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:

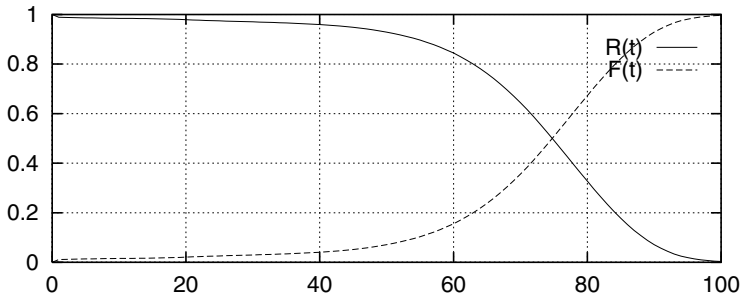
$$MTTF = \theta = \int_0^{\infty} e^{-\lambda t} dt = \frac{1}{\lambda}. \tag{1.12}$$

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.

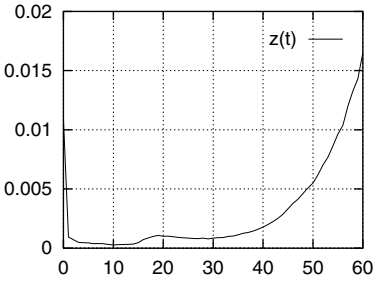
1.2.2 Failure rate over product lifetime

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:

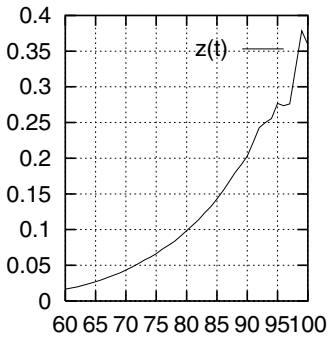
- Region 1, with decreasing failure rate (infant mortality).
  Failures in this region are termed *infant mortalities*; they are attributed to poor quality as a result of variations in the production process.
- Region 2, with constant failure rate; $z(t) = \lambda$ (working life).
  This region represents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



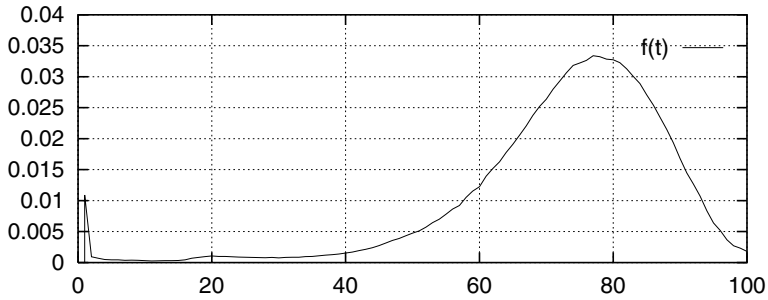
(a) $R(t)$, $F(t)$ vs years



(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

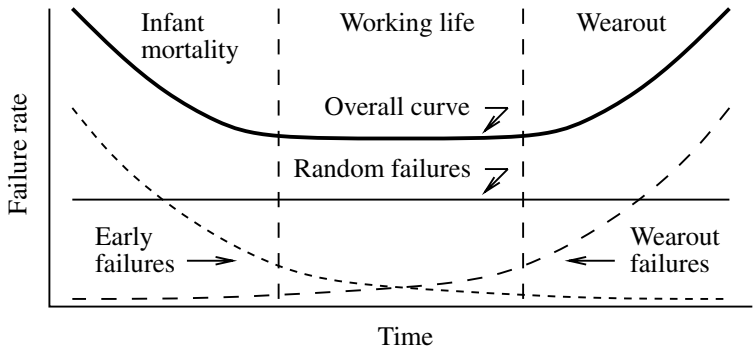


Figure 1.2 Bathtub curve

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.

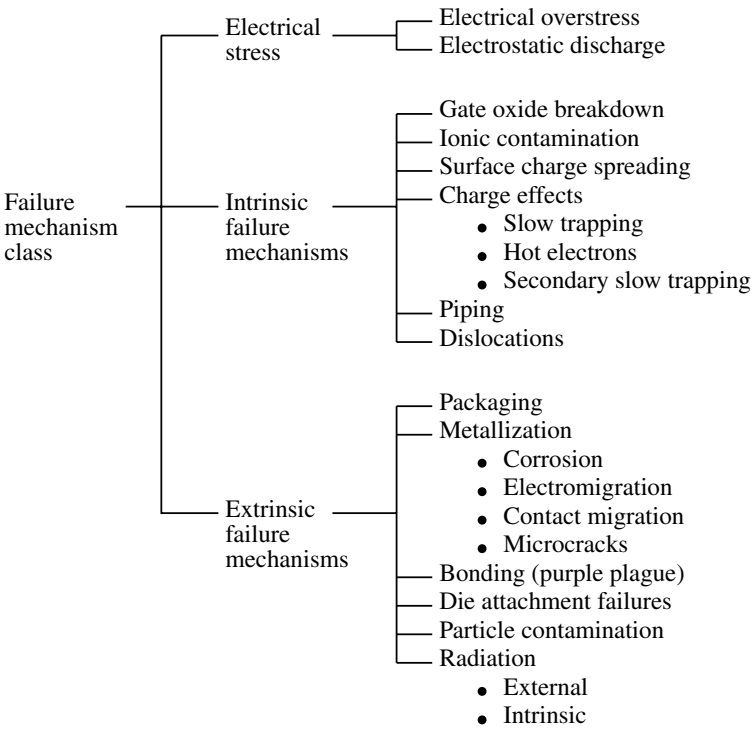


Figure 1.3 Classification of failure mechanisms

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).