Chapter 2
CMOS Reliability Overview

2.1 Introduction

For over four decades, scientists have been scaling devices to increasingly smaller feature sizes (Lewyn et al. 2009; International technology roadmap for semiconductors 2011). This trend is driven by a seemingly unending demand for ever-better performance and by fierce global competition. The steady CMOS technology downscaling is needed to meet requirements on speed, complexity, circuit density, power consumption and ultimately cost required by many advanced applications. However, going to these ultra-scaled CMOS devices also brings some drawbacks.

This chapter discusses the most important effects designers have to deal with in order to manufacture reliable integrated circuits in nanometer CMOS processes. The intent of this chapter is not to give an in-depth description of the physics behind each failure mechanism, but to provide the reader with a basic understanding of the most important unreliability effects and how these effects evolve with technology. First, Sect. 2.2 briefly outlines how various unreliability effects came into play in the course of history. Next, Sect. 2.3 reviews the most important spatial unreliability effects in modern CMOS technologies. These effects are related to process variations and are visible right after production. A difference is made between systematic and random effects. Time-dependent unreliability effects are then discussed in Sect. 2.4. These effects are divided into aging and transient effects.

2.2 The Origin of CMOS Unreliability

Device reliability was first studied in the early sixties, when increasingly complex integrated systems were developed and fabricated. Conferences such as the first international reliability physics symposium (IRPS 1962, Chicago) were the first attempts to bring engineers and scientists together from all over the world to study the physics behind various failure effects (Physics of failure in electronics 1962).

E. Maricau and G. Gielen, *Analog IC Reliability in Nanometer CMOS*,
Analog Circuits and Signal Processing, DOI: 10.1007/978-1-4614-6163-0_2,

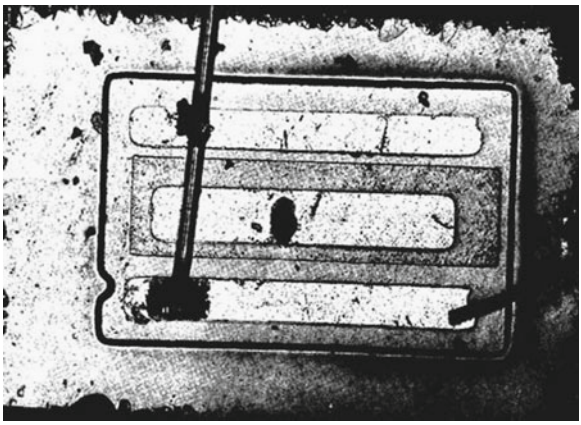

Fig. 2.1 Photomicrograph of an early silicon mesa transistor on which the emitter bond has separated due to 'purple plague' (Phillips et al. 1962). This phenomenon, also known as 'purple death', was an important reliability problem in the late 1960s and the early 1970s. An intermetallic reaction between the golden bond wires and the aluminum bond pads formed a brittle bright purple compound of $AuAl_2$ which led to the creation of voids in the metal lattice

During the 1970s, effects such as corrosion, bonding issues (e.g. the 'purple plague' as depicted in Fig. 2.1) and ionic contamination were the most common causes of circuit failure. All these issues were however related to the way how integrated circuits were packaged and mounted on a printed circuit board (PCB). Only in the late seventies and early eighties the first real integrated circuit reliability issues became visible. Oxide thickness scaling increased the gate-oxide electric field, and transistor wearout effects such as hot carrier injection (HCI) started to affect device performance within the lifetime of a circuit (Takeda et al. 1983; Hu et al. 1985). Initially, the application of an arbitrary voltage stress resulted in an identical parameter shift for matched devices. Therefore these temporal unreliability effects were at first considered as deterministic. However, when the oxide dielectric reached atomic-scale dimensions, this resulted in the first stochastic temporal unreliability effect: time-dependent dielectric breakdown (Solomon 1977). Further, matched devices were, in the early eighties, considered identical in terms of electrical performance. In the second half of that decade, however, when device dimensions entered the nanometer scale, stochastic errors and variations at atomic level became apparent at device level and sensitive analog circuits were the first to suffer from process variability effects (Lakshmikumar et al. 1986; Pelgrom et al. 1989). Device mismatch became a big issue (especially analog) designers had to deal with in order to guarantee good accuracy and high yield.

To overcome scaling limitations of devices fabricated in ultra-scaled CMOS processes, changes in device structures, processing materials and processing conditions have been introduced. These changes have drastically increased the complexity of nanometer CMOS technologies. Examples of these new techniques include:

Table 2.1 Evolution of nanometer CMOS characteristics

Year	L_g (nm)	V_{DD} (V)	V_{TH} (V)	EOT (nm)	E_{ox} (MV/cm)
1995	350	3.3	0.58–0.70	10.0–12.0	2.17–2.72
1998	250	1.8–2.5	0.47–0.52	6.0–7.0	1.83–3.38
2003	180	1.8	0.39–0.43	4.5–5.5	2.49–3.13
2001	130	1.2	0.35–0.40	3.5–4.0	2.00–2.43
2004	90	1.0–1.2	0.25–0.40	1.6–3.0	2.00–5.93
2007	65	1.0–1.2	0.20–0.35	1.5–2.0	3.25–6.66
2009	45	1.0–1.1	0.20–0.35	1.0–1.4	4.64–9.00
2011	32	0.9–1.0	0.20–0.35	0.8–1.1	5.00–10.0

(Iwai 1999; Bult 2000; Bravaix et al. 2009; Wu et al. 2009; Europractice 2012; International technology roadmap for semiconductors 2011)

strained silicon channels to increase the transistor drive current, the introduction of high-k oxides and metal gates to allow further gate oxide scaling combined with reduced gate leakage, and Cu-interconnect with low-k dielectrics to ensure lower RC-delays (Horstmann et al. 2009; International technology roadmap for semiconductors 2011). However, the introduction of these new materials also increased the impact of already existing but before then unimportant aging effects, such as electromigration (EM) and negative bias temperature instability (NBTI), and even created new problems such as positive bias temperature instability (PBTI) (Lewyn et al. 2009; Groeseneken et al. 2010). Table 2.1 gives an overview of typical technology parameters for the most recent CMOS nodes. The table clearly shows how the average oxide electric field increases with each new technology node, aggravating all transistor wearout effects. All of these phenomena can have a large impact on the reliability of a circuit, right after production or during its operational lifetime. Therefore, a good understanding of the impact of each effect on the electrical behavior of a single transistor and eventually on the performance of an entire circuit is mandatory.

