12

Scaling Trends of On-Chip Power Distribution Noise

A scaling analysis of the voltage drop across the on-chip power distribution networks is performed in this chapter. The design of power distribution networks in high performance integrated circuits has become significantly more challenging with recent advances in process technology. Insuring adequate signal integrity of the power supply has become a primary design issue in high performance, high complexity digital integrated circuits. A significant fraction of the on-chip resources is dedicated to achieve this objective.

State-of-the-art circuits consume higher current, operate at higher speeds, and have lower noise tolerance with the introduction of each new technology generation. CMOS technology scaling is forecasted to continue for at least another ten years [279]. The scaling trend of noise in high performance power distribution grids is, therefore, of practical interest. In addition to the constraints on the noise magnitude, electromigration reliability considerations limit the maximum current density in on-chip interconnect. The scaling of the peak current density in power distribution grids is also of practical interest. The results of this scaling analysis depend upon various assumptions. Existing scaling analyses of power distribution noise are reviewed and compared along with any relevant assumptions. The scaling of the inductance of an on-chip power distribution network as discussed here extends the existing material presented in the literature. Scaling trends of on-chip power supply noise in ICs packaged in high performance flip-chip packages are the focus of this investigation.

The chapter is organized as follows. Related existing work is reviewed in Section 12.1. The interconnect characteristics assumed in the analysis are discussed in Section 12.2. The model of the on-chip power

Table 12.1. Ideal scaling of CMOS circuits [280]

Parameter	Scaling factor
Device dimensions	$1/S$
Doping concentrations	S
Voltage levels	$1/S$
Current per device	$1/S$
Gate load	$1/S$
Gate delay	$1/S$
Device area	$1/S^2$
Device density	S^2
Power per device	$1/S^2$
Power density	1
Total capacitance	SS_C^2
Total power	S_C^2
Total current	SS_C^2

distribution noise used in the analysis is described in Section 12.3. The scaling of power noise is described in Section 12.4. Implications of the scaling analysis are discussed in Section 12.5. The chapter concludes with a summary.

12.1 Prior work

Ideal scaling of CMOS transistors was first described by Dennard *et al.* in 1974 [280]. Assuming a scaling factor S, where $S > 1$, all transistor dimensions uniformly scale as $1/S$, the supply voltage scales as $1/S$, and the doping concentrations scale as S. This "ideal" scaling maintains the electric fields within the device constant and ensures a proportional scaling of the I–V characteristics. Under the ideal scaling paradigm, the transistor current scales as $1/S$, the transistor power decreases as $1/S^2$, and the transistor density increases as S^2. The transistor switching time decreases as $1/S$, the power per circuit area remains constant, while the current per circuit area scales as S. The die dimensions increase by a chip dimension scaling factor S_C. The total capacitance of the on-chip devices and the circuit current both increase by SS_C^2 while the circuit power increases by S_C^2. The scaling of interconnect was first described by Saraswat and Mohammadi [281]. These ideal scaling relationships are summarized in Table 12.1.

Several research results have been published on the impact of technology scaling on the integrity of the IC power supply [132], [220], [282], [283]. The published analyses differ in the assumptions concerning the on-chip and package level interconnect characteristics. The analyses can be classified according to several categories: whether resistive IR or inductive $L\frac{dI}{dt}$ noise is considered, whether wire-bond or flip-chip packaging is assumed, and whether packaging or on-chip interconnect parasitic impedances are assumed dominant. Traditionally, the package-level parasitic inductance (the bond wires, lead frames, and pins) has dominated the total inductance of the power distribution system while the on-chip resistance of the power lines has dominated the total resistance of the power distribution system. The resistive noise has therefore been associated with the resistance of the on-chip interconnect and the inductive noise has been associated with the inductance of the off-chip packaging [283], [284], [285].

Scaling behavior of the resistive voltage drop in a wire bonded integrated circuit of constant size has been investigated by Song and Glasser in [282]. Assuming that the interconnect thickness scales as $1/S$, the ratio of the supply voltage to the resistive noise, *i.e.*, the signal-to-noise ratio (SNR) of the power supply voltage, scales as $1/S^3$ under ideal scaling (as compared to $1/S^4$ under constant voltage scaling). Song and Glasser proposed a multilayer interconnect stack to address this problem. Assuming that the top metal layer has a constant thickness, scaling of the power supply signal-to-noise ratio improves by a power of S as compared to standard interconnect scaling.

Bakoglu [132] investigated the scaling of both resistive and inductive noise in wire-bonded ICs considering the increase in die size by S_C with each technology generation. Under the assumption of ideal interconnect scaling (*i.e.*, the number of interconnect layers remains constant and the thickness of each layer is reduced as $1/S$), the SNR of the resistive noise decreases as $1/S^4S_C^2$. The SNR of the inductive noise due to the parasitic impedances of the packaging decreases as $1/S^4S_C^3$. These estimates of the SNR are made under the assumption that the number of interconnect levels increases as S. This assumption scales the on-chip capacitive load, average current, and, consequently, the SNR of both the inductive and resistive noise by a factor of S. Bakoglu also considered an improved scaling situation where the number of chip-to-package power connections increases as SS_C^2, effectively assuming flip-chip packaging. In this case, the resistive SNR_R scales as 1 assuming

that the thickness of the upper metal levels is inversely scaled as S. The inductive SNR_L scales as $1/S$ under the assumption that the effective inductance per power connection scales as $1/S^2$.

