# Chapter 11
# Decoupling Capacitance

The on-going miniaturization of integrated circuit feature sizes has placed significant requirements on the on-chip power and ground distribution networks. Circuit integration densities rise with each nanoscale technology generation due to smaller devices and larger dies. The on-chip current densities and the total current also increase. Simultaneously, the higher switching speed of smaller transistors produces faster current transients in the power distribution network. Supplying high average currents and continuously increasing transient currents through the high impedance on-chip interconnects results in significant fluctuations of the power supply voltage in scaled CMOS technologies.

Such a change in the supply voltage is referred to as power supply noise. Power supply noise adversely affects circuit operation through several mechanisms, as described in Chap. 1. Supplying sufficient power current to high performance ICs has therefore become a challenging task. Large average currents result in increased $IR$ noise and fast current transients result in increased $L\ dI/dt$ voltage drops ($\Delta I$ noise) [23].

Decoupling capacitors are often utilized to manage this power supply noise. Decoupling capacitors can have a significant effect on the principal characteristics of an integrated circuit, i.e., speed, cost, and power. Due to the importance of decoupling capacitors in current and future ICs, significant research has been developed over the past several decades, covering different areas such as hierarchical placement of decoupling capacitors, sizing and placing of on-chip decoupling capacitors, resonant phenomenon in power distribution systems with decoupling capacitors, and static on-chip power dissipation due to leakage current through the gate oxide.

In this chapter, a brief review of the background of decoupling capacitance is provided. In Sect. 11.1, the concept of a decoupling capacitance is described and an historical retrospective is described. A practical model of a decoupling capacitor is also described. In Sect. 11.2, the impedance of a power distribution system with decoupling capacitors is presented. Target specifications of the impedance



I.P. Vaisband et al., *On-Chip Power Delivery and Management*,
DOI 10.1007/978-3-319-29395-0_11

of a power distribution system are reviewed. Antiresonance phenomenon in a system with decoupling capacitors is intuitively explained. A hydraulic analogy of the hierarchical placement of decoupling capacitors is also presented. Intrinsic and intentional on-chip decoupling capacitances are discussed and compared in Sect. 11.3. Different types of on-chip decoupling capacitors are qualitatively analyzed in Sect. 11.4. The advantages and disadvantages of several types of widely used on-chip decoupling capacitors are also discussed in Sect. 11.4. Enhancing the efficiency of on-chip decoupling capacitors with a switching voltage regulator is presented in Sect. 11.5. Finally, some conclusions are offered in Sect. 11.6.

## 11.1 Introduction to Decoupling Capacitance

Decoupling capacitors are often used to maintain the power supply voltage within specification so as to provide signal integrity while reducing electromagnetic interference (EMI) radiated noise. In this book, the use of decoupling capacitors to mitigate power supply noise is evaluated. The concept of a decoupling capacitor is described in this section. An historical retrospective is presented in Sect. 11.1.1. A description of a decoupling capacitor as a reservoir of charge is discussed in Sect. 11.1.2. Decoupling capacitors are shown to be an effective way to provide sufficient charge to a switching current load within a short period of time. A practical model of a decoupling capacitor is presented in Sect. 11.1.3.

### *11.1.1 Historical Retrospective*

About 600 BC, Thales of Miletus recorded that the ancient Greeks could generate sparks by rubbing balls of amber on spindles [196]. This is the triboelectric effect [197], the mechanical separation of charge in a dielectric (insulator). This effect is the basis of the capacitor.

In October 1745, Ewald Georg von Kleist of Pomerania invented the first recorded capacitor: a glass jar coated inside and out with metal. The inner coating was connected to a rod that passed through the lid and ended in a metal sphere, as shown in Fig. 11.1 [198]. By layering the insulator between two metal plates, von Kleist dramatically increased the charge density. Before Kleist's discovery became widely known, a Dutch physicist, Pieter van Musschenbroek, independently invented a similar capacitor in January 1746 [199]. It was named the Leyden jar, after the University of Leyden where van Musschenbroek worked.

Benjamin Franklin investigated the Leyden jar and proved that the charge was stored on the glass, not in the water as others had assumed [200]. Originally, the units of capacitance were in "jars." A jar is equivalent to about 1 nF. Early capacitors were also known as *condensors*, a term that is still occasionally used today. The

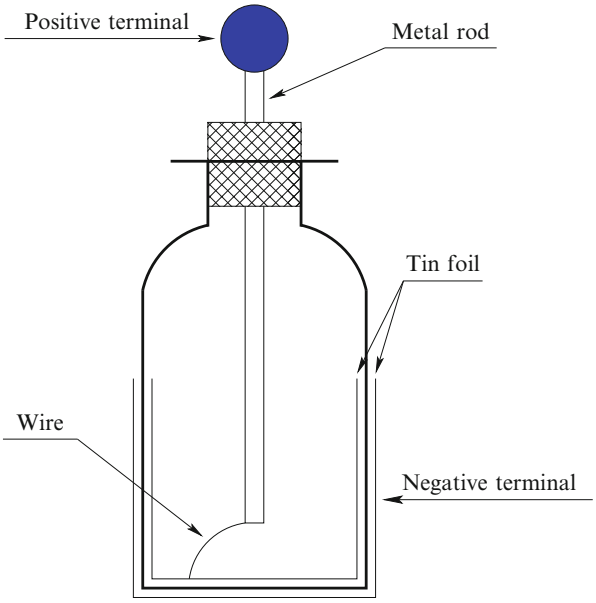


**Fig. 11.1** Leyden jar originally developed by Ewald Georg von Kleist in 1745 and independently invented by Pieter van Musschenbroek in 1746. The charge is stored on the glass between two tin foils (capacitor plates) [198]

term condensor was coined by Alessandro Volta in 1782 (derived from the Italian *condensatore*), referencing the ability of a device to store a higher density of electric charge than a normal isolated conductor [200].

### *11.1.2 Decoupling Capacitor as a Reservoir of Charge*

A capacitor consists of two electrodes, or plates, each of which stores an equal amount of opposite charge. These two plates are conductive and are separated by an insulator (dielectric). The charge is stored on the surface of the plates at the boundary with the dielectric. Since each plate stores an equal but opposite charge, the net charge across the capacitor is always zero.

The capacitance $C$ of a capacitor is a measure of the amount of charge $Q$ stored on each plate for a given potential difference (voltage $V$) which appears between the plates,

$$C = \frac{Q}{V}. \tag{11.1}$$

The capacitance is proportional to the surface area of the conducting plate and inversely proportional to the distance between the plates [201]. The capacitance

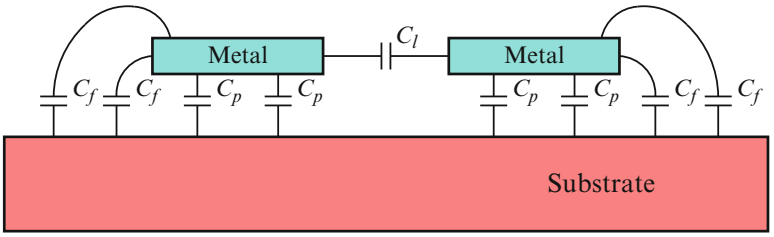


**Fig. 11.2** Capacitance of two metal lines placed over a substrate. Three primary components compose the total capacitance of the on-chip metal interconnects. $C_l$ denotes the lateral flux (side) capacitance, $C_f$ denotes the fringe capacitance, and $C_p$ denotes the parallel plate capacitance

is also proportional to the permittivity of the dielectric substance that separates the plates. The capacitance of a parallel plate capacitor is

$$C \approx \frac{\epsilon A}{d}, \tag{11.2}$$

where $\epsilon$ is the permittivity of the dielectric, $A$ is the area of the plates, and $d$ is the spacing between the plates. Equation (11.2) is only accurate for a plate area much greater than the spacing between the plates, $A \gg d^2$. In general, the capacitance of the metal interconnects placed over the substrate is composed of three primary components: a parallel plate capacitance, fringe capacitance, and lateral flux (side) capacitance [202], as shown in Fig. 11.2. Accurate closed-form expressions have been developed by numerically fitting a model that describes parallel lines above the plane or between two parallel planes [203–208].

As opposite charge accumulates on the plates of a capacitor across an insulator, a voltage develops across the capacitor due to the electric field formed by the opposite charge. Work must be done against this electric field as more charge is accumulated. The energy stored in a capacitor is equal to the amount of work required to establish the voltage across the capacitor. The energy stored in the capacitor is

$$E_{\text{stored}} = \frac{1}{2}CV^2 = \frac{1}{2}\frac{Q^2}{C} = \frac{1}{2}VQ. \tag{11.3}$$

From a physical perspective, a decoupling capacitor serves as an intermediate storage of charge and energy. The decoupling capacitor is located between the power supply and current load, i.e., electrically closer to the switching circuit. The decoupling capacitor is therefore more efficient in terms of supplying charge as compared to a remote power supply. The amount of charge stored on the decoupling capacitor is limited by the voltage and the capacitance. Unlike a decoupling capacitor, the power supply can provide an almost infinite amount of charge. A hydraulic model of a decoupling capacitor is illustrated in Fig. 11.3. Similar to water stored in a water tank and connected to the consumer through a system of pipes, the charge on the decoupling capacitor stored between the conductive plates

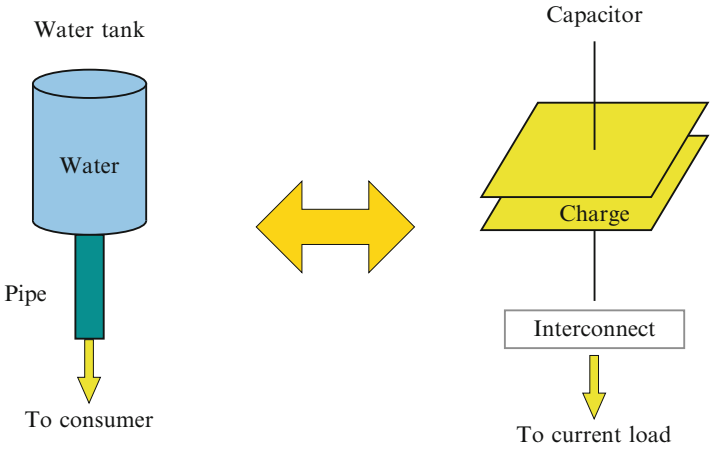


**Fig. 11.3** Hydraulic model of a decoupling capacitor as a reservoir of charge. Similar to water stored in a water tank and connected to the consumer through a system of pipes, charge on the decoupling capacitor is stored between the conductive plates connected to the current load through a hierarchical interconnect system

is connected to the current load through a hierarchical interconnect system. To be effective, the decoupling capacitor should satisfy two requirements. First, the capacitor should have sufficient capacity to store a significant amount of energy. Second, to supply sufficient power at high frequencies, the capacitor should be able to release and accumulate energy at a high rate.

### *11.1.3 Practical Model of a Decoupling Capacitor*

Decoupling capacitors are often used in power distribution systems to provide the required charge in a timely manner and to reduce the output impedance of the overall power delivery network [51]. An ideal decoupling capacitor is effective over the entire frequency range: from DC to the maximum operating frequency of a system. Practically, a decoupling capacitor is only effective over a certain frequency range. The impedance of a practical decoupling capacitor decreases linearly with frequency at low frequencies (with a slope of $-20$ dB/dec in a logarithmic scale). As the frequency increases, the impedance of the decoupling capacitor increases linearly with frequency (with a slope of 20 dB/dec in a logarithmic scale), as shown in Fig. 11.4. This increase in the impedance of a practical decoupling capacitor is due to the parasitic inductance of the decoupling capacitor. The parasitic inductance is referred to as the effective series inductance (ESL) of a decoupling capacitor [126].The impedance of a decoupling capacitor reaches the minimum impedance at the frequency $\omega = \frac{1}{\sqrt{LC}}$. This frequency is known as the resonant

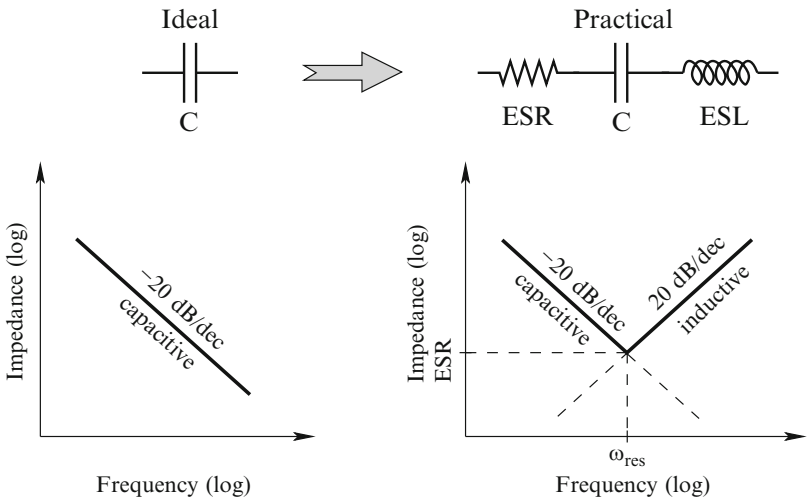


**Fig. 11.4** Practical model of a decoupling capacitor. The impedance of a practical decoupling capacitor decreases linearly with frequency, reaching the minimum at a resonant frequency. Beyond the resonant frequency, the impedance of the decoupling capacitor increases linearly with frequency due to the ESL. The minimum impedance is determined by the ESR of the decoupling capacitor

frequency of a decoupling capacitor. Observe that the absolute minimum impedance of a decoupling capacitor is limited by the parasitic resistance, i.e., the effective series resistance (ESR) of a decoupling capacitor.The parasitic resistance of a decoupling capacitor is due to the resistance of the metal leads and conductive plates and the dielectric losses of the insulator. The ESR and ESL of an on-chip metal-oxide-semiconductor (MOS) decoupling capacitor are illustrated in Fig. 11.5. Note that the parasitic inductance of the decoupling capacitor is determined by the area of the current loops, decreasing with smaller area, as shown in Fig. 11.5b [209].

