Testing of Digital Systems

N. K. Jha and S. Gupta

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014, Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2003

Printed in the United Kingdom at the University Press, Cambridge

Typeface Times 10.5/14pt. *System* LaTeX 2_ε [DBD]

A catalogue record of this book is available from the British Library

ISBN 0 521 77356 3 hardback

Contents

1 Introduction

by Ad van de Goor

We introduce some basic concepts in testing in this chapter. We first discuss the terms fault, error and failure and classify faults according to the way they behave over time into permanent and non-permanent faults.

We give a statistical analysis of faults, introducing the terms failure rate and mean time to failure. We show how the failure rate varies over the lifetime of a product and how the failure rates of series and parallel systems can be computed. We also describe the physical and electrical causes for faults, called failure mechanisms.

We classify tests according to the technology they are designed for, the parameters they measure, the purpose for which the test results are used, and the test application method.

We next describe the relationship between the yield of the chip manufacturing process, the fault coverage of a test (which is the fraction of the total number of faults detected by a given test) and the defect level (the fraction of bad parts that pass the test). It can be used to compute the amount of testing required for a certain product quality level.

Finally, we cover the economics of testing in terms of time-to-market, revenue, costs of test development and maintenance cost.

1.1 Faults and their manifestation

This section starts by defining the terms failure, error and fault; followed by an overview of how faults can manifest themselves in time.

1.1.1 Failures, errors and faults

A system **failure** occurs or is present when the *service* of the system differs from the specified service, or the service that should have been offered. In other words: the system *fails* to do what it has to do. A failure is caused by an *error*.

There is an **error** in the system (the system is in an erroneous state) when its *state* differs from the state in which it should be in order to deliver the specified service. An *error* is caused by a *fault*.

A **fault** is present in the system when there is a *physical difference* between the 'good' or 'correct' system and the current system.

Example 1.1 A car cannot be used as a result of a flat tire. The fact that the car cannot be driven safely with a flat tire can be seen as the *failure*. The failure is caused by an *error*, which is the erroneous state of the air pressure of the tire. The *fault* that caused the erroneous state was a puncture in the tire, which is the physical difference between a good tire and an erroneous one.

Notice the possibility that a fault does not (immediately) result in a failure; e.g., in the case of a very slowly leaking tire. □

1.1.2 Fault manifestation

According to the way faults manifest themselves in time, two types of faults can be distinguished: *permanent* and *non-permanent* faults.

1.1.2.1 Permanent faults

The term **permanent fault** refers to the presence of a fault that affects the functional behavior of a system (chip, array or board) *permanently*. Examples of permanent, also called *solid* or *hard*, faults are:

- Incorrect connections between integrated circuits (ICs), boards, tracks, etc. (e.g., missing connections or shorts due to solder splashes or design faults).
- Broken components or parts of components.
- Incorrect IC masks, internal silicon-to-metal or metal-to-package connections (a manufacturing problem).
- Functional design errors (the implementation of the logic function is incorrect).

Because the permanent faults affect the logic values in the system permanently, they are easier to detect than the non-permanent faults which are described below.

1.1.2.2 Non-permanent faults

Non-permanent faults are present only part of the time; they occur at random moments and affect the system's functional behavior for finite, but unknown, periods of time. As a consequence of this random appearance, detection and localization of non-permanent faults is difficult. If such a fault does not affect the system during test, then the system appears to be performing correctly.

The non-permanent faults can be divided into two groups with different origins: *transient* and *intermittent* faults.

Transient faults are caused by environmental conditions such as cosmic rays, α-particles, pollution, humidity, temperature, pressure, vibration, power supply fluctuations, electromagnetic interference, static electrical discharges, and ground loops.

Transient faults are hard to detect due to their obscure influence on the logic values in a system. Errors in random-access memories (RAMs) introduced by transient faults are often called **soft errors**. They are considered non-recurring, and it is assumed that no permanent damage has been done to the memory cell. Radiation with α-particles is considered a major cause of soft errors (Ma and Dressendorfer, 1989).

Intermittent faults are caused by non-environmental conditions such as loose connections, deteriorating or ageing components (the general assumption is that during the transition from normal functioning to worn-out, intermittent faults may occur), critical timing (hazards and race conditions, which can be caused by design faults), resistance and capacitance variations (resistor and capacitor values may deviate from their specified value initially or over time, which may lead to timing faults), physical irregularities, and noise (noise disturbs the signals in the system).

A characteristic of intermittent faults is that they behave like permanent faults for the duration of the failure caused by the intermittent fault. Unfortunately, the time that an intermittent fault affects the system is usually very short in comparison with the application time of a test developed for permanent faults, which is typically a few seconds. This problem can be alleviated by continuously repeating the test or by causing the non-permanent fault to become permanent. The natural transition of non-permanent faults into permanent faults can take hours, days or months, and so must be accelerated. This can be accomplished by providing specific environmental *stress conditions* (temperature, pressure, humidity, etc.). One problem with the application of stress conditions is that new faults may develop, causing additional failures.

1.2 An analysis of faults

This section gives an analysis of faults; it starts with an overview of the frequency of occurrence of faults as a function of time; Section 1.2.2 describes the behavior of the failure rate of a system over its lifetime and Section 1.2.3 shows how the failure rate of series and parallel systems can be computed. Section 1.2.4 explains the physical and electrical causes of faults, called *failure mechanisms*.

1.2.1 Frequency of occurrence of faults

The frequency of occurrence of faults can be described by a theory called *reliability theory*. In-depth coverage can be found in O'Connor (1985); below a short summary is given.

The point in time t at which a fault occurs can be considered a random variable u. The probability of a failure *before* time t, $F(t)$, is the *unreliability* of a system; it can be expressed as:

$$F(t) = P(u \leq t). \tag{1.1}$$

The **reliability** of a system, $R(t)$, is the probability of a correct functioning system at time t; it can be expressed as:

$$R(t) = 1 - F(t), \tag{1.2}$$

or alternatively as:

$$R(t) = \frac{\text{number of components surviving at time } t}{\text{number of components at time 0}}. \tag{1.3}$$

It is assumed that a system initially will be operable, i.e., $F(0) = 0$, and ultimately will fail, i.e., $F(\infty) = 1$. Furthermore, $F(t) + R(t) = 1$ because at any instance in time either the system has failed or is operational.

The derivative of $F(t)$, called the **failure probability density function** $f(t)$, can be expressed as:

$$f(t) = \frac{dF(t)}{dt} = -\frac{dR(t)}{dt}. \tag{1.4}$$

Therefore, $F(t) = \int_0^t f(t)dt$ and $R(t) = \int_t^\infty f(t)dt$.

