Chapter 13
Efficient Placement of Distributed On-Chip Decoupling Capacitors

Decoupling capacitors are widely used to manage power supply noise [281] and are an effective way to reduce the impedance of power delivery systems operating at high frequencies [28, 29]. A decoupling capacitor acts as a local reservoir of charge, which is released when the power supply voltage at a particular current load drops below some tolerable level. Since the inductance scales slowly [129], the location of the decoupling capacitors significantly affects the design of the power/ground networks in high performance integrated circuits such as microprocessors. At higher frequencies, a distributed system of decoupling capacitors are placed on-chip to effectively manage the power supply noise [279].

The efficacy of decoupling capacitors depends upon the impedance of the conductors connecting the capacitors to the current loads and power sources. As described in [277], a maximum parasitic impedance between the decoupling capacitor and the current load (*or* power source) exists at which the decoupling capacitor is effective. Alternatively, to be effective, an on-chip decoupling capacitor should be placed such that both the power supply and the current load are located inside the appropriate effective radius [277]. The efficient placement of on-chip decoupling capacitors in nanoscale ICs is the subject of this chapter. Unlike the methodology for placing a single lumped on-chip decoupling capacitor presented in Chap. 12, a system of *distributed* on-chip decoupling capacitors is described in this chapter. A design methodology to estimate the parameters of the distributed system of on-chip decoupling capacitors is also presented, permitting the required on-chip decoupling capacitance to be allocated under existing technology constraints.

This chapter is organized as follows. Technology limitations in nanoscale integrated circuits are reviewed in Sect. 13.1. The problem of placing on-chip decoupling capacitors in nanoscale ICs while satisfying technology constraints is formulated in Sect. 13.2. The design of a distributed on-chip decoupling capacitor network is presented in Sect. 13.3. Various design tradeoffs are discussed in Sect. 13.4. A design methodology for placing distributed on-chip decoupling

I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_13

capacitors is presented in Sect. 13.5. Related simulation results for typical values of on-chip parasitic resistances are discussed in Sect. 13.6. Some specific conclusions are summarized in Sect. 13.7.

13.1 Technology Constraints

On-chip decoupling capacitors have traditionally been designed as standard gate oxide CMOS capacitors [288]. As technology scales, leakage current through the gate oxide of an on-chip decoupling capacitor has greatly increased [289–291]. Moreover, in modern high performance ICs, a large portion (up to 40 %) of the circuit area is occupied by the on-chip decoupling capacitance [292, 293]. Conventional gate oxide on-chip decoupling capacitors are therefore prohibitively expensive from an area and yield perspective, as well as greatly increasing the overall power dissipated on-chip [294].

To reduce the power consumed by an IC, MIM capacitors are frequently utilized as decoupling capacitors. The capacitance density of a MIM capacitor in a 90 nm CMOS technology is comparable to the maximum capacitance density of a CMOS capacitor and is typically 10–30 $\mathrm{fF}/\mu\mathrm{m}^2$ [256, 259, 270]. A maximum magnitude of an on-chip decoupling capacitor therefore exists for a specific distance between a current load and a decoupling capacitor (as constrained by the available on-chip metal resources). Alternatively, a minimum achievable impedance per unit length exists for a specified capacitance density of an on-chip decoupling capacitor placed at a specific distance from a circuit module, as illustrated in Fig. 13.1.

Observe from Fig. 13.1 that the available metal area for the second level of a distributed on-chip capacitance is greater than the fraction of metal resources dedicated to the first level of a distributed on-chip capacitance. Capacitor C_2 can therefore be larger than C_1. Note also that a larger capacitor can only be placed farther from the current load. Similarly, the metal resources required by the first level of interconnection (connecting C_1 to the current load) is smaller than the metal resources dedicated to the second level of interconnections. The impedance Z_2 is therefore smaller than Z_1.

13.2 Placing On-Chip Decoupling Capacitors in Nanoscale ICs

Decoupling capacitors have traditionally been allocated into the white space (those areas not occupied by the circuit elements) available on the die based on an unsystematic or ad hoc approach [283, 284], as shown in Fig. 13.2. In this way, decoupling capacitors are often placed at a significant distance from the current load. Conventional approaches for placing on-chip decoupling capacitors result

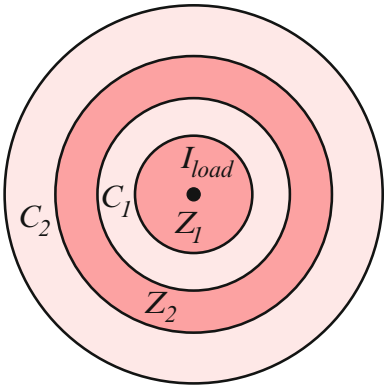

Fig. 13.1 Fundamental limits of on-chip interconnections. Two levels of a distributed on-chip decoupling capacitance are allocated around a current load. The interconnect impedance is inversely proportional to the fraction of metal area dedicated to the interconnect level, decreasing as the decoupling capacitor is farther from the current source ($Z_1 > Z_2$). The decoupling capacitance increases as the capacitor is farther from the current load due to the increased area ($C_1 < C_2$). The two levels of interconnection and distributed decoupling capacitance are shown, respectively, in *dark pink* and *light pink*

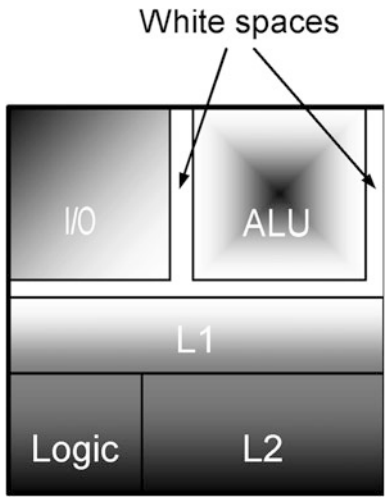

Fig. 13.2 Placement of on-chip decoupling capacitors using a conventional approach. Decoupling capacitors are allocated into the *white space* (those areas not occupied by the circuits elements) available on the die using an unsystematic or ad hoc approach. As a result, the power supply voltage drops below the minimum tolerable level for remote blocks (shown in *dark gray*). Low noise regions are *light gray*

in oversized capacitors. The conventional allocation strategy, therefore, results in increased power noise, compromising the signal integrity of an entire system, as illustrated in Fig. 13.3. This issue of power delivery cannot be alleviated by simply increasing the size of the on-chip decoupling capacitors. Furthermore, increasing the size of more distant on-chip decoupling capacitors results in wasted area, increased power, reduced reliability, and higher cost. A design methodology is therefore required to account for technology trends in nanoscale ICs, such as increasing frequencies, larger die sizes, higher current demands, and reduced noise margins.

