Chapter 2
Reliability, Faults, and Fault Tolerance

A clear understanding of several concepts and terminology related to reliability is needed to proceed with the understanding of the methodologies which are applied to guarantee optimal operability of VLSI systems, fault tolerance, and circuit architectures implementing them. Basic terms such as reliability, fault tolerance, faults, and fault modeling are introduced and explained in detail. The chapter is organized as follows. In Section 2.1, the general concepts of reliability, fault tolerance, and yield are explained. Faults and fault models are presented in Section 2.2. A realistic transistor fault model adapted to current CMOS technology is presented in Section 2.3.

2.1 Reliability and Fault Tolerance

Reliability is defined according to *IEEE* as the ability of a system or component to perform its required functions under stated conditions and for a specified period of time. The process yield of a manufacturing process is defined as the fraction, or percentage, of acceptable parts among all parts that are fabricated [17]. A system failure occurs or is present when the *service* provided by the system differs from the specified service or the service that should have been offered. In other words, the system *fails* to perform what it is expected to.

In classical theory [18, 19], the reliability $R(t)$ is defined as the probability of a system to operate correctly during the time interval $[0,\ t]$, given that it has been operative at time 0. Let $F(t) = P\{T \leq t\}$ be the probability that a failure occurs at a time T, smaller than or equal to t, then

$$F(t) = \int_{-\infty}^{t} f(t)dt, \tag{2.1}$$

where $f(t)$ represents the probability density function (PDF) of the random variable, time to failure. $R(t)$ represents the probability that a system has not failed by time t, which is expressed as $R(t) = P\{T > t\}$, and consequently

M. Stanisavljević et al., *Reliability of Nanoscale Circuits and Systems*,
DOI 10.1007/978-1-4419-6217-1_2, © Springer Science+Business Media, LLC 2011

$$R(t) = 1 - F(t). \tag{2.2}$$

The failure rate λ represents the probability that a failure occurs within a time interval $[t_1,\ t_2]$, given that it has not occurred prior to t_1. In electronic systems, λ can legitimately be considered constant, and in this case,

$$R(t) = e^{-\lambda t}. \tag{2.3}$$

Finally, the mean time to failure (MTTF) is expressed as the expected value of time to failure and is derived as

$$\text{MTTF} = \int_0^\infty R(t)dt, \tag{2.4}$$

and upon constant failure rate,

$$\lambda = \frac{1}{\text{MTTF}}. \tag{2.5}$$

The so-called bathtub curve which is shown in Fig. 2.1 is widely accepted to represent a realistic model of the failure rate of electronic equipment and systems over time [20]. The bathtub curve consists of three characteristic zones. Failure rates follow a decreasing pattern during the early times of operation, where infant mortality deteriorates the system, typically due to oxide defects, particulate masking defects, or contamination-related defects. Failure rate remains constant over the major part of the system operation life. Failures are random, mostly manifesting themselves as soft errors. Wearout occurs in the final stage of the system lifetime, where failure rate increases, typically due to electromigration-related defects, oxide wearout, or hot carrier injection.

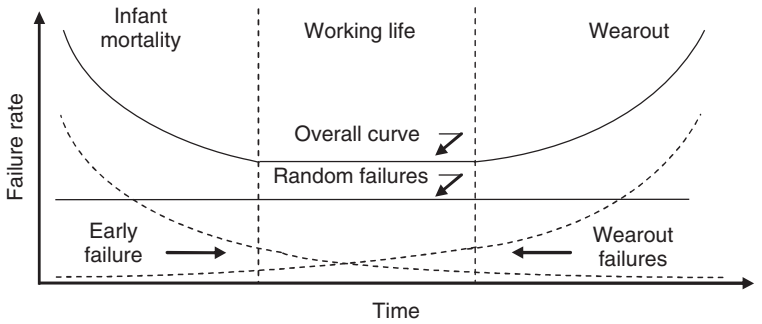


Fig. 2.1 Bathtub curve; the time axis is not to scale ([21], with kind permission of Springer Science and Business Media). The hard curve shows cumulative contributions of its three components that are presented as the dotted curves named Early failure and Wearout failures and the solid line named Random failures