The **failure rate**, $z(t)$, is defined as the conditional probability that the system fails during the time-period $(t, t + \Delta t)$, given that the system was operational at time t.

$$z(t) = \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \cdot \frac{1}{R(t)} = \frac{dF(t)}{dt} \cdot \frac{1}{R(t)} = \frac{f(t)}{R(t)}. \tag{1.5}$$

Alternatively, $z(t)$ can be defined as:

$$z(t) = \frac{\text{number of failing components per unit time at time } t}{\text{number of surviving components at time } t}. \tag{1.6}$$

$R(t)$ can be expressed in terms of $z(t)$ as follows:

$$\int_0^t z(t)dt = \int_0^t \frac{f(t)}{R(t)}dt = -\int_{R(0)}^{R(t)} \frac{dR(t)}{R(t)} = -\ln \frac{R(t)}{R(0)},$$

$$\text{or, } R(t) = R(0)e^{-\int_0^t z(t)dt}. \tag{1.7}$$

The **average lifetime** of a system, θ, can be expressed as the mathematical expectation of t to be:

$$\theta = \int_0^\infty t \cdot f(t)dt. \tag{1.8}$$

For a non-maintained system, θ is called the **mean time to failure (*MTTF*)**:

$$MTTF = \theta = -\int_0^{\infty} t \cdot \frac{dR(t)}{dt} dt = -\int_{R(0)}^{R(\infty)} t \cdot dR(t).$$

Using partial integration and assuming that $\lim_{T\to\infty} T \cdot R(T) = 0$:

$$MTTF = \lim_{T\to\infty} \left\{ -t \cdot R(t) \mid_0^T + \int_0^T R(t)dt \right\} = \int_0^{\infty} R(t)dt. \tag{1.9}$$

Given a system with the following reliability:

$$R(t) = e^{-\lambda t}, \tag{1.10}$$

the failure rate, $z(t)$, of that system is computed below and has the constant value λ:

$$z(t) = \frac{f(t)}{R(t)} = \frac{dF(t)}{dt} / R(t) = \frac{d(1 - e^{-\lambda t})}{dt} / e^{-\lambda t} = \lambda e^{-\lambda t} / e^{-\lambda t} = \lambda. \tag{1.11}$$

Assuming failures occur randomly with a constant rate λ, the *MTTF* can be expressed as:

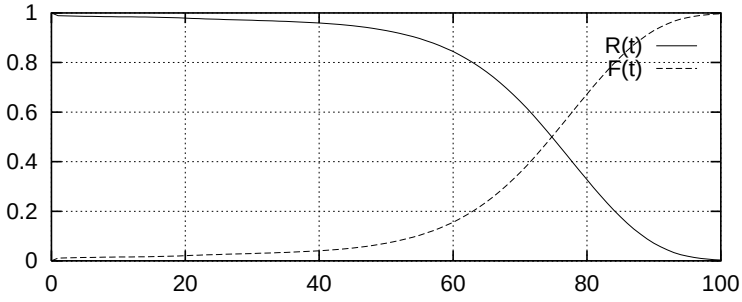
$$MTTF = \theta = \int_0^{\infty} e^{-\lambda t} dt = \frac{1}{\lambda}. \tag{1.12}$$

For illustrative purposes, Figure 1.1 shows the values of $R(t)$, $F(t)$, $f(t)$ and $z(t)$ for the life expectancy of the Dutch male population averaged over the years 1976–1980 (Gouda, 1994). Figure 1.1(a) shows the functions $R(t)$ and $F(t)$; the maximum age was 108 years, the graph only shows the age interval 0 through 100 years because the number of live people in the age interval 101 through 108 was too small to derive useful statistics from. Figures 1.1(b) and 1.1(c) show $z(t)$ and Figure 1.1(d) shows $f(t)$ which is the derivative of $F(t)$. Notice the increase in $f(t)$ and $z(t)$ between the ages 18–20 due to accidents of inexperienced drivers, and the rapid decrease of $z(t)$ in the period 0–1 year because of decreasing infant mortality.

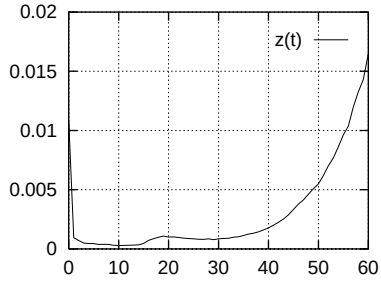
1.2.2 Failure rate over product lifetime

A well-known graphical representation of the failure rate, $z(t)$, as a function of time is shown in Figure 1.2, which is known as the **bathtub curve**. It has been developed to model the failure rate of mechanical equipment, and has been adapted to the semiconductor industry (Moltoft, 1983). It can be compared with Figure 1.1(d). The bathtub curve can be considered to consist of three regions:

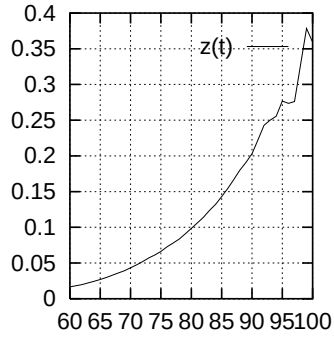
- Region 1, with decreasing failure rate (infant mortality).
  Failures in this region are termed *infant mortalities*; they are attributed to poor quality as a result of variations in the production process.
- Region 2, with constant failure rate; $z(t) = \lambda$ (working life).
  This region represents the 'working life' of a component or system. Failures in this region are considered to occur randomly.



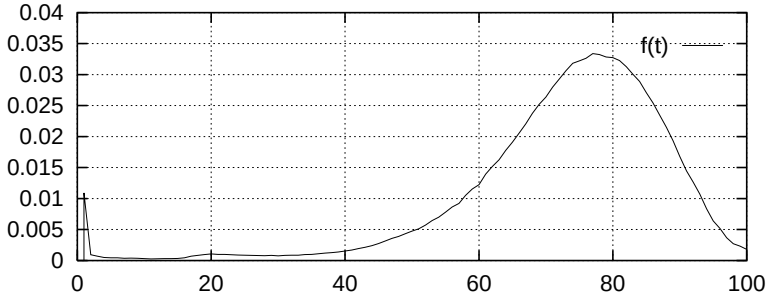
(a) $R(t)$, $F(t)$ vs years



(b) $z(t)$ vs younger years



(c) $z(t)$ vs older years



(d) $f(t)$ vs years

Figure 1.1 Life expectancy of a human population

- Region 3, with increasing failure rate (wearout).
  This region, called 'wearout', represents the end-of-life period of a product. For electronic products it is assumed that this period is less important because they will not enter this region due to a shorter economic lifetime.

From Figure 1.2 it may be clear that products should be shipped to the user only after they have passed the infant mortality period, in order to reduce the high field repair cost. Rather than ageing the to-be-shipped product for the complete infant mortality

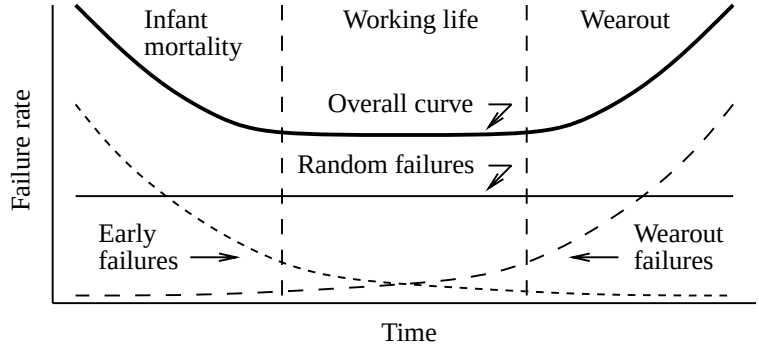


Figure 1.2 Bathtub curve

period, which may be several months, a shortcut is taken by increasing the failure rate. The failure rate increases when a component is used in an 'unfriendly' environment, caused by a **stress condition**. An important stress condition is an increase in temperature which accelerates many physical–chemical processes, thereby accelerating the ageing process. The accelerating effect of the temperature on the failure rate can be expressed by the experimentally determined **equation of Arrhenius**:

$$\lambda_{T_2} = \lambda_{T_1} \cdot e^{(E_a(1/T_1 - 1/T_2)/k)}, \tag{1.13}$$

where:

T_1 and T_2 are absolute temperatures (in Kelvin, K),
λ_{T_1} and λ_{T_2} are the failure rates at T_1 and T_2, respectively,
E_a is a constant expressed in electron-volts (eV), known as the **activation energy**,
and k is Boltzmann's constant ($k = 8.617 \times 10^{-5}$ eV/K).