A detailed overview of modeling and mitigation of package-level inductive noise is presented by Larsson [283]. The SNR of the inductive noise is shown to decrease as $1/S^2S_C$ under the assumption that the number of interconnect levels remains constant and the number of chip-to-package power/ground connections increases as SS_C. The results and key assumptions of the power supply noise scaling analyses are summarized in Table 12.2.

The effect of the flip-chip pad density on the resistive drop in power supply grids has been investigated by Arledge and Lynch in [220]. All other conditions being equal, the maximum resistive drop is proportional to the square of the pad pitch. Based on this trend, a pad density of 4000 pads/cm^2 is the minimum density required to assure an acceptable on-chip IR drop and I/O signal density at the 50 nm technology node [220].

Nassif and Fakhouri describe an analytic expression relating the maximum power distribution noise to the principal design and technology characteristics [286]. The expression is based on a lumped model similar to the model depicted in Fig. 12.3. The noise is shown to increase rapidly with technology scaling based on the ITRS predictions [287]. Assuming constant inductance, a reduction of the power grid resistance and an increase in the decoupling capacitance are predicted to be the most effective approaches to decreasing the power distribution noise.

12.2 Interconnect characteristics

The power noise scaling trends depend substantially on the interconnect characteristics assumed in the analysis. The interconnect characteristics are described in this section. The assumptions concerning the scaling of the global interconnect are discussed in Section 12.2.1. Variation of the grid inductance with interconnect scaling is described in Section 12.2.2. Flip-chip packaging characteristics are discussed in Section 12.2.3. The impact of the on-chip capacitance on the results of the analysis is discussed in Section 12.2.4.

Table 12.2. Scaling analyses of power distribution noise

Scaling analysis	Noise type	Noise scaling	SNR scaling	Analysis assumptions	
Glasser and Song [282]	On-chip IR noise	S^2	$1/S^3$	Ideal interconnect scaling	Wire-bond package, fixed die size
		S	$1/S^2$	Thickness of the top metal remains constant	
Bakoglu [132]	On-chip IR noise	$S^3S_C^2$	$1/S^4S_C^2$	Ideal interconnect scaling, wire-bond package (the number of power connections is constant)	Current and capacitance scale as $S^2S_C^2$ (due to the scaling of the number of metal levels by S) as compared to SS_C^2
		$1/S$	1	Reverse interconnect scaling ($\propto S$), the number of power connections scale as SS_C^2 (flip-chip)	
	Package $L\frac{dI}{dt}$ noise	$S^3S_C^3$	$1/S^4S_C^3$	Number of power connections remains constant, inductance per connections increases as S_C	
		1	$1/S$	Number of power connections scale as SS_C^2 (flip-chip), inductance per connection scales as $1/S^2$	
Larsson [283]	Package $L\frac{dI}{dt}$ noise	SS_C	$1/S^2S_C$	Wire-bond package, number of package connections increases as SS_C	
Mezhiba and Friedman	On-chip IR noise	1	$1/S$	Metal thickness remains constant	Area array flip-chip package, pad pitch scales as $1/\sqrt{S}$ (*i.e.*, the number of power connections scales as SS_C^2)
		S	$1/S^2$	Ideal interconnect scaling	
	On-chip $L\frac{dI}{dt}$ noise	S	$1/S^2$	Metal thickness remains constant	
		1	$1/S$	Ideal interconnect scaling	

12.2.1 Global interconnect characteristics

The scaling of the cross-sectional dimensions of the on-chip global power lines directly affects the power distribution noise. Two scenarios of global interconnect scaling are considered here.

In the first scenario, the thickness of the top interconnect layers (where the conductors of the global power distribution networks are located) is assumed to remain constant. Through several recent technology generations, the thickness of the global interconnect layers has not been scaled in proportion to the minimum local line pitch due to power distribution noise and interconnect delay considerations. This behavior is in agreement with the 1997 edition of the International Technology Roadmap for Semiconductors (ITRS) [288], [289], where the minimum pitch and thickness of the global interconnect are assumed constant.

In the second scenario, the thickness and minimum pitch of the global interconnect layers are scaled in proportion to the minimum pitch of the local interconnect. This assumption is in agreement with the more recent editions of the ITRS [10], [287]. The scaling of the global interconnect in future technologies is therefore expected to evolve in the design envelope delimited by these two scenarios.

The number of metal layers and the fraction of the metal resources dedicated to the power distribution network are also assumed constant. The ratio of the diffusion barrier thickness to the copper interconnect core is assumed to remain constant with scaling. The increase in resistivity of the interconnect due to electron scattering at the interconnect surface interface (significant at line widths below 45 nm [10]) is neglected for relatively thick global power lines.

Under the aforementioned assumptions, in the constant metal thickness scenario, the effective sheet resistance of the global power distribution network remains constant with technology scaling. In the scenario of scaled metal thickness, the grid sheet resistance increases with technology scaling by a factor of S.