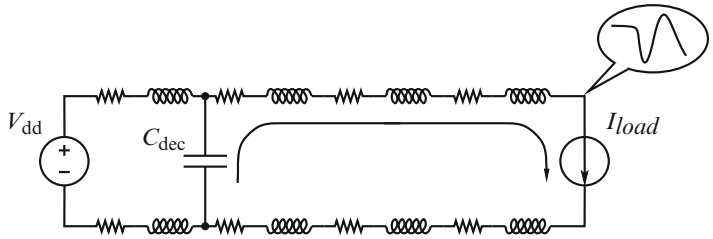

Fig. 13.3 A conventional on-chip decoupling capacitor. Typically, a large decoupling capacitor is placed farther from the current load due to physical limitations. Current flowing through the long power/ground lines results in large voltage fluctuations across the terminals of the current load

To be effective, a decoupling capacitor should be placed physically close to the current load. This requirement is naturally satisfied in board and package applications, since large capacitors are much smaller than the dimensions of the circuit board (or package) [223]. In this case, a lumped model of a decoupling capacitor provides sufficient accuracy [295].

The size of an on-chip decoupling capacitor, however, is directly proportional to the area occupied by the capacitor and can require a significant portion of the on-chip area. The minimum impedance between an on-chip capacitor and the current load is fundamentally affected by the magnitude (and therefore the area) of the capacitor. Systematically partitioning the decoupling capacitor into smaller capacitors solves this issue. A system of distributed on-chip decoupling capacitors is illustrated in Fig. 13.4.

In a system of distributed on-chip decoupling capacitors, each decoupling capacitor is sized based on the impedance of the interconnect segment connecting the capacitor to the current load. A particular capacitor only provides charge to a current load during a short period. The rationale behind this scheme can be explained as follows. The capacitor closest to the current load is engaged immediately after the switching cycle is initiated. Once the first capacitor is depleted of charge, the next capacitor is activated, providing a large portion of the total current drawn by the load. This procedure is repeated until the last capacitor becomes active. Similar to the hierarchical placement of decoupling capacitors presented in [28, 136], this technique provides an efficient solution for providing the required on-chip decoupling capacitance based on specified capacitance density constraints. A system of distributed on-chip decoupling capacitors should therefore be utilized to provide a low impedance, cost effective power delivery network in nanoscale ICs.

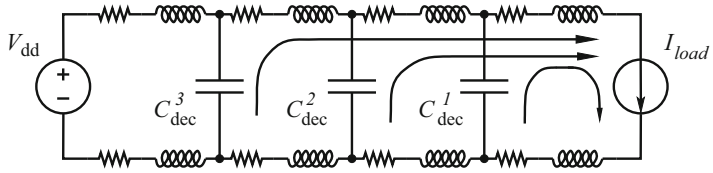

Fig. 13.4 A network of distributed on-chip decoupling capacitors. The magnitude of the decoupling capacitors is based on the impedance of the interconnect segment connecting a specific capacitor to a current load. Each decoupling capacitor is designed to only provide charge during a specific time interval

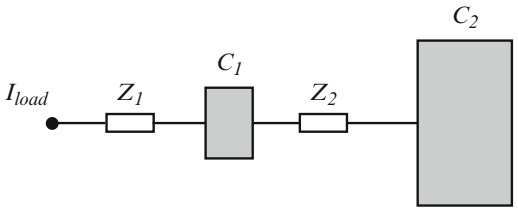

Fig. 13.5 A physical model of a system of distributed on-chip decoupling capacitors. Two capacitors are assumed to provide the required charge drawn by the load. Z_1 and Z_2 denote the impedance of the metal lines connecting, respectively, C_1 to the current load and C_2 to C_1

13.3 Design of a Distributed On-Chip Decoupling Capacitor Network

As described in Sect. 13.2, a system of distributed on-chip decoupling capacitors is an efficient solution for providing the required on-chip decoupling capacitance based on the maximum capacitance density available in a particular technology. A physical model of the technique is illustrated in Fig. 13.5. For simplicity, two decoupling capacitors are assumed to provide the required charge drawn by the current load. Note that as the capacitor is placed farther from the current load, the magnitude of an on-chip decoupling capacitor increases due to relaxed constraints. In the general case, the described methodology can be extended to any practical number of on-chip decoupling capacitors. Note that Z_1 is typically limited by a specific technology (determined by the impedance of a single metal wire) and the magnitude of C_1 (the area available in the vicinity of a circuit block).

A circuit model of a system of distributed on-chip decoupling capacitors is shown in Fig. 13.6. The impedance of the metal lines connecting the capacitors to the current load is modeled as resistors R_1 and R_2. A triangular current source is assumed to model the current load. The magnitude of the current source increases linearly, reaching the maximum current $I_{\max}$ at rise time t_r, i.e., $I_{load}(t) = I_{\max}\frac{t}{t_r}$. The maximum tolerable ripple at the load is 10 % of the power supply voltage.

Note from Fig. 13.6 that since the charge drawn by the current load is provided by the on-chip decoupling capacitors, the voltage across the capacitors during discharge

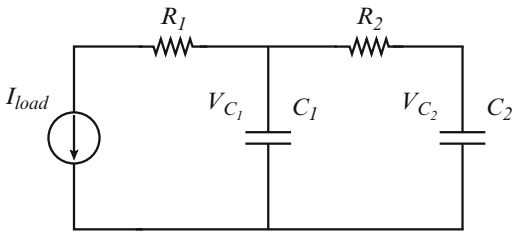

Fig. 13.6 A circuit model of an on-chip distributed decoupling capacitor network. The impedance of the metal lines is modeled, respectively, as R_1 and R_2

drops below the initial power supply voltage. The required charge during the entire switching event is thus determined by the voltage drop across C_1 and C_2.

The voltage across the decoupling capacitors at the end of the switching cycle ($t = t_r$) can be determined from Kirchhoff's laws [287]. Writing KVL and KCL equations for each of the loops (see Fig. 13.6), the system of differential equations describing the voltage across C_1 and C_2 at t_r is

$$\frac{dV_{C_1}}{dt} = \frac{V_{C_2} - V_{C_1}}{R_2 C_1} - \frac{I_{load}}{C_1}, \tag{13.1}$$

$$\frac{dV_{C_2}}{dt} = \frac{V_{C_1} - V_{C_2}}{R_2 C_2}. \tag{13.2}$$

Simultaneously solving (13.1) and (13.2) and applying the initial conditions, the voltage across C_1 and C_2 at the end of the switching activity is