From Arrhenius' equation it can be concluded that the failure rate is exponentially dependent on the temperature. This is why temperature is a very important stress condition (see example below). Subjecting a component or system to a higher temperature in order to accelerate the ageing process is called **burn-in** (Jensen and Petersen, 1982). Practical results have shown that a burn-in period of 50–150 hours at 125 °C is effective in exposing 80–90% of the component and production-induced defects (e.g., solder joints, component drift, weak components) and reducing the initial failure rate (infant mortality) by a factor of 2–10.

Example 1.2 Suppose burn-in takes place at 150 °C; given that $E_a = 0.6$ eV and the normal operating temperature is 30 °C. Then the acceleration factor is:

$$\lambda_{T_2}/\lambda_{T_1} = e^{0.6(1/303 - 1/423)/8.617 \times 10^{-5}} = 678,$$

which means that the infant mortality period can be reduced by a factor of 678. □

1.2.3 Failure rate of series and parallel systems

If all components of a system have to be operational in order for the system to be operational, it is considered to be a **series system**. Consider a series system consisting of n components, and assume that the probability of a given component to be defective is independent of the probabilities of the other components. Then the reliability of the system can be expressed (assuming $R_i(t)$ is the reliability of the ith component) as:

$$R_s(t) = \prod_{i=1}^{n} R_i(t). \tag{1.14}$$

Using Equation (1.7), it can be shown that:

$$z_s(t) = \sum_{i=1}^{n} z_i(t). \tag{1.15}$$

A **parallel system** is a system which is operational as long as at least one of its n components is operational; i.e., it only fails when *all* of its components have failed. The unreliability of such a system can be expressed as follows:

$$F_p(t) = \prod_{i=1}^{n} F_i(t). \tag{1.16}$$

Therefore, the reliability of a parallel system can be expressed as:

$$R_p(t) = 1 - \prod_{i=1}^{n} F_i(t). \tag{1.17}$$

1.2.4 Failure mechanisms

This section describes the physical and electrical causes for faults, called **failure mechanisms**. A very comprehensive overview of failure mechanisms for semiconductor devices is given in Amerasekera and Campbell (1987), who identify three classes (see Figure 1.3):

1 Electrical stress (in-circuit) failures:
These failures are due to poor design, leading to electric overstress, or due to careless handling, causing static damage.
2 Intrinsic failure mechanisms:
These are inherent to the semiconductor die itself; they include crystal defects, dislocations and processing defects. They are usually caused during wafer fabrication and are due to flaws in the oxide or the epitaxial layer.
3 Extrinsic failure mechanisms:
These originate in the packaging and the interconnection processes; they can be attributed to the metal deposition, bonding and encapsulation steps.

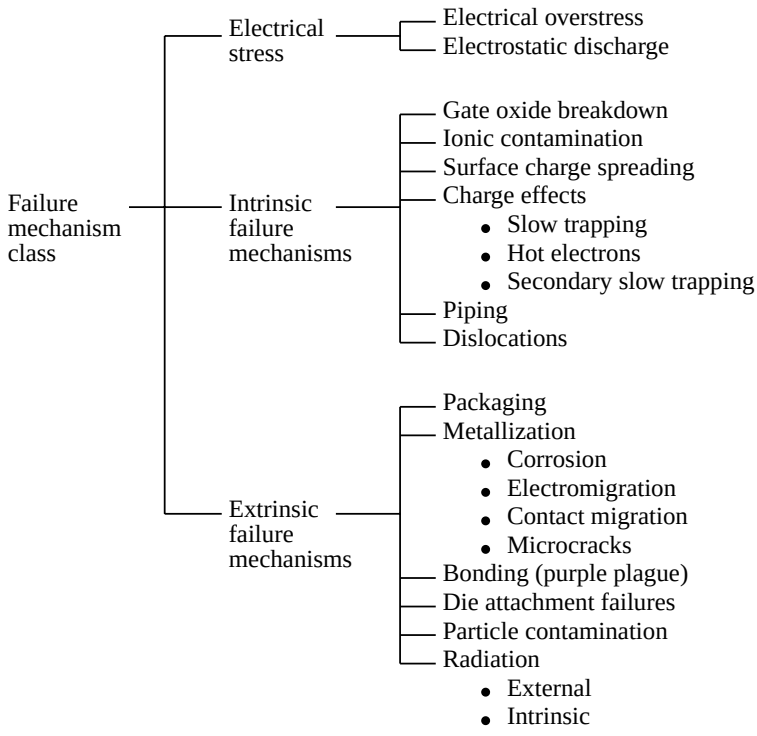


Figure 1.3 Classification of failure mechanisms

Over time, the die fabrication process matures, thereby reducing the intrinsic failure rate, causing the extrinsic failure rate to become more dominant. However, it is very difficult to give a precise ordering of the failure mechanisms; some are dominant in certain operational and environmental conditions, others are always present but with a lower impact.

An important parameter of a failure mechanism is E_a, the activation energy, describing the temperature dependence of the failure mechanism. E_a typically varies between 0.3 and 1.5 eV. Temperatures between 125 °C and 250 °C have been found to be effective for burn-in, without causing permanent damage (Blanks, 1980). The exact influence of the temperature on the failure rate (i.e., the exact value of E_a) is very hard to determine and varies between manufacturers, batches, etc. Table 1.1 lists experimentally determined values for the activation energies of the most important failure mechanisms, which are described next.

Corrosion is an electromechanical failure mechanism which occurs under the condition that moisture and DC potentials are present; Cl^- and Na^+ ions act as a catalyst. Packaging methods (good sealing) and environmental conditions determine the corrosion process to a large extent; CMOS devices are more susceptible due to their low power dissipation.

Table 1.1. *Activation energies of some major failure mechanisms*

Failure mechanism	Activation energy E_a
Corrosion of metallization	0.3–0.6 eV
Electrolytic corrosion	0.8–1.0 eV
Electromigration	0.4–0.8 eV
Bonding (purple plague)	1.0–2.2 eV
Ionic contamination	0.5–1.0 eV
Alloying (contact migration)	1.7–1.8 eV

Electromigration occurs in the Al (aluminum) metallization tracks (lines) of the chip. The electron current flowing through the Al tracks causes the electrons to collide with the Al grains. Because of these collisions, the grains are dislocated and moved in the direction of the electron current. Narrow line widths, high current densities, and a high temperature are major causes of electromigration, which results in open lines in places where the current density is highest.