12.2.2 Scaling of the grid inductance

The inductive properties of power distribution grids are investigated in [69], [70]. It is shown that the inductance of the power grids with alternating power and ground lines behaves analogously to the grid resistance. That is, the grid inductance increases linearly with the grid length and decreases inversely linearly with the number of lines

in the grid. This linear behavior is due to the periodic structure of the alternating power and ground grid lines. The long range inductive coupling of a specific (signal or power) line to a power line is cancelled out by the coupling to the ground lines adjacent to the power line, which carry current in the opposite direction [68], [70]. As described in Chapter 9, inductive coupling in periodic grid structures is effectively a short range interaction. Similar to the grid resistance, the grid inductance can be conveniently expressed as a dimension independent grid sheet inductance $L_\square$ [70], [276]. The inductance of a specific grid is obtained by multiplying the sheet inductance by the grid length and dividing by the grid width. As described in Section 9.6.4, the grid sheet inductance can be estimated as

$$L_\square = 0.8P\left(\ln\frac{P}{T+W}+\frac{3}{2}\right)\frac{\mu H}{\square}, \tag{12.1}$$

where W, T, and P are the width, thickness, and pitch of the grid lines, respectively. The sheet inductance is proportional to the line pitch P. The line density is reciprocal to the line pitch. A smaller line pitch means a higher line density and more parallel paths for the current to flow. The sheet inductance is however relatively insensitive to the cross-sectional dimensions of the lines, as the inductance of the individual lines is similarly insensitive to these parameters. Note that while the sheet resistance of the power grid is determined by the metal conductivity and the net cross-sectional area of the lines, the sheet inductance of the grid is determined by the line pitch and the ratio of the pitch to the line width and thickness.

In the constant metal thickness scenario, the sheet inductance of the power grid remains constant since the routing characteristics of the global power grid do not change. In the scaled thickness scenario, the line pitch, width, and thickness are reduced by S, increasing the line density and the number of parallel current paths. The sheet inductance therefore decreases by a factor of S, according to (12.1).

12.2.3 Flip-chip packaging characteristics

In a flip-chip package, the integrated circuit and package are interconnected via an area array of solder bumps mounted onto the on-chip I/O pads [103]. The power supply current enters the on-chip power distribution network from the power/ground pads. A view of the on-chip area array of power/ground pads is shown in Fig. 12.1.

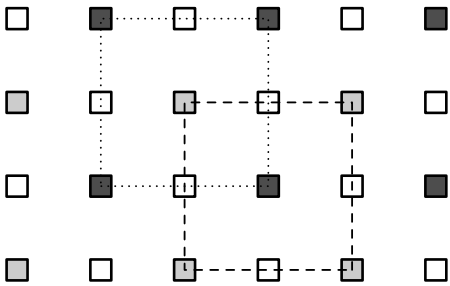

Fig. 12.1. An area array of on-chip power/ground I/O pads. The power pads are colored dark gray, the ground pads are colored light gray, and the signal pads are white. The current distribution area of the power pad (*i.e.*, the power distribution cell) in the center of the figure is delineated by the dashed line. The current distribution area of the ground pad in the center of the figure is delineated by the dotted line.

One of the main goals of this work is to estimate the significance of the *on-chip* inductive voltage drop in comparison to the on-chip resistive voltage drop. Therefore, all of the power/ground pads of a flip-chip packaged IC are assumed to be equipotential, *i.e.*, the variation in the voltage levels among the pads is considered negligible as compared to the noise within the on-chip power distribution network. For the purpose of this scaling analysis, a uniform power consumption per die area is assumed. Under these assumptions, each power (ground) pad supplies power (ground) current only to those circuits located in the area around the pad, as shown in Fig. 12.1. This area is referred to as a power distribution cell (or power cell). The edge dimensions of each power distribution cell are proportional to the pitch of the power/-ground pads. The size of the power cell area determines the effective distance of the on-chip distribution of the power current. The power distribution scaling analysis becomes independent of die size.

An important element of this analysis is the scaling of the flip-chip technology. The rate of decrease in the pad pitch and the rate of reduction in the local interconnect half-pitch are compared in Fig. 12.2, based on the ITRS [10]. At the 150 nm line half-pitch technology node, the pad pitch P is 160 μm. At the 32 nm node, the pad pitch is forecasted to be 80 μm. That is, the linear density of the pads doubles for a fourfold reduction in circuit feature size. The pad size and pitch P scale, therefore, as $1/\sqrt{S}$ and the area density ($\propto 1/P^2$) of the pads increases as S with each technology generation. Interestingly, one of

the reasons given for this relatively infrequent change in the pad pitch (as compared with the introduction of new CMOS technology generations) is the cost of the test probe head [10]. The maximum density of the flip-chip pads is assumed to be limited by the pad pitch. Although the number of on-chip pads is forecasted to remain constant, some recent research has predicted that the number of on-chip power/-ground pads will increase due to electromigration and resistive noise considerations [220], [290].

