Springer Series in Reliability Engineering

T. Tinga

Principles of Loads and Failure Mechanisms

Applications in Maintenance, Reliability and Design

Springer

Springer Series in Reliability Engineering

Series Editor

Hoang Pham

For further volumes:
http://www.springer.com/series/6917

T. Tinga

Principles of Loads and Failure Mechanisms

Applications in Maintenance, Reliability and Design

Springer

T. Tinga
Netherlands Defence Academy
Den Helder
The Netherlands

ISSN 1614-7839
ISBN 978-1-4471-4916-3 ISBN 978-1-4471-4917-0 (eBook)
DOI 10.1007/978-1-4471-4917-0
Springer London Heidelberg New York Dordrecht

Library of Congress Control Number: 2012956164

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The failure of structures or (parts of) systems is a common problem in practice. In all sectors of industry, from transport (air, rail, water, and road) to process industry, energy generation and high tech manufacturing industry, complex capital assets are used and must be maintained to assure their failure-free operation. In many cases the failed parts can be repaired or replaced, after which the system can fulfil its intended function again. In other cases, failures must be prevented against all costs, as the consequences for the system or its surroundings can be significant or even disastrous. Examples of the latter are the failure of critical aero engine parts leading to aircraft crashes or failing safety systems in a nuclear plant.

For many other failures, a deliberation must be made between the costs of preventive replacements and the costs and collateral damage of a failure. However, to be able to make a solid consideration, it is important to understand why components fail and how they fail. Moreover, to effectively plan preventive measures, it is also crucial to be able to calculate when a component will fail. Especially those questions will be answered by this book.

The large variety of failure mechanisms, i.e. the ways in which parts or systems fail, will be treated extensively. Knowledge on these mechanisms, and especially the effect of the governing loads on failure, are essential to understand why, how and when components fail and how this can be prevented.

Goal—The main goal of this book is to make the reader aware of the relation between the operation of any asset or system, the resulting loads on (parts of) the system and the possible consequences in terms of failures. In addition to the awareness, it also aims to provide quantitative relations enabling more detailed analyses of failing systems.

Therefore, an overview of the most important failure mechanisms is provided, treating for each mechanism the basic failure processes on a material level, but also discussing the quantitative dependence on applied loads. The book does not claim to be complete in providing all the details on all specific mechanisms. For that purpose references to other books and papers are provided. However, the focus here is on providing an overview over the possible mechanisms and showing the similarities in the load-to-failure relations.

For students, Part I of the book provides a basic introduction to the field of loads and failure mechanisms and creates the important awareness for their interdependence. The contents fit well in courses on physics of failure, structural integrity, design, maintenance, and reliability.

For practitioners from industry, the book offers an accessible and rather complete overview of possible failure mechanisms and associated loads. This will assist them in solving practical problems on failures in all kinds of machines and systems. This is supported by the lists of common failures and the decision scheme in Chap. 8.

For researchers, Part II of the book points out the interesting links between the multiple disciplines associated to failure and maintenance challenges. Where research on failure mechanisms is mostly limited to the fields of materials science and mechanical, thermal and electrical engineering, the chapters in Part II of this book show that there are interesting connections with many other disciplines like statistics, stochastics and reliability engineering, design, monitoring, and prognostics, but also non-technical disciplines like business and economics.

Outline—The book is organized as follows. First, it is divided into two parts. In Part I the basic principles of loads and failure mechanisms are treated. This part is particularly suitable for use in an introductory course on the physics of failure. Part II discusses how the detailed knowledge on loads and failure mechanisms can be applied in maintenance, reliability and design. In this part, existing practical methodologies are combined with recently developed concepts to provide innovative solutions to optimize maintenance and design processes. This part is therefore particularly interesting for both engineers/managers in industry and researchers/academics active in the fields of maintenance and reliability engineering and maintenance management, as well as for students taking more advanced courses on maintenance management and optimization.

Within each part, the following topics will be treated.

Part I—Basic Principles of Loads and Failure Mechanisms

In Chap. 1 the balance between loads and load-carrying capacity will be discussed, which is the basic concept of the present book. Also, the difference between external and internal loads and the different load types will be introduced. Chapter 2 will provide an overview of a large number of external loads that can act on systems or parts. For each load type, the generic load will be defined and the various specific sources for that load will be discussed. In Chap. 3 the internal loads, i.e. the loads acting on the material level, will be treated. The external loads introduced in Chap. 2 will be translated into internal loads, which govern the failure behaviour. Finally, in Chap. 4 a rather extensive overview of failure mechanisms will be provided, ranging from mechanical mechanisms like fatigue, creep and wear to thermal and electric failure mechanisms.

Part II—Applications in Maintenance, Reliability and Design

Chapter 5 introduces the basic concepts of maintenance, providing their definitions, but also giving an overview of the various maintenance policies. Many of the topics in consecutive chapters will build on the basics in Chap. 5. Chapter 6 then provides a detailed discussion of load and usage-based maintenance policies,

where the knowledge on failure mechanisms from Part I is demonstrated to yield a large potential in improving maintenance efficiency. In Chap. 7 the link between failure mechanisms and reliability engineering methods is discussed, showing that combining these fields offers a large potential for maintenance improvements. Chapter 8 focuses on the analysis of failures, either before or after they have occurred. Useful methodologies are discussed and case studies demonstrate the practical application. Also, a decision scheme to support the determination of the failure mechanism for an actual failure is provided here. Finally, Chap. 9 treats the various aspects of the design process where knowledge of the failure mechanisms may play a role. Topics like life cycle management, design philosophies and probabilistic design are discussed here.

Examples—Especially, in Part I of the book, a considerable number of examples have been provided within the text. In these (worked) examples the theory treated in the text is translated into a practical problem, showing how the theory can be applied and what typical values the quantities introduced in the text could attain. These examples are typically useful for students in practising and rehearsing the treated topics.

Acknowledgements—Writing this book has been quite an effort and would not have been possible without the help of many people. The initial idea of writing a book on failure mechanisms came from my former colleague Cees Tromp, who already experienced that not many books on this topic exist. Based on his previous work and contributions from Theo Popma, we started to write two readers in Dutch, which could be used in the courses at the Netherlands Defence Academy. After the invitation by Springer to write a full book on the topic, these have been transformed into the present book. Moreover, it has been extended with a part on applications based on our present research program.

During the process of writing the book, many people have provided input or have reviewed parts of the text. I wish to sincerely thank the following people for their valuable input and feedback: Alex de Goeij, Henk Jan ten Hoeve, Axel Homborg, Marc Masen, Dick Meuldijk, Ed Reddering, Cyp van Rijn, Rob Ross, Wieger Tiddens, Arjen Vollebregt, Sander Wanningen, and Martijn Woldman. Finally, I also thank Yvonna and our kids Jasper and Nynke for their moral support during the writing process.

Contents

Part I
Principles of Loads and Failure Mechanisms

In this first part of the book the basic principles of loads and failure mechanisms will be treated. After an introduction on the balance between loads and capacity, the details of external and internal loads are treated. In the final chapter all relevant failure mechanisms are discussed. The second part of this book will then focus on the applications of these basic principles in maintenance, reliability and design.

Chapter 1
Introduction: The Basics of Failure

1.1 Failure

Before discussing the details of failures, it is important to define what failure is, since that term is not unambiguous. Throughout this book, failure will be considered as reaching such a state that the intended function of the part or system can no longer be fulfilled. Therefore, failure does not always imply the real physical failure of a part, like fracture or melting, but could also be the result of extensive deformation leading to rubbing or seizure of a rotating part.

Moreover, it depends on what level is considered. Failure of a specific part or subsystem does not automatically imply that the complete system fails. A plant equipped with several pumps does not stop when only one pump fails. In that case, a failure occurs on the subsystem level (pump), but no failure occurs on the system level (plant). This difference between component and system reliability of failures will appear at several instances in the book.

Another important issue when describing failures is that a division should be made between failure mode and failure mechanism. The *failure mode* is the manner in which a system or component functionally fails, that is, describing to what extent a certain function cannot be fulfilled anymore. This means that the failure modes are generally related to the performance requirements of the system. A non-performance can be detected either during operation from an observed decrease in the system performance (e.g. a reduced capacity of a pump) but also from periodic inspections (mostly executed when the system is not operating), indicating that certain limits with respect to system or component condition have been exceeded. Examples of the latter are excessive elongation of a gas turbine blade, unacceptable crack lengths in an aircraft structure or large areas of coating delamination due to corrosion on the hull of a ship.

Further, failure modes can be defined on several levels of hierarchy. For example, one of the failure modes of a car is a non-working engine. The engine also has several failure modes, like a lack of fuel, a broken crankshaft or a defect

T. Tinga, *Principles of Loads and Failure Mechanisms*,
Springer Series in Reliability Engineering,
DOI: 10.1007/978-1-4471-4917-0_1, © Springer-Verlag London 2013

engine management computer. In this way, after several steps, the failure on the system level can be attributed to failure modes of individual components. A structured method to decompose a system in this way will be introduced in Chap. 8.

A failure mode on the component level can then still have several potential causes, including both non-physical and physical failures. Non-physical failures are mostly due to human errors, for example, application of wrong types of fuel or lubricants, or due to contamination (dust, fouling). In these cases, the failure mode (i.e. functional failure) is observed, without any actual physical failure occurring. Physical failures, on the other hand, are due to a physical or chemical process or mechanism yielding degradation of the component and ultimately leading to the physical failure. This process is called the *failure mechanism*.

Whether this failure mechanism will be active and whether it will lead to failure within a certain operational period depends on the magnitude of the applied loads, as will be discussed in the next subsection. However, it should be noted that also true physical failures can be caused by human errors, for example, when a system is operated in a wrong way. This means that finding the *root cause* of a certain failure mode requires determination of both the failure mechanism and the origin of the loads that activate the mechanism, as will be discussed in detail in Chap. 8.

Finally, it should be noted that only a limited number of different failure mechanisms exist. And although a specific failure mode of a component can be caused by several of these failure mechanisms (e.g. corrosion, fatigue, wear), the number of possible mechanisms is much smaller than the almost infinite number of failure modes that can be defined for all possible systems, subsystems and components. Therefore, reducing an observed failure mode to the root cause in terms of failure mechanism and governing load significantly simplifies the solution process, but also requires detailed understanding of all failure mechanisms. That is the main motivation and aim of the present book.

1.2 Balance

In a general sense, a failure process is always a result of an imbalance between the load on a system and that system's load-carrying capacity. The *load* is the extent to which a part or system is exposed to external influences. Generally, a load is associated to a mechanical load, like an applied force or a deformation. However, the concept of load is much broader than that, since also a thermal, electrical or chemical load can be concerned. The *load-carrying capacity* of a part or system is the extent to which it can resist or withstand the external influences without failing. Examples are the strength of a structure (i.e. mechanical load) or the corrosion resistance of a material (chemical load). In the end, the question whether a system will fail can be answered by looking at the relative magnitude of the load compared to the load-carrying capacity. If the load exceeds the capacity of the system, it will fail. This is visualized schematically in Fig. 1.1.

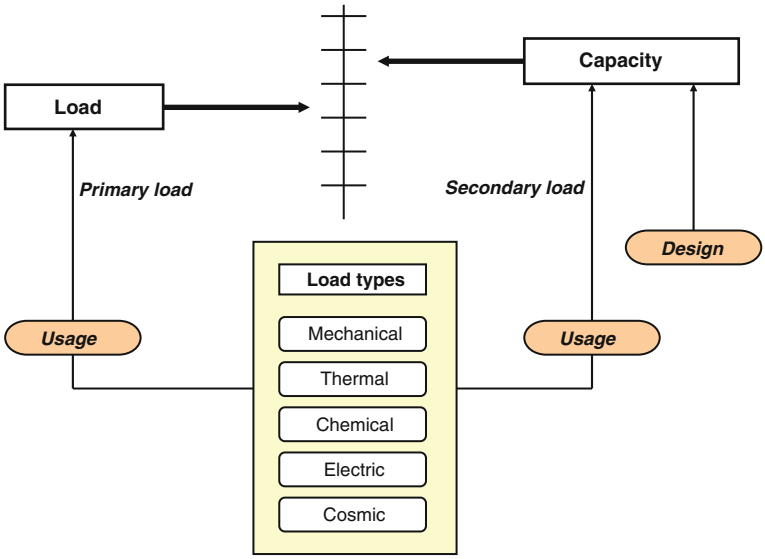

Fig. 1.1 Schematic representation of the relation between load and load-carrying capacity

Both the loading and the capacity of a system can be influenced, but in very different ways. The load-carrying capacity of a system is largely determined during the design phase, where materials and manufacturing methods are selected and decisions on shape and dimensions are made. Therefore, once a system has been manufactured, the possibilities for increasing the load-carrying capacity are limited. Only modifications or redesigns, which are often quite elaborate and costly, can then solve the problem.

The loading of the system, on the other hand, is mainly determined by the way the system is used or operated. Quantities like rotational speed, heat production or applied voltage, resulting from a specific usage, will determine the loads that occur in or on the system. Thus, the operator has control over the loads on the system and can use that control to prevent failures. These loads resulting from the usage of the system are to be compared directly to the capacity and are called *primary loads*. However, in many cases, the loads also affect the load-carrying capacity. For example, the strength of a material will decrease when the temperature rises (thermal load), and the mechanical properties of plastics will deteriorate after a prolonged exposure to UV radiation (radiative load). The latter loads are called *secondary loads* as they only affect the balance between loads and capacity in an indirect manner.

1.3 External and Internal Loads

Failure is a local process that arises at the material level. For instance, cracks in a part originate when in the crystal lattice of the material the bonds between several atoms break up. Similarly, electric breakdown of an insulator occurs when a locally large electric field forces electrons to separate from their atoms, resulting in a discharge. Therefore, to be able to assess whether the local load-carrying capacity is sufficient, that is, whether failure will occur or not, also the load on that local level must be known. But in many instances, only the global loads that are applied to the system externally are known. In those cases, the *external loads* must be translated into the local *internal loads*. For example, the heat production in a total system (global) will result in a certain temperature at a specific location within that system (local). And the local stress level at a specific location in a structure will be determined by the complete set of externally applied forces. The translation of external loads into internal loads will be discussed in detail in Chap. 3.

1.4 Load Types and Failure Mechanisms

In Sect. 1.2, the term load has been defined in a very general sense, because a considerable number of different types of loads exist. At the same time, also the capacity of the system is a versatile concept, since it is dictated by a range of failure mechanisms. Figure 1.2 provides an overview of all these different (primary) loads (left-hand column) and the associated failure mechanisms (right-hand column). The centre column of the figure indicates which internal load parameter is associated to each external load type.

The figure shows that some failure mechanisms are only affected by one single load parameter, while for others several load parameters play a role. For instance, for the failure mechanism corrosion, the degradation rate is determined by a combination of thermal (temperature), chemical (acidity, pH value) and electric (current) loads. On the other hand, the different load types will also act as secondary loads in decreasing the capacity. Especially for electric failures, the secondary loads are often responsible for such a reduction in capacity that failure due to the governing primary load takes place.

1.5 Character of the Load

Apart from the division into different load types, as was discussed in the previous section, loads can also be divided according to the character of the load. This means that a division can be made between static and dynamic loads on the one hand and between deterministic and stochastic loads on the other hand.

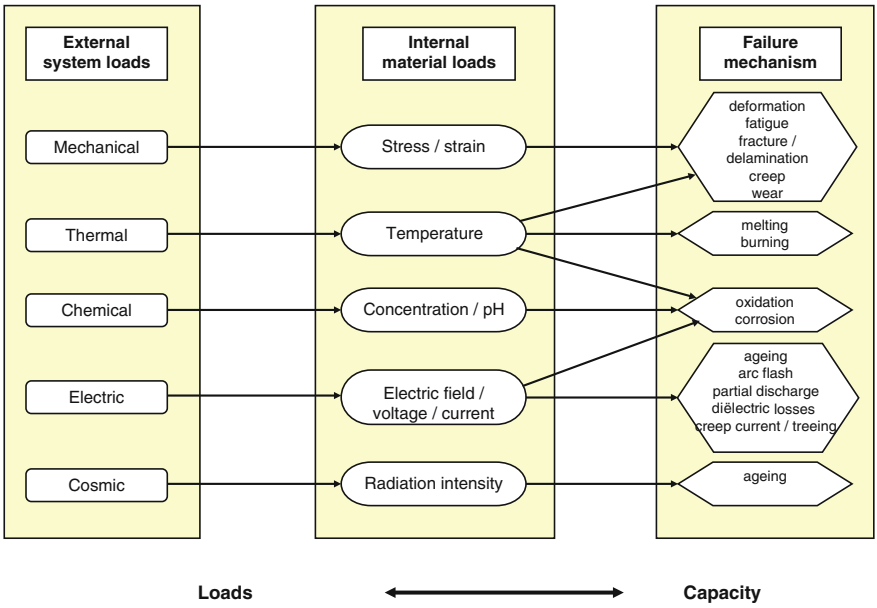

Fig. 1.2 Overview of primary load types and the associated failure mechanisms

The difference between static and dynamic loads originates from the amount of variation of the load in time. When a load is constant in time, it is called a static load. An example is the loading of a bridge resulting from its own weight. A load that varies in time, either in magnitude or in direction, is called a dynamic load. The loading of the bridge caused by passing traffic is therefore a dynamic load, that is, the weight of the cars is moving along the bridge and varies in time.

The difference between deterministic and stochastic loads originates from the extent to which the load is a priori known or can be predicted unambiguously. When the load is fully known, it is called deterministic. Referring to the previous example, the loading of the bridge due to its weight is deterministic: the weight is known and the load can be calculated. If, on the other hand, the load is unknown, for example, because it varies in an unpredictable way, it is called a stochastic load. The loading of the bridge by the passing traffic is a good example: the number of passing cars at any moment is hard to predict.

There are several ways to overcome the problem of a (partly) unknown load. Firstly, the upper limit of the load can be used in calculations. Although the exact load is unknown, this upper limit can in most cases be estimated rather easily. In the bridge example, the exact number of cars is unknown, but the maximum number of cars on the bridge at any moment can be estimated. A second way to handle the stochastic loads is the application of a distribution function. Such a function indicates the probability that a certain load (magnitude) will occur, which can be used to calculate the probability of failure.

The maximum loads on the Dutch sea walls arise when a spring tide and a storm coincide. The actual load is thus governed by the position of the Moon and Sun (deterministic), the wind force, direction and the duration of the storm (stochastic). These variables together determine the probability that the dike will fail. The sea walls have been designed such that a failure will occur only once in 10,000 years, which is considered to be an acceptable risk.

The same principle of accepted risk is applied in the design of aircraft and ships, which are loaded by wind gusts and waves, respectively. Also in this case, the actual load and the resulting material stress is a stochastic variable. For both wind gusts and waves, statistical distributions are available, which can be used to calculate the extreme loads with very small probability of occurrence during the service life of the aircraft or ship. The design calculations are based on these extreme loads, which means that material selection and decisions on dimensions are such that the structure is able to withstand these loads. This probabilistic approach to failure or design will be treated in more detail in Chaps. 7 and 9.

1.6 Perspectives

From the previous sections, it is apparent that exceedence of the load-carrying capacity by the actual loads yields failure of the system. In the next chapters, the different load types and failure mechanisms will be discussed in detail, thus providing insight into the possibilities to prevent failures in practice. However, that can be done from three different perspectives: design, operation or failure analysis.

From the perspective of the designer, the process generally starts with characterizing all the expected external loads, which are then translated into the expected internal loads on the material level. Comparison of these loads with the capacity of the system then determines whether the design is acceptable or should be modified. In the latter case, the capacity will be increased by selecting other materials or changing dimensions.

Also from the perspective of the operator, the loads are leading, since they are governed by the operation of the system. However, in case of a pending failure, not the design but the usage of the system will have to be changed, in order to reduce the loads.

The final perspective concerns the failure analysis that is executed after a failure has occurred. In this case, the opposite way is travelled: from the observed failure mechanism, one tries to determine what internal load has caused the failure and from what external load it originated. The goal is then to determine whether the failure was due to a design error (capacity too low) or due to an operational fault

Fig. 1.3 The usage of a system determines the service life consumption

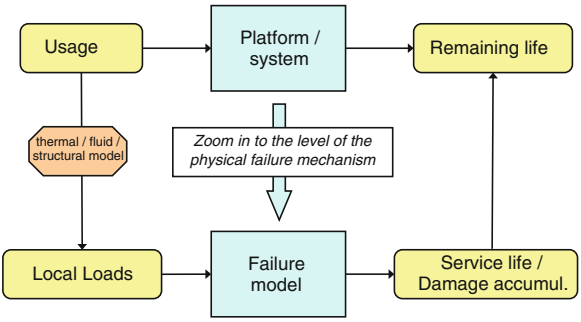

Fig. 1.4 Zooming into the level of the physical failure mechanisms provides insight into the quantitative relation between usage and service life consumption

(load too high). The outcome is in many cases very relevant for insurance companies, as it puts the blame at either the operator or the manufacturer of the failed system.

This book focuses specifically on the operator perspective, with the aim of providing the reader with sufficient insight into the effect of the use of a system on the different failure mechanisms. This insight is essential to understand failures that take place in practice or to prevent these failures to occur. On the one hand, that improves the safety and the availability of a system, since unexpected failures can be prevented, while on the other hand the maintenance process can be optimized, which reduces the life-cycle costs of the system.

The latter aspect of maintenance is illustrated using Fig. 1.3. For any asset or system, the rate at which the service life is consumed is determined by the specific usage of that asset. The remaining service life of a system is important information for an operator. However, whereas in many cases the usage of the asset is known, the remaining life is not, since the relation between these two is a 'black box'.

At this point, knowledge on the loads and failure mechanisms is required to get insight into that relation between usage and service life consumption. This is shown in Fig. 1.4. With detailed knowledge on the different loads (as a function of usage) and failure mechanisms, that will be provided in the next chapters in this book, the maintenance intervals of the asset can be adapted to the specific usage of a system. In that way, maintenance can be performed just-in-time and over-maintenance (high costs) or under-maintenance (many failures) can be prevented. This usage-based approach to maintenance will be discussed in more detail in Chap. 6.

1.7 Summary

In this chapter, the definition of a failure is provided and the difference between a failure mode and failure mechanism has been explained. Then, the concepts of load and load-carrying capacity have been introduced and their mutual relation is discussed. It has been shown that failures occur when the load on the system exceeds its capacity. Moreover, the difference between internal and external loads

is treated and the large variety of different load types and failure mechanisms is illustrated. The difference between firstly static and dynamic loads and secondly deterministic and stochastic loads has been explained, and the different perspectives to look at failures have been presented: design, operator and failure analysis. Finally, the possible role of knowledge on failure mechanisms in maintenance has been briefly introduced.

Chapter 2
External Loads

2.1 Introduction

In the previous chapter, it has been indicated that a large number of load types exist. The most important load types are shown in Fig. 1.2. Each of these load types can originate from a number of specific sources, but the resulting generic load is unique for a certain load type, as is depicted in Fig. 2.1. For example, a mechanical load will always yield a force or a moment (generic load), but the specific source of that force can be diverse: centrifugal load caused by a rotation, gravity, torque provided by a drive shaft or thermal expansion.

In this chapter, the different external load types are discussed and the possible sources are indicated. Using practical examples, it is shown for each source how the loads can be calculated.

2.2 Mechanical Loads

The mechanical loading of a structure is caused by forces. That can be either forces that act directly on a structure, like a concentrated or distributed force or a moment, or forces caused by an indirect load like pressure or weight.

A *concentrated force* is a force that acts on a single point, while a *distributed force* is exerted on a certain area or volume. A towing line connected to a car, for example, exerts a concentrated force on the towing ring, whereas a book resting on a table exerts a distributed force on the table due to its weight. A *moment* is caused by a force that is acting on a line not passing through the centre of gravity or rotation point of a body. The magnitude of the moment (M) is determined by the product of the force and the perpendicular distance between the working line and the line through the rotation point (moment arm).

T. Tinga, *Principles of Loads and Failure Mechanisms*,
Springer Series in Reliability Engineering,
DOI: 10.1007/978-1-4471-4917-0_2, © Springer-Verlag London 2013

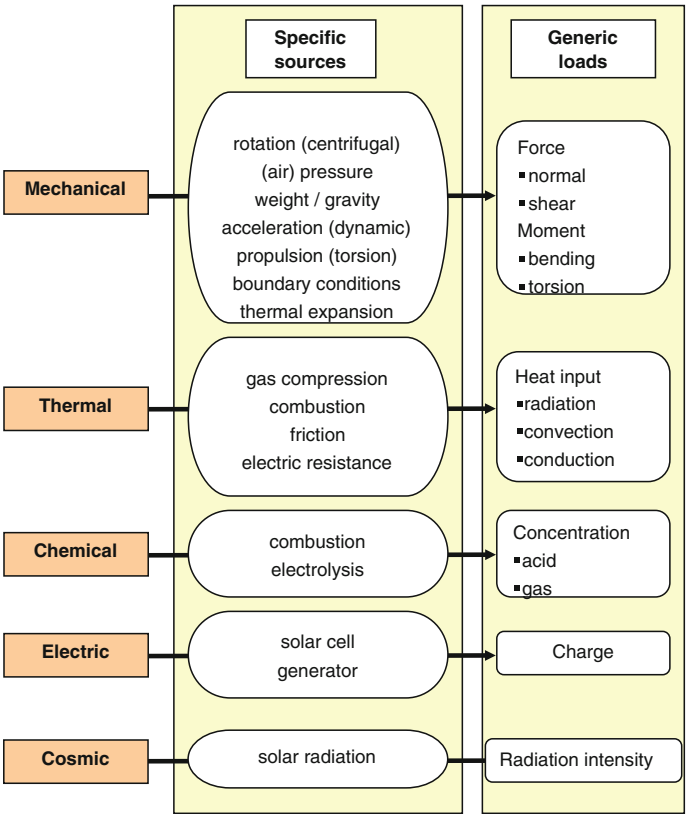

Fig. 2.1 Overview of load types, associated specific sources and resulting generic loads

Instead of acting directly on a body, forces can also originate from indirect loads, for example, caused by

- environment → (air) pressure
- weight → gravity
- accelerations → mass force
- drive/propulsion/guidance
- thermal expansion
- internal stress

In the next subsections, these sources will be discussed, and they will be illustrated using explicit examples.

A complete structure is generally loaded by a considerable set of concentrated and distributed forces, each acting at different locations. The combination of all these forces determines the total mechanical loading of the structure. As a result, one specific point in the structure can be loaded in tension or compression, bending or torsion. The translation of all the combined external forces into the local force at

a specific point is covered by the theory of statics. The basic principles of this theory will be treated in Sect. 2.2.7.

2.2.1 Environment and Pressure

It is very common in practice that a body experiences a load from its environment. In most cases, this concerns a pressure load on a body residing in a gas or liquid environment. Examples are the loading of a pressure vessel by the large internal gas pressure or the buoyancy force exerted on the hull of a ship. Also, the loads on an aircraft wing by lift and drag are examples of this type of load.

A pressure load delivers a distributed load f (N/m^2) acting on (part of) the surface of a body. The total exerted force F (N) can be calculated by integrating the distributed force over the surface area A (m^2)

$$F = \int_A f dA \tag{2.1}$$

Example 2.1 (Pressure Loads on an Aircraft Wing) During flight, the loading of an aircraft wing is determined by the pressure loads on the wing and the weight distribution of the plane. Figure 2.2 shows an aircraft wing profile in an airflow with speed v and an angle of attack α. The flowing air causes two forces on the wing: the lift force L perpendicular to the airflow and the drag D in the direction of v. The magnitudes of L and D depend on the speed v and the angle α.

The basic lift distribution $l(y)$ across the span B of the wing is shown schematically in Fig. 2.3. It is assumed here that this distribution is known. The actual lift distribution at specified flight conditions (aircraft weight, flight altitude, air speed, etc.) is obtained by multiplying the basic lift distribution $l(y)$ with the intensity factor A. In case of a stationary flight condition, the magnitude of A can be determined from the balance between aircraft weight W and the total lift L_w of the wings:

$$W = L_w = A \int_{-B/2}^{+B/2} l(y) dy \tag{2.2}$$

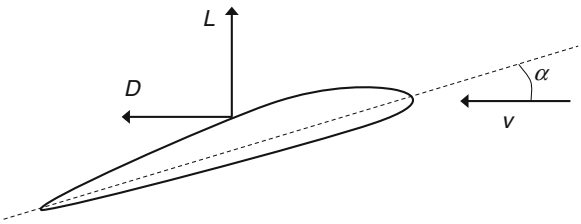

Fig. 2.2 Lift and drag forces acting on an aircraft wing

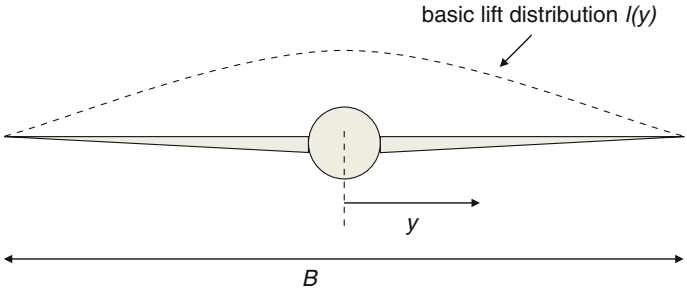

Fig. 2.3 Basic lift distribution across the span of the wings

In Fig. 2.4, a simplified basic lift distribution for an aeroplane is given. The mass of the plane, including fuel and payload, is $m = 45{,}200$ kg. The following values are assumed for the widths b_i of the sections: $b_1 = 1$ m; $b_2 = 7$ m; $b_3 = 3$ m; $b_4 = 2$ m, and for the corresponding basic lift forces: $l_1 = 0.3$ N/m; $l_2 = 1.0$ N/m; $l_3 = 0.7$ N/m; $l_4 = 0.4$ N/m. Using the gravity constant $g = 9.81$ m/s^2, the intensity factor A can be calculated from (2.2). This yields the value $A = 21{,}714$.

Since the mass of the aircraft is concentrated at the fuselage, the downward force (gravity) acts on that centre part of the plane. On the other hand, the upward force caused by the lift distribution acts on the wings of the aircraft. This difference in acting points results in bending of the wings, causing relatively high (bending) loads at the wing root.

In reality, the plane will not constantly be in a stationary situation with an exact balance between lift and weight. Due to air turbulence and gust, the lift distribution will change frequently (e.g. due to a change in entry angle α, Fig. 2.2), yielding an imbalance between the two forces and a resulting up- or downwards motion of the plane.

2.2.2 Weight

The weight of a structure generally causes a distributed load that depends on the mass distribution. The gravity force W is proportional to the mass of each volume element, with the gravity constant $g = 9.81$ m/s^2 as the proportionality constant,

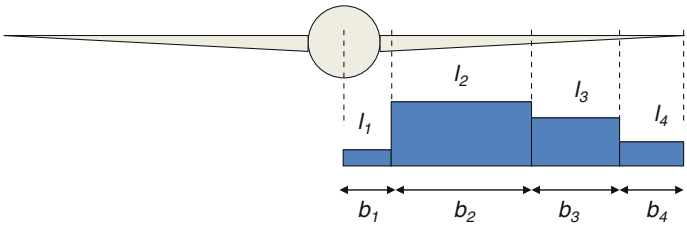

Fig. 2.4 Block wise basic lift distribution across the span of one wing

and acts on the centre of gravity of the element. The total gravity force, or weight, is obtained by integrating the distributed load over the complete volume

$$W = \int_V g \, dm \tag{2.3}$$

Example 2.2 (Gravity Loading of an Aircraft Wing) The wing of the aircraft from the previous example is not only loaded by the upward lift forces, but also by the downward gravity loading on the wings. This means that both the lift distribution and mass distribution across the span of the wing play a role. In this example, the mass distribution shown in Fig. 2.5 will be assumed.

The given mass distribution $m(y)$ (kg/m) is multiplied by the gravity constant g (m/s^2) to obtain the weight distribution $W(y)$ (N/m)

$$W(y) = m(y)g \tag{2.4}$$

Using values of $c_1 = 9$ m; $c_2 = 4$ m; $\xi_1 = 500$ kg/m; $\xi_2 = 150$ kg/m, the total gravity force or weight for one wing is obtained through

$$W = \sum_i m_i g = \sum_i c_i \xi_i g = 50 \text{ kN} \tag{2.5}$$

2.2.3 Acceleration

In Sect. 2.2.1, it was mentioned that gust loads cause an aircraft to move up- and downwards. According to Newton's law,

$$F = ma \tag{2.6}$$

the associated accelerations yield additional mass forces. A similar phenomenon is present in rotating machinery, in which rotating parts are subject to centripetal accelerations. To keep the parts in their rotating motion, the resulting centrifugal force (acting in an outward radial direction) must be balanced by an inward

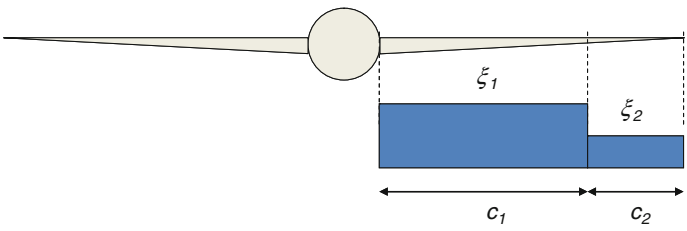

Fig. 2.5 Mass distribution across the span of the wing

centripetal force. The centrifugal force acting on a mass m rotating with an angular velocity ω (rad/s) in a circular motion with radius r is given by

$$F_{cf} = m\omega^2 r \tag{2.7}$$

Both these forces caused by different types of accelerations will be illustrated in the next two examples. In Example 2.5 in the next subsection, another case of mass forces is illustrated.

A final remark on this topic concerns the principle of balancing. In many rotating and reciprocating machines, the loads on the parts are reduced considerably by balancing the forces acting on these parts by equal (but oppositely directed) forces. During operation, defects in the machine can cause (partial) distortion of this balance, which may lead to loads that are considerably higher than the expected loads.

Example 2.3 (Mass Forces on an Aircraft Wing) In a stationary flight, the lift and gravity forces are in balance, that is, $L_0 = mg = W$. The ratio n of lift force over gravity force therefore equals 1. During a gust (e.g. a vertical upward wind blast), the lift is temporarily increased, resulting in a n-value greater than 1 and an upward movement of the aircraft. This means that the additional lift $\Delta L = L_0 \Delta n$ yields a vertical acceleration of the plane equal to

$$\ddot{z} = \frac{L_0\,\Delta n}{m} = \frac{mg\,\Delta n}{m} = g\,\Delta n \tag{2.8}$$

Due to this acceleration, additional forces are acting on the wings with a magnitude of

$$F(y) = m(y)\ddot{z} = m(y)g\,\Delta n \tag{2.9}$$

Example 2.4 (Centrifugal Force on a Gas Turbine Blade) Figure 2.6 schematically shows the shaft of a gas turbine, to which a blade is attached. For a rotating element with mass dm located at a distance r from the centre of the axis, which is moving at a rotational speed ω, the centripetal acceleration is $\omega^2 r$. This acceleration yields a centrifugal force $\mathrm{d}F_{\mathrm{cf}}$ on the element equal to

$$\mathrm{d}F_{cf} = \omega^2 r\,\mathrm{d}m \tag{2.10}$$

Assuming that the cross-sectional area A of the blade is constant across its length and that the mass density of the material is ρ, the mass of the element is

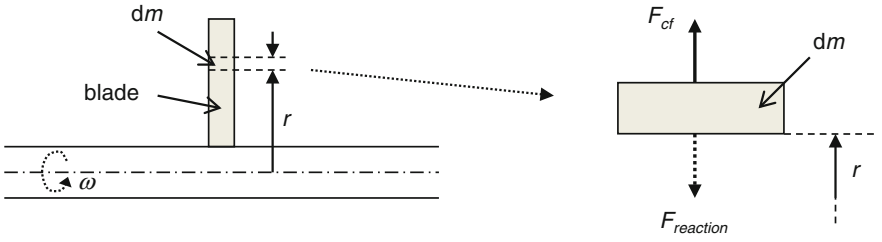

Fig. 2.6 Loading of a turbine blade by the centrifugal force

$dm = \rho A dr$. Therefore, the increase of the centrifugal force within the element is

$$dF_{cf} = \rho A \omega^2 r \, dr \tag{2.11}$$

The total centrifugal force at any radius is then obtained by integrating this expression to r and using the boundary condition that the centrifugal force is zero at the blade tip, that is, $F_{\mathrm{cf}}\,(r = R_1) = 0$:

$$F_{\mathrm{cf}} = \frac{1}{2}\rho A \omega^2 \left(R_1^2 - r^2\right) \tag{2.12}$$

The centrifugal force thus decreases quadratically with r and has a maximum value at the root of the blade ($r = R_0$).

2.2.4 Drive/Propulsion/Guidance

In many machines, subsystems are driven by motors, where forces and moments are transferred with the purpose of getting or keeping parts of the subsystem in motion. On the other hand, the integrity of the system must be preserved, which means that at other parts of the machine motions must be prevented. Examples of components providing this restriction are bearings and movement limiters. In all these cases, mechanical loads are exerted on the different parts.

In the examples below, the loading of a piston engine crankshaft is discussed, as well as the support of the shaft by journal bearings. In the second example, the loads on the elements of a roller bearing are treated.

Example 2.5 (Piston Engine Crankshaft) Figure 2.7 schematically shows a cross section of one cylinder of a piston engine. The piston performs an oscillating

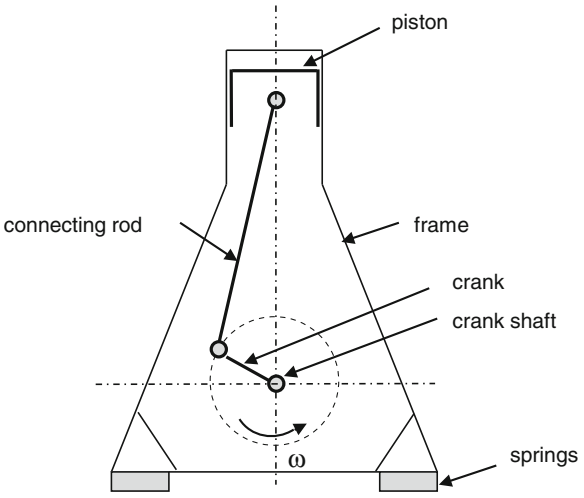

Fig. 2.7 One-cylinder piston engine

motion that is transferred by the connecting rod and the crank, resulting in a rotation of the crankshaft.

The loads acting in the machine are the gas pressure in the cylinder, the mass forces of the moving piston and connecting rod and the centrifugal forces on the rotating parts.

Gas Forces

The gas force F_{gas} acting on the piston is transferred through the piston pin to the connecting rod, while the frictionless cylinder wall exerts a (reaction) force F_{wall} on the side of the piston, see Fig. 2.8. As all forces acting on the piston must be in balance, it follows that

$$F_{\text{wall}} = F_{\text{gas}} \tan \beta \tag{2.13}$$

The forces F_{gas} and F_{wall} are transferred to the crankshaft and firstly yield the loads on the main bearing, which ensures the proper positioning of the crankshaft, and secondly the torque T_{gas} on the shaft. The forces on the bearing exactly equal the forces on the piston (force equilibrium, see Fig. 2.8a), and the torque is given by

$$T_{\text{gas}} = F_{\text{wall}} s \tag{2.14}$$

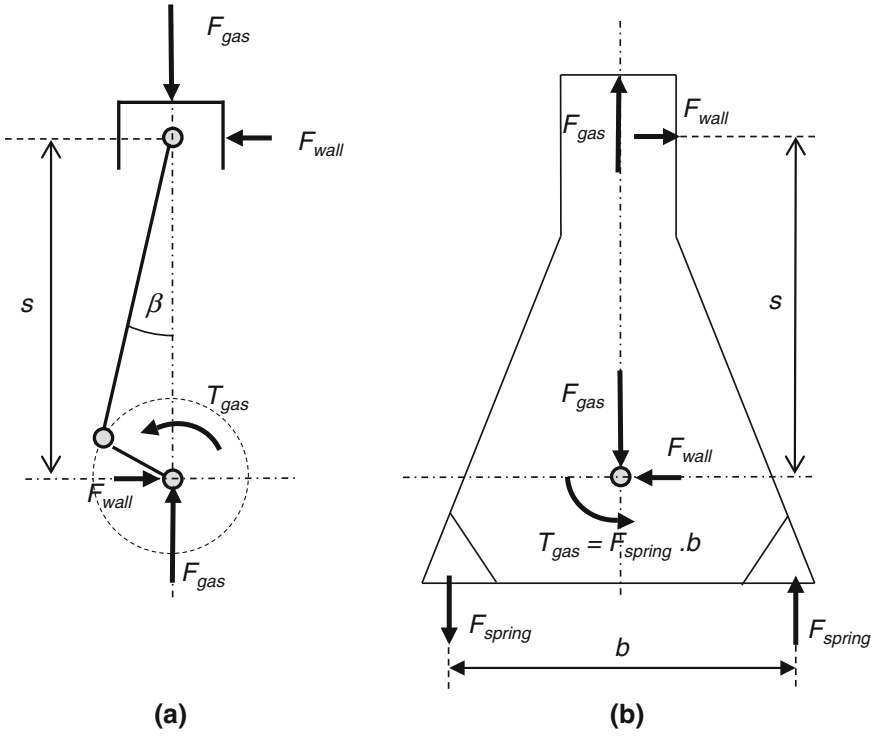

Fig. 2.8 Overview of loads on **a** driving mechanism and **b** frame

where s is the distance between the centre of the crankshaft and the piston pin. Note that the vertical forces (gas force and its reaction) are exactly on the line through the centre of the crankshaft, which means that no moments or torques are initiated by these forces.

To ensure structural integrity, all forces and moments on the driving mechanism (Fig. 2.8a) must be counterbalanced by forces on the frame of the engine. Figure 2.8b shows that force equilibrium can be attained within the frame, but the torque T_{gas} is transferred to the foundation of the engine by the springs.

Mass Forces

The motion of the driving mechanism is governed by an oscillating (vertical) movement of the piston and a rotating motion of the crankshaft. These motions also occur at the ends of the connecting rod. To be able to calculate the mass forces on the rod, the mass of this body is split into two separate mass points positioned at the piston pin and the crank pin, respectively. The total oscillating mass then equals the mass of the piston plus the *oscillating* mass of the connecting rod. The total rotating mass consists of the mass of the crank and the *rotating* mass of the connecting rod.

Oscillating Mass Force

Calculation of the oscillating mass force requires the acceleration of the piston to be known, which is calculated next (Fig. 2.9).

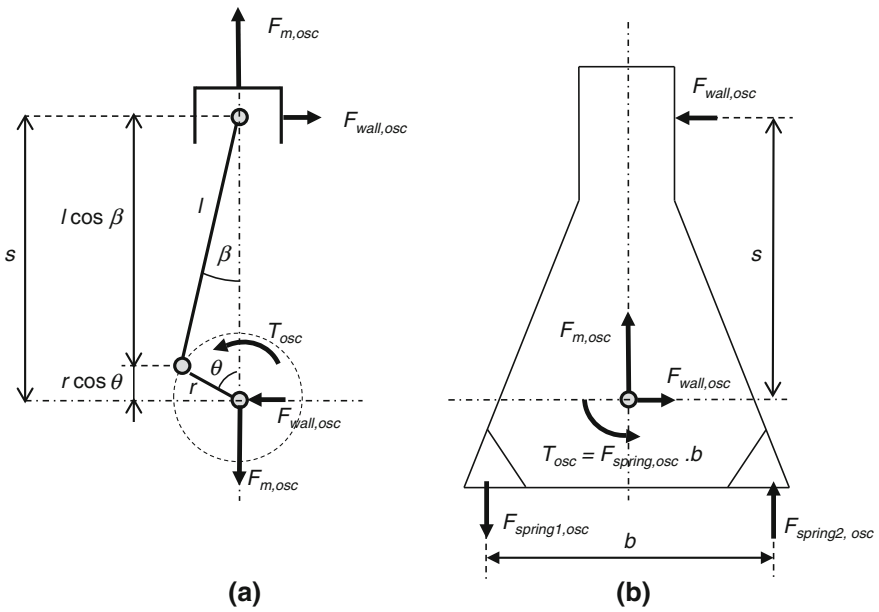

Fig. 2.9 Oscillating mass force acting on **a** the driving mechanism and **b** the frame

The position of the piston is given by the coordinate s defined as

$$s = r\cos\theta + l\cos\beta \tag{2.15}$$

Using $r\sin\theta = l\sin\beta$, $\cos\beta = \sqrt{1 - sin^2\,\beta}$ and $\lambda = \frac{r}{l}$, Eq. (2.15) yields

$$s = l\left(\lambda\cos\theta + \sqrt{1 - \lambda^2\sin^2\theta}\right) \tag{2.16}$$

Then, the time derivative of s is

$$\dot{s} = l\left(-\lambda\sin\theta - \frac{\lambda^2}{2}\frac{\sin 2\theta}{\sqrt{1 - \lambda^2\sin^2\theta}}\right)\dot{\theta} \tag{2.17}$$

And assuming a uniform motion, $\dot{\theta} = \text{constant} = \omega$, the acceleration (2nd time derivative of s) is given by

$$\ddot{s} = l\left[-\lambda\cos\theta - \frac{\lambda^2}{2}\left(\frac{2\cos 2\theta}{\sqrt{1 - \lambda^2\sin^2\theta}}\right) + \frac{\lambda^2}{2}\frac{\sin^2 2\theta}{\left(1 - \lambda^2\sin^2\theta\right)^{3/2}}\right]\omega^2 \tag{2.18}$$

By neglecting the terms with λ^2 (since $\lambda^2 \ll 1$), the oscillating mass force can be approximated as

$$F_{m,\mathrm{osc}} = m_{\mathrm{osc}}\omega^2 r(\cos\theta + \lambda\cos 2\theta) = m_{\mathrm{osc}}\omega^2 r(\cos(\omega t) + \lambda\cos(2\omega t)) \tag{2.19}$$

This mass force contains two harmonics that vary with the single and double rotational speed. The oscillating mass force is transferred to the main bearing in the same way as the gas force and also contributes to the torque

$$T_{\mathrm{osc}} = F_{\mathrm{wall,\,osc}}s \tag{2.20}$$

where

$$F_{\mathrm{wall,\,osc}} = F_{m,\,\mathrm{osc}}\tan\beta \tag{2.21}$$

A closer look at the loads on the frame due to the gas force and mass force, respectively, yields some remarkable observations (see Fig. 2.8). For the gas force, equilibrium exists between the forces (F_{gas}) on the cylinder head and on the main bearing. For the frame, also a force equilibrium exists, while the torque is transferred to the mounts of the frame. For the oscillating mass force, both the forces and the torque are transferred to the mounts.

Rotating Mass Force

The rotation of the mass m_{rot} yields a centrifugal force consisting of a horizontal and vertical component:

$$F_{\mathrm{mrot},\,h} = m_{\mathrm{rot}}\omega^2 r\sin\theta \tag{2.22}$$

$$F_{\mathrm{mrot},v} = m_{\mathrm{rot}}\omega^2 r\cos\theta \tag{2.23}$$

The rotating mass force is also transferred to the mounts by the springs, see Fig. 2.10.

Summary of Loads

The total loading of one single cylinder of a piston engine thus consists of the following contributions:

- The crank shaft is loaded by a torque, given by

$$T = T_{\mathrm{gas}} - T_{\mathrm{osc}} \tag{2.24}$$

  As a reaction force, this torque is also acting on the frame to provide the moment equilibrium.
- The main bearing is loaded by a combination of forces, where the horizontal and vertical components are given by

$$F_h = F_{\mathrm{wall}} - F_{\mathrm{wall,\,osc}} + F_{\mathrm{rot},\,h} \tag{2.25}$$

$$F_v = F_{\mathrm{gas}} - F_{m,\,\mathrm{osc}} - F_{\mathrm{rot},\,v} \tag{2.26}$$

- On the frame, the unbalanced mass forces are acting

$$F_h = -F_{\mathrm{wall,\,osc}} + F_{\mathrm{rot},h} \tag{2.27}$$

$$F_v = F_{\mathrm{gas}} - F_{m,\,\mathrm{osc}} - F_{\mathrm{rot},\,v} \tag{2.28}$$

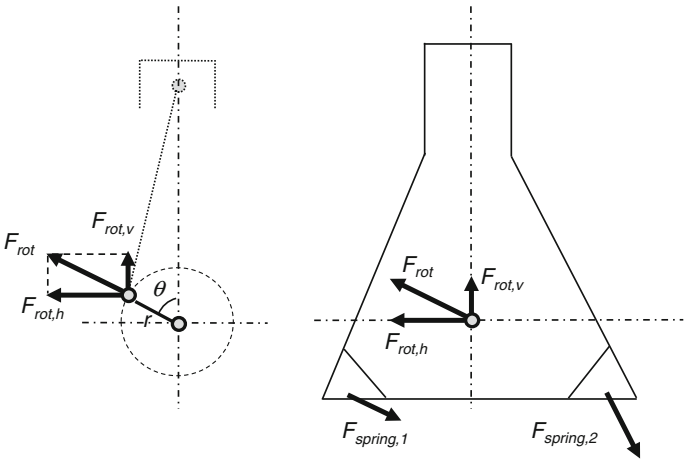

Fig. 2.10 Rotating mass force acting on the driving mechanism (*left*) and the frame (*right*)

Inline Piston Engine

An n-cylinder inline engine can be considered as a series of n connected single cylinder engines. To reduce the variations in torque applied to the crankshaft, the fuel in the separate cylinders is ignited equidistant in time. This is accomplished by varying the crank angles, as is shown for a three-cylinder engine in Fig. 2.11. The mass forces of the three cylinders and the corresponding moments around y- and x-axes are partly in balance. This will be illustrated next for the centrifugal forces.