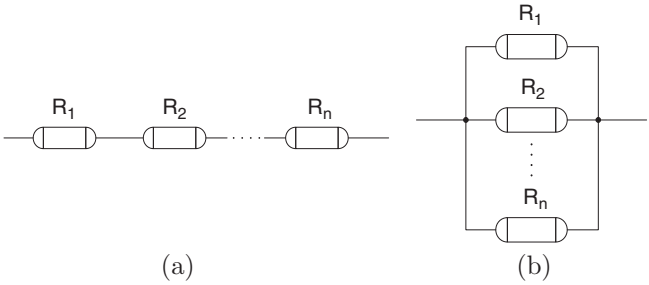


Fig. 2.2 Electrical component configurations: **(a)** serial and **(b)** parallel

Some major architectural configurations of electronic systems are very common, and the analysis of their reliability behavior forms the foundation of the analysis of any complex system. In the serial configuration, depicted in Fig. 2.2a, several blocks, n, with failure rates $R_1(t), \ldots, R_n(t)$ considered independent of each other are cascaded. The correct operation of the system depends on the reliability of each block and is mathematically expressed as

$$R_{\text{system}} = R_1(t) \cdot R_2(t) \cdot \cdots \cdot R_n(t) = \prod_{i=1}^{n} R_i(t). \tag{2.6}$$

In the parallel configuration, depicted in Fig. 2.2b, malfunction of all composing blocks is necessary to cause the system to fail. Naming the probability of failure or unreliability of the components $Q_i = 1 - R_i$ and omitting the expression of time (t) for clarity, the probability of failure of the system is expressed as

$$Q_{\text{system}} = \prod_{i=1}^{n} Q_i. \tag{2.7}$$

The reliability of the system composed of parallel implementation is expressed as

$$R_{\text{system}} = 1 - Q_{\text{system}} = 1 - \prod_{i=1}^{n} (1 - R_i) \tag{2.8}$$

and can be higher than the reliability of individual components if redundancy is applied. Realistic designs are typically composed of hybrid arrangement of parallel and serial configurations, where the system reliability can be obtained by iterative decomposition of the network into its series and parallel components and step-by-step solving.

Finally, a system in a k-out-of-n configuration consists of n components. Only k components need to function properly to enable the full system to operate.

A system which has the ability to deliver the expected service operation despite the occurrence of faults or the presence of defects is named fault tolerant. Fault tolerance of microelectronic systems is presented in detail in Chapter 4.

2.2 Faults and Fault Models

The following three terms are crucial and related to system failure and thus need to be clearly defined, which are named *defect*, *error*, and *fault* [17].

A defect in an electronic system is the unintended difference between the implemented hardware and its intended design. Some typical defects of VLSI chips include [22]

- process defects, taking the form of missing contact windows, parasitic transistors, oxide breakdown, etc.;
- material defects, due to bulk defects (cracks, crystal imperfections), surface impurities, etc.; and
- age defects, taking the form of dielectric breakdown, electromigration, etc.

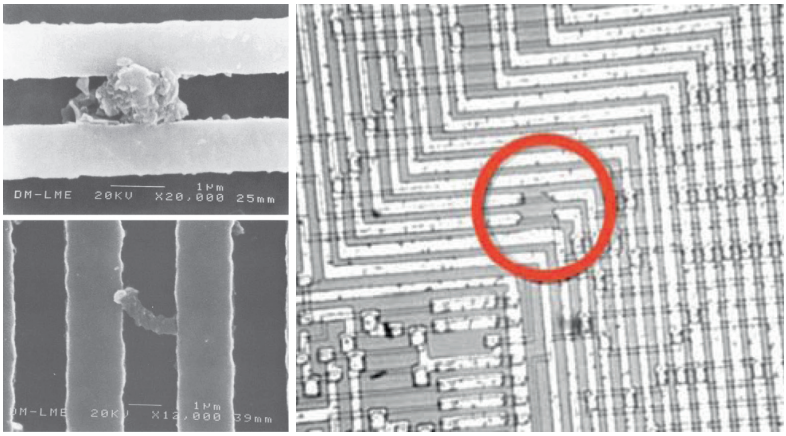
Defects can be also classified by the statistical effect they produce:

- Systematic, defects that have the same impact across large dimensions, such as die or wafer, and that can be modeled in a systematic way. These defects are usually the result of process–design interaction.
- Random (stochastic), all types of defects that cannot be controlled or modeled in a predictable and systematic way. They include random particles in the resist or in the materials, inserted or removed, or defects in the crystal structure itself that alter the intended behavior of the material and results in excessive leakage or in a shift in the device threshold (V_{th}), eventually causing the failure of the device. The failure modes resulting from these defects are

  1. Opens
  2. Shorts
  3. Leakage
  4. V_{th} shift
  5. Variability in mobility (μ)

Random defects do not necessarily result in a complete failure of the device, but in a significant deterioration of its performance [23].

Some classical microphotographs of defects are presented in Fig. 2.3. The patterns are easily recognizable and are presented as illustrative cases. Visual inspection cannot be applied in the detection of defects in modern digital systems, consisting of hundreds of billions of transistors and their interconnections routed over nine metal layers. Test techniques are applied [17], which form a discipline of its own. The application of test techniques and enabling the testability of complex digital systems (design for testability) imposes the designers additional constraints in terms of classical circuit specifications (area, delay). In addition, any fault-tolerant technique



(a) (b)

Fig. 2.3 Defect images. **(a)** Bridging defects with low-resistance electrical behavior on the *top* and high-resistance electrical behavior on the *bottom* microphotograph ([24], with kind permission of Springer Science and Business Media) and **(b)** open defect inside the *circle* ([25], with kind permission of Springer Science and Business Media)

which is implemented at the hardware level must be proven compliant with the test methodologies, which may be a difficult task.

The existence or emergence of defects reduces yield.

A wrong output signal produced by a defective system is called an error. An error is an effect whose cause is some defect. Errors can be classified into three main groups, namely permanent, intermittent, and transient errors, according to their stability and concurrence [26].

- Permanent errors are caused by irreversible physical changes in a chip. The most common sources for this kind of errors are the manufacturing processes. Permanent errors can also occur during the usage of the circuit, especially when the circuit is old and starts to wear out. Common to all permanent errors is that once they have occurred, none will vanish and consequently the test to detect them can be repeated, conducting to identical results. Permanent errors are also known under the denomination of hard errors.
- Intermittent errors are occasional error bursts that usually repeat themselves every now and then and are not continuous as permanent errors. These errors are caused by unstable or aging hardware and are activated by an environmental change such as a temperature or voltage change. Intermittent errors often precede the occurrence of a permanent error; for instance an increased resistance in a wire may be observed before it cracks, creating an open circuit. Intermittent errors are very hard to detect because they may only occur under certain environmental constraints or in the presence of some specific input vector combination.
- Transient errors are temporal single malfunctions caused by some temporary environmental conditions which can be an external phenomenon such as radiation or noise originating from other parts of the chip. Transient errors do not leave any

permanent marks on the chip and therefore they are also called soft errors (SE). A common manifestation of transient error is a change of the binary value of a single bit (e.g., a bit flip in memory cell). Another term, single-event upset (SEU), is used for soft error, which describes the fact that malfunctions (upsets) are commonly caused by single events such as an absorbed radiation. The occurrence of transient errors is commonly random and therefore hard to detect.

Error sources can be classified according to the phenomenon causing the error. Such origins are for instance related to the manufacturing process, physical changes during operation, internal noise caused by other parts of the circuit, and external noise originating from the chip environment.

A fault is a representation of a defect at the abstracted functional level. A fault is present in the system when physical difference is observed between the "good" or "correct" system and the actual system. Discussions presented in this book mostly relate to permanent faults caused by physical defects.

The most common faults in a chip are spots and bridging faults caused by silicon impurities, lithography, and process variations [27]. These faults cause permanent errors in a circuit. The probability of these defects is likely to increase with technology scaling, as larger numbers of transistors are integrated in a single chip and the size of chips increases, while devices and wires sizes decrease. This results in a decreasing yield, and consequently higher price, of functioning chips. The move toward nanoscale circuits also raises a list of new problems originating from the manufacturing process. As the fabrication dimensions shrink, the proportional extent of deviations becomes larger and their effects more severe. Lithography deviation is the main cause of gate length deviations. Moreover, fluctuations of the doping profile in turn cause deviations of the transistor threshold voltage. These effects, together with the increase of resistive vias and contacts, eventually result in large operation speed deviation. Simultaneously, the operation frequency of integrated circuits is expected to increase. The worst-case scenario consisting of a series configuration of "slow" devices may lead to timing violations and therefore to malfunction of the circuit. This is considered an intermittent error because the circuit may correctly operate most of the time; this would not be the case for permanent errors.