Bonding is the failure mechanism which consists of the deterioration of the contacts between the Au (gold) wires and the Al pads of the chip. It is caused by interdiffusion of Au–Al which causes open connections.

Ionic contamination is caused by mobile ions in the semiconductor material and is a major failure mechanisms for MOS devices. Na^+ ions are the most mobile due to their small radius; they are commonly available in the atmosphere, sweat and breath. The ions are attracted to the gate oxide of a FET transistor, causing a change in the threshold voltage of the device.

Alloying is also a form of Al migration of Al into Si (silicon) or Si into Al. Depending on the junction depth and contact size, the failure manifests itself as a shorted junction or an open contact. As device geometries get smaller, alloying becomes more important, because of the smaller diffusion depths.

Radiation (Ma and Dressendorfer, 1989) is another failure mechanism which is especially important for dynamic random-access memories (DRAMs). Trace impurities of radioactive elements present in the packaging material of the chip emit α-particles with energies up to 8 MeV. The interaction of these α-particles with the semiconductor material results in the generation of electron–hole pairs. The generated electrons move through the device and are capable of wiping out the charge stored in a DRAM cell, causing its information to be lost. This is the major cause of soft errors in DRAMs. Current research has shown that high-density static random-access memories (SRAMs) also suffer from soft errors caused by α-particles (Carter and Wilkins, 1987).

1.3 Classification of tests

A test is a procedure which allows one to distinguish between good and bad parts. Tests can be classified according to the technology they are designed for, the parameters they measure, the purpose for which the test results are used, and the test application method.

1.3.1 Technology aspect

The type of tests to be performed depends heavily on the technology of the circuit to be tested: analog, digital, or mixed-signal.

Analog circuits have the property that the domain of values of the input and output signals is analog; i.e., the signals can take on *any* value in a given range (this range is delimited by a lower and an upper bound (e.g., in the case of voltage levels those bounds may be determined by the supply voltage, resulting in a range of 0 to +5 V)). Analog tests aim at determining the values of the analog parameters such as voltage and current levels, frequency response, bandwidth, distortion, etc. The generation of the test input stimuli, the processing of these stimuli by the circuit, as well as the determination of the values of the test response signals are inherently imprecise due to the analog nature of the signals (infinite accuracy does not exist). Therefore, the determination whether a circuit satisfies its requirements is not based on a single value, but on a range of values (i.e., an interval), for each of the test response signals.

Digital circuits have the property that the domain of values of the input and output signals is binary (usually referred to as *digital*); i.e., the signals can only take on the value 'logic 0' or 'logic 1'. Tests for digital circuits determine the values of the binary test response signals, given binary test stimuli which are processed digitally by the to-be-tested circuit; this can be done precisely due to the binary nature of the signal values. These tests are called *logical tests* or *digital tests*.

Mixed-signal circuits have the property that the domain of values of the input signals is digital (analog) while the domain of the values of the output signals is analog (digital), e.g., digital-to-analog converter (DAC) and analog-to-digital converter (ADC) circuits. Testing mixed-signal circuits is based on a combination of analog and digital test techniques.

This book mainly focuses on testing digital circuits.

1.3.2 Measured parameter aspect

When testing digital circuits, a classification of tests can be made based on the nature of the type of measurement which is performed on the value of the binary signal. When the measurement aims at verifying the logical correctness of the signal value one speaks about logical tests; when it concerns the behavior of the signal value in time, or its voltage level and/or drive capability, one speaks about electrical tests.

Logical tests: **Logical tests** aim at finding faults which cause a change in the logical behavior of the circuit: deviations from the good device are only considered faults when a response signal level is a 'logic 0' instead of the expected 'logic 1', or vice versa. These faults may be anywhere in the circuit and are not considered to be time-dependent (i.e., they are permanent faults).

Electrical tests: **Electrical tests** verify the correctness of a circuit by measuring the values of electrical parameters, such as voltage and current levels, as well as their behavior over time. They can be divided into *parametric* tests and *dynamic* tests.

Parametric tests are concerned with the *external behavior* of the circuit; i.e., voltage/current levels and delays on the input and output pins of the chip. The specifications of the signal values on the input and output pins of a chip have a time-independent part (voltage and current levels) and a time-dependent part (rise and fall times). The qualification of a chip via the verification of the time-independent properties of the signal values on the input and output pins is called **DC parametric testing**, whereas the qualification via the verification of the time-dependent properties is called **AC parametric testing**. For a comprehensive treatment of DC and AC parametric tests the reader is referred to Stevens (1986) and van de Goor (1991).

A special case of DC parametric testing is the $\boldsymbol{I_{DDQ}}$ **test** method (Hawkins and Soden, 1986); this test method has been shown to be capable of detecting logical faults in CMOS circuits and can be used to measure the reliability of a chip. I_{DDQ} is the quiescent power supply current which is drawn when the chip is not switching. This current is caused by sub-threshold transistor leakage and reverse diode currents and is very small; on the order of tens of nA. I_{DDQ} tests are based on the fact that defects like shorts and abnormal leakage can increase I_{DDQ} by orders of magnitude. It has been shown that the I_{DDQ} test method is effective in detecting faults of many fault models (see Chapter 2 for a description of fault models).

Dynamic tests are aimed at detecting faults which are time-dependent and internal to the chip; they relate to the speed with which the operations are being performed. The resulting failures manifest themselves as logical faults. Delay tests, which verify whether output signals make transitions within the specified time, and refresh tests for DRAMs, belong to the class of dynamic tests.

The main emphasis of this book is on logical tests; however, a chapter has been included on I_{DDQ} fault testing because of its elegance and fault detection capabilities,

and a chapter on delay fault testing (which is a form of dynamic test) has been included because of its importance owing to the increasing speed of modern circuits.

1.3.3 Use of test results

The most obvious use of the result of a test is to distinguish between good and bad parts. This can be done with a test which **detects faults**. When the purpose of testing is repair, we are interested in a test which **locates faults**, which is more difficult to accomplish.

Testing can be done during normal use of the part or system; such tests are called **concurrent tests**. For example, byte parity, whereby an extra check bit is added to every eight information bits, is a well known concurrent error detection technique capable of detecting an odd number of errors. Error correcting codes, where, for example, seven check bits are added to a 32-bit data word, are capable of correcting single-bit errors and detecting double-bit errors (Rao and Fujiwara, 1989). Concurrent tests, which are used extensively in fault-tolerant computers, are not the subject of this book. **Non-concurrent tests** are tests which cannot be performed during normal use of the part or system because they do not preserve the normal data. They have the advantage of being able to detect and/or locate more complex faults; these tests are the subject of this book.