Figure 2.12 shows the rotating mass forces of the three cylinders. Vectorial summation of these forces shows that the resultant force is zero, that is, the mass forces are balanced. The moment of these forces relative to the y–z plane (see also Fig. 2.11a) equals

$$M_{\text{rot},\,i} = F_{\text{rot},\,i}a \tag{2.29}$$

The direction of these moments in the y–z–plane is shown in Fig. 2.12b. The vectorial sum of the moments is

$$M_{\text{rot, tot}} = F_{\text{rot}}a\sqrt{3} \tag{2.30}$$

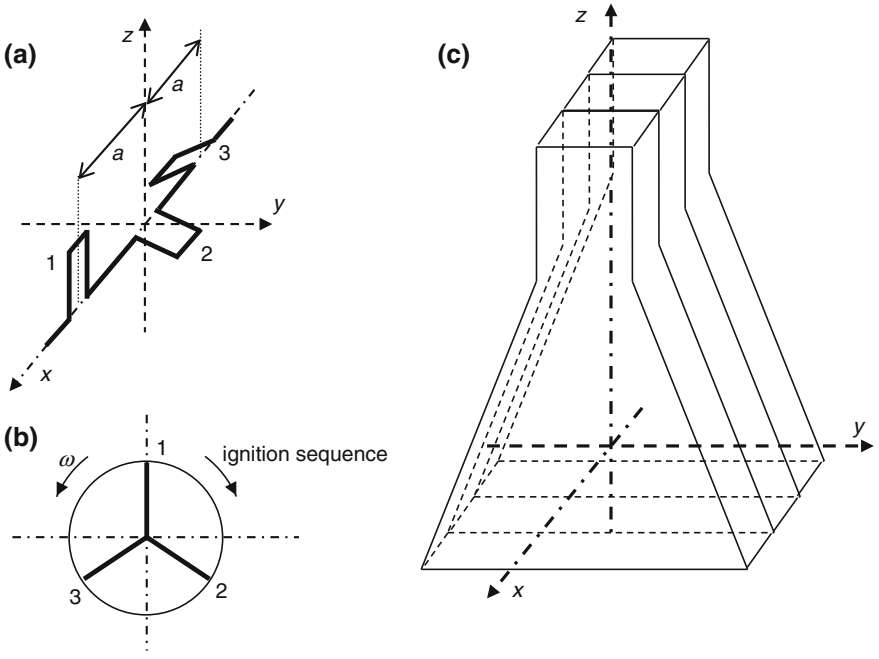

Fig. 2.11 Three-cylinder engine. **a** Crankshaft. **b** Diagram showing the ignition times of the separate cylinders. **c** The frame

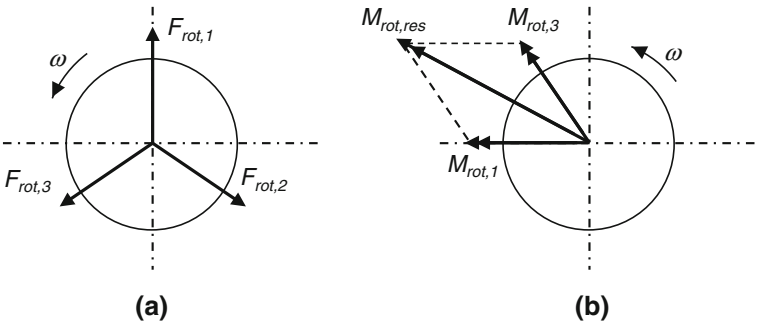

Fig. 2.12 Rotating mass forces **a** and torques **b** on the crankshaft, projected to the y–z plane

where $F_{\mathrm{rot}} = F_{\mathrm{rot},\,i}$. This rotating moment is transferred through the frame to the mounts and causes vibrations in the supporting structures. To prevent this undesirable dynamic load, generally counter-masses are placed on the cranks of all cylinders. The mass forces generated by these masses eliminate the respective rotating forces and thus reduce the loading of the crank shaft. An alternative approach is the application of counter weights at the ends of the crankshaft that eliminate the rotating moment.

Example 2.6 (Ball Bearing) Figure 2.13 schematically shows a ball bearing, for which the shaft and inner race are loaded by a force W. This load is transferred to the outer race by the rolling elements.

The distribution of the load W over the different rolling elements will be derived using Fig. 2.14. Due to the elasticity of the bearing, the centre of the axis will undergo a small displacement δ, see Fig. 2.14a.

The displacement of the axis results in a distribution of the force W over the elements in the lower half of the bearing (Fig. 2.14b). For a displacement δ of the axis, the deformation of the lower rolling element is $\delta_0 = \delta$. For the other

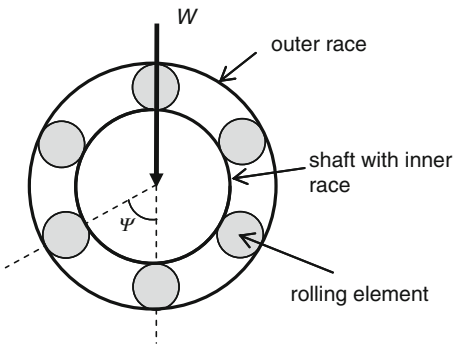

Fig. 2.13 Schematic representation of a ball bearing

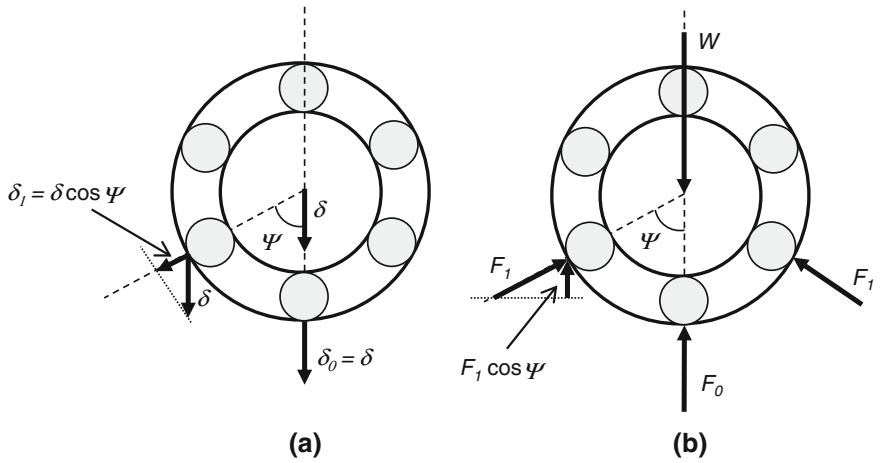

Fig. 2.14 Elastic deformation of a ball bearing; elastic displacements (*left*) and distribution of loads (*right*)

elements, the deformation is $\delta_1 = \delta_0 \cos \psi$. Assume that the rolling elements can be considered as a nonlinear springs with the following relation between force and displacement

$$F = c\delta^{1.1} \tag{2.31}$$

Then, the force on the lower rolling element is

$$F_0 = c(\delta_0)^{1.1} = c\delta^{1.1} \tag{2.32}$$

and on the other elements

$$F_1 = c(\delta_1)^{1.1} = c(\delta \cos \psi)^{1.1} \tag{2.33}$$

From the vertical force equilibrium, it follows that

$$W = F_0 + 2F_1 \cos \psi = F_0\left(1 + 2(\cos \psi)^{2.1}\right) \tag{2.34}$$

In this example, the angle ψ between the elements is 60°, which means that $F_0 = 0.682\ W$ and $F_1 = 0.318\ W$.

2.2.5 Thermal Expansion

In many cases, the thermal expansion of structures or parts yields mechanical loads in the structure. So although the origin of the load is thermal, that is, a heating or cooling process that causes a rise or drop in temperature, the nature of resulting forces is mechanical.

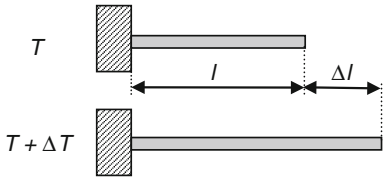

Fig. 2.15 Thermal expansion of an unrestrained bar

An increase in temperature of a certain material generally yields a thermal expansion of the body. The magnitude of the thermal expansion for an unrestrained body with original length l subjected to a temperature increase ΔT equals (Fig. 2.15)

$$\Delta l = \alpha \, \Delta T \, l \tag{2.35}$$

where α is the coefficient of thermal expansion, which is a material property. As the unrestrained bar in Fig. 2.15 can expand freely, no mechanical loads will be generated. That will be the case when the expansion is prohibited, for example, when the bar would be fixed in between two other parts. The generation and calculation of thermal stresses will be discussed in the next chapter.

2.2.6 Internal Stress

Another common type of mechanical load is due to internal stresses that are already present in a structure. Examples of this type of load are deliberately applied pre-stresses and undesirable residual stresses due to plastic deformation. The origin of the latter type of internal stresses will be discussed in Sect. 3.2.6. The purpose of deliberately applied pre-stresses is to prevent that certain parts are overloaded or not loaded at all. The pre-stresses are applied during assembly of the system by the use of an external force. A well-known example is a bolt that is pre-stressed to avoid that it releases during use at high temperature due to thermal expansion. Another example concerns the tie-rods in a diesel engine, as will be shown in Example 2.7.

Example 2.7 (Tie-rods in a Diesel Engine) In diesel engines, as well as in other piston engines, very high pressures occur in the cylinders during combustion. In the 1970s, peak pressures were in the order of 100 bar, but in modern engines the values have increased to around 180 bar. These high pressures tend to lift the cylinder head. To avoid cylinder leakage, the head is attached to the frame of the engine with tie-rods (see Fig. 2.16), in which a considerable pre-stress is introduced. The loading of these tie-rods and the required pre-stress are discussed in this example. As the pressure in the cylinder increases during the compression phase of the cycle and reaches a maximum value during combustion, the loading of the tie-rods is a dynamic load (that varies in time).

To avoid cylinder leakage, a pre-stressing force F_v is applied to the tie-rods. The associated reaction force is a compressive force F_v on the cylinder liner. When

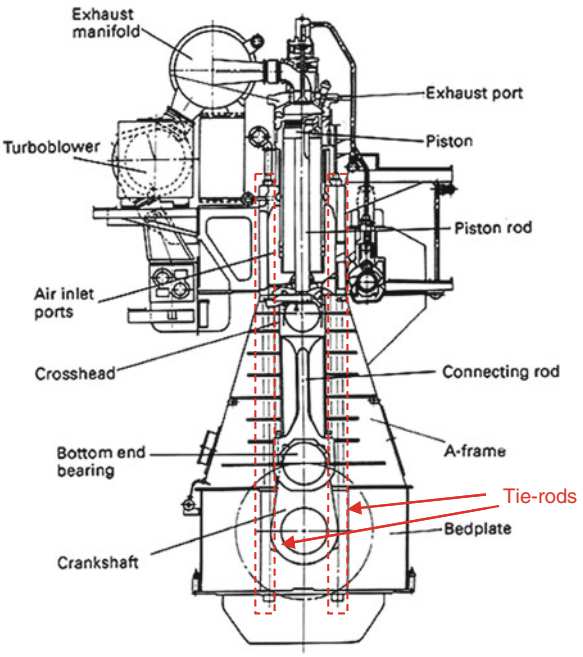

Fig. 2.16 Tie-rods attach the cylinder head to the frame of a diesel engine

the stiffness of the tie-rod and cylinder liner is k_{tr} and k_{cl}, respectively, then the force in the tie-rod for a certain elongation u_{tr} equals

$$F_{tr} = k_{tr}u_{tr} = \frac{EA}{l}u_{tr} \tag{2.36}$$

where the stiffness is expressed in terms of the elastic modulus E, cross-sectional area A and length l of the tie-rod. In the liner, the compressive force equals

$$F_{cl} = k_{cl}u_{cl} \tag{2.37}$$

where u_{cl} is the elongation of the liner.

As the forces in the tie-rod and the liner both equal the pre-stressing force F_v, the elongations can be expressed as

$$u_{tr} = \frac{F_v}{k_{tr}} \quad \text{and} \quad u_{cl} = -\frac{F_v}{k_{cl}} \tag{2.38}$$

In Fig. 2.17, the application of the pre-stressing force is visualized. The force–displacement curve for the liner has been copied and translated to the right-hand side of the figure, as will be explained later.

Due to the loading by a gas force F_{gas}, the cylinder head is lifted over a small distance Δu (Fig. 2.18). As a result, both the tie-rod and the cylinder liner are subject to an elongation Δu that causes additional forces in both parts equal to

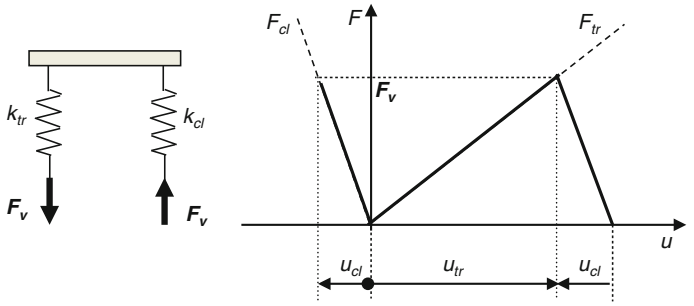

Fig. 2.17 Pre-stress in a cylinder head and the associated stiffness diagram

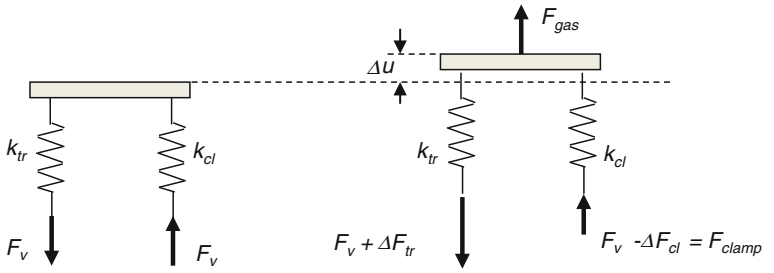

Fig. 2.18 Pre-stressed cylinder head and the loading due to the gas force

$$\Delta F_{\mathrm{tr}} = k_{\mathrm{tr}} \Delta u \tag{2.39}$$

$$\Delta F_{\mathrm{cl}} = k_{\mathrm{cl}} \Delta u \tag{2.40}$$

These forces must be in equilibrium with the gas force F_{gas}

$$F_{\mathrm{gas}} = \Delta F_{\mathrm{tr}} + \Delta F_{\mathrm{cl}} \tag{2.41}$$

From these equations, it follows that

$$\Delta F_{\mathrm{tr}} = \frac{k_{\mathrm{tr}}}{k_{\mathrm{cl}}} \Delta F_{\mathrm{cl}} \quad ; \quad \Delta F_{\mathrm{cl}} = \frac{k_{\mathrm{cl}}}{k_{\mathrm{tr}}} \Delta F_{\mathrm{tr}} \tag{2.42}$$

$$\Delta F_{\mathrm{cl}} = \frac{k_{\mathrm{cl}}}{k_{\mathrm{tr}} + k_{\mathrm{cl}}} F_{\mathrm{gas}} \quad ; \quad \Delta F_{\mathrm{tr}} = \frac{k_{\mathrm{tr}}}{k_{\mathrm{tr}} + k_{\mathrm{cl}}} F_{\mathrm{gas}} \tag{2.43}$$

which means that the clamping force F_{clamp} is given by

$$F_{\mathrm{clamp}} = F_v - \Delta F_{\mathrm{cl}} = F_v - \frac{k_{\mathrm{cl}}}{k_{\mathrm{tr}} + k_{\mathrm{cl}}} F_{\mathrm{gas}} \tag{2.44}$$

and the total loading of the tie-rod is

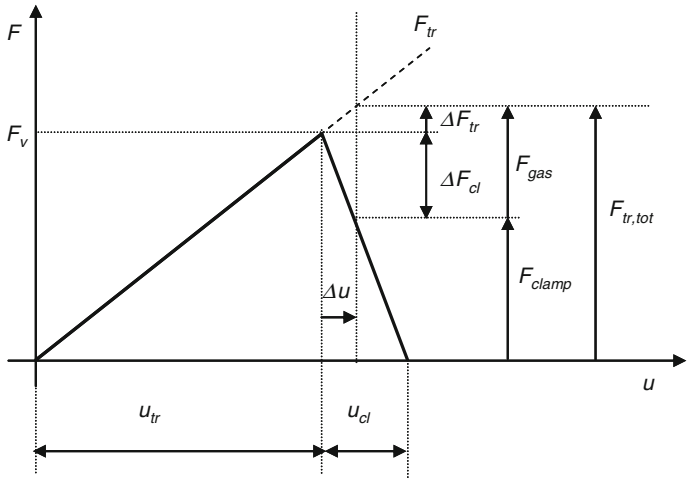

Fig. 2.19 Stiffness diagram of a pre-stressed cylinder head

$$F_{\mathrm{tr,\,tot}} = F_v + \Delta F_{\mathrm{tr}} = F_v + \frac{k_{\mathrm{tr}}}{k_{\mathrm{tr}} + k_{\mathrm{cl}}} F_{\mathrm{gas}} = F_{\mathrm{clamp}} + F_{\mathrm{gas}} \tag{2.45}$$

The Eqs. (2.42) to (2.45) are visualized in the stiffness diagram in Fig. 2.19. This diagram shows that the load on the tie-rods can be reduced by choosing a low stiffness (i.e. using rods with a large length); especially the dynamic part of the load, ΔF_{tr}, is reduced considerably in a flexible tie-rod, which is particularly important to avoid fatigue failures. To prevent the lifting of the cylinder head, the clamping force F_{clamp} should always have a positive value ($F_{\mathrm{clamp}} > 0$), which requires a large pre-stressing force F_v. The magnitude of this force largely determines the cross section of the tie-rods in a specific engine.

2.2.7 Internal Forces

Generally a complete structure is loaded by a number of concentrated and distributed forces acting on different locations. The combined action of all these forces eventually determines the total mechanical load on the system or structure. As a result, a certain location in the structure can be loaded in tension or compression, in bending or in torsion. The translation of the set of individual forces into their combined action is governed by the laws of *statics*. In simple one- or two-dimensional structures (e.g. beams, trusses), the resulting normal force, transverse force and bending moment on certain cross sections of the structure can be calculated rather easily. For more complex, three-dimensional structures generally numerical analyses like finite element analysis (FEA) are applied.

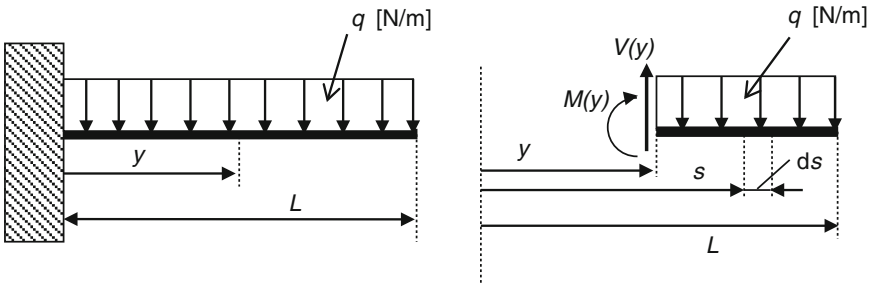

Fig. 2.20 Calculation of the internal forces in a one side fixed beam

Transverse Force and Bending Moment

When a load is applied in a direction perpendicular to the structure, a certain distribution of the transverse force V and bending moment M will exist in the body. This will be illustrated using the simple example of a beam that is fixed on one side and loaded by a uniformly distributed load q normal to the beam, see Fig. 2.20.

To calculate the internal forces, a fictitious section is made at a distance y from the fixation. The effect of the removed part of the beam is represented by a transverse force $V(y)$ and a bending moment $M(y)$ at the section plane. Since for the remaining part of the beam the forces and moments must be in equilibrium, the magnitudes of $V(y)$ and $M(y)$ can be calculated. The transverse force balances the resultant of the distributed load q

$$V(y) = \int_y^L q \, \mathrm{d}s \tag{2.46}$$

and the bending moment should be in equilibrium with the total moment generated by the distributed load

$$M(y) = -\int_y^L (s - y) q \, \mathrm{d}s \tag{2.47}$$

The calculation of internal forces and moments in more realistic components and structures is illustrated further in two examples: an aircraft wing and a gas turbine rotor blade.

Example 2.8 (Internal Forces in an Aircraft Wing) During stationary flight conditions, the distributed load $q_0(y)$ on an aircraft wing is a combination of (upward) lift forces and (downward) gravity forces

$$q_0(y) = -A\, l(y) + m(y) g \tag{2.48}$$

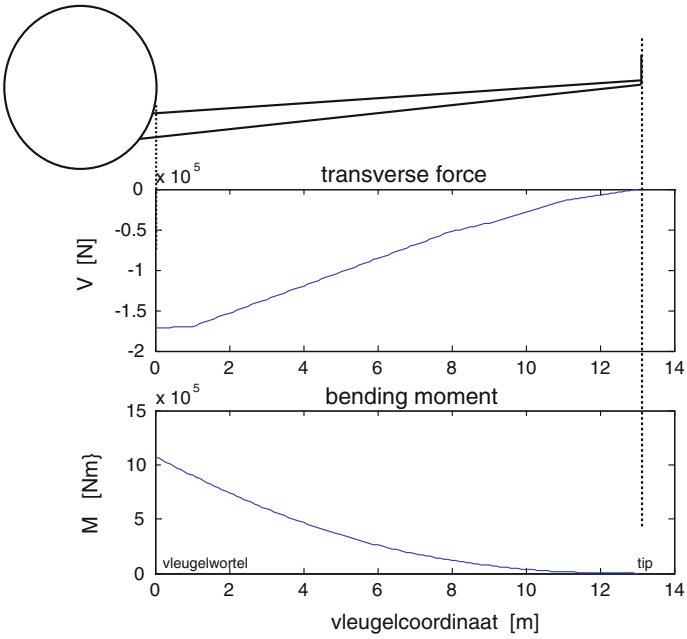

Fig. 2.21 Variations of transverse force and bending moment across the wing

Assume that the lift force and mass distribution are given by the block wise distributions in Figs. 2.4 and 2.5 and $A = 21{,}714$. Using Eqs. (2.46) and (2.47), the variations of the transverse force and bending moment across the wing can be calculated. The result is shown in Fig. 2.21.

Example 2.9 (Normal Force in a Gas Turbine Rotor Blade) The centrifugal force acting on a rotating gas turbine blade (see Sect. 2.2.3) yields a tensile force in the blade. This load is the normal force N, which is directed normal to the blade cross section.

A centrifugal force equal to $\omega^2 r \mathrm{d}m$ acts on the element with height $\mathrm{d}r$ and mass $\mathrm{d}m$ (see Fig. 2.22). This force must be balanced by the increase of the normal force $\mathrm{d}N$ across the height $\mathrm{d}r$. When the cross-sectional area of the blade equals A along the whole length of the blade and the mass density is ρ, then

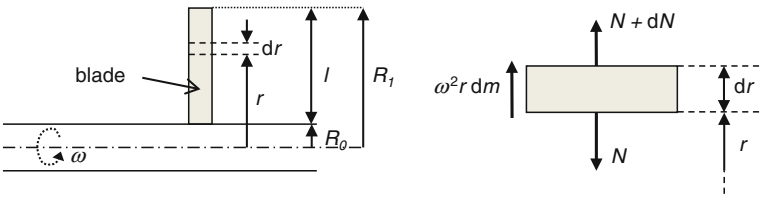

Fig. 2.22 Internal normal force in a rotating turbine blade

$$\mathrm{d}m = \rho A \, \mathrm{d}r \tag{2.49}$$

and

$$\mathrm{d}N = -\rho A \, \mathrm{d}r \, \omega^2 r \tag{2.50}$$

Integrating this expression to the radius r with the boundary condition $N(r = R_1) = 0$ (i.e. the normal force reduces to zero at the tip of the blade) yields

$$N(r) = \frac{1}{2}\rho A \omega^2 \left(R_1^2 - r^2\right) \tag{2.51}$$

The normal force $N(r)$ thus varies parabolically with r and attains its maximum value at the blade root ($r = R_0$).

2.3 Thermal Loads

Thermal loads on a system are caused by internal or external heat generation, resulting in an increase or decrease in the temperature. The externally generated heat can enter the body by convection or radiation. Combined with the internally generated heat, it is then distributed over the body by conduction.

The thermal load is expressed as the amount of heat q (W/m^2) that is entering or leaving a certain area of the body per unit time. Also for this type of load, several sources exist. In the next subsections, the mechanisms of conduction, convection and radiation are treated and the different sources of thermal loads are discussed.

2.3.1 Conduction

Heat conduction is the mechanism that enables the redistribution of heat within a body. Heat can also be transferred to another body by conduction, but only when the two bodies are in thermal contact. In all cases, the magnitude and direction of the heat flow will depend on the temperature difference between (different parts of) the bodies. Heat will always flow from the higher to the lower temperature, and the magnitude of the heat flow will be larger when the temperature difference per unit length, that is, the temperature gradient $\mathrm{d}T/\mathrm{d}x$, is larger:

$$q = k \, \frac{\mathrm{d}T}{\mathrm{d}x} \tag{2.52}$$

where k is the heat conduction coefficient (W/mK). This coefficient is a material property with high values ($\sim$100–400 W/mK) for well conducting materials like metals and much lower values for materials like glass ($\sim$1 W/mK). The higher the conduction coefficient, the larger the heat flow will be for a given temperature gradient.

2.3.2 Convection

A second way in which heat can be transferred between to media is by convection. This process takes place at the interface of two media, for example, at the interface between the combustion gas and the metal parts of an engine. Also in this case, the magnitude and the direction of the heat flow depend on the temperature difference between the two media:

$$q = h\,(T_2 - T_1) \tag{2.53}$$

where h is the heat transfer coefficient (W/m^{2}K), and T_1 and T_2 are the temperatures of both media. The value of h depends on the two materials, but also on the flow conditions at the interface where the heat transfer takes place. For example, the heat transfer from a gas to a metal surface is much higher for a turbulent flow ($\sim$1,500 W/m^{2}K) than for a laminar flow ($\sim$150 W/m^{2}K).

2.3.3 Radiation

Finally, heat can also be transferred by radiation. According to the Stefan–Boltzmann radiation law, the amount of heat radiated by a black body is proportional to the fourth power of the (absolute) temperature T (K):

$$q = \sigma T^4 \tag{2.54}$$

In this radiation law, the constant $\sigma = 5.67\ 10^{-8}$ W/m^{2}K^4 is the Stefan–Boltzmann constant.

A non-black body will only radiate a certain fraction of the heat that a black body provides. This fraction depends on the emissivity ε of the material, which attains a value between 0 and 1. Further, in real conditions, a body will not only radiate, but will also receive radiation from other bodies and its environment. The nett amount of heat that a body radiates in an environment with temperature T_{env} is therefore given by

$$q = \sigma\varepsilon\,(T^4 - T_{\text{env}}^4) \tag{2.55}$$

A typical value for the emissivity of a metal surface is $\varepsilon = 0.75$, while application of a ceramic coating to the metal reduces the emissivity to $\varepsilon = 0.35$.

Example 2.10 (Thermal Loading of a Gas Turbine Blade) Gas turbine blades face a severe thermal loading, since they operate in a very high temperature gas flow. A schematic representation of a turbine blade and its surrounding is shown in Fig. 2.23. To ensure that the metal temperature (T_{wall}) stays within acceptable limits, the blade is internally cooled by gas with a temperature $T_2 = 800$ °C.

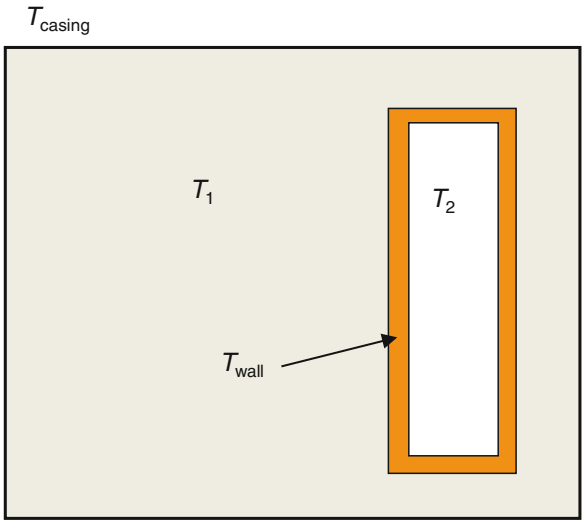

Fig. 2.23 Metal and gas temperatures in the surrounding of a turbine blade

The temperature of the hot gas in the gas path of the engine is $T_1 = 1{,}500$ °C, and the temperature of the metal casing is $T_{\text{casing}} = 500$ °C.

The thermal load on the turbine blade consists of several contributions:

1. Convection from the hot gas to the blade wall: $q_{c1} = h_1(T_1 - T_{\text{wall}})$.
2. Convection from the wall to the cooling gas: $q_{c2} = h_2(T_{\text{wall}} - T_2)$.
3. Radiation from the blade to the external surrounding: $q_{r1} = \sigma\varepsilon T_{\text{wall}}^4$.
4. Radiation from the casing to the blade: $q_{r2} = \sigma\varepsilon T_{\text{casing}}^4$.
5. Radiation inside the blade (from one side to other side): $q_{r3} = \sigma\varepsilon T_{\text{wall}}^4$.

Assuming that the momentary wall temperature is $T_{\text{wall}} = 950$ °C, the heat transfer coefficients are $h_1 = 1{,}500$ W/m^{2}K and $h_2 = 150$ W/m^{2}K and the emissivity of the blade material $\varepsilon = 0.75$, the total thermal load on the turbine blade can be calculated. Taking the heat flows into the blade in a positive sense and the outgoing heat flows negative, this yields

$$\begin{aligned} q_{tot} &= q_{c1} - q_{c2} - q_{r1} + q_{r2} + q_{r3} - q_{r3} \\ &= 1500(1500 - 950) - 150(950 - 800) \\ &\quad + 0.75 \cdot 5.67 \cdot 10^{-8}\left[(950 + 273)^4 - (500 + 273)^4\right] \\ &= 7.23 \cdot 10^5 \text{ W/m}^2 \end{aligned}$$

Note that the incoming radiative heat contribution q_{r3} on one wall at the inside of the blade is completely balanced by an equally large outgoing heat flow.

In the situation used in this example, the calculated heat flow is positive, which means that there is a nett heat flow *into* the turbine blade. This means that the wall temperature will increase and a new situation will arise: the incoming contribution

q_{c1} decreases, whereas the out flowing contributions q_{c2} and q_{r1} increase, resulting in a much lower thermal load. This process will continue until a steady-state situation is reached in which the nett heat flow $q_{tot} = 0$.

2.3.4 External Heat Generation

Different sources of heat generation exist. The most relevant sources are discussed in this subsection. External heat generation generally leads to an increase in temperature of the gas or fluid that surrounds a system. By convection or radiation, the heat is transferred from the environment to the system, resulting in a thermal loading of the system.

Compression

When a gas is compressed, the temperature will increase. The relation between gas pressure (p), volume (V) and temperature (T) in an ideal gas is given by the Boyle–Gay Lussac's gas law:

$$\frac{pV}{T} = \text{constant} \tag{2.56}$$

This shows that increasing the pressure of a constant volume of gas will yield a proportional increase in the (absolute) temperature. In practical cases, the increased gas temperature generally yields a large heat input to the surrounding structure (e.g. cylinder wall in a piston engine) by convection or radiation.

Combustion

The combustion process of fuels is a complex process that generates a considerable amount of heat. To quantify the amount of heat, measurements are required or a combustion model must be developed and applied. This process also elevates the temperature of the gas, providing a large thermal loading of the surrounding structures.

2.3.5 Internal Heat Generation

In addition to the external sources of heat generation, also a number of sources for internal heat generation can be identified. Opposed to the external sources, for which the heat must be transferred to the body under consideration by convection, conduction or radiation, the internal sources deliver a direct heat flow q in the body. The most common sources, mechanical friction and electrical losses, are discussed next.

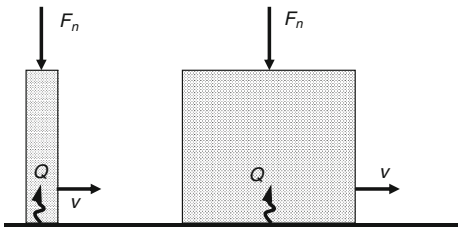

Fig. 2.24 Heat generation in two similar bodies with different contact areas

Friction

When two parts are moving along each other, friction will occur at the interface. Friction is a force (F_f) that opposes the motion and its magnitude is determined by the friction coefficient μ and the (normal) force F_n applied to both parts

$$F_f = \mu F_n \tag{2.57}$$

The magnitude of the friction coefficient depends on the materials of both parts and the surface roughness at the interface, as will be discussed in more detail in Sect. 4.6.1. By opposing the motion, friction also causes dissipation of energy, resulting in heat generation in one or both parts. The heat generation Q (W) is proportional to the relative speed v of the two bodies

$$Q = F_f v \tag{2.58}$$

Note that this equation yields the amount of generated heat per unit of time, but the amount of heat is independent of the contact area. This means that two similar bodies sliding along a flat surface with the same speed and normal loads, but with contact areas A and $5A$, respectively (see Fig. 2.24), generate the same amount of heat Q. However, the heat flow q (W/m^2) is a factor 5 higher in the body with the smaller contact area.

Electric Losses

When an electric current flows in a conductor, losses will occur due to the resistance. These electrical losses are converted into heat. The amount of heat P (W) generated per unit of time in a conductor (that satisfies the Ohm's law) is determined by the resistance R of the conductor and either the potential difference (or voltage) V or the current I:

$$P = VI = \frac{V^2}{R} = I^2 R \tag{2.59}$$

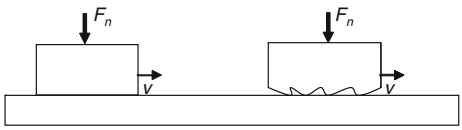

Fig. 2.25 Heat generation in two similar bodies with different surface roughness

Example 2.11 (Flash Temperature) As was mentioned above, the local heat flow due to friction largely depends on the contact area. An extreme case of high heat flows (and associated high temperatures) occurs when the contact between a part and the underlying surface is only provided by a few small roughness peaks. This is schematically shown in Fig. 2.25, where two parts with the same nominal contact area are presented.

However, whereas the left-hand part has a perfectly smooth surface, the right-hand side part has quite a rough surface profile. The real contact area of the smooth part is equal to the nominal contact area, while for the rough part it is only a fraction of the nominal area. Both parts move with the same speed and are loaded with the same normal force, so the total amount of heat generated in the contact is equal. However, the local thermal load q per unit contact area is much higher in the rough part. This means that temperature increase at the roughness peaks is much higher than the average temperature increase in the smooth part. This high local temperature is called the flash temperature and in some cases even leads to melting of the material at the roughness peaks.

2.4 Electric Loads

The external electric loads on a system are generally quantified in terms of the voltage or current that are applied to the system. The magnitude of voltage (ΔV) is expressed in Volt (V), while the unit of current (I) is Ampere (A). On an engineering level, electrical systems are generally analysed in terms of these quantities, but to understand the internal loading and failure of electric systems, the origin of these quantities is treated first.

Electric loading of a system can only occur when at some place in or near a system electric charge is present. The symbol generally used for electric charge is Q, while its magnitude is expressed in terms of Coulomb (C). It then depends on the property or the function of the system (i.e. conductor or insulator) whether the absolute amount of charge or its distribution in space and/or time is governing the electric load on the system.

The collection of all point charges in a body generates an electric field. As the electric field is characterized by both its magnitude and direction, it is mathematically described by the vector field $\vec{E}$. Further, for an electrostatic situation, where no time variation of the electric field occurs, at any location in the field the electric potential V can be calculated. This potential defines the amount of potential energy that a test charge would have when it is present at that specific location in the electric field. The potential V at some point can therefore be calculated as the energy that is required to move a small electric unit charge along an arbitrary path from a (remote) location with zero potential to the specified point. In mathematical terms, this is expressed as the line integral along a path C with length l

$$V = -\int_C \vec{E} \cdot d\vec{l} \tag{2.60}$$

The unit of potential is Volt (V), which is equivalent to Joules per Coulomb (J/C). The potential energy of a charge Q in a point with potential V is therefore

$$U_p = VQ \tag{2.61}$$

The other way around, each individual point charge contributes to the electric field and thus to the potential in each location. The contribution of the point charge Q to the potential at a distance r from the charge is

$$V = \frac{1}{4\pi\varepsilon_0}\frac{Q}{r} \tag{2.62}$$

where ε_0 is the dielectric constant or permittivity (equal to 8.854×10^{-12} Fm^{-1}). Finally, the magnitude and direction of the electric field can then be obtained by calculating the gradient of the potential

$$\vec{E} = -\vec{\nabla}V = -\begin{pmatrix} \frac{\mathrm{d}V}{\mathrm{d}x} \\ \frac{\mathrm{d}V}{\mathrm{d}y} \\ \frac{\mathrm{d}V}{\mathrm{d}z} \end{pmatrix} \tag{2.63}$$

This implies that the unit of electric field strength must be V/m. The foregoing discussion demonstrates that any distribution of electric charge in a body yields an electric field and an associated distribution of electric potential. The engineering quantities voltage and current are closely related to these definitions, as will be shown next.

The voltage between two locations is defined as the potential difference between these points. This means that the voltage represents the energy that is required to move a small electric unit charge along a certain path from one to the other location. Again, in a static electric field, voltage is independent of the path followed. In general, the voltage is thus the ratio of the amount of work (W) and the magnitude of the charge

$$\Delta V = V_1 - V_2 = \frac{W}{Q} \tag{2.64}$$

The unit of voltage is equal to the unit of potential, that is, Volt (V) or Joules per Coulomb (J/C).

Whereas the voltage is related to a static situation with a certain distribution of charge yielding potential differences, electric current is related to moving charges. The point charges in a body not only cause the electric field, but are also affected by the existing electric field. More specifically, the electric field exerts a force on each charged particle, resulting in a motion of that particle either in the direction (positive charge) or opposite (negative charge) to the direction of the field;

especially in solid materials, the charge carriers (mostly electrons) cannot travel in a straight line, but are bouncing from atom to atom. However, the net motion will be opposite to the direction of the electric field. The net velocity associated with this motion is called the drift velocity of the charge carriers and is proportional to the magnitude of the electric field (E)

$$v_{\text{drift}} = \mu E \tag{2.65}$$

The proportionality constant μ represents the mobility of the charge carriers (m^2/Vs), which largely depends on the type of material and especially the aggregation state of the material (gas, fluid or solid).

When the drift velocity of the charge carriers is known, the amount of charge moving in a material and thus the current I can be assessed as

$$I = nAv_{\text{drift}}Q \tag{2.66}$$

In addition to the drift velocity, the current is proportional to the cross-sectional area A and the number n (per unit volume) and charge Q of the moving particles.

Example 2.12 (Electron Drift Velocity) A typical material used for conductors is copper, with a mass density $\rho = 8.94$ g/cm^3 and an atomic weight of 63.546 g/mol. If a copper wire with a 2 mm diameter is carrying a current $I = 2$ A, the drift velocity of the electrons can be calculated. Firstly, the number and magnitude of charge carriers must be determined. In a metal, electrons will be the charge carriers, and in copper each atom has one free electron. One cubic centimetre of copper has a mass of 8.94 g and thus contains $8.94/63.546 = 0.1407$ mol Cu atoms. By multiplying with Avogadro's number (6.02×10^{23}), the number of atoms per cm^3 is obtained. The charge of an electron equals 1.6×10^{-19} C. Finally, the cross-sectional area of the wire is given by $A = \pi r^2 = \pi(0.1)^2 = 0.0314$ cm^2. Using Eq. (2.66), the drift velocity can be calculated

$$v_{\text{drift}} = \frac{I}{nAQ} = \frac{2}{(0.1407 \cdot 6.02 \cdot 10^{23})(0.0314)(1.6 \cdot 10^{-19})} = 4.7 \times 10^{-3} \text{ cm/s}$$

This example shows that even a considerable electric current only yields a very low drift velocity of the electrons in the metal conductor.

Finally, the relation between voltage and current is defined by Ohm's law

$$I = \frac{\Delta V}{R} \tag{2.67}$$

showing that the component's resistance R determines the magnitude of the current at a certain applied voltage. Combining Ohm's law with Eqs. (2.65)–(2.66) and assuming that the voltage ΔV applied to a resistor with length l yields an uniform electric field $E = \Delta V/l$, the resistance R can be obtained from

$$R = \frac{1}{nQ\mu}\frac{l}{A} = \rho\frac{l}{A} \tag{2.68}$$

This demonstrates that the resistance in a material depends on the dimensions (A and l) and the resistivity ρ of the material. The conductivity, being the reciprocal of the resistivity, is proportional to the amount of free charge carriers per unit volume (nQ) and their mobility μ.

2.4.1 Sources

Similar to the mechanical loads, also electric loads can arise from different sources. The generic load in this case is the specific distribution of charge over a body, resulting in a potential difference or electric current. Specific sources for this type of load are the different methods to transfer or generate electric charge. The most widely applied principle is electromagnetic induction, where a changing magnetic field moves the charge carriers (i.e. electrons) and thus generates potential differences or electric currents. This principle is applied in, for example, generators. Alternative sources are the photovoltaic effect, as applied in solar cells, electrochemical reactions (e.g. fuel cells) and electrostatic sources. These different ways of generating electric loads are discussed next.

Electromagnetic Induction

Induction is caused by the interaction between electric and magnetic fields. Charge carriers that move in a magnetic field experience a force that is proportional to the magnitude of the magnetic field B and the velocity v. The direction of the force is perpendicular to the plane spanned by the vectors $\vec{B}$ and $\vec{v}$, as defined by the following expression

$$\vec{F} = q\vec{v} \times \vec{B} \tag{2.69}$$

Due to the experienced force, the charge carriers (e.g. electrons in a conducting winding) start to move, yielding a potential difference or electric current. Since the relative orientations of the magnetic field and direction of motion are important, the concept of magnetic flux is used. The magnetic flux through a loop (winding) is the surface integral of the magnetic field over the surface of that loop. This means that the flux is maximum when the loop is perpendicular to the magnetic field and zero when both are aligned. The resulting electromotive force or voltage in a loop with surface S (and contour C) moving in a magnetic field with magnitude B is obtained from the gradient of the magnetic flux Φ according to

$$\Delta V = -\frac{\mathrm{d}\Phi}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\int_S \vec{B}\cdot\mathrm{d}\vec{a}\right) = \int_C (\vec{v}\cdot\vec{B})\times\mathrm{d}\vec{s} \tag{2.70}$$

and thus also appears to be proportional to the speed of the loop.

This principle of induction is applied in a generator to move charge and create a potential difference or electric current by rotating a set of windings in a stationary magnetic field. All present generators are derived from the dynamo, which was the first electrical machine capable of converting mechanical rotation into electrical power. In a dynamo, a commutator or collector is applied, which reverses the direction of the electric field (and current) each half revolution. This is accomplished by sliding carbon brushes connected to the housing along different segments of the commutator. The result is a (slightly pulsed) direct current (DC).

Without a commutator, the dynamo becomes an alternator, which delivers an alternating current (AC). If such a generator is used to power an electric grid, the rotational speed must be constant and synchronized with the electrical frequency of the grid. In most generators, no permanent magnet is applied, but a rotating field winding, fed by direct current, moves relative to a stationary winding that produces the alternating current. The direct current for the field winding is obtained from the generated power, for example, through brushes contacting the rotor or by applying rectifiers.

Photovoltaic Effect

Another principle to move charge carriers and create a potential difference or electric current is the photovoltaic effect. This effect is observed in materials where energy from incident light (photons) is used to excite charge carriers into the conduction band (see Sect. 3.4). In a semiconducting material, a so-called *p*–*n* junction can be constructed, which makes that electrons are excited in the *n* side of the junction and holes are created on the *p* side [1]. Due to the built-in electric field, the negatively charged electrons are forced to the external electrical circuit, while the positively charged holes do the same in opposite direction, thus creating the electric current.

This principle is applied in a solar cell (Fig. 2.26), where incident sunlight is utilized to generate electricity. A solar cell is made of silicon, depending on the

Fig. 2.26 Solar cell (*Source* Wikimedia Commons)

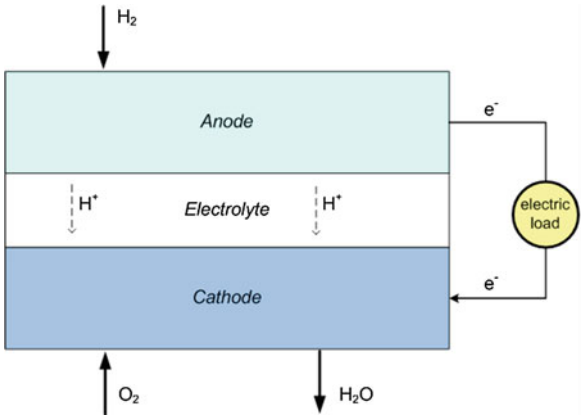

Fig. 2.27 Schematic representation of a fuel cell

type of cell in either polycrystalline, monocrystalline or amorphous form that has been fabricated to produce a *p–n* junction. Since a typical single solar cell produces a voltage of only around 0.5 V, several cells are usually combined into a solar panel to increase the delivered voltage and/or current.

Electrochemical Reactions

Free moving electric charge carriers can also be generated by chemical reactions. This principle is, for example, applied in fuel cells, which generate electricity through a chemical reaction. A cell consists of an anode, an electrolyte and a cathode (see Fig. 2.27), and two chemical reactions occur at the two interfaces between the three sections. The typical fuel used in these cells is hydrogen gas (H_2), which is oxidized at the anode using a catalyst (e.g. fine platinum powder). The oxidation of hydrogen yields positively charged ions and electrons:

$$2H_2 \rightarrow 4H^+ + 4e^-$$

The ions travel through the electrolyte to the cathode, but the electrons cannot and are thus forced to the electrical connection with the cathode, where they create the required current.

At the cathode, the hydrogen ions react with oxygen and the electrons to form water, which is drained as waste product. The cathode reaction is

$$4H^+ + O_2 + 4e^- \rightarrow 2H_2O$$

Also at the cathode, a catalyst is applied, which is often nickel. Depending on the design of the cell, the electrolyte material can be one of several options, for example, a H^+-conducting polymer membrane, the ceramic yttrium-stabilized zirconia or an aqueous alkaline solution.

A typical single fuel cell generates a voltage of 0.6–0.7 V, and the delivered current depends on the surface areas of anode and cathode. If higher voltages or currents are required, several cells can be combined in a fuel cell stack, where

parallel connections deliver higher current and series connections yield higher voltages.

Electrostatic Sources

The final source of electric loads discussed here is the electrostatic source. Contrary to the previous sources, which typically generate motion of charge carriers in conductors (thus resulting in electric current), electrostatic electricity is generally related to a static surface charge imbalance. This means that the amount of charge varies across the surface of a body, which therefore only occurs in insulating materials. In a conducting material, all differences in surface charge will immediately be balanced since the charge carriers can move freely.

However, in practice, the unbalance in surface charge will generally be quite small, unless processes occur that accumulate charge at some location. An example of such a process is the triboelectric effect. Certain materials become electrically charged when they are rubbed against each other. One of the materials then gets a negative charge, while the other material obtains an equally large positive charge. Examples of material combinations exhibiting this effect are amber rubbed with wool and glass rubbed with silk. Although very high charges and associated voltages can be attained in this way, the discharge will generally not generate a large current. But still electrostatic discharge is a major problem for many sensitive electronic devices failing due to this load (see Sect. 4.9.4).

Piezoelectric and Thermoelectric Sources

Finally, also piezoelectric and thermoelectric sources exist. In the former case, the mechanical deformation of specific solids produces a certain amount of electric charge. As an example, piezoelectric materials are commonly applied as the ignition source for cigarette lighters. A thermoelectric source produces a voltage when a temperature difference exists between two ends of a conductor. This effect is widely applied for temperature measurements, for example, in thermocouples.

2.5 Chemical Loads

Chemical loading of a system occurs when certain substances, for example, acids or certain gasses that cause degradation of the system, are in contact with the system. The degradation can be either due to a direct chemical reaction, as is the case for aggressive materials attacking metal or plastic parts in which they are contained, or by an electrochemical reaction where the chemical reaction only occurs when also an electric current can be established. Most common corrosion processes are based on electrochemical reactions.

The concentration of the concerning substance generally determines the magnitude of this type of load, although for some reactions a small amount of material is sufficient to start the reaction and a higher concentration does not affect the

degradation rate anymore. On the other hand, external parameters like temperature and the presence of other materials (catalysts) may considerably affect the reaction rate. For acids, the concentration of hydrogen ions is governing the load, which is generally expressed as a pH value.

For electrochemical corrosion reactions, the presence of an electrolyte and an electric connection between anode and cathode (see Sect. 4.10 for details on corrosion) are required in addition to the presence of a galvanic couple (two metals). Each of these three constituents may act as the rate-limiting factor in the corrosion reaction. Together they govern the resulting electric current that determines how many corrosion reactions can take place and thus determines the degradation rate.

In some cases, the chemical loads originate from a biological source. An example of this type of load is microbiologically induced corrosion (MIC), where micro-organisms create the environment that enhance certain corrosion processes.

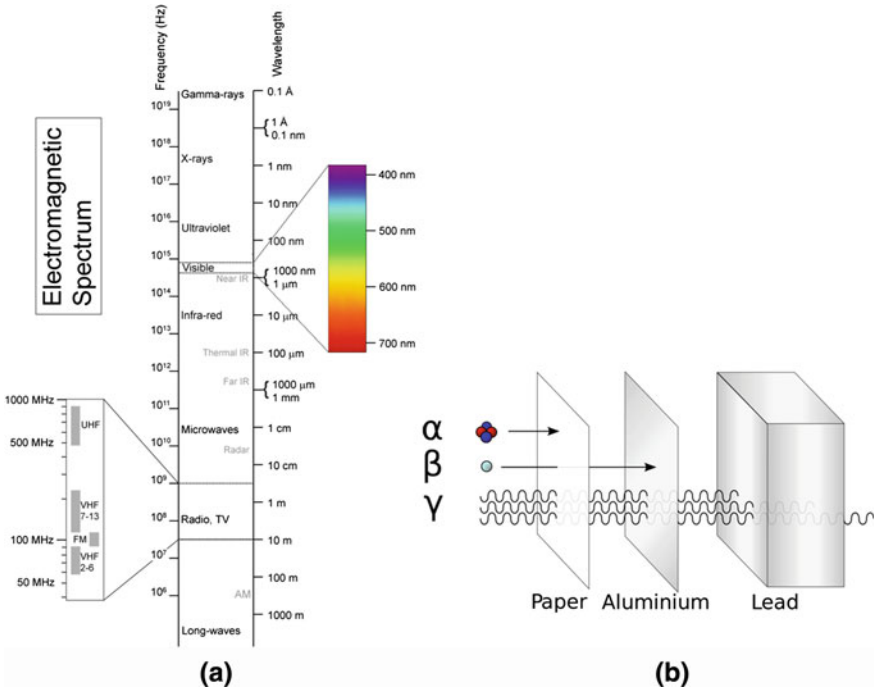

Fig. 2.28 **a** Electromagnetic spectrum. **b** Different types of radiation and their abilities to penetrate solid materials (Wikipedia)

2.6 Radiative Loads

Radiation can also cause degradation of a material, which ultimately could lead to failure of a system. In most cases, this is a secondary load: the material properties and thus the load-carrying capacity decrease over time. Two different types of radiation exist: radiation consisting of particles and electromagnetic radiation consisting of photons. The damage caused by the different types of radiation, both in structural materials and biological tissues (e.g. the human body) depends on the energy of the particles or photons. Particles like alpha particles, beta particles and neutrons generally have considerable energies and are therefore able to ionize atoms. Electromagnetic radiation is also able to ionize atoms, provided that the energy is sufficiently high. This is the case for radiation with very high frequencies (and thus small wavelengths) like X-rays and γ-rays, see Fig. 2.28a.