The impedance of a decoupling capacitor depends upon a number of characteristics. For instance, as the capacitance is increased, the capacitive curve moves down and to the right (see Fig. 11.4). Since the parasitic inductance for a particular capacitor is fixed, the inductive curve remains unaffected. As different capacitors are selected, the capacitive curve moves up and down relative to the fixed inductive curve. The primary way to decrease the total impedance of a decoupling capacitor for a specific semiconductor package is to increase the value of the capacitor [211]. Note that to move the inductive curve down, lowering the total impedance characteristics, a number of decoupling capacitors should be connected in parallel. In the case of identical capacitors, the total impedance is reduced by a factor of 2 for each doubling in the number of capacitors [136].

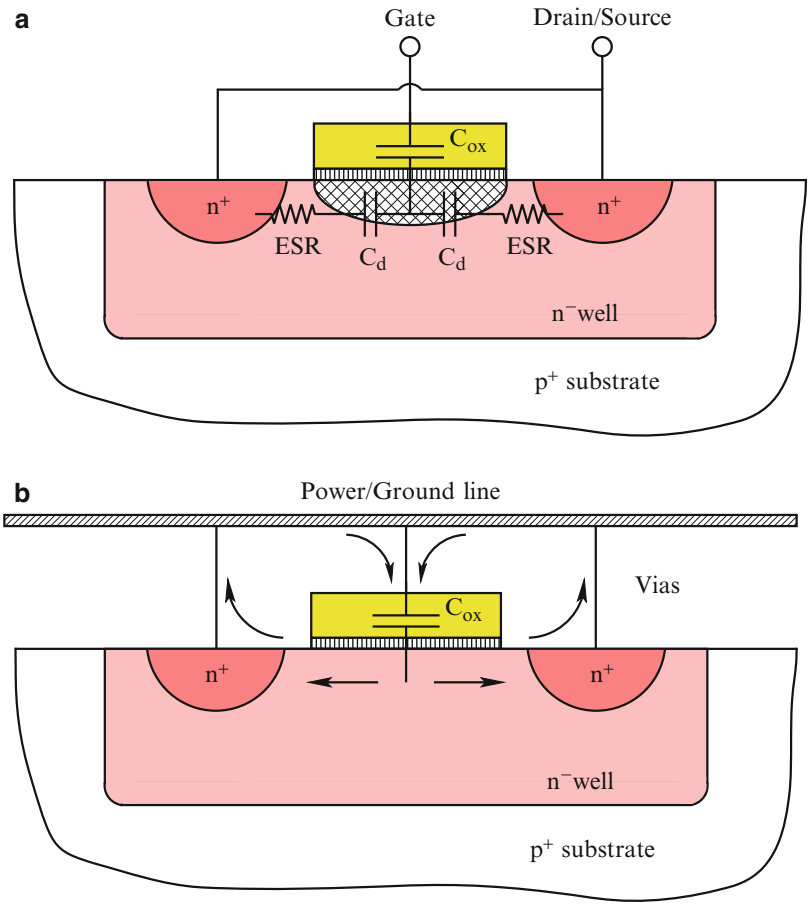


**Fig. 11.5** Physical structure of an on-chip MOS decoupling capacitor. (**a**) ESR of an MOS-based decoupling capacitor. The ESR of an on-chip MOS decoupling capacitor is determined by the doping profiles of the $n^+$ regions and $n^-$ well, the size of the capacitor, and the impedance of the vias and gate material [210]. (**b**) ESL of an MOS-based decoupling capacitor. The ESL of an on-chip MOS decoupling capacitor is determined by the area of the current return loops. The parasitic inductance is lowered by shrinking the area of the current return loops

## 11.2 Impedance of Power Distribution System with Decoupling Capacitors

As described in Sect. 11.1.2, a decoupling capacitor serves as a reservoir of charge, providing the required charge to the switching current load. Decoupling capacitors are also used to lower the impedance of the power distribution system. The impedance of a decoupling capacitor decreases rapidly with frequency, shunting the high frequency currents and reducing the effective current loop of a power distribution

network. The impedance of the overall power distribution system with decoupling capacitors is the subject of this section. In Sect. 11.2.1, the target impedance of a power distribution system is described. It is shown that the impedance of a power distribution system should be maintained below a target level to guarantee fault-free operation of the entire system. The antiresonance phenomenon is presented in Sect. 11.2.2. A hydraulic analogy of a system of decoupling capacitors is described in Sect. 11.2.3. The analogy is drawn between a water supply system and the hierarchical placement of decoupling capacitors at different levels of a power delivery network.

### *11.2.1 Target Impedance of a Power Distribution System*

To ensure a small variation in the power supply voltage under a significant current load, the power distribution system should exhibit a small impedance as seen from the current load within the frequency range of interest [114]. A circuit network representing the impedance of a power distribution system as seen from the terminals of the current load is shown in Fig. 11.6.

The impedance of a power distribution system is with respect to the terminals of the load circuits. In order to ensure correct and reliable operation of an IC, the impedance of a power distribution system should be maintained below a certain upper bound $Z_{\text{target}}$ in the frequency range from DC to the maximum operating frequency $f_0$ of the system [212–214]. The maximum tolerable impedance of a power distribution system is henceforth referred to as the target impedance. Note that the maximum operating frequency $f_0$ is determined by the switching time of the on-chip signal transients, rather than by the clock frequency. The shortest signal switching time is typically an order of magnitude smaller than the clock period. The maximum operating frequency is therefore considerably higher than the clock frequency.

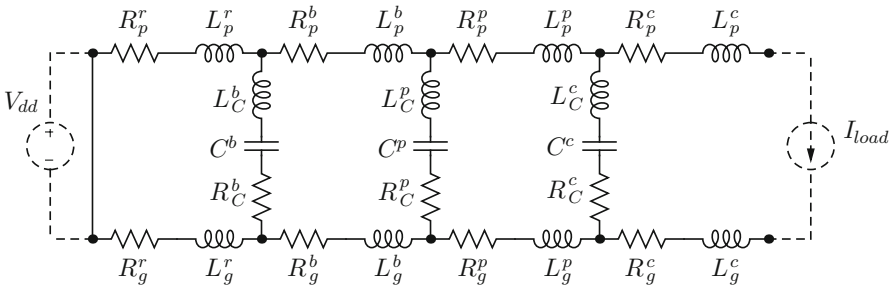


**Fig. 11.6** A circuit network representing the impedance of a power distribution system with decoupling capacitors as seen from the terminals of the current load. The ESR and ESL of the decoupling capacitors are also included. Subscript *p* denotes the power paths and subscript *g* denotes the ground path. Superscripts *r*, *b*, *p*, and *c* refer, respectively, to the voltage regulator, board, package, and on-chip power delivery networks

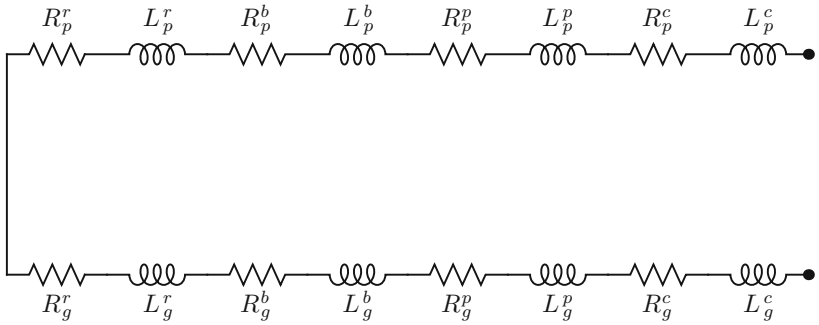


**Fig. 11.7** A circuit network representing the impedance of a power distribution system without decoupling capacitors

One primary design objective of an effective power distribution system is to ensure that the output impedance of the network is below a target output impedance level. It is therefore important to understand how the output impedance of the circuit, shown schematically in Fig. 11.6, depends upon the impedance of the comprising circuit elements. A power distribution system with no decoupling capacitors is shown in Fig. 11.7. The power source and load are connected by interconnect with resistive and inductive parasitic impedances. The magnitude of the impedance of this network is

$$|Z_{\text{tot}}(\omega)| = |R_{\text{tot}} + j\omega L_{\text{tot}}|\,, \tag{11.4}$$

where $R_{\text{tot}}$ and $L_{\text{tot}}$ are the total resistance and inductance of the power distribution system, respectively,

$$R_{\text{tot}} = R_{\text{tot}}^p + R_{\text{tot}}^g, \tag{11.5}$$

$$R_{\text{tot}}^p = R_p^r + R_p^b + R_p^p + R_p^c, \tag{11.6}$$

$$R_{\text{tot}}^g = R_g^r + R_g^b + R_g^p + R_g^c, \tag{11.7}$$

$$L_{\text{tot}} = L_{\text{tot}}^p + L_{\text{tot}}^g, \tag{11.8}$$

$$L_{\text{tot}}^p = L_p^r + L_p^b + L_p^p + L_p^c, \tag{11.9}$$

$$L_{\text{tot}}^g = L_g^r + L_g^b + L_g^p + L_g^c. \tag{11.10}$$

The variation of the impedance with frequency is illustrated in Fig. 11.8. To satisfy a specification at low frequency, the resistance of the power delivery network should be sufficiently low, $R_{\text{tot}} < Z_{\text{target}}$. Above the frequency $f_{L_{\text{tot}}} = \frac{1}{2\pi}\frac{R_{\text{tot}}}{L_{\text{tot}}}$, however, the impedance of the power delivery network is dominated by the inductive reactance $j\omega L_{\text{tot}}$ and increases linearly with frequency, exceeding the target impedance at the frequency $f_{\text{max}} = \frac{1}{2\pi}\frac{Z_{\text{target}}}{L_{\text{tot}}}$.

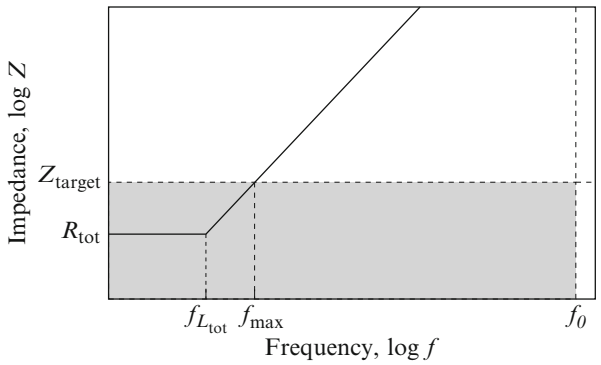


**Fig. 11.8** Impedance of a power distribution system without decoupling capacitors. The *shaded area* denotes the target impedance specifications of the overall power distribution system

The high frequency impedance should be reduced to satisfy the target specifications. Opportunities for reducing the inductance of the power and ground paths of a power delivery network are limited [25, 215–218]. The inductance of the power distribution system is mainly determined by the board and package interconnects [219–221]. The feature size of the board and package level interconnect depends upon the manufacturing technology. The output impedance of a power distribution system is therefore highly inductive and is difficult to lower [134].

The high frequency impedance is effectively reduced by placing capacitors across the power and ground interconnections. These shunting capacitors effectively terminate the high frequency current loop, permitting the current to bypass the inductive interconnect, such as the board and package power delivery networks [222–225]. The high frequency impedance of the system as seen from the current load terminals is thereby reduced. Alternatively, at high frequencies, the capacitors decouple the high impedance paths of the power delivery network from the load. These capacitors are therefore referred to as decoupling capacitors [226, 227]. Several stages of decoupling capacitors are typically utilized to maintain the output impedance of a power distribution system below a target impedance [136, 228], as described in Sect. 11.2.3.