Figure 2.2 illustrates how nanometer CMOS reliability issues can be categorized into spatial and temporal unreliability effects. Spatial unreliability effects are immediately visible right after production and are fixed in time. Spatial unreliability effects can be random (e.g. random dopant fluctuations (RDF), line edge roughness (LER), etc) or systematic (e.g. gradient effects, etc.). The effects depend on the circuit layout, the neighboring environment, process conditions and the impact the geometry and structure of the circuit and can lead to yield loss. This yield loss can be functional or parametric, i.e. resulting in malfunctioning circuits or circuits with degraded performance respectively. Temporal unreliability effects, on the other hand, are time-varying and change depending on operating conditions such as the operating voltage, temperature, switching activity, presence and activity of neighboring circuits. A difference is made between wearout or aging effects (e.g. hot carrier injection (HCI), NBTI, etc.) and transient effects (e.g. electromagnetic interference (EMI), single event upsets (SEU), etc.). In the following sections, these effects are discussed in more detail.

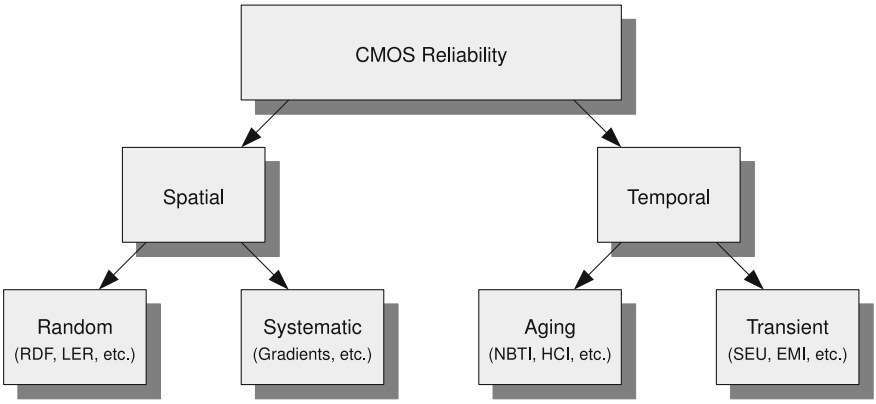

Fig. 2.2 A CMOS circuit can fail from spatial or temporal unreliability effects. The former are visible right after production and can be random or systematic. The latter become a potential problem during the operational lifetime of the circuit and present themselves as an aging effect or a transient effect

2.3 Spatial Unreliability

Spatial unreliability or process variability is an increasing problem in nanometer CMOS IC production. The problem results from the increasing complexity needed to fabricate nanometer CMOS devices, combined with the scaling towards atomistic device dimensions (<180 nm CMOS). Typically, parametric yield is used as a metric to express the impact of these effects on the performance of the circuit right after production.[1] A high yield implies low spatial unreliability.

Two major sources of process variability are distinguished: local or intradie and global or interdie effects. Local variability results in parametric variations of identically designed transistors across a short distance, typically within the same circuit. This is also referred to as device mismatch. Global variability refers to variations between devices that are separated by a long distance or that are fabricated at a different time. Typically global variability is variability from die to die, wafer to wafer or lot to lot. Global variability causes a shift in the mean value of design parameters such as channel length or doping density. Since most spatial unreliability problems result from local variability effects, global variability is not discussed here. Local variability originates from systematic and random reliability effects. The former includes variability caused by optical proximity correction, layout-induced strain and well-proximity effects. The latter includes random dopant fluctuation (RDF) effects, line edge and width roughness (LER and LWR), fixed charges in the gate dielectric and interface roughness. Systematic variability is typically addressed through

[1] Parametric circuit failures are related to process variations and are circuits that do function but with a performance outside the desired range. Catastrophic circuit failures result from process errors or defects and are described by the functional yield. The latter are not covered in this work.

careful layout design, compensating circuit techniques and with advanced manufacturing flows. Solving random variability issues, on the other hand, requires innovative process and design techniques and accurate device models. For technology generations below 90 nm CMOS, the impact of random variability is becoming increasingly important (Lewyn et al. 2009). Both the systematic and random variability effects are discussed in more detail in the next sections.

2.3.1 Systematic Effects

While most device-related sources of spatial unreliability are random, a large fraction of the variation of the interconnect is a function of layout characteristics (i.e. design dependent). These sources of variability have a large systematic component. With the aggressive scaling to smaller feature sizes, this component has become larger primarily due to resolution limitations. The inability to scale the wavelength of the light source for lithography has led to an increase of systematic variations, especially in circuit areas with high interconnect and device density (Agarwal and Nassif 2007). To mitigate these problems, a lot of research has gone into more advanced manufacturing flows such as double-patterning technologies (DPT), optical-proximity correction (OPC), extreme ultraviolet lithography (EUVL) and into design techniques such as the use of extremely regular circuit layout (Strojwas 2011).

2.3.2 Random Effects

Random spatial unreliability results from physical phenomena such as random dopant effects, line edge and width roughness, fixed charges in the gate dielectric and oxide thickness variation resulting from interface roughness (Agarwal and Nassif 2007). Random effects typically affect the mismatch between closely spaced identically designed devices. At device level these effects mainly result in variations of the gate length (L), the threshold voltage (V_{TH}) and the current factor (β) (Zhao et al. 2007), and they can to first order be modeled with Pelgrom's model (Pelgrom et al. 1989):

$$\sigma(\delta V_{\mathrm{TH}}) \approx \frac{A_{\mathrm{VTH}}}{\sqrt{WL}} \tag{2.1}$$

where $\sigma(\delta V_{\mathrm{TH}})$ is the standard deviation on the threshold voltage mismatch between two identically sized transistors, WL is the size of the active area of one transistor and A_{VTH} is a technology-dependent constant commonly expressed in $\mathrm{mV}\,\mu\mathrm{m}$.[2]

[2] Pelgrom's model expresses the standard deviation on the *difference* between the threshold voltages of two matched transistors. The standard deviation on the threshold voltage of a single transistor can be found by dividing A_{VTH} by $\sqrt{2}$.

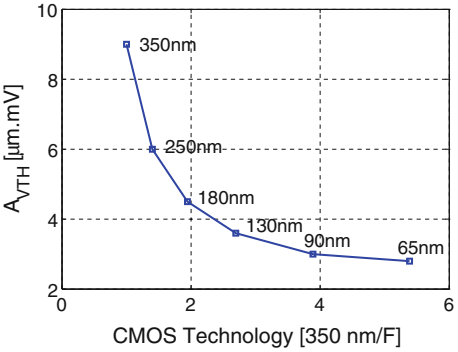

Fig. 2.3 Measured A_{VTH} values for minimum-length pMOS devices as a function of 350 nm over the minimum feature size F. Extrapolating the curve, significant improvements are not expected beyond the 65 nm technology node. Data taken from Lewyn et al. (2009)

The actual relationship is more complex than (2.1) (Hong et al. 2011). Currently, IC foundries supply Monte-Carlo (MC) simulation models to accurately simulate the impact of process variations on transistor and circuit performance. Pelgrom's formula is however still used by designers for initial circuit design. It is therefore interesting to look at A_{VTH} trends in advanced technology nodes (see Fig. 2.3). From Fig. 2.3 it is clear how A_{VTH} does not improve much beyond the 90 nm technology node. Where for older technologies (>180 nm) transistor matching was primarily determined by lithographic accuracy that scaled well with technology, other factors that do not scale well are now taking over (Lewyn et al. 2009). These effects are discussed in the following sections.