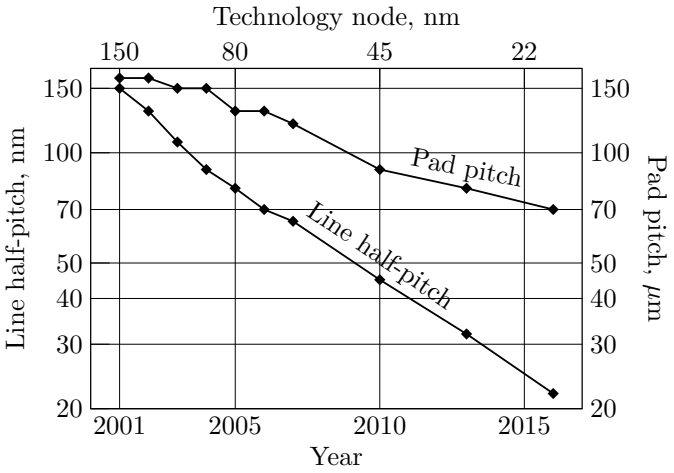

Fig. 12.2. Decrease in flip-chip pad pitch with technology generations as compared to the local interconnect half-pitch.

12.2.4 Impact of on-chip capacitance

On-chip capacitors are used to reduce the impedance of the power distribution grid lines as seen from the load terminals. A simple model of an on-chip power distribution grid with a power load and a decoupling capacitor is shown in Fig. 12.3. The on-chip loads are switched within tens of picoseconds in modern semiconductor technologies. The frequency spectrum of the load current therefore extends well beyond 10 GHz. The on-chip decoupling capacitors shunt the load current at the highest frequencies. The bulk of the power current bypasses the on-chip distribution network at these frequencies. At the lower frequencies, however, the capacitor impedance is relatively high and the bulk of the

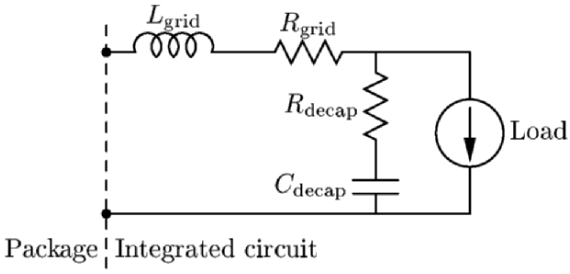

Fig. 12.3. A simplified circuit model of the on-chip power distribution network with a power load and a decoupling capacitance.

current flows through the on-chip power distribution network. The decoupling capacitors therefore serve as a low pass filter for the power current.

Describing the same effect in the time domain, the capacitors supply the (high frequency) current to the load during a switching transient. To prevent excessive power noise, the charge on the decoupling capacitor should be replenished by the (lower frequency) current flowing through the power distribution network before the next switching of the load, *i.e.*, typically within a clock period. The effect of the on-chip decoupling capacitors is therefore included in the model by assuming that the current transients within the on-chip power distribution network are characterized by the clock frequency of the circuit, rather than by the switching times of the on-chip load circuits. Estimates of the resistive voltage drop are based on the average power current, which is not affected by the on-chip decoupling capacitors.

12.3 Model of power supply noise

The following simple model is utilized in the scaling analysis of on-chip power distribution noise. A power distribution cell is modeled as a circle of radius r_c with a constant current consumption per area I_a, as described by Arledge and Lynch [220]. The model is depicted in Fig. 12.4. The total current of the cell is $I_{\text{cell}} = I_a \cdot \pi r_c^2$. The power network current is distributed from a circular pad of radius r_p at the center of the cell. The global power distribution network has an effective sheet resistance $\rho_\square$. The incremental voltage drop dV_R across the elemental circular resistance $\rho_\square \, dr/2\pi r$ is due to the current $I_a(\pi r_c^2 - \pi r^2)$ flowing

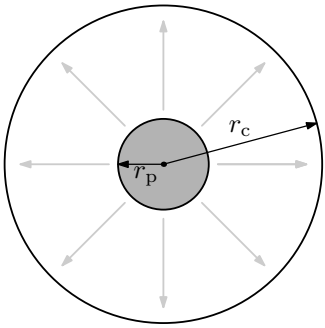

Fig. 12.4. A model of the power distribution cell. Power supply current spreads out from the power pad in the center of the cell to the cell periphery, as shown by the arrows.

through this resistance toward the periphery of the cell. The voltage drop at the periphery of the power distribution cell is

$$\begin{aligned}\Delta V_R &= \int_{r_p}^{r_c} dV_R = \int_{r_p}^{r_c} I(r) \cdot dR(r) \\ &= \int_{r_p}^{r_c} \pi(r_c^2 - r^2) I_a \cdot \rho_\square \frac{dr}{2\pi r} \\ &= I_a \pi r_c^2 \rho_\square \cdot \frac{1}{2\pi} \left(\ln \frac{r_c}{r_p} + \frac{r_p^2}{2r_c^2} - \frac{1}{2} \right) \\ &= I_{\text{cell}} \rho_\square \cdot C \left(\frac{r_c}{r_p} \right). \end{aligned} \tag{12.2}$$

The resistive voltage drop is proportional to the product of the total cell current I_{cell} and the effective sheet resistance $\rho_\square$ with the coefficient C dependent only on the r_c/r_p ratio. The ratio of the pad pitch to the pad size is assumed to remain constant. The coefficient C, therefore, does not change with technology scaling.

