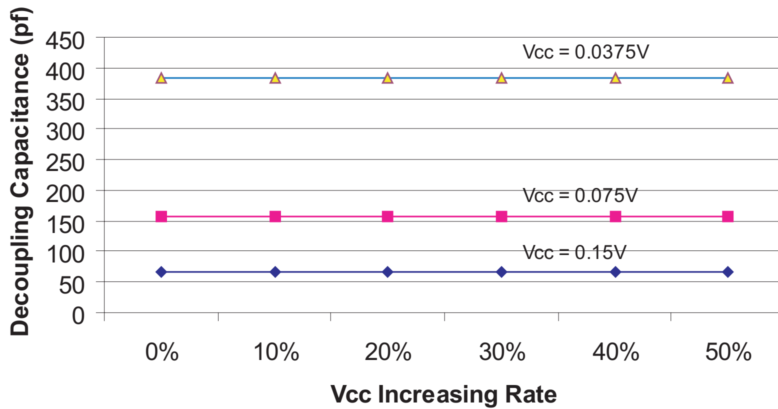



(c)

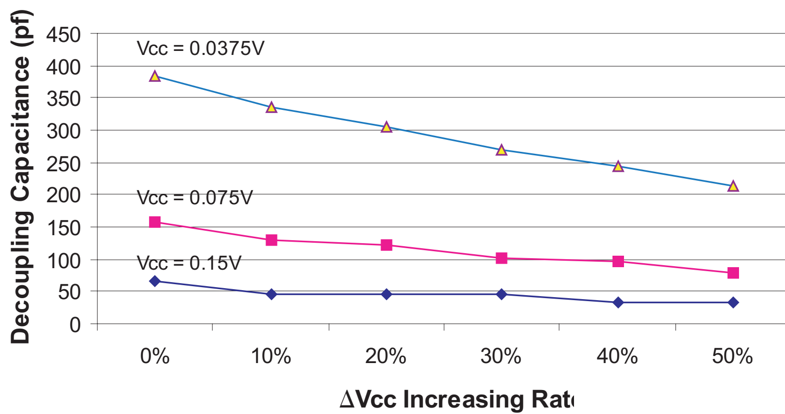
**Figure 1-14.** Sensitivity study of on-chip decoupling capacitances [83]. *(Figure continues on next page)*




(d)




(e)




(f)

**Figure 1-14** (*continued*).

c01.qxd 12/16/2003 11:21 AM Page 23



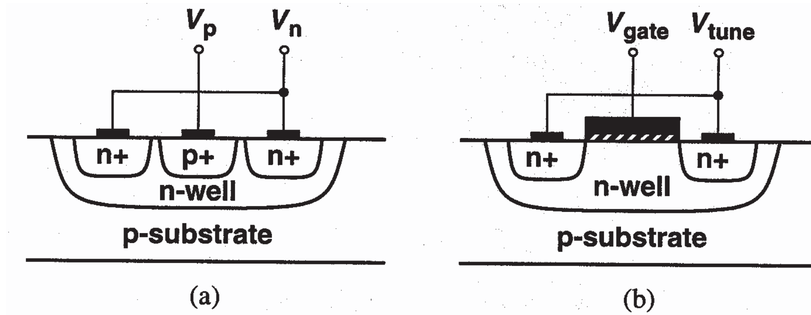



**Figure 1-15.** Decoupling capacitor [43]. (a) PN Junction. (b) MOS varactor.

when the chip becomes faster and larger in size. The characteristic impedance is $Z_0 = \sqrt{L/C}$. Adding decoupling capacitors will increase the capacitance but does not affect the inductance of the power planes. As a result, $Z_0$ is reduced, and current spikes generate smaller voltage drops because $\Delta V = Z_0 \Delta I$