Random Dopant Fluctuations

Variations of device parameters such as the transistor V_{TH} partly result from fluctuations in the amount and location of dopant atoms in the transistor channel (Takeuchi et al. 2007; Kuhn et al. 2008). Since the number of dopant atoms in the channel of scaled transistors is always decreasing, the impact of the variation associated with the atoms increases. Figure 2.4 illustrates the decreasing average number of dopant atoms as a function of technology. Note how the number of dopants decreases by almost three orders of magnitude when going from a 1 µm (with around 1e4 dopant atoms) to a 32 nm technology (with less than 100 dopant atoms). Random dopant fluctuations (RDF) are assumed to be the major contributor to device mismatch of identically designed devices. For example, in (Kuhn 2007) it is shown how the simulated RDF is responsible for around 65 % of the total NMOS V_{TH} variation of a 65 nm CMOS technology. Similar results were obtained for a 45 nm PMOS transistor, where RDF was responsible for 60 % of the $\sigma(V_{TH})$. The effect of RDF on the transistor threshold variation is frequently represented by (Stolk et al. 1998):

$$\sigma(V_{TH}) \propto \frac{t_{ox}}{\varepsilon_{ox}} \frac{\sqrt[4]{N}}{\sqrt{W_{eff} L_{eff}}} \tag{2.2}$$

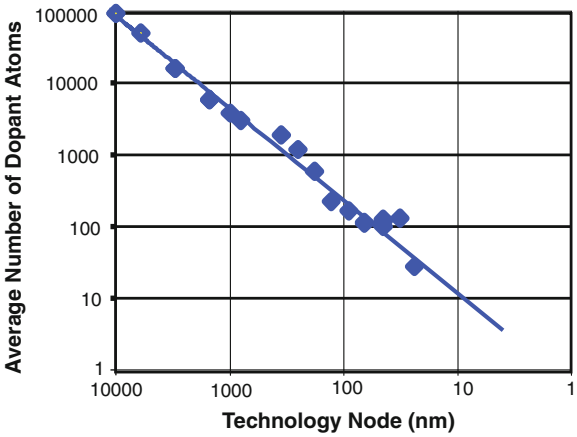

Fig. 2.4 Average number of dopant atoms in the channel of a transistor as a function of the technology node (Kuhn et al. 2008)

with t_{ox} the gate oxide thickness, ε_{ox} the oxide permittivity, N the number of channel dopants and W_{eff} and L_{eff} the transistor effective width and length. Equation (2.2) shows that V_{TH} matching improves with technology scaling (both t_{ox}/ε_{ox} and N reduce with smaller feature sizes). However, the device area ($W_{eff} L_{eff}$) also decreases with each new technology generation. Therefore the net result of RDF is a significant increase in process variability for scaled CMOS technologies.

Line Edge/Width Roughness

Line edge and line width roughness (LER and LWR respectively) result from subwavelength lithography. Since the 0.25 μm technology generation, the semiconductor industry has used subwavelength lithography to pattern transistors. Fabrication processes initially used the wavelength of light ($\lambda = 248$ nm) to pattern minimum feature sizes of 250 and 180 nm transistors. Then the value of λ decreased to 193 for 130 nm transistors and it has remained there ever since, even for 65 nm and smaller transistors. As shown in Fig. 2.5, this lithographic gap causes the LER and LWR effect (saha 2010). Although LER and LWR occur in both the front-end and the back-end of a CMOS process, LER and LWR in the poly-gate patterning are the primary concern (Kim et al. 2004). This results in both an increase of the subthreshold current as well as a variation of the V_{TH} (Asenov et al. 2003; Fukutome et al. 2006). Further, assuming the variations on the line edge are not correlated, $\sigma(\text{LWR}) = \sqrt{2}\sigma(\text{LER})$. Asenov et al. (2003) studied the combined effect of LER and RDF on current fluctuations. They demonstrated that these two sources of transistor variability are statistically independent. Experiments have shown that LER is on the order of 5 nm and does not scale with technology. Also, LER has a much stronger channel length dependence compared to RDF. LER is expected to replace

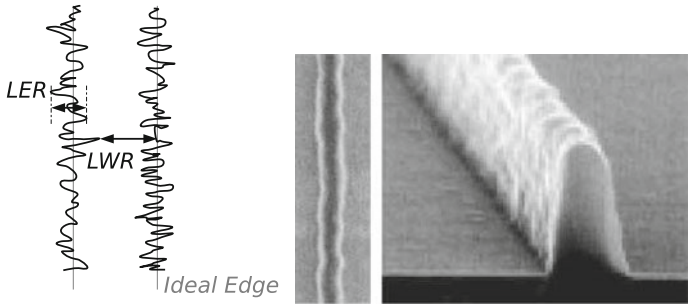

Fig. 2.5 Line edge roughness and line width roughness are a major source of process variations and result in an increase of the subthreshold current and variations on the V_{TH}. Photo: (Mack 2006)

RDF as the dominant source of transistor mismatch as device scaling continues. The transitional channel length is around 45 nm and depends on the actual device architecture and on the lithographic process used to fabricate the devices.

Gate Dielectric Variations

The gate dielectric can suffer from various non-idealities and defects such as variations in the oxide thickness, fixed charges in the oxide and interface traps. These physical effects result in parametric variations in the drive current, the gate tunneling current and/or the V_{TH} (Kuhn et al. 2008; Saha 2010).

Asenov et al. (2003) have shown that V_{TH} fluctuations induced by local oxide thickness variations become comparable to voltage fluctuations introduced by RDF for sub-30 nm CMOS technologies. Moreover, the variability due to oxide thickness fluctuations is statistically independent from the V_{TH} variations introduced by RDF. HKMG devices allow larger physical oxide thicknesses, but still suffer from large oxide thickness variations due to the roughness of the interface between the silicon and the high-k layer and between the high-k layer and the metal gate (saha 2010).

High-k gate dielectrics also suffer from fixed charges in the gate-oxide layer. These charges can affect the carrier mobility and the V_{TH}. As a consequence, variations in the location of these charges may affect the distribution of the mobility and the V_{TH} (Kaushik et al. 2006). Further, electron mobility degradation and V_{TH} instability due to fast transient charging (FTC) in interface traps is an increasing concern for high-k dielectrics (Kuhn et al. 2008).