### *11.2.2 Antiresonance*

Decoupling capacitors are a powerful technique to reduce the impedance of a power distribution system over a significant range of frequencies. A decoupling capacitor, however, reduces the resonant frequency of a power delivery network, making the system susceptible to resonances. Unlike the classic self-resonance in a series circuit formed by a decoupling capacitor combined with a parasitic resistance and inductance [138, 229] or by an on-chip decoupling capacitor and

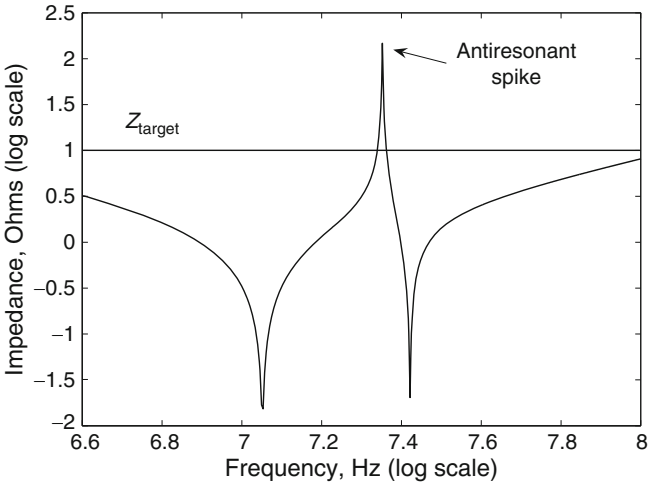


**Fig. 11.9** Antiresonance of the output impedance of a power distribution network. Antiresonance results in a distinctive peak, exceeding the target impedance specification

the parasitic inductance of the package (i.e., chip-package resonance) [230, 231], antiresonance occurs in a circuit formed by two capacitors connected in parallel. At the resonant frequency, the impedance of the series circuit decreases in the vicinity of the resonant frequency, reaching the absolute minimum at the resonant frequency determined by the ESR of the decoupling capacitor. At antiresonance, however, the circuit impedance drastically increases, producing a distinctive peak, as illustrated in Fig. 11.9. This antiresonant peak can result in system failures as the impedance of the power distribution system becomes greater than the maximum tolerable impedance $Z_{\text{target}}$. The antiresonance phenomenon in a system with parallel decoupling capacitors is the subject of this section.

To achieve a low impedance power distribution system, multiple decoupling capacitors are placed in parallel. The effective impedance of a power distribution system with several identical capacitors placed in parallel is illustrated in Fig. 11.10. Observe that the impedance of the power delivery network is reduced by a factor of 2 as the number of capacitors is doubled. Also note that the effective drop in the impedance of a power distribution system diminishes rapidly with each additional decoupling capacitor. It is therefore desirable to utilize decoupling capacitors with a sufficiently low ESR in order to minimize the number of capacitors required to satisfy a target impedance specification [136].

A number of decoupling capacitors with different magnitudes is typically used to maintain the impedance of a power delivery system below a target specification over a wide frequency range. Capacitors with different magnitudes connected in parallel, however, result in a sharp antiresonant peak in the system impedance [29]. The antiresonance phenomenon for different capacitive values is illustrated in Fig. 11.11. The antiresonance of parallel decoupling capacitors can be explained as

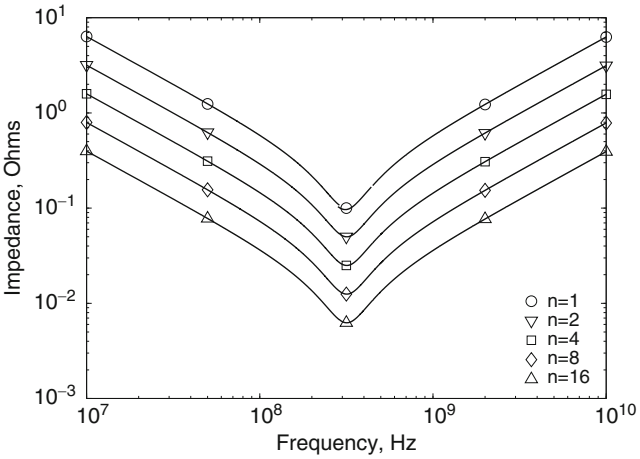


**Fig. 11.10** Impedance of a power distribution system with $n$ identical decoupling capacitors connected in parallel. The ESR of each decoupling capacitor is $R = 0.1\,\Omega$, the ESL is $L = 100\,\text{pH}$, and the capacitance is $C = 1\,\text{nF}$. The impedance of a power distribution system is reduced by a factor of 2 as the number of capacitors is doubled

follows. In the frequency range from $f_1$ to $f_2$, the impedance of the capacitor $C_1$ has become inductive whereas the impedance of the capacitor $C_2$ remains capacitive (see Fig. 11.11). Thus, an $LC$ tank is formed in the frequency range from $f_1$ to $f_2$, producing a peak at the resonant frequency located between $f_1$ and $f_2$. As a result, the total impedance drastically increases and becomes greater than the target impedance, causing a system to fail.

The magnitude of the antiresonant spike can be effectively reduced by lowering the parasitic inductance of the decoupling capacitors. For instance, as discussed in [136], the magnitude of the antiresonant spike is significantly reduced if board decoupling capacitors are mounted on low inductance pads. The magnitude of the antiresonant spike is also determined by the ESR of the decoupling capacitor, decreasing with larger parasitic resistance. Large antiresonant spikes are produced when low ESR decoupling capacitors are placed on inductive pads. A high inductance and low resistance result in a parallel $LC$ circuit with a high quality factor $Q$,

$$Q = \frac{L}{R}. \tag{11.11}$$

In this case, the magnitude of the antiresonant spike is amplified by $Q$. Decoupling capacitors with a low ESR should therefore always be used on low inductance pads (with a low ESL).

Antiresonance also becomes well pronounced if a large variation exists between the capacitance values. This phenomenon is illustrated in Fig. 11.12. In the case of two capacitors with distinctive nominal values ($C_1 \gg C_2$), a significant gap between

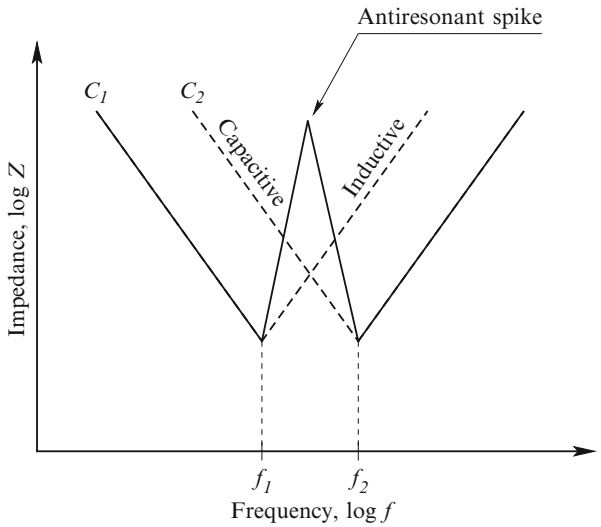


**Fig. 11.11** Antiresonance of parallel capacitors, $C_1 > C_2$, $L_1 = L_2$, and $R_1 = R_2$. A parallel $LC$ tank is formed in the frequency range from $f_1$ to $f_2$. The total impedance drastically increases in the frequency range from $f_1$ to $f_2$ (*the solid line*), producing an antiresonant spike

two capacitances results in a sharp antiresonant spike with a large magnitude in the frequency range from $f_1$ to $f_2$, violating the target specification $Z_{\text{target}}$, as shown in Fig. 11.12a. If another capacitor with nominal value $C_1 > C_3 > C_2$ is added, the antiresonant spike is canceled by $C_3$ in the frequency range from $f_1$ to $f_2$. As a result, the overall impedance of a power distribution system is maintained below the target specification over a broader frequency range, as shown in Fig. 11.12b. As described in [232], the high frequency impedance of two parallel decoupling capacitors is only reduced by a factor of 2 (or 6 dB) as compared to a single capacitor. It is also shown that adding a smaller capacitor in parallel with a large capacitor results in only a small reduction in the high frequency impedance. Antiresonances are effectively managed by utilizing decoupling capacitors with a low ESL and by placing a greater number of decoupling capacitors with progressively decreasing magnitude, shifting the antiresonant spike to the higher frequencies (out of the range of the operating frequencies of the circuit) [233].

### *11.2.3 Hydraulic Analogy of Hierarchical Placement of Decoupling Capacitors*

As discussed in Sect. 11.1.2, an ideal decoupling capacitor should provide a high capacity and be able to release and accumulate energy at a sufficiently high rate. Constructing a device with both high energy capacity and high power capability

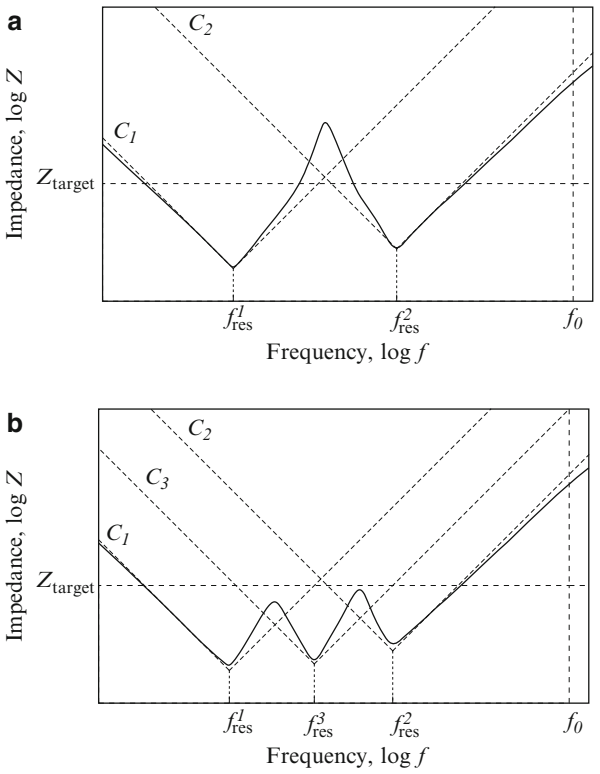


**Fig. 11.12** Antiresonance of parallel capacitors. (**a**) A large gap between two capacitances results in a sharp antiresonant spike with a large magnitude in the frequency range from $f_1$ to $f_2$, violating the target specification $Z_{\text{target}}$. (**b**) If another capacitor with magnitude $C_1 > C_3 > C_2$ is added, the antiresonant spike is canceled by $C_3$ in the frequency range from $f_1$ to $f_2$. As a result, the overall impedance of the power distribution system is maintained below the target specification over the desired frequency range

is, however, challenging. It is expensive to satisfy both of these requirements in an ideal decoupling capacitor. Moreover, these requirements are typically contradictory in most practical applications. The physical realization of a large decoupling capacitance requires the use of discrete capacitors with a large nominal capacity and, consequently, a large form factor. The large physical dimensions of the capacitors have two implications. The parasitic series inductance of a physically large capacitor is relatively high due to the increased area of the current loop within the capacitors. Furthermore, due to technology limitations, the large physical size of the capacitors prevents placing the capacitors sufficiently close to the current load. A greater physical separation increases the inductance of the current path from the capacitors to the load. A tradeoff therefore exists between the high capacity and low parasitic inductance of a decoupling capacitor for an available component technology.

Gate switching times of a few tens of picoseconds are common in modern high performance ICs, creating high transient currents in the power distribution system. At high frequencies, only those on-chip decoupling capacitors with a low ESR and a low ESL can effectively maintain a low impedance power distribution system. Placing a sufficiently large on-chip decoupling capacitor requires a die area many times greater than the area of a typical circuit. Thus, while technically feasible, a single-tier decoupling solution is prohibitively expensive. A large on-chip decoupling capacitor is therefore typically built as a series of small decoupling capacitors connected in parallel. At high frequencies, a large on-chip decoupling capacitor exhibits a distributed behavior. Only on-chip decoupling capacitors located in the vicinity of the switching circuit can effectively provide the required charge to the current load within the proper time. An efficient approach to this problem is to hierarchically place multiple stages of decoupling capacitors, progressively smaller and closer to the load.

Utilizing hierarchically placed decoupling capacitors produces a low impedance, high frequency power distribution system realized in a cost effective way. The capacitors are placed in several stages: on the board, package, and circuit die. Arranging the decoupling capacitors in several stages eliminates the need to satisfy both the high capacitance and low inductance requirements in the same decoupling stage [30].

The hydraulic analogy of the hierarchical placement of decoupling capacitors is shown in Fig. 11.13. Each decoupling capacitor is represented by a water tank. All of the water tanks are connected to the main water pipe connected to the consumer (current load). Water tanks at different stages are connected to the main pipe through the local water pipes, modeling different interconnect levels. The goal of the water supply system (power delivery network) is to provide uninterrupted water flow to the consumer at the required rate (switching time). The amount of water released by each water tank is proportional to the tank size. The rate at which the water tank is capable of providing water is inversely proportional to the size of the water tank and directly proportional to the distance from the consumer to the water tank.