The properties of the grid inductance are analogous to the properties of the grid resistance, as discussed in Section 12.2.1. Therefore, analogous to the resistive voltage drop ΔV_R discussed above, the inductive voltage drop ΔV_L is proportional to the product of the sheet inductance $L_\square$ of the global power grid and the magnitude of the cell transient current dI_{cell}/dt,

$$\Delta V_L = L_\square \frac{dI_{\text{cell}}}{dt} \cdot C\left(\frac{r_c}{r_p}\right). \tag{12.3}$$

12.4 Power supply noise scaling

An analysis of the on-chip power supply noise is presented in this section. The interconnect characteristics assumed in the analysis are described in Section 12.2. The power supply noise model is described in Section 12.3. Ideal scaling of the power distribution noise in the constant thickness scenario is discussed in Section 12.4.1. Ideal scaling of the noise in the scaled thickness scenario is analyzed in Section 12.4.2. Scaling of the power distribution noise based on the ITRS projections is discussed in Section 12.4.3.

12.4.1 Analysis of constant metal thickness scenario

The scaling of a power distribution grid over four technology generations according to the constant metal thickness scenario is depicted in Fig. 12.5. The minimum feature size is reduced by $\sqrt{2}$ with each generation. The minimum feature size over four generations is therefore reduced by four, *i.e.*, $\left(\sqrt{2}\right)^4 = 4$, while the size of the power distribution cell (represented by the size of the square grid) is halved ($\sqrt{4} = 2$). As the cross-sectional dimensions of the power lines are maintained constant in this scenario, both the sheet resistance $\rho_\square$ and sheet inductance $L_\square$ of the power distribution grid remain constant with scaling under these conditions.

The cell current I_{cell} is the product of the area current density I_a and the cell area πr_c^2. The current per area I_a scales as S; the area of the cell is proportional to P^2 which scales as $1/S$. The cell current I_{cell}, therefore, remains constant (*i.e.*, scales as 1). The resistive drop ΔV_R, therefore, scales as $I_{\text{cell}} \cdot \rho_\square \propto 1 \cdot 1 \propto 1$. The resistive SNR_R^{I} of the power supply voltage, consequently, decreases with scaling as

$$\text{SNR}_R^{\text{I}} = \frac{V_{\text{dd}}}{\Delta V_R} \propto \frac{1/S}{1} \propto \frac{1}{S}. \tag{12.4}$$

This scaling trend agrees with the trend described by Bakoglu in the improved scaling regime [132]. Faster scaling of the on-chip current as described by Bakoglu is offset by increasing the interconnect thickness by S which reduces the sheet resistance $\rho_\square$ by S. This trend is more

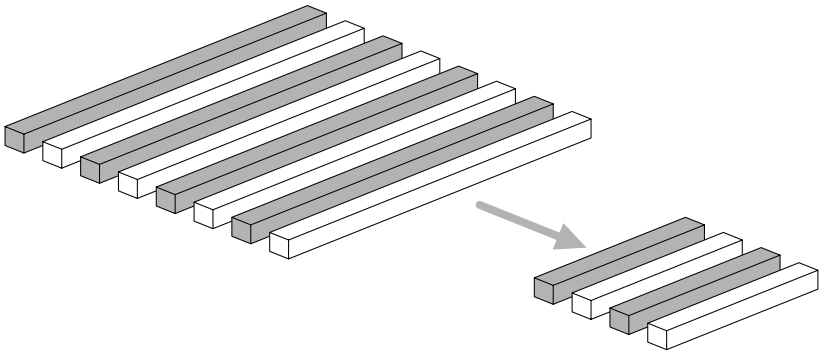

Fig. 12.5. The scaling of a power distribution grid over four technology generations according to the constant metal thickness scenario. The cross-sectional dimensions of the power lines remain constant. The size of the power distribution cell, represented by the size of the square grid, is halved.

favorable as compared to the $1/S^2$ dependence established by Song and Glasser [282]. The improvement is due to the decrease in the power cell area of a flip-chip IC by a factor of S whereas a wire-bonded die of constant area is assumed in [282].

The transient current dI_{cell}/dt scales as $I_{\text{cell}}/\tau \propto 1/(1/S) \propto S$, where $\tau \propto 1/S$ is the transistor switching time. The inductive voltage drop ΔV_L, therefore, scales as $L_{\square} \cdot dI_{\text{cell}}/dt \propto 1 \cdot S$. The inductive SNR_L^{I} of the power supply voltage decreases with scaling as

$$\text{SNR}_L^{\text{I}} = \frac{V_{\text{dd}}}{\Delta V_L} \propto \frac{1/S}{S} \propto \frac{1}{S^2}. \tag{12.5}$$

The relative magnitude of the inductive noise therefore increases by a factor of S faster as compared to the resistive noise. Estimates of the inductive and resistive noise described by Bakoglu also differ by a factor of S [132].

12.4.2 Analysis of the scaled metal thickness scenario

The scaling of a power distribution grid over four technology generations according to the scaled metal thickness scenario is depicted in Fig. 12.6. In this scenario, the cross-sectional dimensions of the power lines are reduced in proportion to the minimum feature size by a factor of four, while the size of the power distribution cell is halved. Under these conditions, the sheet resistance $\rho_{\square}$ of the power distribution grid

increases by S, while the sheet inductance $L_\square$ of the power distribution grid decreases by S with technology scaling.