$$\begin{aligned} V_{C_1}|_{t=t_r} = \frac{1}{2(C_1+C_2)^3 t_r}\Bigg[& 2C_1^3 t_r + C_1^2 t_r (6C_2 - I_{\max} t_r) \\ & - C_2^2 t_r \left(2C_2 \left(I_{\max} R_2 - 1\right) + I_{\max} t_r\right) \\ & + 2C_1 C_2 \left(C_2^2 \left(1 - e^{-\frac{(C_1+C_2)t_r}{C_1 C_2 R_2}}\right) I_{\max} R_2^2 \right. \\ & \left. + C_2 \left(3 - I_{\max} R_2\right) t_r - I_{\max} t_r^2\right)\Bigg], \end{aligned} \tag{13.3}$$

$$\begin{aligned} V_{C_2}|_{t=t_r} = \frac{1}{2(C_1+C_2)^3 t_r}\Bigg[& 2C_1^3 t_r + C_2^2 t_r \left(2C_2 - I_{\max} t_r\right) \\ & + 2C_1 C_2 t_r \left(C_2 \left(3 + I_{\max} R_2\right) - I_{\max} t_r\right) \\ & + C_1^2 \left(2C_2^2 \left(e^{-\frac{(C_1+C_2)t_r}{C_1 C_2 R_2}} - 1\right) I_{\max} R_2^2 \right. \\ & \left. + 2C_2 \left(3 + I_{\max} R_2\right) t_r - I_{\max} t_r^2\right)\Bigg], \end{aligned} \tag{13.4}$$

where $I_{\max}$ is the maximum magnitude of the current load and t_r is the rise time.

Note that the voltage across C_1 and C_2 after discharge is determined by the magnitude of the decoupling capacitors and the parasitic resistance of the metal line(s) between the capacitors. The voltage across C_1 after the switching cycle, however, depends upon the resistance of the P/G paths connecting C_1 to a current load and is

$$V_{C_1} = V_{load} + I_{\max} R_1, \tag{13.5}$$

where V_{load} is the voltage across the terminals of a current load. Assuming $V_{load} \geq 0.9V_{\rm dd}$ and $V_{C_1}^{\max} = V_{\rm dd}$ (meaning that C_1 is infinitely large), the upper bound for R_1 is

$$R_1^{\max} = \frac{V_{\rm dd}(1-\alpha)}{I_{\max}}, \tag{13.6}$$

where α is the ratio of the minimum tolerable voltage across the terminals of a current load to the power supply voltage ($\alpha = 0.9$ in this chapter). If $R_1 > R_1^{\max}$, no solution exists for providing sufficient charge drawn by the load. In this case, the circuit block should be partitioned, reducing the current demands ($I_{\max}$).

Note that expressions for determining the voltage across the decoupling capacitors are transcendental functions. No closed-form solution, therefore, exists. From (13.3) and (13.4), the design space can be graphically obtained for determining the maximum tolerable resistance R_2 and the minimum magnitude of the capacitors, maintaining the voltage across the load equal to or greater than the minimum allowable level. The voltage across C_1 after discharge as a function of C_1 and R_2 is depicted in Fig. 13.7.

Observe from Fig. 13.7 that the voltage across capacitor C_1 increases exponentially with capacitance, saturating for large C_1. The voltage across C_1, however, is almost independent of R_2, decreasing slightly with R_2 (see Fig. 13.7a). This behavior can be explained as follows. As a current load draws charge from the decoupling capacitors, the voltage across the capacitors drops below the initial level. The charge released by a capacitor is proportional to the capacitance and the change in voltage. A larger capacitance therefore results in a smaller voltage drop. From Fig. 13.6, note that as resistance R_2 increases, capacitor C_2 becomes less effective (a larger portion of the total current is provided by C_1). As a result, the magnitude of C_1 is increased to maintain the voltage across the load above the minimum tolerable level. Similarly, a larger C_2 results in a smaller C_1. As C_2 is increased, a larger portion of the total current is provided by C_2, reducing the magnitude of C_1. This phenomenon is well pronounced for small R_2, diminishing with larger R_2, as illustrated in Fig. 13.7b.

In general, to determine the parameters of the system of distributed on-chip decoupling capacitors, the following assumptions are made. The parasitic resistance of the metal line(s) connecting capacitor C_1 to the current load is known. R_1 is determined by technology constraints (the sheet resistance) and by design constraints (the maximum available metal resources). The minimum voltage level at the load is $V_{load} = 0.9V_{\rm dd}$. The maximum magnitude of the current load $I_{\max}$ is

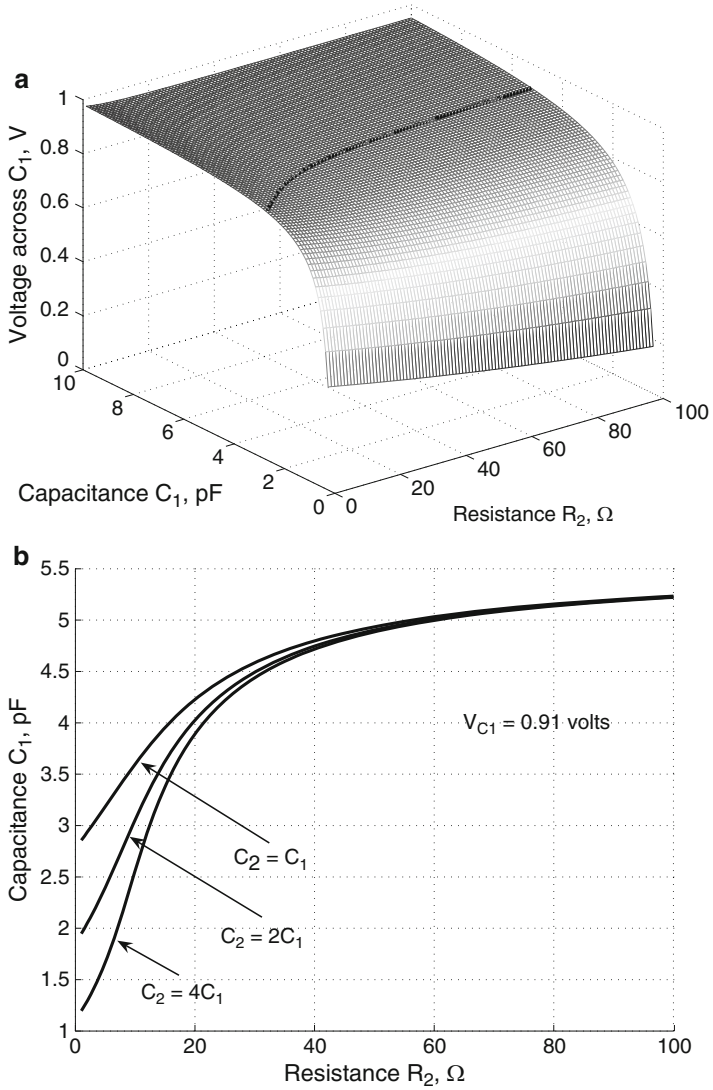

Fig. 13.7 Voltage across C_1 during discharge as a function of C_1 and R_2: $I_{\max} = 0.01$ mA, $V_{\mathrm{dd}} = 1$ V, and $t_r = 100$ ps; (**a**) assuming $C_1 = C_2$ and $R_1 = 10\,\Omega$, the minimum tolerable voltage across C_1, resulting in $V_{load} \geq 0.9V_{\mathrm{dd}}$, is 0.91 V (shown as a *black equipotential line*), (**b**) design space for determining C_1 and R_2 resulting in the voltage across C_1 equal to 0.91 V