Design for testability (DFT) is a technique which allows non-concurrent tests to be performed faster and/or with a higher fault coverage by including extra circuitry on the to-be-tested chip. DFT is used extensively when testing sequential circuits. When a test has to be designed for a given fault in a combinational circuit with n inputs, the search space for finding a suitable test stimulus for that fault consists of 2^n points. In case of a sequential circuit containing f flip-flops, the search space increases to 2^{n+f} points, because of the 2^f states the f flip-flops can take on. In addition, since we cannot directly control the present state lines of a sequential circuit nor directly observe the next state lines, a justification sequence is required to get the circuit into the desired state and a propagation sequence is required to propagate errors, if any, from the next state lines to primary outputs. This makes test generation a very difficult task for any realistic circuit. To alleviate this problem, a DFT technique can be used which, in the test mode, allows all flip-flops to form a large shift register. As a shift register, the flip-flops can be tested easily (by shifting in and out certain 0–1 sequences); while at the same time test stimuli for the combinational logic part of the circuit can be shifted in (scanned in), and test responses can be shifted out (scanned out), thus reducing the sequential circuit testing problem to combinational circuit testing. This particular form of DFT is called *scan design*.

When the amount of extra on-chip circuitry for test purposes is increased to the extent that test stimuli can be generated and test responses observed on-chip, we speak of **built-in self-test (BIST)**, which obviates the requirement for an au-

tomatic test equipment (ATE) and allows for at-speed (at the normal clock rate) testing.

1.3.4 Test application method

Tests can also be classified, depending on the way the test stimuli are applied to the part and the test responses are taken from the part.

An ATE can be used to supply the test stimuli and observe the test responses; this is referred to as an **external test**. Given a board with many parts, the external test can be applied in the following ways:

1 Via the normal board connectors:
This allows for a simple interface with the ATE and enables the board to be tested at normal speed; however, it may be difficult (if not impossible) to design tests which can detect all faults because not all circuits are easy to reach from the normal connectors. At the board level, this type of testing is called **functional testing**; this is the normal way of testing at this level.

2 Via a special fixture (a board-specific set of connectors):
A special board-specific connector is used to make all signal lines on the board accessible. High currents are used to drive the test stimuli signals in order to temporarily overdrive existing signal levels; this reduces the rate at which tests can be performed to about 1 MHz. This type of testing is called **in-circuit testing**. It enables the location of faulty components.

1.4 Fault coverage requirements

Given a chip with potential defects, an important question to be answered is how extensive the tests have to be. This section presents an equation relating the *defect level* of chips to the *yield* and fault coverage (Williams and Brown, 1981; Agrawal *et al*., 1982; McCluskey and Buelow, 1988; Maxwell and Aitken, 1993). This equation can be used to derive the required test quality, given the process yield and the desired defect level.

The **defect level**, ***DL***, is the fraction of bad parts that pass all tests. Values for *DL* are usually given in terms of defects per million, DPM; desired values are less than 200 DPM, which is equal to 0.02% (Intel, 1987).

The **process yield**, ***Y***, is defined as the fraction of the manufactured parts that is defect-free. The exact value of *Y* is rarely known, because it is not possible to detect all faulty parts by testing all parts for all possible faults. Therefore, the value of *Y* is usually approximated by the ratio: number-of-not-defective-parts/total-number-of-parts; whereby the number-of-not-defective-parts is determined by counting the parts which pass the used test procedure.

The **fault coverage**, ***FC***, is a measure to grade the quality of a test; it is defined as the ratio of: actual-number-of-detected-faults/total-number-of-faults (where the faults are assumed to belong to a particular fault model, see Chapter 2). In practice, it will be impossible to obtain a complete test (i.e., a test with $FC = 1$) for a VLSI part because of: (a) imperfect fault modeling: an actual fault (such as an open connection in an IC) may not correspond to a modeled fault or vice versa, (b) data dependency of faults: it may not be sufficient to exercise all functions (such as ADD and SUB, etc. in a microprocessor) because the correct execution of those functions may be data-dependent (e.g., when the carry function of the ALU is faulty), and (c) testability limitations may exist (e.g., due to pin limitations not all parts of a circuit may be accessible). This implies that if the circuit passes the test, one cannot guarantee the absence of faults.

Assume that a given chip has exactly n stuck-at faults (SAFs). An SAF is a fault where a signal line has a permanent value of logic 0 or 1. Let m be the number of faults detected ($m \leq n$) by a test for SAFs. Furthermore, assume that the probability of a fault occurring is independent of the occurrence of any other fault (i.e., there is no clustering), and that all faults are equally likely with probability p. If A represents the event that a part is free of defects, and B the event that a part has been tested for m defects while none were found, then the following equations can be derived (McCluskey and Buelow, 1988).

The fault coverage of the test is defined as: $FC = m/n$. (1.18)

The process yield is defined as: $Y = (1-p)^n = P(A)$. (1.19)

$$P(B) = (1-p)^m. \tag{1.20}$$

$$P(A \cap B) = P(A) = (1-p)^n. \tag{1.21}$$

$$\begin{aligned} P(A|B) &= P(A \cap B)/P(B) = (1-p)^n/(1-p)^m \\ &= (1-p)^{n(1-m/n)} = Y^{(1-FC)}. \end{aligned} \tag{1.22}$$

The defect level can be expressed as:

$$DL = 1 - P(A|B) = 1 - Y^{(1-FC)}. \tag{1.23}$$

Figure 1.4 (Williams and Brown, 1981) shows curves for DL as a function of Y and FC; it is a nonlinear function of Y. However, for large values of Y, representing manufacturing processes with high yield, the curve approaches a straight line.

Example 1.3 Assume a manufacturing process with an yield of $Y = 0.5$ and a test with $FC = 0.8$; then $DL = 1 - 0.5^{(1-0.8)} = 0.1295$ which means that 12.95% of all shipped parts are defective. If a DL of 200 DPM (i.e., $DL = 0.0002 = 0.02\%$) is required, given $Y = 0.5$, the fault coverage has to be: $FC = 1 - (\log(1 - DL)/\log Y) = 0.999\,71$, which is 99.971%. □

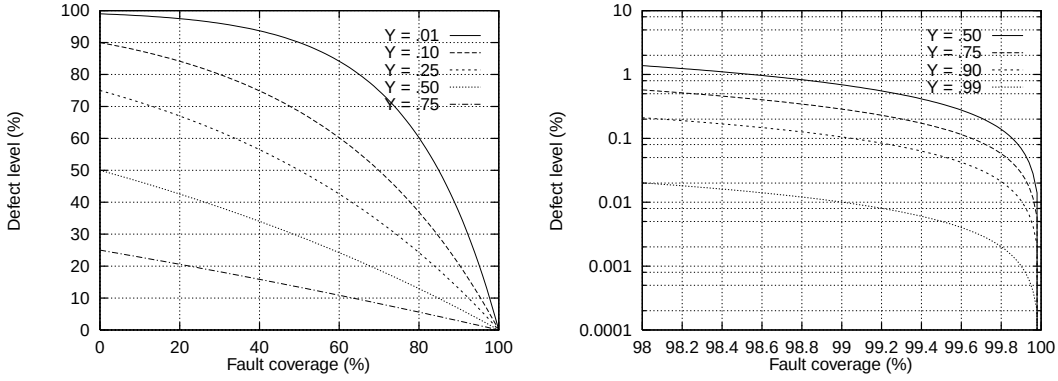


Figure 1.4 Defect level as a function of fault coverage

1.5 Test economics

This section discusses the economics of testing; it starts with a subsection highlighting the repair cost during product phases; thereafter the economics and liability of testing are considered. The third subsection gives an analysis of the cost/benefit of test development, and the last subsection discusses the field maintenance cost.