The diminishing value of reliability of very deep submicron technologies is an established fact. Moreover, it is widely accepted that nanoelectronic-based systems will rely on a significantly lower reliability rate than what was known so far. More details of challenges and faults in nanodevices are given in the following chapter.

If error detection and recovery do not take place in a timely manner, a *failure* can occur that is manifested by the inability of the system to provide a specified service. *Fault tolerance* is the capability of a system to recover from a fault or error without exhibiting failure. A fault in a system does not necessarily result in an error; a fault may be latent in that it exists but does not result in an error; the fault must be sensitized by a particular system state and input conditions to produce an error. The techniques related to fault-tolerant systems include fault avoidance, fault masking, detection of erroneous or compromised system operation, containment of error propagation, and recovery to normal system operations [28].

Actual defects in a circuit cannot be directly considered in the design and validation of the circuit and therefore special fault models are needed. Fault models are simplifications of the phenomena caused by defects on the circuit and were first introduced by Eldred in the late 1950s [29]. Fault models have been developed at each level of abstraction, i.e., behavioral, functional, structural, switch, and geometric levels.

In this book, we limit our discussions to switch-level and geometric fault models. The higher level abstraction models do not offer a satisfying level of accuracy, which is required to study and apply fault-tolerant techniques further assessed. This comment also covers *stuck-at* (permanent connection of the gate input or the output to supply lines) and *von Neumann* fault models (which consists of transient bit-flip faults at the gates and interconnects [13]) which belong to structural fault models. Even though the stuck-at fault model is the most popular and widely used model in industry, which has the ability to detect a majority of physical defects, it is not adequate for accurate reliability evaluation in modern technologies [30, 31]. These referenced papers show that approximating the gate probabilities of failure by (bounding) constants introduces sizable errors, leading to overdesign. Moreover, stuck-at fault models will not be suitable for future nanodevices as demonstrated on the example of single-electron transistor (SET) circuits by Beiu et al. [31].

Switch-level fault models are defined at the transistor level. The most prominent fault models in this category are the *stuck-off/stuck-open* and *stuck-on/stuck-short* fault models. If a transistor is permanently in non-conducting state due to a fault, it is considered to be stuck-off or stuck-open. Similarly, if a transistor is permanently in conducting state, it is considered to be stuck-on or stuck-short. These fault models are specially suited for the CMOS technology.

Geometric fault models assume that the layout of the chip is known. For example, knowledge of line widths, inter-line and inter-component distances, and device geometries is used to develop these fault models. At this level, problems related to the manufacturing process can be detected. The layout information, for example, can be used to identify lines or components that are most likely to be shorted due to process defects. The *bridging* fault model leads to accurate detection of realistic defects. With shrinking geometries of VLSI chips, this model becomes increasingly important.

A new model for CMOS technologies that combines the benefits of switch-level and geometric fault models has been developed and is presented in Section 2.3. The model exhibits much better accuracy than typical switch-level models, while exhibiting a complexity comparable to switch-level models. Moreover, a simple fault model for SET has been developed and is used in simulations and results presented in Section 6.4.

2.3 Transistor Fault Model

A major step in any design automation process consists of simulation. In order to perform a simulation for reliability, an accurate and realistic fault model is necessary. Considering permanent errors as the main and most intricate source

of unreliability, physical defects and fault modes are modeled with a netlist fault description. There are various ways of modeling physical defects, at various levels of abstraction, as presented in Section 2.2. Geometrical models that are close to the physical layout are complex and impractical in large-scale simulations. However, they are the most accurate. Statistical models related to physical defects distribution are not hard to embed into circuit-based analysis. The stuck-at approach which is traditionally used in fault coverage analysis is not sufficient to handle the analysis of various faults in nanometer-scale devices. The following two basic approaches are a starting point for our model, namely inductive fault analysis (IFA) [32] and transistor-level fault modeling [33], both of which have complex implementations.