0.01 A, the rise time t_r is 100 ps, and the power supply voltage V_{dd} is 1 V. Note that the voltage across C_2 after discharge as determined by (13.4) is also treated as a design parameter. Since the capacitor C_2 is directly connected to the power supply (a shared power rail), the voltage drop across C_2 appears on the global power line,

compromising the signal integrity of the overall system. The voltage across C_2 at t_r is therefore based on the maximum tolerable voltage fluctuations on the P/G line during discharge (the voltage across C_2 at the end of the switching cycle is set to 0.95 V).

The system of equations to determine the parameters of an on-chip distributed decoupling capacitor network as depicted in Fig. 13.6 is

$$V_{load} = V_{C_1} - I_{\max} R_1, \tag{13.7}$$

$$V_{C_1} = f(C_1, C_2, R_2), \tag{13.8}$$

$$V_{C_2} = f(C_1, C_2, R_2), \tag{13.9}$$

$$\frac{I_{\max} t_r}{2} = C_1 \left(V_{\mathrm{dd}} - V_{C_1}\right) + C_2 \left(V_{\mathrm{dd}} - V_{C_2}\right), \tag{13.10}$$

where V_{C_1} and V_{C_2} are the voltage across C_1 and C_2 and determined, respectively, by (13.3) and (13.4). Equation (13.10) states that the total charge drawn by the current load is provided by C_1 and C_2. Note that in the general case with the current load determined a priori, the total charge is the integral of $I_{load}(t)$ from zero to t_r. Solving (13.7) for V_{C_1} and substituting into (13.8), C_1, C_2, and R_2 are determined from (13.8), (13.9), and (13.10) for a specified $V_{C_2}(t_r)$, as discussed in the following section.

13.4 Design Tradeoffs in a Distributed On-Chip Decoupling Capacitor Network

To design a system of distributed on-chip decoupling capacitors, the parasitic resistances and capacitances should be determined based on design and technology constraints. As shown in Sect. 13.3, in a system composed of two decoupling capacitors (see Fig. 13.6) with known R_1; R_2, C_1, and C_2 are determined from the system of Eqs. (13.7), (13.8), (13.9) and (13.10). Note that since this system of equations involves transcendental functions, a closed-form solution cannot be determined. To determine the system parameters, the system of Eqs. (13.7), (13.8), (13.9) and (13.10) is solved numerically [296].

Various tradeoff scenarios are discussed in this section. The dependence of the system parameters on R_1 is presented in Sect. 13.4.1. The design of a distributed on-chip decoupling capacitor network with the minimum magnitude of C_1 is discussed in Sect. 13.4.2. The dependence of C_1 and C_2 on the parasitic resistance of the metal lines connecting the capacitors to the current load is presented in Sect. 13.4.3. The minimum total budgeted on-chip decoupling capacitance is also determined in this section.

Table 13.1 Dependence of the parameters of a distributed on-chip decoupling capacitor network on R_1

	$R_2 = 5\,(\Omega)$		$R_2 = 10\,(\Omega)$	
R_1 (Ω)	C_1 (pF)	C_2 (pF)	C_1 (pF)	C_2 (pF)
1	1.35	7.57	3.64	3.44
2	2.81	5.50	4.63	2.60
3	4.54	3.64	5.88	1.77
4	6.78	1.87	7.56	0.92
5	10.00	0	10.00	0

$V_{\text{dd}} = 1\,\text{V}$, $V_{load} = 0.9\,\text{V}$, $t_r = 100\,\text{ps}$, and $I_{\max} = 0.01\,\text{A}$

13.4.1 Dependence of System Parameters on R_1

The parameters of a distributed on-chip decoupling capacitor network for typical values of R_1 are listed in Table 13.1. Note that the minimum magnitude of R_2 exists for which the parameters of the system can be determined. If R_2 is sufficiently small, the distributed decoupling capacitor network degenerates to a system with a single capacitor (where C_1 and C_2 are combined). For the parameters listed in Table 13.1, the minimum magnitude of R_2 is $4\,\Omega$, as determined from numerical simulations.

Note that the parameters of a distributed on-chip decoupling capacitor network are determined by the parasitic resistance of the P/G line(s) connecting C_1 to the current load. As R_1 increases, the capacitor C_1 increases substantially (see Table 13.1). This increase in C_1 is due to R_1 becoming comparable to R_2, and C_1 providing a greater portion of the total current. Alternatively, the system of distributed on-chip decoupling capacitors degenerates to a single oversized capacitor. The system of distributed on-chip decoupling capacitors should therefore be carefully designed. Since the distributed on-chip decoupling capacitor network is strongly dependent upon the first level of interconnection (R_1), C_1 should be placed as physically close as possible to the current load, reducing R_1. If such an allocation is not practically possible, the current load should be partitioned, permitting an efficient allocation of the distributed on-chip decoupling capacitors under specific technology constraints.

13.4.2 Minimum C_1

In practical applications, the size of C_1 (the capacitor closest to the current load) is typically limited by technology constraints, such as the maximum capacitance density and available area. The magnitude of the first capacitor in the distributed system is therefore typically small. In this section, the dependence of the distributed on-chip decoupling capacitor network on R_1 is determined for minimum C_1. A target magnitude of 1 pF is assumed for C_1. The parameters of a system of distributed

Table 13.2 Distributed on-chip decoupling capacitor network as a function of R_1 under the constraint of a minimum C_1

R_1	$V_{C_2} \neq$ const				$V_{C_2} = 0.95$ V	
(Ω)	R_2 (Ω)	C_2 (pF)	R_2 (Ω)	C_2 (pF)	R_2 (Ω)	C_2 (pF)
1	2	5.59	5	8.69	4.68	8.20
2	2	6.68	5	11.64	3.46	8.40
3	2	8.19	5	17.22	2.28	8.60
4	2	10.46	5	31.70	1.13	8.80
5	2	14.21	5	162.10	–	–

$V_{\text{dd}} = 1$ V, $V_{load} = 0.9$ V, $t_r = 100$ ps, $I_{\max} = 0.01$ A, and $C_1 = 1$ pF

on-chip decoupling capacitors as a function of R_1 under the constraint of a minimum C_1 are listed in Table 13.2. Note that V_{C_2} denotes the voltage across C_2 after discharge.