Other Sources

Other sources of random process variability include: variations associated with patterning proximity effects such as optical proximity correction (OPC), variations associated with polish such as shallow-trench isolation and its effect on gates and

interconnections and variations associated with strain such as high-stress capping layers and embedded silicon germanium layers (Kuhn et al. 2008; Saha 2010).

2.4 Temporal Unreliability

Temporal unreliability becomes apparent after a circuit has been produced, when it is used in a certain environment, at a given temperature and workload and over a period of time. The impact of these effects on the circuit can be permanent or temporary. Aging effects cause a gradual degradation of the circuit (which does not always directly result in reduced circuit performance) and at least part of the damage is permanent. Transient effects only temporarily distort the circuit performance and the circuit performs back as before once the noise source is removed.

2.4.1 Aging Effects

Integrated circuit aging phenomena were first observed during the seventies and the eighties. At that time, research effort was mainly focused towards understanding these effects, rather than solving circuit reliability problems. In the nineties, however, circuit aging became more and more an issue due to the aggressive scaling of the device geometries and the increasing electric fields. At that time, measurements on individual transistors were used to determine circuit design margins in order to guarantee reliability. After the turn of the century, the introduction of new materials to further scale CMOS technologies introduced additional failure mechanisms and made existing aging effects more severe. This section reviews the most important integrated-circuit aging phenomena observed in sub-90 nm CMOS technologies: hot carrier injection (HCI), time-dependent dielectric breakdown (TDDB), bias temperature instability (BTI) and electromigration (EM).

Hot Carrier Injection

In general, hot carriers are particles that obtain a very high kinetic energy from being accelerated in a high electric field. These energetic carriers can be injected into 'forbidden' regions of the device, such as the gate oxide, instead of following their intended trajectory. When injected into such a region they can get trapped or cause the generation of interface states. These defects in turn lead to shifts in the electrical characteristics of the transistor such as a shift of the V_{TH}, the current factor β and the output conductance g_{o}. The degradation of integrated circuits due to hot carrier injection (HCI) first became a problem in the mid-eighties due to the continuous scaling of transistor dimensions without accompanying supply voltage reduction (Takeda et al. 1983; Tam et al. 1984; Hu et al. 1985). In the mid-nineties,

the circuit operating voltage was dropped to reduce power consumption and graded drain junctions were introduced to solve reliability problems. Hence, HCI became less of an issue. Also, measurements on advanced HKMG CMOS transistors revealed how high-k stacks appear to be more resilient to HCI stress than SiO_2 stacks (Amat et al. 2009). Nevertheless, HCI can still be a problem since supply voltage scaling is slowing down in recent years because of the non-scalability of the subthreshold slope (Wang et al. 2007; Maricau et al. 2008; Bravaix et al. 2009). HCI is primarily a problem in nMOS devices (Lunenborg 1996; Parthasarathy 2006). Nevertheless, although pMOS devices are less sensitive to HCI, the effect can enhance other aging effects such as negative bias temperature instability (NBTI) (Parthasarathy 2006).

As illustrated in Fig. 2.6, four different hot carrier injection mechanisms can be distinguished (Takeda et al. 1983): channel hot electron (CHE) injection, drain avalanche hot carrier (DAHC) injection, secondary generated hot electron (SGHE) injection and substrate hot electron (SHE) injection.

1. When the gate voltage is approximately equal to the drain voltage, the **channel hot electron** (**CHE**) injection effect is at its maximum. So-called 'lucky electrons' gain sufficient energy to surmount the Si/SiO_2 barrier at the drain end of the channel, without loosing energy due to collisions with atoms in the channel (see Fig. 2.6a). For low gate voltages, the field does not attract electrons to the

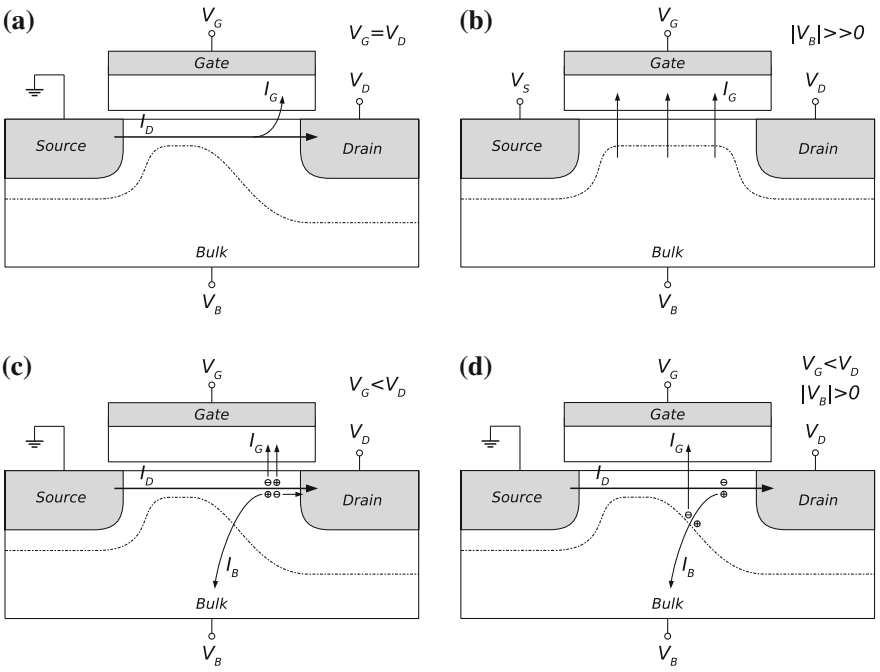

Fig. 2.6 Four different hot carrier injection mechanisms can be distinguished. **a** CHE. **b** SHE. **c** DAHC. **d** SGHE

gate electrode and for high drain voltages the electric field at the drain leads to avalanche multiplication resulting in drain avalanche hot carrier generation. As holes are much 'cooler' (i.e. heavier) than electrons, the channel hot carrier effect in nMOS devices is shown to be more significant than in pMOS devices (Hu et al. 1985).