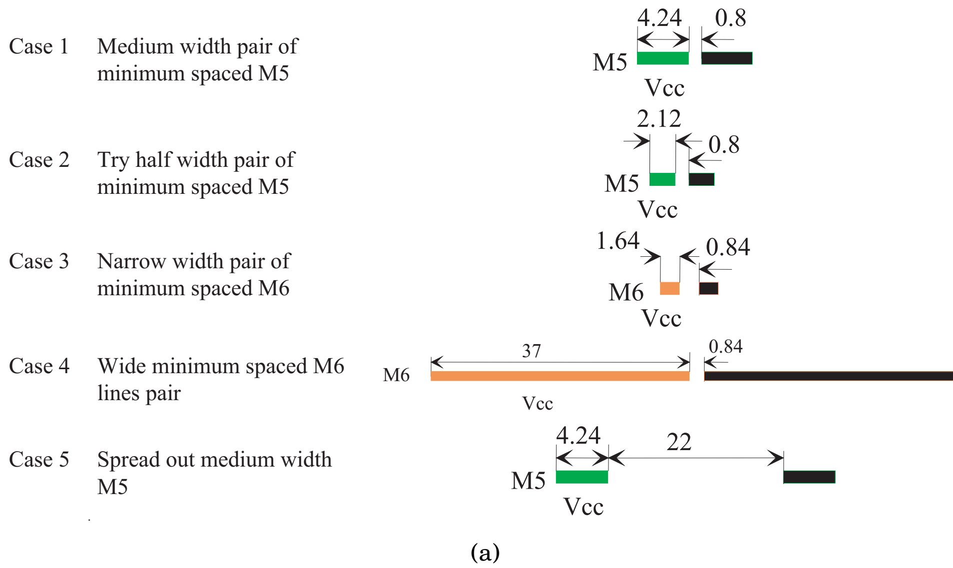
Low impedance of the power network helps the pulse response and curbs the instantaneous fluctuations. The impedance $Z_0$ can be further reduced by lowering the inductance $L$ of the power network. This section presents a metal wire design method to reduce the inductance by carefully selecting the sizes and spaces of power lines.

Figure 1-16(a) shows five different combinations of the widths and spaces for two adjacent $V_{cc}$ and $V_{ss}$ lines [21]. The inductance and resistance of these five combinations are shown in Figure 1-16(b) and Figure 1-16(c) for 10,000 μm long power lines. The inductance is calculated by using a two-dimensional model with the current loops between adjacent $V_{ss}$ and $V_{cc}$ lines. The first-order estimation of the unit-length loop inductance for two adjacent $V_{cc}$ and $V_{ss}$ lines is as follows:

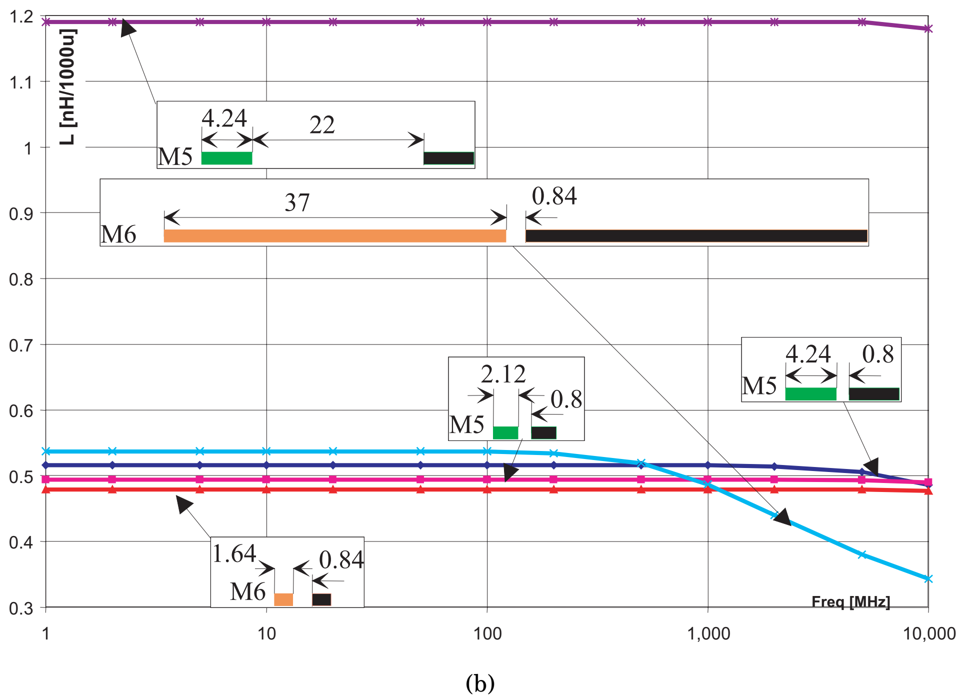
$$L = \mu \frac{s}{w} \tag{1-5}$$

In Equation (1-5), $\mu$ is the permeability of the dielectric material between adjacent $V_{cc}$ and $V_{ss}$ lines, $s$ the space between the $V_{cc}$ and $V_{ss}$ lines, and $w$ the width of $V_{cc}$ or $V_{ss}$ lines. The $V_{cc}$ and $V_{dd}$ nets are interchangeable in this book. Usually, $V_{cc}$ is used for the analog signal and $V_{dd}$ for digital design.

The inductance becomes large when the line space is big, which




(a)

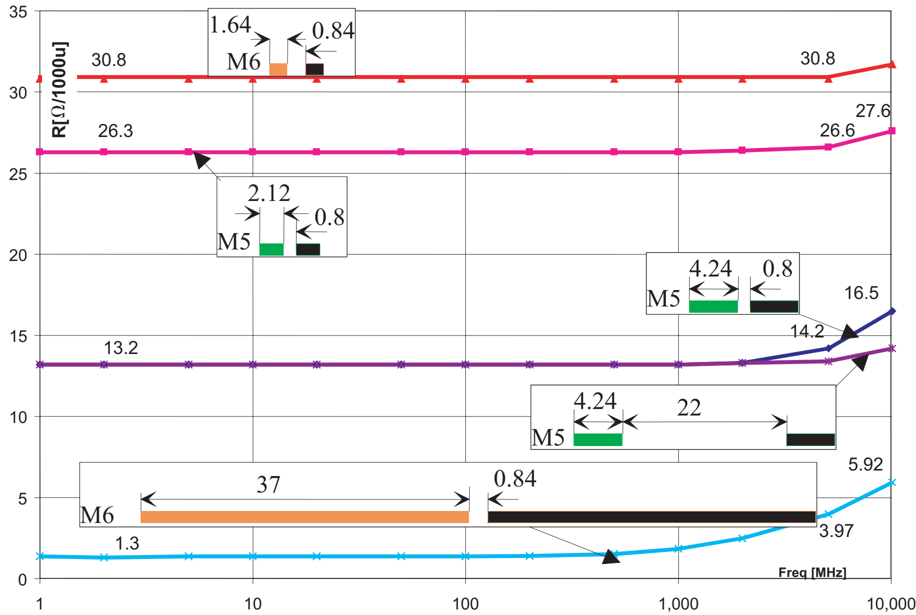



(b)

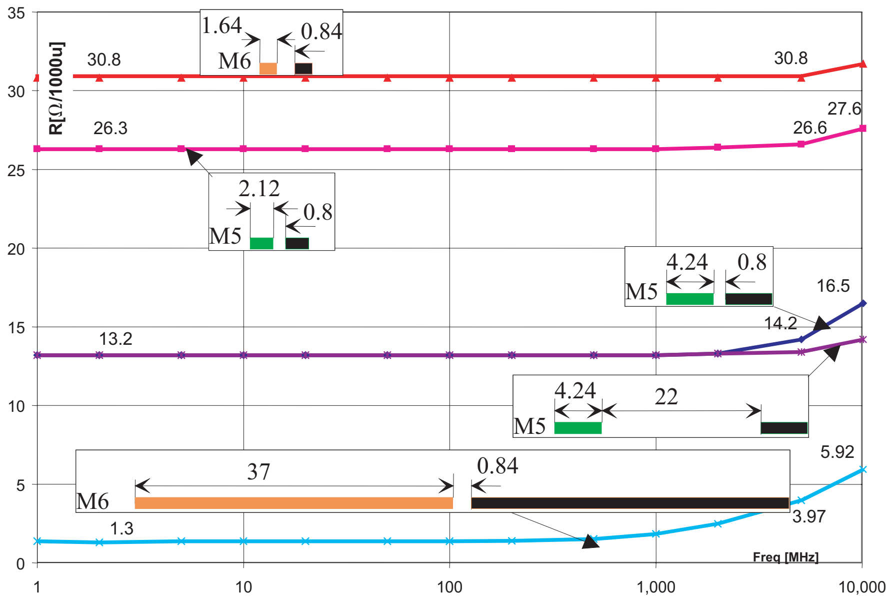
**Figure 1-16.** Characterization results of $V_{dd}/V_{ss}$ metal structures [21]. (a) $V_{cc}$ and $V_{ss}$ cases. (b) On-chip inductance characterizations.