Note that two scenarios are considered in Table 13.2 to evaluate the dependence of a distributed system of on-chip decoupling capacitors on R_1 and R_2. In the first scenario, the distributed on-chip decoupling capacitor network is designed to maintain the minimum tolerable voltage across the terminals of a current load. In this case, the magnitude of C_2 increases with R_1, becoming impractically large for large R_2. In the second scenario, an additional constraint (the voltage across C_2) is applied to reduce the voltage fluctuations on the shared P/G lines. In this case, as R_1 increases, C_2 slightly increases. In order to satisfy the constraint for V_{C_2}, R_2 should be significantly reduced for large values of R_1, meaning that the second capacitor should be placed close to the first capacitor. As R_1 is further increased, R_2 becomes negligible, implying that capacitors C_1 and C_2 should be merged to provide the required charge to the distant current load. Alternatively, the system of distributed on-chip decoupling capacitors degenerates to a conventional scheme with a single oversized capacitor [297].

Note that simultaneously satisfying both the voltage across the terminals of the current load and the voltage across the last decoupling capacitor is not easy. The system of on-chip distributed decoupling capacitors in this case depends upon the parameters of the first decoupling stage (R_1 and C_1). If C_1 is too small, no solution exists to satisfy $V_{load}^{\min}$ and $V_{C_2}^{\min}$. Sufficient circuit area should therefore be allocated for C_1 early in the design process to provide the required on-chip decoupling capacitance in order to satisfy specific design and technology constraints.

13.4.3 Minimum Total Budgeted On-Chip Decoupling Capacitance

As discussed in Sect. 13.4.1 and 13.4.2, the design of a system of distributed on-chip decoupling capacitors is greatly determined by the parasitic resistance of the metal lines connecting C_1 to the current load and by the magnitude of C_1. Another

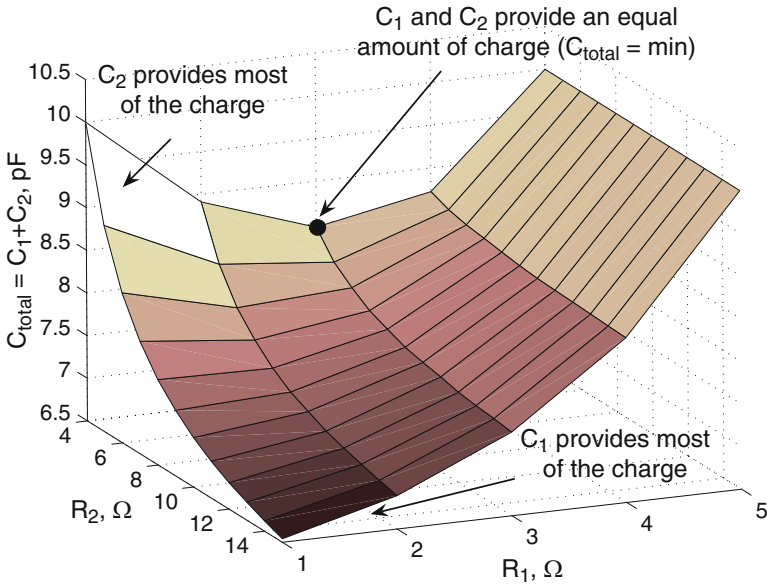

Fig. 13.8 The total budgeted on-chip decoupling capacitance as a function of the parasitic resistance of the metal lines, R_1 and R_2: $I_{\max} = 10\,\text{mA}$, $V_{\text{dd}} = 1\,\text{V}$, $V_{load} = 0.9\,\text{V}$, and $t_r = 100\,\text{ps}$. In the system of distributed on-chip decoupling capacitors, an optimal ratio $\frac{R_2}{R_1}$ exists, resulting in the minimum total budgeted on-chip decoupling capacitance

important design constraint is the total budgeted on-chip decoupling capacitance. Excessive on-chip decoupling capacitance results in increased circuit area and greater leakage currents. Large on-chip decoupling capacitors can also compromise the reliability of the overall system, creating a short circuit between the plates of a capacitor [294]. It is therefore important to reduce the required on-chip decoupling capacitance while providing sufficient charge to support expected current demands.

To estimate the total required on-chip decoupling capacitance, $C_{total} = C_1 + C_2$ is plotted as a function of R_1 and R_2, as depicted in Fig. 13.8. Note that if R_2 is large, C_2 is ineffective and the system of distributed on-chip decoupling capacitors behaves as a single capacitor. Observe from Fig. 13.8 that C_{total} increases with R_1 for large R_2. In this case, C_1 is oversized, providing most of the required charge. C_1 should therefore be placed close to the current load to reduce the total required on-chip decoupling capacitance.

Similarly, if R_2 is reduced with small R_1, C_2 provides most of the charge drawn by the current load. The distributed on-chip decoupling capacitor network degenerates to a conventional system with a single capacitor. As R_1 increases, however, the total required on-chip decoupling capacitance decreases, reaching the minimum (see Fig. 13.8 for $R_1 = 3\,\Omega$ and $R_2 = 4\,\Omega$). In this case, C_1 and C_2 each provide an equal amount of the total charge. As R_1 is further increased (C_1 is placed farther from the current load), C_1 and C_2 increase substantially to compensate for the

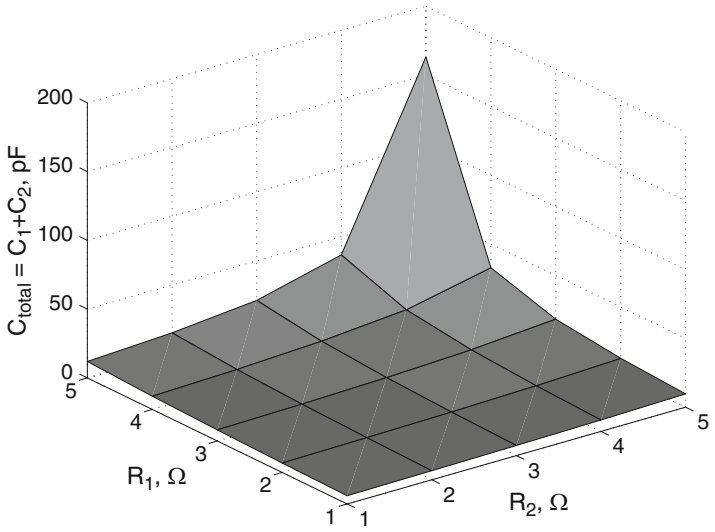

Fig. 13.9 The total budgeted on-chip decoupling capacitance as a function of the parasitic resistance of the metal lines, R_1 and R_2: $I_{\max} = 10\,\text{mA}$, $V_{\text{dd}} = 1\,\text{V}$, $V_{load} = 0.9\,\text{V}$, and $t_r = 100\,\text{ps}$. C_1 is fixed and set to 1 pF. The total budgeted on-chip decoupling capacitance increases with R_1 and R_2. As the parasitic resistance of the metal lines is further increased beyond $4\,\Omega$, C_{total} increases substantially, becoming impractically large

increased voltage drop across R_1. In the system of distributed on-chip decoupling capacitors, an optimal ratio $\frac{R_2}{R_1}$ exists which requires the minimum total budgeted on-chip decoupling capacitance.