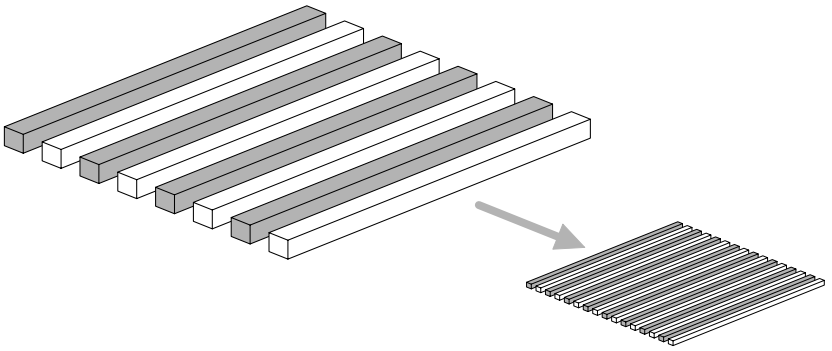

Fig. 12.6. The scaling of a power distribution grid over four technology generations according to the scaled metal thickness scenario. The cross-sectional dimensions of the power lines are reduced in proportion to the minimum feature size by a factor of four. The size of the power distribution cell, represented by the size of the square grid, is halved.

Analogous to the constant metal thickness scenario, the cell current I_{cell} remains constant. The resistive drop ΔV_R, therefore, scales as $I_{\text{cell}} \cdot \rho_\square \propto 1 \cdot S \propto S$. The resistive SNR_R^{II} of the power supply voltage, consequently, decreases with scaling as

$$\text{SNR}_R^{\text{II}} = \frac{V_{\text{dd}}}{\Delta V_R} \propto \frac{1/S}{S} \propto \frac{1}{S^2}. \tag{12.6}$$

As discussed in the previous section, the transient current dI_{cell}/dt scales as $I_{\text{cell}}/\tau \propto S$. The inductive voltage drop ΔV_L, therefore, scales as $L_\square \cdot dI_{\text{cell}}/dt \propto 1/S \cdot S \propto 1$. The inductive SNR_L^{II} of the power supply voltage decreases with scaling as

$$\text{SNR}_L^{\text{II}} = \frac{V_{\text{dd}}}{\Delta V_L} \propto \frac{1/S}{1} \propto \frac{1}{S}. \tag{12.7}$$

The rise of the inductive noise is mitigated if ideal interconnect scaling is assumed and the thickness, width, and pitch of the global power lines are scaled as $1/S$. In this scenario, the density of the global power lines increases as S and the sheet inductance $L_\square$ of the global power distribution grids decreases as $1/S$, mitigating the inductive noise and

SNR_L by S. The sheet resistance of the power distribution grid, however, increases as S, exacerbating the resistive noise and SNR_R by a factor of S. Currently, the parasitic resistive impedance dominates the total impedance of on-chip power distribution networks. Ideal scaling of the upper interconnect levels will therefore increase the overall power distribution noise. However, as CMOS technology approaches the nanometer range and the inductive and resistive noise becomes comparable, judicious tradeoffs between the resistance and inductance of the power networks will be necessary to achieve the minimum noise level (see Chapter 11) [206], [207].

12.4.3 ITRS scaling of power noise

Although the ideal scaling analysis allows the comparison of the rates of change of both resistive and inductive voltage drops, it is difficult to estimate the *ratio* of these quantities for direct assessment of their relative significance. Furthermore, practical scaling does not accurately follow the concept of ideal scaling due to material and technological limitations. An estimate of the ratio of the inductive to resistive voltage drop is therefore conducted in this section based on the projected 2001 ITRS data [10].

Forecasted demands in the supply current of high performance microprocessors are shown in Fig. 12.7. Both the average current and the transient current are rising exponentially with technology scaling. The rate of increase in the transient current is more than double the rate of increase in the average current as indicated by the slope of the trend lines depicted in Fig. 12.7. This behavior is in agreement with ideal scaling trends. The faster rate of increase in the transient current as compared to the average current is due to rising clock frequencies. The transient current of modern high performance processors is approximately one teraampere per second (10^{12} A/s) and is expected to rise, reaching hundreds of teraamperes per second. Such a high magnitude of the transient current is caused by switching hundreds of amperes within a fraction of a nanosecond.

In order to translate the projected current requirements into supply noise voltage trends, a case study interconnect structure is considered. The square grid structure shown in Fig. 12.8 is used here to serve as a model of the on-chip power distribution grid. The square grid consists of interdigitated power and ground lines with a $1\,\mu\mathrm{m} \times 1\,\mu\mathrm{m}$ cross section and a $1\,\mu$m line spacing. The length and width of the grid are equal to