c01.qxd 12/16/2003 11:21 AM Page 25






(c)

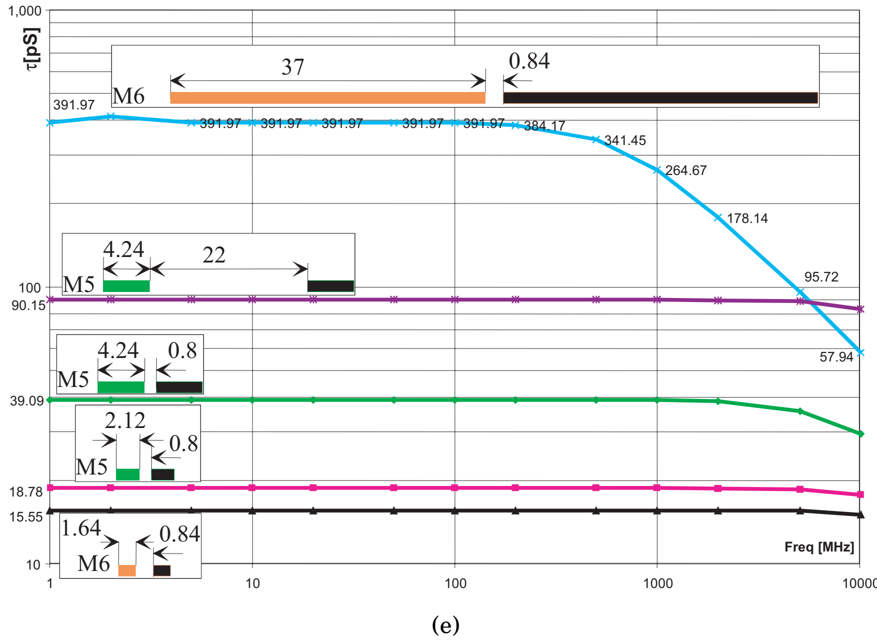



(d)

**Figure 1-16** *(continued)*. (c) Resistance characterizations. (d) Impedance calculation. *(Continued on next page)*

c01.qxd 12/16/2003 11:21 AM Page 26






(e)

**Figure 1-16** *(continued)*. (e) $L/R$ delay.

is opposite to the case of line-to-line capacitance coupling. Case 5 has far more inductance than any other cases, since it has a large line-to-line space. More magnetic coupling is caused by two conductors in the far distance and that is one of difficulties in accurate inductance modeling.

The inductance is reduced at high frequencies because time varying currents tend to concentrate near the surface of the conductors at high frequencies; this is known as the *skin effect* [6].

As a consequence of this electromagnetic induction phenomenon, the magnitude of the current density drops exponentially with the distance away from the surface. The distance at which the current density becomes a fraction $1/e$ of its value at the surface is called *skin depth,* which is calculated by

$$\sigma_s = \sqrt{\frac{\rho}{\pi \mu f}} \tag{1-6}$$

In Equation (1-6), $f$ is the frequency, and $\mu$ and $\rho$ are the permeability and resistivity of the material. Making the thickness of the

c01.qxd 12/16/2003 11:21 AM Page 27



conductor larger than approximately $2\sigma_s$ will not reduce the effective resistance of the line.

Figure 1-16(c) shows the resistance plots over the frequency for the five line configurations shown in Figure 1-16(a). The skin effects are observed at the higher frequencies with the increased resistances for all configurations. Case 4, shown in Figure 1-16(c), which has the largest width, shows the skin effect at the lowest frequency due to its large width.

The impedance of a power line is calculated as follows:

$$|Z(f)| = \sqrt{R^2 + (2\pi f L)^2} \tag{1-7}$$

In Equation (1-7), $f$ is the clock frequency and $R$ and $L$ are unit-length line resistance and unit-length line inductance. Figure 1-16(d) shows the impedance as the frequency functions of the $V_{cc}$ and $V_{ss}$ line configurations shown in Figure 1-16(a).