A power supply is typically treated as an infinite amount of charge. Due to large physical dimensions, the power supply cannot be placed close to the current load (the consumer). The power supply therefore has a long response time. Unlike the power supply, an on-chip decoupling capacitor can be placed sufficiently close to the consumer. The response time of an on-chip decoupling capacitor is significantly shorter as compared to the power supply. An on-chip decoupling capacitor is therefore able to respond to the consumer demand in a much shorter period of time but is capable of providing only a small amount of water (or charge). Allocating decoupling capacitors with progressively decreasing magnitudes and closer to the current load, an uninterrupted flow of charge can be provided to the consumer. In the initial moment, charge is only supplied to the consumer by the on-chip decoupling capacitor. As the on-chip decoupling capacitor is depleted, the package decoupling capacitor is engaged. This process continues until the power supply is activated. Finally, the power supply is turned on and provides the necessary charge with relatively relaxed timing constraints. The voltage regulator, board, package, and

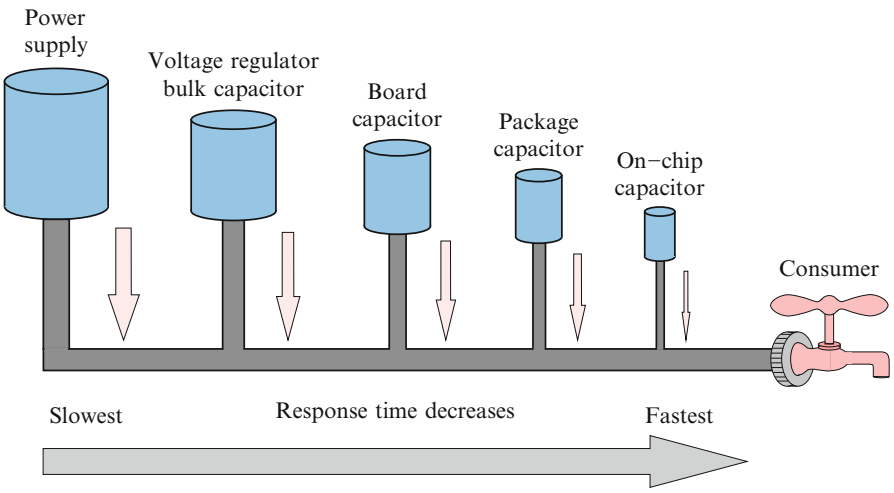


**Fig. 11.13** Hydraulic analogy of the hierarchical placement of decoupling capacitors. The decoupling capacitors are represented by the water tanks. The response time is proportional to the size of the capacitor and inversely proportional to the distance from a capacitor to the consumer. The on-chip decoupling capacitor has the shortest response time (located closest to the consumer), but is capable of providing the least amount of charge

on-chip decoupling capacitors therefore serve as intermediate reservoirs of charge, relaxing the timing constraints for the power delivery supply.

A hierarchy of decoupling capacitors is utilized in high performance power distribution systems in order to extend the frequency region of the low impedance characteristics to the maximum operating frequency $f_0$. The impedance characteristics of a power distribution system with board, package, and on-chip decoupling capacitors (see Fig. 11.6) are illustrated in Fig. 11.14. By utilizing the hierarchical placement of decoupling capacitors, the antiresonant spike is shifted outside the range of operating frequencies (beyond $f_0$). The overall impedance of a power distribution system is also maintained below the target impedance over the entire frequency range of interest (from DC to $f_0$).

### Fully Compensated System

A special case in the impedance of an *RLC* circuit formed by a decoupling capacitor and the parasitic inductance of the P/G lines is achieved when the zeros of a tank circuit impedance cancel the poles, making the impedance purely resistive and independent of frequency,

$$R_L = R_C = R_0 = \sqrt{\frac{L}{C}}, \tag{11.12}$$

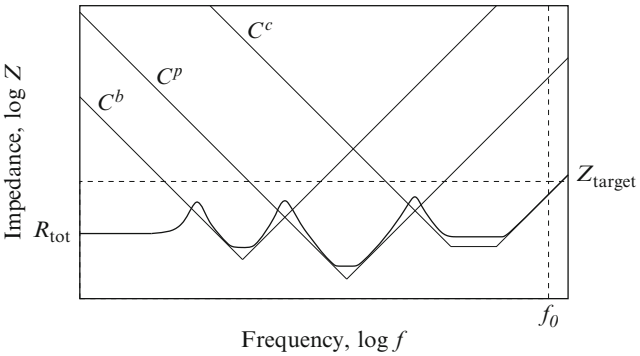


**Fig. 11.14** Impedance of a power distribution system with board, package, and on-chip decoupling capacitances. The overall impedance is shown with a *black line*. The impedance of a power distribution system with three levels of decoupling capacitors is maintained below the target impedance (*dashed line*) over the frequency range of interest. The impedance characteristics of the decoupling capacitors are shown by the *thin solid lines*

$$\frac{L}{R_L} = R_C C, \tag{11.13}$$

where $R_L$ and $R_C$ are, respectively, the parasitic resistance of the P/G lines and the ESR of the decoupling capacitor. In this case, the impedance of the *RLC* tank is fully compensated. Equations (11.12) and (11.13) are equivalent to two conditions, i.e., the impedance at the lower frequencies is matched to the impedance at the high frequencies and the time constants of the inductor and capacitor currents are also matched. A constant, purely resistive impedance, characterizing a power distribution system with decoupling capacitors, is achieved across the entire frequency range of interest, if each decoupling stage is fully compensated [137, 234]. The resistance and capacitance of the decoupling capacitors in a fully compensated system are completely determined by the impedance characteristics of the power and ground interconnect and the location of the decoupling capacitors.

The hierarchical placement of decoupling capacitors exploits the tradeoff between the capacity and the parasitic inductance of a capacitor to achieve an economically effective solution. The total decoupling capacitance of a hierarchical scheme $C_{\text{total}} = C^b + C^p + C^c$ is larger than the total decoupling capacitance of a single-tier solution, where $C^b$, $C^p$, and $C^c$ are, respectively, the board, package, and on-chip decoupling capacitances. The primary advantage of utilizing a hierarchical placement is that the inductive limit is imposed only on the final stage of decoupling capacitors which constitutes a small fraction of the total required decoupling capacitance. The constraints on the physical dimensions and parasitic impedance of the capacitors in the remaining stages are therefore significantly reduced. As a result, cost efficient electrolytic and ceramic capacitors can be used to provide medium size and high capacity decoupling capacitors [30].

## 11.3 Intrinsic vs Intentional On-Chip Decoupling Capacitance

Several types of on-chip capacitances contribute to the overall on-chip decoupling capacitance. The *intrinsic* decoupling capacitance is the inherent capacitance of the transistors and interconnects that exists between the power and ground terminals. The thin gate oxide capacitors placed on-chip to solely provide power decoupling are henceforth referred to as an *intentional* decoupling capacitance. The intrinsic decoupling capacitance is described in Sect. 11.3.1. The intentional decoupling capacitance is reviewed in Sect. 11.3.2.

### *11.3.1 Intrinsic Decoupling Capacitance*

An intrinsic decoupling capacitance (or symbiotic capacitance) is the parasitic capacitance between the power and ground terminals within an on-chip circuit structure. The intrinsic capacitance is comprised of three types of parasitic capacitances [235].

One component of the intrinsic capacitance is the parasitic capacitance of the interconnect lines. Three types of intrinsic interconnect capacitances are illustrated in Fig. 11.15. The first type of interconnect capacitance is the capacitance $C_1^i$ between the signal line and the power/ground line. Capacitance $C_2^i$ is the capacitance between signal lines at different voltage potentials. The third type of intrinsic interconnect capacitance is the capacitance $C_3^i$ between the power and ground lines (see Fig. 11.15).

Parasitic device capacitances, such as the drain junction capacitance and gate-to-source capacitance, also contribute to the overall intrinsic decoupling capacitance where the terminals of the capacitance are connected to power and ground. For example, in the simple inverter circuit depicted in Fig. 11.16, if the input is one (high) and the output is zero (low), the NMOS transistor is turned on, connecting $C_p$ from $V_{dd}$ to $G_{nd}$, providing a decoupling capacitance to the other switching circuits,

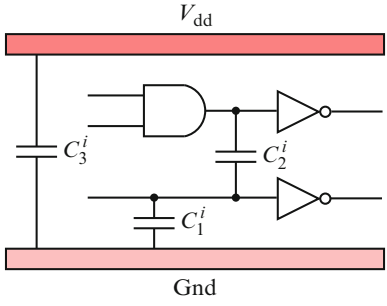


**Fig. 11.15** Intrinsic decoupling capacitance of the *interconnect lines*. $C_1^i$ denotes the capacitance between the signal line and the power/ground line. $C_2^i$ denotes the capacitance between signal lines. $C_3^i$ denotes the capacitance between the power and ground lines

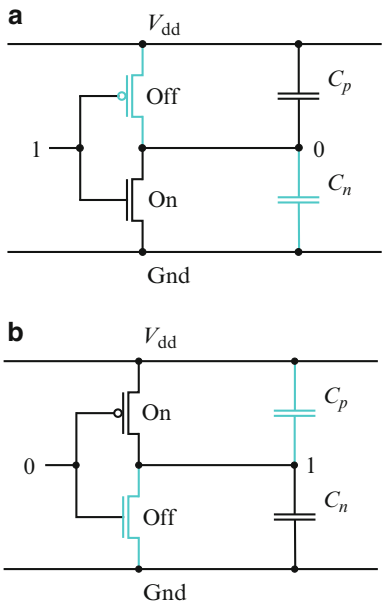


**Fig. 11.16** Intrinsic decoupling capacitance of a non-switching circuit; (**a**) inverter input is high, (**b**) inverter input is low

as illustrated in Fig. 11.16a. Alternatively, if the input is zero (low) and the output is one (high), the PMOS transistor is turned on, connecting $C_n$ from $G_{nd}$ to $V_{dd}$, providing a decoupling capacitance to the other switching circuits, as illustrated in Fig. 11.16b.

Depending upon the total capacitance $(C_p + C_n)$ and the switching factor $SF$, the decoupling capacitance from the non-switching circuits is [236]

$$C_{\text{circuit}} = \frac{P}{V_{dd}^2 f} \frac{(1 - SF)}{SF}, \tag{11.14}$$

where $P$ is the circuit power, $V_{dd}$ is the power supply voltage, and $f$ is the switching frequency. The time constant for $C_{\text{circuit}}$ is determined by $R_{\text{PMOS}} C_n$ or $R_{\text{NMOS}} C_p$ and usually varies in a 0.18 μm CMOS technology from about 50–250 ps [236].

The contribution of the transistor and interconnect capacitance to the overall decoupling capacitance is difficult to determine precisely. The transistor terminals as well as the signal lines can be connected either to power or ground, depending upon the internal state of the digital circuit at a particular time. The transistor and interconnect decoupling capacitance therefore depends on the input pattern and the internal state of the circuit. The input vectors that produce the maximum intrinsic decoupling capacitance in a digital circuit are described in [237].

Another source of intrinsic capacitance is the *p-n* junction capacitance of the diffusion wells. The N-type wells, P-type wells, or wells of both types are implanted into a silicon substrate to provide an appropriate body doping for the PMOS and NMOS transistors. The N-type wells are ohmically connected to the power supply

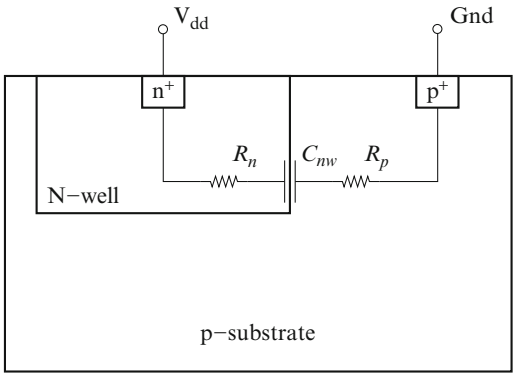


**Fig. 11.17** N-well junction intrinsic decoupling capacitance. The capacitor $C_{nw}$ is formed by the reverse-biased *p-n* junction between the N-well and the p-substrate

while the P-type wells are connected to ground to provide a proper body bias for the transistors. The N-well capacitor is the reverse-biased *p-n* junction capacitor between the N-well and p-substrate, as shown in Fig. 11.17. The total on-chip N-well decoupling capacitance $C_{nw}$ is determined by the area, perimeter, and depth of each N-well. Multiplying $C_{nw}$ by the series and contact resistance in the N-well and p-substrate, the time constant $(R_p + R_n + R_{\text{contact}})C_{nw}$ for an N-well capacitor is typically in the range of 250–500 ps in a 0.18 µm CMOS technology [236]. The parasitic capacitance of the wells usually dominates the intrinsic decoupling capacitance of ICs fabricated in an epitaxial CMOS process [177, 238]. The overall intrinsic on-chip decoupling capacitance consists of several components and is

$$C_{\text{intrinsic}} = C_{\text{inter}} + C_{pn} + C_{\text{well}} + C_{\text{load}} + C_{gs} + C_{gb}, \tag{11.15}$$

where $C_{\text{inter}}$ is the interconnect capacitance, $C_{pn}$ is the *p-n* junction capacitance, $C_{\text{well}}$ is the capacitance of the well, $C_{\text{load}}$ is the load capacitance, $C_{gs}$ is the gate-to-source (drain) capacitance, and $C_{gb}$ is the gate-to-body capacitance.