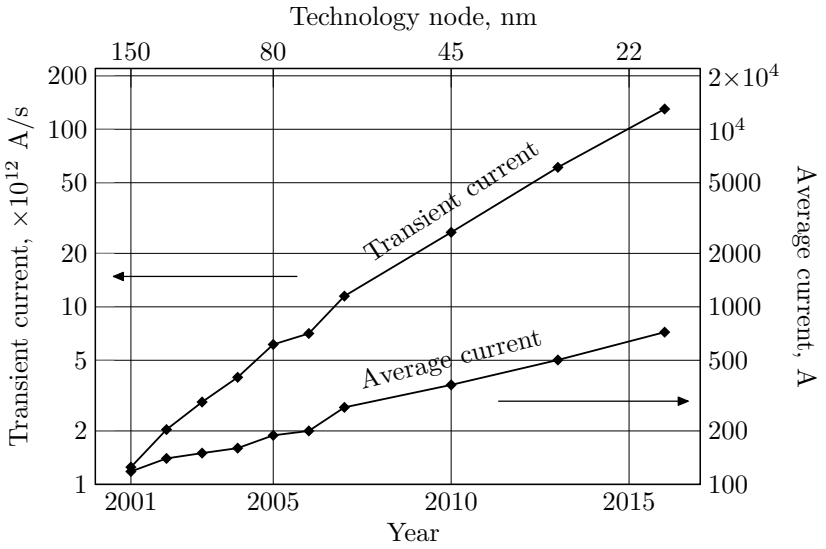

Fig. 12.7. Increase in power current demands of high performance microprocessors with technology scaling, according to the ITRS. The average current is the ratio of the circuit power to the supply voltage. The transient current is the product of the average current and the on-chip clock rate, $2\pi f_{\text{clk}}$.

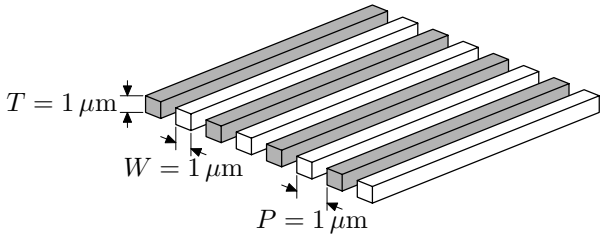

Fig. 12.8. Power distribution grid used to estimate trends in the power supply noise.

the size of a power distribution cell. The grid sheet inductance is 1.8 picohenrys per square, and the grid sheet resistance is 0.16 ohms per square. The size of the power cell is assumed to be twice the pitch of the flip-chip pads, reflecting that only half of the total number of pads are used for the power and ground distribution as forecasted by the ITRS for high performance ASICs.

The electrical properties of this structure are similar to the properties of the global power distribution grid covering a power distribution cell with the same routing characteristics. Note that the resistance and

inductance of the *square* grid are independent of grid dimensions [276] (as long as the dimensions are severalfold greater than the line pitch). The average and transient currents flowing through the grid are, however, scaled from the IC current requirements shown in Fig. 12.7 in proportion to the area of the grid. The current flowing through the square grid is therefore the same as the current distributed through the power grid within the power cell. The power current enters and leaves from the same side of the grid, assuming the power load is connected at the opposite side. The voltage differential across this structure caused by the average and transient currents produces, respectively, on-chip resistive and inductive noise. The square grid has the same inductance to resistance ratio as the global distribution grid with the same line pitch, thickness, and width. Hence, the square grid has the same inductive to resistive noise ratio. The square grid model also produces the same rate of increase in the noise because the current is scaled proportionately to the area of the power cell.

The resulting noise trends under the constant metal thickness scenario are illustrated in Fig. 12.9. As discussed in Section 12.2, the area of the grid scales as $1/S$. The current area density increases as S. The total average current of the grid, therefore, remains constant. The resistive noise also remains approximately constant, as shown in Fig. 12.9. The inductive noise, alternatively, rises steadily and becomes comparable to the resistive noise at approximately the 45 nm technology node. These trends are in reasonable agreement with the ideal scaling predictions discussed in Section 12.4.1.

The inductive and resistive voltage drops in the scaled metal thickness scenario are shown in Fig. 12.10. The increase in inductive noise with technology scaling is limited, while the resistive noise increases by an order of magnitude. This behavior is similar to the ideal scaling trends for this scenario, as discussed in Section 12.4.2.

Note that the structure depicted in Fig. 12.8 has a lower inductance to resistance ratio as compared to typical power distribution grids because the power and ground lines are relatively narrow and placed adjacent to each other, reducing the area of the current loop and increasing the grid resistance [69], [276]. The width of a typical global power line varies from tens to a few hundreds of micrometers, resulting in a significantly higher inductance to resistance ratio. The results shown in Figs. 12.9 and 12.10 can be readily extrapolated to different

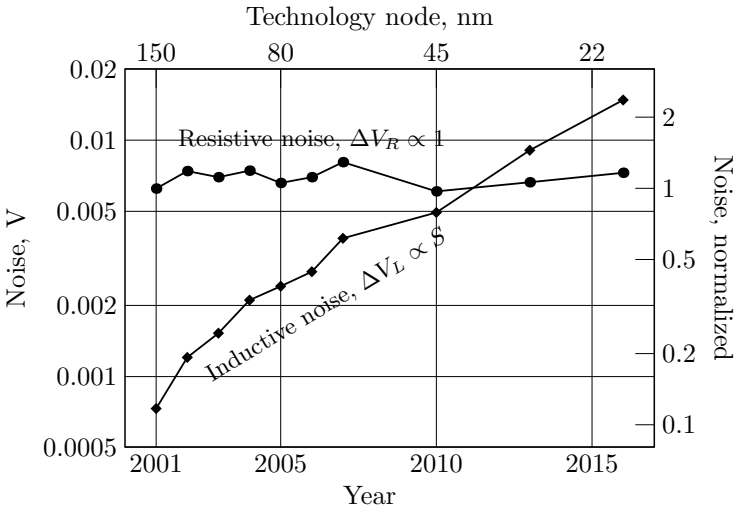

Fig. 12.9. Scaling trends of resistive and inductive power supply noise under the constant metal thickness scenario.