At the high frequency, the impedance is rising, especially for Case 5, due to the inductance effect, as shown in Figure 1-16(d). Case 4, shown in Figure 1-16(a) with the largest wire width and small line space, has the smallest impedance.

The *inductance delay* due to the line inductance and line resistance is calculated as follows:

$$\tau = L/R \tag{1-8}$$

The $L/R$ delay characterizes the importance of the inductance in power network modeling. Figure 1-16(e) shows the $L/R$ delay results; Case 2 and Case 3, with small line widths and small line spaces, have the smallest $L/R$ delay, as small as 15–19 ps for a 10000 μm long power line.

If the $L/R$ delay is much smaller than the $RC$ delay per unit length, the line inductance $L_{vcc}$ or $L_{vss}$ can be ignored in the on-chip power network model. In this condition, the $RC$ network is accurate enough to model the on-chip power network.

Based on the experimental results shown in Figure 1-16(e), we can conclude that narrow and dense lines are preferred in the power network design for metal inductance reduction. However other effects, like the $IR$ drop, need to be considered as well.

Just considering how to reduce the inductance effect through wire sizing is not very useful since the inductance is still dominated by the package in modern chips. But we can use dense and narrow lines for reducing both on-chip inductance and resistance. An

c01.qxd 12/16/2003 11:21 AM Page 28



example is shown in Figure 1-17. The inductance is obviously reduced based on our experiments.

The resistance of these narrow lines combined is equal to, or less than, a wide line. The example in Figure 1-17 shows a practical guideline used in the Intel microprocessor power network design.

## 1.6 PROCESS SCALING IMPACTS

We have considered two scenarios for the technology scaling in microprocessor chips. Scenario A scales the existing chip to a new process with a scaling factor $S$ with little logic change. In Scenario A, die size is reduced by $S^2$. Scenario B scales the existing chip to a new process with lots of new logics implemented.

In Scenario B, the die size is assumed to be unchanged when using the new process due to more transistors employed in the new design. Table 1-3 shows the impact on the microprocessor power distribution of using the above two scaling scenarios for the microprocessor chips. The detailed derivations are given below.

### Scenario A

The line width and space are both reduced by $S$, assuming the line thickness change is negligible in process shrinking. The unit-length resistance is increased by $1/S$. The unit-length capacitance is reduced in $S$ by assuming that the plate capacitance is reduced by $1/S^2$ but the coupling capacitance increases by $1/S$ due to the smaller line space.