2. **Substrate hot electron** (**SHE**) or substrate hot hole (SHH) injection is the result of a high positive or a high negative bias at the bulk of the transistor. This leads to carriers in the substrate driven to the Si/SiO_2 interface, gaining kinetic energy and potentially surmounting the energy barrier at the channel/gate-oxide interface to be injected into the oxide. In contrast to the other hot carrier generation mechanisms, this effect is uniformly distributed along the channel instead of being concentrated near the drain of the transistor (see Fig. 2.6b). This generation mechanism is especially present in circuits where stacked devices, typically implying a non-zero bulk bias, are used (e.g. current-source differential pairs and cascode circuits).
3. At stress conditions with high drain voltage and low gate voltage, electron-hole pairs can be created due to impact ionization of the channel current near the drain of the transistor. Each of these electrons and holes can then accelerate in the channel electric field and can potentially surmount the Si/SiO_2 barrier to get trapped or to create interface states. This phenomenon is known as avalanche multiplication and results in **drain avalanche hot carrier generation** (**DAHC**) (see Fig. 2.6c). Additionally, some of the generated carriers lead to a bulk current. The DAHC injection mechanism causes the most stringent device degradation because a large amount of hot electrons are injected into the gate oxide at the same time.
4. **Secondary generated hot electron injection** (**SGHE**) involves the generation of hot carriers from impact ionization with a secondary carrier that was created by an earlier impact ionization incident. This earlier generated carrier can be generated under DAHC conditions or from photons generated in the high field region near the drain (i.e. bremsstrahlung radiation). Under the influence of the field generated by the substrate's bulk bias, the first carriers are accelerated and potentially generate secondary carriers. These secondary carriers also accelerate in the bulk bias field towards the surface region where they further gain kinetic energy to overcome the surface energy barrier (see Fig. 2.6d). SGHE is observed as a rather small effect with limited contribution to the transistor degradation.

As explained above, each of the four hot carrier mechanisms occurs at different transistor operating conditions. Typically, DAHC and CHE effects are much worse than SHE and SGHE effects and therefore limit the device and circuit lifetime. For transistors with a minimum gate length of $0.35\,\mu m$, DAHC has the worst effect on transistor performance and is at its maximum when $2V_{GS} = V_{DS}$. For smaller transistor dimensions, on the other hand, CHE dominates the hot carrier degradation effect (Takeda et al. 1983).

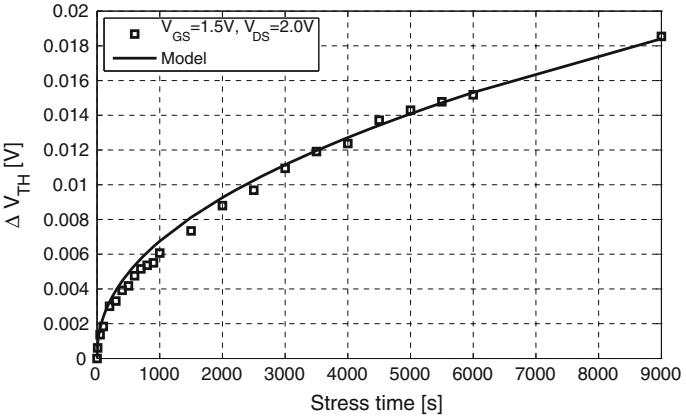

Fig. 2.7 Hot carrier injection (HCI) is typically modeled with a power law time dependence. Here the (measured and modeled) V_{TH} shift of a 65 nm nMOS transistor, stressed with $V_{GS} = 1.5\,V$ and $V_{DS} = 2.0\,V$, is depicted (Maricau et al. 2008)

HCI is typically modeled with a power law dependence on the stress time (see Fig. 2.7) (Hu et al. 1985; Kufluoglu and Ashraful Alam 2004; Maricau et al. 2008):

$$\Delta V_{TH} = A_{HCI} t^{n_{HCI}} \tag{2.3}$$

where ΔV_{TH} represents the HCI-induced V_{TH} shift and n_{HCI} is the time exponent which is typically around 0.5 (Hu et al. 1985). The trapping generation of the carriers increases exponentially with increasing oxide electric field (E_{ox}). Besides the oxide electric field and the maximum lateral electric field (E_{lat}), HCI dependence on temperature (T) and transistor length (L) has also been reported (Hu et al. 1985; Wang et al. 2007; Maricau et al. 2008):

$$A_{HCI} \propto \frac{1}{\sqrt{L}} \exp\left(\alpha_{HCI,1} E_{ox}\right) \exp\left(-\frac{\alpha_{HCI,2}}{E_{lat}}\right) \tag{2.4}$$

with $\alpha_{HCI,1}$ and $\alpha_{HCI,2}$ technology-dependent parameters. In addition to the average effect, predicted by (2.3) and (2.4), HCI also introduces an extra source of variability, due to the randomly generated traps in the gate dielectric or at the substrate/dielectric interface. Further, this effect has been shown to be more pronounced for sub-65nm technologies (Magnone et al. 2011).

Time-Dependent Dielectric Breakdown

The correct operation of a MOS transistor relies on the insulating properties of the dielectric layer below the gate electrode of the transistor (Stathis 2001). Each dielectric material has a maximum electric field it can sustain. When a larger electric

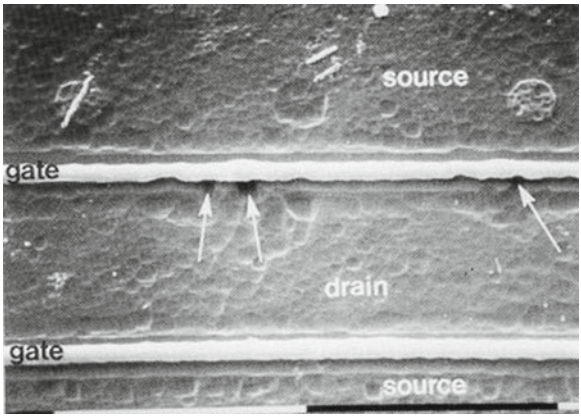

Fig. 2.8 Multiple breakdown spots at the drain junction of an nMOS transistor. Note the thermal damage to the silicon. *Source* Yazdani (2011)

field is applied, this leads to hard breakdown (HBD). HBD is an extremely local phenomenon, characterized by a loss of the gate oxide insulating properties and allowing a large gate current to flow.[3]

At lower electric fields, the insulator can wearout after some time and finally break down completely. This is called time-dependent dielectric breakdown (TDDB) (Fig. 2.8).

Prior to oxide TDDB, a degradation process of the dielectric takes place that initiates the generation of traps in random positions inside the oxide and at the interface. A stress-induced leakage current (SILC) is produced during this degradation stage (Kamohara et al. 1998; Young et al. 2012). If the dielectric degradation increases, a critical trap density is reached and BD occurs. Due to this behavior HBD is a stochastic phenomenon and can be described using a Weibull probability distribution (Wolters and van der Schoot 1985; Wu and Su 2005):

$$F(t_{\mathrm{BD}}) = 1 - \exp\left[-\left(\frac{t_{\mathrm{BD}}}{\alpha_{\mathrm{BD}}}\right)^{\beta_{\mathrm{BD}}}\right] \tag{2.5}$$

with $F(t_{BD})$ the cumulative density function for the time-to-BD. α_{BD} and β_{BD} are technology-dependent parameters.