In addition to the energy, the damaging effect of the radiative load also depends on the penetration depth. The relatively large α-particles heavily interact with other materials, so they generally only travel a few centimetres in air or a few millimetres in low density materials. On the other hand, gamma rays consisting of photons with neither mass nor electric charge are very difficult to stop and therefore deeply penetrate into materials (Fig. 2.28b).

Finally, the magnitude of a radiative load is determined by the duration of the exposure and (for lower frequency electromagnetic radiation) the reflectivity of the surface, which determines the fraction of the incoming radiation to be absorbed.

2.7 Summary

In this chapter, an overview has been given of all relevant load types. For each type, the generic load was described and a number of specific sources for that generic load were discussed. In most cases, also quantitative relations were given that enable the calculation of the magnitude of the load.

References

1. Callister, W.D., Rethwisch, D.G.: Materials Science and Engineering, 8th edn. Wiley, Hoboken (2011)

Further Reading

1. Callister, W.D., Rethwisch, D.G.: Materials Science and Engineering, 8th edn. Wiley, Hoboken (2011)
2. Moran, M.J., Shapiro, H.N.: Fundamentals of Engineering Thermodynamics, 5th edn. Wiley, Hoboken (2006)
3. Purcell, E.M.: Electricity and Magnetism, 2nd edn. McGraw-Hill, New York (1985)

Chapter 3
Internal Loads

3.1 Introduction

In the previous chapter, a considerable number of external loads and their specific sources have been discussed. In the present chapter, it will be demonstrated that these external loads yield internal loads on the material level. These internal loads can be considered as failure parameters, since they are governing the failure mechanisms that will be discussed in Chap. 4. The translation of external to internal loads requires knowledge on the shape and dimensions of a body or part, but also on the material properties.

Figure 3.1 provides an overview of the internal loads (final column) for the different load types that have been discussed in the previous chapter. Also, the associated material properties are shown. In the next subsections, the internal loads will be discussed in detail and appropriate examples will be used to demonstrate how the loads can be calculated.

3.2 Mechanical Loads

Externally applied mechanical loads cause stress and deformation (strain) in materials. In Sect. 2.2.7, it was explained that for relatively simple one- or two-dimensional structures, the total set of externally applied forces and moments yields the resulting internal forces and moments on the cross sections of the structure constituents (e.g. beams). In the present section, it will be discussed how these internal forces can be translated into stresses and strains on the material level. It will be shown that the latter depend on the actual shape and dimensions of the cross sections and the properties of the materials used.

To start with, relatively simple one- or two-dimensional structures will be analysed after which the more general three-dimensional stress states will be treated, as well as the principle of equivalent stress. Then, elastic and plastic

T. Tinga, *Principles of Loads and Failure Mechanisms*,
Springer Series in Reliability Engineering,
DOI: 10.1007/978-1-4471-4917-0_3, © Springer-Verlag London 2013

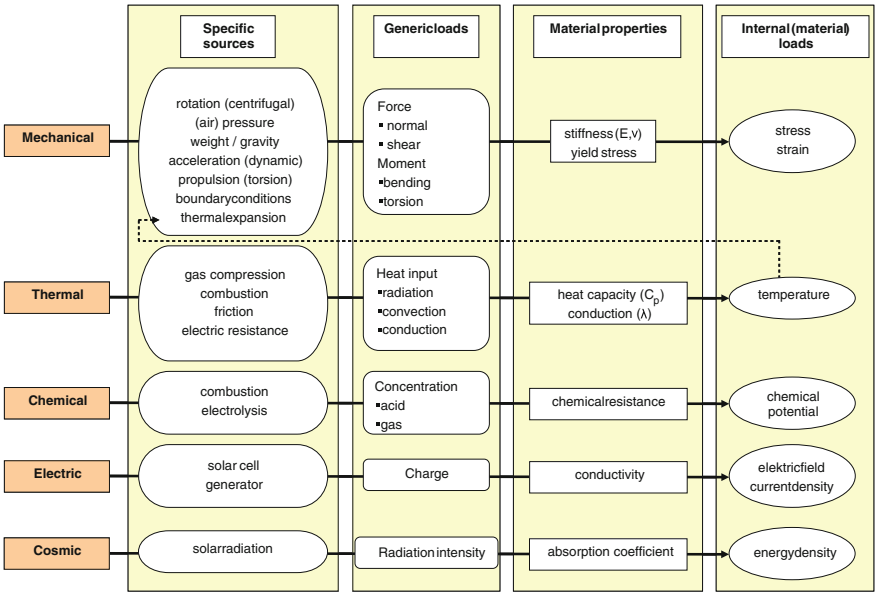

Fig. 3.1 Overview of external and internal loads for different load types

deformation will be discussed, and the section ends with some more advanced topics like thermal stress, stress concentration, contact stress and stresses at crack tips.

3.2.1 One- and Two-Dimensional Stress States

3.2.1.1 Normal and transverse forces

In a prismatic beam with homogeneous material properties subjected to a tensile force F aligned to the length of the bar (Fig. 3.2), the internal load is distributed homogeneously over the cross-sectional area A. This yields a normal stress σ equal to

$$\sigma = \frac{F}{A} \tag{3.1}$$

If a transverse force V is applied to the same beam, now in a direction perpendicular to the length of the bar, a uniformly distributed shear stress develops. The magnitude of this shear stress τ again depends on the cross-sectional area and is given by

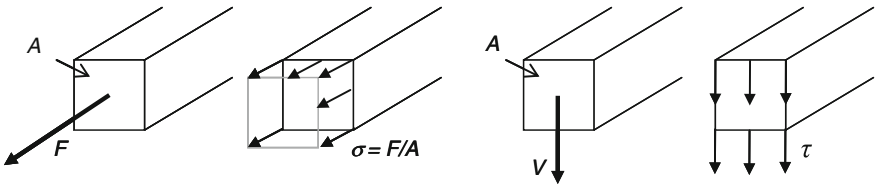

Fig. 3.2 Normal tensile force and transverse force applied to a beam and the resulting uniform stress distributions

$$\tau = \frac{V}{A} \tag{3.2}$$

It is clear that an equal *external* load, for example, a certain tensile force F, applied to two different beams will yield different *internal* loads. In a slender beam with a small cross-sectional area, the stress will be much larger than in a large beam, although the external loads are equal.

Example 3.1 (Stress in a Turbine Blade) In example 2.4, the normal force on an element in a gas turbine blade was calculated as

$$\mathrm{d}F_{cf} = \omega^2 r \, \mathrm{d}m$$

and the total centrifugal force at a distance r from the central line is

$$F_{cf} = \frac{1}{2}\rho A \omega^2 \left(R_1^2 - r^2\right)$$

To translate this centrifugal force into an internal load, that is, the centrifugal stress, the cross-sectional area A of the blade must be considered. The stress is thus equal to

$$\sigma_{cf} = \frac{F_{cf}}{A} = \frac{1}{2}\rho \omega^2 \left(R_1^2 - r^2\right) \tag{3.3}$$

3.2.1.2 Torsion

In parts that are designed to transfer a rotational motion, like shafts and axes, generally a torsional load is applied. Well-known examples are the cranck-shaft in a piston engine, transferring the torsional moments generated by the cylinders to the driven rotating machinery, and the shafts connecting an engine with, for example, (ship/aircraft) propellers or a generator. The loading condition in this kind of parts is schematically shown in Fig. 3.3.

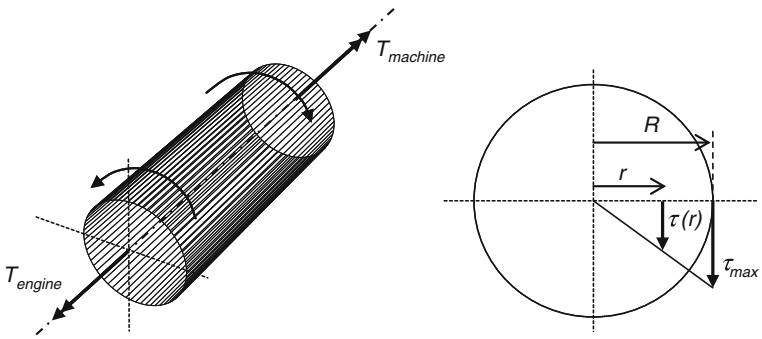

Fig. 3.3 Torsional loading of a shaft (*left*) and the resulting stress distribution (*right*)

Due to the loading, stresses develop in the material in the plane of the shaft cross section. The torsional shear stress τ at a specific radius r can be calculated by relating the applied torque T to the resistance against torsion I_p

$$\tau = \frac{Tr}{I_p} \tag{3.4}$$

It appears that the shear stress increases linearly with the radius r, see also Fig. 3.3, which means that the shear stress attains its maximum value at the outer surface of the shaft. Moreover, the stress distribution is rotation symmetric. The resistance against torsion is called the polar moment of inertia I_p. For a hollow shaft with inner diameter d and outer diameter D, the value of I_p is given by

$$I_p = \frac{\pi}{32}\left(D^4 - d^4\right) \tag{3.5}$$

Example 3.2 (Shear Stress in a Steam Turbine Shaft) In nuclear power plants, electricity is generated using a steam turbine that drives a generator. For a typical plant, the power transferred by the shaft of the steam turbine is in the order of 1,000 MW. The torque T in the shaft is directly related to the power P through the angular velocity ω

$$P = T\omega \tag{3.6}$$

The angular velocity of the steam turbine is equal to that of the generator. Since the frequency of the electricity in most countries is 50 Hz, both the generator and steam turbine operate at this frequency. Therefore, the torque in the shaft can be calculated:

$$T = \frac{P}{\omega} = \frac{1,000 \times 10^6}{2\pi \times 50} = 3.18 \times 10^6\,\text{Nm}$$

The shear stress in the shaft depends on its radius. If the maximum allowable stress in the shaft is 50 MPa (i.e. 20 % of the material yield stress), then the minimal radius of the shaft can be calculated:

$$\tau = \frac{Tr}{I_p} = \frac{Tr}{\frac{1}{2}\pi r^4} = \frac{2T}{\pi r^3}$$

$$r^3 = \frac{2T}{\pi \tau_{\text{all}}} = \frac{2 \times 3.18 \times 10^9}{\pi \times 50 \times 10^6} = 0.0405 \text{ m}^3 \quad \Rightarrow \; r = 0.343 \text{ m}$$

3.2.1.3 Bending

In parts that are subject to bending loads, also a non-uniform stress distribution exists. Contrary to the previously discussed normal, transverse and torsional loads, a bending load always causes both tensile and compressive stresses. On the one side of the structure's neutral line, the stresses are tensile, while on the other side, they are compressive. Exactly on the neutral line, the bending stress is zero. This is shown schematically for a beam with a rectangular cross section in Fig. 3.4. It appears that the stress increases linearly with the distance from the neutral line.

The bending stress at a certain distance y from the neutral line can be expressed as

$$\sigma_b = \frac{My}{I_x} \tag{3.7}$$

This means that the maximum bending stress for the beam occurs at the largest distance from the neutral line (i.e. the surface of the beam, in Fig. 3.4 at $h/2$), which equals

$$\sigma_{\max} = \frac{\frac{1}{2}hM}{I_x} \tag{3.8}$$

The resistance against bending I_x is called the (area) moment of inertia. For a rectangular cross section, I_x is given by

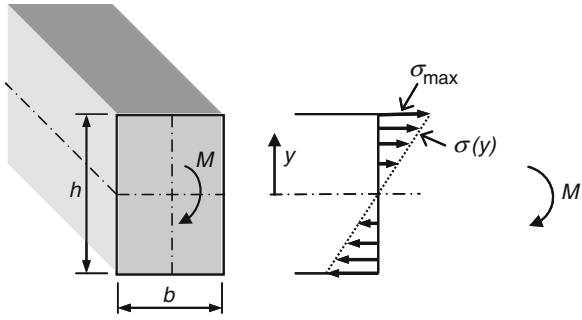

Fig. 3.4 Bending load on a beam with a rectangular cross section (*left*) and the resulting stress distribution (*right*)

$$I_x = \frac{1}{12} bh^3 \tag{3.9}$$

where b and h are the two dimensions of the cross section as indicated in Fig. 3.4.

From the descriptions in this subsection, it can be concluded that in all cases, the internal load (i.e. stress) is obtained from the external load (i.e. force, moment) by dividing it by the resistance against deformation, Eqs. (3.1), (3.2), (3.4) and (3.7).

3.2.2 Three-Dimensional Stress States

In the previous subsection, one- and two-dimensional structures have been analysed, in which the forces and moments cause relatively simple uniaxial (only normal stresses) or biaxial (combination of normal and shear stresses) stress states that are also uniform on a cross section. In more realistic structures, the stress state is more complex and is generally three-dimensional, while it also varies across different locations.

For an arbitrary loaded body, the material in any plain cross section is subjected to a stress vector that can be resolved into three components: one normal stress and two mutually perpendicular shear stresses. Therefore, 6×3 stress components are acting on the six faces of an infinitesimal cubic volume element, see Fig. 3.5a. In this figure, only the components on the visible faces of the cube are shown. The stresses acting on the remaining faces are identical but with opposite sign, which provides the required force equilibrium. Since the moments acting on the volume element must also be in equilibrium, it follows that $\sigma_{xy} = \sigma_{yx}$, $\sigma_{xz} = \sigma_{zx}$ and $\sigma_{yz} = \sigma_{zy}$.

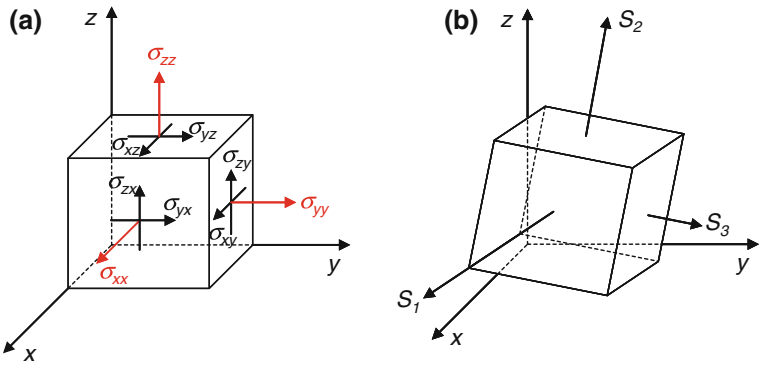

Fig. 3.5 **a** Volume element showing all stress components of the three-dimensional stress tensor. **b** The element has been rotated such that only the principal stresses remain

In mathematical terms, the stress state is described by the Cauchy stress tensor σ_{ij}, which is a second-order tensor. This means that two indices (or a two-dimensional matrix) are required to denote all stress components. In a three-dimensional Cartesian coordinate system, this yields 3×3 components. The first index (taking the value 1, 2, 3 or x, y, z) represents the direction in which the stress component is acting, and the second index indicates the (normal direction of) the plane on which the stress component acts. The stress component σ_{xx} is thus the component acting in the x-direction on the y–z plane.

The three components with two identical indices are called the normal stresses (σ_{xx}, σ_{yy} and σ_{zz}, also denoted as σ_x, σ_y and σ_z), whereas the other six components represent the shear stresses (σ_{xy}, σ_{yx} and σ_{xz},…, also denoted as τ_{xy}, τ_{yx} and τ_{xz}, …). As was mentioned above, the required equilibrium in moments makes that the six shear stress components are pair-wise identical ($\tau_{ij} = \tau_{ji}$). As a result, the stress state in any point is completely described by six independent stress components.

3.2.3 Principal Stress

For a given stress state, the magnitude of the individual stress components depends on the orientation of the coordinate system. The value of the normal stress, for example, changes from σ_y to σ_z when the coordinate system is rotated 90° around the x-axis. It appears to be possible to rotate the coordinate system in such a way that only normal stresses remain, and all shear stresses reduce to zero, see Fig. 3.5b. The stresses S_1, S_2 and S_3 associated with this state are called the principal stresses and are the maximum values the normal stresses can attain for a given stress state. Mathematically, the principal stresses are the eigenvalues of the Cauchy stress tensor.

For a two-dimensional stress state, the calculation of the principal stresses can be visualized using Mohr's circle of stress distribution. A common two-dimensional stress state is a plane stress situation, which exists in thin sheets or at the outer surface of bodies where the out-of-plane stress is zero (see Fig. 3.6a). In a plot of the Mohr's circle, see Fig. 3.6b, the vertical and horizontal axes represent the shear (τ) and normal stress (σ), respectively. The planes 1 and 2 of the cubic element in Fig. 3.6a can be mapped onto two individual points in the σ–τ plot: (σ_1, τ) and (σ_2, $-\tau$). It can be shown [1] that rotation of the cubic element over an angle ϕ around an axis normal to the stress-free plane yields a rotation over an angle 2ϕ of the associated points in the σ–τ plot. Apparently, the image points of the planes normal to the stress-free plane are on a circle with centre point m with coordinates $[(\sigma_1 + \sigma_2)/2, 0]$ and radius

$$r = \sqrt{\left(\frac{\sigma_1 - \sigma_2}{2}\right)^2 + \tau^2} \tag{3.10}$$

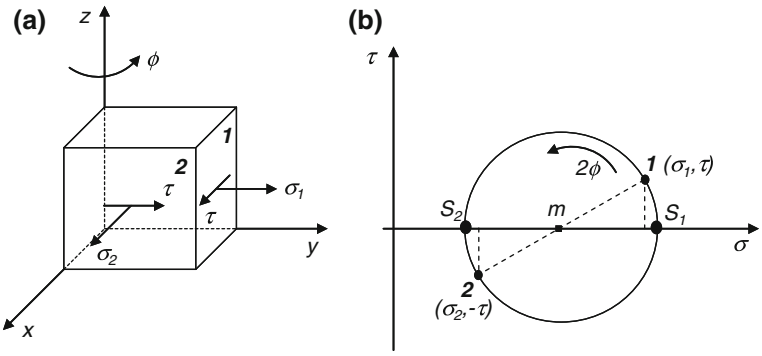

Fig. 3.6 **a** Stress components for a plane stress state. **b** Graphical representation of the stresses in planes 1 and 2 in a Mohr's circle

Moreover, it is also possible to rotate the cubic element such that the shear stresses on the two planes disappear, and only normal stresses remain. In this situation, the two points are on the horizontal axis in the Mohr's circle (see Fig. 3.6b).

By definition, this is a principal stress state, and thus, the principal stresses S_1 and S_2 can rather easily be determined from the plot:

$$S_1 = \frac{\sigma_1 + \sigma_2}{2} + \sqrt{\left(\frac{\sigma_1 - \sigma_2}{2}\right)^2 + \tau^2} \tag{3.11}$$

$$S_2 = \frac{\sigma_1 + \sigma_2}{2} - \sqrt{\left(\frac{\sigma_1 - \sigma_2}{2}\right)^2 + \tau^2} \tag{3.12}$$

The orientation of the principal stress plane of S_1 can be obtained from Fig. 3.6b

$$\tan(2\phi_1) = \frac{\tau}{\frac{1}{2}(\sigma_1 - \sigma_2)} \tag{3.13}$$

The maximum principal stress is an important quantity when considering the initiation and propagation of cracks. The magnitude of the maximum principal stress will govern the moment in time (i.e. the number of cycles) at which cracks occur, while its orientation will determine the direction of crack propagation. Cracks will generally propagate in a plane normal to the maximum principal stress. This will be discussed in more detail in Sect. 4.4.6.

3.2.4 Equivalent Stress

In addition to the maximum principal stress, which is governing the behaviour of cracks in materials, also the maximum shear stress $\tau_{\max}$ is an important quantity for the failure behaviour of materials. The distortion of a material caused by the shear stress components appears to govern the mechanism of ductile fracture (see 4.2).

However, it is difficult to compare the various components of the stress tensor (load) with the scalar values that define the material strength, like the yield strength or tensile strength (capacity). Therefore, an equivalent stress can be defined that translates the various components of the stress tensor in any stress state into one representative scalar value. The von Mises equivalent stress σ_{Mises}, based on the assumption that the maximum shear stress causes ductile fracture, is defined as

$$\sigma_{\text{Mises}} = \sqrt{\frac{1}{2}(S_1 - S_2)^2 + \frac{1}{2}(S_2 - S_3)^2 + \frac{1}{2}(S_3 - S_1)^2} \tag{3.14}$$

where S_1, S_2 and S_3 are the three principal stresses. When the principal stresses are not available, the equivalent stress can also be obtained from the generic stress tensor σ_{ij}

$$\sigma_{\text{Mises}} = \sqrt{\frac{1}{2}(\sigma_{11} - \sigma_{22})^2 + \frac{1}{2}(\sigma_{22} - \sigma_{33})^2 + \frac{1}{2}(\sigma_{33} - \sigma_{11})^2 + 3\left(\tau_{12}^2 + \tau_{23}^2 + \tau_{13}^2\right)} \tag{3.15}$$

The thus obtained scalar value σ_{Mises} can be directly used to compare the internal load with the capacity (e.g. tensile stress) of a component. In a design process, the ratios $\frac{\sigma_{\text{Mises}}}{\sigma_{\text{yield}}}$ and $\frac{\sigma_{\text{Mises}}}{\sigma_{\text{tensile}}}$ are often used to asses the structural integrity of a certain design. The ratios should at least be smaller than unity, but in most cases, additional safety factors are applied, for example $\frac{\sigma_{\text{Mises}}}{\sigma_{\text{yield}}} \leq 0.75$.

3.2.5 Elastic Deformation

A material that is subjected to an internal or external load will undergo deformation. As long as the amount of deformation is limited, it will disappear again when the load is removed. This is called elastic deformation, which is a fully reversible process. When the load is increased beyond the yield stress of the material, inelastic or plastic deformation will occur, which is permanent and cannot be removed by reducing the load. In the present subsection, the elastic deformation will be discussed, while plastic deformation will be treated in the next subsection.

A bar that is loaded by a force F in longitudinal direction will show an elongation Δl that is directly proportional to F and the original length l. Moreover, the elongation Δl appears to be inversely proportional to the cross-sectional area A of the bar and also depends on its material properties. The elongation is expressed as

$$\Delta l = \frac{Fl}{EA} \tag{3.16}$$

where the material constant E is called the elastic or Young's modulus. Analogous to the stiffness of a spring, the material stiffness is defined as the ratio between the applied force and the resulting deformation

$$k = \frac{F}{\Delta l} = \frac{EA}{l} \tag{3.17}$$

The stiffness of the bar thus appears to depend on both the geometry of the bar (A and l) and the material properties (E).

The deformation of a material is quantified by the strain, which is represented by the symbol ε. The strain is defined as the relative elongation

$$\varepsilon = \frac{\Delta l}{l} \tag{3.18}$$

Just as the stress is the internal or local representation of the externally applied load (e.g. a force), the strain is the local representation of the elongation. This means that for a certain bar, an external force F may yield an overall deformation Δl. But inside the component, both the stress and strain values may locally deviate from the average values $\frac{F}{A}$ and $\frac{\Delta l}{l}$.

Moreover, just as the stiffness of a component defines the relationship between the applied external load and resulting overall deformation, the local quantities stress and strain are also related. This relation is called Hooke's law and is given by

$$\sigma = \varepsilon E \tag{3.19}$$

The local stiffness is no longer dependent on the geometry of the component, but only depends on the material stiffness represented by the elastic modulus E.

Finally, the elongation of a component in the loaded direction is accompanied by a contraction in the directions transverse to the loading. This is called the Poisson's effect, and the Poisson's ratio ν quantifies the effect. For a uniaxial load in z-direction, the Poisson's contraction ε_x is given by

$$\varepsilon_x = -\nu\varepsilon_z \tag{3.20}$$

For most metals, the Poisson's ratio has a value in the range of 0.2 to 0.4. For a perfectly incompressible material, the value would be 0.5. Rubber shows behaviour that is close to perfect incompressibility and therefore has a Poisson's ratio close to 0.5.

3.2.5.1 Shear Deformation

For a transverse deformation caused by a shear stress, a relationship between stress and strain similar to Hooke's law for a normal load exists. A rectangular surface element is deformed by a shear stress τ and is changed into a parallelogram, see Fig. 3.7. The relationship between the shear stress τ and the shear angle γ is

$$\tau = \gamma G \tag{3.21}$$

The proportionality constant in this case is the shear modulus G, which is again a material constant.

Note that the elastic modulus E and shear modulus G are not always completely independent. For a homogeneous isotropic material, the following relationship between the elastic constants exists

$$E = 2G(1 + \nu) \tag{3.22}$$

which means that two elastic constants are sufficient to fully describe the elastic behaviour of a material.

3.2.5.2 Torsion

In Sect. 3.2.1, it was shown that application of a torque T to a shaft with a circular cross section yields a shear stress that increases linearly with the radius (see Fig. 3.3). This load distribution causes the shaft to deform, where the central axis of the shaft remains straight, while successive cross-sectional planes of the shaft rotate relative to each other. This deformation can be described by the torsion angle θ. For a shaft with length l that is clamped at one side and loaded by a torque T, the torsion angle θ at the free end of the shaft is given by

$$\theta = \frac{lT}{GI_p} \tag{3.23}$$

where I_p is the polar moment of inertia, defined in Sect. 3.2.1 as

$$I_p = \frac{\pi}{32}\left(D^4 - d^4\right) \tag{3.5}$$

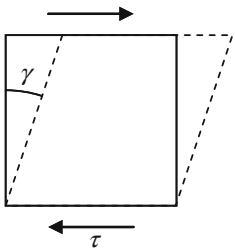

Fig. 3.7 Shear angle γ resulting from a shear stress τ

The torsional stiffness of a shaft, which is the ratio between applied load and resulting deformation, is thus given by

$$k_{\text{torsion}} = \frac{T}{\theta} = \frac{GI_p}{l} \tag{3.24}$$

3.2.5.3 Bending

When a one-side clamped beam with length l is loaded by a transverse force F at the other side, see Fig. 3.8, the end of the beam shows a transverse displacement f.

In every point x, the beam will show a curvature κ equal to

$$\kappa(x) = \frac{\mathrm{d}^2 y(x)}{\mathrm{d}x^2} = \frac{M(x)}{EI} \tag{3.25}$$

with $y(x)$ the local transverse displacement, $M(x)$ the local bending moment, E the material elastic modulus and I the (area) moment of inertia. It can be derived that the integrated effect of all these local curvatures results in a displacement f at the end of the beam equal to

$$f = \frac{Fl^3}{3EI} \tag{3.26}$$

which means that the bending stiffness of this beam is given by

$$k_{\text{bend}} = \frac{F}{f} = \frac{3EI}{l^3} \tag{3.27}$$

Just as for the normal and torsional loads, again the stiffness of the component depends on both geometrical (I, l) and material properties (E).

In conclusion, the (elastic) deformation of mechanically loaded parts or components is determined by the loading on the one hand (force, moment or torque) and on the resistance against deformation (material properties, E, and shape, dimensions like l, A, I, I_p) on the other hand.

Example 3.3 (Elastic Deformation in a Turbine Blade) In example 3.1, it was derived that the centrifugal force in a turbine blade at distance r from the central line is given by

$$\sigma(r) = \tfrac{1}{2}\rho\,\omega^2\left(R_1^2 - r^2\right)$$

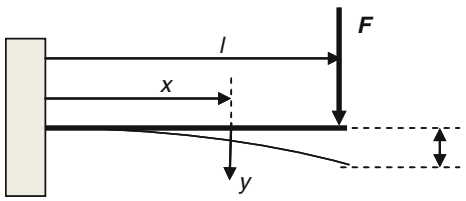

Fig. 3.8 Bending of a clamped beam

Applying Hooke's law, this yields a strain distribution equal to

$$\varepsilon(r) = \frac{\frac{1}{2}\rho\,\omega^2\left(R_1^2 - r^2\right)}{E} \tag{3.28}$$

The elongation of the blade can then be calculated by integrating the local strains from blade root to blade tip, giving

$$\Delta l = \int_{R_0}^{R_1} \varepsilon(r)\mathrm{d}r = \frac{\frac{1}{2}\rho\,\omega^2}{E}\left[R_1^2(R_1 - R_0) - \frac{1}{3}\left(R_1^3 - R_0^3\right)\right] \tag{3.29}$$

3.2.6 Plastic Deformation

As was mentioned in the previous subsection, plastic or inelastic deformation will occur in a material when the applied load exceeds the material yield stress. From that moment, the linear relationship between stress and strain (i.e. Hooke's law) is no longer valid, and after the removal of the applied load, a certain amount of deformation will remain in the material, possibly resulting in residual stresses.

The elasto-plastic behaviour of a material can conveniently be explained using the results of a tensile test. In such a tensile test, a cylindrical test specimen is subjected to a steadily increasing tensile load, while the elongation of the specimen is measured. By plotting the stress σ versus the strain ε, the experimentally determined tensile behaviour of the material can be visualized in a so-called tensile curve. An example of such a curve for low-alloy steel is shown in Fig. 3.9.

Initially, the stress and strain increase proportionally according to Hooke's law. At a certain stress level, the curve starts to deviate from the linear trend, which indicates the end of the elastic regime. A further increase in the stress yields a fast change in slope until a (almost) horizontal curve develops. This is the yielding part of the curve, in which the material behaves (almost) perfectly plastic, that is, a very small increase in stress yields an (almost) infinite increase in plastic strain. The stress level associated with (the onset of) this part of the curve is called the yield stress.

At a certain amount of plastic strain, the stress starts to increase again, caused by hardening of the material. On the microstructural level, processes take place that increase the material strength and stiffness, which makes that the load carried can increase again. After some time, the maximum load that can be carried by the material is reached, which is called the (ultimate) tensile strength, (U)TS. After this stress level has been reached, the cylindrical specimen starts to constrict. This means that local plastic deformation reduces the cross-sectional area of the bar, resulting in an even higher effective stress and more severe plastic deformation. The control system of the testing machine then reduces the applied load and the nominal stress decreases, until fracture of the bar occurs at the fracture stress.

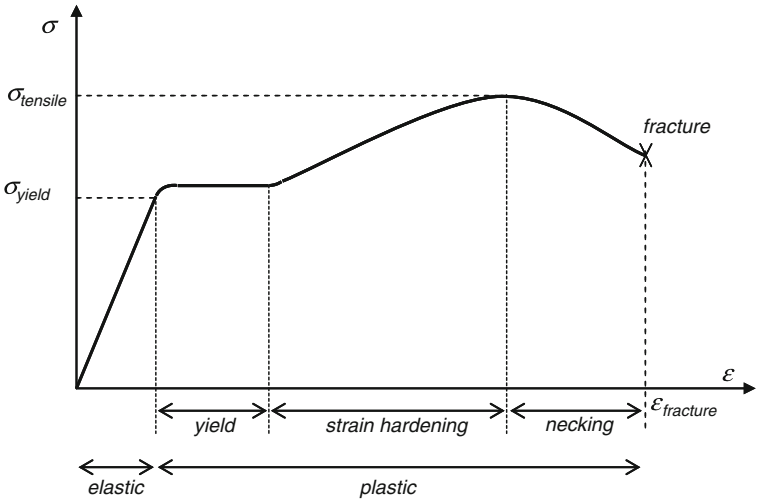

Fig. 3.9 Tensile curve for a low-alloy steel

The associated strain level is called the fracture strain and is a measure of the material ductility. Materials with a large fracture strain are called ductile, while brittle materials already fail at low strains. Comparison of a series of tensile curves with different types of steel shows that with an increase in tensile strength (in most cases associated with an increase in hardness), the stress range between the yield point and the fracture point decreases (in a relative sense). This is due to the fact that in high-strength steels, the possibility for plastic deformation reduces, which means that the materials get more brittle. Therefore, the ratio of tensile and yield stress of a material is a useful indication of the material ductility.

Not all materials exhibit exactly the same behaviour as the low-alloy steel shown in Fig. 3.9. For many materials, the perfectly plastic yield part of the curve does not appear. An example of a material without such a yielding part is the aluminium alloy that is shown in Fig. 3.10. For these types of materials, there is no clear transition from elastic to plastic deformation. Therefore, the yield stress in this case is defined as the stress level at which 0.2 % plastic strain has developed. This quantity is indicated by the symbol $\sigma_{0.2}$ and can be obtained from the tensile curve by drawing a straight line parallel to the elastic part of the curve and starting at 0.2 % strain. The intersection of this line with the tensile curve provides the $\sigma_{0.2}$-value.

3.2.6.1 Residual Stresses

A material that has been deformed elastically will be free of stress after the load has been reduced, but this is not always the case for plastic deformations. In the latter case, the deformed shape does not fit with the initial situation (before

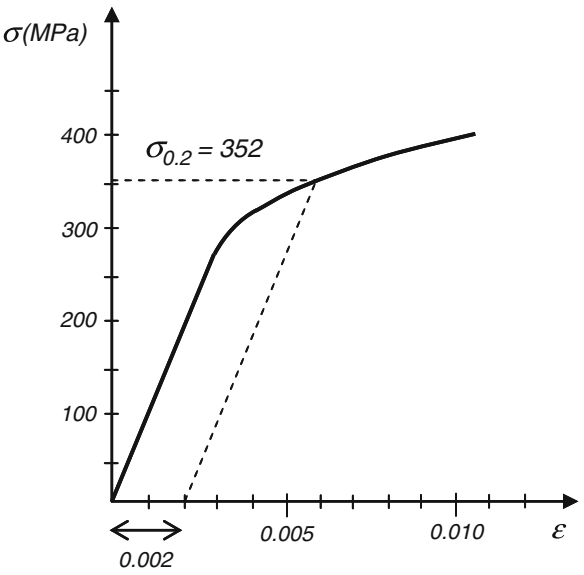

Fig. 3.10 Tensile curve for an aluminium alloy

loading), which means that residual stress may develop in the material. This will be illustrated in the following example on a mechanically overloaded beam and in the next section using the example of an overload due to thermal expansion.

Example 3.4 (Residual Stress: Bending Overload) In the outer-most fibre of a beam (with rectangular cross section $b \times h$), the yield stress $\sigma_{0.2}$ is exceeded when the applied bending moment M exceeds the elastic limit moment M_e. Using Eqs. (3.8) and (3.9), the magnitude of M_e can be derived to be

$$M_e = \frac{1}{6}\sigma_{0.2} b\, h^2 \tag{3.30}$$

If the material exhibits perfectly plastic behaviour (i.e. the stress reaches, but never exceeds $\sigma_{0.2}$), the resulting stress distribution in the beam cross section is shown in Fig. 3.11a.

Further increasing the applied bending moment yields an extension of the plastic region until finally the situation in Fig. 3.11b is reached. The load associated with the latter situation is called the fully plastic moment M_p. The magnitude of this moment, causing the stress distribution in Fig. 3.11b, can be obtained by multiplying the force (F) with the arm (d) for both upper and lower half of the beam

$$M_p = 2Fd = 2\left[\sigma_{0.2} b \times \frac{1}{2}h\right] \times \frac{1}{4}h = \frac{1}{4}\sigma_{0.2} b\, h^2 = \frac{3}{2}M_e \tag{3.31}$$

Figure 3.12a shows the situation where at both sides of the neutral line, the yield stress is exceeded over a height of $h/4$. The magnitude of the bending moment M_{load} associated with this stress distribution is given by

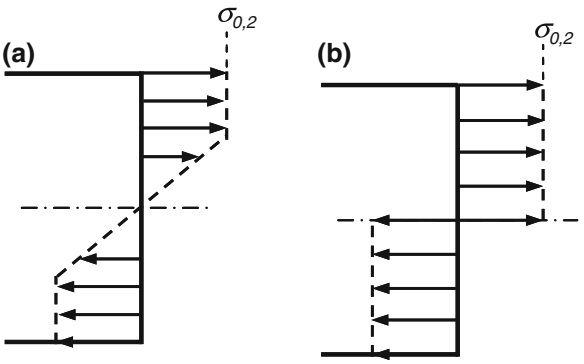

Fig. 3.11 Stress distribution in a beam cross section for **a** loading partly in the plastic regime and **b** loading fully in the plastic regime

$$M_{\text{load}} = 2\left[\sigma_{0.2} b \times \frac{1}{4}h \times \frac{3}{8}h + \frac{1}{2}\sigma_{0.2} b \times \frac{1}{4}h \times \frac{2}{3} \times \frac{1}{4}h\right] = \frac{11}{48}\sigma_{0.2} b\, h^2 \tag{3.32}$$

When the load is removed, the plastic deformation will remain, but the elastic deformation will be recovered, associated with a linear decrease in the stress level. This reverse process is shown in Fig. 3.12b. The moment M_{unload} associated with this stress distribution is

$$M_{\text{unload}} = 2\left[\gamma\sigma_{0.2} b \times \frac{1}{2}h \times \frac{2}{3} \times \frac{1}{2}h\right] = -\frac{1}{6}\gamma\sigma_{0.2} b\, h^2 \tag{3.33}$$

where γ is a yet unknown scaling factor. After the removal of the external load, no bending moment is acting on the beam anymore. Therefore, the sum of the moments of the loading and unloading processes should be equal to zero,

$$M = M_{\text{load}} + M_{\text{unload}} = 0 \tag{3.34}$$

which yields the value of γ

$$\gamma = \frac{\frac{11}{48}}{\frac{1}{6}} = \frac{11}{8} \tag{3.35}$$

Apparently, internal tensile and compressive stresses remain after the plastic deformation. The maximum residual stress exists in the outermost fibre, see Fig. 3.13.

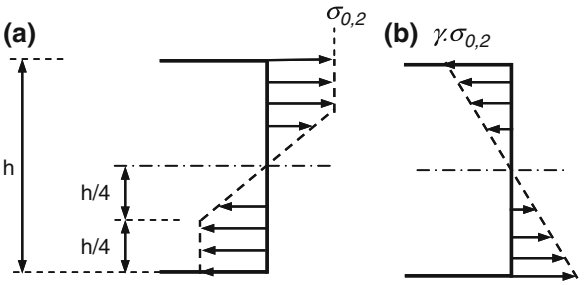

Fig. 3.12 Stress distribution in a beam cross section **a** as a result of externally loading (partly in the plastic regime) the beam and **b** as a result of unloading the beam

3.2.7 Thermal Stress

Thermal stresses develop in a structure when thermal expansion is opposed by some other mechanism that prevents the (full) expansion. Contrary to many other types of stress, thermal stresses are thus not caused by a mechanical load, but by a thermal load, that is, a change in temperature.

When the temperature is increased, most materials exhibit thermal expansion. A bar with initial length l, which is free to expand and is subjected to a temperature increase ΔT (Fig. 3.14), will show a thermal elongation Δl equal to

$$\Delta l = \alpha \, \Delta T \, l \tag{3.36}$$

where α is the coefficient of thermal expansion. A typical value of α for steel is 12×10^{-6}/K. The thermal strain ε_{th} is obtained by the division of the elongation of the bar with the original length

$$\varepsilon_{\text{th}} = \frac{\Delta l}{l} = \alpha \, \Delta T \tag{3.37}$$

Since the bar can expand freely, no thermal stresses will develop in this case.

A different situation occurs when the thermal expansion is obstructed, for example by constraining the bar between two walls (Fig. 3.15a). Also, in this case, the thermal strain is expressed by Eq. (3.37). However, due to the dimensional constraints, the length of the bar cannot change, which means that the thermal expansion must be compensated by an elastic deformation in the opposite direction

$$\varepsilon_{\text{el}} = \frac{\sigma}{E} = -\varepsilon_{\text{th}} = -\alpha \, \Delta T \tag{3.38}$$

A compressive stress (and resulting deformation) thus develops in the bar, which cancels out the thermal expansion. This problem can also be considered as the superposition of a thermal and a mechanical problem, as is visualized in Fig. 3.15. The reaction force F exerted by the wall is the source for the mechanical deformation.

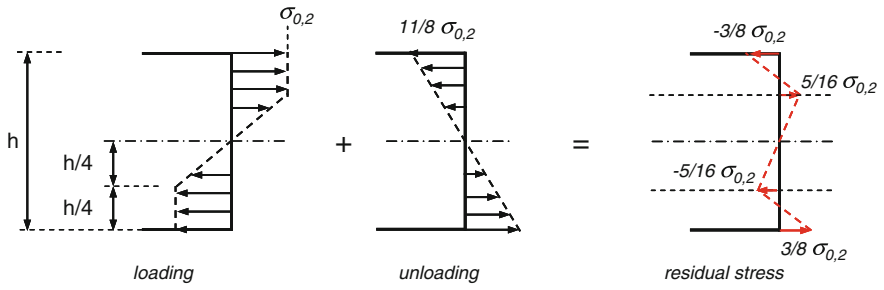

Fig. 3.13 The development of residual stresses after an overload

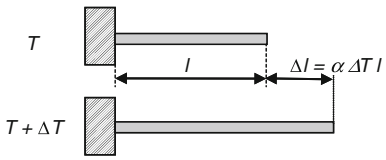

Fig. 3.14 Free thermal expansion of a bar

The balancing of thermal strain by an elastic deformation is possible as long as the resulting stress does not exceed the material yield stress. For thermal loads beyond this point, plastic deformation will occur in the bar, and residual stresses will remain after cooling the bar down to room temperature. For a typical steel (S355) with $\sigma_{0,2} = 355$ N/mm^2, $E = 2.1 \times 10^5$ N/mm^2 and $\alpha = 12 \times 10^{-6}$/K, this means that $\Delta T < 141$ °C to prevent plastic deformation.

An infinitely stiff constraint as in the previous example is not very common in practice. The resulting stresses are generally much more limited since the elasticity of the opposing structure consumes part of the deformation. Moreover, in many machines operating at high temperature, like gas turbines and diesel engines, free thermal expansion is facilitated as much as possible to prevent high thermal stresses to develop. This can, for example, be accomplished by only applying an axial bearing at one location in the machine, offering the shaft the freedom to expand in longitudinal direction.

Equation (3.37) shows that in a structure, differences in thermal strain (and thus thermal stresses) can develop when either the temperature change (ΔT) or the coefficient of thermal expansion (α) varies across the structure. Temperature differences can occur due to different heating rates (e.g. thick vs. thin regions in a component) or as a result of cooling methods (e.g. cooling channels inside gas turbine blades). Variations in coefficient of thermal expansion occur when two different materials are joined together. A well-known example is a ceramic (thermal barrier) coating applied to a metal part. The ceramic expands much less than the metal, so thermal stresses will develop at the ceramic–metal interface.

Example 3.5 (Residual Stresses Due to Thermal Expansion) As explained in the previous subsection, the thermal expansion of a constrained bar already leads to plastic deformation at a modest temperature rise. And when that happens, residual stresses are present in the material after cooling down the bar. Assuming perfectly plastic behaviour, the process of successive heating and cooling is schematically shown in Fig. 3.16.

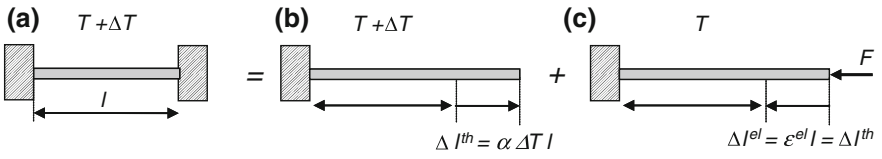

Fig. 3.15 Elastic deformation caused by thermal expansion of a constrained bar

The trajectory 1–2 represents the free thermal expansion of the bar at a temperature increase ΔT. Since the bar is constrained, this expansion must be balanced by an elasto-plastic deformation, making the total strain equal to zero. This deformation is represented by trajectory 2–3. Due to the perfectly plastic behaviour of the material, the magnitude of the stress cannot exceed the yield stress $\sigma_{0.2}$. The horizontal part of the trajectory 2–3 represents the plastic deformation, and the point 3 is the situation in which the heated and constrained bar resides.

The trajectory 3–4 then indicates the free contraction of the bar upon cooling down to the initial temperature, which is exactly the opposite of trajectory 1–2. The contraction is again compensated in 4–5 by elastic deformation to make the total strain zero again. It should be noted that the final stress value (point 5) is larger than zero. This means that a residual stress has developed in the material.

For comparison, the dashed lines in the figure indicate the process for a purely elastic material behaviour. In that case, the stress will be exactly zero after cooling down and no residual stresses are present.

3.2.8 Stress Concentration

Calculation of the global or average stresses in a part or structure generally provides a good indication whether the capacity is sufficient to carry the loads. However, due to irregularities in shape, stress concentrations can locally lead to very high stresses, which may endanger the structural integrity.

An example is shown in Fig. 3.17, where a part with different widths at both ends is uniaxially loaded in tension. Using the finite element method, the distortion of the bar and the stress distribution have been calculated. The results in Fig. 3.17 show that although the transition from the wide to the narrow side is rounded to avoid a sharp edge, the maximum stress near the transition is considerably higher than the average stress. This means that a stress concentration is present that raises the stress level above the expected average stress level.

The stress concentration factor K_t is defined as the ratio between the maximum stress $\sigma_{\max}$ and the (nominal) average stress σ_{avg}

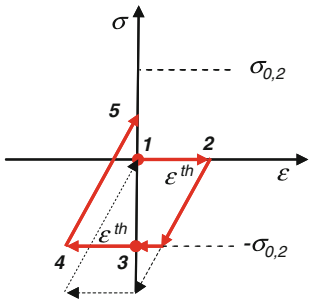

Fig. 3.16 Stress–strain behaviour of a constrained bar during successive heating and cooling

$$K_t = \frac{\sigma_{\max}}{\sigma_{\text{avg}}} \tag{3.39}$$

For the specific case in Fig. 3.17, the value of K_t is shown to depend on the ratio of the widths (w/h) and on the relative size of the corner radius (r/h), see Fig. 3.18. The stress concentration thus increases for larger width transitions and smaller corner radii.

In addition to the width transition shown in Fig. 3.17, also other geometrical discontinuities instigate stress concentrations. Examples are sharp edges, small corner radii, holes and cracks. At the edges of holes in structures, K_t can rise up to a value of 3, while the tip of a crack can be considered as a very small corner radius, which leads to high stress concentrations. The stress distribution around cracks will be treated in more detail in Sect. 3.2.10.

For a large number of general situations, values of K_t have been calculated for a range of dimensions. These solutions are available in handbooks [2], websites or software codes.

3.2.9 Contact Stress

In many practical applications, regular contact occurs between different parts of a system, giving rise to contact stresses. The magnitude of these contact stresses depends on the loads applied to the individual parts, but also and especially on the size of the contact area. Specifically, in spherical or cylindrical parts, the contact area is often small, leading to a relatively high contact stress.

A well-known example is a rolling-element bearing, in which rolling elements (balls, cylinders) facilitate the relative rotation of the inner and outer races. In these bearings, contact occurs between the rolling elements and the races, leading to contact stresses. For a ball element, a point contact will occur, while a cylinder element will show a line contact.

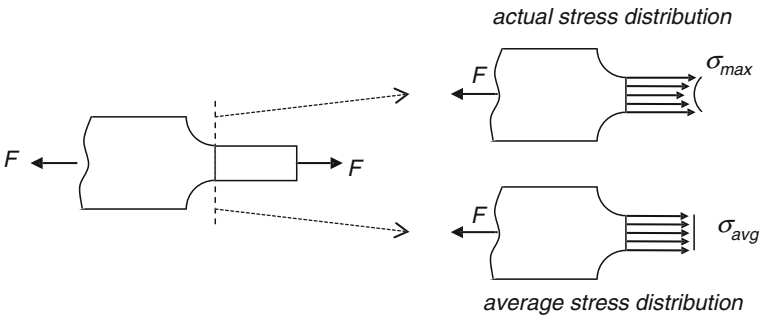

Fig. 3.17 Stress concentration at a geometrical discontinuity

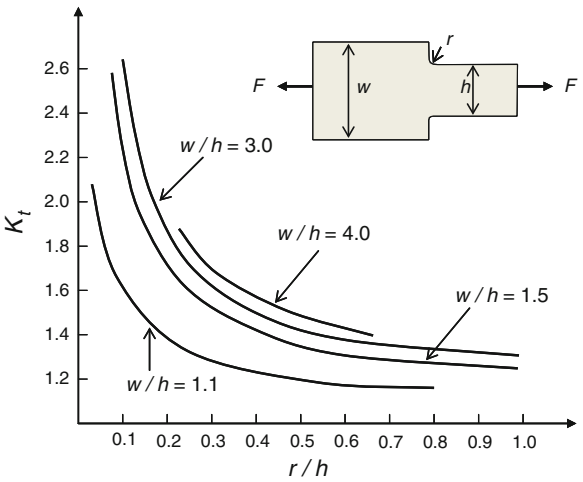

Fig. 3.18 Values of the stress concentration factor for different dimensions

A non-conforming contact is defined as a contact in which the shapes of the bodies are dissimilar enough that under zero load, they only touch at a point (or possibly along a line). This occurs for spherical surfaces with opposite radii (e.g. rolling element and inner ring of bearing) and for large differences in radius. In the non-conforming case, the contact area is small compared with the sizes of the objects and the stresses are highly concentrated in this area, see Fig. 3.19. For these cases, the Hertzian theory on contact problems can be applied. The most relevant equations from that theory will be discussed next.

The width of the contact area in between two cylinder-shaped bodies can be calculated as

$$b = C_1 \sqrt{\frac{F}{l r_{\text{eff}}}} \tag{3.40}$$

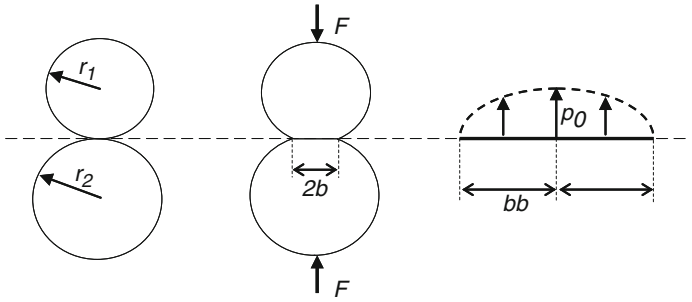

Fig. 3.19 Two bodies with opposite radii constitute a non-conforming contact. Application of a normal load F yields a contact area width $2b$ and the stress distribution shown on the *right*

where F is the applied normal force and l the length of the cylinder. The constant C_1 depends on the elastic properties of the materials used

$$C_1 = \sqrt{\frac{4}{\pi E^*}} = \sqrt{\frac{4}{\pi}\left[\frac{1-\nu_1}{E_1} + \frac{1-\nu_2}{E_2}\right]} \tag{3.41}$$

For a typical steel, $C_1 = 3.34 \times 10^{-3}$ mmN$^{-0.5}$. The effective radius is defined as

$$r_{\text{eff}} = \frac{1}{r_1} \pm \frac{1}{r_2} \tag{3.42}$$

For opposite curvatures of the bodies, the +sign is valid, while for equally directed curvatures, the −sign should be used.

The maximum contact pressure (p_0) is given by

$$p_0 = C_2\sqrt{\frac{r_{\text{eff}}F}{2l}} \tag{3.43}$$

The constant C_2 again depends on the elastic properties of the materials used

$$C_2 = \sqrt{\frac{2E^*}{\pi}} \tag{3.44}$$

For a typical steel, the value equals $C_2 = 270$ mm^{-1}N$^{0.5}$.