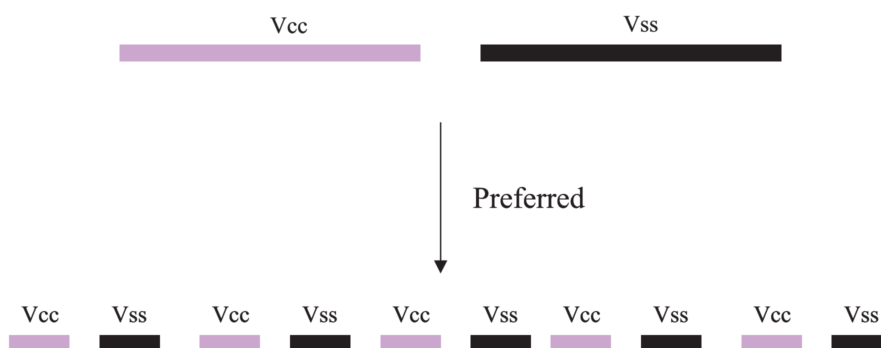



**Figure 1-17.** Design guidelines for on-chip power lines.

c01.qxd 12/16/2003 11:21 AM Page 29



**Table 1-3.** Technology scaling model for microprocessor power distribution

| | Design Parameters | Scenario A | Scenario B |
|---|---|---|---|
| Dimensions | Die size | $S^2$ (down) | Unchanged |
| | Transistor count | Unchanged | $1/S$ (up) |
| | Metal width | $S$ (down) | $S$ (down) |
| | Metal space | $S$ (down) | $S$ (down) |
| | Metal thickness | Unchanged | Unchanged |
| | Global metal length | $S$ (down) | Unchanged |
| | Decoupling capacitance bound | $S^2$ (down) | Unchanged |
| | Area % of decoupling capacitor | Unchanged | Unchanged |
| *RLC* Parameters | Metal resistance | Unchanged | $1/S$ (up) |
| | Metal capacitance | $S^2$ (down) | $S$ (down) |
| | Loop inductance | $S$ (down) | Unchanged |
| | Clock frequency | $1/S^2$ (up) | $1/S^2$ (up) |
| | Toggling transistors per cycle | Unchanged | $1/S$ (up) |
| | Average gate capacitance | $S^2$ (down) | $S^2$ (down) |
| | Total gate capacitance | $S^2$ (down) | $S$ (down) |
| | Total signal connections | Unchanged | $1/S^2$ (up) |
| | Total wire capacitance | $S^2$ (down) | $1/S$ (up) |
| | Total toggling capacitance | $S^2$ (down) | Unchanged |
| Power Consumption | Power consumption (total) | $S^2$ (down) | Unchanged |
| | Supply current (total) | $S$ (down) | $1/S$ (up) |
| | Current density on power line | Unchanged | $1/S^2$ (up) |
| Voltage Drop | Supply voltage | $S$ (down) | $S$ (down) |
| | IR drop | $S$ (down) | $1/S^2$ (up) |
| | $L \cdot Di/Dt$ drop | $S^2$ (down) | $1/S$ (up) |

The die size is reduced by $S^2$, and the length of power lines is scaled in $S$. The line resistance for the power network is not changed, and the line capacitance for the power network or long signal lines is reduced by $S^2$.

Based on Equation (1-5), the unit-length inductance between two adjacent $V_{cc}$ and $V_{ss}$ lines is not changed, because the line space ($s$) and line width ($w$) are both reduced by $S$. The total line inductance is reduced in $S$, due to the power line length scaled in $S$.

Chip clock frequency is assumed to increase by $1/S^2$, which is a simplification of the fact that the microprocessor frequency will roughly double every two years for the next process generation. In Scenario A, the logic of the chip is changed very little and the number of toggling transistors per clock cycle is kept unchanged.

The channel length and width of each device are both scaled down in $S$. The average gate capacitance is down by $S^2$. So the total

c01.qxd 12/16/2003 11:21 AM Page 30



gate capacitance is down by $S^2$. Since the total wire capacitance of signals is also down in $S^2$, with unchanged transistor numbers and signal connections, the total toggling capacitance ($C_{toggle} = C_{gate} + C_{wire}$) of the chip is reduced by $S^2$. The supply voltage is scaled in $S$ at each process generation, as shown in Figure 1-1(a).

The power consumption can be estimated as: $0.5 \cdot f \cdot V^2_{dd} \cdot C_{toggle}$, where $f$ = clock frequency, $V_{dd}$ = supply voltage, and $C_{toggle}$ = total toggling capacitance of the chip. The power consumption is reduced by $S^2$ based on the above assumptions for the frequency, supply voltage, and the total toggling capacitance per clock cycle.

The current of the power distribution network is calculated by the power consumption divided by the supply voltage. Since the power is down by $S^2$ and $V_{dd}$ is down by $S$, the current is thus down by $S$. Since the line width is down by $S$ and current down by $S$, the current density of the power line is not changed.

The $IR$ drop is down by $S$, since the line resistance is not changed but the current is reduced in $S$. The $L \cdot di/dt$ voltage drop is reduced by $S^2$ because the line inductance $L$ is scaled down by $S$; $di$ (current) is reduced by $S$ for the same $dt$ period.

Based on Equation (1-3), we got the bound of the total on-chip decoupling capacitance with 10 times the total toggling capacitance to achieve 10% $V_{dd}$ noise bound. Because the total toggling capacitance is reduced by $S^2$, the upper bound of the total decoupling capacitance needed in the chip is also reduced by $S^2$.

Since the die size is reduced by $S^2$ in Scenario A, the percentage of die size used for the on-chip decoupling capacitance is not changed in this scenario.

### Scenario B

The die size is assumed to be not changed in this scenario, so the global line length is not changed. The line resistance of the power network is increased by $1/S$. The line capacitance of the power network, or long signals, is reduced in $S$, since the unit-length capacitance is down in $S$, as derived in Scenario A.