Note that in the previous scenario, the magnitude of the on-chip decoupling capacitors has not been constrained. In practical applications, however, the magnitude of the first decoupling capacitor (placed close to the current load) is limited. To determine the dependence of the total required on-chip decoupling capacitance under the magnitude constraint of C_1, C_1 is fixed and set to 1 pF. $C_{total} = C_1 + C_2$ is plotted as a function of R_1 and R_2, as shown in Fig. 13.9. In contrast to the results depicted in Fig. 13.8, the total budgeted on-chip decoupling capacitance required to support expected current demands increases with R_1 and R_2. Alternatively, C_2 provides the major portion of the total charge. Thus, the system behaves as a single distant on-chip decoupling capacitor. In this case, C_1 is too small. A larger area should therefore be allocated for C_1, resulting in a balanced system with a reduced total on-chip decoupling capacitance. Also note that as R_1 and R_2 further increase (beyond $4\,\Omega$, see Fig. 13.9), the total budgeted on-chip decoupling capacitance increases rapidly, becoming impractically large.

Comparing Figs. 13.8 and 13.9, note that if C_1 is constrained, a larger total decoupling capacitance is required to provide the charge drawn by the current load. Alternatively, the system of distributed on-chip decoupling capacitors under a magnitude constraint of C_1 behaves as a single distant decoupling capacitor. As

a result, the magnitude of a single decoupling capacitor is significantly increased to compensate for the *IR* voltage drop across R_1 and R_2. The system of distributed on-chip decoupling capacitors should therefore be carefully designed to reduce the total budgeted on-chip decoupling capacitance. If the magnitude of C_1 is limited, C_2 should be placed close to the current load to be effective, reducing the total required on-chip decoupling capacitance. Alternatively, the parasitic impedance of the P/G lines connecting C_1 and C_2 should be reduced (e.g., utilizing wider lines and/or multiple lines in parallel) [73].

13.5 Design Methodology for a System of Distributed On-Chip Decoupling Capacitors

An overall methodology for designing a distributed system of on-chip decoupling capacitors is illustrated in Fig. 13.10. General differential equations for voltages $V_{C_1}(t)$ and $V_{C_2}(t)$ across capacitors C_1 and C_2 are derived based on Kirchhoff's laws. The maximum parasitic resistance $R_1^{\max}$ between C_1 and the current load is determined from (13.6) for specific parameters of the system, such as the power supply voltage V_{dd}, the minimum voltage across the terminals of the current load V_{load}, the maximum magnitude of the current load $I_{\max}$, and the rise time t_r. If $R_1 > R_1^{\max}$, no solution exists for the system of distributed on-chip decoupling capacitors. Alternatively, the voltage across the terminals of a current load always drops below the minimum acceptable level. In this case, the current load should be partitioned to reduce $I_{\max}$, resulting in $R_1 < R_1^{\max}$.

Simultaneously solving (13.1) and (13.2), the voltage across C_1 and C_2 is estimated at the end of a switching cycle ($t = t_r$), as determined by (13.3) and (13.4). The parameters of the distributed on-chip decoupling capacitor network C_1, C_2, and R_2, are determined from (13.7), (13.8), (13.9) and (13.10). Note that different tradeoffs exist in a system of distributed on-chip decoupling capacitors, as discussed in Sect. 13.4. If the voltage across the terminals of a current load drops below the minimum tolerable level, the total budgeted on-chip decoupling capacitance should be increased. The system of Eqs. (13.7), (13.8), (13.9) and (13.10), is solved for an increased total on-chip decoupling capacitance, resulting in different C_1, C_2, and R_2 until the criterion for the maximum tolerable power noise $V_{load} > V_{load}^{\min}$ is satisfied, as shown in Fig. 13.10.

Note that the system of distributed on-chip decoupling capacitors permits the design of an effective power distribution system under specified technology constraints. The techniques presented in this chapter are also applicable to future technology generations. The methodology also provides a computationally efficient way to determine the required on-chip decoupling capacitance to support expected current demands. In the worst case example presented in this chapter, the simulation time to determine the parameters of the system of on-chip distributed decoupling capacitors is under one second on a Pentium III PC with one gigabyte of RAM. A methodology for efficiently placing on-chip decoupling capacitors can also be

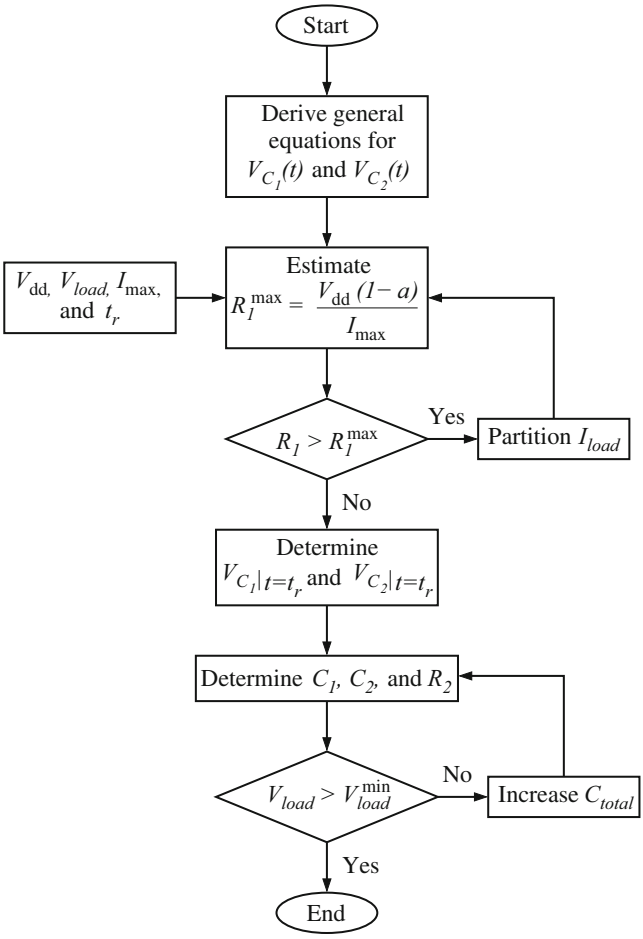

Fig. 13.10 Design flow for determining the parameters of a system of distributed on-chip decoupling capacitors

integrated into a standard IC design flow. In this way, the circuit area required to allocate on-chip decoupling capacitors is estimated early in the design process, significantly reducing the number of iterations and the eventual time to market.