1.5.1 Repair cost during product phases

During the lifetime of a product, the following phases can be recognized: component manufacture (chips and bare boards), board manufacture, system manufacture, and working life of a product (when it is in the field). During each phase, the product quality has to be ensured. The relationship of the test and repair costs during each of these phases can be approximated with the *rule-of-ten* (Davis, 1982) (see Figure 1.5): if the test and repair cost in the component manufacturing phase is R, then in the board manufacture phase it is $10R$, in the system manufacturing phase it is $100R$, and during the working life phase it is $1000R$. This is due to the increase in the difficulty level of locating the faulty part, the increase in repair effort (travel time and repair time) and the larger volume of units involved.

1.5.2 Economics and liability of testing

Section 1.5.1 explained the rule-of-ten for the test and repair cost. From this it is obvious that by eliminating faults in the early phases of a product, great savings can be obtained. In addition, a good test approach can reduce the development time, and therefore the time-to-market, as well as the field maintenance cost. However, test development also takes time and therefore increases the development time. The

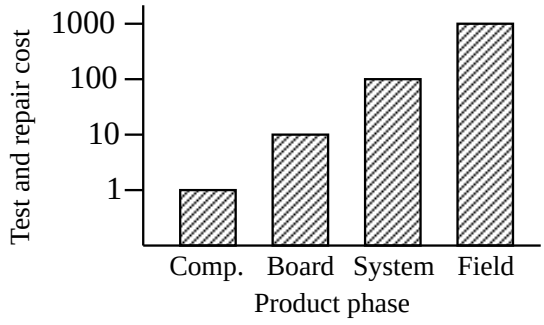


Figure 1.5 Rule-of-ten

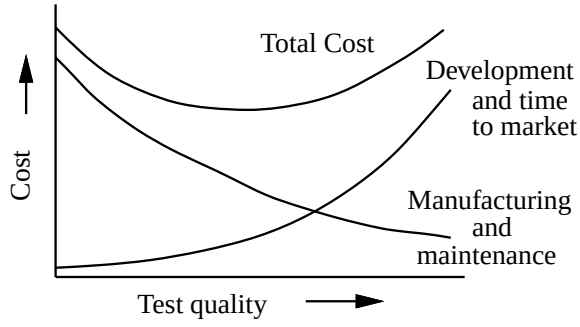


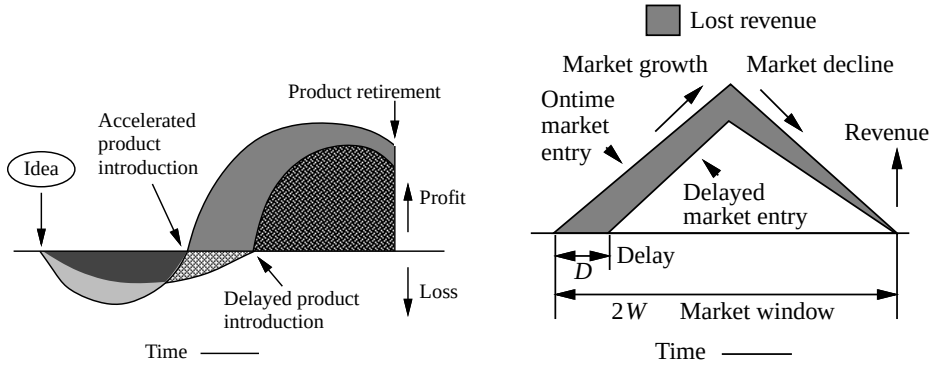
Figure 1.6 Influence of test quality on cost

right amount of test development effort therefore is a tradeoff (see Figure 1.6). With increasing test quality: (a) the cost of test development, as well as the elapsed test development time (which can be translated into a later time-to-market), increases, and (b) the manufacturing test and repair cost decreases. Therefore, the total cost of the product (during all its product phases) is minimized for some test quality which is not the highest test quality.

Another important reason for testing is the liability aspect when defects during the working life period cause breakdowns which may lead to personal injury, property damages and/or economic loss. Well-known examples from the car industry (accidents) and the medical practice (overdose of radiation) show civil and punitive lawsuits where the equipment manufacturer has been liable for large amounts (millions of dollars) of money.

Looking at the lifetime of a product, one can recognize several economic phases (see Figure 1.7(a)): the development phase, the market growth phase, and the market decline phase.

During the development phase, an associated cost is incurred, resulting in an initial loss. Initially, the development team is small and grows when the product is better defined such that more disciplines can participate. Later on, the development effort



(a) Exact (b) Approximate

Figure 1.7 Product revenue over lifetime

decreases until the development has been completed. The area of the curve under the zero-line of Figure 1.7(a) represents the total product development cost.

After development, the product is sold, resulting in a profit. Initially, the market share is low and increases with increasing product acceptance (market growth); thereafter, due to competition/obsolescence, the market share decreases (market decline) until the product is taken out of production (product retirement). The total product revenue is the difference in the two areas marked 'profit' and 'loss', respectively.

For typical digital systems, such as mainframes, workstations, and personal computers, the development period may be on the order of one to two years, while the *market window* (consisting of the market growth and market decline parts) is on the order of four years (for personal computers it is closer to two years!). Usually, the development cost is a very small part of the total lifetime cost of a system (on the order of 10%) and is rather constant for a given system. Ignoring this cost and approximating the market growth and market decline areas by straight lines, the life-cycle model of Figure 1.7(a) may be approximated with that of Figure 1.7(b).

Digital technology evolves at a rate of about $1.6X$/year, which means that the speed (in terms of clock frequency) increases, and the cost (in terms of chip area) decreases, at this rate. The result is that when a product is being designed with a particular technology in mind, the obsolescence point is more or less fixed in time. Therefore, when a delay (D) is incurred in the development process, this will have a severe impact on the revenue of the product (see Figure 1.7).

From Figure 1.7(b), and assuming that the maximum value of the market growth is M and reached after time W, the revenue lost due to a delay, D, (hatched area) can be computed as follows. The expected revenue (ER) in case of on-time market entry is: $ER = \frac{1}{2} \cdot 2W \cdot M = W \cdot M$. The revenue of the delayed product (RDP) is: $RDP = \frac{1}{2} \cdot (2W - D)\left(\frac{(W-D)}{W} \cdot M\right)$.

The lost revenue (LR) is: $LR = ER - RDP$

$$= W \cdot M - \frac{2W^2 - 3D \cdot W + D^2}{2W} \cdot M$$
$$= ER \cdot \frac{D(3W - D)}{2W^2}. \tag{1.24}$$

It should be noted that a delay 'D' may be incurred due to excessive test development, but more likely, due to insufficient test development because of more repairs during later product phases and a longer development phase due to the longer time to get the product to the required quality level.

1.5.3 Cost/benefit of test development

Many factors contribute to the cost of developing a test (Ambler *et al.*, 1992). These factors can be divided into per unit cost and cost incurred due to changes in the product development process (i.e., schedule and performance consequences).