Based on Equation (1-5), the unit-length inductance between two adjacent $V_{cc}$ and $V_{ss}$ lines is not changed due to the line space ($s$) and the line width ($w$), both reduced by $S$. The total line inductance is not changed because the global line length is not changed.

The chip clock frequency is supposed to increase by $1/S^2$ about every two years for each process generation. In Scenario B, new

logic features are implemented, assuming employment of $1/S$ more transistors in the design. Therefore, the total toggling transistors per cycle increases by $1/S$.

The gate channel length and channel width are both scaled down by $S$, so each gate capacitance is down by $S^2$ and the total gate capacitance is down by $S$. The total signal number is increased by $1/S^2$, for $1/S$ more transistors used in the design. This implies that the total wire capacitance of signals in this chip is increased by $1/S$, based on the unit line capacitance in this scenario being reduced by $S$.

If we assume that the total wire capacitance is almost equal to the total gate capacitance across a chip (and that is the case we found in a microprocessor chip), we get the unchanged total toggling capacitance, $C_{toggle}$ ($C_{toggle} = C_{gate} + C_{wire}$). The supply voltage is reduced in $S$ at each process generation.

The average power consumption is calculated by $0.5 \cdot f \cdot V_{dd}^2 \cdot C_{toggle}$, where $f$ = clock frequency, $V_{dd}$ = supply voltage, and $C_{toggle}$ = toggling capacitance. The power consumption is unchanged in this scenario. The current through the power distribution network is calculated by the power consumption divided by the supply voltage. Since the power is unchanged and $V_{dd}$ is down by $S$, the total current increases by $1/S$.

Because the wire width is down by $S$ and current increases by $S$, the current density of the power network increases by $1/S^2$. The IR drop increases by $1/S^2$, due to the line resistance increasing by $1/S$ and the supply current also increases by $1/S$. The $L \cdot di/dt$ noise increases by $1/S$ since $L$ not changed; $di$ (current) increases by $1/S$ for the same $dt$ period.

Because the total toggling capacitance per cycle is unchanged, the upper bound of the total on-chip decoupling capacitance is also unchanged, based on Equation (1-3). Since the die size is not changed in Scenario B, the area percentage used for the on-chip decoupling capacitance is also unchanged.

Although the scaling models show unchanged power consumption in Scenario B, for most new microprocessors we see more aggressive transistor number increase or more parallelism used for higher performance. This observation results in more power consumption in new microprocessors. For example, Alpha 21264 (0.35 μm) has 1.63 times more transistors than Alpha 21164 (0.50 μm) (> 1/0.7 = 1.42 scaling factor assumed in Scenario B), and the power consumption is increased from 50 W to 72 W [1].

c01.qxd 12/16/2003 11:21 AM Page 32



Process scaling factor $S$ in Table 1-3 is the ratio of the minimum feature sizes between two process generations. $S$ is about 0.7 [10]. For example, an 0.18 μm process is scaled to 0.13 μm for a scaling factor $S$ of about 0.72 (0.13/0.18 = 0.72).

## 1.7 SUMMARY

This chapter discusses the modeling issues of on-chip power grids. It provides the primary models and characterization results for the resistance, capacitance, and inductance associated with metal lines and vias to route the power distribution network on the chip. The power distribution network, in general, can be characterized as a low-pass $RLC$ filter for the frequency domain analysis.

In addition, the resonant frequency should be removed from the working frequency of the circuit; otherwise, this $RLC$ network will generate a lot of noise. We describe the inductance effects for the on-chip power grid. Usually, very dense and narrow width $V_{ss}$ and $V_{cc}$ lines are interleaved with each other to reduce the inductance.

In general, as a designer of a power grid, you want to increase the capacitance while reducing the resistance and inductance. The latter two parameters are associated with the $IR$ drop and $L \cdot di/dt$ noise.

The capacitance increase for a power grid is implemented by adding intentional decoupling capacitors. In addition, decoupling capacitors are inserted at the noisy nodes of the power distribution network. A CAD algorithm has been proposed to automate this decoupling capacitor insertion process [83].

Finally, we predict future design directions by providing technology scaling models related to power distribution performance and voltage drop based on two different chip improvement scenarios.