Silicon-on-insulator (SOI) CMOS circuits lack diffusion wells and therefore do not contribute to the intrinsic on-chip decoupling capacitance. A reliable estimate of the contribution of the interconnect and transistors to the on-chip decoupling capacitance is thus particularly important in SOI circuits. Several techniques for estimating the intrinsic decoupling capacitance are presented in [140, 239]. The overall intrinsic decoupling capacitance of an IC can also be determined experimentally. In [240], the signal response of a power distribution system versus frequency is measured with a vector network analyzer. An *RLC* model of the system is constructed to match the observed response. The magnitude of the total on-chip decoupling capacitance is determined from the frequency of the resonant peaks in the response of the power system. Alternatively, the total on-chip decoupling capacitance can be experimentally determined from the package-chip resonance, as described in [231]. The intentional decoupling capacitance placed on-chip during the design process is known within the margins of the process variations. Subtracting the intentional capacitance from the measured overall capacitance yields an estimate of the on-chip intrinsic capacitance.

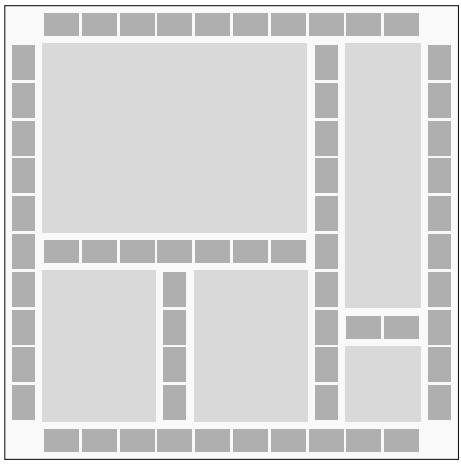

**Fig. 11.18** Banks of on-chip decoupling capacitors (*the dark gray rectangles*) placed among circuit blocks (*the light gray rectangles*)

### *11.3.2 Intentional Decoupling Capacitance*

Intentional decoupling capacitance is often added to a circuit during the design process to increase the overall on-chip decoupling capacitance to a satisfactory level. The intentional decoupling capacitance is typically realized as a gate capacitance in large MOS transistors placed on-chip specifically for this purpose. In systems with mixed memory and logic, however, the intentional capacitance can also be realized as a trench capacitance [241, 242].

Banks of MOS decoupling capacitors are typically placed among the on-chip circuit blocks, as shown in Fig. 11.18. The space between the circuit blocks is often referred to as "white" space, as this area is primarily used for global routing and does not contain any active devices. Unless noted otherwise, the term "on-chip decoupling capacitance" commonly refers to the intentional decoupling capacitance. Using more than 20 % of the overall die area for intentional on-chip decoupling capacitance is common in modern high speed integrated circuits [149, 243].

An MOS capacitor uses the thin oxide layer between the N-well and polysilicon gate to provide the additional decoupling capacitance needed to mitigate the power noise, as shown in Fig. 11.19. An optional fuse (or control gate) is typically provided to disconnect the thin oxide capacitor from the rest of the circuit in the undesirable situation of a short circuit due to process defects. As the size and shape of MOS capacitors vary, the $R_n C_{ox}$ time constant typically ranges from 40 to 200 ps in a 0.18 μm CMOS process. Depending upon the switching speed of the circuit, typical on-chip MOS decoupling capacitors are effective for $RC$ time constants below 200 ps [236].

An MOS capacitor is formed by the gate electrode on one side of the oxide layer and the source-drain inversion channel under the gate on the other side of the oxide layer. The resistance of the channel dominates the ESR of the MOS capacitor. Due to the resistance of the transistor channel, the MOS capacitor is modeled as

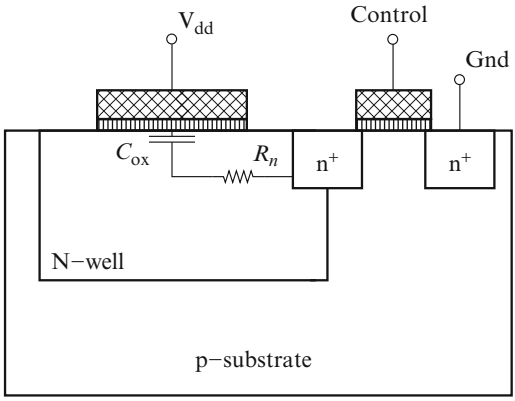


**Fig. 11.19** Thin oxide MOS decoupling capacitor

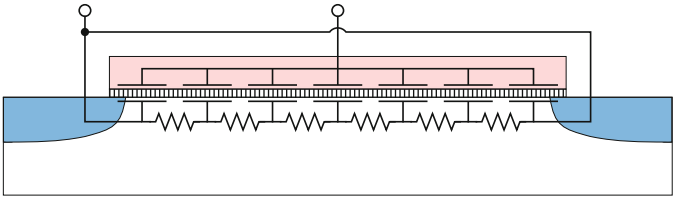

**Fig. 11.20** Equivalent *RC* model of an MOS decoupling capacitor

a distributed *RC* circuit, as shown in Fig. 11.20. The impedance of the distributed *RC* structure shown in Fig. 11.20 is frequency dependent, $Z(\omega) = R(\omega) + \frac{1}{j\omega C(\omega)}$. Both the resistance $R(\omega)$ and capacitance $C(\omega)$ decrease with frequency. The low frequency resistance of the MOS capacitor is approximately one twelfth of the source-drain resistance of the MOS transistor in the linear region [244]. The low frequency capacitance is the entire gate-to-channel capacitance of the transistor. At high frequencies, the gate-to-channel capacitance midway between the drain and source is shielded from the capacitor terminals by the resistance of the channel, decreasing the effective capacitance of the MOS capacitor. The higher the channel resistance per transistor width, the lower the frequency at which the capacitor efficiency begins to decrease. Capacitors with a long channel (with a relatively high channel resistance) are therefore less effective at high frequencies as compared to short-channel capacitors. A higher series resistance of the on-chip MOS decoupling capacitor, however, is beneficial in damping the resonance of a die-package *RLC* tank circuit [244].

Long channel transistors, however, are more area efficient. In transistors with a minimum length channel, the source and body contacts dominate the transistor area, while the MOS capacitor stack occupies a relatively small fraction of the total area. For longer channels, the area of the MOS capacitor increases while the area overhead of the source/drain contacts remain constant, increasing the capacitance per total area [30]. A tradeoff therefore exists between the area efficiency and the ESR of the MOS decoupling capacitor. Transistors with a channel length 12 times

greater than the minimum length are a good compromise [244]. In this case, the *RC* time constant is smaller than the switching time of the logic gates, which typically are composed of transistors with a minimum channel length, while the source and drain contacts occupy a relatively small fraction of the total area.

## 11.4 Types of On-Chip Decoupling Capacitors

Multiple on-chip capacitors are utilized in ICs to satisfy various design requirements. Four types of widely utilized on-chip decoupling capacitors are the subject of this section. Polysilicon-insulator-polysilicon (PIP) capacitors are presented in Sect. 11.4.1. Three types of MOS decoupling capacitors, accumulation, depletion, and inversion, are described in Sect. 11.4.2. Metal-insulator-metal (MIM) decoupling capacitors are reviewed in Sect. 11.4.3. In Sect. 11.4.4, lateral flux decoupling capacitors are described. The design and performance characteristics of the different on-chip decoupling capacitors are compared in Sect. 11.4.5.

### *11.4.1 Polysilicon-Insulator-Polysilicon (PIP) Capacitors*

Both junction and MOS capacitors use diffusion for the lower electrodes. The junction isolating the diffused electrode exhibits substantial parasitic capacitance, limiting the voltage applied across the capacitor. These limitations are circumvented in PIP capacitors, which employ two polysilicon electrodes in combination with either an oxide or an oxide-nitride-oxide (ONO) dielectric [245], as illustrated in Fig. 11.21. Since typical CMOS and BiCMOS processes incorporate multiple polysilicon layers, PIP capacitors do not require any additional masking steps. The gate polysilicon can serve as the lower electrode of the PIP capacitor, while the resistor polysilicon (doped with a suitable implant) can form the upper electrode. The upper electrode is typically doped with either an N-type source/drain (NSD) or P-type source/drain (PSD) implant. The implant resulting in the lowest sheet resistance is preferable, since heavier doping reduces the ESR and minimizes voltage modulation due to polysilicon depletion [245].

PIP capacitors require additional process steps. Even if both of the electrodes consist of existing depositions, the capacitor dielectric is unique to this structure and consequently requires a process extension. The simplest way to form this dielectric is to eliminate the interlevel oxide (ILO) deposition that normally separates the two polysilicon layers and add a thin oxide layer on the lower polysilicon electrode. With this technique, a capacitor can be built between the two polysilicon layers as long as the second polysilicon layer is not used as an interconnection.

Silicon dioxide has a relatively low permittivity. A higher permittivity, and therefore a higher capacitance per unit area, is achieved using a stacked ONO dielectric (see Fig. 11.21b). Observe from Fig. 11.21 that the PIP capacitors normally

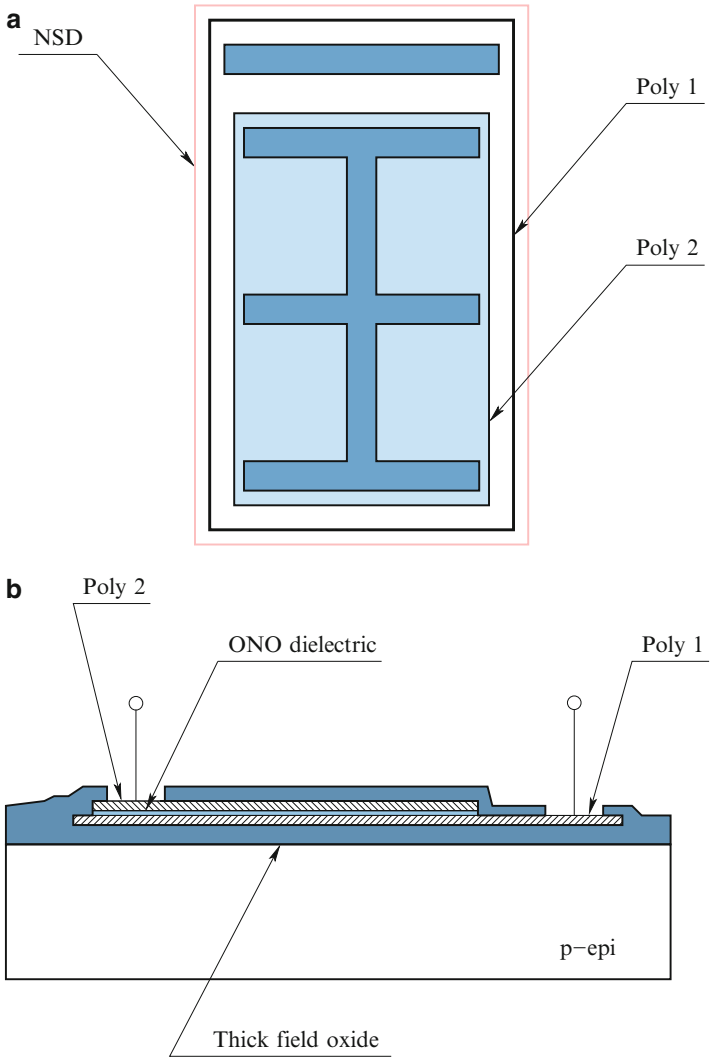


**Fig. 11.21** PIP oxide-nitride-oxide (ONO) capacitor. The entire capacitor is enclosed in an N-type source/drain region, reducing the sheet resistance of the polysilicon layer; (**a**) layout, (**b**) cross section

reside over the field oxide. The oxide steps should not intersect the structure, since those steps cause surface irregularities in the lower capacitor electrode, resulting in localized thinning of the dielectric, thereby concentrating the electric field. As a result of the intersection, the breakdown voltage of the capacitor can be severely compromised.

Selecting the dielectric material in a PIP capacitor, several additional issues should be considered. Composite dielectrics experience hysteresis effects at high

frequencies (above 10 MHz) due to the incomplete redistribution of static charge along the oxide-nitride interface. Pure oxide dielectrics are used for PIP capacitors to achieve a relatively constant capacitance over a wide frequency range. Oxide dielectrics, however, typically have a lower capacitance per unit area. Low capacitance dielectrics are also useful for improving matching among the small capacitors.

Voltage modulation of the PIP capacitors is relatively small, as long as both electrodes are heavily doped. A PIP capacitor typically exhibits a voltage modulation of 150 ppm/V [245]. The temperature coefficient of a PIP capacitor also depends on voltage modulation effects and is typically less than 250 ppm/°C [246].

### *11.4.2 MOS Capacitors*

An MOS capacitor consists of a metal-oxide-semiconductor structure, as illustrated in Fig. 11.22. A top metal contact is referred to as the gate, serving as one plate of the capacitor. In digital CMOS ICs, the gate is often fabricated as a heavily doped $n^+$-polysilicon layer, behaving as a metal. A second metal layer forms an ohmic contact to the back of the semiconductor and is called the bulk contact. The semiconductor layer serves as the other plate of the capacitor. The bulk resistivity is typically 1–10 $\Omega \cdot$cm (with a doping of $10^{15}$ $cm^{-3}$).