[3] Besides TDDB, which is a time-dependent wearout effect, oxide BD can also result from electrical overstress (EOS), electrostatic discharge (ESD) or under the presence of weak spots in the oxide. EOS and ESD involve the application of a high voltage being applied across the oxide. This causes a dramatic increase of the gate current, localized heating and a meltdown of the silicon. Early life BD failures due to weak spots in the oxide are essentially similar to TDDB, but happen within the first year of the circuit operational life. This work focuses aging effects, therefore EOS, ESD and early life failures are not discussed here.

During a breakdown degradation process, different BD modes can be distinguished. Depending on the thickness of the gate oxide, one or more modes occur. The most harmful mode, the Hard-BD (HBD), provokes the complete loss of the oxide dielectric properties with gate currents in the mA range at standard operation voltages. However, HBD is in nanometer CMOS technologies only a significant reliability threat at elevated operating voltages (i.e. $V_{GS} > 1.2V$ for EOT $= 0.9$ nm) (Degraeve et al. 2008; Pae et al. 2010).

For oxide thicknesses below 5 nm (i.e. sub-180 nm CMOS), HBD can be preceded by Soft-BD (SBD). SBD can be observed as a partial loss of the dielectric properties, resulting in a small increase of the gate current and a significant increase of the gate current noise (Gielen et al. 2008). Finally, in ultra-thin oxides (approximately below 2.5 nm thickness), SBD is followed by Progressive-BD (PBD), until final HBD. PBD is detected as a slow increase of the gate current over time (see Fig. 2.9).

When looking at the impact of BD on the transistor electrical characteristics, it has been shown that the degradation process prior to BD (Martín-Martínez et al. 2007) and the BD spot location (Fernández 2007) can vary largely for transistors of the same size and therefore have a strong influence on the channel current. The transistor geometry also has a strong impact on this current. Although right after SBD a very limited effect is observed (Kaczer et al. 2004), a significant influence on the transistor characteristics is produced at longer times (Kaczer et al. 2004; Cester et al. 2004). This can be modeled as a local mobility reduction in the BD region (Cester et al. 2004). Another important aspect of gate oxide breakdown is the fact that one BD does not necessarily imply circuit failure (Kaczer et al. 2002).

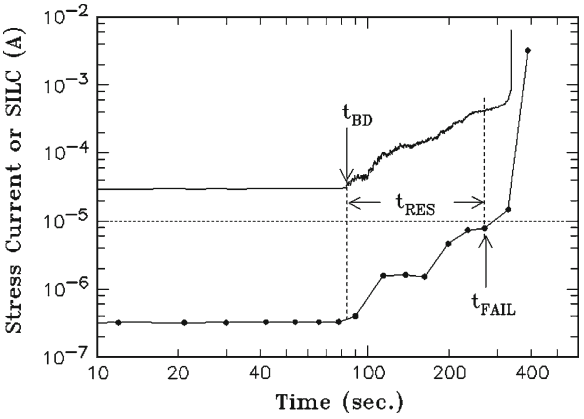

Fig. 2.9 Evolution of the gate current under constant voltage stress (CVS with $V_G = 2.75$ V and T = 140 °C, *top curve*) and the stress-induced leakage (SILC) at 1V (*bottom curve*). The stressed devices are 8 μm^2 nFETs with a $t_{ox} = 1.35$ nm. t_{BD} indicates the first soft breakdown (SBD) event, t_{FAIL} marks the initiation of the hard breakdown (HBD) effect. The residual time (t_{RES}) between soft and hard breakdown depicts the progressive breakdown state. *Source* (Sune et al. 2006)

Bias Temperature Instability

Bias Temperature Instability (BTI) recently gained a lot of attention due to its increasingly adverse impact in nanometer CMOS technologies (Schroder and Babcock 2003). BTI is typically observed as a V_{TH} shift after a bias voltage has been applied to a MOS gate at elevated temperature. For example, when measured over a lifetime of 5 years and under normal operating conditions, V_{TH} shifts of up to 30 mV can be expected for transistors processed in a sub-45 nm technology (Kaczer et al. 2010). BTI-induced degradation of the carrier mobility has also been measured (Schroder and Babcock 2003).

Two different BTI phenomena can be observed: negative BTI (NBTI) and positive BTI (PBTI). NBTI occurs in pMOS transistors when a negative bias voltage is applied. This effect is a significant reliability threat in both older SiO_2 and SiON technologies and is still a problem in newer HKMG technologies (Degraeve et al. 2008). The PBTI effect affects nMOS transistors and results in a similar wearout behavior as NBTI, but has only been observed in HKMG nMOS devices. There, the impact of PBTI on the transistor characteristics can be similar to or even larger than the NBTI effect (Grasser et al. 2010). Currently, there still is no consensus about the microscopical origins of both BTI phenomena. Most authors argue that the NBTI effect results from a combination of hole trapping in oxide defects and generation of interface states at the channel oxide interface (Schroder and Babcock 2003; Kaczer et al. 2008; Grasser and Kaczer 2009). PBTI is believed to come from electron trapping in preexistant oxide traps, combined with a trap generation process (Crupi et al. 2005; Ioannou et al. 2009). Further, initial research on next generation CMOS structures such as multi-gate devices (MuGFETs, FinFETs, etc.) indicates that BTI remains a problem in future CMOS technologies (Groeseneken et al. 2008; Wang et al. 2011; Feijoo et al. 2012).

When time-dependent voltage stress is applied, a peculiar property of the BTI mechanism is revealed: the so-called relaxation or recovery of the degradation immediately after the stress voltage has been reduced (see Fig. 2.10) (Kaczer et al. 2008). This phenomenon greatly complicates the evaluation of BTI, its modeling, and the extrapolation of its impact on circuits. It currently appears that BTI degradation does not fully recover when the stress is removed, hence leaving a permanent residual degradation. BTI degradation can therefore be modeled as a combination of a permanent and a recoverable degradation component (Grasser and Kaczer 2009; Maricau et al. 2011):

$$\Delta V_{TH} \propto \left[\underbrace{\exp\left(\alpha_1 V_{GS}\right) t^{n_P}}_{\text{Permanent Part}} + \underbrace{V_{GS}^{\alpha_2}\left(C_R + n_R \log_{10}(t)\right)}_{\text{Recoverable Part}} \right] \exp\left(-\frac{E_a}{kT}\right) \tag{2.6}$$

where ΔV_{TH} is a function of the transistor gate-oxide electric field (E_{ox}) and the temperature (T). Further, α_1, α_2 are technology-dependent voltage scaling factors, E_a is the activation energy, C_R, n_P and n_R are the time exponents for the permanent