The resulting shear stress distribution is shown in Fig. 3.20. In the upper graph on the right-hand side, the contact pressure is plotted, and in the lower graph, the variation in the shear stress component with depth is presented. It appears that the shear stress component τ, which governs the failure process, has its maximum at a depth of $0.78b$.

Example 3.6 (Contact Stresses in a Cylindrical Roller Bearing) A specific cylindrical roller bearing has the following characteristics:

- Basic dynamic load rating: $C = 35{,}500$ N
- Inner race diameter: $D_i = 37.5$ mm
- Outer race diameter: $D_o = 55.5$ mm
- Cylinder diameter and length: $D_c = 9$ mm and $l_c = 10$ mm
- Number of cylinders: 13.

Analogous to the calculation in example 2.6, it can be shown that the force F_0 acting on the most severely loaded cylinder equals 35 % of the bearing load F. For a load $F = 0.1\ C = 3{,}550$ N, the contact area on the cylinder in contact with the inner race can be calculated with Eq. (3.40), using $C_1 = 3.34 \times 10^{-3}$ mmN$^{-0.5}$. This yields a half contact area width $b = 0.17$ mm.

The maximum contact pressure is then $p_0 = 1.34 \times 10^3$ N/mm^2 (using $C_2 = 270$ mm^{-1}N$^{0.5}$). Due to this high contact pressure, large stresses will develop in the inner race underneath the surface. For failure, particularly the shear

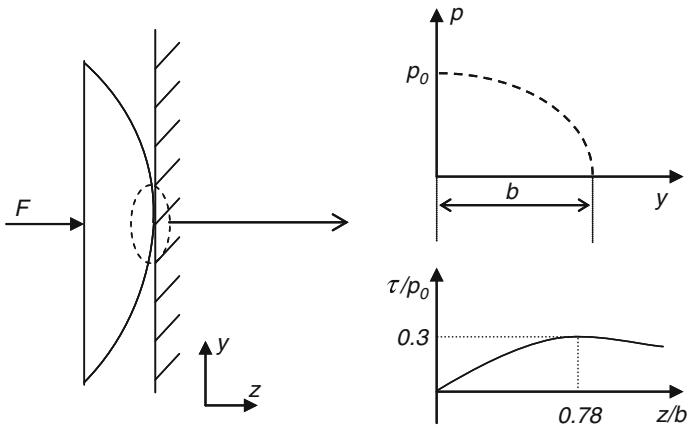

Fig. 3.20 Distribution of the contact pressure and the shear stress below the contact area for a Hertzian contact

stresses are important, for which it was shown that the maximum value exists at some depth below the surface. For the component τ, the maximum value exists at a depth of $0.78b$. At each passage of the cylinder, the stress will vary between 0 and $0.3p_0$. For the shear stress component τ_{xy}, the variations are even higher: the stress fluctuates between $-0.25p_0$ and $0.25p_0$ at a depth of $0.5b$. This stress variation will lead to the initiation of cracks at some depth below the surface, in this particular bearing at a depth of 85–130 μm (see also section on surface fatigue, 4.6.7).

3.2.10 Stress Around Cracks

After prolonged operation of a system or structure, cracks can develop in the material due to several failure mechanisms. Moreover, most materials already contain initial defects from the manufacturing stage, and also, welding processes can cause cracks and pores to initiate. The cracks weaken the structure that might fail at a nominal load far below the yield strength of the material. It is therefore important to understand that how cracks affect the stress distribution in a structure.

Depending on the loading conditions, cracks can be divided into three modes. These are the tensile (mode I), shear (mode II) and tearing (mode III) modes of cracking, as is shown in Fig. 3.21. In this subsection, only the most common mode I cracks are discussed, and for the formulations of the other modes, the reader should refer to textbooks on fracture mechanics [3].

Based on the theory of elasticity, the stress distribution can be derived for a tensile-loaded infinitely large flat plate, in which an ellipsoidal-shaped opening is present. To simulate a real crack, the height of the ellipsoid is reduced to zero, see Fig. 3.22.

For the two-dimensional (plane stress) situation in Fig. 3.22, the distribution of the different stress components can be derived, yielding the following equations

$$\sigma_x = \frac{S\sqrt{\pi a}}{\sqrt{2\pi r}} \cos\frac{\theta}{2}\left(1 - \sin\frac{\theta}{2}\sin\frac{3\theta}{2}\right) - S \tag{3.45}$$

$$\sigma_y = \frac{S\sqrt{\pi a}}{\sqrt{2\pi r}} \cos\frac{\theta}{2}\left(1 + \sin\frac{\theta}{2}\sin\frac{3\theta}{2}\right) \tag{3.46}$$

$$\tau_{xy} = \frac{S\sqrt{\pi a}}{\sqrt{2\pi r}} \cos\frac{\theta}{2}\left(\sin\frac{\theta}{2}\cos\frac{3\theta}{2}\right) \tag{3.47}$$

where S is the nominal stress that is applied to the plate (at a large distance from the crack), a is the half crack length and r and θ define the distance and direction to the crack tip.

The equations demonstrate that the stress level is dominated by the so-called stress intensity factor K

$$K = S\sqrt{\pi a} \tag{3.48}$$

The stress intensity factor is the basic concept of the theory of fracture mechanics. It should be noted that the intensity of the stress field caused by the crack not only depends on the applied stress, but also depends on the length of the crack.

Equation (3.48) is valid for the ideal case of an infinite plate with a central crack under tensile loading. For any other situation, like bending loads, finite plate width or edge cracks, the stress intensity factor is modified by a geometry factor F

$$K = FS\sqrt{\pi a} \tag{3.49}$$

The value of F can be obtained from stress intensity factor solution handbooks (e.g. [4]) for a range of standard problems or can be calculated using a finite

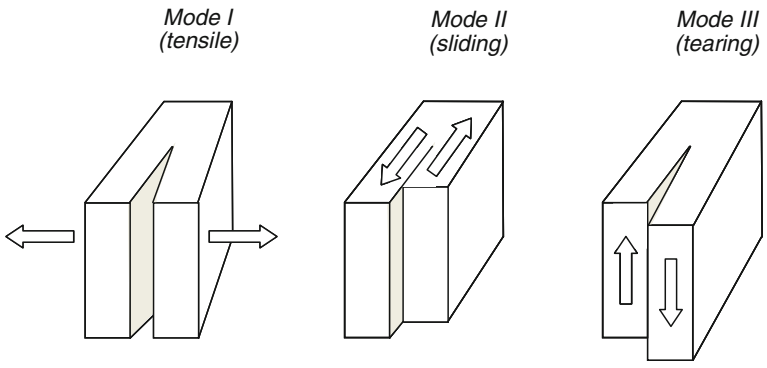

Fig. 3.21 Schematic representation of the three different cracking modes

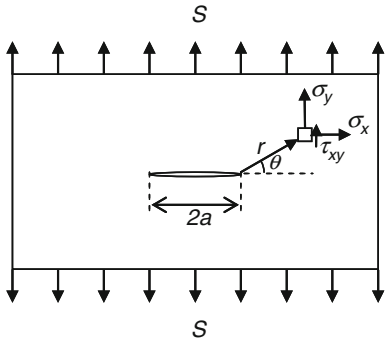

Fig. 3.22 Definition of stress components in an infinite plate with a central crack

element analysis for any other situation, as will be discussed in more detail in Sect. 4.4.6.

The derivation of Eqs. (3.45) to (3.47) is based on elasticity theory, which implies linear behaviour of the materials. However, the equations show that at the crack tip ($r \to 0$), the stress will become infinitely large. In a real situation, this cannot happen, since a plastic zone will develop in front of the crack, in which the material no longer behaves in a linear way. For a relatively small plastic zone ($r_p \ll a$), Eqs. (3.45) to (3.47) are a good approximation. The size of the plastic zone r_p can be approximated by substituting $\theta = 0$, $\sigma_y = \sigma_{0.2}$ and $r = r_p$ in Eq. (3.46):

$$r_p = \frac{1}{\pi}\left(\frac{K}{\sigma_{0.2}}\right)^2 \tag{3.50}$$

For the values $S = 100$ N/mm^2, $\sigma_{0.2} = 400$ N/mm^2 (aluminium) and $a = 10$ mm, this yields a plastic zone size $r_p = 0.6$ mm.

Finally, many structures are periodically inspected to check whether cracks are present and to identify their lengths. In many cases, this is done visually, but especially in the tip region, it is very hard to differentiate between cracked and sound material, which makes assessment of the crack length difficult. The crack opening distance is, however, easier to measure, and since it is directly related to the crack length, this quantity can be used to indirectly determine the crack length. The relationship between crack opening distance v and the crack length a is

$$v = 2\frac{S}{E}a \tag{3.51}$$

with E the elastic modulus of the material. For example, at a load of $S = 100$ N/mm^2, a crack of $a = 10$ mm in a material with a yield strength $\sigma_{0.2} = 70{,}000$ N/mm^2 will show a crack opening $v = 0.06$ mm.

3.3 Thermal Loads

In the previous chapter, the heat flow (q) was shown to be the generic *external* thermal load, which is enabled by the mechanisms of conduction, convection and radiation. The *internal* load parameter associated with this heat flow is the temperature, commonly indicated by the symbol T. The same heat flow applied to different bodies will generally yield quite different temperature values inside the bodies. But since the temperature governs the failure mechanisms (e.g. overheating, melting, creep), it is important to understand how the (change in) temperature can be obtained for a given external thermal load. That will be the topic of the present section.

Temperature is a so-called state parameter, which always has a value. Whereas a body can be stress free if no mechanical loads are applied, it can never be temperature free. All bodies will have a temperature that is larger or equal to absolute zero, which is defined as $T = 0\ \mathrm{K} = -273.15\ ^\circ\mathrm{C}$.

The temperature of a body can only change when heat is supplied to the body or extracted from the body. As was discussed in the previous chapter, a heat flow q can deliver or extract a certain amount of heat through convection, conduction or radiation. Another possibility is the establishment of a heat flow due to internal heat generation inside the body (e.g. due to friction or electric losses).

The magnitude of the temperature rise caused by a certain heat input Q is governed by the heat capacity c_p and the mass m of the body

$$\Delta T = \frac{Q}{m\, c_p} \tag{3.52}$$

The heat capacity is a material property that indicates how much heat is required to raise the temperature of a unit mass of that material by one degree. Typical values range from $c_p = 100$ to 500 J/kgK for metals up to 4,000 J/kgK for water.

In practical situations, many heat flows will take heat into and out of the body simultaneously. The net effect of all these heat flows will then determine whether the temperature increases or decreases. For example, an internally cooled gas turbine blade will be heated from the outside by the hot gasses passing the blade, while it is simultaneously cooled on the inside by convection to the cooling air and by conduction to the metal disc to which the blade is attached. When the heat flows from the outside and the inside are constant in time, a steady-state blade temperature will establish. In this situation, the ingoing and outgoing flows exactly balance each other and the net heat flow is zero, as will be demonstrated in the next example.

Example 3.7 (Thermal Loading of a Gas Turbine Blade) In the previous chapter, the external thermal loads on a gas turbine blade have been calculated (Example 2.10). The schematic representation of the turbine blade and its surrounding is again shown in Fig. 3.23. To ensure that the metal temperature (T_{wall}) stays within

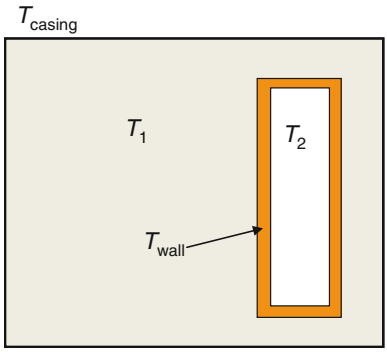

Fig. 3.23 Metal and gas temperatures in the surrounding of a turbine blade

acceptable limits, the blade is internally cooled by gas with a temperature $T_2 = 800$ °C. The temperature of the hot gas in the gas path of the engine is $T_1 = 1{,}500$ °C, and the temperature of the metal casing is $T_{casing} = 500$ °C.

The thermal load on the turbine blade appeared to consist of several contributions due to convection and radiation. The total thermal load on the turbine blade was calculated to be $q_{tot} = 7.23 \times 10^5$ W/m^2.

In the situation used in this example, the calculated heat flow is positive, which means that there is a net heat flow *into* the turbine blade. This means that the wall temperature will increase, and a new situation will arise: the incoming contribution q_{c1} decreases, whereas the outflowing contributions q_{c2} and q_{r1} increase, resulting in a much lower thermal load. This process will continue until a steady-state situation is reached in which the net heat flow $q_{tot} = 0$.

When the heat capacity of the turbine blade is known, the increase in the wall temperature in a certain time period can be calculated. This first requires the calculation of the total amount of heat Q that is flowing into the blade:

$$Q = q_{tot} \times \Delta t \times A$$

with A the total surface area of the blade. Using values of $A = 0.01$ m^2, heat capacity $c_p = 300$ J/kgK and a time period of 1 s, the total amount of heat equals $Q = 7{,}230$ J. Then, using Eq. (3.52), the temperature increase for a mass of 0.4 kg will be

$$\Delta t = \frac{Q}{mc_p} = \frac{7,230}{0.4 \times 300} = 60.3\,^{\circ}\text{C}$$

Note that this temperature increase would considerably change the various heat flows, which means that in the next second, the temperature rise will be quite different.

Example 3.8 (Thermal Equilibrium in a Gas Turbine Blade) In the previous example, the heat flows in a turbine blade have been identified. However, it was assumed that the temperature distribution in the blade wall is homogeneous, that is, the temperature over the wall cross section is constant. In a real situation, the

conduction of heat from outside to inside will cause a temperature gradient across the wall thickness.

This is schematically shown in Fig. 3.24, where the variation in temperature from the outside gas temperature (T_{gas}), going across the wall thickness (the grey region) to the cooling gas temperature (T_{cool}), is visualized.

The figure shows that three temperature gradients exist: two discrete drops in temperature at the transitions from gas to metal (ΔT_1 and ΔT_3) and one gradual decrease in temperature inside the metal part (ΔT_2). At the interface between gas and metal, heat transfer by convection occurs. Note that the seemingly discrete step in temperature exactly at the interface in reality will be a temperature gradient in the thin boundary layer in the gas close to the blade surface.

When all the heat flows are constant in time, an equilibrium situation will establish and the inflow of heat will be exactly balanced by the outflow of heat. The wall temperature for this situation will be determined now.

The relationship between heat flow and temperature drop at the gas–metal interfaces (for ΔT_1 and ΔT_3) is

$$q_{conv} = h\left(T_{gas} - T_{wall}\right) \Leftrightarrow \Delta T = \frac{q_{conv}}{h} \tag{3.53}$$

where h is the heat transfer coefficient, which attains different values at the outside and inside of the blade. Inside the metal, conduction arises, yielding the following expression for ΔT_2

$$q_{cond} = k\frac{\Delta T}{\Delta x} \Leftrightarrow \Delta T = \frac{q\,\Delta x}{k} \tag{3.54}$$

with k the heat conduction coefficient and Δx the wall thickness of the blade.

The two gas temperatures (T_{gas} and T_{cool}) are known, and the sum of the three temperature drops must cover the difference between the two gas temperatures. Finally, in the equilibrium situation, the ingoing and outgoing flows exactly balance each other, which means that also the flow that connects the in and outflow, that is, the conduction inside the blade, must be equal in magnitude. Thus, the following requirements must be fulfilled

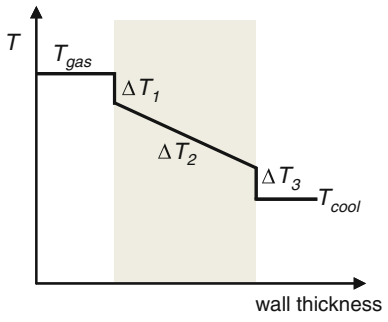

Fig. 3.24 Variation in temperature across a turbine blade wall

$$\begin{aligned} \Delta T_1 + \Delta T_2 + \Delta T_3 &= T_{\text{gas}} - T_{\text{cool}} \\ q_{\text{in}} &= q_{\text{cond}} = q_{\text{out}} \end{aligned} \tag{3.55}$$

Solving the set of equations then yields expressions for the (metal) wall temperatures at the hot gas side and at the cooling air side

$$T_{\text{wall, hot}} = T_{\text{gas}} - \frac{T_{\text{gas}} - T_{\text{cool}}}{1 + h_{\text{gas}}\left(\frac{\Delta x}{k} + \frac{1}{h_{\text{cool}}}\right)} \tag{3.56}$$

$$T_{\text{wall, cool}} = T_{\text{cool}} + \frac{T_{\text{gas}} - T_{\text{cool}}}{1 + h_{\text{cool}}\left(\frac{\Delta x}{k} + \frac{1}{h_{\text{gas}}}\right)} \tag{3.57}$$

If very high values are selected for the heat transfer coefficients h_{gas} and h_{cool}, the equations show that $T_{\text{wall,hot}} = T_{\text{gas}}$ and $T_{\text{wall,cool}} = T_{\text{cool}}$. This means that for a situation with very effective heat transfer from gas to metal, the blade wall will attain the respective gas temperatures and the complete temperature difference between hot gas and cooling air will be accommodated inside the blade.

On the other hand, when the value of the conduction coefficient k is very high, the temperature gradient inside the metal will be small and the temperature difference will be covered at the two interfaces with the gas. This illustrates that the ratio between the values of h and k determines what equilibrium will be established. The absolute values of the parameters then determine the magnitude of the heat flow through the metal part.

3.4 Electric Loads

The external electric loads have been discussed in the previous chapter. It was shown that the presence of electric charge and the specific distribution of this charge in space and time constitute the generic electric load. But again, for the failure mechanism at the material level, not the external load, being the absolute value of the charge, potential or current is leading, but the internal load quantified by its gradient or density. To better understand the relationship between electric loads and capacity, first the electric properties of materials on the atom level will be discussed briefly and the specific characteristics of semi-conducting materials will be treated. After that, the internal electric loads, that is, electric field strength and current density, will be discussed.

3.4.1 Electric Properties

All atoms in a material consist of a very small nucleus, composed of protons and neutrons, which is encircled by moving electrons. Both protons and electrons have an electric charge with magnitude 1.602×10^{-19} C, but for electrons, this is a

negative charge, while protons have a positive charge. Electric conduction is accommodated by electrons moving through the conductor, but not all electrons can move freely and thus contribute to the conduction. The number of free electrons per atom depends on the arrangement of electron states (certain energy levels) and the occupation of these states by electrons [5]. A detailed treatment of this subject is beyond the scope of this book, but the essentials will be discussed next.

In a solid material consisting of a large number of atoms, the single electron states are merged together into electron energy bands. The number and position of the energy bands depend on the material type. The electrons associated with the atoms occupy part of the available bands. The highest occupied state at 0 K is called the Fermi energy E_f, and only electrons with energies exceeding this E_f can contribute to electric conduction and are then called free electrons. At very low temperatures (near 0 K), conduction is impossible in many materials, since all electrons are in states $\leq E_f$. However, electrons can be excited to higher-energy states by, for example, raising the temperature or by delivering energy in other ways, for example, by incident radiation. Whether this is possible largely depends on the energy gap between the filled states and the nearest empty states. In principle, four different band structures are possible, as is schematically shown in Fig. 3.25.

For conducting materials like metals, empty states are available in the same band adjacent to the highest filled states (Fig. 3.25a) or the next (empty) band partially overlaps with the filled band (Fig. 3.25b). In both cases, electrons can easily jump to empty states, which explains the high conductivity of these materials. In insulators, a large band gap exists between the filled band (which is called the valence band in this case) and the empty (conduction) band (Fig. 3.25c). The band gap is typically larger than 2 eV, and electrons can thus not jump to free states. Therefore, no free electrons are available in the material as mobile charge carriers. Finally, the band structure of semi-conducting materials is in between that of conductors and insulators, since the band gap is relatively small (<2 eV).

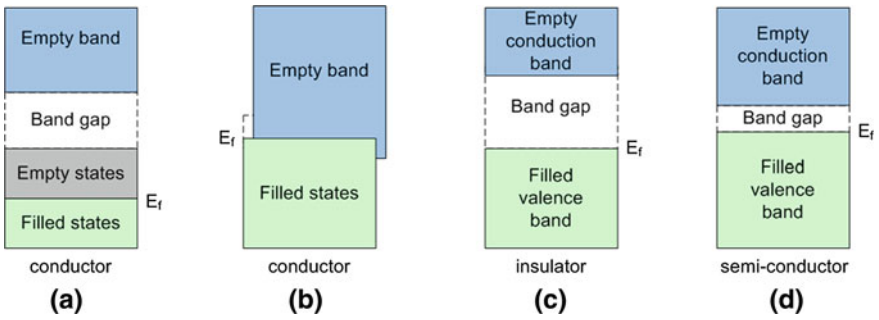

Fig. 3.25 The different possible energy band structures in **a** and **b** conducting, **c** insulating and **d** semi-conducting materials

Thermal activation can in these materials accommodate the creation of free electrons and thus increases the conductivity of the material.

3.4.2 Semiconductors

Semi-conducting materials are very important for electrical applications, since their properties form the basis for many electrical devices. The main reason is that the electric properties of some semiconductors are very sensitive for even very small concentrations of impurities, which enables the creation of very specific effects in these materials. Those materials for which the properties largely depend on the impurity atoms are called extrinsic semiconductors. Before their behaviour is discussed, the intrinsic semiconductors will be treated.

An intrinsic semiconductor, for example, silicon or germanium, has a band structure as shown in Fig. 3.25d, with a small band gap between the valence and conduction band. For each electron that is excited into the conduction band, a vacancy or hole is created in the valence band. If electrons from other states in the valence band jump into the empty state, the hole is actually moving and thus is a positive charge carrier. Therefore, the conduction process in an intrinsic semiconductor is accommodated by electrons and holes, each with their own mobility μ (see also Sect. 2.4), so the conductivity σ is given by

$$\sigma = n|e|\mu_e + p|e|\mu_h \tag{3.58}$$

where the concentration of holes p, that is, the number of holes per unit volume, exactly equals the concentration of electrons n, which is therefore called the intrinsic carrier concentration.

Almost all commercial semiconductors are however extrinsic semiconductors, in which impurity atoms provide excess electrons or holes that increase the conductivity. Only very small concentrations (in the order of one impurity atom per 10^{12} atoms) are sufficient to drastically modify the electric properties. If an impurity is added that provides an additional electron, the material is called an n-type semiconductor, while a p-type semiconductor contains impurity atoms that deliver an excess hole (or actually a deficit of one electron). An example of the former type is silicon (Si) doped with phosphorus (P), and a typical p-type material is silicon doped with boron (B). Since in these types of materials, the number of electrons and holes is not equal anymore; the conductivity is now governed by the majority charge carriers. For example, in a p-type semiconductor, it is given by

$$\sigma = p|e|\mu_h \tag{3.59}$$

where the carrier density now is determined by the dopant concentration. Note however that the mobility of the holes decreases when the impurity concentration is increased too much.

The most simple application of *p*- and *n*-type semiconductors is the creation of a *p*–*n* junction which acts as a rectifier. Such a junction is created from a single piece of semi-conducting material (e.g. silicon) where the doping is different on both sides of the junction. In the *p*-type region of the junction, holes are the charge carriers, whereas electrons accommodate the conduction in the *n*-side of the junction. Application of a potential difference across the junction with the positive pole of the battery connected to the *p*-side causes both the holes and the electrons to move towards the junction. At the junction, they recombine and cancel, resulting in a considerable flow of charge carriers and a corresponding low resistivity. This situation is called the forward bias of the junction.

If the battery is connected in the opposite way (see Fig. 3.26), the reverse bias situation is obtained. In that case, both the holes and electrons are drawn away from the junction and almost no free charge carriers remain in the junction region. Now, the junction is highly insulative.

3.4.3 Electric Field and Current Density

For the potential, it was already shown in the previous chapter that its gradient determines the electric field strength (E)

$$\vec{E} = -\vec{\nabla} V = -\begin{pmatrix} \frac{\mathrm{d}V}{\mathrm{d}x} \\ \frac{\mathrm{d}V}{\mathrm{d}y} \\ \frac{\mathrm{d}V}{\mathrm{d}z} \end{pmatrix} \tag{3.60}$$

Since the capability of material to withstand a certain electric load is expressed in terms of breakdown field strength (E_{bd}, see also 4.9.2) and not in applied

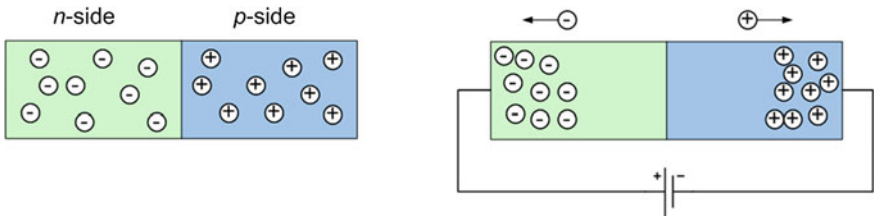

Fig. 3.26 A *p*–*n* rectifying junction with the distributions of holes and electrons for no applied potential (*left*) and reverse bias (*right*)

voltage, the electric field is the true internal load parameter. This means that the internal electric load on an insulator, externally loaded by a certain potential difference, will increase as the thickness of the insulating layer is reduced, since the resulting electric field gets more intense.

For the electric current, a similar approach is followed. Not the absolute value of the current, but the local current density is the internal load parameter that governs failure. The current density (J) is defined as

$$J = \frac{I}{A} \tag{3.61}$$

with A the cross-sectional area of the part (e.g. wire) that conducts the current. Analogous to the mechanical stress being the local representation of the applied force F, taking into account the specific cross-sectional area of the part loaded by F, the current density is the local representation of the electric current I. A current flowing in a thin wire will yield a much higher current density than the same current flowing in a thick wire. The thin wire will therefore fail sooner than the thick wire.

Finally, the relationship between voltage (ΔV) and current (I) is normally expressed in terms of Ohm's law on an engineering level

$$\Delta V = IR \tag{3.62}$$

where the resistance R is the proportionality constant. However, the same relationship can also be expressed in terms of the internal loads current density (J) and electric field strength (E)

$$J = \frac{E}{\rho} = \sigma E \tag{3.63}$$

In this case, the proportionality constant is the internal (local) representation of resistance, that is, the resistivity ρ or its reciprocal σ (conductivity). Values of the

Table 3.1 Resistivity values for some typical materials

Material class	Material	Resistivity ρ [Ωm]
Conductors	Silver	1.47×10^{-8}
	Copper	1.72×10^{-8}
	Lead	22×10^{-8}
Semiconductors	Graphite	3.5×10^{-5}
	Germanium	0.6
	Silicon	2,300
Insulators	Glass	10^{10}–10^{14}
	Mica	10^{11}–10^{15}
	Quartz	75×10^{16}

resistivity for some typical conductors, insulators and semiconductors are shown in Table 3.1.

3.4.4 Dielectric Behaviour

A dielectric material is an electrically insulating material that may exhibit an electric dipole structure. This means that a separation of positive and negative charge carriers occurs on an atomic level. The specific electric properties associated with this behaviour make these materials suitable to be applied in capacitors.

A capacitor generally consists of two parallel conducting plates to which a voltage is applied. As a result, one of the plates gets positively charged, while the other attains a negative charge, and an electric field is established in between the plates. A capacitor is characterized by its capacitance C, which indicates the amount of charge Q that is stored for a given potential difference ΔV:

$$C = \frac{Q}{\Delta V} \tag{3.64}$$

The unit of capacitance is coulomb per volt (C/V) or farad (F). Further, the capacitance is proportional to the dimensions of the parallel plates, that is, the surface area A and the spacing l of the plates. If a vacuum exists between the plates, the capacitance is given by

$$C = \varepsilon_0 \frac{A}{l} \tag{3.65}$$

The constant ε_0 is a universal constant (8.85×10^{-12} F/m) and is called the permittivity of vacuum. However, if a dielectric material is inserted in the region between the plates, the capacitance increases with a factor ε_r, which is the dielectric constant of the material. The electric field strength generated in between the two plates is thus also increased by the presence of the dielectric material. The values of the dielectric constant for some typical materials are shown in Table 3.2.

Table 3.2 Dielectric constant for some typical materials [5]

Material class	Material	Dielectric constant ε_r
Ceramics	Mica	5.4–8.7
	Porcelain	6.0
	Fused silica	3.8
Polymers	Nylon	3.6
	Polystyrene	2.6
	Polyethylene	2.3

3.5 Chemical Loads

In the previous chapter, the chemical loads have been introduced. A division was made between chemical loads causing a direct chemical reaction and loads that are responsible for electrochemical (corrosion) reactions. In this subsection, only the latter type of (electro-)chemical loads will be treated since the corrosion process is a relevant and very common failure process. The direct chemical attack of materials, which is especially relevant for the process industry, is a much more specific topic that is outside the scope of this book.

3.5.1 Electrochemical Reactions

As was mentioned before, most common corrosion processes are driven by electrochemical reactions, where electric charge is transferred in a (mostly aqueous) solution. An example is the reaction between zinc and hydrochloric acid

$$Zn + 2HCl \rightarrow ZnCl_2 + H_2 \tag{3.66}$$

From the reaction, it can be observed that the zinc and hydrogen atoms (Zn, H) and the ions Zn^{2+} and H^+ are the reacting species, while the Cl^- ions do not contribute. This chemical reaction can be split in two half reactions (after removing the chloride ions from the equations): the anodic or oxidation reaction

$$Zn \rightarrow Zn^{2+} + 2e^- \tag{3.67}$$

and the cathodic or reduction reaction

$$2H^+ + 2e^- \rightarrow H_2 \tag{3.68}$$

In the overall reaction, the metal zinc is thus transformed into zinc chloride, which is soluble in water, and hydrogen gas. This means that the metal steadily dissolves and thus degrades. The electrons released in the anodic reaction migrate to the adjoining surface, where they can react with the hydrogen ions (cathodic reaction). The ions (Zn^{2+} and H^+) can only move if they are carried by a fluid (generally water) or condensed vapour. This fluid is called the electrolyte. Finally, it is evident that during the reaction, electric charge (two electrons per oxidized zinc atom) is transferred, which is the reason to denote it as an electrochemical reaction. Note that in an equilibrium situation, the number of electrons generated in the anodic reaction exactly equals the number of electrons absorbed in the cathodic reaction. However, if some external source provides excess electrons, the rate of corrosion (as expressed by the anodic reaction) is reduced, while the rate of the cathodic reaction increases. This will be shown later (Sect. 4.10) to be the basis of cathodic protection.

3.5.2 Electrode Potentials

Due to the electrochemical nature of the reactions in (3.67) and (3.68), each half reaction is associated with a certain single electrode potential, which is characteristic for that specific half reaction. Moreover, these half-cell electrode potentials constitute the driving force for the corrosion reaction. It has been derived [6] that the change in free energy ΔG of a corroding system equals

$$\Delta G = -nFE \tag{3.69}$$

where n is the number of electrons exchanged in the reaction (e.g. 2 in the Zn + HCl reaction) and F is Faraday's constant, equal to 96,487 C. The electrochemical potential E of the reaction is the combined effect of the anode (e_a) and cathode (e_c) electrode potentials

$$E = e_a + e_c \tag{3.70}$$

The change in free energy determines whether a spontaneous reaction can occur. If in any chemical reaction, the reaction products have a lower free energy than the reactants, a spontaneous reaction can take place and ΔG will have a negative value. For corrosion reactions, this means that the sign of the free energy change determines the direction of the reaction, that is, whether the oxidation or reduction reaction prevails.

To determine the value (and sign) of ΔG for a corrosion reaction, the values of the half-cell electrode potentials must be known. These values can be obtained from the electromotive force (emf) series, a listing of electrode potentials at standard conditions [7]. A short overview of some common corrosion reactions and their electrode potentials is provided in Table 3.3. Since the absolute values of the half-cell potentials cannot be measured directly, a potential difference measurement with a standard reference electrode is used to determine the values. The standard hydrogen electrode (SHE) is commonly used for that purpose, which is the reason that its single electrode potential is set to 0.000 V in the emf series (see also Table 3.3).

Table 3.3 Standard electrode potentials [7]

	Reaction	Standard potential e^0 (volts vs. SHE)
Noble	$Au^{3+} + 3e^- \rightarrow Au$	+1.498
	$Pt^{2+} + 2e^- \rightarrow Pt$	+1.118
	$Cu^{2+} + 2e^- \rightarrow Cu$	+0.342
	$2H^+ + 2e^- \rightarrow H_2$	0.000
	$Fe^{2+} + 2e^- \rightarrow Fe$	−0.447
	$Zn^{2+} + 2e^- \rightarrow Zn$	−0.762
	$Al^{3+} + 3e^- \rightarrow Al$	−1.662
Active	$K^+ + e^- \rightarrow K$	−2.931

Finally, the materials having reactions with positive electrode potentials are designated noble (the more noble they are, the higher the potential is), while the negative electrode potentials represent the active materials. When an electrochemical cell is constructed with two electrically connected metal electrodes submersed in an electrolyte, the potential difference measured across the electrodes can be predicted from the electrode potentials in the emf series.

Example 3.9 (Potential Difference in Electrochemical Cell) For the reaction

$$\mathrm{Zn} + \mathrm{Cu}^{2+} \rightarrow \mathrm{Zn}^{2+} + \mathrm{Cu} \tag{3.71}$$

the electrode potentials for the two half reactions can be obtained from Table 3.3. Note that the zinc half reaction is given in opposite direction in Table 3.3, which means that the sign of the electrode potential must be reversed. Therefore, if a Zn electrode is combined with a Cu electrode and the electrolyte has a standard concentration, then the voltage across the cell will be

$$E = e_a + e_c = -e_{\mathrm{Zn/Zn^{2+}}} + e_{\mathrm{Cu/Cu^{2+}}} = 0.762\ \mathrm{V} + 0.342\ \mathrm{V} = 1.10\ \mathrm{V}$$

The positive value of the potential means that according to (3.69), the change in free energy ΔG has a negative sign and the reaction will occur spontaneously, that is, zinc metal will go into solution and copper will be formed on the Cu electrode from the Cu^{2+} ions in the electrolyte. If the potential would have been negative, the opposite reaction would have been the spontaneous reaction.

The half-cell electrode potentials in Table 3.3 are valid for standard conditions, where the temperature is 25 °C (298 K) and the concentration of ions in the electrolyte is 1 M (mol/l). However, in practical situations, these conditions will rarely occur. In that case, the Nernst equation can be applied to calculate the actual electrode potential. Considering the generic half reaction for a metal M

$$\mathrm{M}^{n+} + ne^- \rightarrow \quad \mathrm{M} \tag{3.72}$$

the Nernst equation is given by

$$e = e^0 - \frac{RT}{nF} \ln[\mathrm{M}^{n+}] \tag{3.73}$$

where e^0 is the standard electrode potential of M, T is the absolute temperature (in K), R is the gas constant (8.314 J/molK) and $[M^{n+}]$ represents the concentration of the ions in the electrolyte. For the standard concentration of 1 M, the second term becomes zero and the potential will equal the standard potential.

Using this equation, the half-cell electrode potentials for any condition can be calculated and thus the electrochemical cell potential for any combination of half cells. In that way, the driving force for corrosion reactions can be obtained and it can be determined whether or not a certain reaction will take place. However, it is important to notice that the magnitude of the potential (as the driving force) is not an indication for the rate of the reaction, that is, a reaction with a large potential is

not necessarily a rapid reaction. The rate depends on the kinetics of the reactions, as will be discussed in the next chapter (Sect. 4.10).

3.5.3 Electrochemical Loads

From the discussion in the previous subsections, it can be concluded that for an (electrochemical) corrosion process, the relevant loads are as follows:

- the electromotive force
- the concentration of ions in the electrolyte
- the temperature.

These loads are governed by both the types of material involved and the environmental conditions. Assuming that the type of material applied in a certain system is a design choice, which determines the load-carrying capacity of the system, the usage of the system governs the other factors: the reduction reaction (required to form an electrochemical cell) and the magnitude of the electromotive force. These factors will determine whether failure occurs and what the degradation rate will be. This will mainly depend on the environment the system is used in, that is, the presence of a more noble material in electric contact (galvanic corrosion), the presence of an aqueous solution or vapour and the pH-value of any present acid. In the next chapter, the different corrosion mechanisms will be treated and determination of the magnitude of corrosion rates will be discussed.

3.6 Radiative Loads

Similar to the electric loads, also for radiative loads, the density of the radiative energy governs the failure of a system rather than the total amount of energy. This means that the energy should be related to either a certain time period or a specific surface area, in order to obtain a value for the energy density.

Moreover, only the part of the radiation that is absorbed by the receiving body contributes to the damage. This is governed by the reflection or absorption coefficient of the material, which depends on the material type, colour and surface finish. A black body will absorb more radiation than a white body, and a shiny and highly reflecting surface will reduce the absorption of radiation considerably.

3.7 Summary

In this chapter, the internal load parameters have been discussed for a range of load types. The characteristic of an internal load is that it represents the local loading condition at some spot in a component, part or structure. And since failure occurs

at the material level, this local loading drives the failure process and the associated internal load parameter is essential in assessing the integrity of the component.

For all internal load parameters, formulations have been provided to obtain their values from the external loads. Generally speaking, this involves relating the external loads to both geometric dimensions and material properties. Finally, several examples have been given to demonstrate this process for real applications.

References

1. Hibbeler, R.C.: Mechanics of Materials, 6th edn. Pearson Education, Upper Saddle River, NJ, USA (2005)
2. Pilkey, W.D.: Peterson's Stress Concentration Factors. Wiley, New York (1997)
3. Janssen, M., Zuidema, J., Wanhill, R.J.H.: Fracture Mechanics, 2nd edn. Delft University Press, Delft (2002)
4. Tada, H., Paris, P.C., Irwin, G.R.: The Stress Analysis of Cracks Handbook. Del Research Corporation, St. Louis, USA (1973)
5. Callister, W.D., Rethwisch, D.G.: Materials Science and Engineering, 8th edn. Wiley, Hoboken, NJ, USA (2011)
6. Jones, D.A.: Principles and Prevention of Corrosion, 2nd edn. Prentice-Hall, Upper Saddle River (1996)
7. Haynes, W.M. (ed.): Handbook of Chemistry and Physics, 91st edn. CRC Press (2010)

Further Reading

1. Callister, W.D., Rethwisch, D.G.: Materials Science and Engineering, 8th edn. Wiley, Hoboken, NJ, USA (2011)
2. Jones, D.A.: Principles and Prevention of Corrosion, 2nd edn. Prentice-Hall, Upper Saddle River (1996)
3. Timoshenko, S.P., Goodier, J.N.: Theory of Elasticity, 3rd edn. McGraw-Hill International Editions (1970)
4. Purcell, E.M.: Electricity and Magnetism, 2nd edn. McGraw-Hill, New York (1985)

Chapter 4
Failure Mechanisms

4.1 Introduction

In the previous chapters, the different load types have been treated extensively and it has been discussed how they can be translated into internal loads or failure parameters. This chapter will discuss the various failure mechanisms, that is, the physical mechanisms underlying the failures of parts, components or structures.

As was discussed in Chap. 1, failure is considered here as reaching such a state that the intended function of the part or system can no longer be fulfilled. Therefore, as was also mentioned, failure does not always imply the real physical failure of a part. It was also discussed in Chap. 1 that failure occurs when the load on a system exceeds its load-carrying capacity. The present chapter will describe for various failure mechanisms how this capacity is defined and which (internal) loads govern the failure process by exceeding this capacity. The following failure mechanisms will be treated:

- Mechanical failures: static overload, deformation, fatigue, creep and wear
- Thermal failures: melting and thermal degradation
- Electric failures: current overload and intrinsic breakdown
- Chemical failures: corrosion
- Radiative failures

After treating these single failure mechanisms, the chapter will be concluded with a subsection on failure processes and interaction between failure mechanisms.

4.2 Static Overload

Static overload is the most intuitive failure mechanism. It occurs when the applied load, quantified by the mechanical stress, exceeds the static strength of the material. The static strength or tensile strength is a material property, which in this case constitutes the load-carrying capacity of the part or system.

T. Tinga, *Principles of Loads and Failure Mechanisms*,
Springer Series in Reliability Engineering,
DOI: 10.1007/978-1-4471-4917-0_4, © Springer-Verlag London 2013

The tensile strength can be determined experimentally by performing a tensile test, in which an increasing force is applied to a geometrically well-defined test specimen (see also 3.2.6). The stress level at which the test bar fails is defined as the tensile strength. This property, as well as the yield strength at which plastic deformation starts, is temperature dependent. For most materials, the strength decreases at higher temperatures.

The design of a system or structure and the choice for a certain material should be such that the expected stress level during operation is lower than the strength of the material. Generally, an additional safety factor is applied to increase system reliability. However, despite the safety factor, still many systems fail due to static overload. This is mostly caused by unexpected operational conditions, like rotational speeds or temperatures (decreasing the strength) that significantly exceed their normal values.

In a tensile test, the load is increased steadily and the plastic deformation can gradually develop in the material. At impact loads, this is not possible and the material may behave differently. To investigate the ability of a material to undergo fast plastic deformation, a (Charpy) impact test on a notched specimen can be performed. By impacting a predefined mass from different heights onto a standardized test specimen, the energy at which the material fractures is determined. This energy thus characterizes the material toughness or ductility. For tough materials that are able to deform plastically at these high strain rates, the amount of energy absorbed will be high and the process is called ductile fracture. On the other hand, brittle fracture occurs in materials that do not allow fast plastic deformation. In that case, cleavage occurs caused by the tensile stress acting on crystallographic planes with a low bonding strength. The transition from brittle to ductile fracture is defined to occur at an energy of 27 J in the impact test.

Materials that are ductile at room temperature may exhibit brittle fracture at low temperatures. The variation of the material toughness (i.e. energy at fracture) with temperature for a typical steel is shown in Fig. 4.1. The temperature associated with an energy of 27 J is denoted the ductile to brittle transition temperature (DBTT). Its value depends on the chemical composition and microstructure of the material. This transition constitutes one of the root causes for the sinking of the

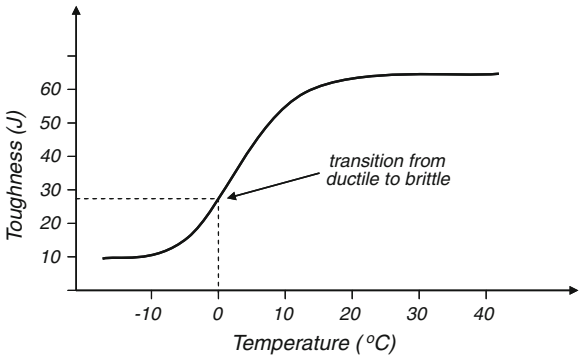

Fig. 4.1 Typical material toughness variation with temperature. The ductile to brittle transition is defined to occur at 27 J

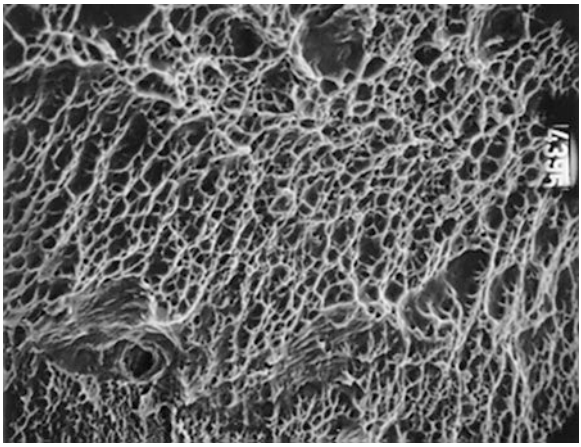

Fig. 4.2 Micrograph of static overload fracture surface (published with kind permission of © NLR 2012. All Rights Reserved)

famous steamship Titanic in the year 1912. The designers of the ship did not acknowledge that the steel used for the hull had a DBTT around 0 °C. The high-impact load caused by the collision with an iceberg resulted in a fast brittle fracture of the hull over a length of almost 100 m.

If the failure has serious consequences, a failure analysis is generally performed to determine the root cause. One of the methods used in such an analysis is the examination of the fracture surface with an optical or electron microscope. Static overload failure of a ductile material can then be recognized by the dimples on the fracture surface. An example of a fracture surface micrograph is shown in Fig. 4.2.

During the deformation process, voids develop in the material due to locally high stresses. This particularly happens near irregularities in the material microstructure, like precipitates, impurities and inclusions (see Fig. 4.3), that provides a stress concentration (see 3.2.8). At increasing load, the voids will grow and cause a gradual degradation of the material in between the voids. At some moment, the remaining cross section of the part is insufficient to carry the load. The voids then coalesce and failure occurs. The voids created during this process can be recognized as the characteristic dimples on the fracture surface (Fig. 4.2).

At brittle fractures, no plastic deformation occurs and therefore no voids are initiated. The cleavage process associated with brittle fracture yields a rather flat and shiny fracture surface.

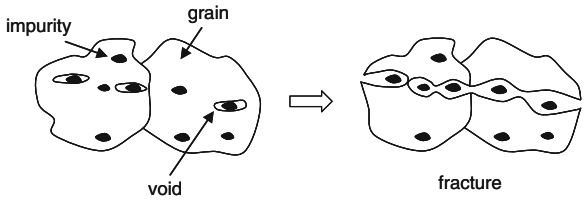

Fig. 4.3 Overload failure due to the formation and coalescence of micro voids

Example 4.1 (Fracture of a Turbine Blade Due to Overload) In example 3.1, the centrifugal stress in a gas turbine blade at some distance r from the engine centre line was derived to be

$$\sigma_{cf} = \frac{F_{cf}}{A} = \frac{1}{2}\rho\omega^2\left(R_1^2 - r^2\right) \tag{4.1}$$

with ρ the material mass density, ω the rotational speed of the blade and R_1 the radius at the blade tip. It was also shown that the stress is maximum at the blade root at $r = R_0$, having a value

$$\sigma_{\mathrm{cf,root}} = \frac{1}{2}\rho\omega^2\left(R_1^2 - R_0^2\right) \tag{4.2}$$

Assuming that for a specific blade $R_0 = 0.2$ m and $R_1 = 0.36$ m, and using the mass density for a nickel-based superalloy, $\rho = 8{,}300$ kg/m^3, the stress at the blade root can be calculated for any rotational speed. For example, for $N = 10{,}000$ rpm, that is, $\omega = 1{,}047$ rad/s, the stress equals 408 MPa.

Further, when also the strength of the material is known, the maximum allowable rotational speed can be calculated. A typical tensile strength for a superalloy is $\sigma_{\mathrm{tensile}} = 1{,}300$ MPa. This means that an overload failure of the blade will occur when the centrifugal stress exceeds this strength, which in this case will happen at $N = 17{,}855$ rpm.

4.3 Deformation

In some cases, the excessive deformation of a part (without fracture) can be the failure mechanism. For many rotating parts, the dimensions must stay within tight tolerances to prevent dragging or seizure. There is a variety of sources for excessive deformation, like elastic or plastic deformation, thermal expansion and creep. The most important internal load parameter associated with deformation is mechanical stress, but also temperature plays an important role. For the creep mechanism (see 4.5) and thermal expansion, temperature is the dominant load parameter that governs the failure mechanism. Further, both elastic and plastic deformation are affected by the temperature through its effect on the material properties like elastic modulus and yield strength. Also for this failure mechanism, the anticipated stress and temperature levels are accounted for in the design of the system, but unexpected operating conditions may still lead to failures.

Example 4.2 (Excessive Deformation of a Turbine Blade) In the example in the previous subsection, the rotational speed at which overload failure of a turbine blade occurs was calculated. Another failure mode for a turbine blade would be excessive deformation (elongation) of the blade leading to rubbing of the casing or even seizure of the blade.

From Eq. (4.1) and Hooke's law, it can be directly derived that the strain as function of the radius is given by

$$\varepsilon(r) = \frac{\sigma(r)}{E} = \frac{1}{2}\frac{\rho}{E}\omega^2\left(R_1^2 - r^2\right) \tag{4.3}$$

with E the material elastic modulus. With this expression, the local deformation of the blade at some radius r is obtained, so the total elongation of the blade can be calculated by integrating the strain from R_0 to R_1. This yields the following expression for the elongation ΔL

$$\Delta L = \int_{R_0}^{R_1} \varepsilon(r)dr = \frac{1}{2}\frac{\rho}{E}\omega^2\left(\frac{1}{3}R_1^3 - R_1^2R_0 + \frac{1}{3}R_0^3\right) \tag{4.4}$$

Using the blade dimensions from the previous example, that is, $R_0 = 0.2$ m and $R_1 = 0.36$ m, and using the elastic modulus for a nickel-based superalloy, $E = 150$ GPa, the elongation as a function of rotational speed is

$$\Delta L = 2.17 \cdot 10^{-10}\omega^2 \tag{4.5}$$

This means that for a rotational frequency of $N = 10{,}000$ rpm, the blade elongation will be 0.24 mm.

Finally, if a spacing of 5 mm is present between the non-rotating tip of the blade and the casing, the rotational speed at which the blade will touch the casing can be calculated. This yields a value of $N = 45{,}817$ rpm. Note that this value is very high since it is assumed here that only the elastic deformation of the blade due to the centrifugal force causes the elongation. In reality, also the thermal expansion, possibly plastic deformation and creep will contribute to the blade elongation. Blade rubbing will therefore already be encountered at much lower rotational speeds.

4.4 Fatigue

In Sect. 4.2, the mechanism of static overload was discussed, showing that failure occurs when the applied stress exceeds the material strength. Early in the twentieth century, however, it was discovered that axles of trains failed after some period of operation, although the axles were designed properly, that is, the stress levels in the axles were always below the material strength. After extensive research, it was discovered that the fatigue mechanism was responsible for these failures.

When a repetitive load is applied to a component, fatigue failure can occur at load levels which are well below the ultimate strength of the material and failure occurs at seemingly safe stress levels. The number of cycles to failure depends on the magnitude of the applied cyclic load and on the capacity of the material, which in this case is characterized by the material fatigue strength. Typically, materials fail due to fatigue after 10^3–10^7 cycles.

In practical applications, many systems and their components are subjected to cyclic loads, for example, caused by rotations or vibrations. But as long as the magnitude of the alternating stress is relatively small due to a robust design of the

system, failure will only occur after a large number of cycles. However, in aerospace applications, weight limitations enforce designers to apply thin structures which are just sufficient to carry the anticipated loads and cyclic loads may become critical for the service life of the structure. For that reason, a lot of research has been done on fatigue in aerospace materials and besides knowledge on the fundamentals of the mechanism, also many numerical models and tools are now available.

In this section, firstly, the fatigue mechanism and its related loads will be discussed. After that, the available methods to calculate the fatigue life of cyclically loaded parts are treated, where generally two different approaches are followed. The first approach considers the material as a continuum, in which fatigue damage accumulates that eventually leads to failure. The fracture mechanics approach, on the other hand, explicitly addresses the cracks in the material as discontinuities and bases the life assessment on the behaviour of these cracks. The continuum approach will be treated in 4.4.4 and 4.4.5 while the fracture mechanics approach will be discussed in 4.4.6.