The capacitance of an MOS capacitor depends upon the voltage applied to the gate with respect to the body. The dependence of the capacitance upon the voltage across an MOS capacitor (a capacitance versus voltage (CV) diagram) is plotted in Fig. 11.23. Depending upon the gate-to-body potential $V_{gb}$, three regions of operation are distinguished in the CV diagram of an MOS capacitor. In the accumulation mode, mobile carriers of the same type as the body (holes for an NMOS capacitor with a p-substrate) accumulate at the surface. In the depletion mode, the surface is devoid of any mobile carriers, leaving only a space charge (depletion layer). In the inversion mode, mobile carriers of the opposite type of the body (electrons for an NMOS capacitor with a p-substrate) aggregate at the surface, inverting the conductivity type. These three regimes are roughly separated by the

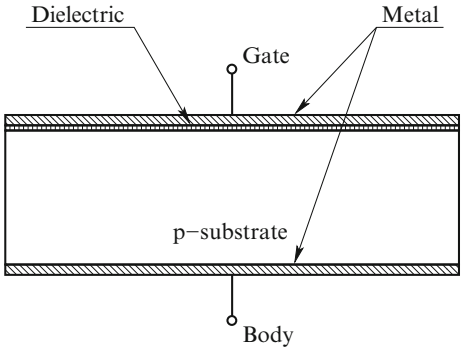


**Fig. 11.22** The structure of an n-type MOS capacitor

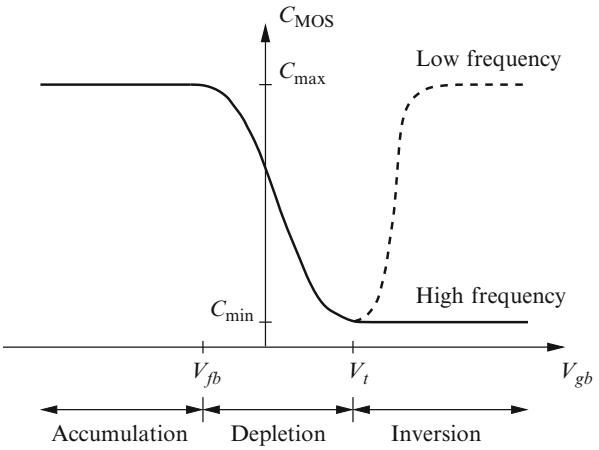


**Fig. 11.23** Capacitance versus gate voltage (CV) diagram of an n-type MOS capacitor. The flat band voltage $V_{fb}$ separates the accumulation region from the depletion region. The threshold voltage $V_t$ separates the depletion region from the inversion region

two voltages (see Fig. 11.23). A flat band voltage $V_{fb}$ separates the accumulation regime from the depletion regime. The threshold voltage $V_t$ demarcates the depletion regime from the inversion regime. Based on the mode of operation, three types of MOS decoupling capacitors exist and are described in the following three subsections.

### Accumulation

In MOS capacitors operating in accumulation, the applied gate voltage is lower than the flat band voltage ($V_{gb} < V_{fb}$) and induces negative charge on the metal gate and positive charge in the semiconductor. The hole concentration at the surface is therefore above the bulk value, leading to surface accumulation. The charge distribution in an MOS capacitor operating in accumulation is shown in Fig. 11.24. The flat band voltage is the voltage at which there is no charge on the plates of the capacitor (there is no electric field across the dielectric). The flat band voltage depends upon the doping of the semiconductor and any residual charge existing at the interface between the semiconductor and the insulator. In the accumulation mode, the charge per unit area $Q_n$ at the semiconductor/oxide interface is a linear function of the applied voltage $V_{gb}$. The oxide capacitance per unit area $C_{ox}$ is determined by the slope of $Q_n$, as illustrated in Fig. 11.25. The capacitance of an MOS capacitor operating in accumulation achieves the maximum value and is

$$C_{MOS_{accum}} = C_{max} = A\,C_{ox} = A\,\frac{\epsilon_{ox}}{t_{ox}}, \tag{11.16}$$

where $A$ is the area of the gate electrode, $\epsilon_{ox}$ is the permittivity of the oxide, and $t_{ox}$ is the oxide thickness.

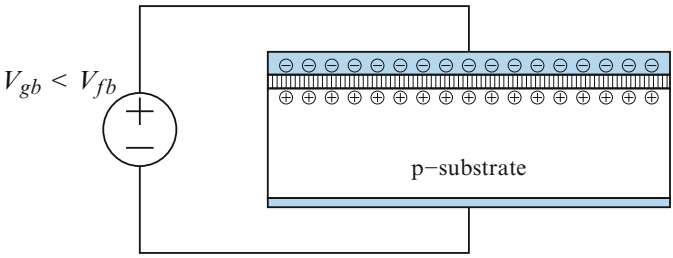


**Fig. 11.24** Charge distribution in an NMOS capacitor operating in accumulation ($V_{gb} < V_{fb}$)

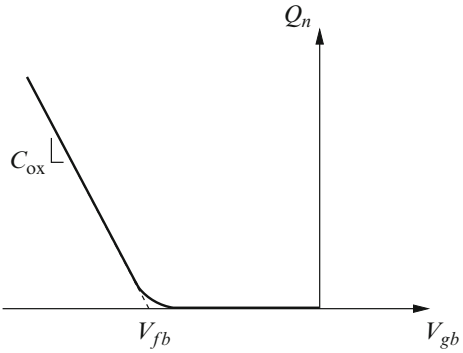


**Fig. 11.25** Accumulation charge density as a function of the applied gate voltage. The capacitance per unit area $C_{ox}$ is determined by the slope of the line

## Depletion

In MOS capacitors operating in depletion, the applied gate voltage is brought above the flat band voltage and below the threshold voltage ($V_{fb} < V_{gb} < V_t$). A positive charge is therefore induced at the interface between the metal gate and the oxide. A negative charge is induced at the oxide/semiconductor interface. This scenario is accomplished by pushing all of the mobile positive carriers (holes) away, exposing the fixed negative charge from the donors. Hence, the surface of the semiconductor is depleted of mobile carriers, leaving behind a negative space charge. The charge distribution in the MOS capacitor operating in depletion is illustrated in Fig. 11.26.

The resulting space charge behaves like a capacitor with an effective capacitance per unit area $C_d$. The effective capacitance $C_d$ depends upon the gate voltage $V_{gb}$ and is

$$C_d(V_{gb}) = \frac{\epsilon_{\text{Si}}}{x_d(V_{gb})}, \tag{11.17}$$

where $\epsilon_{\text{Si}}$ is the permittivity of the silicon and $x_d$ is the thickness of the depletion layer (space charge). Observe from Fig. 11.26 that the oxide capacitance per unit area $C_{ox}$ and depletion capacitance per unit area $C_d$ are connected in series. The capacitance of a MOS structure in the depletion region is therefore

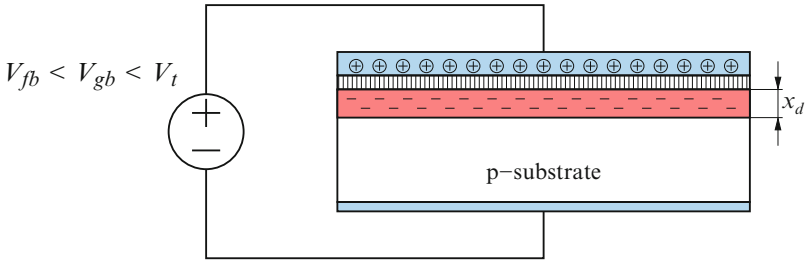


**Fig. 11.26** Charge distribution in an NMOS capacitor operating in depletion ($V_{fb} < V_{gb} < V_t$). Under this bias condition, all of the mobile positive carriers (holes) are pushed away, depleting the surface of the semiconductor, resulting in a negative space charge with thickness $x_d$

$$C_{\text{MOS}_{\text{deplet}}} = A \frac{C_{\text{ox}} C_d}{C_{\text{ox}} + C_d}. \tag{11.18}$$

Note that the thickness of the silicon depletion layer becomes wider as the gate voltage is increased, since more holes are pushed away, exposing more fixed negative ionized dopants, leading to a thicker space charge layer. As a result, the capacitance of the depleted silicon decreases, reducing the overall MOS capacitance.

### Inversion

In MOS capacitors operating in inversion, the applied gate voltage is further increased above the threshold voltage ($V_t < V_{gb}$). The conduction type of the semiconductor surface is inverted (from p-type to n-type). The threshold voltage is referred to as the voltage at which the conductivity type of the surface layer changes from p-type to n-type (in the case of an NMOS capacitor). This phenomenon is explained as follows. As the gate voltage is increased beyond the threshold voltage, holes are pushed away from the $Si/SiO_2$ interface, exposing the negative charge. Note that the density of holes decreases exponentially from the surface into the bulk. The number of holes decreases as the applied voltage increases. The number of electrons at the surface therefore increases with applied gate voltage and becomes the dominant type of carrier, inverting the surface conductivity. The charge distribution of an MOS capacitor operating in inversion is depicted in Fig. 11.27.

Note that the depletion layer thickness reaches a maximum in the inversion region. The total voltage drop across the semiconductor also reaches the maximum value. Further increasing the gate voltage, the applied voltage drops primarily across the oxide layer. If the gate voltage approaches the threshold voltage, the depleted layer capacitance per unit area $C_d^{\text{min}}$ reaches a minimum [247]. In this case, the overall MOS capacitance reaches the minimum value and is

$$C_{\text{MOS}_{\text{inv}}} = C_{\text{MOS}}^{\text{min}} = A \frac{C_{\text{ox}} C_d^{\text{min}}}{C_{\text{ox}} + C_d^{\text{min}}}, \tag{11.19}$$

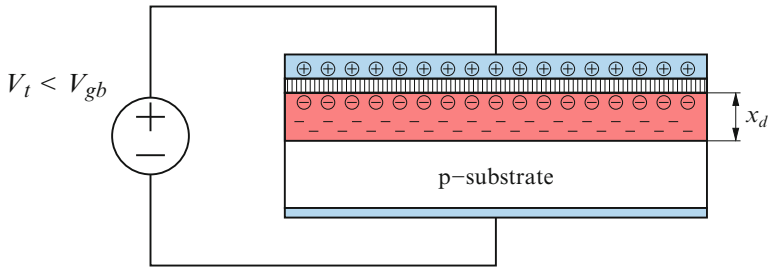


**Fig. 11.27** Charge distribution of an NMOS capacitor operating in inversion ($V_t < V_{gb}$). Under this bias condition, a negative charge is accumulated at the semiconductor surface, inverting the conductivity of the semiconductor surface (from p-type to n-type)

where

$$C_d^{\min} = \frac{\epsilon_{\mathrm{Si}}}{x_d^{\max}}. \tag{11.20}$$

Note that at low frequencies (quasi-static conditions), the generation rate of holes (electrons) in the depleted silicon surface layer is sufficiently high. Electrons are therefore swept to the $Si/SiO_2$ interface, forming a sheet charge with a thin layer of electrons. The inversion layer capacitance under quasi-static conditions therefore reaches the maximum value. At high frequencies, however, the generation rate is not sufficiently high, prohibiting the formation of the electron charge at the $Si/SiO_2$ interface. In this case, the thickness of the silicon depletion layer reaches the maximum. Hence, the inversion layer capacitance reaches the minimum.

An MOS transistor operated as a capacitor has a substantial ESR, most of which is associated with the lower electrode. This parasitic resistance can be reduced by using a fairly short channel length (25 μm or less) [245]. If the source and drain diffusions are omitted, the backgate contact is typically placed entirely around the gate.

A layout and cross section of an MOS capacitor formed in a BiCMOS process are illustrated in Fig. 11.28. Since the N-type source/drain layer follows the gate oxide growth and polysilicon deposition, the lower plate should consist of some other diffusion (typically deep-$n^+$). Deep-$n^+$ has a higher sheet resistance than the N-type source/drain layer (typically 100 Ω/□), resulting in a substantial parasitic resistance of the lower plate. The heavily concentrated n-type doping thickens the gate oxide by 10–30 % through dopant-enhanced oxidation, resulting in higher working voltages but a lower capacitance per unit area. The deep-$n^+$ is often placed inside the N-well to reduce the parasitic capacitance to the substrate. The N-well can be omitted, however, if the larger parasitic capacitance and lower breakdown voltage of the deep-$n^+$/p-epi junction can be tolerated.