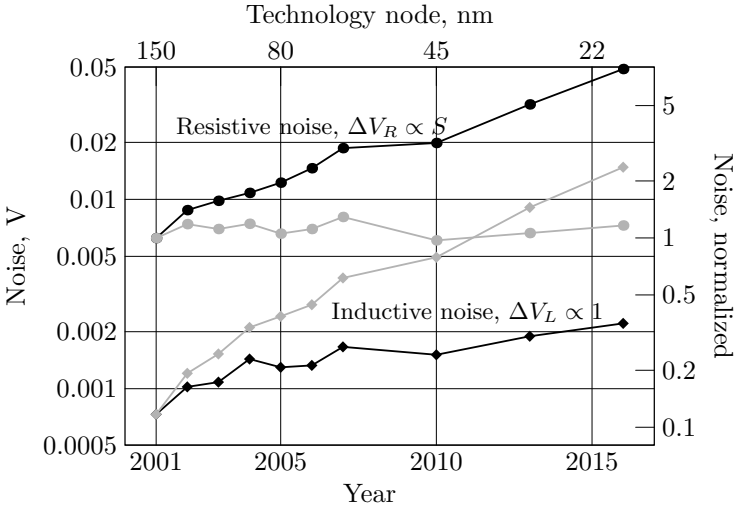

Fig. 12.10. Scaling trends of resistive and inductive power supply noise under the scaled metal thickness interconnect scaling scenario. The trends of the constant metal thickness scenario are also displayed in light gray for comparison.

grid configurations, using the expression for the grid sheet inductance (12.1).

Several factors offset the underestimation of the relative magnitude of the inductive noise due to the relatively low inductance to resistance ratio of the model shown in Fig. 12.8. If the global power distribution grid is composed of several layers of interconnect, the lines in the lower interconnect levels have a smaller pitch and thickness, significantly reducing the inductance to resistance ratio at high frequencies [291]. The transient current is conservatively approximated as the product of the average current I_{avg} and the angular clock frequency $2\pi f_{\text{clk}}$. This estimate, while serving as a useful scaling parameter, tends to overestimate the absolute magnitude of the current transients, increasing the ratio of the inductive and resistive voltage drops.

12.5 Implications of noise scaling

As described in the previous section, the amplitude of both the resistive and inductive noise relative to the power supply voltage increases with technology scaling. A number of techniques have been proposed to mitigate the unfavorable scaling of power distribution noise. These techniques are briefly summarized below.

To maintain a constant supply voltage to resistive noise ratio, the effective sheet resistance of the global power distribution grid should be reduced. There are two ways to allocate additional metal resources to the power distribution grid. One option is to increase the number of metalization layers. This approach adversely affects fabrication time and yield and, therefore, increases the cost of manufacturing. The ITRS forecasts only a moderate increase in the number of interconnect levels, from eight levels at the 130 nm line half-pitch node to eleven levels at the 32 nm node [10]. The second option is to increase the fraction of metal area per metal level allocated to the power grid. This strategy decreases the amount of wiring resources available for global signal routing and therefore can also necessitate an increase in the number of interconnect layers.

The sheet inductance of the power distribution grid, similar to the sheet resistance, can be lowered by increasing the number of interconnect levels. Furthermore, wide metal trunks typically used for power distribution at the top levels can be replaced with narrow interdigitated power/ground lines. Although this configuration substantially lowers the grid inductance, it increases the grid resistance and, consequently, the resistive noise [276].

Alternatively, circuit techniques can be employed to limit the peak transient power current demands of the digital logic. Current steering logic, for example, produces a minimal variation in the current demand between the transient response and the steady state response. In synchronous circuits, the maximum transient currents typically occur during the beginning of a clock period. Immediately after the arrival of a clock signal at the latches, a signal begins to propagate through the blocks of sequential logic. Clock skew scheduling can be exploited to spread in time the periods of peak current demand [36].

The constant metal thickness scaling scenario achieves a lower overall power noise until the technology generation is reached where the inductive and resistive voltage drops become comparable. Beyond this node, a careful tradeoff between the resistance and inductance of the power grid is necessary to minimize the on-chip power supply noise. The increase in the significance of the inductance of the power distribution interconnect is similar to that noted in signal interconnect [61], [292]. The trend, however, is delayed by several technology generations as compared to signal interconnect. As discussed in Section 12.2, the high frequency harmonics are filtered out by the on-chip decoupling capacitance and the power grid current has a comparatively lower frequency content as compared to the signal lines.

12.6 Summary

A scaling analysis of power distribution noise in flip-chip packaged integrated circuits is presented in this chapter. Published scaling analyses of power distribution noise are reviewed and various assumptions of these analyses are discussed. The primary conclusions can be summarized as follows.

- Under the constant metal thickness scenario, the relative magnitude (*i.e.*, the reciprocal of the signal-to-noise ratio) of the resistive noise increases by the scaling factor S, while the relative magnitude of the inductive noise increases by S^2
- Under the scaled metal thickness scenario, the scaling trend of the inductive noise improves by a factor of S, but the relative magnitude of the resistive noise increases by S^2
- The importance of on-chip inductive noise increases with technology scaling

- Careful tradeoffs between the resistance and inductance of power distribution networks in nanometer CMOS technologies will be necessary to achieve minimum power supply noise levels