The transistor-level fault modeling is applied at an abstraction level above the physical layout and can be classified as a switch-level fault model. It usually incorporates only stuck-on and stuck-off models of transistors for representing faults. These models represent a very reduced set of possible physical defects and therefore they are not sufficient. On the other hand, the IFA approach, which is a geometric fault model, has some drawbacks, mainly related to high computational complexity of the used tools, complete dependency on geometrical characteristics, and difficulty of properly handling analog layouts. Our model provides better accuracy that is comparable to IFA models and is operated with a time complexity comparable to switch-level models.

A hierarchical transistor fault model is developed in order to overcome shortfalls of transistor-level fault modeling using some results from the IFA approach and also to cover a range of impacts as wide as possible that device faults have on the circuit behavior. The fault model consists of two layers (Fig. 2.4). The upper layer (LY2) models various physical defects such as missing spot, unwanted spot, Gate oxide short (GOS) with channel, floating gate coupled to a conductor, and bridging

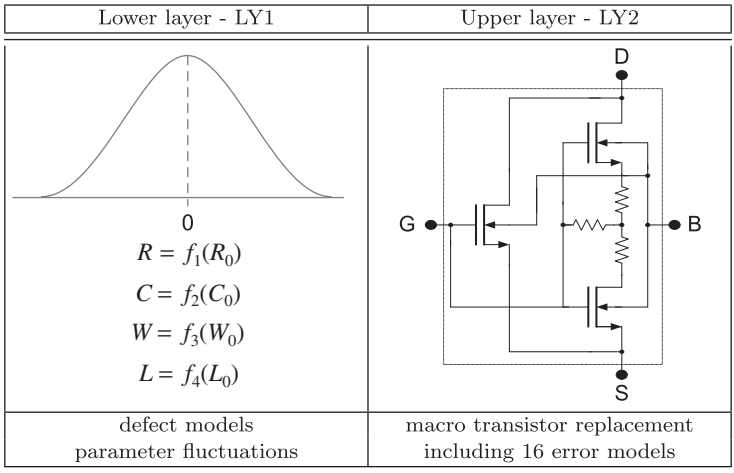


Fig. 2.4 Two-layer fault model

faults [34, 35]. Some of the physical defects are depicted in Fig. 2.3. The models have been developed from structural and lithography defects, and each defect model is described in terms of electrical parameters of its components. Thus, for simulation purposes, physical defects are translated into equivalent electrical linear devices such as resistors, capacitors and nonlinear devices such as diodes and scaled transistors. A total of 16 possible defects are considered for each transistor, which are listed in Table 2.1. The number of implemented defective transistor equivalent circuits is nine, while seven of them are available in two implementations, i.e., for high and low values of defect model parameters. All defective transistor equivalent circuits (for open drain, open source, floating gate, drain–source short, drain–gate short, gate–source short, drain–bulk short, source–bulk short, and gate oxide short) are depicted in Fig. 2.5a–i. Opens and shorts are modeled as a resistance which is placed in parallel with a capacitance on the spot of a defect [33, 36]. The floating gate (Fig. 2.5c) is modeled as a capacitive divider between the gate terminal and source [35, 37]. Gate oxide short (GOS; Fig. 2.5i) is modeled by dividing the gate area into three equivalent transistors: two are in a series configuration and are placed in parallel with the third one, with a common node at the location of the physical gate oxide short spot [35, 38].

Table 2.1 List of transistor failures modeled in the upper layer (LY2)