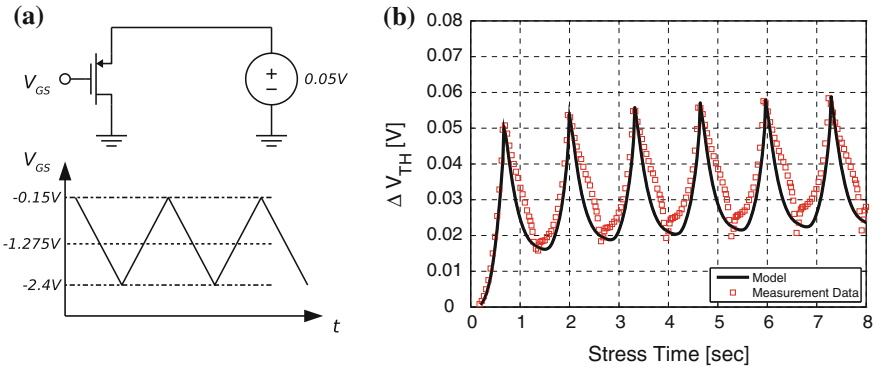

Fig. 2.10 The time-dependent V_{TH} shift of a pMOS transistor subjected to a triangle-shaped stress voltage (see **a**). Part of the NBTI damage is recovered, every time the stress is reduced (see **b**) (Maricau et al. 2011)

and recoverable part and k is the Boltzmann constant. Note how Eq. (2.6) is only valid for a fixed stress voltage, an accurate BTI model for time-varying stress voltages is given in Chap. 3. Also, it is important to note that BTI is shown not to be frequency-dependent (i.e. at least for measurements up to 3 GHz) (Sasse 2008; Ramey et al. 2009). Further, BTI drain bias dependency has also been observed (Schlünder et al. 2005; Luo et al. 2007).

BTI effects in large micrometer-sized transistors are typically considered deterministic (Wang et al. 2007; Maricau et al. 2011). The application of a given stress on matched transistors therefore results in an identical shift of the transistor parameters. Scaling transistors down to nanometer dimensions, however, gradually changed these deterministic effects into stochastically distributed failure mechanisms due to an ever-increasing impact of individual trapping and detrapping events (Kaczer et al. 2010, 2011) (see Fig. 2.11). At device level this results in a time-dependent shift of the transistor parameters ($\Delta V_{TH} = f(t)$) augmented with a time-dependent increase of the standard deviation on these parameters ($\sigma(V_{TH}) = g(t)$). Initially matched transistors, processed in ultra-scaled nanometer CMOS technologies, can therefore cause circuit performance failure resulting from increased time-dependent transistor mismatch (Gielen et al. 2011).

Electromigration

Electromigration (EM) is an aging effect taking place in interconnect wires, contacts and vias in an integrated circuit (Tu 2003). The effect causes material transport by gradual movement of the ions in a conductor due to the momentum transfer between conducting electrons and the diffusing metal atoms. EM is important in applications where high direct current densities are used. Integrated circuits are very prone to this

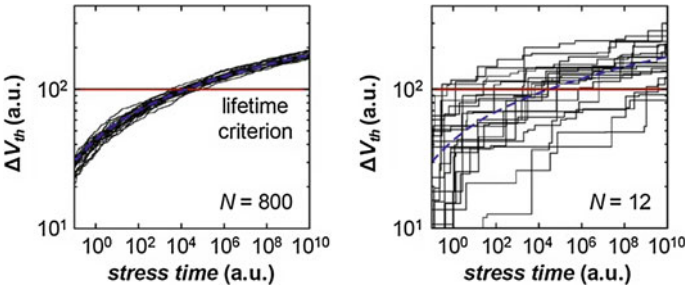

Fig. 2.11 Due to the discrete nature of trapping and detrapping events, the time-dependent BTI V_{TH} shift becomes stochastic for small devices in sub-45 nm CMOS. The lifetime of a large transistor with 800 defects (*left*) is much better defined than the lifetime of a small transistor with only 12 defects (*right*) (Kaczer et al. 2011)

effect, since current densities in excess of 1e5A/cm^2 are being measured (Tu 2003; Lewyn et al. 2009). A typical household extension cord carries only about 1e2A/cm^2 as it is limited by Joule heating rather than electromigration. The electromigration phenomenon is already known for over 100 years, but first became a practical problem in 1966 when the first integrated circuits became commercially available. Figure 2.12 illustrates how the gradual shift of the metal can create a void (open) or a hillock (short) which can potentially cause circuit failure.

In a homogeneous crystalline structure, there is hardly any momentum transfer between the conducting electrons and the metal ions. However, at the grain boundaries, this homogeneity does not exist and the conducting electrons have a large impact on the metal ions. This causes atoms to become separated from the grain boundaries and to be transported in the direction of the current, along the grain boundaries (Jerke and Lienig 2004). The mean time-to-failure (MTTF) of a wire, when subjected to electromigration, can be expressed by Black's law (Black 1969):

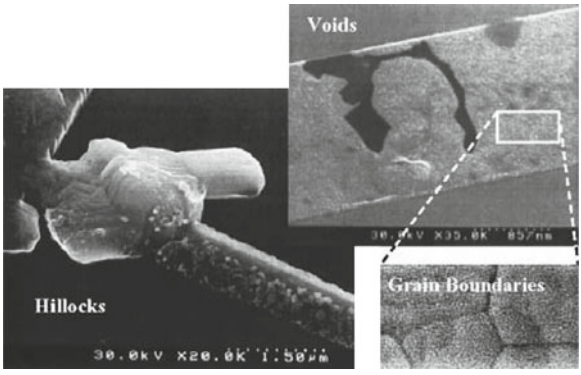

Fig. 2.12 Hillock and void formations in wires due to electromigration (Jerke and Lienig 2004)

$$\mathrm{MTTF} = \frac{A}{J^n} \exp\left(\frac{E_\mathrm{a}}{kT}\right) \tag{2.7}$$

with A a constant dependent on the cross-sectional area of the interconnect wire, J the current density, E_a the activation energy (e.g. 0.7 eV for aluminum), k the Boltzmann constant, T the temperature and n a scaling factor (typically $n = 2$). Note how, besides the current density, the temperature also strongly affects the lifetime of the wire. For an interconnect to remain reliable at high temperatures, the maximum current density must decrease. EM is a very layout dependent phenomenon. The MTTF of a wire does not only depend on the width of the wires, but particular attention must also be paid to vias and contact holes. Since the ampacity of a (tungsten) via is less than of a metal wire, a via is more prone to EM compared to a wire with the same dimensions. Where needed, multiple vias must therefore be used. Also, these vias must be organized such that the current is distributed evenly through all the vias. Additionally, 90-degree corner bends in wires must be avoided, since the current density in such bends is higher than that in oblique angles (Jerke and Lienig 2004; Lienig and Jerke 2005).