Regardless of how an MOS capacitor is constructed, the two capacitor electrodes are never entirely interchangeable. The lower plate always consists of a diffusion with substantial parasitic junction capacitance. This junction capacitance

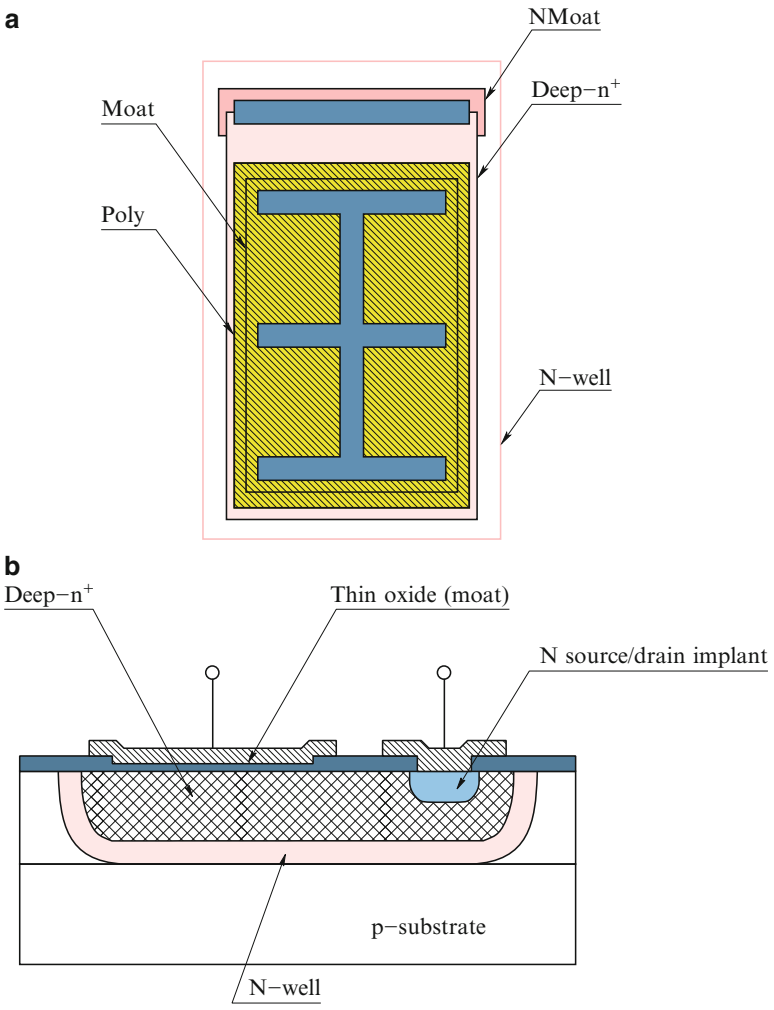


**Fig. 11.28** Deep-$n^+$ MOS capacitor constructed in a BiCMOS process; (**a**) layout, (**b**) cross section

is eliminated by connecting the lower plate of the capacitor to the substrate potential. The upper plate of the MOS capacitor consists of a deposited electrode with a relatively small parasitic capacitance. The lower plate of an MOS capacitor should therefore be connected to the driven node (with the lower impedance). Swapping the two electrodes of an MOS capacitor can load a high impedance node with a high parasitic impedance, compromising circuit performance.

The major benefit of MOS capacitors is the natural compatibility with CMOS technology. MOS capacitors also provide a high capacitance density [248], providing a cost effective on-chip decoupling capacitance. MOS capacitors result in

relatively high matching: the gate oxide capacitance is typically controlled within 5 % error [246]. MOS capacitors, however, are non-linear devices that exhibit strong voltage dependence (more than 100 ppm/V [249]) due to the variation of both the dielectric constant and the depletion region thickness within each plate. The performance of the MOS capacitors is limited at high frequencies due to the large diffusion-to-substrate parasitic capacitance. As technology scales, the leakage currents of MOS capacitors also increase substantially, increasing the total power dissipation. High leakage current is the primary issue with MOS capacitors.

An MOS on-chip capacitance is typically realized as accumulation and inversion capacitors. Note that capacitors operating in accumulation are more linear than capacitors operating in inversion [250]. The MOS capacitance operating in accumulation is almost independent of frequency. Moreover, MOS decoupling capacitors operating in accumulation result in an approximately 15× reduction in leakage current as compared to MOS decoupling capacitors operating in inversion [251]. MOS decoupling capacitors operating in accumulation should therefore be the primary form of MOS decoupling capacitors in modern high performance ICs.

### *11.4.3 Metal-Insulator-Metal (MIM) Capacitors*

A MIM capacitor consists of two metal layers (plates) separated by a deposited dielectric layer. A cross section of a MIM capacitor is shown in Fig. 11.29. A thick oxide layer is typically deposited on the substrate, reducing the parasitic capacitance to the substrate. The parasitic substrate capacitance is also lowered by utilizing the top metal layers as plates of a MIM capacitor. For instance, in comb MIM capacitors [252], the parasitic capacitance to the substrate is less than 2 % of the total capacitance.

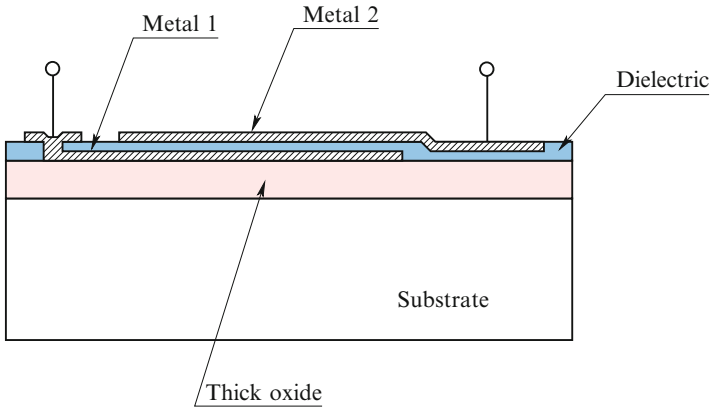


**Fig. 11.29** Cross section of a MIM capacitor. A thick oxide ($SiO_2$) layer is typically deposited on the substrate to reduce the parasitic capacitance to the substrate

Historically, MIM capacitors have been widely used in RF and mixed-signal ICs due to the low leakage, high linearity, low process variations (high accuracy), and low temperature variations [253–255] of MIM capacitors. Conventional circuits utilize $SiO_2$ as a dielectric deposited between two metal layers. Large MIM capacitors therefore require significant circuit area, prohibiting the use of MIM capacitors as decoupling capacitors in high complexity ICs. The capacitance density can be increased by reducing the dielectric thickness and employing high-$k$ dielectrics. Reducing the dielectric thickness, however, results in a substantial increase in leakage current which is highly undesirable.

MIM capacitors with a capacitance density comparable to MOS capacitors (8–10 $fF/\mu m^2$) have been fabricated using $Al_2O_3$ and $AlTiO_x$ dielectrics [256], $AlTaO_x$ [257], and $HfO_2$ dielectric using atomic layer deposition (ALD) [258]. A higher capacitance density (13 $fF/\mu m^2$) is achieved using laminate ALD $HfO_2 - Al_2O_3$ dielectrics [259, 260]. Laminate dielectrics also result in higher voltage linearity and reliability. Recently, MIM capacitors with a capacitance density approximately two times greater than the capacitance density of MOS capacitors have been fabricated [261]. A capacitance density of 17 $fF/\mu m^2$ is achieved using a $Nb_2O_5$ dielectric with $HfO_2 - Al_2O_3$ barriers.

Unlike MOS capacitors, MIM capacitors require high temperatures for thin film deposition. Integrating MIM capacitors into a standard low temperature ($\leq$400 °C) back-end high complexity digital process is therefore a challenging problem [262]. This problem can be overcome by utilizing MIM capacitors with plasma enhanced chemical vapor deposition (PECVD) nitride dielectrics [263, 264]. Previously, MIM capacitors were unavailable in CMOS technology with copper metallization. Recently, MIM capacitors have been successfully integrated into CMOS and BiCMOS technologies with a copper dual damascene metallization process [265–267]. In [268], a high density MIM capacitor with a low ESR using a plug-in copper plate is described, making MIM capacitors highly efficient for use as a decoupling capacitor.

MIM capacitors are widely utilized in RF and mixed-signal ICs due to low voltage coefficients, good capacitor matching, precision control of capacitor values, small parasitic capacitance, high reliability, and low defect densities [269]. MIM capacitors also exhibit high linearity over a wide frequency range. Additionally, a high capacitance density with lower leakage currents has recently been achieved, making MIM capacitors the best candidate for decoupling power and ground lines in modern high performance, high complexity ICs. For instance, for a MIM capacitor with a dielectric thickness $t_{ox} = 1$ nm, a capacitance density of 34.5 $fF/\mu m^2$ has been achieved [270].

### *11.4.4 Lateral Flux Capacitors*

The total capacitance per unit area can be increased by using more than one pair of interconnect layers. Current technologies offer up to ten metal layers, increasing the capacitance nine times through the use of a sandwich structure. The capacitance

is further increased by exploiting the lateral flux between the adjacent metal lines within a specific interconnect layer. In scaled technologies, the adjacent metal spacing (on the same level) shrinks faster than the spacing between the metal layers (on different layers), resulting in substantial lateral coupling.

A simplified structure of an interdigitated capacitor exploiting lateral flux is shown in Fig. 11.30. The two terminals of the capacitor are shown in *light pink* and *dark pink*. Note that the two plates built in the same metal layer alternate to better exploit the lateral flux. Ordinary vertical flux can also be exploited by arranging the segments of a different metal layer in a complementary pattern [271], as illustrated in Fig. 11.31. Note that a higher capacitance density is achieved by using a lateral flux together with a vertical flux (parallel plate structure).

An important advantage of using a lateral flux capacitor is reducing the bottom plate parasitic capacitance as compared to an ordinary parallel plate structure. This reduction is due to two reasons. First, the higher density of the lateral flux capacitor results in a smaller area for a specific value of total capacitance. Second, some of

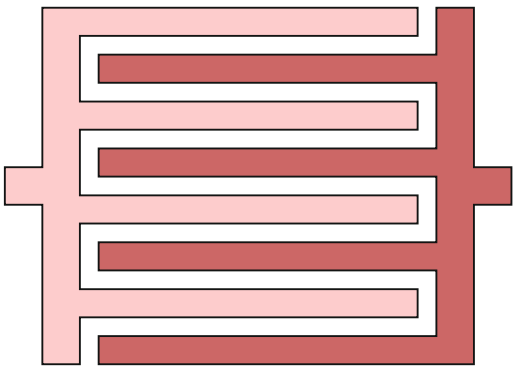

**Fig. 11.30** A simplified structure of an interdigitated lateral flux capacitor (top view). Two terminals of the capacitor are shown in *light pink* and *dark pink*

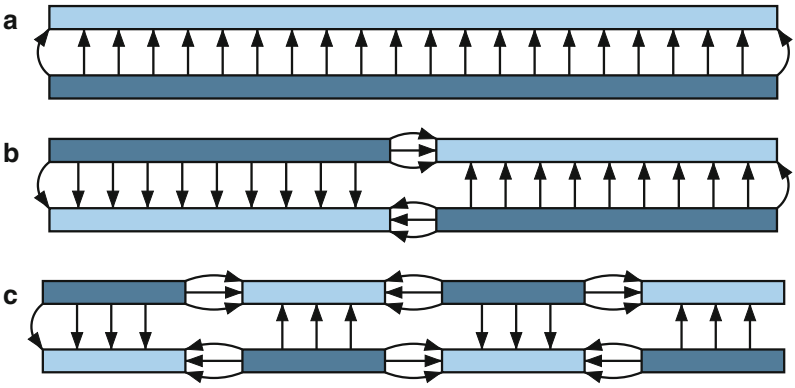


**Fig. 11.31** Vertical flux versus lateral flux; (**a**) standard parallel plate structure, (**b**) structure divided by two cross-connected metal layers, (**c**) structure divided by four cross-connected metal layers

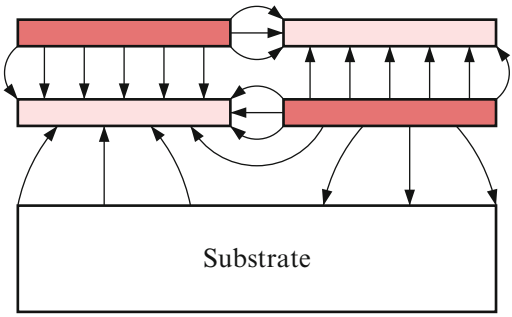


**Fig. 11.32** Reduction of the bottom plate parasitic capacitance through flux stealing. Shades of *pink* denote the two terminals of the capacitor

the field lines originating from one of the bottom plates terminate on the adjacent plate rather than the substrate, further reducing the bottom plate capacitance, as shown in Fig. 11.32. Such phenomenon is referred to as flux stealing. Thus, some portion of the bottom plate parasitic capacitance is converted into a useful plate-to-plate capacitance. Three types of enhanced lateral flux capacitors with a higher capacitance density are described in the following three subsections.

### Fractal Capacitors

Since the lateral capacitance is dependent upon the perimeter of the structure, the maximum capacitance can be obtained with those geometries that maximize the total perimeter. Fractals are therefore good candidates for use in lateral flux capacitors. A fractal is a structure that encloses a finite area with an infinite perimeter [272]. Although lithography limitations prevent fabrication of a real fractal, quasi-fractal geometries with feature sizes limited by lithography have been successfully fabricated in fractal capacitors [273]. It has been demonstrated that in certain cases, the effective capacitance of fractal capacitors can be increased by more than ten times.

The final shape of a fractal can be tailored to almost any form. The flexibility arises from the characteristic that a wide variety of geometries exists, determined by the fractal initiator and generator [272]. It is also possible to use different fractal generators during each step. Fractal capacitors of any desired form can therefore be constructed due to the flexibility in the shape of the layout. Note that the capacitance per unit area of a fractal capacitor depends upon the fractal dimensions. Fractals with large dimensions should therefore be used to improve the layout density [273].