Acronym	Failure type
DHO	*D*rain *H*ard *O*pen, resulting in stuck-off fault
DSO	*D*rain *S*oft *O*pen, resulting in partial stuck-off fault
SHO	*S*ource *H*ard *O*pen, resulting in stuck-off fault
SSO	*S*ource *S*oft *O*pen, resulting in partial stuck-off fault
FLG	*FL*oating *G*ate resulting in disconnected input
DSHS	*D*rain *S*ource *H*ard *S*hort, resulting in stuck-on fault
DSSS	*D*rain *S*ource *S*oft *S*hort, resulting in partial stuck-on fault
DGHS	*D*rain *G*ate *H*ard *S*hort, resulting in input–output bridging fault
DGSS	*D*rain *G*ate *S*oft *S*hort, resulting in partial input–output bridging fault
GSHS	*G*ate *S*ource *H*ard *S*hort, resulting in input stuck-at fault
GSSS	*G*ate *S*ource *S*oft *S*hort, resulting in partial input stuck-at fault
DBHS	*D*rain *B*ulk *H*ard *S*hort, resulting in excessive current flowing through the substrate
DBSS	*D*rain *B*ulk *S*oft *S*hort, resulting in partial excessive current flowing through the substrate
SBHS	*S*ource *B*ulk *H*ard *S*hort, resulting in current flowing through the substrate only for non-common sources
SBSS	*S*ource *B*ulk *S*oft *S*hort, resulting in small current flowing through the substrate only for non-common sources
GOS	*G*ate *O*xide *S*hort, resulting in an excessive current flowing through the gate oxide insulator

The lower abstraction model layer (LY1) consists of defective transistor circuit model parameters (e.g., resistances R, capacitances C, and geometric parameters gate length L, gate width W for gate-oxide short model) whose variation can have a significant influence on the defect model. Here, each parameter is modeled with the Normal distribution – $\mathcal{N}(\mu, \sigma)$, with a nominal mean value (μ) and a given

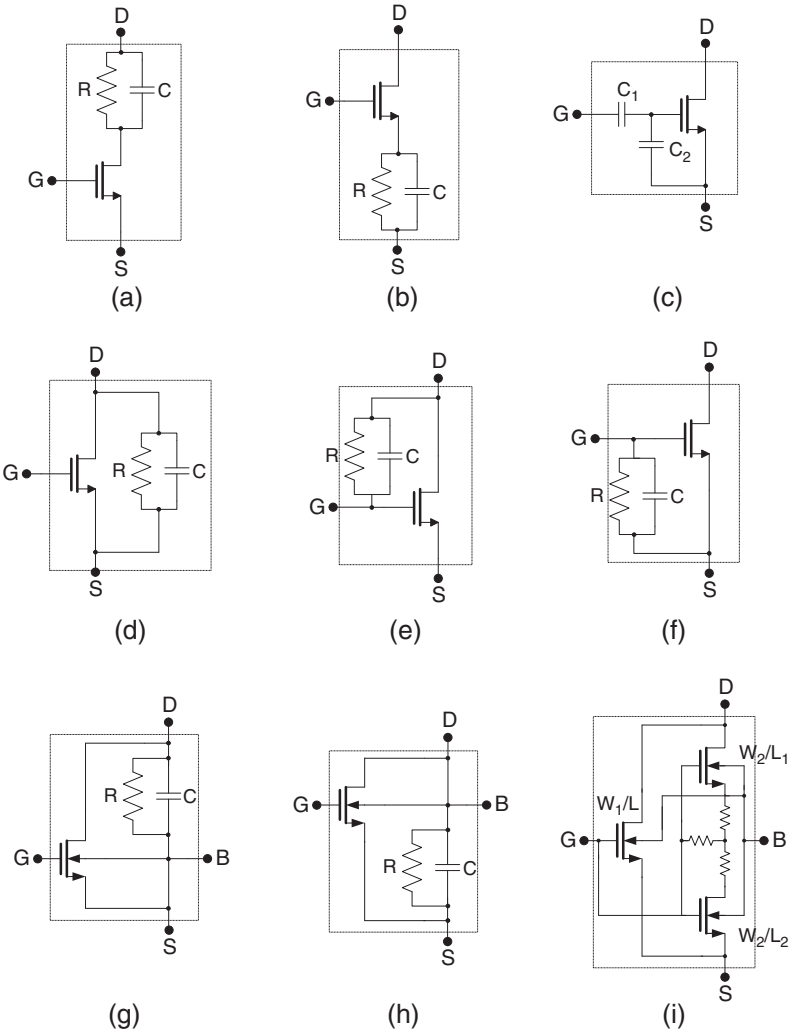


Fig. 2.5 Transistor equivalent defect models: **(a)** open drain, **(b)** open source, **(c)** floating gate, **(d)** drain–source short, **(e)** drain–gate short, **(f)** gate–source short, **(g)** drain–bulk short, **(h)** source–bulk short and **(i)** gate oxide short

standard deviation (σ). Nominal parameter values of R have been chosen according to [35, 37] to 1 and 5 kΩ resistance for hard and soft short defects, respectively, and to 100 and 0.5 MΩ for hard and soft opens, respectively.