4.4.1 Definitions for Cyclic Loads

A cyclic load can be quantified using a number of parameters, as is shown in Fig. 4.4. The most important parameters are the stress range $\Delta\sigma$ and the stress amplitude σ_a that indicate the magnitude of the cycle. Additionally, the mean stress σ_m plays an important role. Its value can also be obtained from the stress ratio R, the ratio between maximum and minimum stress during a cycle. In Fig. 4.4b, two examples are shown: at $R = -1$, the mean stress equals zero and both tensile ($\sigma > 0$) and compressive stresses ($\sigma < 0$) occur. At $R = 0$, all stresses are tensile and σ_m attains a non-zero value.

4.4.2 Fatigue Mechanism

On the level of the microstructure, materials mainly deform by crystal planes sliding along each other. The slip planes that are activated are very specific sets of

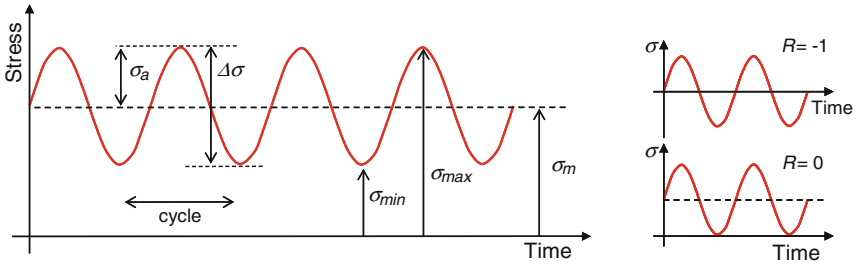

Fig. 4.4 Definition of parameters to quantify cyclic loads

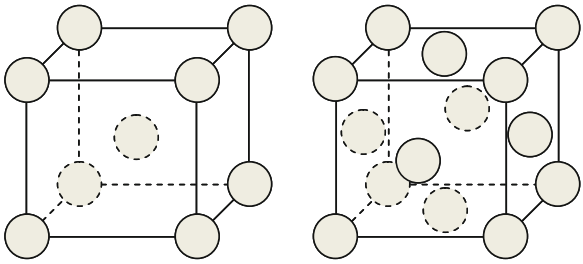

Fig. 4.5 Two cubic lattice structures: FCC (*left*) and BCC (*right*)

planes that depend on the crystal lattice. The two most common crystal lattices, face-centred cubic (FCC) and body-centred cubic (BCC), are shown in Fig. 4.5.

For materials with a FCC structure, the available slip planes and slip directions are shown in Fig. 4.6. A slip system consists of a combination of a slip plane and a slip direction. Combining the planes and directions shown in Fig. 4.6 yields 12 (4 planes/3 directions) octahedral and 12 (6 planes/2 directions) cubic slip systems. Deformation of the crystal can only occur by sliding along one of these slip systems.

As soon as slip has occurred along one slip system, the material will generally harden, which means that it becomes more difficult to create additional deformation on the same slip plane. This principle is the basis for the fatigue mechanism on a micro scale. This is illustrated in Fig. 4.7. During the first upward loading, a surface step is created. The following downward loading will cause slip in the opposite direction, but not exactly on the same slip plane, for the reason explained above. After some more cycles, a micro crack is formed as an intrusion

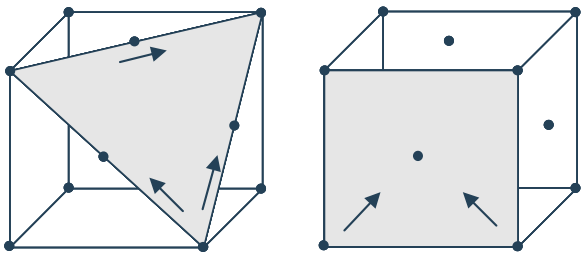

Fig. 4.6 Octahedral (*left*) and cubic (*left*) slip systems in an FCC structure

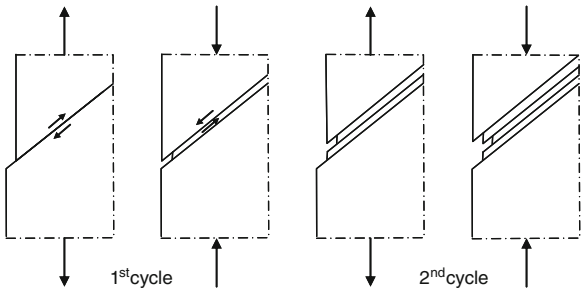

Fig. 4.7 Illustration of the mechanism for the initiation of a fatigue crack

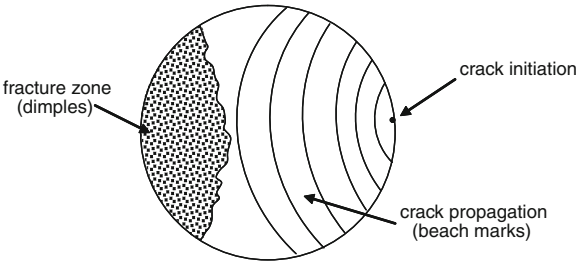

Fig. 4.8 Schematic representation of a fatigue fracture surface

into the material. This micro crack will act as a stress concentration, and the crack can start to propagate from there.

After the initiation (or nucleation) of this micro crack, continued cyclic loading of the part causes extension of the crack, which is called crack propagation. As the cracked fraction of the cross-sectional area of the part cannot carry any load, the stress in the remaining cross section increases with increasing crack length. At some moment, the local stress exceeds the material strength and fracture due to static overload will occur. This is schematically shown in Fig. 4.8, where a crack initiated in a bar with circular cross section. The fracture zone at the left-hand side represents the static overload failure, showing the characteristic dimples (see Sect. 4.2) on the fracture surface.

The crack propagation part on the right-hand side of the fracture surface in Fig. 4.8 also shows characteristic features, which are called beach marks or striations. These very fine parallel lines on the fracture surface have been created by the stepwise propagation of the crack front caused by the cyclic load. During failure analysis, the number and spacing of the striations may provide information on the number and magnitude of the load cycles that caused the fracture. A micrograph with a characteristic fatigue fracture surface is shown in Fig. 4.9.

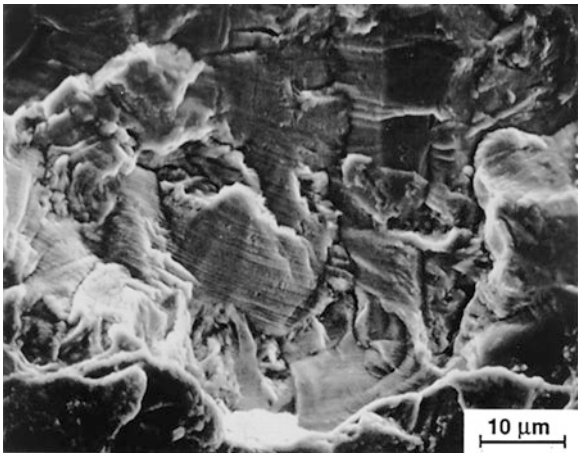

Fig. 4.9 Micrograph of a characteristic fracture surface for fatigue failure (published with kind permission of © NLR 2012. All Rights Reserved)

4.4.3 Low-Cycle and High-Cycle Fatigue

A distinction is often made between low-cycle fatigue (LCF) and high-cycle fatigue (HCF). There are two ways to distinguish between these two mechanisms. The first way is to just consider the number of cycles to failure. In most cases, the transition between LCF and HCF is defined to be at 10^6 cycles to failure, although other authors define the transition at 10^4 cycles [1]. This illustrates that the distinction between LCF and HCF based on the number of cycles to failure is rather arbitrary.

A better way to distinguish the two types of fatigue is to consider the character of the loading that is applied. If the maximum stress during the load cycles does not exceed the material yield stress, that is, the deformation is fully elastic, the process is called high-cycle fatigue. If also some amount of plastic strain develops during the cyclic loading, caused by the stress exceeding the yield stress in part of the cycle, it is called low-cycle fatigue. In the latter case, the stress range $\Delta\sigma$ is not an appropriate quantity to define the magnitude of the load, since during plastic deformation, the stress hardly increases. For that reason, the load in case of low-cycle fatigue is often defined in terms of strain range $\Delta\varepsilon$.

High-cycle fatigue is often caused by vibration of the components. This generally leads to relatively small loads, but due to the high frequency of the cyclic load, the required large number of cycles leading to fatigue failure can be reached within relatively short periods of time. For example, gas turbine compressor blades can start to vibrate when they are excited at frequencies close to their natural frequency, that is, typically 500–1,000 Hz. This means that 10^6 cycles can be accumulated in only a couple of minutes. This illustrates the specific problem of high-cycle fatigue: occurrence of the phenomenon does not leave much time to react, which means that in most cases HCF failure (e.g. fracture) will take place. For that reason, much effort is taken in the design process of critical systems to prevent HCF to happen.

An example of such a critical system is a gas turbine, in which especially the slender compressor blades are prone to vibrations and therefore susceptible for HCF. In gas turbine design, the Campbell diagram (see Fig. 4.10) is applied to prevent HCF failures. This diagram consists of two sets of lines. The dashed lines represent the natural frequencies of the component that is considered, which are plotted as a function of rotor speed. The lines represent the various vibrational modes, that is, 1st flapping mode (1F), 1st torsional mode (1T). The natural frequencies slightly increase with higher rotor speeds, since the components generally get somewhat stiffer when they are extended.

The second set of lines is called the engine orders (EO). These lines represent the different excitation frequencies that are expected to be present during operation. The first engine order (1EO) is equal to the rotor speed, the 2EO is the double rotor speed, etc. If in the engine inlet the bearing is centred by three struts, the rotating blades will pass three struts each revolution. Upon passing, the aerodynamic flow will be slightly different and the blade will feel a small pulse. This repetitive pulsing will provide an excitation at a frequency of three times the rotor

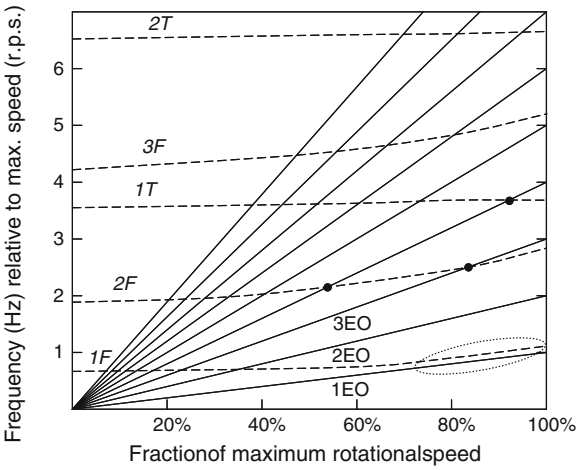

Fig. 4.10 Campbell diagram used to prevent high-cycle fatigue

frequency, which is the 3EO. Since it is very probable that the frequencies associated with the lower EO are present, the design of the components must be such that the natural frequencies are not equal or close to those EO.

The Campbell diagram visualizes this process. Intersections between the two sets of lines at low rotor speeds are no problem, since they only occur during startup and shutdown of the engine, and the amplitudes of vibrations are small. However, in the region between 80 and 100 % rotor speed, no intersections should be present, since they would represent operating conditions at which HCF may occur. The designer should then modify the shape, dimensions or material properties of the component to move the natural frequency away from the engine order.

Even when a system is designed properly using methods as described above, HCF may still occur during operation. If for some reason the natural frequency of the component changes while in service, an intersection with one of the EO may still be possible. A well-known example in aerospace applications is so-called foreign object damage (FOD), where objects from outside the aircraft (e.g. stones from the runway or ingested birds during flight) enter the engine and hit the rotating parts. The resulting damage may change the natural frequency of the part, and HCF can be initiated.

4.4.4 Life Assessment for Constant Amplitude Loads

For constant amplitude loads, where the magnitude of the alternating stress is constant in time, the number of cycles to failure for a certain load level can be obtained from an *S–N* curve or Wöhler curve. This curve provides the relation between stress amplitude (S) and the number of cycles to failure (N_f), and is available in handbooks for all common materials or can be determined experimentally. An example of an *S–N* curve is shown in Fig. 4.11, in which the curves for two different types of materials are given. Note that a certain *S–N* curve is

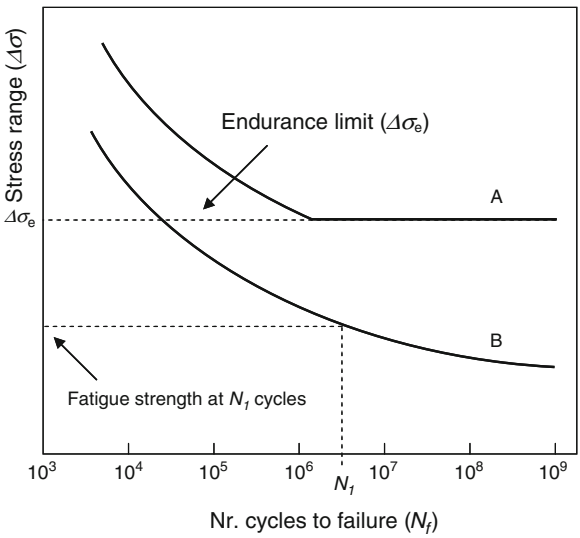

Fig. 4.11 Typical *S–N* curves for materials with (*A*, e.g. steel) and without (*B*, e.g. aluminium) an endurance limit

determined experimentally for a specific temperature and *R* value. If these values are different in the component to be analysed, another curve has to be used or a correction must be performed.

For many materials, for example, steel and titanium alloys, an endurance limit exists. For loads lower than this limit, no fatigue failure will occur, see Fig. 4.11. However, corrosion or an acid environment can reduce this limit or even make it disappear. For other materials, with aluminium as a well-known example, no endurance limit exists and fatigue should be accounted for at all stress levels.

The *S–N* curve can also be expressed as a numerical relation between applied stress range $\Delta\sigma$ and number of cycles to failure N_f, as proposed by Basquin [2]:

$$\frac{\Delta\sigma}{2} = A\left(2N_f\right)^b \tag{4.6}$$

where A and b are constants to be determined from *S–N* data. Moreover, it was mentioned in the previous subsection that for low-cycle fatigue, the load is often quantified in terms of strain range $\Delta\varepsilon$. In that case, the material fatigue resistance must also be provided as a $\varepsilon{-}N$ curve in stead of a *S–N* curve. The associated numerical equation is called the Coffin-Manson relation [3, 4]:

$$\frac{\Delta\varepsilon_p}{2} = \varepsilon_f\left(2N_f\right)^c \tag{4.7}$$

where $\Delta\varepsilon_p$ is the *plastic* strain range and c and ε_f are empirical constants.

To determine the number of cycles to failure for a combination of a cyclic and constant stress (= non-zero mean stress), the Smith diagram can be used. In this diagram, the sum of the mean and alternating stress ($\sigma_m \pm \sigma_a$) is plotted versus the mean stress (σ_m), see Fig. 4.12, resulting in a contour of constant service life

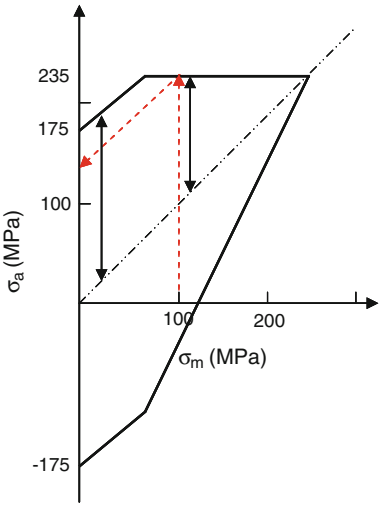

Fig. 4.12 Smith diagram for a typical steel alloy and a given number of cycles to failure (e.g. 10^6 cycles)

(e.g. 10^6 cycles). This type of diagram shows that for an increasing value of the mean stress, the allowable alternating stress decreases. The dashed lines show that for this alloy at a mean stress level of 100 MPa, an additional alternating stress of 135 MPa is allowed to reach the indicated service life. The sum of the mean stress and alternating stress then equals 235 MPa, which is the yield strength of the alloy. The upper right corner of the contour represents the combination of a mean stress equal to the yield strength and a negligible alternating stress.

The service life can thus be calculated from these diagrams when the magnitude of the applied stresses is known. The expected stress levels (σ_m and σ_a) can, for example, be calculated with a finite element (FE) analysis, but generally must be corrected for a number of life-extending or life-limiting effects, like stress concentrations, size factors, surface roughness and internal stresses. Additionally, safety factors are generally applied to prevent unexpected failures at loads that slightly deviate from the expected loads. These factors will be discussed next.

4.4.4.1 Stress Concentrations

Discontinuities (holes) and shape transitions in a design can raise the stress level considerably, as was discussed in Sect. 3.2.8. This means that the nominal stress that is obtained from the FE analysis must be multiplied by the stress concentration factor K_t to get the real stress value

$$\sigma_{\text{real}} = K_t \sigma_{\text{nom}} \tag{4.8}$$

which mostly considerably reduces the fatigue life time of the component.

4.4.4.2 Component Size

From the description of the fatigue mechanism in 4.4.2, it can be concluded that the cyclic loading causes the initiation of cracks, which eventually lead to failure of the complete part. Since the initiation of cracks is stress-driven, the location of crack initiation will always be the location with the highest stress. For that reason, the fatigue mechanism is extremely sensitive for stress concentrators like impurities, inclusions, pores and other irregularities. The presence of such material faults drastically reduces the service life of the complete part.

The consequence of this sensitivity for local irregularities is that material test results obtained from small test specimens cannot directly be applied to much larger components. The probability that in the test section of a small specimen a pore or inclusion is present is quite small, which means that the measured fatigue behaviour is representative for the pure material. However, if the same material is applied to a much larger component, for example, a large diameter shaft, the probability that an irregularity is present somewhere in the shaft is much larger, so the predicted service life must be reduced by the size factor. An example of the relation between size factor (k_2) and shaft diameter is shown in Fig. 4.13.

4.4.4.3 Surface Quality

As much as the initiation of fatigue cracks is sensitive for internal material defects, it is sensitive for surface defects. Also surface defects, for example, grooves, pits or other notches, act as stress concentrators and therefore enhance the crack initiation process. This means that the fatigue service life depends on the quality of the surface finish. This effect can also be quantified using a correction factor, as is shown for various surface roughness values R_a in Fig. 4.14.

4.4.4.4 Internal Stress

Since the total effective stress level governs the fatigue mechanism, internal stresses present in a part may significantly affect the part's service life. If a part already contains a compressive internal stress, an applied tensile stress must be

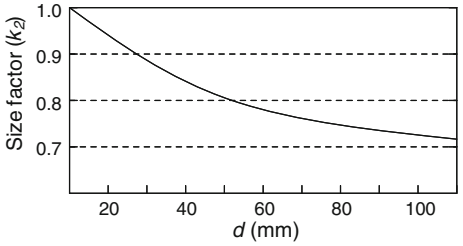

Fig. 4.13 Size effect as function of shaft diameter (d) for a bending load

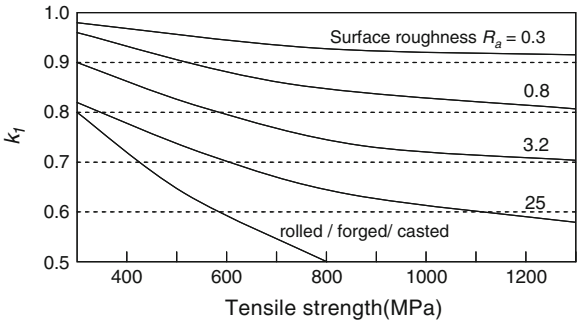

Fig. 4.14 Surface quality coefficient as function of surface roughness (R_a)

much higher before a critical stress range is reached. This effect is utilized in some applications, for example, metal springs, by introducing compressive stresses in the material on purpose. This is achieved by, for example, the shot-peening process, where small hard particles are shot onto the metal surface with high speeds. The resulting plastic deformation of the surface layer yields a compressive stress state in the part, which may increase the fatigue service life by 10–20 %.

4.4.4.5 Safety Factors

Finally, when all previously discussed factors are taken into account and a rather reliable prediction of the fatigue service life can be calculated, it may still be necessary to apply an additional safety factor to cover all remaining uncertainties. Especially in critical applications, like aerospace or nuclear power generation, the additional safety factors can be quite large to lower the risk of failure.

4.4.5 Life Assessment for Variable Amplitude Loads

For variable amplitude loads, where the magnitude of the applied stress range varies in time, the life cannot directly be obtained from the diagrams discussed in 4.4.4. In those cases, the cumulative damage rule proposed by Palmgren [5] and Miner [6] must be applied. This rule defines the fatigue damage D after n cycles at a constant stress amplitude σ_a, assuming that failure at this σ_a occurs after N cycles, as

$$D = \frac{n}{N} \tag{4.9}$$

The fatigue damage parameter D thus represents the fraction of the total life that already has been consumed. For example, 100 cycles at a stress range that yields failure after 1,000 cycles provide a damage number $D = 0.1$, which means that 10 % of the service life has been consumed after these 100 cycles.

For a variable amplitude load consisting of p blocks of n_i $(1 < i < p)$ constant amplitude cycles with different stress amplitudes $\sigma_{a,i}$ and associated number of cycles to failure N_i, the cumulative fatigue damage will be equal to

$$D = \sum_{i=1}^{p} D_i = \sum_{i=1}^{p} \frac{n_i}{N_i} \tag{4.10}$$

Reaching the value $D = 1$ means that the total life has been consumed and failure will occur. Therefore, if some load sequence with duration T has generated an amount of damage D_T in a component, then the total service life L of the component is obtained as

$$L = \frac{1}{D_T} T \tag{4.11}$$

It should be noted that this linear damage rule assumes that the fatigue life is independent of the order in which the different (blocks of) loads are applied. However, experiments have shown that crack propagation does depend on the time sequence of the loads. These so-called sequence effects will be discussed in Sect. 4.4.6 on fracture mechanics.

4.4.5.1 Spectrum Loading

Whereas the presented damage rule can be applied rather easily for a limited number of blocks with constant amplitude loads, real load sequences generally contain large numbers of different stress levels and only very small numbers of constant amplitude cycles per block. To make application of the damage rule still feasible, methods have been developed to reduce the amount of data [7]. Two ways to reduce the load sequences are as follows:

- remove cycles smaller than a certain threshold value, that do not contribute significantly to the fatigue damage accumulation
- represent the sequence of minima and maxima by a statistical distribution of loads

4.4.5.2 Counting Methods

The reduction methods developed for this task are called ‘counting methods’. The first step for all these methods is the division of the load axis in a discrete number of intervals, see Fig. 4.15. It is then a simple procedure to count the number of maximum and minimum loads within each interval, which then can be plotted as a histogram (Fig. 4.16a).

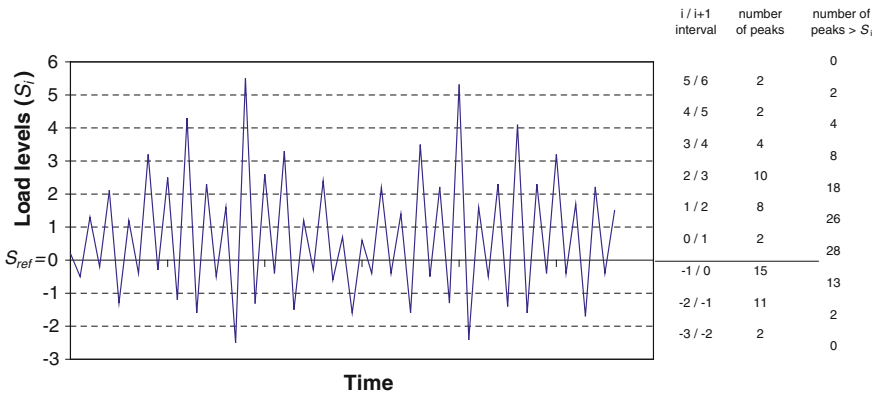

Fig. 4.15 Example of a load sequence, the associated number of peaks and number of level exceedances

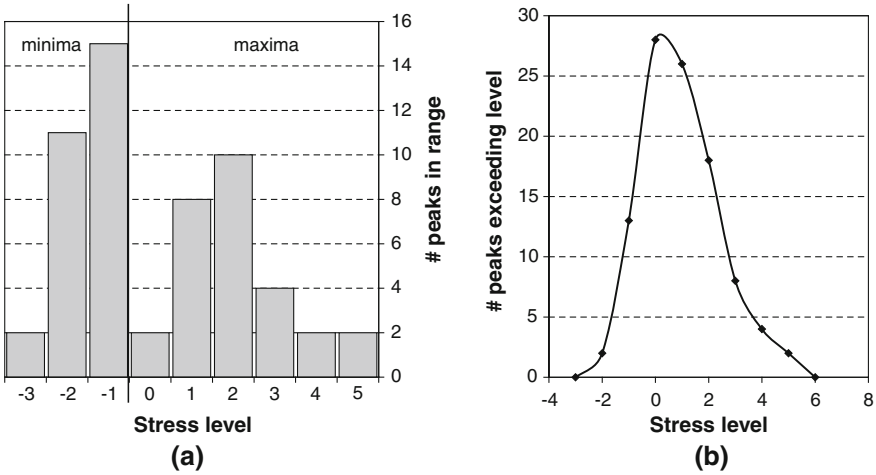

Fig. 4.16 Two ways to characterize the load sequence in Fig. 4.15: histogram and exceedance diagram

Another way to represent the history is to count the number of maximum (minimum) values above (below) a certain level i, which yields a so-called exceedance diagram (Fig. 4.16b).

Instead of counting the number of peaks, for fatigue, it may be more useful to determine how many times load variations from one minimum value to the subsequent maximum occurred. This can be achieved by counting the number of load ranges ($\Delta S = S_{\max} - S_{\min}$) and plotting them in a histogram or exceedance diagram. Note that for all these representations, information about the sequence in which the peaks occurred is lost.

The discussed counting results are all one-dimensional, that is, each load occurrence is characterized by a single value (S or ΔS). However, a load variation

Table 4.1 Matrix representation of a load sequence

		S_{max} in interval					
		0 → 1	1 → 2	2 → 3	3 → 4	4 → 5	5 → 6
S_{min} in interval	−1 → 0	1	5	6	3		
	−2 → −1	1	2	4	1	2	1
	−3 → −2		1				1

(which quantifies the fatigue damage) is characterized by two values: S_{max} and S_{min} or ΔS and S_{mean}. In a two-dimensional count system, this additional information can be stored. The result is a matrix as shown in Table 4.1, where each number indicates how many times a load variation occurred between the corresponding S_{max} and S_{min} values. Also in this representation, the sequence of the variations is lost.

In the methods discussed until now, all minima and maxima are counted and consequentially all variations are taken into account, including the small variations. It is debatable whether these small variations must be considered, since they play a minor role in the fatigue damage process but for large sequences are a considerable burden for the analysis process. In the remainder of this section, this topic will be discussed by treating several specific counting methods that address this problem.

4.4.5.3 Mean Crossing Peak Count Method

One way to filter out the small load variations is the use of the mean crossing peak count method. In this method, only one peak value (the most extreme one) is counted between two successive crossings of the mean value. This is illustrated in Fig. 4.17 for a small sample of a load sequence. In traditional methods, all peaks (1–14) are counted, as indicated in Fig. 4.17a. However, in the mean crossing peak count method, only peaks 1, 2, 5, 8, 11 and 14 are counted, which means that the small variations, for example, 3–4, are omitted.

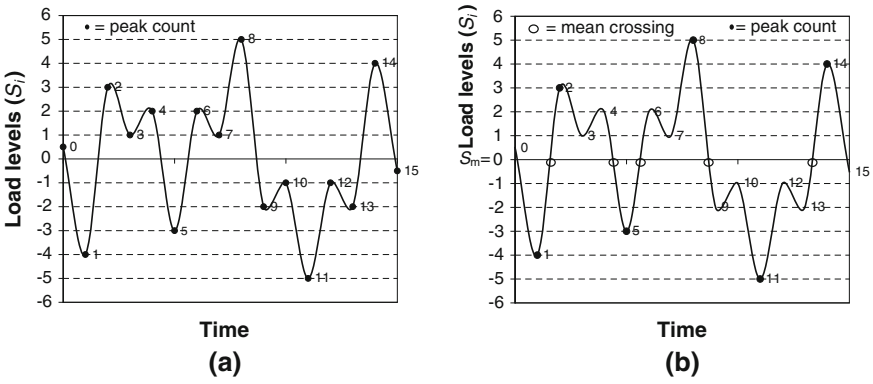

Fig. 4.17 Illustration of the mean crossing peak count method

4.4.5.4 Level Crossing Count Methods

The principle of the level crossing count method is shown in Fig. 4.18. Each time the load passes a certain positive level ($S > S_i$) in the upward direction a count is made. This eliminates the majority of the small variations, but a variation like 6–7 in Fig. 4.18a is not removed. This can be achieved by extending the method with an additional requirement: a crossing of level i is only counted if the load has returned to a significantly lower level i', see Fig. 4.18b.

4.4.5.5 Range Pair Count Methods (Rain Flow Count)

The final method discussed here is the range pair count method, which is also called rain flow or pagoda roof counting. Figure 4.19 shows two methods to analyse a load variation ABCD. In method *I,* three successive load ranges are counted separately. However, if the small interruption BC would be absent, the load range (AD) would be much larger. This is taken into account by the range pair method (method *II*).

Application of the range pair method is illustrated in Fig. 4.20. Consider four successive peak values, such as ABCD in Fig. 4.19. A count is made only when peaks B and C are in between peaks A and D. The load variation BCB' is then counted (stored) as a range pair (BC and CB'), and before the analysis proceeds, B and C are removed from the series of peak values. This procedure is illustrated in Fig. 4.20: in the first step, four range pairs are detected, stored and removed, and in the second step, another range pair is counted. The remaining sequence is called the residue, which can be counted with a traditional counting method. Using this method, part of the sequence information is contained and, moreover, the largest variations in the sequence are always obtained, which provides a conservative approximation of the load sequence.

The name rain flow counting refers to the pictures in Fig. 4.21, which illustrate the counting method in another way. The procedure is as follows:

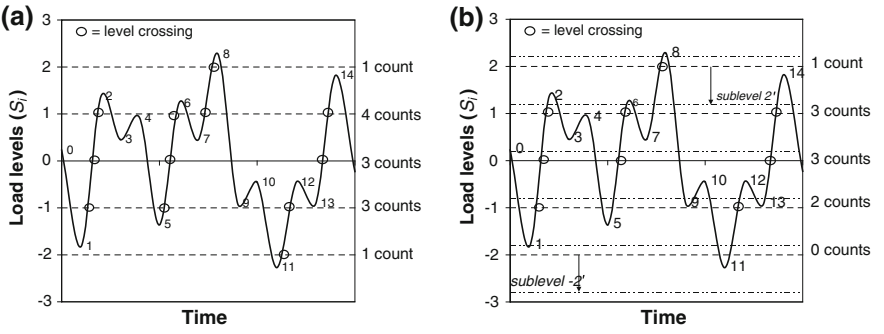

Fig. 4.18 Illustration of the level crossing count method

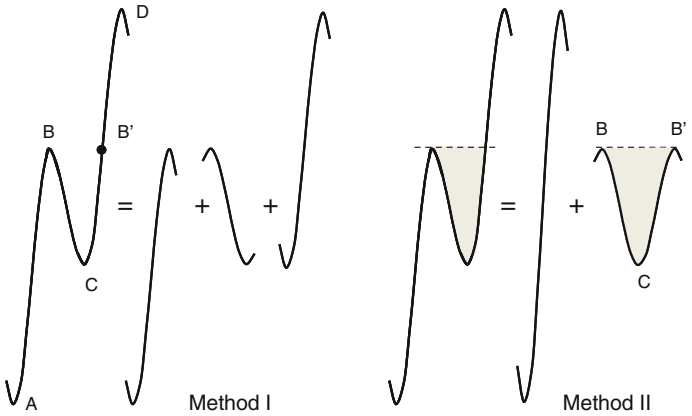

Fig. 4.19 Different ways of analysing load sequences

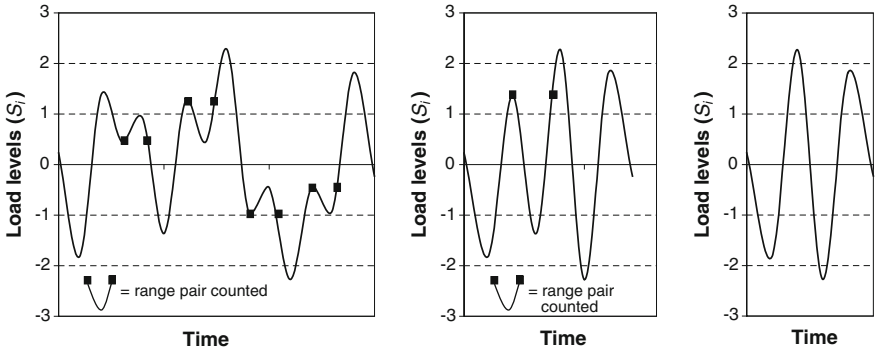

Fig. 4.20 Application of the range pair counting method to a load sequence

- start with analysing the tensile peaks (positive values) and consider each peak as a source of water that drips down.
- determine the number (and magnitude) of half cycles by looking for terminations of the flow, occurring when either:
  - it reaches the end of the time history (e.g. half cycle C)
  - it merges with a flow started at an earlier *tensile peak* (e.g. half cycle B)
  - it flows opposite a *tensile peak* of greater magnitude (e.g. half cycle A)
- repeat this procedure for the compressive peaks
- pair up half cycles of identical magnitude (but opposite sense) to count the number of complete cycles. Typically, there are some residual half cycles

The result for the sequence in Fig. 4.21 is given in Table 4.2. Note that a lot of algorithms have been developed to obtain the range pairs from a load sequence in only a single run.

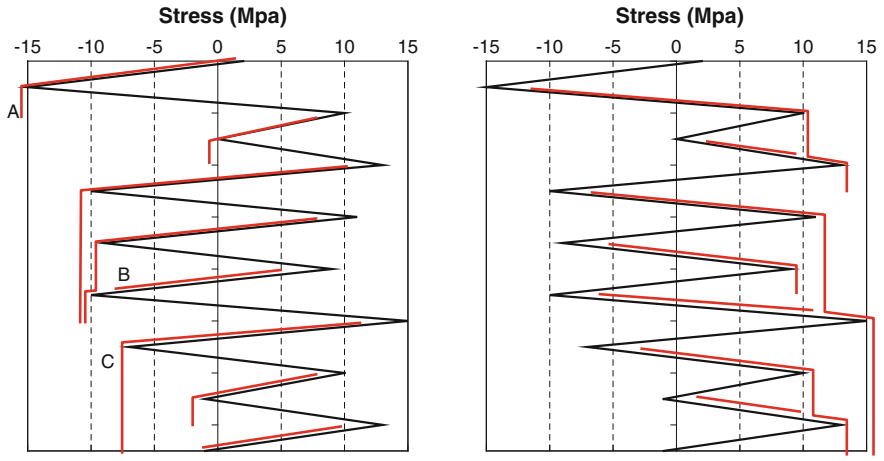

Fig. 4.21 Illustration of the range pair count/rain flow method

Table 4.2 Result of range pair count of the sequence in Fig. 4.21

Stress (MPa)	Whole cycles	Half cycles
10	1	0
11	1	0
14	0	1
17	0	1
18	0	1
19	0	1
20	0	1
21	0	1
22	1	0
23	0	1
25	0	1
28	0	1

4.4.6 Fracture Mechanics

In the previous subsections, the fatigue service life calculations were based on the continuum damage approach, that is, the evolution of the damage parameter D is calculated without considering the details of the damage. In the field of fracture mechanics, another approach is followed, which is based on a detailed analysis of the behaviour of cracks.

The initiation and propagation of cracks is a problem that occurs quite often in practice. Although most structures are designed such that no cracks occur during the service life, in some applications the initiation of cracks is inevitable, for example, due to high cyclic loads, incidental damage, material defects or

inappropriate welding. Well-known examples are aircraft structures, which are designed under strict weight limitations, but are also subjected to high cyclic loads. If there is a potential for crack initiation in a structure, it is important to know where the cracks may occur, what the maximum allowable crack length is and how fast the cracks grow. Based on this information, appropriate inspection schemes can be developed to monitor the cracks during the service life. The principles of fracture mechanics (see for example [8]) are used to calculate critical crack lengths and crack propagation rates. This section treats the most important aspects of that field.

4.4.6.1 Stress Intensity Factor

Cracks commonly initiate at irregularities in the material or near stress concentrations. Examples of material irregularities are impurities and porosity, whereas stress concentrations develop near holes or shape transitions in the geometry of the material. Once a crack has initiated, it can be opened in three different ways by the applied load (Fig. 4.22):

- mode I: opening in tension
- mode II: in-plane shearing mode
- mode III: out-of-plane shearing or tearing mode

In this section, only the most common mode I loading will be discussed. Relations for the other modes can be found in [7, 8].

The growth rate of cracks depends on the local stresses near the crack tip. These stresses can be quantified by the stress intensity factor (SIF), denoted by the symbol K. This factor actually defines the severity of the stress field, which can be related directly to the crack growth rate. The SIF concept constitutes the core of the fracture mechanics theory and will be elaborated next.

Using the theory of elasticity, the stress distribution around an elliptical hole in an infinite plate loaded in tension can be determined. To simulate a real crack, the minor axis of the ellipse is reduced to zero, as is shown in Fig. 4.23.

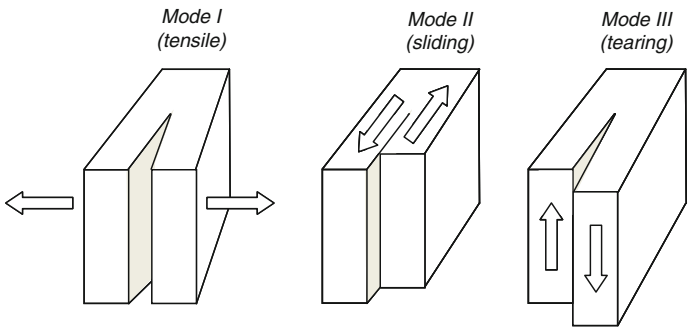

Fig. 4.22 Three different crack opening modes

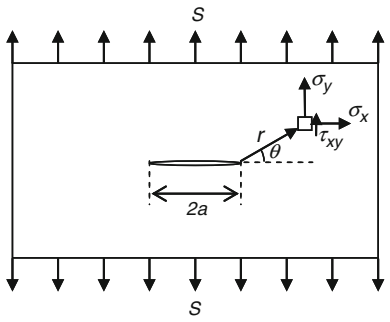

Fig. 4.23 Stress distribution in an infinite plate containing a central crack

The stress distribution is then given by the three stress components:

$$\sigma_x = \frac{S\sqrt{\pi a}}{\sqrt{2\pi r}}\cos\frac{\theta}{2}\left(1 - \sin\frac{\theta}{2}\sin\frac{3\theta}{2}\right) - S \tag{4.12}$$

$$\sigma_y = \frac{S\sqrt{\pi a}}{\sqrt{2\pi r}}\cos\frac{\theta}{2}\left(1 + \sin\frac{\theta}{2}\sin\frac{3\theta}{2}\right) \tag{4.13}$$

$$\tau_{xy} = \frac{S\sqrt{\pi a}}{\sqrt{2\pi r}}\cos\frac{\theta}{2}\left(\sin\frac{\theta}{2}\cos\frac{3\theta}{2}\right) \tag{4.14}$$

where S is the remotely applied stress, a the half-crack length, and r and θ the distance and direction to the crack tip. This shows that the stress level is governed by the SIF

$$K = S\sqrt{\pi a} \tag{4.15}$$

and the coordinates r and θ. Further, it can be observed that the severity of the local stress (the factor K) is determined by both the nominal load S and the crack length a. Note that the unit of the SIF K is MPa$\sqrt{\text{m}}$. Equation (4.15) is valid for a tensile-loaded infinite plate with a central crack. For other situations, like bending, finite plate width or different geometries and edge cracks, the value for K can be obtained from handbooks [9, 10] or be calculated using FE methods. In general,

$$K = FS\sqrt{\pi a} \tag{4.16}$$

where the geometry factor F is obtained from literature or calculations.

For some standard problems, analytical expressions have been derived for F, as is the case for an infinite sheet with an infinite row of collinear cracks (see Fig. 4.24)

$$F = \sqrt{\frac{\tan(\pi a/W)}{\pi a/W}} \tag{4.17}$$

In this case, the factor F depends on the dimensionless ratio of the crack length (a) and the width (W) of a section. A plot of F vs. a/W is shown in Fig. 4.24. For other problems, no analytical solutions are available, but numerical results presented in plots or tables can be obtained from handbooks.

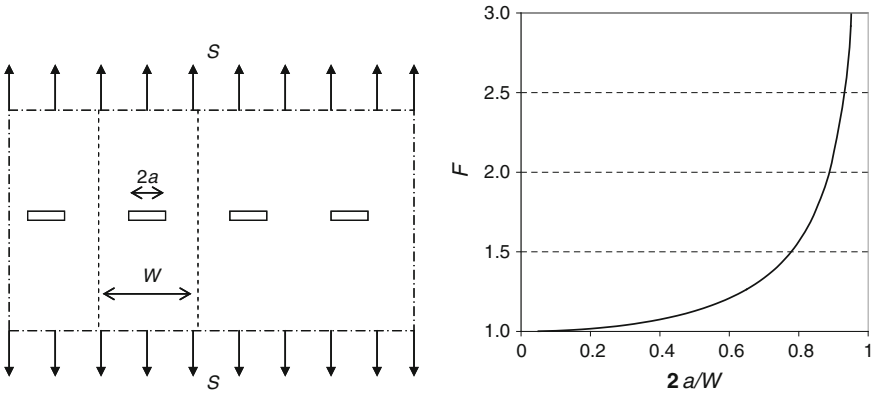

Fig. 4.24 Variation of geometry factor F for an infinite sheet with an infinite row of collinear cracks of length $2a$

For complex real structures, appropriate handbook solutions are not available and numerical analyses (e.g. FE) must be performed to obtain the SIF solution. An example of a FE model used for such a calculation on an aircraft structural part is shown in Fig. 4.25.

4.4.6.2 Crack Tip Plasticity and Crack Closure

For the derivation of Eqs. (4.12–4.14), the theory of elasticity was used, implying linear behaviour of the material. The equations show that at the crack tip ($r \rightarrow 0$), the stresses would become infinite. In reality, a plastic zone will develop, where the linear behaviour is no longer valid. For a relatively small plastic zone ($r_p \ll a$), Eqs. (4.12–4.14) are a good approximation. The magnitude of r_p can be estimated by substitution of $\theta = 0$, $\sigma_y = \sigma_{0.2}$ and $r = r_p$, in (4.13), giving

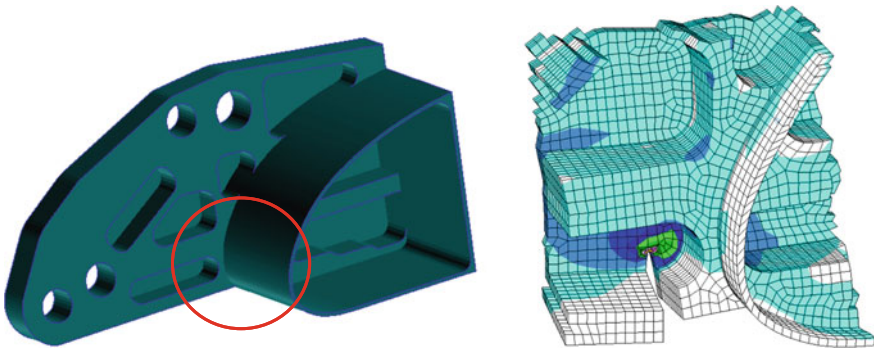

Fig. 4.25 Finite element model of a complex component used to determine the SIF solution (published with kind permission of © NLR 2012. All Rights Reserved)

$$r_p = \frac{1}{\pi}\left(\frac{K}{\sigma_{0.2}}\right)^2 \tag{4.18}$$

For values $S = 100$ N/mm^2, $\sigma_{0.2} = 400$ N/mm^2 (aluminium) and $a = 10$ mm, the plastic zone radius r_p equals 0.6 mm. Since this plastic zone also affects the crack growth rate, r_p is sometimes used as an addition to the crack length used in the SIF

$$K = FS\sqrt{\pi(a + r_p)} \tag{4.19}$$

Moreover, the size of the plastic zone depends on the applied stress cycle. If during a load sequence a high load occurs, a large plastic zone will develop. But a series of smaller load cycles after the peak load will have much smaller plastic zones, which means that the crack has to grow through the large zone of the peak load. Generally, this reduces the crack growth rate, which is known as 'crack retardation'. Wheeler has modelled this effect by defining a factor Φ:

- if r_p for the current cycle $< r_{pp}$ for the peak load: $\Phi = \left(\frac{r_p}{r_{pp}}\right)^m$
- if r_p for the current cycle $> r_{pp}$ for the peak load: $\Phi = 1$

Then, this factor is used to modify the constant amplitude (ca) crack growth rate

$$\left(\frac{da}{dN}\right)_i = \Phi_i \left(\frac{da}{dN}\right)_{ca} \tag{4.20}$$

Since plastic deformation occurs during the loading part of a cycle, after unloading the crack faces do not exactly match anymore. Therefore, also during unloading, plastic deformation (in compression) will occur, but less than during loading. As a result, a certain amount of residual plasticity will reside in the wake of the growing crack, leading to a compressive stress state in unloaded condition. These compressive stresses close the crack and only after a certain amount of tensile loading it will open again. This is illustrated in Fig. 4.26, where an opening

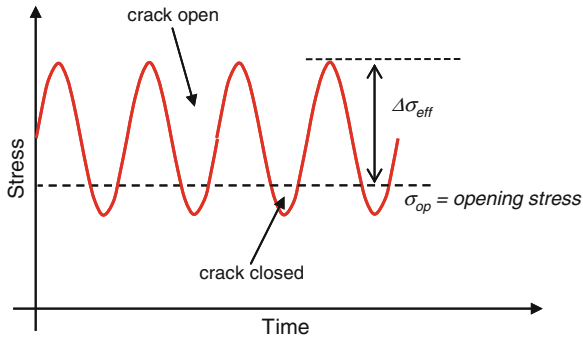

Fig. 4.26 Effect of crack closure on effective stress range

stress (σ_{op}) is defined as the stress level at which the crack starts to open (when the residual stress is overcome).

It is clear that the effective stress range ($\Delta\sigma_{\text{eff}}$) is now smaller than the nominal stress range, which means that the crack growth rate will also be lower. This effect can be included in the relations by defining the opening factor f:

$$f = \frac{K_{op}}{K_{\max}} \tag{4.21}$$

Finally, the effective SIF range ΔK_{eff} is used in the crack growth laws to incorporate this effect

$$\Delta K_{\text{eff}} = F\Delta\sigma_{\text{eff}}\sqrt{\pi a} \tag{4.22}$$

4.4.6.3 Crack Opening

When structures are inspected, a minimal crack opening is required to make cracks detectable. The crack opening v at the centre of the crack can be derived to equal

$$v = 2\frac{S}{E}a \tag{4.23}$$

with E the material elastic modulus. For a stress $S = 100$ N/mm^2, yield stress $\sigma_{0.2} = 70{,}000$ N/mm^2 and $a = 10$ mm, the crack opening is $v = 0.06$ mm. This illustrates that the visibility of cracks largely depends on the question whether the crack is loaded or not. Moreover, the crack opening is directly related to the crack length.

4.4.6.4 Crack Propagation

The crack initiation is typically predicted using ε–N curves or S–N curves, as discussed in one of the previous sections. Once a crack has initiated, it will start to grow as soon as the load exceeds a certain threshold stress (quantified by ΔK_{th}). The crack propagation rate depends on the magnitude of the load and on the capacity of the material to withstand the load, the crack growth resistance. The latter is a material property that can be determined by performing a crack growth test. A typical example of such a test result is shown in Fig. 4.27, where the crack propagation rate, defined as the increase in the crack length a per load cycle N (da/dN), is plotted versus the SIF range $\Delta K = K_{\max} - K_{\min}$.

The crack propagation curve contains three regions: the first region is the slow crack growth region, in which the rate initially is high, but is reduced quite soon. The second part covers the largest fraction of the total crack growth process, and the rate is described by the well-known Paris law

$$\frac{da}{dN} = C(\Delta K)^n \tag{4.24}$$

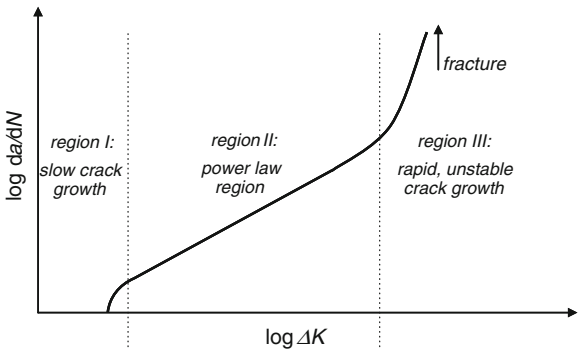

Fig. 4.27 Typical *crack growth curve* showing the three different regimes

On a double logarithmic scale, this equation provides a linear relation between crack growth rate and SIF. Finally, the third region of rapid unstable crack growth is reached, where the rate increases significantly and ultimately final failure is observed.

More sophisticated crack growth laws have been developed, which extend the classical Paris law. An example is the following relation

$$\frac{da}{dN} = C\frac{(1-f)^n \Delta K^n \left(1-\frac{\Delta K_{th}}{\Delta K}\right)^p}{(1-R)^n \Delta K^n \left(1-\frac{\Delta K}{(1-R)K_c}\right)^q} \tag{4.25}$$

which describes the complete curve shown in Fig. 4.27. In this relation, the effect of the threshold stress (ΔK_{th}), stress ratio (R), fracture toughness (K_c) and crack closure (f) is included.

4.4.6.5 Failure Criteria

A crack propagating in a structure may lead to failure of the structure in several ways:

- **overload failure**: the crack front has grown to an extent where the remaining cross section of the structure is unable to carry the load. This is quantified by the ultimate tensile strength (σ_{uts}) of the material.
- **functional impair**: the crack leads to effects that impair the proper functioning of the structure, for example, a crack in a fuel tank leads to leakage of the fuel.
- **fracture toughness**: the SIF K exceeds the fracture toughness K_{Ic} which leads to static failure of the structure.

The final failure mode shows that the fracture toughness (K_{Ic}) of a material can be seen as the critical value of K, which leads to static failure by a mode I crack. K_{Ic} is therefore a material property that indicates the sensitivity of the material for cracks under static loading. Knowing the value of K_{Ic} for a material allows the calculation of a critical crack length (a_{cr}) for a certain load S, using Eq. (4.16):

$$a_{\mathrm{cr}} = \frac{1}{\pi}\left(\frac{K_{\mathrm{Ic}}}{FS}\right)^2 \tag{4.26}$$

Note that for thin sheets, the failure may not be a mode I failure, but shearing will also play a role. This means that a mixed mode I/III failure occurs, for which the critical SIF (K_c) is in general higher than the fracture toughness K_{Ic}. This means that thin sheets in general are more crack tolerant than thick sections.