In older technologies, aluminum was commonly used as a conducting material for interconnect wires. Aluminum has a good conductivity and a good adherence to the silicon substrate. However, aluminum is very susceptible to electromigration. Research indicated how adding 1–2 % of copper to aluminum increased the resistance to EM about 50 times. This effect is attributed to the fact that copper inhibits the diffusion of atoms along the grain boundaries (Tu 2003). Due to the further scaling of CMOS technologies, a need for a better interconnect conductor than Al(Cu) (having a lower resistance-capacitance delay) was needed. Therefore the industry has turned to full-copper interconnect wires. Copper has a much higher melting point than aluminum and therefore atomic diffusion should be much slower in copper than in aluminum. So, electromigration should be much less in copper interconnects. Surprisingly, the benefit is not as big as expected, and when compared to Al(Cu) wires, copper wires have a lower MTTF. As a solution, tin has been found very effective in retarding electromigration in copper (Tu 2003). However, electromigration still remains a major problem in nanoscale CMOS circuits today (Zhang et al. 2010).

2.4.2 Transient Effects

Transient unreliability effects distort the normal operation of a circuit for a limited time period. Typically the term signal integrity (SI) is used to describe how the quality of a signal in an electronic system changes under these transient effects. Is the signal properly transferred from one subcircuit to the next and what is the quality of the signal at the circuit output? A good-quality signal guarantees high-speed reliable data transfer within a system and between different systems. A signal waveform in an integrated circuit can be distorted by two types of unwanted signals: noise and interference.

Noise

Noise is an unwanted and random perturbation of a signal and results from active or passive devices (e.g. thermal noise, flicker noise, popcorn noise, etc.) within the circuit itself. Noise is bounded by physical limitations and influenced by the fabrication process (technology, device selection, processing quality, etc.). Since device noise determines the minimum detectable signal level, the operation of analog circuits in particular is very prone to these noise sources. Noise is typically modeled as an input-referred noise source, determined from circuit noise analysis and quantified using the noise figure (NF) and signal-to-noise ratio (SNR) parameters. Noise is the ultimate limit to performance in electronic circuits.

Electromagnetic Interference

Electromagnetic interference (EMI) is defined as the influence of unwanted signals generated by a source circuit and picked up by a receptor or victim circuit, affecting its signal performance and quality. The coupling path between the source and the victim circuit can be conductive, capacitive, magnetic or radiative. A coupling path can also consist of two or more of these coupling mechanisms working together. As opposed to noise, an electromagnetic signal has a source or origin external to the (sub)circuit it affects. The source signal can be deterministic (man-made) or random (natural). Examples of natural electromagnetic interference sources are atmospheric noise (e.g. produced by lightning during thunderstorms) and cosmic noise. Man-made interference signals can be functional signals, which are generated during the normal operation of the source circuits, or accidental signals. Examples of man-made interference are on-chip crosstalk and simultaneous switching noise (functional EMI caused by other circuits that are part of the system) and mobile devices, engine ignitions and microwave ovens (accidental EMI caused by unrelated external sources) (Redoute 2009; Loeckx 2010).

1. **On-chip crosstalk** between two circuits or circuit elements (e.g. interconnect wires) is defined as a deviation from the ideal signal waveform propagating in the victim circuit, caused by the influence of signal transitions in the source circuit. Basically, three types of parasitic coupling may result in crosstalk: electric field coupling, magnetic field coupling and common impedance coupling. The latter occurs when multiple current paths share the same conductor. If the source circuit generates noise in the conductor, in the form of alternating current, the voltage over the finite-impedance conductor is modulated. Therefore, the current going to the victim circuit is also modulated and the operation of the victim circuit might be affected. Electric field coupling, on the other hand, results from capacitive coupling between different interconnect nets. In the same way, magnetic coupling can be modeled by coupled inductors (Redoute 2009).
2. **Simultaneous switching noise (SSN)** is a particular case of common impedance crosstalk when subcircuits on the same IC share the same power distribution bus.

This phenomenon is also known as power/ground noise, ground bounce, substrate noise or dI/dt noise when the power/ground signal is connected with a bondwire. Simultaneous switching of multiple digital gates produces large transient current spikes which flow through the power and ground lines of the chip (Redoute 2009). In case of a mixed-signal circuits, SSN is the primary source of substrate noise, where interference generated by the digital circuit influences the operation of a neighboring and potentially sensitive analog circuit (Donnay and Gielen 2003; Stefanou 2011).

3. **Energetic particles** such as alpha and gamma particles result in ionized radiation of the semiconductor material and potentially cause a non-destructive change in the state of CMOS devices. Naturally occurring alpha particles, impinging on the transistors in a circuit, generate electron-hole pairs in several picoseconds. The charges generated in or near the depletion region are separated by the oxide electric field. These particles are particularly dangerous for storage devices or memories as they initiate state changes, resulting in a soft error or single event upset (SEU).
4. **Radiated EMI** is a form of electromagnetic interference where a remote source (e.g. another circuit, a cellphone antenna, a running engine or a microwave oven) becomes an unintentional transmitter of electromagnetic waves that are picked up by the victim circuit. The receiver antenna in the victim circuit can be formed by PCB traces, connection cables or wire loops in the integrated circuit. Once picked up, the irradiated signal can disturb the normal operation of the victim circuit.

Typically, a circuit is subjected to various sources of electromagnetic interference at the same time. The power and frequency spectrum of these interference sources can also vary with the environment, the temperature and the circuit workload (e.g. in case of interconnect crosstalk or substrate noise). It is therefore, at design time, very hard to predict the impact of unwanted interference signals on the operation of the circuit. To guarantee reliable circuit operation, electromagnetic compatibility (EMC) regulations for both emission of (EME) and susceptibility to (EMS) interference signals are used. Each circuit, depending on the field of application, must comply to these rules. The international electrotechnical commission (IEC), for example, is one of the international standards organizations which are addressing the need for standardized IC EMC test methods, such as the IEC 61000 standard (IEC 61000 structure 2012).

2.5 Conclusions

This chapter has briefly reviewed the major unreliability effects affecting the correct operation of circuits integrated in a nanometer CMOS technology. A distinction between spatial and temporal unreliability effects has been made. The former are visible right after fabrication and include random dopant effects, line edge roughness and oxide thickness variations as dominant phenomena in sub-65 nm CMOS.

Temporal unreliability effects are time-dependent and include aging effects (e.g. bias temperature instability and electromigration) and transient effects (e.g. noise and electromagnetic interference). The description of the effects in this chapter, though being far from complete, has summarized the most important aspects of CMOS reliability and additional references have been provided for the interested reader. The remainder of this work focuses on temporal transistor aging effects, although spatial random effects are also included in the circuit reliability simulator proposed in Chap. 5.

http://www.springer.com/978-1-4614-6162-3

Analog IC Reliability in Nanometer CMOS

Maricau, E.; Gielen, G.
2013, XVI, 198 p., Hardcover
ISBN: 978-1-4614-6162-3