An extraction of actual or realistic values of these parameters requires an access to the fabrication process parameters and test parameters that are usually kept confidential within the process manufacturer. However, some of the parameters may be extracted by means of building and measuring different testing structures on test

chips. Some results have been presented in the comprehensive literature related to bridging faults [36, 37], resistive opens and shorts [39], and transistor gate geometrical parameters [40]. One possible test structure for extracting drain/source open resistance is illustrated in Fig. 2.6 and consists of an array of multiple transistors connected in series, and uniformly distributed over the chip, with the possibility of measuring the current flowing through each line. Here, IDDQ testing (which relies on measuring the supply current (I_{DD}) in the quiescent state) with the respective data from the process manufacturer regarding the probability of drain/source open could provide a means of extracting the nominal value of the resistance parameter.

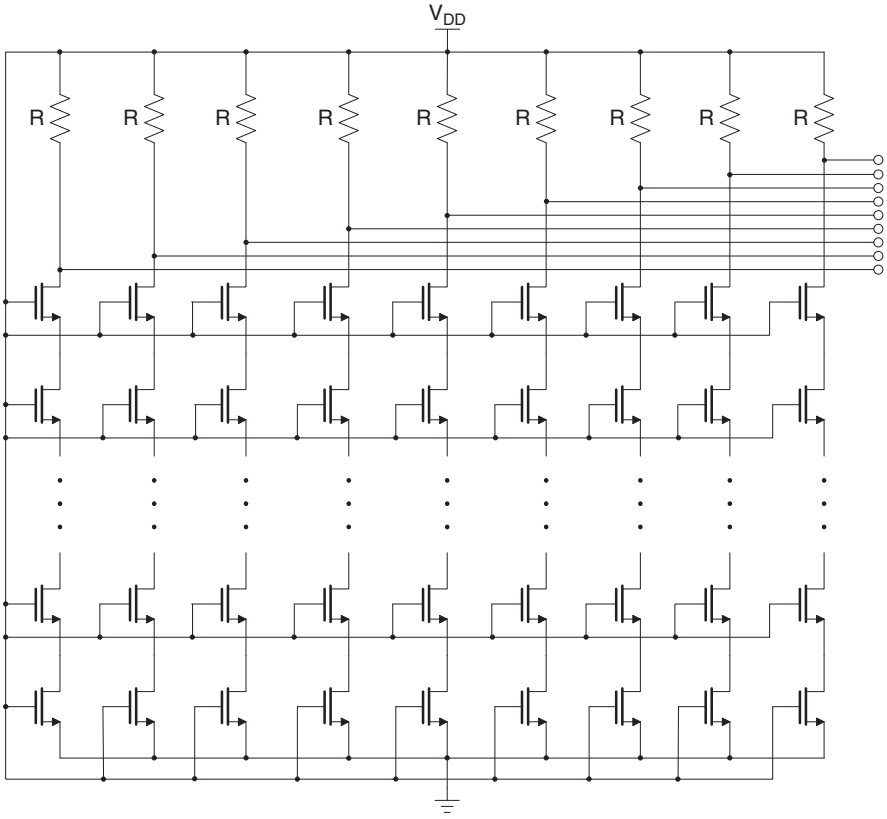


Fig. 2.6 Test structure for measuring drain/source open resistance parameter

The layer that represents the mapping of interconnection defects into their electrical models (open spots and bridging faults) [36] is not included in the defect models and simulations. Modeling of interconnection defects at system level is highly dependent on geometrical characteristics of the layout, where maintaining the correspondence between the physical and electrical parameters remains a problem that

needs to be solved. In the transistor-level simulations, this layer can be excluded, considering that more than 80% [41] of signal errors in modern circuits are due to global signals stuck-at supply or ground.

The transistor-level model presented in this section will be widely used in reliability simulations throughout the course of this book.