4.4.6.6 Performing a Crack Growth Analysis

A lot of commercial computer codes have been developed to perform crack growth analyses, both standalone (e.g. [11, 12]) and integrated in FE codes (e.g. [13]). The benefit of these commercial codes is mainly the availability of databases with material properties (crack growth curves, fracture toughness) and SIF solutions (the geometry factor F for various geometries, dimensions and crack lengths). However, the basic algorithm in all these codes is rather simple, as will be described next.

A crack growth analysis consists of the following steps:

1. Preparation:
   a. gather information about crack growth properties for the present material
   b. select the appropriate SIF solution
   c. define the initial crack length a_0
   d. define the failure criteria (critical crack length, K_{Ic} or tensile strength)
   e. define the load sequence in terms of load blocks (N cycles at load range $\Delta\sigma$)

   Actual calculation:

2. calculate the value of K for the present load and present value of a
3. calculate da/dN for that K using an appropriate crack growth law (e.g. Paris law)
4. multiply da/dN with the number of cycles in the load block to obtain da
5. update the crack length: $a = a + \mathrm{d}a$
6. repeat steps 2–5 until one of the failure criteria is met:
   a. critical crack length a_{crit} is reached *or*
   b. K equals K_{Ic} *or*
   c. σ in remaining cross section exceeds σ_{uts}

Using this algorithm, the propagation of a crack in a certain structure for a given load sequence can be calculated and the associated life time can be obtained. This information is generally used to determine the service life of a system and/or the inspection scheme (intervals, inspection methods).

4.5 Creep

Creep is a phenomenon that causes inelastic deformation in a material at elevated temperature, although the applied stress level is still in the elastic regime, that is, below the material yield stress. It is a time-dependent process, where the creep strain rate depends on both the temperature and the magnitude of the applied stress. As the name suggests, the process is rather slow, especially when the temperature level is moderate. This also means that the experimental determination of a material's creep behaviour is quite expensive, since a typical creep test at one combination of stress and temperature lasts for one half to a full year. During the test, the elevated temperature must be maintained by a furnace or inductance heating.

Figure 4.28 shows a schematic representation of the creep strain evolution during a creep test at constant temperature and stress. For most materials, three regions can be identified in the creep curve: the primary creep region shows a relatively high creep rate at the start of the deformation process. It rather quickly transitions to the secondary creep region, where the creep strain rate is constant for a prolonged period of time. Finally, in the tertiary regime shortly before final rupture, the creep rate increases again.

4.5.1 Creep Mechanism

Also a creep failure can be recognized during failure analysis by specific features on the fracture surface. The micrograph in Fig. 4.29 shows a faceted fracture surface, which is typical for creep and other high-temperature failures. The origin of the facets is the grain boundaries in the material, which are the interfaces between the different crystal grains.

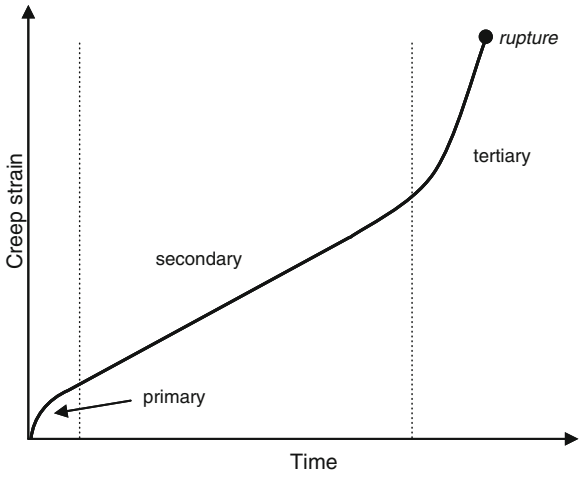

Fig. 4.28 Typical creep curve at a certain stress and temperature level, showing the three different regimes

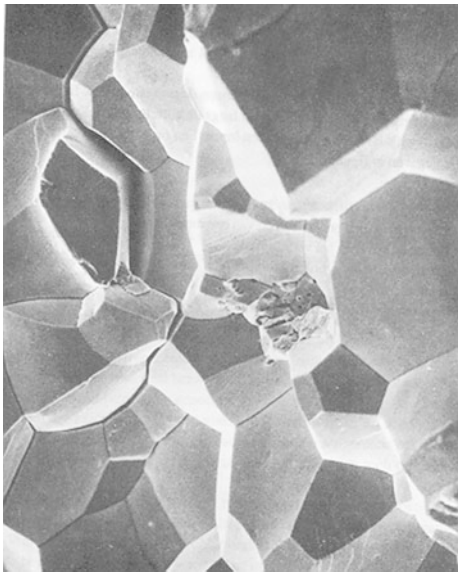

Fig. 4.29 Typical fracture surface after creep failure (published with kind permission of © NLR 2012. All Rights Reserved)

At high temperatures, these grain boundaries are the weakest regions in a material, resulting in the initiation of creep damage at the grain boundaries by the formation of voids. The voids extend along the grain boundaries and ultimately intergranular failure occurs when the material fractures across the voids. At the fracture surface, the interfaces of the grains are clearly visible then, as can be seen in Fig. 4.29.

4.5.2 Life Assessment

The creep life time is generally defined as the time to rupture (fracture), but can also be related to a certain maximum allowable amount of deformation. In the latter case, excessive elongation may cause rubbing of a (rotating) part to a counter surface, leading to functional failure of the system.

Since the secondary region of the creep curve covers the majority of the time, the creep strain rate in that region mainly governs the life time of a part. The creep strain rate $\dot{\varepsilon}_{cr}$ in the secondary regime is generally expressed as a power law of stress (σ) and temperature (T), which is called the Norton creep law

$$\dot{\varepsilon}_{cr} = AT^n\sigma^m \tag{4.27}$$

where A, n and m are material constants that can be obtained from experiments. Most materials are extremely sensitive for changes in temperature, much more than for stress level changes, that is, $m < n$. For typical gas turbine blade materials, a temperature increase of only 25° (e.g. from 900 to 925 °C) yields a reduction in the creep life by a factor three.

The Norton creep law in (4.27) is just a phenomenological description of the creep behaviour in the secondary regime of the creep curve. For many applications, this is an approximation that is sufficiently accurate to perform life assessments on parts subjected to creep loads. However, for critical applications like gas turbine blades, more accurate creep models are required. For these materials, micromechanical material models have been developed that are based on the underlying mechanisms at the microstructural level, like the motion of dislocations and the effects of precipitates in the material [14, 15]. The models are able to predict the creep behaviour not only in the secondary regime, but also in the primary and tertiary regime.

The rupture life of a component can be obtained from a creep rupture curve, showing for a certain temperature the time to failure for various applied stress levels. Generally, the curves for several temperatures are provided in one plot. It is also possible to combine all combinations of stress and temperature in one curve by using the Larson-Miller parameter [16]. As time to rupture t and temperature T are combined into one parameter P_{LM} according to

$$P_{\mathrm{LM}} = T(C + \log t) \tag{4.28}$$

a plot of P_{LM} versus stress represents the material creep behaviour over a range of stress and temperature levels, as is shown in Fig. 4.30. Every point on this curve thus represents an infinite number of life time—temperature combinations. The constant C in the parameter depends on the material that is considered.

The method just described can be utilized to calculate the creep life at a combination of stress and temperature that is constant in time. But just as with variable amplitude fatigue loads, a variable stress and/or temperature requires application of a cumulative damage rule to calculate the life time. Analogous to the Palmgren–Miner rule for fatigue, a damage rule has been proposed for creep, which is called the Robinson rule [17]. This rule states that at a variable load consisting of p time periods with different stresses and/or temperatures, the cumulative creep damage equals

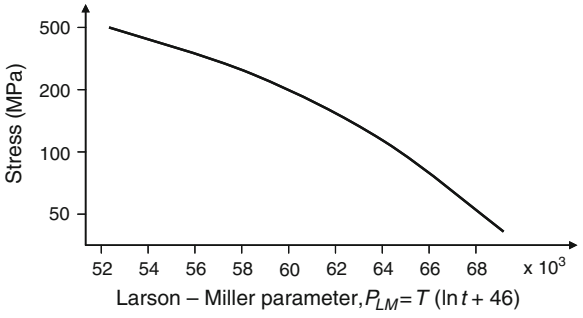

Fig. 4.30 Larson-Miller plot of creep behaviour

$$D = \sum_{i=1}^{p} D_i = \sum_{i=1}^{p} \frac{\Delta t_i}{t_{r,i}} \tag{4.29}$$

where $t_{r,i}$ represents the creep rupture time at the stress/temperature combination in period i with duration Δt_i. When the total amount of damage reaches unity, failure will occur.

4.6 Wear

Wear is the general term for the failure mechanisms associated with the relative motion of two or more parts that are in physical contact, where the movement yields a loss of material at the surface of at least one of the bodies. Wear can also be caused by a medium (gas, fluid) flowing along a part. The mechanism thus yields a progressive degradation which ultimately leads to failure of the part. The failure is triggered by either dimensional anomalies (e.g. unacceptable clearance, bad fit), functional failure (e.g. excessive friction or vibrations) or the initiation of other failure mechanisms (e.g. static overload).

In non-technical reports and papers on failure and maintenance, the term wear is often used to indicate all physical mechanisms leading to the degradation of parts or systems that ultimately yield failure. Also mechanisms like creep and fatigue are thus contained in the term, often called 'wear and tear'. In this book, the term wear is more purely used to indicate the physical mechanism of wear, as defined in the first lines of this subsection.

The topic of wear is closely related to the concept of friction, for which the basics have already been discussed in Sect. 2.3.5. (internal heat generation). These two phenomena interact since most types of wear are only possible if there is a certain amount of friction, while the other way around the wear process considerably affects the friction in the contact area. Friction and wear therefore constitute the two main topics of the scientific field of tribology [18, 19]. The third topic in this field is lubrication, which is applied to reduce both friction and wear.

Although the wear mechanism is confined to the degradation of parts that are in contact and are moving relative to each other, still several types of wear mechanisms can be identified, as will be discussed in separate subsections. Before the treatment of the wear mechanisms, the concepts of friction and lubrication will be elaborated. This section will be finished by discussing the life assessment method for the wear mechanism.

4.6.1 Friction

Friction is a force that occurs at the interfaces in a contact between two or more bodies. It can be both beneficial (e.g. brakes, traction drive mechanisms) and undesirable (e.g. bearings, contact between piston and cylinder). As was discussed

in Sect. 2.3.5, the friction force, that is the force opposing the motion of one body relative to another body, depends on the normal force F_n on the sliding body and the friction coefficient μ

$$F_f = \mu F_n \tag{4.30}$$

Two important factors determining the magnitude of the friction coefficient in a contact are the surface roughness and the relative material hardness.

The surface roughness plays an important role since it determines the real contact area between the two bodies. Unless the surfaces are very flat and polished, for hard materials, like most metals, the real contact area is generally considerably smaller than the nominal contact area. Only the peaks of the surface roughness profile are in real contact, while the intermediate valleys are not contributing to the friction. But even for softer materials, like plastics and elastomers, the real contact area is only about 50 % of the nominal contact area. This difference between actual and nominal contact area has three important consequences: (1) the applied normal force must be carried by only a fraction of the nominal surface area, which means that local stress levels may be considerably higher than expected. As a result, roughness peaks may deform plastically, generally resulting in an increased surface area; (2) the heat generation due to friction is caused solely at the roughness peaks. As was discussed in Sect. 2.3.5, this may considerably localize the heat flows and lead to local failure of the lubricant or even melting of the roughness peaks; (3) wear of the material will take place primarily at these roughness peaks, which means that a certain volume loss at the materials interface may be associated with a much larger thickness decrease than anticipated. In addition to the real/nominal contact area effects, a high surface roughness may also lead to asperity interlocking. In that case, motion cannot take place between the two bodies without deformation of the asperities, which leads to a high friction coefficient.

The second important factor affecting the friction coefficient is the material hardness. It determines, together with the magnitude of the applied load, which of two basic mechanisms will mainly determine the friction coefficient for relatively smooth surfaces: ploughing or adhesion. Note that for rough surfaces, the above mentioned asperity locking mechanism largely determines the friction. If the hardness of the two materials in contact differs more than 20 %, the roughness peaks of the harder material penetrate into the softer material and cause severe deformation. This mechanism is called ploughing and depending on the magnitude of the applied load, either elastic or plastic deformation of the softer material results. As this mechanism is associated with relatively large penetration depths (δ), the resulting friction is considerable. This can be visualized with a spherical indenter with radius R that is sliding along a counter surface, see Fig. 4.31.

The applied normal force F_n is balanced by the contact pressure p in the contact area A_r. The frictional force $F_f = \mu F_n$ originates from the stress σ at the frontal area A_f. Therefore, the friction coefficient μ can be deduced to be [19]

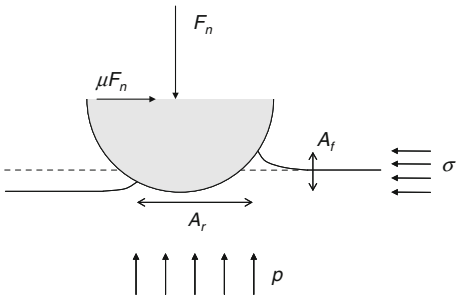

Fig. 4.31 Sliding spherical indenter representing ploughing friction

$$\mu = \frac{\sigma}{p}\frac{A_f}{A_r} \tag{4.31}$$

In many practical situations, plastic deformation will occur during ploughing. In that case, both contact pressure p and the stress σ at the frontal surface will be close to the material yield stress, that is, $\sigma/p \sim 1$. This means that the friction coefficient solely depends on the ration between the two areas. Finally, the ratio A_f/A_r, and thus the friction coefficient μ, can be shown to be a function of the relative penetration depth δ/R.

Adhesive forces are the main contributor to friction in contacts where no penetration (ploughing) of the mating surface occurs. This happens when the two materials have similar hardness or when the surfaces are smoothly finished, that is, the surface roughness is very low. In that case, atomic or intermolecular forces cause attraction between the surfaces. When two identical materials are in contact, these forces are called cohesive forces, while for dissimilar materials adhesive forces are acting at the interface. Since cohesive forces are generally stronger than adhesive forces, the friction between two identical materials is higher than between two dissimilar materials.

For the latter, the magnitude of the adhesive forces depends on the metallurgical compatibility of the materials. This quantity is associated with the surface and interface energies of the two materials. When two materials a and b are brought into tight contact, the two free surfaces of the individual materials disappear, which releases the surface energies γ_a and γ_b. But a new interface must be formed, which requires an interface energy γ_{ab}. If two identical materials are pressed together, the interface energy will be zero. For all other material combinations, the interface energy will be larger than zero, and the metallurgical compatibility quantifies how favourable the formation of such an interface would be. Materials are metallurgically compatible when $\gamma_{ab} \sim \frac{1}{4}(\gamma_a + \gamma_b)$ and incompatible when $\gamma_{ab} \sim \frac{1}{2}(\gamma_a + \gamma_b)$. In the latter case, the interface energy is comparable to the average surface energy of the two materials and the formation of an interface does not provide a very large energetic benefit.

The adhesive (or cohesive) forces cause a shear stress τ at the interface that opposes the relative motion of the two bodies that are in contact. Therefore, the friction force F_f is determined by this shear stress and the real contact area A_r

$$F_f = \tau A_r \tag{4.32}$$

Adhesive forces in a contact may be so high that they cause material transfer from one to the other surface. This is the origin of adhesive wear, as will be discussed in Sect. 4.6.4.

4.6.2 Lubrication

Lubricants play an important role in reducing friction and wear and thereby increasing the lifetime of parts and systems. Lubrication in moving contacts can be provided by a large number of materials, ranging from mineral or synthetic oils and greases to solid lubricants like graphite and the polymer PTFE. Discussion of the characteristics and properties of all these lubricants is outside the scope of this book. However, the aspect of lubrication that is particularly important for the different wear mechanisms treated later in this section is the differentiation in lubrication regimes. In sliding or rolling contacts, which are typically present in all types of bearings, three regimes can exist [19]: boundary lubrication (BL), hydrodynamic lubrication (HL) and mixed lubrication (ML).

In any contact between two materials, the applied mechanical load must be transferred through the contact. If a lubricant is present in the contact, but the load is still predominantly transferred by mechanical contact between the two materials, the regime is called BL. The function of the lubricant is then to reduce the friction and wear.

In many rotating contacts, as in ball and journal bearings, the mechanical contact between the two materials can be completely prevented by the creation of a lubricant film in the contact. This requires the build-up of a hydrodynamic pressure in the contact that can balance the load applied to the bearing. This is only possible when the rotational speed is sufficiently high and the film layer exceeds a critical thickness ($h_{\min}$). However, in that case, the wear rate is reduced to practically zero and the friction coefficient decreases from a value in the order of 0.1 for BL to a value lower than 0.01.

The relation between the (rotational) speed (v) and the resulting friction coefficient (and lubricant film thickness h) for a specific contact/bearing can be visualized in a so-called Stribeck curve, as is shown in Fig. 4.32.

The curve clearly shows the three lubrication regimes: for low bearing speeds, the pressure build-up is insufficient to separate the different parts and BL is active. For high speeds, the critical film thickness is exceeded and HL is possible. At intermediate speeds, part of the load is carried by the film, but also a fraction is transferred by mechanical contact.

Finally, lubricants can play an important role in condition monitoring, as the presence of wear particles in lubricating oil may indicate the wear of specific parts in a system. This will be discussed in more detail in 6.5.1.

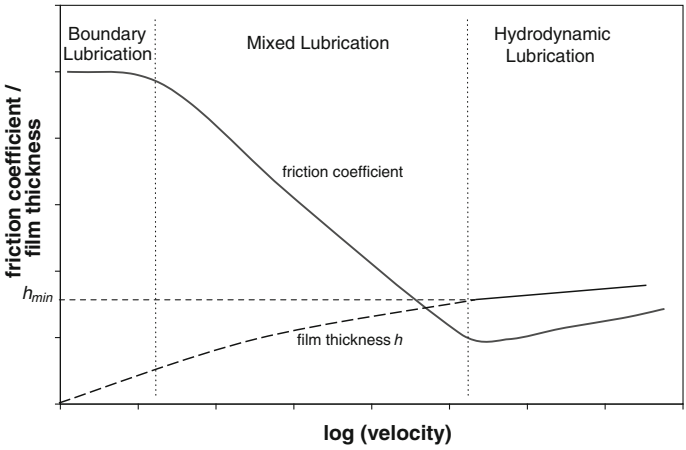

Fig. 4.32 Stribeck curve showing the relation between bearing speed and friction coefficient

4.6.3 Wear Mechanisms

As was defined in the introduction of this section, wear generally occurs when two parts are in contact and are moving relative to each other. This type of wear is called two-body wear and can be divided into four basic wear mechanisms: adhesive wear, abrasive wear, corrosive wear and surface fatigue (ASTM standard G40-10b). In addition to these two-body wear mechanisms, erosion also causes wear. In that case, a medium (gas, fluid) is flowing along the part and causes material loss. This type of wear is called single-body wear. The five-wear mechanisms will be discussed in the next subsections.

4.6.4 Adhesive Wear

Adhesive wear can be observed when relatively smooth surfaces are in frictional contact and it generally occurs when the two materials have a similar hardness. Adhesive wear yields material transfer from one to the other surface, where the transferred material is called wear debris. The amount of adhesive wear depends on the intensity of adhesion between the two mating surfaces. As was discussed in 4.6.1, this is determined by the metallurgical compatibility of the two materials. The more similar the materials, the higher the adhesive forces generally are.

The friction force was shown (in 4.6.1) to depend on the shear stress τ at the interface. For non-wearing conditions, this shear stress is completely governed by the friction at the interface, that is, the shear stress equals the interface stress τ_{12}. However, when the friction increases, at some instance, the shear stress at the interface will exceed the shear strength τ_1 or τ_2 of the weaker of the two materials.

From that moment, relative motion of the two bodies will no longer take place at the interface, but inside the weaker material. This means that material from the weaker material is transferred to the stronger material and the weaker material starts to wear, see Fig. 4.33.

This type of wear is generally associated with very high levels of friction and thus often leads to the direct stopping of machines or systems. Moreover, the high friction causes a considerable amount of heat generation (see 2.3.5), which also might lead to direct failure of the part or system. Adhesive wear can be prevented by a proper material selection (no similar materials in contact) or by lubrication.

4.6.5 Abrasive Wear

Abrasive wear occurs in a contact when two conditions are met: a considerable difference in hardness (>20 %) exists between the two materials in contact, and the hard material has a rough surface. In that case, the asperities of the harder material penetrate into the softer counter surface and wear is occasioned by scratching and the removal of softer material. This mechanism can be observed when a hard material is sliding against a softer material, which is called two-body abrasion, but also when hard particles are present in the contact area between two softer materials. The latter process is called three-body abrasion and may be caused by sand or dust particles from an outside source, as well as by hard (oxidized) wear particles generated inside the wearing system.

The abrasive wear rate depends on several factors, like the hardness (difference) and the shape of the asperities or particles. Also the applied load is important, since it determines the degree of penetration, which is often quantified by the ratio between penetration depth h and the radius of the 'indenter' (i.e. asperity or

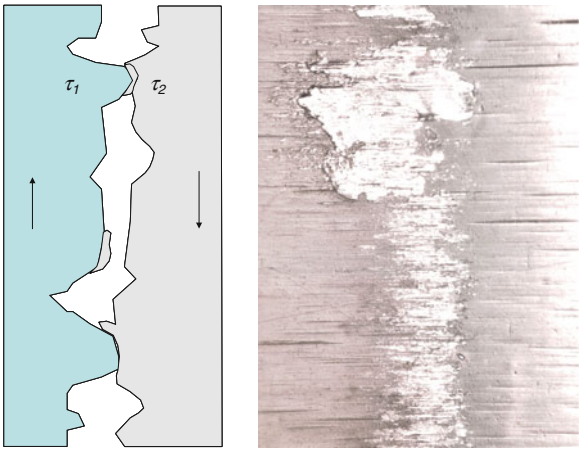

Fig. 4.33 Schematic representation of the material transfer between roughness peaks (*left*) and the result of an adhesive wear process (*right*) [20]

particle) R. For low h/R ratios, ploughing will occur, where only elastic or plastic deformation takes place at the surface, but no large amounts of materials are removed. For a higher degree of penetration, the mechanism will transition from ploughing to cutting. In the latter case, material will be removed and the abrasive wear rate will be much higher. Two wear scars due to abrasive wear by sand particles are shown in Fig. 4.34. The photograph shows the scratches caused by the particles.

The most effective way of reducing three-body abrasion is appropriate sealing or filtering in order to prevent abrasive particles to enter the contact. For two-body abrasion, reduction in the hardness difference between the materials and reducing the surface roughness of the harder material will decrease the wear rate. For both types of abrasive wear, increasing the hardness of the wearing material will generally also reduce the amount of wear.

4.6.6 Corrosive Wear

When rubbing takes place in a corrosive environment, surface reactions occur and reaction products are formed on one or both surfaces. These reaction products generally poorly adhere to the surface and are easily removed upon rubbing. Removal of the reaction products, however, brings bare unprotected material into contact with the corrosive environment again, resulting in a rapid formation of new reaction products. In this way, the repetitive interaction between corrosion and a sliding or rubbing motion may results in relatively high wear rates. Even parts from corrosion resistant stainless steels can show a rapid decrease in their thickness when the oxide layer that normally passivates the material is repeatedly removed.

This type of wear can be prevented by either reducing the effect of the corrosive environment or minimizing the mechanical rubbing action. In many cases, applying a lubricant can provide improvements for both of these aspects.

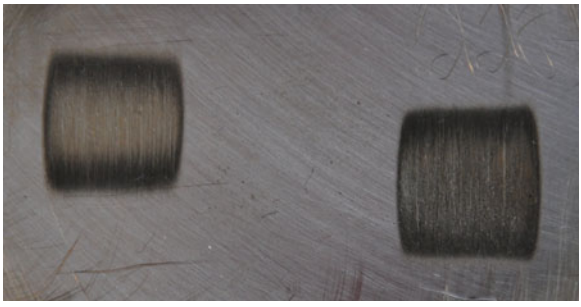

Fig. 4.34 Wear scars due to abrasive wear by sand particles in a rectangular contact

Fig. 4.35 Surface fatigue damage in the rolling elements of a wind turbine bearing

4.6.7 Surface Fatigue

While the previous three mechanisms mainly occur in sliding contacts, fatigue wear is a mechanism that is very common in rolling contacts. As was discussed in Sect. 3.2.9, a cylindrical or spherical element pressed onto a flat plate causes a subsurface stress distribution in that plate. During rolling, the stress at some fixed location in the plate will vary in time due to the passing of the rolling element. A repetitive passing of rolling elements, for example, in a bearing, will cause a cyclic load that could initiate a crack. Since the maximum stress magnitude is observed at some distance below the surface, subsurface cracks will develop in the material. When the cracks propagate, they will eventually reach the surface and a piece of material will break out. This piecewise loss of material causes the wear of the material. An example of surface fatigue damage in a wind turbine bearing is shown in Fig. 4.35.

Note that, contrary to the previous three wear mechanisms, for fatigue wear, no direct contact is required between the two parts. Only the transmission of the (cyclic) stress is sufficient to cause this type of wear. In well-designed bearings, direct contact between rolling elements and inner and outer race is prevented by (the hydrostatic pressure in) the lubricant film that is present in the contact area. This means that adhesive and abrasive wear are normally negligible, but as the stress is transmitted by the lubricant film, fatigue wear may exhibit after some time. Note that it normally takes a large number of load cycles before fatigue wear sets in, while abrasive and adhesive wears produce progressive damage from the start of sliding.

4.6.8 Erosion

Erosion is basically a single-body wear mechanism, which takes place when a fluid or gas is flowing along a part's surface. Measurable wear rates generally require high flow velocities and/or high temperatures. The various forms of erosion will be discussed next [19].

4.6.8.1 Gas Erosion

The mechanism of gas erosion becomes active when gas at very high velocity and generally high temperature flows across a solid part. As material is removed by the passing gas, dimensional anomalies may occur after a prolonged period of exposure to the gas flow. An example of this type of erosion is the creation of wear tracks in cylinder valves of combustion engines, where hot combustion gas is passing continuously at high velocity.

4.6.8.2 Fluid Erosion

A similar wear process may be active when a fluid is passing across a part. Due to the higher density of fluids, material will already be removed at lower flow velocities than for gas erosion. Moreover, the wear rate largely depends on the angle of incidence of the fluid flow on the part. A specific form of fluid erosion in this respect is liquid impingement erosion, where droplets of fluid carried along by a gas collide with the solid surface. The impinging droplets cause a pulsed surface loading of the solid part, which may lead to material loss due to surface fatigue (see 4.6.7). This is a typical problem in steam turbines, where condensed steam hits the metal turbine parts, as well as on aircraft structures hit by rain drops during flight.

4.6.8.3 Cavitation Erosion

Another specific form of erosion is cavitation erosion, where imploding gas bubbles cause pulsed loading of metal surfaces. Cavitation occurs in a fluid when the pressure locally drops below the vapour pressure. The vaporization of the fluid yields formation of gas bubbles, which may shortly after formation collapse again due to an increased pressure level at a nearby location. Cavitation is a common problem near ship propellers, caused by inappropriate flow conditions around the rotating propeller blades. The resulting cavitation erosion may severely damage the blades by the formation of pits, which is sometimes even enhanced by corrosion. Also in centrifugal pumps, the mechanism is a common source of damage.

4.6.8.4 Particle Erosion

Whereas at the previous forms of erosion, the wear is purely caused by the gas or fluid, quite often the medium also contains solid particles. In that case, the erosion wear rate will be considerably higher, since actually a two-body wear mechanism (i.e. abrasive wear, see 4.6.5) becomes active. A common example of particle erosion is the wear of air-breathing engines (e.g. gas turbines) in sandy environments, where sand particles in the ingested air cause abrasive wear of the engine

parts. But also the high wear rates in dredging machinery, like pumps and pipes, caused by the mixture of water and sand are due to this mechanism. The wear rate for this mechanism depends on a number of factors, including the size and shape of the particles, the impingement angle and impact velocity.

4.6.9 Life Assessment

Although a large number of different types of wear exists, the number of numerical models to predict wear rates is quite limited. This is partly due to the large numbers of influence factors in any moving contact situation, like material hardness, surface roughness, friction coefficient, temperature, amount of lubrication. For that reason, mainly phenomenological relations have been derived from test programs on specific material combinations, but these relations cannot easily be generalized to other material types or material combinations.

However, one general relation is available that is applied very widely on all types of wear problems. This relation proposed by Archard [21] is given by

$$V = kF_n\Delta s \tag{4.33}$$

It states that the volume loss V is proportional to the normal load F_n applied to the parts in contact and the travelled distance Δs. The proportionality constant k is called the specific wear rate, which depends on all the factors mentioned before (material hardness, surface roughness, friction coefficient, temperature, lubrication, etc.). This demonstrates that the Archard law is not a physical law describing the actual mechanism, but also is a phenomenological relation that specifically addresses the role of normal load and travelled distance. Nevertheless, the Archard law is applied very often and the obtained values of specific wear rate are a good indication of the class in which the wear problem is situated (e.g. mild wear, severe wear).

Example 4.3 (Specific Wear Rate Calculation) In a wear test, a cylindrical pin (radius $r = 1.0$ cm, mass density $\rho = 5{,}000$ kg/m^3) is pressed against a larger rotating cylinder ($R = 30$ cm) with a pressure $p = 10$ N/mm^2, see Fig. 4.36. The large cylinder rotates with 20 revolutions per minute. After one hour, the mass reduction in the pin is determined to be 0.2 g.

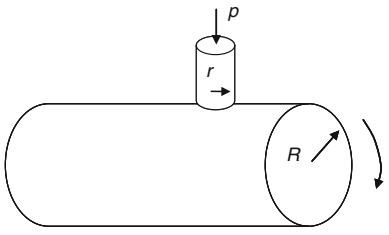

Fig. 4.36 Wear test with *cylindrical pin* sliding along a large rotating cylinder

The specific wear rate for this contact situation can be calculated using Archard's law. The volume loss V can be obtained from the mass reduction by dividing by the mass density:

$$V = \frac{m}{\rho} = \frac{0.2 \cdot 10^{-3}\ \text{kg}}{5000\ \text{kg/m}^3} = 4.0 \cdot 10^{-8}\ \text{m}^3 \tag{4.34}$$

The normal force F_n acting on the cylindrical pin is calculated by multiplying the applied pressure p by the contact area:

$$F_n = pA = p \cdot \pi r^2 = 10^7 \text{N/m}^2 \cdot \pi (0.01\ \text{m})^2 = 3.14 \cdot 10^3\ \text{N} \tag{4.35}$$

Finally, the sliding distance Δs can be calculated from the rotational frequency N and the cylinder circumference:

$$\Delta s = 2\pi RN\Delta t = 2\pi \cdot 30 \cdot 10^{-2}\text{m} \cdot 20\,\text{rpm} \cdot\ 60\ \text{min}\ = 2.26 \cdot 10^3\text{m} \tag{4.36}$$

Using these three quantities in the Archard equation provides the value of the specific wear rate: $k = 5.63\ 10^{-16}\ \text{m}^2/\text{N}$.

4.7 Melting

Melting of a material is the most basic form of thermal failure. It is a quite common failure mode in electrical components, when the heat generated by the electric current cannot be cooled away sufficiently fast. As was discussed in Sect. 3.3, the temperature increase in a part due to the input of a given amount of heat depends on the heat capacity c_p of the material. The latter property therefore provides the translation of the external load (heat) to the internal load (temperature), which eventually determines whether or not failure occurs. The load-carrying capacity of the material associated with this failure mechanism is the melting temperature, which is a material constant depending on the composition of the material.

When a part has molten, it clearly cannot fulfil its intended function anymore, so failure has occurred. However, in many cases, the material properties, like strength or stiffness, already reduce considerably when the temperature approaches the melting temperature. Therefore, failure often occurs due to other failure mechanisms like static overload, and the thermal load can then be considered as a secondary load that deteriorates the load-carrying capacity of the part or structure. This will be discussed in the next subsection.

4.8 Thermal Degradation

In materials that are used at elevated or high temperatures, the material properties may deteriorate gradually in time. As was discussed in Chap. 1, this is a secondary load which reduces the load-carrying capacity of a material. The deterioration in most cases concerns the mechanical properties like strength and stiffness, but also the electrical properties of a component may degrade. The latter will be discussed in Sect. 4.9.2, while the present subsection focuses on the mechanical properties. As the mechanical properties of materials are determined by the microstructure, that is, the crystal structure at a microscopic level, deterioration of the mechanical properties can generally be explained by the microstructural changes appearing during thermal degradation.

Microstructural degradation is a change of the material microstructure due to a (prolonged) exposure to high temperatures, resulting in a decrease in the material properties. Since diffusion rates in materials increase significantly with rising temperatures, the microstructure in many materials is not stable anymore at elevated temperature. Especially, materials in which microstructures are specifically designed to obtain the optimal mechanical properties are sensitive for this type of degradation. Generally, several different phases are produced by specific heat treatments, but these processes can rather quickly be reversed during operation at (too) high temperatures.

Nickel-based superalloys, which are applied in gas turbine blades, constitute a class of materials that is sensitive to microstructural degradation. The superior high-temperature mechanical properties of these materials are attributed to the very specific microstructure. The nickel matrix material contains a high volume fraction ($\sim$70 %) of Ni_3Al precipitates, appearing as cuboidal particles with similar sizes in a more or less regular pattern (see Fig. 4.37). Since the deformation of these materials must take place in the more ductile matrix phase, which only is present in the very narrow channels around the precipitates, a very high mechanical strength is accomplished.

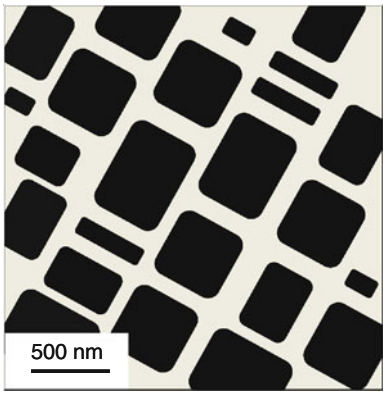

Fig. 4.37 Schematic representation of the typical microstructure of a nickel-based superalloy, consisting of cube-like precipitates in a nickel matrix

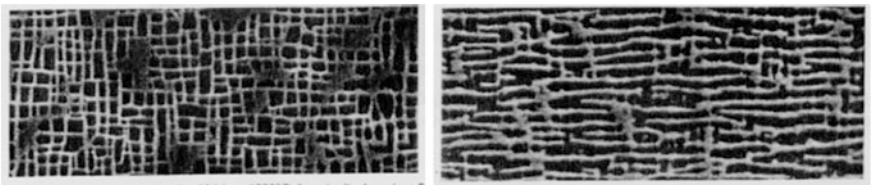

Fig. 4.38 Comparison of the microstructures of the initial material (*left*) and after a certain period of use at high temperature (*right*). The precipitates have been removed by chemical etching and thus appear as the dark phase in these micrographs

However, during use at high temperature (~800–1,000 °C), the shape of the precipitates will change due to diffusion processes [22]. If no mechanical loads are present, the precipitates will become spheroidal and their size will increase. In the presence of a mechanical stress, as is the case in rotating gas turbine parts, the cuboidal precipitates evolve into elongated plates directed normal to the direction of the applied stress. This process is called rafting and is illustrated by the two micrographs in Fig. 4.38.

Both the coarsening of the microstructure and the rafting process affect the mechanical properties of the alloy. Since the regions of the matrix phase increase in width, deformation becomes easier and the material strength decreases. Also the creep rates increase and the fatigue resistance diminishes.

4.9 Electric Failures

Electronic components and devices nowadays have a major impact on the functioning of systems in all sectors of industry. Only very few systems are still operated without any electronic control system. Modern cars, aircraft, trains, climate systems, industry process control, traffic control and many more systems only operate with well-functioning electronic parts. Therefore, the failure of electronic components also significantly affects the reliability of all these systems, which also indicates the importance of understanding their failure behaviour.

However, compared to mechanical systems, the failure behaviour of electronic systems is rather complex. There are mainly three reasons for this complex behaviour. The first reason is the complexity of the systems, which are generally composed of very large numbers of individual components. Secondly, many of these parts are loaded in a similar way, yielding comparable probabilities of failure for many components. This means that it is often impossible to identify a limited number of critical components, which dominate the system failure behaviour. The third reason is the difficulty of measuring or estimating the loading of individual components. It is therefore hard to establish a quantitative relation between applied load and service life time, which makes predictions of system failures challenging.

This situation is quite different from mechanical systems, where typically only a few parts and failure mechanisms are dominant, which can also rather easily be predicted from the estimated or measured usage and/or loads. The complexity of the failure behaviour of electronic systems therefore often leads to the conclusion that failures in these systems are unpredictable and occur completely randomly.

However, a better understanding of the failure behaviour on the component level can also for these systems lead to an improved predictability of failures. As will be explained in Chap. 8, this requires a thorough analysis of the different failure modes, and more importantly, the associated failure mechanisms. From the overview of typical failures in electronic parts and systems in Chap. 8, it can be concluded that only a fraction of the failures is due to electric loads. Many failures, like fracture of leads and melting due to overheating, are caused by mechanical and thermal loads and the associated mechanisms that have been discussed in Sects. 4.2–4.8. In Sect. 4.9.6, a number of approaches to include these loads in the life assessment of electronic systems will be discussed.

However, this section will start with specifically treating the electric failures that are caused by electric loads, where it should be remarked that failure in an electric sense can be roughly divided into two categories: either the conductive capacity of a component intended for conduction or the insulative capacity of an insulator is reduced or completely lost.

4.9.1 Current Overload

The major electric failure mechanism of conducting parts is the failure due to an excessive electric current that leads to overheating. The eventual failure can be due to melting, evaporation or cracking (due to a thermally induced mechanical load) of the conducting part. Although strictly speaking this is not an electric failure mechanism, that is, the current does not directly affect the electric properties of the material, the failure is due to an electric load.

As was mentioned before, the internal load in terms of current density is governing this failure mechanism. The capacity of an electric component, that is, the maximum allowable current density, will depend both on the electric and thermal properties of the part. In Sect. 2.3.5 on external thermal loads, it was shown that the amount of heat dissipated by an electric current depends on the resistance of the conducting part

$$P = I^2 R \tag{4.37}$$

This equation can be transformed to the local quantities current density (J) and resistivity (ρ) by substituting the respective definitions into Eq. (4.37)

$$P = (JA)^2 \frac{\rho l}{A} = J^2 \rho (lA) \tag{4.38}$$

showing that the heat generated per unit volume is the product of J and ρ. The maximum allowable current density is determined by the cooling capacity of the system, that is, the capacity to quickly remove the generated heat from the component. Generally, the cooling capacity is proportional to the surface area of the component ($\sim$radius r), while it is shown above that the heat generation is proportional to the cross-sectional area of the component ($\sim r^2$). This means that a thin wire most often allows a higher current density than a thick wire. This is also demonstrated in Example 4.4 below.

Similar problems can occur at connections between two conductors. If the connection is not sound (in terms of conductivity), a locally high resistance exists and heat is generated leading to a hot spot that can deteriorate the connection.

Finally, an excessively high current is sometimes caused by a loose connection which only leaves a small cross-sectional area to carry the complete current. The remaining connection then heats up easily, melts and breaks up disrupting the current. A useful application of this principle is found in fuses, where the weak link in the network is designed to exist at a convenient and easily accessible location.

As will be shown in Chap. 8, other causes of conduction failures are bad contacts due to corrosion and mechanical failures of wires and leads due to fatigue of overloads. However, these failures are not caused by electric loads.

Example 4.4 (Maximum Current Density in Wire) A copper wire with 5 mm diameter carries a current I for a short period of time (10 min). The amount of heat that can be removed per unit surface area is $P_c = 10$ kW/m^2. The resistivity of copper $\rho = 1.72 \times 10^{-8}$ Ωm, the heat capacity $C_p = 385$ J/kgK and the melting temperature $T_m = 1{,}085$ °C. The mass density $\rho_{md} = 8{,}940$ kg/m^3. The maximum allowable current density in the wire can now be calculated.

To achieve that the wire carries the current for ten minutes without melting, the amount of heat generated beyond the maximum amount of heat that can be removed by cooling must be limited

$$Q_{\text{eff}} = Q_{\text{generated}} - Q_{\text{cooling}} = \left(P_{\text{generated}} - P_{\text{cooling}}\right)\Delta t < Q_{\text{allowable}} \tag{4.39}$$

The allowable amount of heat is determined by the allowable temperature rise ΔT, the mass density ρ_{md} and the heat capacity C_p

$$Q_{\text{allowable}} = mC_p\Delta T = (\rho_{md}Al)C_p\Delta T \tag{4.40}$$

where the allowable temperature rise ΔT from room temperature to melting point equals 1060 °C. The generated amount of heat depends on the current density and is given by

$$Q_{\text{generated}} = P_{\text{generated}}\Delta t = J^2\rho Al\Delta t \tag{4.41}$$

Finally, the amount of heat that can be removed by cooling depends on the surface area of the wire, and is therefore given by

$$Q_{\text{cooling}} = P_{\text{cooling}}\Delta t = P_c(2\pi rl)\Delta t \tag{4.42}$$

Combining the latter three equations enables the calculation of the maximum allowable current density:

$$Q_{\text{eff}} = Q_{\text{allowable}} \Leftrightarrow \Delta t\left(J^2\rho lA - 2P_c\pi rl\right) = \rho_{md}AlC_p\Delta T \tag{4.43}$$

which yields the following expression for the current density

$$J = \sqrt{\frac{\rho_{md}\pi r^2 C_p\Delta T + 2P_c\pi r\Delta t}{\rho\pi r^2\Delta t}} \tag{4.44}$$

Applying all the numerical values for the different parameters yields a current density for the 5 mm wire of $J = 2.88 \times 10^7$ A/m^2, which corresponds to a current $I = 565$ A. If a thicker wire is used, for example, 8 mm diameter, then the allowable current increases significantly to $I = 1284$ A, while the allowed current density decreases slightly to $J = 2.55 \times 10^6$ A/m^2. The reduction in allowable current density is caused by the fact that the cooling is proportional to the surface area, while the heat generation is proportional to the volume of the wire. If no cooling would be applied, the allowed current density is independent of the wire diameters. Therefore, this example demonstrates that not the current is governing the failure, but the current density.

4.9.2 Intrinsic Breakdown

When very high electric fields are applied, an insulating material may break down. This means that suddenly large numbers of electrons are excited into the conduction band (see Sect. 3.4) and the material looses its insulating capacity. This process is called intrinsic breakdown or dielectric breakdown. Due to the electric field, the freed electrons start to move and cause a large electric current. This yields a considerable dissipation of heat, which might result in degradation, melting or even vaporization of the dielectric material. The dielectric strength or breakdown strength (E_{bd}) of a material represents the maximum magnitude of the

Table 4.3 Dielectric constant and dielectric strength for some typical materials [23]

Material class	Material	Dielectric constant ε_r	Dielectric strength (V/mm)
Ceramics	Mica	5.4–8.7	40,000–80,000
	Porcelain	6.0	1,600–16,000
	Fused silica	3.8	10,000
Polymers	Nylon	3.6	16,000
	Polystyrene	2.6	20,000–28 ,000
	Polyethylene	2.3	18,000–20,000

electric field that can be resisted without breakdown. The dielectric strength of some typical materials is shown in Table 4.3.

In some cases, breakdown of an insulator is due to local weak spots. A conductive needle-like electrode will enhance the local electric field at its tip which may exceed the local electric breakdown strength of the surrounding insulating material igniting a complete breakdown.

A similar breakdown process that often occurs in semi-conducting materials is avalanche breakdown. The freed electrons are accelerated by the electric field. As a result their energy can reach a level such that collisions with other bound electrons leads to the creation of new electron–hole pairs. This creates an avalanche effect in the semiconductor, which eventually leads to breakdown of the material.

Intrinsic breakdown is observed both in high voltage applications (i.e. power electronics, power generation/distribution) and low voltage (micro)electronics. In the former type of applications, the high voltages must be separated by a thick layer of insulating material to lower the electric field strength and prevent the exceedance of the material breakdown strength. However, although in (micro) electronics the applied voltages are much lower, also the distance between separate components is smaller. This means that also in these applications, large electric fields exist and breakdown may lead to failure of the components and circuits. Well-known components suffering from breakdown are diodes for which the applied voltage exceeds the reverse breakdown strength. Also capacitors breakdown, often due to degradation of the dielectric properties by electrolyte dry-out, contamination or gas formation inside the dielectric material. Finally, printed circuit board (PCB)s suffer from the growth of conductive anodic filaments into the board, resulting in current leakage, reduced dielectric strength and short circuits between traces.

Although the final failure mechanism in insulators is the intrinsic breakdown, there are many other mechanisms that affect this process [24]. In that case the applied electric field is considered as the primary load, but the capacity of the system can be decreased by secondary loads. Mostly the degradation process is a very slow process taking long periods of time, whereas breakdown can occur suddenly when the breakdown strength of the material has decreased below the applied electric field. The service life of such a component is therefore mainly determined by the degradation process, of which only some depend on the applied electric loads. The most common degradation or ageing mechanisms preceding breakdown are discussed next.

4.9.2.1 Thermal Ageing

For many materials the breakdown strength is temperature dependent, that is, E_{bd} is lower at higher temperatures. Moreover, prolonged exposure to elevated temperatures often causes an accelerated deterioration of the material and a resulting decrease in the breakdown strength. A high temperature in an insulator can be

caused by external heat sources, but in most cases is caused by internally generated heat due to a small leakage current (Ohmic losses, see Eq. (4.37), or dielectric losses). In the former case, a thermal load causes the degradation but in the latter case both the ageing and the eventual breakdown are due to an electric load.

In case of a rather quick temperature increase, which occurs when no thermal equilibrium is established, the degradation is also quite fast and the process is called a thermal runaway.

4.9.2.2 Partial Discharge

Breakdown of an insulator not always occurs across the complete component, but can also take place locally, for example, at the surface or at some internal location [25]. In that case the material breakdown strength is exceeded, either by a locally low material breakdown strength or a locally high electric field, while the overall strength is sufficient to carry the applied field. Consecutive discharges can however damage the material, for example, by (spark) erosion or pitting, which eventually may yield complete breakdown of the material. Also electric treeing can be a result of partial discharges.

4.9.2.3 Electric Treeing

Repeated partial discharge in some materials causes carbonization of the insulating material. This yields a conducting carbon track inside the insulator, which acts as a needle electrode. Since the electric field is intensified at the sharp tip of this electrode, exceeding the local breakdown strength is more likely. Moreover, electrons are attracted from the surrounding towards the tip, which yields additional growth of branched carbonization tracks. The process continues until the electric field exceeds the breakdown strength of the remaining insulation and complete breakdown occurs.

4.9.2.4 Water Treeing

The combination of an electric field, humidity and the presence of certain elements in the material may lead to the development of water trees [26]. In that case a network of nanotracks is created inside the material, in which ions can move and transport charge. But since the mobility of the ions is limited, also the conductivity of the tracks is low. Therefore, water trees may even completely bridge the insulation without causing breakdown. However, water trees generally have a (much) lower E_{bd} than the original insulating material and therefore often cause breakdowns.

4.9.2.5 Surface Discharge Processes

Whereas the previous mechanisms occur in the interior of insulating materials, partial discharge can also take place at the outer surface. The leakage currents that develop are called creepage currents. These currents, and the associated thermal losses, may lead to treeing along the surface. The resulting structures develop in the direction of the electric field. When also moisture and pollution are present, a mix of electrochemical reactions and discharges erodes the surface leading to the development of tracks. Ultimately, a breakdown along the surface will occur.

4.9.2.6 Interface Discharge

When two insulating materials are connected and their interface is aligned with the electric field, discharge may occur along the interface [27]. Often external effects, like thermomechanical stresses due to fast temperature changes or applied pressures, enhance this process. The result is an extensive network of small creepage tracks that can act like needle electrodes and may lead to fast breakdowns in high voltage equipment.

4.9.3 Breakdown in Gas and Vacuum

In many electrical applications two components (conductors) with different potential levels are separated by a gas or (in power electronics applications) by a vacuum. This can be either an explicitly arranged gas or vacuum insulation, as in gas insulated systems (GIS), or due to the fact that the conductor is surrounded by gas (in most cases air), as for example high voltage overhead lines used to transport electricity.

The breakdown process in gases is an avalanche type breakdown similar to that in semiconductors. The mean free path of a molecule in a gas is the average distance it can travel before a collision with another molecule occurs. In air at ambient pressure the mean free path is 96 nm. Since electrons are much smaller than molecules, their mean free path in a gas is typically five times larger. If an electric field exists between two conducting plates, electrons travelling in the gas in between the electrodes are accelerated by the field. If the energy an electron gained during acceleration exceeds a certain threshold, collision of the electron with a molecule will lead to ionization of the molecule and freeing of new electrons. This creates a chain reaction leading to avalanche breakdown of the gas and the formation of an arc between the two electrodes.

This breakdown only occurs when on the one hand sufficient numbers of molecules are available to provide new electrons and on the other hand the mean free path and electric field strength are large enough for the electrons to gain the amount of energy required for ionization. The gas pressure has a large influence on this process, since the mean free path is inversely proportional to the gas pressure,

while the number of molecules available for ionization increases with gas pressure, but also is proportional to the separation distance of the electrodes. The relation between breakdown voltage V_{bd} and distance d between the electrodes and gas pressure p is given by Paschen's law

$$V_{bd} = \frac{Apd}{\ln(pd) + B} \tag{4.45}$$

where the constants A and B depend on the type of gas that is used. For air at atmospheric pressure, $A = 43.6 \times 10^6$ V/(atm·m) and $B = 12.8$. The relation can be visualized in a plot of breakdown voltage (V_{bd}) against the product of gas pressure and distance (pd), which is known as the Paschen curve. The curve for air is shown in Fig. 4.39. It appears that a minimum breakdown voltage exists at a specific combination of gas pressure and electrode separation distance, that is, $pd = e^{1-B}$. For larger values of pd, the number of molecules encountered by electrons travelling between the electrodes is larger, but the mean free path is smaller, which means that only a fraction of the electrodes can gain sufficient energy for ionization. Therefore, a higher voltage is required to create an arc. For smaller values of pd, electrons can travel relatively large distances and can gain sufficient energy, but the number of ionizations is insufficient to create an arc.

Discharges not only occur in large volumes of gas in between two conductors, but can also take place in gas-filled cavities in liquid or solid insulators. These irregularities in the material may lead to a locally increased magnitude of the electric field, which yields a concentration of electrons at one side of the cavity. When the breakdown strength of the gas in the cavity is exceeded, an arc will be created.