In addition to the capacitance density, the quality factor $Q$ is important in RF and mixed-signal applications. In fractal capacitors, the degradation in quality factor is minimal, since the fractal structure naturally limits the length of the thin metal sections to a few micrometers, maintaining a reasonably small ESR. Hence, smaller dimension fractals should be used to achieve a low ESR. Alternatively, a tradeoff exists between the capacitance density and the ESR in fractal capacitors.

Existing technologies typically provide tighter control over the lateral spacing of the metal layers as compared to the vertical thickness of the oxide layers (both from wafer to wafer and across the same wafer). Lateral flux capacitors shift the burden of matching from the oxide thickness to the lithography. The matching characteristics are therefore greatly improved in lateral flux capacitors. Furthermore, the pseudo-random nature of the lateral flux capacitors compensate for the effects of nonuniformity in the etching process.

Comparing fractal and conventional interdigitated capacitors, note the inherent parasitic inductance of an interdigitated capacitor. Most fractal geometries randomize the direction of the current flow, reducing the ESL. In an interdigitated capacitor, however, the current flows in the same direction for all of the parallel lines. Also in fractal structures, the electric field concentrates around the sharp edges, increasing the effective capacitance density (about 15 %) [273]. Nevertheless, due to simplicity, interdigitated capacitors are widely used in ICs.

### Woven Capacitors

A woven structure is also utilized to achieve high capacitance density. A woven capacitor is depicted in Fig. 11.33. Two orthogonal metal layers are used to construct the plates of the capacitor. Vias connect the metal lines of a specific capacitor plate at the overlap sites. Note that in a woven structure, the current in the adjacent lines flows in the opposite direction. The woven capacitor has therefore much less inherent parasitic inductance as compared to an interdigitated capacitor [202, 274]. In addition, the ESR of a woven capacitor contributed by vias is smaller than the ESR of an interdigitated capacitor. A woven capacitor, however, results in a smaller capacitance density as compared to an interdigitated capacitor with the same metal pitch due to the smaller vertical capacitance.

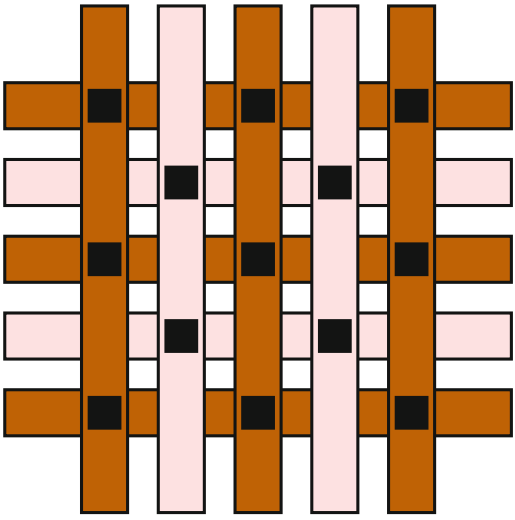

**Fig. 11.33** Woven capacitor. The two terminals of the capacitor are shown in *light pink* and *brown*. The vias are illustrated by the *black colored squares*

**Vertical Parallel Plate (VPP) Capacitors**

Another way to utilize a number of metal layers in modern CMOS technologies is to construct conductive vertical plates out of vias in combination with the interconnect metal. Such a capacitor is referred to as a vertical parallel plate (VPP) capacitor [275]. A VPP capacitor consists of metal slabs connected vertically using multiple vias between the vertical plates. This structure fully exploits lateral scaling trends as compared to fractal structures [274].

### *11.4.5 Comparison of On-Chip Decoupling Capacitors*

On-chip decoupling capacitors can be designed in ICs in a number of ways. The primary characteristics of four common types of on-chip decoupling capacitors, discussed in Sects. 11.4.1, 11.4.2, 11.4.3 and 11.4.4, are listed in Table 11.1. Note that typical MIM capacitors provide a lower capacitance density (1–10 fF/$\mu m^2$) than MOS capacitors. Recently, a higher capacitance density (13 fF/$\mu m^2$) of MIM capacitors has been achieved using laminate ALD $HfO_2$–$Al_2O_3$ dielectrics [259, 260]. A capacitance density of 34.5 fF/$\mu m^2$ has been reported in [270] for a MIM capacitor with a dielectric thickness of 1 nm.

Note that the quality factor of the MOS and lateral flux capacitors is limited by the channel resistance and the resistance of the multiple vias. Decoupling capacitors with a low quality factor produce wider antiresonant spikes with a significantly reduced magnitude [276]. It is therefore highly desirable to limit the quality factor of the on-chip decoupling capacitors. Note that in the case of a low ESR (high quality factor), an additional series resistance should be provided, lowering the magnitude of the antiresonant spike. This additional resistance, however, is limited by the target impedance of the power distribution system [28].

**Table 11.1** Four common types of on-chip decoupling capacitors in a 90 nm CMOS technology

| Feature | PIP capacitor | MOS capacitor | MIM capacitor | Lateral flux capacitor |
|---|---|---|---|---|
| Capacitance density (fF/$\mu m^2$) | 1–5 | 10–20 | 1–30 | 10–20 |
| Bottom plate capacitance (%) | 5–10 | 20–30 | 2–5 | 1–5 |
| Linearity (ppm/V) | 50–150 | 300–500 | 10–50 | 50–100 |
| Quality factor | 5–15 | 1–10 | 50–150 | 10–50 |
| Parasitic resistance (m$\Omega$) | 500–2000 | 1000–10,000 | 50–250 | 100–500 |
| Leakage current (A/$cm^2$) | $10^{-10}$–$10^{-9}$ | $10^{-2}$–$10^{-1}$ | $10^{-9}$–$10^{-8}$ | $10^{-10}$–$10^{-9}$ |
| Temperature dependence (ppm/°C) | 150–250 | 300–500 | 50–100 | 50–100 |
| Process complexity | Extra steps | Standard | Standard | Standard |

The parasitic resistance is another important characteristic of on-chip decoupling capacitors. The parasitic resistance characterizes the efficiency of a decoupling capacitor. Alternatively, both the amount of charge released by the decoupling capacitor and the rate with which the charge is restored on the decoupling capacitor are primarily determined by the parasitic resistance [277]. The parasitic resistance of PIP capacitors is mainly determined by the resistive polysilicon layer. MIM capacitors exhibit the lowest parasitic resistance due to the highly conductive metal layers used as the plates of the capacitor. The increased parasitic resistance of the lateral flux capacitors is due to the multiple resistive vias, connecting metal plates at different layers [274]. In MOS capacitors, both the channel resistance and the resistance of the metal plates contribute to the parasitic resistance. The performance of MOS capacitors is therefore limited by the high parasitic resistance.

Observe from Table 11.1 that MOS capacitors result in prohibitively large leakage currents. As technology scales, the leakage power is expected to become the major component of the total power dissipation. Thick oxide MOS decoupling capacitors are often used to reduce the leakage power. Thick oxide capacitors, however, require a larger die area for the same capacity as a thinner oxide capacitance. Note that the leakage current in MOS capacitors increases exponentially with temperature, further exacerbating the problem of heat removal. Also note that leakage current is reduced in MIM capacitors as compared to MOS capacitors by about seven orders of magnitude. The leakage current of MIM capacitors is also fairly temperature independent, increasing twofold as the temperature rises from 25 °C to 125 °C [265].

Note that PIP capacitors typically require additional process steps, adding extra cost. From the information listed in Table 11.1, MIM capacitors and stacked lateral flux capacitors (fractal, VPP, and woven) are the best candidates for decoupling the power and ground lines in modern high performance, high complexity ICs.

## 11.5 On-Chip Switching Voltage Regulator

The efficiency of on-chip decoupling capacitors can be enhanced by an on-chip switching voltage regulator [278]. The decoupling capacitors reduce the impedance of the power distribution system by serving as an energy source when the power voltage decreases, as discussed in previous sections. The smaller the power voltage variation, the smaller the energy transferred from a decoupling capacitor to the load.

Consider a group of $N$ decoupling capacitors $C$ placed on a die. Where connected in parallel between the power and ground network, the capacitors behave as a single capacitor $NC$. As the power supply level decreases from the nominal level $V_{dd}$ to a target minimum $V_{dd} - \delta V$, the non-switching decoupling capacitors release only a small fraction $k$ of the stored charge into the network,

$$\frac{\delta Q}{Q_0} = \frac{NCV_{dd} - NC(V_{dd} - \delta V)}{NCV_{dd}} = \frac{\delta V}{V_{dd}} \equiv k. \tag{11.21}$$

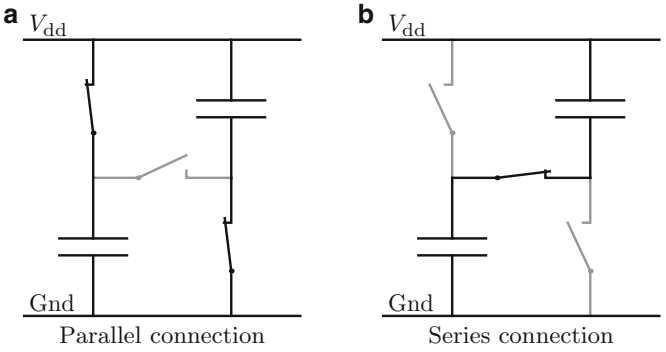


**Fig. 11.34** Switching decoupling capacitors from a parallel to a series connection. (**a**) Parallel connection. (**b**) Series connection

A correspondingly small fraction of the total energy $E$ stored in the capacitors is transferred to the load,

$$\frac{\delta E}{E_0} = \frac{V_{dd}^2 - (V_{dd} - \delta V)^2}{V_{dd}^2} \approx \frac{2\delta V}{V_{dd}} = 2k. \tag{11.22}$$

Switching the on-chip capacitors can increase the charge (and energy) transferred from the capacitors to the load as the power voltage decreases below the nominal voltage level [278]. Rather than a fixed connection in parallel as in the traditional non-switching case, the connection of capacitors to the power and ground networks can be changed from parallel to series using switches, as shown in Fig. 11.34 for the case of two capacitors. When the rate of variation in the power supply voltage is relatively small, the capacitors are connected in parallel, as shown in Fig. 11.34a, and charged to $V_{dd}$. When the instantaneous power supply variation exceeds a certain threshold, the capacitors are reconnected in series, as shown in Fig. 11.34b, transforming the circuit into a capacitor of $C/N$ capacity carrying a charge of $CV_{dd}$. In this configuration, the circuit can release

$$\delta Q_{sw} = CV_{dd}\left(1 - \frac{1-k}{N}\right) \tag{11.23}$$

amount of charge before the drop in the voltage supply level exceeds the noise margin $\delta V = kV_{dd}$. This amount of charge is greater than the charge released in the non-switching case, as determined by (11.21), if $k < \frac{1}{N+1}$. The effective charge storage capacity of the on-chip decoupling capacitors is thereby enhanced. The area of the on-chip capacitors required to lower the peak resonant impedance of the power network to a satisfactory level is decreased.

This technique is employed in the UltraSPARC III microprocessor, as described by Ang, Salem, and Taylor [278]. In addition to the 176 nF of non-switched on-chip

decoupling capacitance, 134 nF of switched on-chip capacitance is placed on the die. The switched capacitance occupies 20 $mm^2$ of die area and is distributed in the form of 99 switching regulator blocks throughout the die to maintain a uniform power supply voltage. The switching circuitry is designed to minimize the short-circuit current when the capacitor is switching. Feedback loop control circuitry ensures stable behavior of the switching capacitors. The switching circuitry occupies 0.4 $mm^2$, a small fraction of the overall regulator area. The regulator blocks are connected directly to the global power distribution grid. In terms of the frequency domain characteristics, the switching regulator lowers the magnitude of the die-package resonance impedance. The switched decoupling capacitors decrease the on-chip power noise by roughly a factor of 2 and increase the operating frequency of the circuit by approximately 20 %.

## 11.6 Summary

A brief overview of decoupling capacitors has been presented in this chapter. The primary characteristics of decoupling capacitors can be summarized as follows.

- A decoupling capacitor serves as an intermediate and temporary storage of charge and energy located between the power supply and current load, which is electrically closer to the switching circuit
- To be effective, a decoupling capacitor should have a high capacity to store a sufficient amount of energy and be able to release and accumulate energy at a sufficient rate
- In order to ensure correct and reliable operation of an IC, the impedance of the power distribution system should be maintained below the target impedance in the frequency range from DC to the maximum operating frequency
- The high frequency impedance is effectively reduced by placing decoupling capacitors across the power and ground interconnects, permitting the current to bypass the inductive interconnect
- A decoupling capacitor has an inherent parasitic resistance and inductance and therefore can only be effective within a certain frequency range
- Several stages of decoupling capacitors are typically utilized to maintain the output impedance of a power distribution system below a target impedance
- Antiresonances are effectively managed by utilizing decoupling capacitors with low ESL and by placing a large number of decoupling capacitors with progressively decreasing magnitude, shifting the antiresonant spike to a higher frequency
- MIM capacitors and stacked lateral flux capacitors (fractal, VPP, and woven) are preferable candidates for decoupling power and ground lines in modern high speed, high complexity ICs