1 Engineering cost, C_E:
This is the time spent in developing the test, and possibly also modifying the design, in order to obtain the required fault coverage. For an application-specific integrated circuit (ASIC), C_E can be calculated as:

$$C_E = (\text{number of Eng. days}) \cdot (\text{cost per day})/(\text{number of ASICs}). \tag{1.25}$$

Example 1.4 If it takes 20 days to develop the test, while the cost/day is \$600 and 2000 ASICs are produced; then

$C_E = 20 \cdot \$600/2000 = \$6/\text{ASIC}.$

2 Increased ASIC cost, C_A:
In order to obtain the required fault coverage, DFT and/or BIST techniques may have to be used. This increases the chip area and reduces the yield such that the cost of the ASIC increases. When the additional DFT/BIST circuitry causes the die area to increase from A_O to A_E, then the cost of the die will increase by the following amount (C_O is the original die cost):

$$C_A = C_O \cdot \{A_E/A_O \cdot Y^{(1-\sqrt{A_E/A_O})} - 1\}. \tag{1.26}$$

Example 1.5 Assume an ASIC cost of \$50 without on-chip test circuits. Furthermore, assume that 50% of this cost is due to the die; the remainder is due to packaging, testing and handling. If the additional circuits increase the die area by 15%, then with an yield of 60% the cost of the die will increase by $C_A = \$4.83$. The cost of the ASIC will increase from \$50 to \$54.83.

3 Manufacturing re-work cost, C_M:
If the ASIC test program has a fault coverage of less than 100% then some ASICs may be defective at the board level. These faulty ASICs have to be located and replaced; this involves de-soldering the defective part and soldering its replacement. When surface mount technology is used, this may be (nearly) impossible such that the board has to be scrapped. C_M can be expressed as follows: C_M = (board repair cost per ASIC) · (defect level 'DL', representing the fraction of faulty components) · (number of ASICs on the board) · (number of boards).

Example 1.6 Consider a single-board system with 25 ASICs which have been produced with a yield of 0.6 and tested with a fault coverage of 0.95, while the board repair cost per ASIC is \$500. Then using Equation (1.23), C_M can be computed as follows:

$C_M = (1 - 0.6^{(1-0.95)}) \cdot 25 \cdot \$500 = \$315/\text{system}.$

4 Time-to-market cost, C_T:
This is a complex issue since testing increases the time-to-market because of test development and (possible) redesign time. On the other hand, testing decreases the time-to-market due to the fact that the manufacturing re-work time and cost will decrease. Given the normal expected revenue, the product delay 'D' and the market window 'W'; the time-to-market cost, C_T, can be approximated using Equation (1.24).

Example 1.7 Suppose we are given a \$10 000 system of which 2000 are projected to be sold over a time period of three years (i.e., $2W = 36$) which is delayed by two months due to insufficient testing (i.e., $D = 2$). Using Equation (1.24), C_T will be:

$C_T = \$20\,000\,000 \cdot \{2(3 \cdot 18 - 2)/2 \cdot 18^2\} = \$3\,209\,877.$

Thus, $C_T/2000 = \$1605/\text{system}$.

5 Cost of speed degradation, C_S:
On-chip test circuitry may cause a degradation in speed due to increased fanin and fanout of circuits and/or due to the addition of one or more logic levels in performance-critical paths. This will make the product less competitive such that its price (and profit margin) may decrease. In order to maintain a constant price/performance ratio, the price of the degraded system has to drop from the price of the original system 'P_O' proportionately with the speed degradation such that ($Perf_D$ = performance of degraded system):

$$C_S = P_O \cdot (1 - Perf_D/Perf_O). \tag{1.27}$$

Example 1.8 Assume a system with a price of \$10 000 (i.e., $P_O = \$10\,000$) which

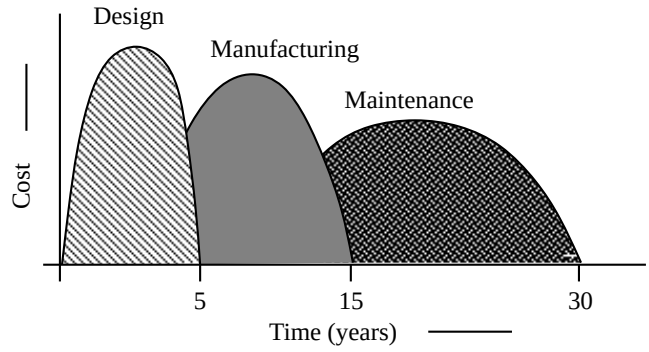


Figure 1.8 Life-cycle cost

incurs a speed degradation of 5% due to on-chip test circuitry, then the decrease in system price will be: $C_S = \$10\,000 \cdot (1 - 0.95) = \$500$.

6 ATE cost, C_T:
The cost of testing a chip, using ATE, is small compared with the costs of points (1)−(5) described above. It can be approximated to a fixed cost per second (e.g., \$0.20/s). Except for special cases where very long test times are required, its contribution to the total cost can be ignored.

1.5.4 Field maintenance cost

Industry analysts estimate that 26% of the revenue and 50% of the profit are generated by the field service centers of equipment vendors (Ambler *et al.*, 1992). In addition, repeated sales depend to a large extent on the quality of service (good/fast maintenance) which reduces the down-time. The importance of field service can furthermore be deduced from the rule-of-ten shown in Figure 1.5.

The life-cycle cost of an industrial electronic system, such as a telephone system or a control or a defense system, is depicted in Figure 1.8 (Ambler *et al.*, 1992). It consists of three parts: design cost (typically on the order of 10%), manufacturing cost (30%), and maintenance cost (60%). The user's 60% maintenance cost, which does not include the cost of down-time, relates to the vendor's 50% profit due to field service.

The maintenance cost can be reduced as follows:

1 Reduction in the number of failures:
This can be done by including fault tolerance techniques (such as error correcting codes for memories and triple modular redundancy for central processing units) in the design.

2 Reduction of the down-time:
This can be accomplished by reducing the travel time and/or the repair time.
The travel time can be reduced by allowing for remote diagnosis (i.e., the use of

telecommunication facilities to diagnose remotely) such that the correct replacement part can be taken along on the repair visit.
The repair time can be reduced by reducing the diagnostic time (through DFT/BIST and/or better test software) and the part replacement time (by having spare parts stocked locally). On-line documentation, product modularity, etc., are other aspects which can reduce the down-time.
Remote repair, where the system is diagnosed remotely and the defective (software) part replaced remotely, is another way of reducing down-time and repair cost. In addition, remote diagnostics/repair allows for centers of competence such that rarely occurring faults may be diagnosed more quickly.