To quantify the breakdown strength of insulators for various cavity sizes, the Paschen curve can be modified to provide the breakdown voltage for different cavity sizes. In this way, the maximum allowable cavity size can be determined for a given electric field strength. For example, when an electric field of 7 kV/mm is present and irregularities in the material cause a factor 2.3 increase in the local

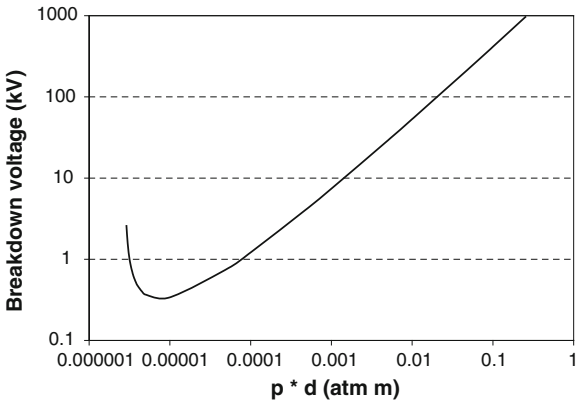

Fig. 4.39 Paschen curve for air at 20 °C

electric field near a cavity (i.e. the local field equals 16 kV/mm), the cavity size should be smaller than 15 μm. After the occurrence of the first discharge in such a cavity, the process of electron concentration at the circumference and successive discharge will repeat. This means that the material is locally exposed to an electron bombardment leading to erosion and pit formation. An electric treeing process as described in the previous subsection may lead to the formation of a carbon track in the material, which eventually may lead to complete breakdown of the insulator. This process of erosion and pit formation can proceed for several years before final breakdown occurs in only hours or minutes.

4.9.4 Electrostatic Discharge

Electrostatic discharge (ESD) is a phenomenon that can easily cause failures in microelectronic devices such as integrated circuits. In most cases, an ESD event is caused by electrostatic electricity (see also 2.4.1) present on parts of human bodies (i.e. hands) which come into contact with electronic devices, for example, during assembly or maintenance. The resulting discharge is relatively small, but may cause failure of parts of the very sensitive electronic device by breakdown or excessive currents. Due to the small magnitude of the discharge, it is mostly undetectable by the human senses and therefore many of the resulting failures are unexpected. Moreover, quite often the ESD event does not yield a direct failure of the device, but causes some amount of degradation that affects the device reliability on the long term.

To prevent failures due to ESD in electronic devices, personnel assembling or repairing these devices should use preventive measures like garment with conducting filaments, conducting wrist and foot straps and anti-static mats.

4.9.5 Electromigration

A failure mechanism that is typical for integrated circuits, and especially affects the metal interconnects that link the transistors and other components, is electromigration. In these very small conducting parts, very high current densities occur. As a result, the large numbers of fast-moving electrons collide with the (ionized) metal atoms and transfer part of their momentum. This causes the atoms to move away from their original position, which may lead to either the creation of a break or gap in the conductor (open circuit) or to the movement of atoms to other interconnects (short circuit).

Already in the 1960s, a phenomenological relation was derived to relate the mean time to failure (MTTF) due to electromigration to the applied current density J and temperature T. This relation is called Black's law [28]

$$\mathrm{MTTF} = \mathrm{AJ}^{-n} e^{\frac{E_a}{kT}} \tag{4.46}$$

where A is a constant based on the cross-sectional area of the interconnect, E_a is an activation energy depending on the applied material, k is Boltzmann's constant and n a scaling factor (usually set to 2 [28]). This relation shows that both an increasing current density and an increasing temperature decrease the service life of the interconnect.

4.9.6 Life Assessment

As was discussed at the beginning of this section, the specific properties of electronic systems make their failure behaviour quite different from the failure of mechanical systems. Due to these properties, the failure of electronic systems is often considered to be a random process which cannot be predicted. Therefore, the traditional way of quantifying the service life of electronics is the statistical approach based on past experience or on vast testing programs. This yields average service life values, generally expressed as mean time between failures (MTBF), with quite large uncertainties. As will be discussed in Chap. 6, this experience-based approach does not include any knowledge on the physics of failure and therefore has some disadvantages. Firstly, the past experience (and associated MTBF values) can only be extrapolated to the future when the usage and loading of the systems will be the same as in the past. And secondly, a considerable amount of data is required to accurately quantify the failure behaviour.

These drawbacks can be circumvented when physical models for the relevant failure mechanisms are applied to predict the service life time of components at given operating conditions. However, while this works quite well for mechanical components with a limited number of critical parts, the large number of individual components makes this approach quite infeasible (yet) for electronic systems.

Therefore, while both the purely experience-based and the purely model-based approaches are not suitable for electronics, a mix of the two approaches seems to be the best solution in this case. That can be achieved by the application of phenomenological models. These models relate the service life (or MTBF) of systems to one or several governing loads, but they are based on a certain amount of failure data at the system level at various conditions rather than on the underlying fundamental failure mechanisms of the individual components. In that way, the randomness of the failure process can be reduced considerably, while at the same time the modelling effort is limited.

Typical governing loads that can be included in these phenomenological models are electric loads (field strength, current density), but also environmental parameters like temperature and humidity and mechanical loads (i.e. vibrations, shocks). Three examples of this type of models for electronic applications are as follows:

- Black's law for electromigration (see 4.9.5)
- Eyring–Peck model for PCBs
- Power law model for intrinsic breakdown of insulators

The latter two models will be discussed in the remainder of this subsection.

4.9.6.1 Eyring–Peck Model for PCBs

The failure of PCBs is often due to the corrosion of the plastic package, where short circuits arise due to metallization processes [29, 30]. From a fairly extensive number of tests, it is recognized that the time to failure for these PCBs depends on both the humidity and the temperature. A phenomenological model is derived, which is called the Eyring–Peck model that relates the humidity H and temperature T in the following way to the time to failure t

$$t = t_{\text{ref}} \left(\frac{85}{H}\right)^3 \exp\left[-\frac{\Delta E}{k}\left(\frac{1}{T} - \frac{1}{358.1}\right)\right] \tag{4.47}$$

In this model, the time to failure is related to the life time (t_{ref}) at a reference condition of 85 % relative humidity and 85 °C. The parameter ΔE represents the activation energy of the corrosion process and k is Boltzmann's constant.

4.9.6.2 Power Law Model for Insulating Material Degradation

As was discussed before, insulating materials gradually degrade when an electric field is present. This degradation starts at locations where the local electric field is higher due to material inhomogeneities or impurities. Moreover, elevated temperatures enhance the degradation process, which means that degradations that normally take years to develop may now occur in only several minutes. Material tests have demonstrated that the following empirical power law relation exists between the service life time T and the applied electric field E

$$\log T = \log C_{\text{bd}} - n \log E \tag{4.48}$$

which is equivalent to

$$T = \frac{C_{\text{bd}}}{E^n} \quad \Leftrightarrow \quad E^n T = C_{\text{bd}} \tag{4.49}$$

with C_{bd} and n material constants. Since the value of the parameter n is rather high, that is, ranging from 9 to 20, depending on the material, the component service life is extremely sensitive for variations in applied electric field strength. For example, for polyethylene ($n = 9$), the service life decreases with a factor 500 when the field strength is doubled.

Analogous to the Miner cumulative damage law (see Sect. 4.4.3), the power law introduced here can also be applied to calculate the service life for a component exposed to different load levels (i.e. field strengths). The parameter C_{bd} can then be considered as the maximum allowable amount of damage and the combinations of service time periods t_i at field strength E_i can be accumulated according to

$$\sum_i t_i E_i^n \leq C_{\mathrm{bd}} \tag{4.50}$$

This means that a certain period of time t_A at field strength E_A is equivalent in terms of life consumption to another period t_B at E_B, according to

$$t_A = t_B \left(\frac{E_B}{E_A} \right)^n \tag{4.51}$$

4.10 Corrosion

Corrosion is probably the most common failure mechanism that is active in a large variety of systems in various sectors, ranging from maritime and offshore to infrastructure, aerospace and chemical industries. The costs of preventing and solving corrosion problems are enormous and a good understanding of the associated mechanisms therefore offers a considerable potential to improve the efficiency of these (maintenance) processes.

While the origin and basics of electrochemical loads have been treated in Sect. 3.5, the basics of corrosion as a failure mechanism are treated in the present section. Firstly, the determination of corrosion rates will be discussed. After that an overview of the most common corrosion mechanisms will be provided and the measures to prevent corrosion will be treated. Finally, the topic of high-temperature oxidation will be addressed.

4.10.1 Corrosion Rates

In the previous chapter, it was shown that a corrosion reaction can take place when three items are present:

- a galvanic couple formed by two (reactive) materials;
- a fluid electrolyte with a sufficiently high concentration of ions;
- an electric connection between the two materials.

When these requirements are met, an electric current will establish, which is directly proportional to the number of corrosion reactions that occur. The magnitude of the current is therefore a good indication for the amount of material

that is corroding away. By understanding the kinetics of corrosion reactions, the established corrosion current can be predicted and, at least for general corrosion processes, a measure for the expected corrosion rate is available, as will be treated next. Note that for localized corrosion processes, the relation between corrosion current and corrosion rate is less direct, since the number of active sites (e.g. pits) is generally unknown. An alternative way of assessing corrosion rates is using experimental data and deriving empirical relations. Some examples of such relations will be discussed at the end of this subsection.

4.10.1.1 Kinetics of Corrosion Reactions

To assess the degradation rate of the material, for example, in terms of mm loss of material per unit time, the number of corrosion events (oxidized atoms that are dissolved in the fluid) per unit of time and unit of area should be considered. Since each corrosion reaction also yields the transfer of a specific number of electrons, the corrosion current is a good indicator of corrosion rate. The proportionality between the electric current I_c, in Ampère (A) or Coulomb per second (C/s), and the amount of mass m that has reacted is provided by Faraday's law:

$$m = \frac{aI_c\Delta t}{nF} \tag{4.52}$$

where a is the atomic weight associated with one reaction, Δt is the time period, n is the number of charge equivalents that is transferred per reaction and F is Faraday's constant. Dividing the reacted mass by time t and the surface area, the corrosion rate r (in kg/m^{2}s) is obtained to be [31]

$$r = \frac{m}{tA} = \frac{i_c a}{nF} \tag{4.53}$$

where $i_c = I_c/A$ is the current density (A/m^2). Finally, the penetration depth d (in m per unit time) can be obtained through dividing by the mass density ρ of the material

$$d = \frac{i_c a}{nF\rho} \tag{4.54}$$

The latter two equations demonstrate that the corrosion current density i_c governs the corrosion rate, so quantification of this quantity enables, for a general corrosion process, prediction of the degradation rate of a material.

In Sect. 4.5.2, it has been shown that any half reaction (oxidation or reduction) is associated with a single electrode potential. In an isolated half reaction at this potential, for example, the reaction

$$2H^+ + 2e^- \rightarrow H_2 \tag{4.55}$$

will be in equilibrium, which means that the rates of the forward (r_f) and backward (r_b) reactions are equal. This is expressed as

$$r_f = r_b = \frac{i_0 a}{nF} \tag{4.56}$$

The corrosion current density in this case is i_0, which is called the exchange current density. Just as the electrode potential, also this exchange current density is a fundamental quantity associated with the specific half reaction. While both quantities cannot be derived from first principles, their values must be obtained from measurements and can be obtained from handbooks for most common materials. Note that the value of i_0 heavily depends on the surface on which the reaction occurs (while the electrode potential is unaffected by the surface). For example, the exchange current density for the reaction in (4.55) on mercury is only 10^{-12} A/cm^2, on iron in the order of 10^{-7} and on platinum even 10^{-3} A/cm^2 [31].

However, in practical situations, the half reactions will not be isolated, the applied potential will differ from the electrode potential and the resulting current density (and associated corrosion rate) will be different. The effect of increasing or decreasing the potential E relative to the electrode potential e is called polarization. The polarization or overpotential is defined as $\eta = E - e$. When the surface potential E becomes more negative, the effect is denoted a cathodic polarization η_c, while an anodic polarization η_a raises the potential in positive direction. Moreover, polarization can be classified into two types. In activation polarization, one of the steps in the oxidation or reduction reaction controls the reaction rate and thus the transfer of charge. In concentration polarization, not the reaction steps but the concentration of ions in the electrolyte or their diffusion rate limits the reaction rate and thus governs the amount of polarization.

For an activation polarization case, it will be shown now how the overpotential affects the two half reactions and determines the resulting current density i_c. For any half reaction, the overpotential η is related to the current density as

$$\eta_a = \beta_a \log \frac{i_a}{i_0} \tag{4.57}$$

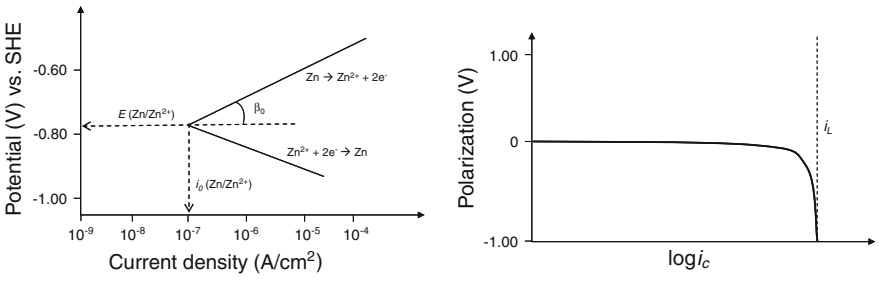

Fig. 4.40 Activation (*left*) and concentration polarization (*right*) plotted versus current density

for anodic polarization, and

$$\eta_c = \beta_c \log \frac{i_c}{i_0} \tag{4.58}$$

for cathodic polarization. In these equations, i_0 is the exchange current density, while i_a and i_c are the actual current densities at the anode or cathode, respectively. The proportionality constants β_a and β_c are called Tafel constants, which determine the magnitude of the current density (and reaction rate) for a certain overpotential. Since an anodic polarization yields an increase in electrode potential, the value of β_a must be positive and, for similar reasons, β_c is negative. The absolute values of β_a and β_c range from 0.03 to 0.2 V per decade of current. The variation of overpotential with current density for the Zn/Zn^{2+} reaction is shown in the left-hand plot of Fig. 4.40.

In case of concentration polarization, the available concentration of ions determines the overpotential. The Nernst equation, that was discussed in Sect. 3.5.2, shows that the electrode potential depends on this concentration. The relation with the current density or reaction rate is given by

$$\eta_{\text{conc}} = \frac{RT}{nF} \ln \left[1 - \frac{i_c}{i_L}\right] \tag{4.59}$$

where i_L is a limiting current density, representing the maximum reaction rate that cannot be exceeded due to a limited diffusion rate of the required species. In Fig. 4.40, the polarization η_{conc} is plotted versus the (logarithm of the) current density i_c. This plot shows that the polarization is negligible until the limiting current density i_L is approached.

The total polarization, being the sum of activation and concentration polarization, can then (for cathodic polarization) be expressed as

$$\eta = \beta_c \log \frac{i_c}{i_0} + \frac{RT}{nF} \ln \left[1 - \frac{i_c}{i_L}\right] \tag{4.60}$$

For anodic polarization of metal dissolution reactions, the concentration polarization is usually absent, so the total polarization reduces to the expression for activation polarization.

From the previous discussion, it can be concluded that the four parameters i_L, i_c, β_a and β_c can be used to describe the kinetics of most common electrochemical corrosion reactions. While until now only half reactions have been discussed, in practice always two half reactions will occur simultaneously. Moreover, for these pairs of half reactions, the amount of oxidation must equal the amount of reduction, thus guaranteeing the conservation of charge. This coupling of half reactions also means that the polarization and resulting current density (i.e. reaction rate) is fully determined by the specific combination of materials and the environmental conditions. This will be demonstrated now for the corrosion reaction of zinc in an acid solution. This involves the two half reactions:

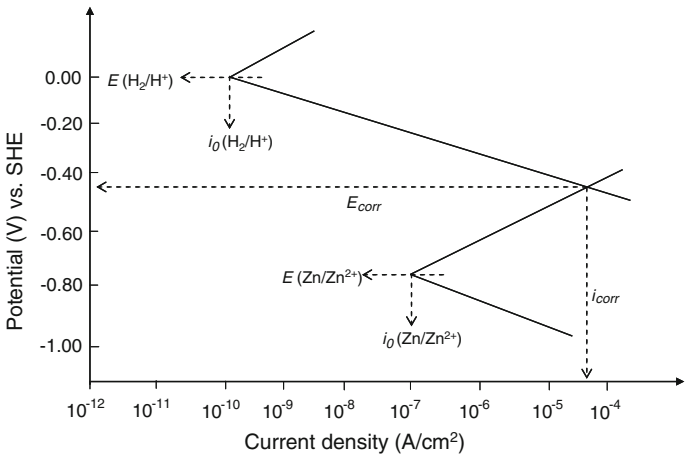

Fig. 4.41 Combination of half reactions providing the corrosion potential (E_{corr}) and current density (i_{corr})

$$Zn \rightarrow Zn^{2+} + 2e^{-} \tag{4.61}$$

$$2H^{+} + 2e^{-} \rightarrow H_2 \tag{4.62}$$

For these two half reactions, the potential can be plotted versus the current density, as is shown in Fig. 4.41.

Since the difference in electrode potential cannot coexist on an electrically conducting material, polarization will shift the potentials to a steady state value, the corrosion potential E_{corr}. This is illustrated in Fig. 4.41 by the straight lines. The associated corrosion current density i_{corr} represents the corrosion rate for this situation.

An important effect that influences the establishment of this balance between the oxidation and reduction reaction is the formation of passive films. These films are formed on specific metals and alloys, for example, stainless steels typically by the addition of >12 % chromium and 8 % nickel to the alloy, and provide an

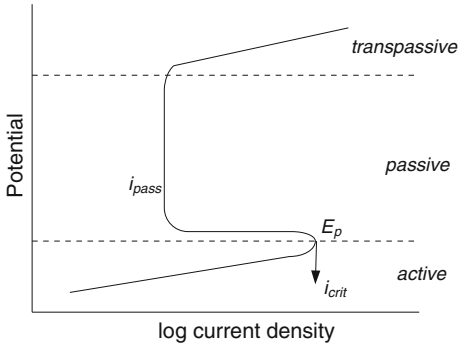

Fig. 4.42 Schematic representation of passivation behaviour

effective barrier against corrosion. The passivation is particularly active at conditions with high anodic polarization. The effect is illustrated in a potential versus current density plot in Fig. 4.42.

At low potentials, typical for acid solutions, the metal is in an active state and the normal behaviour is observed, where the corrosion rate (proportional to the current density) increases with potential. However, at the primary passive potential $E_{p,}$ the passive film becomes stable and the corrosion rate (i_{pass}) drops to a value that might be six orders of magnitude lower than in the active state (i_{corr}). At even higher potential, the passive layer breaks down and the reaction rate increases again. However, these high potential values are rarely observed in practical situations. The presence of a passive film on the surface of a metal or alloy thus reduces the general corrosion rate considerably. However, local breakdown of the passive film may facilitate local corrosion mechanisms like pit formation if repassivation of the film does not occur sufficiently fast.

To summarize this subsection, the discussion on corrosion reaction kinetics demonstrates that if the details of the reaction are known, the degradation rate can be predicted using the equations provided. However, this means that all details in terms of contributing half reactions, conditions and concentrations must be fully known, which is often not the case in practical situations. Therefore, an alternative approach can be followed in which experimental data are used to derive empirical relations for the corrosion rates or service life time. This will be explained in the next subsection.

4.10.1.2 Empirical Corrosion Rate Assessments

Instead of predicting the corrosion rate based on electrochemical principles, as treated in the previous subsection, corrosion rates are also often assessed using empirical methods. This means that corrosion rates are monitored in experiments at different conditions or on various materials. If a sufficient amount of experimental data has been collected, the corrosion rates for that material–conditions combination can be obtained. When also data has been collected at other conditions, expressions may be derived that relate these conditions (e.g. temperature, salinity, humidity, pH) to the corrosion rate in a quantitative sense.

Corrosion testing is performed by exposing well-defined and carefully prepared test specimens to corroding environments. This can be either normal environments in which the test duration is similar to normal component lifetimes, or more aggressive environments in which accelerated tests are performed. In the latter case, the prediction of lifetimes at normal environmental conditions may be difficult, but comparing different materials or coatings is generally possible.

The corrosion rates are generally assessed by measuring the weight loss of exposed specimens at several moments in time. The weight loss W is then used to determine the penetration rate CR (e.g. mm per year) through division by the mass density ρ, the surface area A and the time period Δt:

$$CR = \frac{W}{A\rho\Delta t} \tag{4.63}$$

The observed penetration rate thus depends on the load-carrying capacity of a component, which in this case is the corrosion resistance of the material. That resistance is generally considered to be very good if the penetration rate is less than 0.1 mm/year, and quite poor if it exceeds 1 mm/year.

Finally, an example of an empirical model describing the time to failure for a corroding system will be discussed. The process is described by a power-law defect depth growth model [29], which has been found to give a reasonable description of progressive corrosion in oil pipelines. This model assumes that a defect with initial depth x_0 is initiated by some event at time t_0. This defect will then grow due to corrosion until a critical depth x_{tol} is reached. A further assumption is that the corrosion proceeds with a rate κ (in mm/yr) that governs the growth of the relative depth $X = x/x_0$ as follows:

$$\frac{dX}{dt} = \frac{\kappa}{x_0} X^q \tag{4.64}$$

The growth rate exponent q and the value of κ must be determined from experiments for a specific application. The time to failure for the component can then be derived [29] to be

$$t_f = t_0 + \frac{x_0}{(q-1)\kappa}\left[1 - \left(\frac{x_0}{x_{\text{tol}}}\right)^{q-1}\right] \tag{4.65}$$

for values of q unequal to 1. This example shows that collecting experimental data on corrosion may enable the definition of an empirical relation that may be more generally applicable than for the considered case only. This approach is very similar to the empirical models for electrical failures discussed in Sect. 4.9.6.

4.10.2 Corrosion Mechanisms

Although the basic principle of the considered corrosion mechanisms is an electrochemical reaction that yields material loss, the various mechanisms do have their own characteristics and are active at different conditions. The most common corrosion mechanisms will now be described briefly.

4.10.2.1 General Corrosion

General corrosion is the most common corrosion mechanism, where the surface of the metal is generally attacked at a more or less constant rate. A typical example of general corrosion is atmospheric corrosion that is observed on uncoated carbon

Fig. 4.43 Formation of corrosion pits on an austenitic stainless steel surface (published with kind permission of TNO)

steel structures that are exposed to the atmosphere. The characteristic brown surface layer is the iron oxide formed in this corrosion reaction. The constituents of the atmosphere that are responsible for this type of corrosion are oxygen, water (humidity) and carbon dioxide. Other constituents present in the atmosphere, such as chlorides (from seawater) or sulphur dioxide (from fuel combustion), may enhance the corrosion process.

Due to its constant degradation rate (and thus predictability) and high detectability, general corrosion is not considered to be a critical form of corrosion. Typical corrosion rates depend on the type of atmosphere: 0.01–0.05 mm/yr for land conditions, 0.05–0.1 mm/yr in a maritime atmosphere and 0.1–0.25 mm/yr in industrial conditions [32]. The local corrosion mechanisms discussed in the next subsections are much more unpredictable and thus more dangerous.

4.10.2.2 Pitting Corrosion

A local breakdown of a passive surface layer may yield pitting corrosion. While the majority of the surface is still passive, pits are formed locally (see Fig. 4.43), which can grow into the material and eventually may cause penetration of a structure or part. Especially, stainless steels are susceptible to this corrosion mechanism, since their corrosion resistance depends on the stability of the passive surface layer. Note that the breakdown of the passive film may be initiated by a mechanical load, for example, scratching of the surface by a sharp tool.

As the corrosion products formed during the pitting generally form a barrier between the pit and the electrolyte, the local environment inside the pit becomes more aggressive and the corrosion rate increases. Pitting corrosion is thus a dangerous mechanism, as the location and moment of pit initiation is unpredictable and the degradation rate may be very high.

Fig. 4.44 Crevice corrosion under a weld in austenitic stainless steel (published with kind permission of TNO)

4.10.2.3 Crevice Corrosion

This mechanism occurs in the small volumes in between two parts, for example, between fasteners like bolts, rivets, washers or nuts and the mating surfaces. Figure 4.44 shows an example of crevice corrosion under a weld. In atmospheric conditions, water may be retained in this small crevice, while the remaining surface is dry. Moreover, as for pitting corrosion, the corrosion products are not removed from the crevice and thus form an aggressive (e.g. very acid) environment in which corrosion rates are very high.

4.10.2.4 Galvanic Corrosion

Galvanic corrosion occurs when two dissimilar materials are in (electric) contact and are both exposed to an electrolyte. The metal with the lowest electrode potential (see Table 3.3) will corrode relatively fast, while the second material will be protected from corrosion. The severity of galvanic corrosion depends on the difference in electrode potential between the two materials and on the ratio of the anode and cathode surface areas. A small piece of an active alloy in contact with a large piece of a more noble alloy will corrode quite fast, but if the large piece of alloy would be the active material, the process would be much slower. Finally, also the magnitude of the area that is in contact with the electrolyte determines the corrosion rate, as the electrolyte is responsible for the transfer of charged ions from cathode to anode.

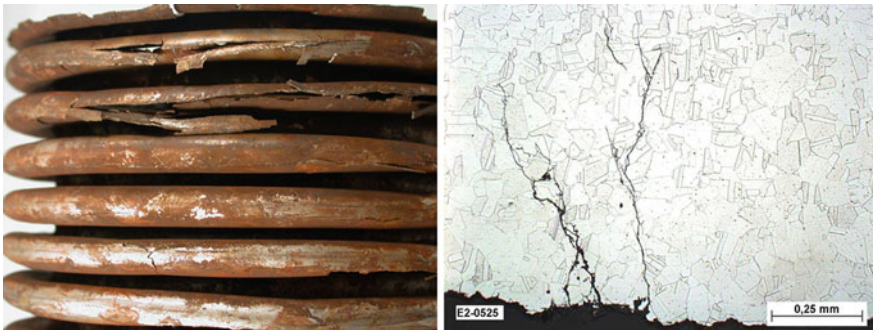

Fig. 4.45 Stress corrosion cracking in austenitic stainless steel component (*left*) and a micrograph showing the development of the cracks (*right*) (published with kind permission of TNO)

4.10.2.5 Stress Corrosion Cracking

In this mechanism, a combination of a mechanical load, that is, a tensile stress, and a chemical load ultimately lead to failure. The corrosion process is accelerated by the mechanical stress, which can be either externally applied or be an internal residual stress due to heat treatments or welding. The mechanism only occurs for certain combinations of alloy types and environments. The most important example is the cracking of stainless steels in chloride solutions at elevated temperatures. The initiation of a crack takes the majority of the time, and after initiation, the cracks can propagate at rates in the order of millimetres per hour. The actual rate depends on the magnitude of the mechanical stress, temperature and severity of the environment (e.g. chloride concentration). An example of stress corrosion cracking is shown in Fig. 4.45, where also a micrograph of the developing cracks is presented.

4.10.2.6 Microbiologically Influenced Corrosion

The final corrosion mechanism discussed here is influenced by micro-organisms. These organisms, which are present in almost all water and soil on the earth, can form biofilms on metal surfaces, which seals the surface from the surrounding fluid. Underneath the biofilm, very specific and generally aggressive environments can develop which cause high corrosion rates. This form of corrosion does not produce a unique type of corrosion. Most MIC is localized, producing pitting, crevice or under deposit corrosion and de-alloying. In addition, it enhances galvanic and erosion corrosion.

However, the micro-organisms can also actively be involved in the corrosion process. This is, for example, the case with the classes of sulphate-reducing bacteria and sulphur-oxidizing bacteria, which accelerate the corrosion process. In general, the result of microbiologically influenced corrosion (MIC) is relatively

deep pits in the metal surface for systems where otherwise no localized corrosion would be expected. This mechanism is typically active on underground storage tanks and in ballast tanks in ships.

Many more, mostly very specific, corrosion mechanisms exist. Their treatment is beyond the scope of this book, so the reader is referred to specialized books on this topic [31–33].

4.10.3 Corrosion Prevention

Now the principles and mechanisms of corrosion have been discussed, the various measures to prevent corrosion can rather easily be derived. As for the prevention of any failure, the solution is either to reduce the load or to increase the capacity of the system.

The load on a corroding system can be reduced by changing the environmental parameters like temperature and humidity. But also inhibitors are commonly used in recirculating systems (e.g. cooling or heating systems) to reduce the aggressiveness of the environment inside the system. Inhibitors are based on various mechanisms, but generally reduce the acidity of fluids or control the amount of dissolved oxygen, which reduces the corrosivity. Finally, for parts susceptible to stress corrosion, the load can be reduced by decreasing the mechanical stress in the part.

Increasing the capacity of the system can be achieved by selecting materials with a higher corrosion resistance, that is, a more noble metal or an alloy that forms a passive surface layer. An alternative solution is the application of a coating that protects the material against corrosion. In general, the function of organic coatings (i.e. paints) is purely the formation of a physical barrier to the corrosive environment, although in some cases inhibitors are embedded in the coating. Metallic coatings may in addition provide sacrificial cathodic protection. The application of a more active material to the surface of a metal ensures the sacrificial degradation of the coating, thus protecting the base material against corrosion. Galvanized zinc coatings applied to steel surfaces for protection against atmospheric corrosion are a well-known example of this type of coatings.

Finally, another effective and widely applied measure for corrosion prevention is cathodic protection. In cathodic protection, the direction of a corrosion reaction is reversed, which switches a material that would normally corrode from the anodic to the cathodic side of the reaction. This can be achieved by the application of an electric current that provides cathodic polarization. The associated lowering of the potential yields a reduction in the anodic reaction rate and thus decreases the corrosion rate. This method is widely applied to pipeline systems and underground storage tanks. An alternative approach for cathodic protection is the connection to a sacrificial anode. The anode of a more active metal will degrade and thus protect the system it is connected to. Zinc anodes attached to steel ship hulls are an example of this type of cathodic protection.

4.10.4 High-Temperature Oxidation and Hot Corrosion

The corrosion mechanisms treated in the previous subsections mostly are associated with materials in contact with aqueous solutions. Only at atmospheric corrosion, no direct contact with a fluid is present, but the condensation of water from humid air does provide the fluid electrolyte.

However, also in dry air, corrosion can occur, although an elevated temperature is required then to activate the reactions. A typical oxidation reaction is given by

$$x\mathrm{M} + \frac{y}{2}\mathrm{O}_2 \rightarrow \mathrm{M}_x\mathrm{O}_y \tag{4.66}$$

The oxide $\mathrm{M}_x\mathrm{O}_y$ forms a surface layer that may protect the underlying material against further oxidation. However, the stability and properties of the oxide layer depend on the type of material and environmental conditions. The most widely adopted kinetics law for oxide film growth is the parabolic rate law [31], stating that the square of the film thickness x is proportional to the time t

$$x^2 = k_f t \tag{4.67}$$

where k_f is the film growth rate constant. The background of this parabolic equation is the diffusion process transporting ions across the oxide scale from the oxide-gas interface to the oxide-metal interface and vice versa. The growth rate of the oxide layer therefore depends on the diffusivity of the oxide (D) and the ion concentration difference across the oxide ($\Delta c/x$):

$$\frac{dx}{dt} = CD\left(\frac{\Delta c}{x}\right) \tag{4.68}$$

Integrating this equation yields the parabolic rate law and the film growth rate constant k_f appears to be equal to $CD\Delta c$, where C is another proportionality constant depending on the specific material. This growth law shows that the degradation rate decreases when a stable oxide layer is formed. In that case, the degradation of the component will be rather limited. However, if a non-adhering oxide scale is formed, no protective layer will be present and the degradation rates will be significantly higher than predicted by the parabolic rate law. Alloying elements that enhance the formation of stable oxide scales can considerably increase the oxidation resistance of alloys. Elements like chromium (Cr), aluminium (Al) and silicon (Si) form protective oxide layers like chromia (Cr_2O_3), alumina (Al_2O_3) or silica (SiO_2) on the metal surfaces. These elements can either be added to the alloy itself or be applied to the metal surface in metallic coatings.

In addition to oxidation that is caused by oxygen, also other gases may cause degradation of metal surfaces. Nitridation, carburization and sulphurization are common degradation process caused by nitrogen, carbon (dioxide) and sulphur, respectively. These processes occur in specific environments, for example, in combustion engines or gas turbines, and on specific alloys.

Finally, the mechanism of hot corrosion will be briefly addressed. Hot corrosion is a special form of high-temperature corrosion that occurs in deposited molten salts. It is a common problem in combustion engines, where sulphur from the fuel forms a salt (e.g. Na_2SO_2 or K_2SO_2). The molten salts are deposited on the components in the vicinity of the combustion process and thereby enhance the corrosion process. The attack rates for hot corrosion are typically an order of magnitude higher than for corrosion processes that are active in the absence of the molten salts.

4.11 Radiative Failures

Failures due to radiative loads are typical for systems that are exposed to cosmic radiation, such as spacecraft, satellites and high-altitude aircraft, or are operating close to other radiation sources, like nuclear reactors or particle accelerators. Further, especially electronic components and integrated circuits are susceptible to this type of load. Basically, two main mechanisms can be distinguished [34].

Particles like alpha particles, beta particles and neutrons, but also very high-energy electromagnetic radiation (e.g. γ-rays), generally have sufficiently high energies to damage the crystal lattice of electronic components. Particularly in semiconductor junctions, the rearrangement of the atoms in the lattice may lead to degraded or undefined behaviour of the device.

Radiation with somewhat lower energy is not able to displace atoms, but causes ionization inside the material. This effect is mostly transient (recoverable soft errors), but could lead to failure if other mechanisms are triggered. Also the performance of the devices may be temporarily decreased.

As for many electric failures, the failure processes associated with radiative loads can hardly be quantified, since many factors determine the precise effects. The main drivers are the type of radiation, the total dose and the radiation flux, but also the operating conditions of the electronic device (e.g. operating voltage) may play a role. Measures to prevent radiative failures include the proper shielding of devices against radiation and application of materials that are less susceptible to radiation (e.g. insulating substrates for chips instead of silicon wafers).

4.12 Failure Processes

In the previous sections in this chapter, a range of failure mechanisms has been discussed, that can each individually lead to failure of a part or system. In addition to those solitary mechanisms, also processes or mechanisms exist that do not directly lead to failure, but may initiate other failure mechanisms. This causes a chain reaction of several mechanism, which is called a failure process. Furthermore,

individual failure mechanism may also interact, which in most cases leads to the acceleration of one or both mechanisms.

In the next subsection, a number of failure sequences will be described, while in Sect. 4.12.2 the interaction between mechanisms will be discussed. Finally, Sect. 4.12.3 will demonstrate how in a rolling element bearing several mechanisms are operative simultaneously, but that their severity depends on the operating conditions.

4.12.1 Failure Sequences

Two common failure sequences will be discussed in this section. The first concerns the (in) stability of a system and the associated failures. The second process is fatigue failure due to vibrations.

4.12.1.1 Stability

The loss of stability of a system is a common cause of failures in different applications. In this subsection, three examples will be discussed.

When a ship looses its stability, for example, due to extremely high waves or damage to the ship, it will generally capsize. This loss of stability initiates a variety of failure mechanisms that eventually may lead to failure of the complete ship (e.g. sinking). The failure mechanisms that can be expected for a capsizing ship are excessive deformation of the hull or other structural parts and static failure (overload) of the structure, as revealed by the development of cracks or even complete fracture of the hull.

Another example of instability leading to failure is the buckling of a structural part loaded in compression. When a compressive load is applied to a slender beam or plate, the part will first start to deform elastically. But when the load is increased even more, at some point, the part becomes instable and an out-of-plane deformation will occur, which is called buckling. This phenomenon can be experienced by compressing a soft-drink can. Up till a certain magnitude, the load can be carried by the can, but suddenly, it will collapse and deform significantly. The magnitude of the load at which the instability occurs is called the buckling load P_{cr}, for which Euler derived the following expression

$$P_{\mathrm{cr}} = \frac{\pi^2 EI}{L^2} \tag{4.69}$$

The critical load depends on the stiffness of the material, as expressed by the elastic modulus E, and the dimensions of the part: length L and moment of inertia I. It is clear that the instability of the structure initiates failure mechanisms like deformation and static overload that will rapidly lead to failure.

The final example of an instability-based failure process concerns systems that are governed by a control system. The control system is aimed to operate the system within certain limits and thus prevents excessive loads on the system to occur. However, such a control system can become instable at certain conditions, which means that the controlled system can move into extreme operating conditions. That often leads to mechanical, thermal or electrical overloading of the system, which may lead to failure.

4.12.1.2 Vibrations and Fatigue

Every dynamic system possesses a natural frequency (f_0). If the system is excited with a frequency close to f_0, the system will start to resonate and large vibrations will develop. The vibrations may lead to mechanical loads (e.g. bending stresses) which eventually cause fatigue failure. The high-cycle fatigue mechanism as discussed in 4.4.3 is a well-known example of this process. The initiation of the vibration is often caused by another failure or event, for example, a foreign object damaging a part in an aero-engine that affects the natural frequency of the part.

4.12.2 Interaction Between Failure Mechanisms

Interaction between different failure mechanisms mostly yields accelerated failure. However, in some cases, the interaction is beneficial to the service life of the component. For both types of interactions, examples will be discussed next.

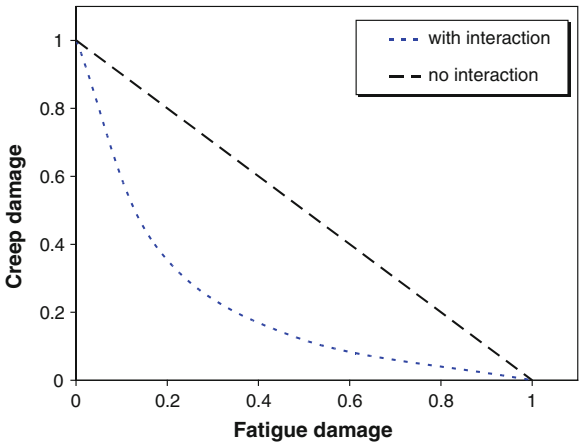

Fig. 4.46 Combinations of creep and fatigue damage that lead to failure, both without and with interaction between the mechanisms

4.12.2.1 Creep and Fatigue

In systems operating at high and variable temperatures, the creep and fatigue mechanisms will quite often act simultaneously. An example of such a system is the hot section of a gas turbine. In Sects. 4.4 and 4.5, the two separate mechanisms have been treated and it was shown how the creep (D_{cr}) and fatigue (D_f) damage can be calculated using Robinson's and Miner's rule, respectively. If no interaction between the mechanisms would be present, the total amount of damage is obtained by a simple addition of the individual contributions

$$D_{tot} = D_{cr} + D_f \tag{4.70}$$

and the resulting life time N (either in terms of cycles or units of time) is given by

$$N = \frac{1}{D_{tot}} \tag{4.71}$$

Plotting the individual damage contributions D_f versus D_{cr} yields a straight line that represents all combinations of creep and fatigue damage leading to failure (i.e. $D_{tot} = 1$), as is shown in Fig. 4.46.

In practice, however, interaction between the mechanisms appears to be present, which makes that the total amount of damage at any instance is larger than the summation of the two individual contributions:

$$D_{tot} = D_{cr} + D_f + f(D_{cr}, D_f) \tag{4.72}$$

where the function f depending on both D_f and D_{cr} describes the interaction. Plotting the $D_{tot} = 1$ line in this case yields the curved line that is also shown in Fig. 4.46. This clearly demonstrates that the interaction in this case reduces the component's service life.

Fig. 4.47 Pump impeller damaged by interaction of erosion and corrosion (published with kind permission of TNO)

4.12.2.2 Corrosion and Fatigue

Corrosion can considerably accelerate the fatigue process. On the one hand, this is caused by corrosion pits acting as initiation locations for fatigue cracks. On the other hand, corrosion can considerably enhance the crack propagation rate due to the formation of corrosion products on both faces of the crack. This affects the SIF K (see Sect. 4.4.6) and thus causes an acceleration of the crack growth.

4.12.2.3 Erosion and Corrosion

The combination of a corrosive fluid and a high flow velocity along a metal surface results in erosion–corrosion. The corrosive fluid causes the degradation of the metal and the formation of a surface layer of corrosion products. In the absence of the high flow velocity, this surface layer would reduce the corrosion rate. However, the flow will erode and remove the surface layer, which exposes the bare metal surface to the corrosive fluid again. An example is shown in Fig. 4.47.

4.12.2.4 Fatigue and Overload

An example of a beneficial interaction between failure mechanisms is the occurrence of an overload during a fatigue crack growth process. As was discussed in Sect. 4.4.6, the high load causes a large plastic zone to develop in front of the crack. A series of smaller load cycles after the peak load will have much smaller plastic zones, which means that the crack has to grow through the large zone of the peak load. Generally, this reduces the crack growth rate, which is known as crack retardation. This means that a crack in a structure which encounters a peak load every now and then may be smaller than a crack in a similar structure that only undergoes regular loading.

4.12.3 Case Study: Failure Mechanisms in a Ball Bearing

Rolling element bearings are relatively simple systems that are applied in a large variety of applications. To suit each specific application, a range of different bearing types are available, such as ball bearings and cylindrical bearings, with or without cage and with single or double rings.

The external load of a bearing consists of the radial force that is exerted on the bearing by the axle or shaft the bearing is attached to. This external force yields

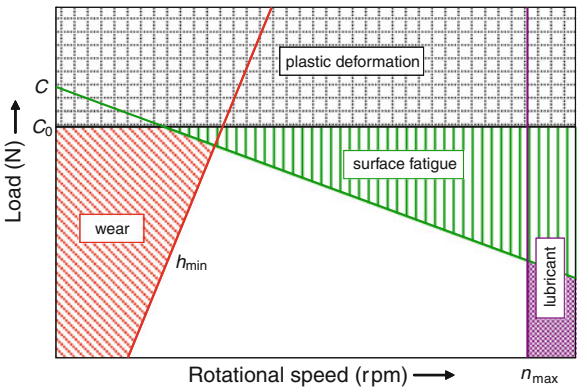

Fig. 4.48 Operating boundaries for a cylindrical bearing

stresses at several locations in the bearing. Since the bearing is a rotating system, the location of maximum stress will move continuously, driven by the rotational speed of the axle and bearing.

To identify all possible failure mechanisms in such a system, generally a system decomposition is executed, which can by either a functional (based on the different functions of the system) or physical decomposition (based on the different physical parts of the system). This decomposition is also the first step of a failure mode and effects analysis (FMEA), as will be discussed in more detail in Chap. 8. The decomposition provides insight into the possible failures.

For a rolling element bearing, this provides the following list of possible failure mechanisms:

- plastic deformation
- surface fatigue
- lubrication problems
- abrasive and adhesive wear

All these failure mechanisms are related to the usage and the loading of the bearing. The two governing parameters are therefore the rotational speed and the force that is exerted on the bearing by the axle. All failure mechanisms are only critical (i.e. lead to failure within the specified service life of the bearing) for a specific range of these two usage parameters, as is shown in Fig. 4.48.

The hatched regions indicate the operational conditions that will yield reduced life times due to the specified failure mechanism, so the lines bounding these regions together define the operating envelope.

For example, adhesive or abrasive wear will occur when the rotational speed is too low to build up a lubricant film that is sufficiently thick to completely separate the rolling elements from the rings and prevent metal to metal contact. Clearly, the required critical film thickness (and thus the required rotational speed) increases if the load on the bearing increases. A very high rotational speed, on the other hand, will generate too much heat and will yield burning or degradation of the lubricant.

The diagram in Fig. 4.48 thus clearly demonstrates how in a real system different failure mechanisms can be active simultaneously, but that the specific operating conditions determine which mechanism is critical and thus will limit the service life time of the complete system.

4.13 Summary

In the present chapter, the most common failure mechanisms have been described. Both an explanation of the underlying physical mechanisms and quantitative relation to asses the life times have been provided for

- Mechanical failures: static overload, deformation, fatigue, creep, wear
- Thermal failures: melting, thermal degradation
- Electric failures: intrinsic breakdown, current overload
- Chemical failures: corrosion, high-temperature oxidation, hot corrosion
- Radiative failures

In addition to these single failure mechanisms, the final section has addressed failure processes and interaction between different mechanisms.

References

1. Schijve, J.: Fatigue of Structures and Materials, 2nd edn. Springer, London (2009)
2. Basquin, O.H.: The exponential law of endurance tests. Proc. ASTM **10**, 625–630 (1910)
3. Manson, S.S.: Behavior of materials under conditions of thermal stress. NACA TN 2933. National Advisory Committee for Aeronautics, Cleveland, OH (1953)
4. Coffin, L.F.: A study of the effects of cyclic thermal stresses on a ductile metal. Trans. ASME **76**, 931–950 (1954)
5. Palmgren, A.: Durability of ball bearings. Z. Ver. Dtsch. Ing. **68**, 339–341 (1924)
6. Miner, M.A.: Cumulative damage in fatigue. J. Appl. Mech. **12**, A159–A164 (1945)
7. Schijve, J., Vlot, A.: Damage and Fatigue Crack Growth of Aircraft Materials and Structures. TU Delft, Delft (1996)
8. Janssen, M., Zuidema, J., Wanhill, R.J.H.: Fracture Mechanics, 2nd edn. Delft University Press, Delft (2002)
9. Tada, H., Paris, P.C., Irwin, G.R.: The stress analysis of cracks handbook. Del Research Corporation, St. Louis, USA (1973)
10. Rooke, D.P., Cartwright, D.J.: Compendium of stress intensity factors. The Hillingdon Press, Uxbridge, UK (1976)
11. NASGRO: www.nasgro.com, Southwest Research Institute. San Antonio, Texas (2012)
12. AFGROW: www.afgrow.net, LexTech, Inc. Centerville, OH (2012)
13. ZENCRACK: www.zentech.co.uk, ZenTech International Ltd. London (2012)
14. Tinga, T., Brekelmans, W.A.M., Geers, M.G.D.: Incorporating strain-gradient effects in a multi-scale constitutive framework for nickel-base superalloys. Phil. Mag. **88**(30–32), 3793–3825 (2008)
15. Tinga, T., Brekelmans, W.A.M., Geers, M.G.D.: Cube slip and non-schmid effects in single crystal nickel-base superalloys. Model. Simul. Mater. Sci. Eng. **18**(015005), 1–31 (2010)

16. Larson, F.R., Miller, J.: A time-temperature relationship for rupture and creep stresses. Trans. ASME **74**, 765–775 (1952)
17. Robinson, E.L.: Effect of temperature variation on the long-time rupture strength of steels. Trans. ASME **74**(5), 777–781 (1952)
18. Halling, J. (ed.): Principles of Tribology. The Macmillan Press LTD, London (1978)
19. Van Beek, A.: Advanced Engineering Design. Lifetime Performance and Reliability. TU Delft, Delft (2006)
20. de Rooij, M.B.: Tribological aspects of Unlubricated Deep Drawing Processes. University of Twente, Enschede (1998)
21. Archard, J.F.: Contact and rubbing of flat surfaces. J. Appl. Phys. **24**, 981–988 (1953)
22. Tinga, T., Brekelmans, W.A.M., Geers, M.G.D.: Directional coarsening in nickel-base superalloys and its effect on mechanical properties. Comput. Mater. Sci. **47**, 471–481 (2009)
23. Callister, W.D., Rethwisch, D.G.: Materials science and engineering, 8th edn. Wiley, Hoboken, NJ, USA (2011)
24. Ross, R.: Electrical failure mechanisms. In: Tinga, T. (ed.) Loads and failure mechanisms, part 2: Advanced mechanisms, numerical methods and applications. Netherlands Defence Academy, Den Helder (2010)
25. Pultrum, E., Gulski, E., Riet, M.J.M.V., Ross, R.: On-line partial discharge detection and classification by pattern recognition on HV teminations In: Jicable, Versailles, France 1999
26. Ross, R.: Inception and propagation mechanisms of water treeing. IEEE Trans. Dielectr. Electr. Insul. **5**(5), 660–680 (1998)
27. Ross, R.: Dealing with interface problems in polymer cable terminations. IEEE Electr. Insul. Mag. **15**(1) (1999)
28. Black, J.R.: Electromigration failure modes in aluminium metallization for semiconductor devices. Proc. IEEE **57**(9), 1587–1594 (1969)
29. Hall, P.L., Strutt, J.E.: Probabilistic physics-of-failure models for component reliabilities using Monte Carlo simulation and Weibull analysis: a parametric study. Rel Eng Syst Saf **80**, 233–242 (2003)
30. Pecht, M., Ko, W.C.: A corrosion rate equation for micro-electronic die metallization. Int. J. Hybrid Microelectron. **13**, 41–51 (1990)
31. Jones, D.A.: Principles and prevention of corrosion, 2nd edn. Prentice-Hall, Upper Saddle River (1996)
32. Gellings, P.J.: Inleiding tot corrosie en corrosiebestrijding. Twente University Press, Enschede (1997)
33. Little, B.J., Lee, J.S.: Microbiologically influenced corrosion. Wiley, Hoboken, New Jersey, USA (2007)
34. Holmes-Siedle, A.G., Adams, L.: Handbook of radiation effects. Oxford University Press, Oxford (2002)

Further Reading

1. Callister, W.D., Rethwisch, D.G.: Materials Science and Engineering, 8th ed. Wiley, Hoboken, NJ, USA (2011)
2. Schijve, J.: Fatigue of Structures and Materials, 2nd ed. Springer, London (2009)
3. Janssen, M., Zuidema, J., Wanhill, R.J.H.: Fracture Mechanics, 2nd edition ed. Delft University Press, Delft (2002)
4. Van Beek, A.: Advanced engineering design. Lifetime Performance and Reliability. TU Delft, Delft (2006)
5. Purcell, E.M.: Electricity and Magnetism, 2nd edn. McGraw-Hill, New York (1985)
6. Jones, D.A.: Principles and Prevention of Corrosion, 2nd edn. Prentice-Hall, Upper Saddle River (1996)

Part II
Applications in Maintenance, Reliability and Design

This second part of the book focuses on the application of the knowledge on loads and failure mechanisms. Many of these applications are in the field of maintenance and reliability engineering, since failures of components and systems govern the reliability of a system and the associated maintenance requirements. But also in the design process, knowing the possible future failures is very valuable.

Chapter 5
Maintenance Concepts

5.1 Introduction

In the first part of this book, the basic principles of loads and failure mechanisms have been treated. This second part of the book focuses on the application of the knowledge on loads and failure mechanisms. Many of these applications are in the field of maintenance and reliability engineering, since failures of components and systems govern the reliability of a system and the associated maintenance requirements.