13.6 Case Study

The dependence of the system of distributed on-chip decoupling capacitors on the current load and the parasitic impedance of the power delivery system is described in this section to quantitatively illustrate the previously presented concepts. Resistive

power and ground lines are assumed to connect the decoupling capacitors to the current load and are modeled as resistors (see Fig. 13.6). The load is modeled as a ramp current source with a 100 ps rise time. The minimum tolerable voltage across the load terminals is 90 % of the power supply. The magnitude of the on-chip decoupling capacitors for various parasitic resistances of the metal lines connecting the capacitors to the current load is listed in Table 13.3. The parameters of the distributed on-chip decoupling capacitor network listed in Table 13.3 are determined for two amplitudes of the current load. Note that the values of R_1 and R_2 are typical parasitic resistances of an on-chip power distribution grid for a 90 nm CMOS technology.

The parameters of the system of distributed on-chip decoupling capacitors are analytically determined from (13.7), (13.8), (13.9) and (13.10). The resulting power supply noise is estimated using SPICE and compared to the maximum tolerable level (the minimum voltage across the load terminals $V_{load}^{\min}$). The maximum voltage drop across C_2 at the end of the switching activity is also estimated and compared to $V_{C_2}^{\min}$. Note that the analytic solution produces an accurate estimate of the on-chip decoupling capacitors for typical parasitic resistances of a power distribution grid. The maximum error in this case study is 0.003 %.

From Table 13.3, note that in the case of a large R_2, the distributed decoupling capacitor network degenerates into a system with a single capacitor. Capacitor C_1 is therefore excessively large. Conversely, if C_2 is placed close to C_1 (R_2 is small), C_2 is excessively large and the system again behaves as a single capacitor. An optimal ratio $\frac{R_2}{R_1}$ therefore exists for specific characteristics of the current load that results in a minimum required on-chip decoupling capacitance. Alternatively, in this case, both capacitors provide an equal portion of the total charge (see Table 13.3 for $R_1 = 0.5\,\Omega$ and $R_2 = 10\,\Omega$). Also note that as the magnitude of the current load increases, larger on-chip decoupling capacitors are required to provide the expected current demands.

The parameters of a distributed on-chip decoupling capacitor network listed in Table 13.3 have been determined for the case where the magnitude of the decoupling capacitors is not limited. In most practical systems, however, the magnitude of the on-chip decoupling capacitor placed closest to the current load is limited by technology and design constraints. A case study of a system of distributed on-chip decoupling capacitors with a limited value of C_1 is listed in Table 13.4. Note that in contrast to Table 13.3, where both R_1 and R_2 are design parameters, in the system with a limit on C_1, R_2 and C_2 are determined by R_1. Alternatively, both the magnitude and location of the second capacitor are determined from the magnitude and location of the first capacitor.

The parameters of the distributed on-chip decoupling capacitor network listed in Table 13.4 are determined for two amplitudes of the current load with R_1 representing a typical parasitic resistance of the metal line connecting C_1 to the current load. The resulting power supply noise at the current load and across the last decoupling stage is estimated using SPICE and compared to the maximum tolerable levels, respectively, $V_{load}^{\min}$ and $V_{C_2}^{\min}$. Note that the analytic solution accurately estimates the parameters of the distributed on-chip decoupling capacitor network, producing a worst case error of 0.0001 %.

Table 13.3 The magnitude of the on-chip decoupling capacitors as a function of the parasitic resistance of the power/ground lines connecting the capacitors to the current load

R_1	R_2	$I_{\max}$	C_1	C_2	V_{load} (mV)		Error	V_{C_2} (mV)		Error
(Ω)	(Ω)	(A)	(pF)	(pF)	$V_{load}^{\min}$	SPICE	(%)	$V_{C_2}^{\min}$	SPICE	(%)
0.5	4.5	0.01	0	9.99999	900	899.999	0.0001	950	949.999	0.0001
0.5	6	0.01	1.59747	6.96215	900	899.986	0.002	950	949.983	0.002
0.5	8	0.01	2.64645	4.97091	900	899.995	0.0006	950	949.993	0.0004
0.5	10	0.01	3.22455	3.87297	900	899.997	0.0003	950	949.996	0.0004
0.5	12	0.01	3.59188	3.17521	900	899.998	0.0002	950	949.997	0.0003
0.5	14	0.01	3.84641	2.69168	900	899.998	0.0002	950	949.997	0.0003
0.5	16	0.01	4.03337	2.33650	900	899.999	0.0001	950	949.998	0.0002
0.5	18	0.01	4.17658	2.06440	900	899.998	0.0002	950	949.998	0.0002
0.5	20	0.01	4.28984	1.84922	900	899.999	0.0001	950	949.998	0.0002
0.5	1.5	0.025	0	24.99930	900	899.998	0.0002	950	949.998	0.0002
0.5	2	0.025	4.25092	17.56070	900	899.999	0.0001	950	949.999	0.0001
0.5	3	0.025	7.97609	11.04180	900	899.999	0.0001	950	949.999	0.0001
0.5	4	0.025	9.67473	8.06921	900	899.999	0.0001	950	949.999	0.0001
0.5	5	0.025	10.65000	6.36246	900	899.999	0.0001	950	949.999	0.0001
0.5	6	0.025	11.2838	5.25330	900	899.999	0.0001	950	949.999	0.0001
0.5	7	0.025	11.72910	4.47412	900	899.999	0.0001	950	949.999	0.0001
0.5	8	0.025	12.05910	3.89653	900	899.999	0.0001	950	949.999	0.0001
0.5	9	0.025	12.31110	3.44905	900	899.980	0.002	950	949.973	0.003
1	4	0.01	0	9.99999	900	899.999	0.0001	950	949.999	0.0001
1	6	0.01	2.16958	6.09294	900	899.990	0.001	950	949.988	0.001
1	8	0.01	3.11418	4.39381	900	899.996	0.0004	950	949.994	0.0006
1	10	0.01	3.64403	3.44040	900	899.997	0.0003	950	949.996	0.0004
1	12	0.01	3.98393	2.82871	900	899.998	0.0002	950	949.997	0.0003
1	14	0.01	4.22079	2.40240	900	899.998	0.0002	950	949.997	0.0003
1	16	0.01	4.39543	2.08809	900	899.998	0.0002	950	949.997	0.0003
1	18	0.01	4.52955	1.84668	900	899.998	0.0002	950	949.998	0.0002
1	20	0.01	4.63582	1.65540	900	899.998	0.0002	950	949.998	0.0002
1	1	0.025	0	24.99940	900	899.998	0.0002	950	949.998	0.0002
1	2	0.025	9.08053	11.37910	900	899.999	0.0001	950	949.999	0.0001
1	3	0.025	11.74820	7.37767	900	899.999	0.0001	950	949.999	0.0001
1	4	0.025	13.02600	5.46100	900	899.999	0.0001	950	949.999	0.0001
1	5	0.025	13.77630	4.33559	900	899.999	0.0001	950	949.999	0.0001
1	6	0.025	14.27000	3.59504	900	899.999	0.0001	950	949.999	0.0001
1	7	0.025	14.61950	3.07068	900	899.999	0.0001	950	949.999	0.0001
1	8	0.025	14.88010	2.67987	900	899.999	0.0001	950	949.999	0.0001
1	9	0.025	15.08180	2.37733	900	899.999	0.0001	950	949.999	0.0001