Summary

- A failure means that a system does not perform; this is caused by an error, which means that the system is in a wrong state; the error is caused by a fault which is the physical difference between the good and the faulty system.
- Permanent faults affect the correct operation of the system all the time; a non-permanent fault only does this part of the time. Non-permanent faults can be classified as transient faults, caused by environmental conditions, or as intermittent faults, caused by non-environmental conditions such as ageing components.
- A system with a reliability $R(t) = e^{-\lambda t}$ has a constant failure rate, i.e., $z(t) = \lambda$, and $MTTF = 1/\lambda$.
- The bathtub curve shows the failure rate over the lifetime of a product. The infant mortality period can be reduced by using burn-in techniques. The activation energy is a measure of the sensitivity of a failure mechanism to increased temperature.
- Failure mechanisms are the physical and/or electrical causes of a fault. The most important failure mechanisms are: corrosion, electromigration, bonding, ionic contamination, alloying and radiation.
- Tests can be classified according to: technology aspects (analog, digital and mixed-signal circuits), measured parameter aspect (logical and electrical tests; where electrical tests can be subdivided into parametric tests (DC parametric and AC parametric), and dynamic tests), use of test results (fault detection versus fault location, concurrent versus non-concurrent tests), and the test application method (functional versus in-circuit).
- DFT techniques use extra chip circuitry to allow for accelerating of, and/or obtaining a higher fault coverage of, a non-concurrent test. When the amount of on-chip circuitry is such that the non-concurrent test can be performed without the use of an ATE, one speaks of BIST.
- A relationship exists between the defect level, the process yield and the fault coverage.

- The cost of repair increases by an order of magnitude for every product phase.
- The optimal test quality is a tradeoff between the time-to-market and the manufacturing and maintenance cost.
- The lifetime cost of an industrial product for the user is dominated by the maintenance cost. This can be reduced by reducing the number of failures and/or reducing the downtime.

Exercises

1.1 Suppose we are given a component operating at a normal temperature of 30 °C. The goal is to age this component such that, when shipped, it already has an effective life of one year. The values for E_a can be taken from Table 1.1 where the midpoint of the interval has to be taken.
   (a) Assume a burn-in temperature of 125 °C; what is the burn-in time assuming corrosion of metallization is the dominant failure mode?
   (b) Same question as (a); now with a burn-in temperature of 150 °C.
   (c) Assume a burn-in temperature of 125 °C; what is the burn-in time assuming alloying is the dominant failure mode?
   (d) Same question as (c); now with a burn-in temperature of 150 °C.

1.2 Suppose we are given a system consisting of two subsystems connected in series. The first subsystem is a parallel system with three components; the second subsystem is a parallel system with two components. What is the reliability, $R(t)$, of this system given that the failure rate, $z(t)$, of each of the five components is λ?

1.3 Given a manufacturing process with a certain yield and a certain required defect level for the shipped parts, compute the required fault coverage:
   (a) Given $DL = 20$ DPM and $Y = 0.6$;
   (b) Given $DL = 20$ DPM and $Y = 0.8$;
   (c) Given $DL = 4$ DPM and $Y = 0.6$;
   (d) Given $DL = 4$ DPM and $Y = 0.8$.

1.4 Suppose a team is designing a product consisting of three boards, each containing 10 ASICs. The process producing the ASICs has an yield of 0.7; while tests have a fault coverage of 95%. The ASIC die price is \$30, the cost of a packaged and tested ASIC (without test circuits) is \$60. The engineering cost is \$600/day and the board-level manufacturing re-work cost is \$500 per ASIC. The targeted product life cycle is 48 months; the system price is \$10 000 including a profit margin of 20% and the expected number of units sold is 5000.
   In order to improve the fault coverage to 99%, on-chip circuitry has to be added with the following consequences: the die area increases by 10%, the speed of the chip decreases by 5%, and the extra test development cost due to the on-chip circuits takes one month (assume a month has 22 engineering days).

(a) Given the projected system:
   i. What is the manufacturing re-work cost (C_M) and what is the expected revenue (*ER*)?
   ii. What is the manufacturing re-work cost when the yield is increased to 0.8?
   iii. What is the manufacturing re-work cost when the yield is increased to 0.8 and the fault coverage is increased to 99%?

(b) Given the system, enhanced with test circuitry.
   i. What are the new values for C_E, C_A, C_M, and C_S?
   ii. What is the value of the lost revenue (*LR*)?
   iii. What is the difference in total profit owing to the inclusion of the test circuitry?

(c) Assume that the test circuitry reduces the maintenance cost by 20%. Furthermore, assume that the original system price ($10 000) was only 40% of the life-cycle cost for the customer due to the maintenance cost (see Figure 1.8). How much is the original life-cycle cost and what is the new life-cycle cost given the system enhanced with test circuitry?

References

Agrawal, V.D., Seth, S.C., and Agrawal, P. (1982). Fault coverage requirements in production testing of LSI circuits. *IEEE J. of Solid-State Circuits*, **SC-17**(1), pp. 57–61.

Ambler, A.P., Abadir, M., and Sastry, S. (1992). *Economics of Design and Test for Electronic Circuits and Systems*. Ellis Horwood: Chichester, UK.

Amerasekera, E.A. and Campbell, D.S. (1987). *Failure Mechanisms in Semiconductor Devices*. John Wiley & Sons: Chichester, UK.

Blanks, H.S. (1980). Temperature dependence of component failure rate. *Microelec. and Rel.*, **20**, pp. 219–246.

Carter, P.M. and Wilkins, B.R. (1987). Influences of soft error rates in static RAMs. *IEEE J. of Solid-State Circuits*, **SC-22** (3), pp. 430–436.

Davis, B. (1982). *The Economics of Automated Testing*. McGraw-Hill: London, UK.

van de Goor, A.J. (1991). *Testing Semiconductor Memories, Theory and Practice*. John Wiley & Sons: Chichester, UK.

Gouda (1994). *Life Table for Dutch Male Population 1976–1980*. Gouda Life Insurance Company (R.J.C. Hersmis): Gouda, The Netherlands.

Jensen, F. and Petersen, N.E. (1982). *Burn-in*. John Wiley & Sons: Chichester, UK.

Hawkins, C.F. and Soden, J.M. (1986). Reliability and electrical properties of gate oxide shorts in CMOS ICs. In *Proc. Int. Test Conference*, Washington D.C., pp. 443–451.

Intel (1987). *Components Quality Reliability Handbook*. Intel Corporation: Santa Clara, CA.

Ma, T.P. and Dressendorfer, P.V. (1989). *Ionizing Radiation Effects in MOS Devices and Circuits*. John Wiley & Sons: New York, NY.

McCluskey, E.J. and Buelow, F. (1988). IC quality and test transparency. In *Proc. Int. Test Conference*, pp. 295–301.

Maxwell, P.C. and Aitken, R.C. (1993). Test sets and reject rates: all fault coverages are not created equal. *IEEE Design & Test of Computers* **10**(1), pp. 42–51.

Moltoft, J. (1983). Behind the 'bathtub' curve – a new model and its consequences. *Microelec. and Rel.*, **23**, pp. 489–500.

O'Connor, P.D.T.O. (1985). *Practical Reliability Engineering*. John Wiley & Sons: Chichester, UK.

Rao, T.R.N. and Fujiwara, E. (1989). *Error-Control Coding for Computer Systems*. Prentice-Hall: Englewood Cliffs, NJ.

Stevens, A.K. (1986). *Introduction to Component Testing*. Addison-Wesley: Reading, MA.

Williams, T.W. and Brown, N.C. (1981). Defect level as a function of fault coverage. *IEEE Trans. on Computers*, **C-30**, pp. 987–988.