$V_{\rm dd} = 1$ V and $t_r = 100$ ps

Table 13.4 The magnitude of the on-chip decoupling capacitors as a function of the parasitic resistance of the power/ground lines connecting the capacitors to the current load for a limit on C_1

R_1	I_{max}	R_2	C_2	V_{load} (mV)		Error	V_{C_2} (mV)		Error
(Ω)	(A)	(Ω)	(pF)	V_{load}^{min}	SPICE	(%)	$V_{C_2}^{min}$	SPICE	(%)
$C_1 = 0.5$ pF									
1	0.005	10.6123	4.05	900	899.999	0.0001	950	949.999	0.0001
2	0.005	9.3666	4.10	900	899.999	0.0001	950	949.999	0.0001
3	0.005	8.1390	4.15	900	899.999	0.0001	950	949.999	0.0001
4	0.005	6.9290	4.20	900	899.999	0.0001	950	949.999	0.0001
5	0.005	5.7354	4.25	900	899.999	0.0001	950	949.999	0.0001
0.5	0.01	4.8606	9.05	900	899.999	0.0001	950	949.999	0.0001
1	0.01	4.3077	9.10	900	900.000	0.0000	950	949.999	0.0001
2	0.01	3.2120	9.20	900	899.999	0.0001	950	949.999	0.0001
3	0.01	2.1290	9.30	900	899.999	0.0001	950	949.999	0.0001
4	0.01	1.0585	9.40	900	899.999	0.0001	950	949.999	0.0001
$C_1 = 1$ pF									
1	0.005	13.2257	3.1	900	899.999	0.0001	950	949.999	0.0001
2	0.005	11.5092	3.2	900	899.999	0.0001	950	949.999	0.0001
3	0.005	9.8686	3.3	900	899.999	0.0001	950	949.999	0.0001
4	0.005	8.2966	3.4	900	899.999	0.0001	950	949.999	0.0001
5	0.005	6.7868	3.5	900	899.999	0.0001	950	949.999	0.0001
0.5	0.01	5.3062	8.1	900	899.999	0.0001	950	949.999	0.0001
1	0.01	4.6833	8.2	900	899.999	0.0001	950	949.999	0.0001
2	0.01	3.4644	8.4	900	899.999	0.0001	950	949.999	0.0001
3	0.01	2.2791	8.6	900	899.999	0.0001	950	949.999	0.0001
4	0.01	1.1250	8.8	900	899.999	0.0001	950	949.999	0.0001

$V_{dd} = 1$ V and $t_r = 100$ ps

Comparing results from Table 13.4 for two different magnitudes of C_1, note that a larger C_1 results in a smaller C_2. A larger C_1 also relaxes the constraints for the second decoupling stage, permitting C_2 to be placed farther from C_1. The first stage of a system of distributed on-chip decoupling capacitors should therefore be carefully designed to provide a balanced distributed decoupling capacitor network with a minimum total required capacitance, as discussed in Sect. 13.4.3.

On-chip decoupling capacitors have traditionally been allocated during a post-layout iteration (after the initial allocation of the standard cells). The on-chip decoupling capacitors are typically inserted into the available white space. If significant area is required for an on-chip decoupling capacitor, the circuit blocks are iteratively rearranged until the timing and signal integrity constraints are satisfied. Traditional strategies for placing on-chip decoupling capacitors therefore result in increased time to market, design effort, and cost.

The methodology for placing on-chip decoupling capacitors presented in this chapter permits simultaneous allocation of the on-chip decoupling capacitors and

the circuit blocks. In this methodology, a current profile of a specific circuit block is initially estimated [298]. The magnitude and location of the distributed on-chip decoupling capacitors are determined based on expected current demands and technology constraints, such as the maximum capacitance density and parasitic resistance of the metal lines connecting the decoupling capacitors to the current load. Note that the magnitude of the decoupling capacitor closest to the current load should be determined for each circuit block, resulting in a balanced system and the minimum required total on-chip decoupling capacitance. As the number of decoupling capacitors increases, the parameters of a distributed on-chip decoupling capacitor network are relaxed, permitting the decoupling capacitors to be placed farther from the optimal location (permitting the parasitic resistance of the metal lines connecting the decoupling capacitors to vary over a larger range). In this way, the maximum effective radii of a distant on-chip decoupling capacitor is significantly increased [277]. A tradeoff therefore exists between the magnitude and location of the on-chip decoupling capacitors comprising the distributed decoupling capacitor network.

13.7 Summary

A design methodology for placing distributed on-chip decoupling capacitors in nanoscale ICs can be summarized as follows.

- On-chip decoupling capacitors have traditionally been allocated into the available white space using an unsystematic approach. In this way, the on-chip decoupling capacitors are often placed far from the current load
- Existing allocation strategies result in increased power noise, compromising the signal integrity of an entire system
- Increasing the size of the on-chip decoupling capacitors allocated with conventional techniques does not enhance power delivery
- An on-chip decoupling capacitor should be placed physically close to the current load to be effective
- Since the area occupied by the on-chip decoupling capacitor is directly proportional to the magnitude of the capacitor, the minimum impedance between the on-chip decoupling capacitor and the current load is fundamentally affected by the magnitude of the capacitor
- A system of distributed on-chip decoupling capacitors has been described in this chapter to resolve this dilemma. A distributed on-chip decoupling capacitor network is an efficient solution for providing sufficient on-chip decoupling capacitance while satisfying existing technology constraints
- An optimal ratio of the parasitic resistance of the metal lines connecting the capacitors exists, permitting the total budgeted on-chip decoupling capacitance to be significantly reduced

- Simulation results for typical values of the on-chip parasitic resistances are also presented, demonstrating high accuracy of the analytic solution. In the worst case, the maximum error is 0.003 % as compared to SPICE
- A distributed on-chip decoupling capacitor network permits the on-chip decoupling capacitors and the circuit blocks to be simultaneously placed within a single design step