Therefore, in this chapter, the various maintenance concepts, strategies and policies that are used in the field will be discussed. A lot of literature is available on maintenance management, and a detailed treatment of all available concepts and strategies is outside the scope of the present book. However, the summary of the most common concepts given in this chapter will provide the reader with the right context for the topics treated in the next chapters.

In the next subsection, the definitions of maintenance, availability and other related concepts will be discussed. Sections 5.3 and 5.4 will elucidate the distinction between a maintenance strategy or concept and a maintenance policy and describe the various existing concepts and policies. Section 5.5 will address the challenges and different approaches in determining preventive maintenance intervals, and finally in Sect. 5.6, the quantification of maintenance performance will be discussed.

5.2 Relation Between Maintenance and Availability

Already as long as systems of any complexity are produced by man, people are concerned with getting the best performance out of the system for an as long as possible operating period and with lowest possible costs. Maintenance of the

T. Tinga, *Principles of Loads and Failure Mechanisms*,
Springer Series in Reliability Engineering,
DOI: 10.1007/978-1-4471-4917-0_5, © Springer-Verlag London 2013

system, ranging from simple checks of oil level and tire pressure to complete overhauls of complex assets, plays a decisive role in that challenge, where obtaining a high system availability is the main goal.

According to the European standard EN13306 [1], maintenance is defined as 'the combination of all technical, administrative and managerial actions during the life cycle of an item intended to retain it in, or restore it to, a state in which it can perform the required function'. Note that, as for the definition of failure given in Chap. 1, the *required function* of the system plays an important role and should thus be defined properly. Also, the extent to which the system should be able to perform the intended functions, that is, the minimally acceptable performance, has to be defined by the user.

Closely related to the concept of maintenance is the availability of a system. Availability is defined as 'the probability that the system will be ready to perform its specified function, in its specified and intended operational environment, when called for at a random point in time (point availability) or during a stated period of time (interval availability)' [2]. Availability (A) is generally quantified as

$$A = \frac{\text{MTBM}}{\text{MTBM} + \text{MDT}} \tag{5.1}$$

where the mean time between maintenance (MTBM) specifies the (average) lengths of the operational periods, and the mean downtime (MDT) is the time required to maintain the system (both scheduled and unscheduled). The latter generally also includes the logistic delay times. Since the average values are used in this definition, A in principle represents the point availability of the system.

If the operating and downtimes are monitored during a certain period, the achieved interval availability A_p can be obtained as [3]

$$A_p = \frac{\sum \text{up times}}{\sum \text{up times} + \sum \text{down times}} \tag{5.2}$$

Note, however, that the mean values (MTBM and MDT) are generally obtained from the same raw data, so the values of A and A_p are (nearly) identical. The real point availability generally varies in time; Eq. (5.1) only provides the average value. This difference between point availability and interval availability will be discussed in more detail in 5.6.2.

To retain the system in the 'available' state, maintenance should be performed. Moreover, since budgets are generally limited, maintenance should be reduced to a minimum. This will lead to a trade-off between the extent to which the system is able to perform the intended functions and the costs associated with achieving this. Presumably, when more maintenance is carried out, the system will be better able to perform the intended functions. But the availability of a system not only depends on the amount of maintenance that is performed. Also, the inherent reliability of the system, maintainability and serviceability play a role, as is shown schematically in Fig. 5.1 [4].

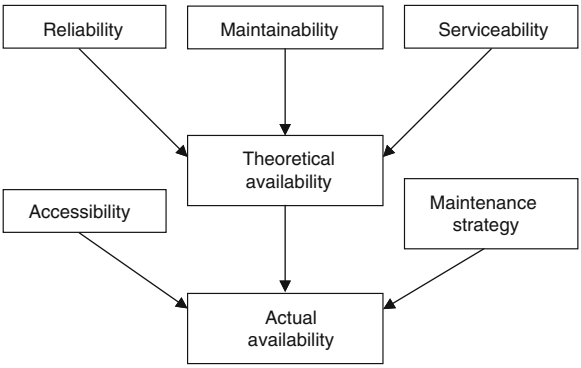

Fig. 5.1 Dependence of system availability on reliability, maintainability and serviceability

The reliability of a system is its ability to perform the required function for a specified period of time under given operating conditions. It is often defined as the probability that the item will not fail before a certain point in time, while no maintenance is performed during that period. The reliability of a system thus heavily depends on both the loads the system is subjected to and its load-bearing capacity. This relation between system reliability and the loads and failure mechanisms discussed in part I will be treated extensively in Chap. 7.

In addition to the reliability, the maintainability affects the availability. The maintainability defines how easily a system can be maintained (i.e. repaired or replaced) in case of a failure. It deals not only with the ease, but also with the accuracy, safety and economy of the performance of maintenance activities. One of the aspects of maintainability is the accessibility of the subsystems and components for maintenance activities, which largely determines the required maintenance times. Figure 5.1 shows that accessibility of the site or asset is posed as an additional factor influencing the availability. This is especially the case when the asset cannot be approached for maintenance tasks during certain (prolonged) periods, as is for example the case for ships, while they are at sea and for certain plants that cannot be accessed for safety reasons.

The final factor influencing the availability is the serviceability or supportability of the system. This concerns designing the system such that it can be supported with minimum (life cycle) costs [2]. Therefore, it should be determined what facilities, equipment and resources (man power, spare parts) are needed when and where and how this can be realized to assure the proper and timely execution of the maintenance tasks. Both the maintainability and supportability of a system are largely determined during the design process. This will be discussed in more detail in Chap. 9.

From Fig. 5.1, it can be observed that the theoretical availability of a system is governed by its reliability, maintainability and supportability, which are determined by both the design and the usage of the system. Moreover, the actual availability of the system depends, in addition to the theoretical availability, also on the applied maintenance strategy and the accessibility of the site or asset.

Finally, even if the system is technically available, most systems also need operators and other supports (e.g. fuel, charged batteries) to be operated. Therefore, the deployability of a system depends on both the (technical) availability and the availability of people and required supports.

From this discussion, it is clear that maintenance plays a decisive role in assuring a certain availability level. Therefore, the design of an effective and efficient maintenance function for an asset is an important process, the result of which is called the maintenance strategy or maintenance concept [5–7]. The maintenance strategy thus determines how the maintenance process will be organized and one of the decisions to be taken concerns the maintenance policies to be applied. Once the maintenance concept has been defined, which commonly takes place at the *strategic* level in a company, it must implemented on both the *tactical* and *operational* level [8]. This means that for each maintenance policy, the associated maintenance activities must be defined. Examples of such activities are inspection, monitoring, routine maintenance, overhaul, rebuilding (modification) and repair. Actions on the tactical level then concern the correct assignment of maintenance resources (e.g. skilled personnel, materials, test equipment) to fulfil the maintenance plan. This mainly entails planning and scheduling activities. Finally, on the operational level, it must be ensured that maintenance tasks are carried out by skilled technicians, within the time scheduled, following the correct procedures and using the proper tools.

In the next section, the different maintenance strategies will be discussed, while Sect. 5.4 will describe the various maintenance policies.

5.3 Maintenance Strategy

As was mentioned in the previous section, the maintenance strategy or maintenance concept is concerned with the design of an effective and efficient maintenance function for an asset and thus determines how the maintenance process will be organized. A lot of definitions and descriptions for maintenance strategy are available in literature [5, 6, 9]. Based on these sources, the following definition is adopted in this work: A maintenance strategy is a mix of maintenance policies, backup equipment, equipment upgrades, and the identification, resource allocation and execution of repair, inspection and replacement decisions. This reflects that a maintenance strategy not only comprises a selection of a proper maintenance policy (see 5.4), but also regards the allocation of facilities and resources.

In the past, management in organizations regarded maintenance as merely a burden of expenses. However, over the course of time, the view on maintenance has changed, and in many companies, it is now seen as a value adding activity. Several concepts have been developed to design the optimal maintenance function in a company. These concepts can be regarded as being wider than the maintenance strategies because they form approaches for organizational design and

sometimes include complete philosophies about how personnel should be motivated. Therefore, these concepts can be seen as a company-wide approach for the design of the maintenance function. However, the relation with maintenance strategies is very close and a maintenance concept is seen by many as a maintenance strategy [10, 11].

Four common maintenance strategies will be elaborated in the next subsections. The first two strategies are reliability centred maintenance (RCM) and risk-based inspection (RBI), which are methodologies to determine which maintenance tasks are most appropriate for a certain asset. The third strategy is integrated logistic support (ILS) and the final strategy that will be discussed is effectiveness centred maintenance (ECM), in which the focus is on the service provided to the customers.

5.3.1 Reliability Centred Maintenance

Reliability Centred Maintenance (RCM) originated in the development of a preventive maintenance plan for the Boeing 747 airplane in the 1960s by the Maintenance Steering Group 1 [11, 12]. In a RCM approach, it is not only considered 'what can be done', but also 'why should it be done'. To facilitate this process, RCM entails asking seven questions on the asset or system under review:

1. What are the functions and associated performance standards of the asset in its present operating context?
2. In what ways does it fail to fulfil its functions?
3. What causes each functional failure?
4. What happens when each failure occurs?
5. In what way does each failure matter?
6. What can be done to predict or prevent each failure?
7. What should be done if a suitable proactive task cannot be found?

The first five steps in the RCM process are in fact a failure mode, effects and criticality analysis (FMECA), a commonly applied analysis method that will be discussed in more detail in Chap. 8. This structured analysis of the asset, which is often facilitated by completing an RCM or FMECA form and using a decision diagram, provides insight into the possible failure modes, their effects and consequences and thus enables to prioritize between different failures.

In the final steps, a suitable maintenance policy is determined for each failure mode. For example, if any evidence can be found that an asset is about to fail, a condition-based policy may be suitable. If no proactive tasks can be identified, the criticality of the asset determines which approach must be followed. For a critical system, a redesign or modification must be performed to prevent the failure, while for non-critical systems, unscheduled (corrective) maintenance is most appropriate.

5.3.2 Risk-Based Inspection

Risk-based inspection (RBI) is a strategy that is commonly applied in the process industry for static containment systems like pipe work and pressure vessels [3]. In this type of installations, the majority of the maintenance tasks concerns non-destructive inspections of the structures. Such periodic condition assessments must guarantee the structural integrity and thus prevent failures with often large consequences.

The RBI concept is applied to prioritize and plan these inspections based on the estimated risk of failure. Risk in this case is determined by both the probability of occurrence of a failure and the consequences of a failure in terms of costs, safety and environmental effects. Both aspects of risk are quantified (in most cases in a subjective way, see also Sect. 8.2 on FMECA analysis) and the multiplication provides the risk number which is used in the prioritization of the inspections.

5.3.3 Integrated Logistics Support

Integrated logistics support (ILS) is based on the life cycle approach, which means that already during the design phase, decisions are made regarding the maintenance activities and logistic support during the operational phase. This strategy is therefore much broader than only maintenance and also includes the definition of required inventories of spare parts, required test and support equipment, personnel, training and support and the gathering and storage of technical data through ICT facilities.

The ILS concept originates from the military, but is now also applied in many other sectors. It is based on the principles of systems engineering [13] and applies two quantitative tools to optimize the maintenance and logistics support: RAMS and LCC. A RAMS analysis contains the quantification of reliability, availability, maintainability and system safety. An LCC analysis quantifies the life cycle costs, that is, all the costs that occur during the complete life cycle, ranging from investment costs, maintenance and logistic support costs to the costs of decommissioning. The life cycle approach indicates the integral aspect of ILS since a higher initial investment (e.g. the purchase of a condition monitoring system) could lead to a better maintainability of the system, lower maintenance costs and therefore lower life cycle costs.

5.3.4 Effectiveness Centred Maintenance

Effectiveness centred maintenance (ECM) focuses on the service provided to the customer. When a system has failed, the only concern of the user is when the

system is working again. This means that the focus of the customer is on the availability of the service instead of the defect rectification time [14]. The four key elements of the ECM approach are (1) the active participation of personnel, (2) an improvement in the quality of equipment, (3) the development of a maintenance strategy and (4) the introduction of a performance measurement system.

By focusing on the effectiveness of maintenance, 'doing the right things' will prevail above 'doing things right'. The ECM approach includes core concepts of quality management, total productive maintenance (TPM) and also RCM. Moreover, ECM and TPM are regarded as people-oriented methodologies, while RCM is seen as an asset-oriented methodology.

5.4 Maintenance Policies

As a large variety of different maintenance policies exists, it is convenient to classify the policies. A suitable distinction has been proposed by Swanson [15], defining three basic policies for maintenance: reactive, proactive and aggressive. Reactive strategies are described as a fire-fighting approach to maintenance, where maintenance activities are provoked by actual failures. Proactive strategies, on the other hand, are developed to prevent system breakdowns using a range of methods to monitor or predict equipment deterioration and undertake minor preventive tasks to restore equipment to a proper condition. Preventive maintenance encompasses maintenance activities that are undertaken after a specified period of time or amount of machine use. Aggressive policies aim at improving the system by modifications or redesign, which also assist to prevent failures.

These three basic policies cover practically all more specific maintenance policies that have been proposed until today. An overview is provided in Fig. 5.2. Note that a lot of classifications have been made, but none of them being as exhaustive and consequent as the classification in Fig. 5.2. For example, in RCM [11], four basic maintenance policies are distinguished: predictive, preventive, corrective and detective maintenance. However, Fig. 5.2 shows that predictive maintenance effectively is a variant of preventive maintenance. The addressed maintenance policies will now be explained in some more detail.

5.4.1 Reactive Maintenance Policies

The two reactive policies are corrective and detective maintenance. Corrective maintenance is characterized as fixing and/or replacing components either when they have failed or when they are found to be failing. There are no interventions until a failure has occurred. The advantage of this policy is that the service lifetime of parts and components is fully utilized, that is, no remaining lifetime is spoiled by replacing components before the actual failure. Detective (i.e. failure-finding)

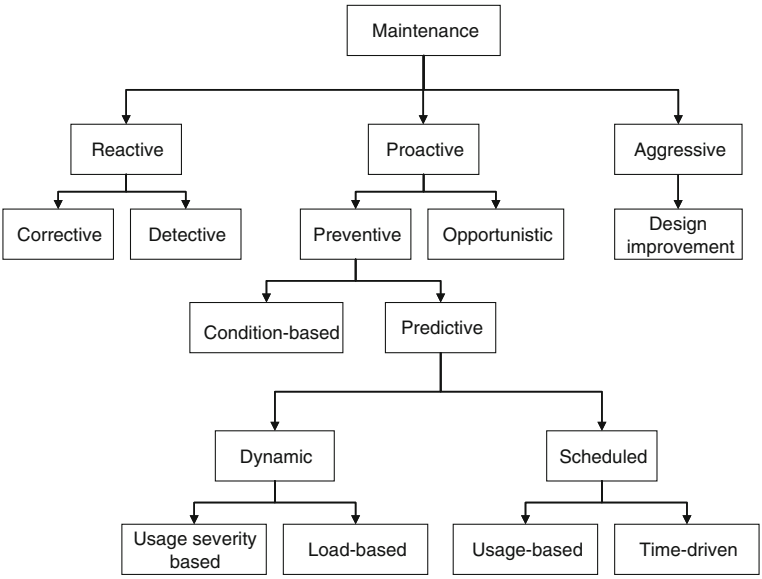

Fig. 5.2 Overview and classification of different maintenance policies

maintenance only applies to hidden or unrevealed failures, which usually only affect protective devices like fire alarms [16]. The device is repaired or replaced only when a (periodic) functional test reveals that it has failed. The actual failure may have occurred long before that moment, but has not triggered a maintenance task.

5.4.2 Proactive Maintenance Policies

The proactive policies can be divided into preventive and opportunistic maintenance policies. In an opportunistic policy, maintenance tasks on a specific (sub)system are triggered by other tasks performed on the same (sub)system. Although the (sub)system does not require the task yet, the clustering of several tasks provides advantages in terms of, for example, combined transport costs. Also, time benefits can be obtained due to the fact that preparative tasks (e.g. removal of covers or hatches) only have to be performed once.

Preventive maintenance policies aim at replacing or repairing components or systems before failure occurs. The resulting prevention of failures is very important for critical systems like aircraft and nuclear plants, but also consequential damage can be prevented. The disadvantage of this policy is that the optimal moment of replacement is often hard to determine. To be on the safe side, components are then often replaced far before the end of their service life, which yields a spoil of component lifetime and consequential high costs. The determination of

the optimal replacement intervals is one of the major challenges in the maintenance field. This challenge will be introduced in Sect. 5.5 and discussed in detail in Chaps. 6 and 7. Based on the method applied to determine the optimal interval, the preventive policies can be subdivided into a condition-based policy and predictive policies.

In condition-based maintenance, the condition of the system is monitored either periodically by inspections or real time by using appropriate sensors. If the monitoring system detects an abnormal situation or the monitored parameter exceeds a predefined threshold, a maintenance task is triggered. In that way maintenance can be performed when it is actually necessary. More details on this policy and the methods deployed to assess the system condition will be discussed in the next chapter (Sect. 6.5). The remaining predictive policies represent other methods to predict the optimal intervals for maintenance. These can be either intervals that are fixed in time (static) or intervals that vary in time (dynamic). The policies with fixed intervals are often called scheduled maintenance, and the length of these intervals is defined in terms of either calendar time or a usage parameter like operating hours or driven kilometres. The dynamic maintenance policies [17] can be both usage (severity) based or load based. All preventive maintenance policies will be discussed in more detail in Sect. 5.5.

5.4.3 Aggressive Maintenance Policies

The aggressive maintenance policies aim at improving the equipment to reduce the number of failures. This is generally achieved by a modification or redesign of the system. A well-known example of an aggressive policy is TPM, which focuses on continuous improvement of the system. The improved component or system performance results in less maintenance needed.

5.5 Preventive Maintenance Interval Determination

The key issue of the preventive methods is the determination of the maintenance intervals, that is, the timing of the various maintenance activities like repair or replacement of parts. If the intervals are too large, failure will occur, while too small intervals lead to over maintenance: The service life of parts is only partially utilized, and the amount of labour hours is unacceptably high. However, the best compromise between these two counteracting processes is not the same for every part or system. The criticality of the component determines whether the interval tends to the *effective* (short) or the *efficient* (long) side. In the former case, where components are critical, large safety factors are applied to avoid failure.

As was discussed in the previous section, several approaches can be followed to determine the preventive maintenance intervals. In Fig. 5.3, a categorization of

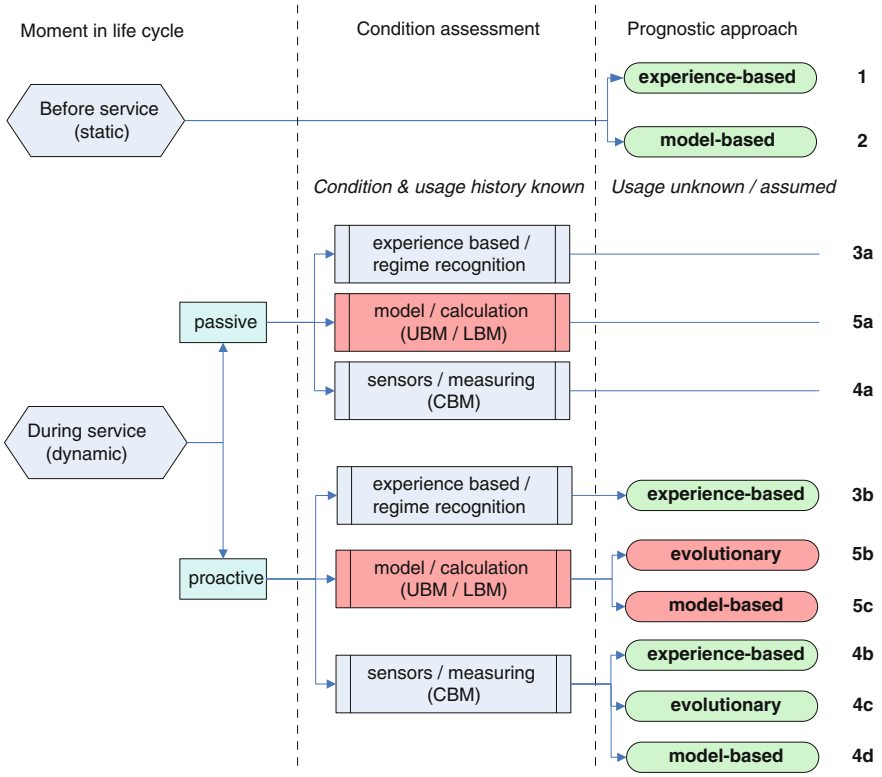

Fig. 5.3 Classification of preventive maintenance policies

these policies is given [17], using three criteria: (1) the moment in the system life cycle at which the intervals are determined, (2) the way in which the system condition is assessed during the service life and (3) the prognostic approach that is followed.

In the present section, these criteria will be discussed and existing policies and methods from literature will be categorized. The rather new concepts of usage and load-based maintenance (UBM and LBM), as indicated by the blocks 5a, b and c in Fig. 5.3, will be treated in detail in Chap. 6. These concepts are closely linked to the principles of loads and failure mechanisms as treated in part I of this book.

5.5.1 Moment in Life Cycle

The first criterion used in Fig. 5.3 to distinguish the various policies is the moment in the life cycle at which the intervals are determined. Traditionally, the manufacturer quantifies the interval during the design phase of the system or component,

using assumptions on the future usage. This leads to a *static* policy in which fixed intervals are applied during the complete service life of the system, disregarding any variations in usage. In the previous section, these policies have been denoted scheduled maintenance policies. More recently, *dynamic* maintenance policies are being developed, where the actual usage or system degradation is taken into account and the required maintenance intervals are regularly updated or even fully determined during the service life. The latter policy can be applied in a *passive* way, where components are replaced or repaired shortly after the condition of the system reaches a critical level. In a more *proactive* variant, a prediction is made for the remaining useful life (RUL) after every condition assessment, which provides more time to plan and prepare the repair or replacement. Note that in this classification, contrary to the scheme in Fig. 5.2, condition-based maintenance is considered as a predictive policy and the main division in preventive/predictive policies is between passive (wait until failure appears to be imminent) and active policies (try to predict when future failures will occur). Condition-based maintenance can be applied in either way, as will be explained later.

5.5.2 Condition Assessment

The second criterion is the method used to determine the system condition during the service life. Obviously, for the static policies (1 and 2 in Fig. 5.3), no condition assessment is performed at all, since the intervals are fully determined before the system enters service. For the dynamic policies, there are currently two approaches.

The most commonly used method is condition monitoring (policy 4 in Fig. 5.3), where appropriate sensors or inspection techniques are used to assess the system condition [18–21]. This can be done in a direct or indirect manner. The indirect method monitors performance parameters, like the flow in a pump, and applies these as an indication for the condition of the system. In the direct method, sensors are installed that directly monitor the condition of the system or component. Examples are delamination sensors in composite structures, crack length sensors, sensors to detect metal particles in lubrication oil and vibration monitoring systems. The difference between indirect and direct methods has been described by denoting the methods as ‘based on process data’ and ‘based on failure data’ [22], or as condition *indicators* (*RCI*) and *predictors* (*RCP*), respectively [23]. Condition monitoring and condition-based maintenance will be treated in detail in Chap. 6.

An alternative method to estimate the system condition is based on the correlation between certain usage profiles and the resulting system degradation (policy 3 in Fig. 5.3). This method is often applied in rotorcraft health and usage monitoring systems (HUMS), where it is called flight regime recognition [24, 25]. For each flight regime, past experience enables the attribution of a relative damage severity. By monitoring the usage, both the present condition and the remaining

life can be estimated. For aircraft engines, similar algorithms have been used fairly extensively [26]. Cycle counting algorithms monitor the number of preset variations in rotational speed or temperature, while exceedance monitoring is applied to register the number of times a certain threshold value has been exceeded. In both cases, engine manufacturers developed (experience-based) algorithms to relate the usage profiles to component life consumption. Finally, also the terrain identification algorithm presented by Heine and Barker [27] belongs to this category. By identifying the terrain type in which a military vehicle operates, an indication of the usage severity is obtained and the component degradation rate can be estimated. These type of usage and load-based maintenance policies will be treated in Sect. 6.4.

5.5.3 Prognostic Approach

The final criterion is the prognostic approach that is followed. In the passive dynamic concept, no prognostic method is used and maintenance is executed shortly after the condition of the system reaches a critical level. For the static concept, the prognostic method is used to calculate the intervals for the complete service life, based on the assumed future usage of the system. And for the proactive dynamic concepts, the prognostics are used to predict the RUL from the moment the condition is assessed. Initially, this prediction will contain considerable uncertainty, but as the end of the service life is approached, the uncertainty is reduced [19], as will be discussed in more detail in Sect. 6.2.

In the last decade, the field of prognostics has attracted considerable research attention [19–21, 28–32], partly boosted by the development of prognostics and health management (PHM) systems for future military aircraft [33, 34]. The hierarchy of prognostic methods as proposed by Lebold and Thurston [35] and Roemer et al. [20, 31], see Fig. 5.4, is used here to categorize the methods. The three basic prognostic approaches are discussed in some more detail: experience-based prognostics, evolutionary methods and model-based prognostics. In that order, the methods enable increasing levels of accuracy, but at the cost of increasing complexity and development efforts.

5.5.3.1 Experience-based Prognostics

The most simple and lowest level prognostic methods are based on historical service failure data and are therefore called experience-based methods. They require a numerical relation describing the data. Parametric families of distributions, like exponential and Weibull, have been used as failure functions for decades now. The field of reliability engineering is based on these failure distributions and many maintenance modelling approaches start from these functions.

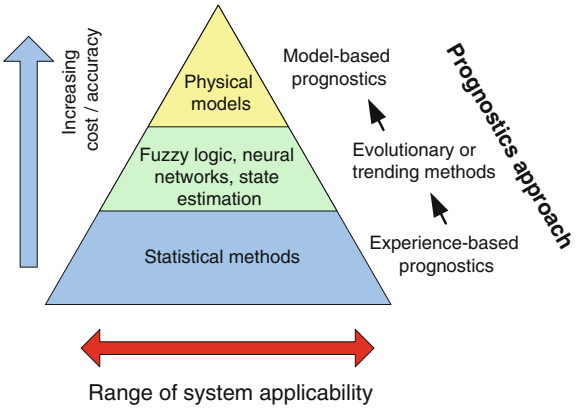

Fig. 5.4 Hierarchy of prognostic approaches [35]

However, these approaches rely on the collection of a sufficiently large set of service failure data. Consequently, during the design phase, when no service failures have occurred yet, these methods can only be applied when test programs on prototypes are performed. During the operational phase, the collection of failure data for a large set of similar systems can quickly yield a useful failure distribution and associated lifetime prediction. However, in many cases, the amount of failure data is quite limited, either due to a small number of systems, or due to the fact that no failures are allowed for safety reasons (e.g. critical parts in aircraft). In those cases, it is very difficult to determine an accurate numerical relation.

Moreover, the relation is based on historical data, which means that a prediction is only accurate when the future usage of the system is similar to the past usage. Due to the lack of an explicit relation between usage and lifetime, these methods are mainly applied in static policies (1 in Fig. 5.3). The prediction of remaining life during service (prognostics) is difficult with these methods, since generally only data about the complete service life is available.

However, Engel and Hess [19, 32] proposed a method in which condition monitoring is used to update the original failure distribution, as will be discussed in Sect. 6.2. Also, the regime recognition type policies (3 in Fig. 5.3), where usage profiles are monitored, are examples of dynamic experience-based concepts. If sufficient experience is available on the relation between usage profiles and service life, a rough estimate of the remaining life may be obtained with these methods. This challenge will be discussed in Chap. 7, where the knowledge on failure mechanisms will be integrated with existing reliability engineering methods.

5.5.3.2 Evolutionary Methods

Evolutionary or trending methods can be applied for prognostics in those cases where the condition of the system is assessed on a regular basis. By extrapolating the trends in measured condition (or at least features that can be correlated to the

condition), the RUL can be predicted. Roemer et al. [20, 31] demonstrated that these methods can be used to predict the effects of fouling on gas turbine compressor performance. As can be observed in Fig. 5.3, this prognostic approach can be used in combination with condition monitoring. Also, the usage and load-based policies can be combined with evolutionary methods, but then the *calculated* evolution of the system degradation is extrapolated.

5.5.3.3 Model-based Prognostics

The most sophisticated prognostic approach is the physical model-based approach. In these methods, the degradation process is simulated using physical models of the component and its failure mechanism(s). As has been demonstrated in part I of this book, an extensive knowledge base is available on these mechanisms and a lot of research has been done in the areas of structural integrity and failure mechanisms [36–42]. However, although the potential of model-based prognostic methods is widely recognized [18, 28, 43, 44], applications in the field of maintenance modelling and prognostics are limited. Roemer and co-workers [20, 21, 30, 31, 45–47] have presented a series of papers in which several physical models are applied to predict failures of bearings, gear tooth, electronic systems and a high power clutch system. Also, the crack severity index proposed by De Jonge and Lummel [48] is based on a physical (crack propagation) model.

The physical methods have a number of clear benefits when compared to the previously discussed prognostic approaches. Firstly, these methods do not rely on the collection of large sets of (failure) data. Once a physical model of the failure mechanism is available, knowledge of the material properties and local loads (as a function of usage) are sufficient to determine the component service life. This means that the method can already be applied in the design phase, where the assumed usage and loads can be used to approximate the service life. Since the usage and loads are not fully known yet, a life prediction with a large uncertainty is obtained.

Secondly, as opposed to the statistical and stochastic methods discussed before, the present method is not based on historical data. Since the quantitative relation between usage and degradation is known, changes in future usage can easily be incorporated in the prognostic analyses. This characteristic offers an opportunity for optimization of maintenance strategies that has not been fully utilized yet, as will be demonstrated in Chap. 6.

5.6 Maintenance Performance

Once a suitable maintenance strategy has been selected and implemented, it is important to assess how well the strategy performs. Indications that the strategy is not performing as well as expected may trigger a modification or even the set-up of

a new strategy. The performance of maintenance can be regarded on several levels. It can be restricted to the maintenance department only or can include the external effectiveness, which measures the environment influenced by the maintenance department. Moreover, the performance can be limited to a single system or may focus on all maintained systems in a company or the complete fleet of assets. In this section, a number of aspects of maintenance performance measurement will be discussed and a calculation method will be provided.

The basic assessment in a maintenance performance measurement concerns the comparison of the achieved performance and the effort made to achieve that. For example, the maintenance performance associated with the sustainment of a fleet of (military) jet engines can be measured by calculating the ratio of percentage engine availability to the total financial and operational costs consumed in achieving this availability at a specific time interval [49]. These two aspects can be referred to as maintenance effectiveness and maintenance efficiency [11]. Maintenance effectiveness defines how well maintenance ensures that an asset is able to perform its intended function. Maintenance efficiency concentrates on how well the resources for maintenance are used. These two factors together constitute the maintenance performance (MP)

$$\text{MP} = \text{Efficiency} \times \text{Effectiveness} \tag{5.3}$$

To be able to quantify the maintenance performance, it is necessary to evaluate the effectiveness and efficiency. For both aspects, parameters are available. The efficiency can be quantified by the maintenance costs (labour, spares and materials) and planning and control, while maintenance effectiveness is determined by aspects like frequency of failures (reliability), lifespan, downtime and the probability of failure in the next period (dependability).

Another important aspect of a performance measurement is that the method must not be based solely on historic information, but should also caption expectations for future performance. This means that both lagging and leading indicators should be included in a proper performance measurement method. For example, availability can be seen as a result of the policy of the previous periods and can therefore be regarded as a lagging indicator for effectiveness. Therefore, a leading indicator should be added to the maintenance effectiveness. It will be shown later that a quantification of the system health is a suitable leading parameter for effectiveness.

5.6.1 Measurement Methodology

Based on the requirements discussed in the previous subsection, a methodology is presented now to assess the maintenance performance, see Fig. 5.5. The basic parameter to quantify the effectiveness of the maintenance strategy is the system availability. It represents the fraction of time the system is available to fulfil its

intended function and thus indicates how well (on average) the maintenance strategy succeeded to prevent failures or to repair failed parts. But as was mentioned before, availability is a lagging indicator, so 'system health' is added as a leading indicator for effectiveness.

The health parameter quantifies the present state of the system, which actually has a close relation to availability. If the system is healthy, a high availability can be achieved rather easily and vice versa. However, a high availability is not a direct effect of good health. Health indicators are used to measure the extent of deterioration or degradation the system has suffered. The purpose of these types of measures is to prevent the asset from being ruined by neglect, because only the urgently needed activities are undertaken [50]. Furthermore, availability focuses on the short-term aspects of the maintenance performance, while system health is a long-term performance measurement.

The health parameter is not defined as the current system health because that would mean that the maintenance performance steadily decreases over the years (the system health generally deteriorates during the system's service life). Therefore, it is formulated as the relative change in system health over the past period. It thus represents the effect of the policy over the last period, but remains a leading indicator because it provides information about the expected performance in the (near) future. As was mentioned before, availability and health are moderately coupled parameters, as is also indicated in Fig. 5.5. Health can be kept high when the availability is low (no usage of the system) or the other way around: The availability is high at the expense of the health (cannibalism of the system).

The efficiency parameters describe what costs are associated with achieving a certain level of effectiveness. If this parameter would be merely based on the expenditures (costs) in the past period, it would only represent policies from the past. It should therefore be complemented by an indicator that provides information about future performance. This is realized by including the parameter 'scheduling'. Scheduling represents the grip on the process, grip on the inventory (are the required items available in stock?), grip on the hours needed for maintenance and the number of work orders that is completed in time. It is important to measure this because there is evidence [51] that one-third of maintenance costs is unnecessarily spent due to overtime costs, bad use of preventive maintenance and bad planning.

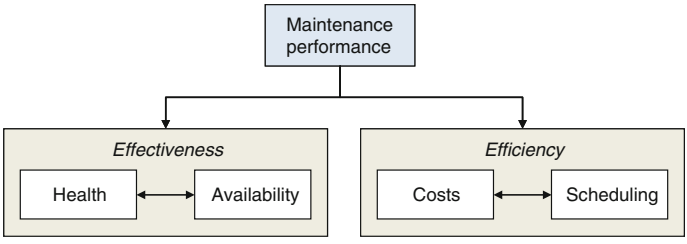

Fig. 5.5 Maintenance performance measurement methodology

5.6.2 Maintenance Performance Indicators

The parameters to be measured for the maintenance performance calculation have been defined in the previous subsection. This subsection will elaborate on the calculation and definition of these parameters. As is warned for in literature on this topic, a pitfall of performance indicators is that people are encouraged to do only what is measured. Moreover, people should not be able to manipulate the performance indicators. For example, it is easy to achieve a high adherence to schedule by scheduling less work. This could be realized by overestimating the amount of work associated with specific work orders. However, the real purpose is a higher productivity, which can often be achieved by challenging people through scheduling more work.

The four parameters availability, health, costs and scheduling will be discussed now in more detail.

5.6.2.1 Availability

The availability quantifies the fraction of the time the equipment is able to fulfil its intended function. It thus represents the effectiveness of the maintenance process to prevent or quickly remedy failures. The availability can be calculated in numerous ways. The most commonly applied variant is the achieved availability, which already was defined in Sect. 5.2 as

$$A = \frac{\text{MTBM}}{\text{MTBM} + \text{DT}} \tag{5.4}$$

where the MTBM specifies the lengths of the operational periods, and the downtime (DT) is the time required to maintain the system (both scheduled and unscheduled). The latter generally also includes the logistic delay times.

However, several considerations in obtaining the underlying data and calculating the availability must be taken into account. Firstly, it should be considered whether the availability calculation should be based on the total available time or only on the time that the system is actually needed. In other words, if a system is only used during day time, does a maintenance task (which yields downtime) executed during the night affect the system availability? Generally this would not be considered as a decrease of availability, so the expression in (5.4) should be modified to

$$A = \frac{\text{MTBM}}{\text{MTBM} + \text{DT}_{\text{op}}} \tag{5.5}$$

where DT_{op} represents only the downtime that occurs during operational periods. This illustrates that smart planning of the maintenance activities (in non-operational periods) can increase the system availability and thus the maintenance

performance. At the same time, using this definition of availability requires a decision on whether or not a scheduled maintenance period is considered as operational time. If it is, this type of maintenance negatively affects the availability, which is not the case when such a period is defined to be a non-operational period.

Secondly, it should be decided whether the availability number should represent the average fleet wide availability or the availability of an individual asset (or subset of assets). If the total MTBM and downtime for all assets in the fleet is summated over a certain period, the average availability is obtained. However, the availability of an individual asset can be quite different from that average, depending on the variation in usage of the assets and the resulting variation in average time to failure. The more detailed the availability calculations are, the more information on the maintenance performance is obtained. For example, it might be possible to compare different maintenance teams within one company.

Thirdly, a suitable length of the period over which the availability is calculated should be selected. If the period is too long, the calculated (interval) availability is the average value over a long period, which does not provide much insight into the maintenance performance. There might be some short periods with a very low system availability (this is essentially a point availability), but these are not signalled because they are compensated for in the remainder of the period.

Fourthly, a decision must be made on the accountability of the maintenance organization for user related failures. Must failures due to misuse of the system be taken into account in calculating the availability, which then automatically decreases the maintenance performance? If it is decided that user related failures should not be included in the calculation, the associated downtime must be subtracted from the total downtime (DT) that is used in the calculations.

Finally, it should be considered to relate the achieved availability to the required availability. If the requested availability is 70 %, but an availability of 80 % is realized, this can be regarded as a waste of maintenance resources. Too much availability (unnecessary availability) can be considered as overproduction. Therefore, in the calculation of the maintenance performance, the achieved availability should be compared to the requested availability, using the following equation:

$$P_{\text{availability}} = 100\,\% - \left|A_{\text{real}} - A_{\text{req}}\right| \tag{5.6}$$

In this expression, A_{real} is the realized availability [calculated with e.g. (5.4)], and A_{req} is the required availability. Since the absolute value of the difference between these two quantities is used, both a value higher and lower than the required availability yields a decreased effectiveness. This is visualized in Fig. 5.6, where the required availability is 70 %.

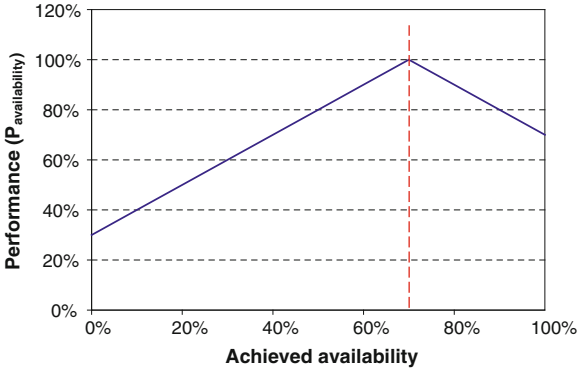

Fig. 5.6 Variation of performance parameter for different achieved availabilities at a required availability of 70 %

5.6.2.2 Health

Health classifies the ability of the system to perform its functions. On the moment of purchase (of a new system), the health will be 100 %. During the service life, the health will usually decrease. However, during modifications, system updates or mid-life updates, the health can increase again. The health of the system can be measured as the total remaining lifetime of the system and should signal any abnormal change in the health of the system. This could also mean that at some stage, the health is higher than 100 %, for example, immediately after a mid-life update.

The health parameter (P_{health}) is defined as the difference in health, in terms of RUL, between the current and the previous period relative to the planned deterioration of the system in that period

$$P_{\text{health}} = \frac{\Delta\text{RUL}_{\text{planned}}}{\Delta\text{RUL}_{\text{actual}}} \tag{5.7}$$

The decrease should be calculated per period. The planned decrease of the system remaining life per period can be estimated as

$$\Delta\text{RUL}_{\text{planned}} = \frac{\text{SL} - \text{RUL}_{\text{end}}}{N} \tag{5.8}$$

where SL is the planned service life of the asset, RUL_{end} is the RUL of the asset at the end of the operational period, and N is the number of periods. If the asset is to be sold when the operational period has ended, it is important that the asset has not been exhausted, but that a certain amount of remaining life is left.

Note that the actual assessment of the remaining life of the asset at any moment may be difficult. The estimation of the remaining life can be made in a subjective or objective manner. A subjective assessment means that an expert estimates the remaining life. An objective way of assessing the remaining life is the use of a condition monitoring technique. It may be clear that the latter method is preferable, but may not be feasible for each asset.

5.6.2.3 Costs

The costs of maintenance constitute the first parameter quantifying the efficiency part of maintenance performance. To obtain a dimensionless parameter, the actual costs are related to the planned costs in the period that is assessed. The planned costs, however, must be derived from objective numbers (like system investment costs) and should be kept constant during the complete system life cycle. Otherwise, the maintenance performance could easily be manipulated by altering the planned costs, which immediately would make comparison with other periods impossible. If the planned costs of maintenance are derived from the (investment) value of the asset, then it is also possible to benchmark with other assets adopting the same approach. Thus, the following expression is used to quantify the cost parameter

$$P_{\text{cost}} = \frac{C_{\text{planned}}}{C_{\text{actual}}} \tag{5.9}$$

where C_{planned} represents the planned maintenance costs and C_{actual} the realized costs. The planned costs can be defined as a certain fraction α of the depreciation of the system in the specific period, given by

$$C_{\text{planned}} = \alpha \frac{C_{\text{investment}} - C_{\text{residual}}}{N} \tag{5.10}$$

where $C_{\text{investment}}$ represents the initial investment costs, C_{residual} the residual value of the asset at the end of the operational life and N the number of periods. The value of the factor α that is considered to be appropriate for the maintenance costs relative to the value of the asset should be determined by experts.

The proposed expression is based on the total costs and is therefore not very specific. This can be enhanced by using different types of costs, for example, costs for corrective and preventive maintenance. Also, a differentiation into labour costs, costs of spare parts and indirect costs (e.g. facilities) could be made. The more the costs are specified, the more insight is gained in the background of badly or well performing maintenance organizations, but at the same time, more effort must be taken to collect the required data.

5.6.2.4 Scheduling

The final parameter used to calculate the maintenance performance is scheduling, which also specifies the efficiency of the process. The scheduling parameter quantifies the amount of control on the maintenance process, which can be measured by the following aspects: (1) the contribution of the planned maintenance activities to the total maintenance time; (2) the fraction of the (maintenance) work orders that is completed within the scheduled period of time; (3) the fraction of the

required items that is available in stock. By combining these three aspects, the following expression is obtained for the scheduling parameter

$$P_{\text{scheduling}} = \frac{1}{3}\left(\frac{\Delta t_{\text{planned}}}{\Delta t_{\text{total}}} + \frac{n_{\text{wo in time}}}{n_{\text{wo}}} + \frac{n_{\text{ri in stock}}}{n_{ri}}\right) \tag{5.11}$$

where $\Delta t_{\text{planned}}$ and Δt_{total} are the time spent on planned maintenance and the total time for maintenance, respectively, and n_{wo} and n_{ri} are the number of work orders and the number of items (i.e. spare parts) requested from the logistic system. The parameter $n_{\text{wo in time}}$ specifies how many work orders have been completed in time, and finally $n_{\text{ri in stock}}$ specifies how many items were in stock at the moment they were requested.

Whereas the cost parameter is a lagging parameter indicating the costs spent in the past, the scheduling parameter is a leading parameter. Although data from the past is used to quantify, the parameter specifies how well controlled the maintenance process is, which provides an indication for the future performance.

5.6.2.5 Maintenance Performance

Finally, the total maintenance performance number can be calculated by multiplying the four individual effectiveness and efficiency parameters according to

$$\text{MP} = P_{\text{availability}} \cdot P_{\text{health}} \cdot P_{\text{costs}} \cdot P_{\text{scheduling}} \tag{5.12}$$

As all the individual parameters range from 0 to $\sim$ 1.5 (0–150 %), also the maintenance performance will attain a value in that range. A maintenance performance of 1 (100 %) indicates that exactly the performance that was anticipated has been realized. A value lower than 1 indicates an underperforming organization.

However, depending on the asset and the type of organization, the different aspects of maintenance performance might not be equally important. In critical applications, like aerospace systems or nuclear plants, where the consequences of failure are significant, the effectiveness of maintenance is generally valued higher than its efficiency. But in a commercial production unit with non-critical installations, the efficiency aspect is more important.

Similarly, the leading (long-term) and lagging (short-term) parameters might be valued differently. For example, in a military organization, the efficiency and long-term parameters are important in peace time and during training periods, but during a deployment, the effectiveness and short-term parameters are much more important: The systems must be available now, regardless the costs and future maintenance performance.

To express these variations in importance of the four parameters, weight factors w_1 to w_4 are added as exponents to the expression for the maintenance performance

$$\text{MP} = \left(P_{\text{availability}}\right)^{w_1} \cdot \left(P_{\text{health}}\right)^{w_2} \cdot \left(P_{\text{costs}}\right)^{w_3} \cdot \left(P_{\text{scheduling}}\right)^{w_4} \tag{5.13}$$

If all weight factors are equal to 1, this expression reduces to the original Eq. (5.12). Values of w_i larger than unity will increase the relative weight of that specific parameter in the maintenance performance assessment.

Example 5.1 (Maintenance Performance Calculation) A maintenance performance calculation will be performed for a fleet of 120 vehicles, for which the basic data are provided in Table 5.1.

The reference numbers for the performance calculation can be partly derived from the data in Table 5.1, and the additional figures provided in Table 5.2. And finally, the achieved results for the past quarter (e.g. Q1 2011) are provided in Table 5.3.

Using this data, the four performance parameters and the overall maintenance performance for this period can be calculated. This calculation yields the following results:

$P_{\text{availability}}$	$1 - \left\lvert \frac{86{,}400}{86{,}400+1{,}200} - 0.95 \right\rvert$	96.37 %
P_{health}	$\frac{0.25}{0.28}$	89.29 %
P_{cost}	$\frac{75{,}000}{70{,}000}$	107.14 %
$P_{\text{scheduling}}$	$\frac{1}{3}\left(\frac{1{,}100}{1{,}200} + 0.95 + \frac{750}{800}\right)$	93.47 %
Maintenance performance (MP)		*86.17 %*

Note that in this calculation, equal weight factors ($w_i = 1$) are used for the four parameters. If for some reason, availability would be more important in this organization (or in this time period), setting $w_1 = 1.5$ would lead to a lower performance, that is, MP = 84.59 %. If costs would be more important, setting $w_3 = 1.5$ would lead to a higher performance, that is, MP = 89.20 %.

Table 5.1 Basic data for a fleet of vehicles

Vehicle fleet basic data	
Total number of vehicles	120
Expected lifetime (years)	40
Planned usage period (years)	30
Number of periods per year	4
Investment costs per vehicle	€ 200,000.00
Expected residual value	€ 50,000.00

Table 5.2 Reference data (per period) for the fleet of vehicles

Reference data	
Fraction (α) of the depreciation planned for maintenance	50 %
Planned maintenance costs	€ 75,000.00
Required availability	95 %
Planned deterioration per vehicle (years)	0.25
Scheduled maintenance hours	1100

Table 5.3 Results for period Q1 2011 for the complete fleet of vehicles

1st quarter 2011	
Total maintenance costs	€ 70,000.00
Operational time: nominal 8 h each day (hours)	86,400
Downtime during operational time D_{op} (hours)	1,200
Average lifetime reduction of the vehicles (years)	0.28
Unscheduled maintenance hours	100
Fraction of work orders completed in time	95 %
# items requested available in stock	750
Total # items requested	800

5.7 Summary

In this chapter, an overview has been given of the basic definitions and concepts of maintenance. Firstly, the definitions of maintenance, availability, reliability, maintainability and serviceability have been provided, and the relations between these concepts were introduced. Then, the different available maintenance strategies and policies have been discussed. After this introduction in the field of maintenance, the challenge of determining preventive maintenance intervals has been discussed. Finally, the concept of maintenance performance was discussed, including a practical methodology to calculate the actual performance.

References

1. EN13306:2001: Maintenance terminology, European standard. In: CEN (European Committee for Standardization), Brussels (2001)
2. Kumar, U.D.: Reliability, Maintenance and Logistic Support; A Life Cycle Approach. Kluwer Academic Publishers, Norwell (2000)
3. Smit, K.: Onderhoudskunde. VSSD, Delft (2010)
4. Bussel, G.J.W., Zaaier, M.B.: Reliability, availability and maintenance aspects of large-scale offshore wind farms a concepts study. In: Imare conference 2001, pp. 119–126
5. Kevin, F.G., Penlesky, R.J.: A framework for developing maintenance strategies. Prod. Inventory Manage. J. (First Quarter), pp. 16–21 (1988)
6. Pintelon, L., Pinjana, S.K.: Evaluating the effectiveness of maintenance strategies. J. Qual. Maintenance Eng. **12**(1), 7–20 (2006)

7. Pintelon, L., Pinjana, S.K.: Bridging the gap between manufacturing and maintenance. In: Oper Manage Change Agent **2**, 587–596 (2004)
8. Crespo Marquez, A.: The Maintenance Management Framework. Models and Methods for Complex Maintenance. Springer Series in Reliability Engineering. Springer-Verlag, London (2007)
9. Kelly, A.: Maintenance Organization and Systems. In: Business centered maintenance. Business Centered Maintenance, Oxford (1997)
10. Bevilacqua, M., Braglia, M.: The analytic hierarchy process applied to maintenance strategy selection. Rel. Eng. Syst. Saf. **70**, 71–83 (2000)
11. Moubray, J.: Reliability-centered maintenance. Industrial Press, New York (1997)
12. Nowlan, F.S., Heap, H.F.: Reliability-Centered Maintenance, vol. AD-A066579. United States Department of Defence (1978)
13. Blanchard, B.S., Fabrycky, W.J.: Systems Engineering and Analysis, 3rd edn. Prentice Hall International, London (1998)
14. Pun, K.F., Chin, K.S., Chow, M.F., Lau, H.C.W.: An effectiveness-centred approach to maintenance management: a case study. J. Qual. Maintenance Eng. **8**(4), 346–368 (2002)
15. Swanson, L.: Linking maintenance strategies to performance. Int. J. Prod. Econ. **70**, 237–244 (2001)
16. Tsang, A.H.C.: Strategic dimensions of maintenance management. J. Qual. Maintenance Eng. **8**(1), 7–39 (2002)
17. Tinga, T.: Application of physical failure models to enable usage and load based maintenance. Rel. Eng. Syst. Saf. **95**(10), 1061–1075 (2010)
18. Jardine, A.K.S., Lin, D., Banjevic, D.: A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mech. Sys. Sig. Proc. **20**, 1483–1510 (2006)
19. Engel, S.J., Gilmartin, B.J., Bongort, K., Hess, A.: Prognostics, the real issue involved with predicting life remaining. Paper presented at the IEEE Aerospace Conference, Big Sky, Montana
20. Byington, C.S., Roemer, M.J., Kacprzynski, G.J., Galie, T.: Prognostic enhancements to diagnostic systems for improved condition-based maintenance. In: IEEE aerospace conference, Big Sky, Montana 2002, pp. 1–11
21. Orsagh, R., Brown, D., Roemer, M.J., Dabney, T., Hess, A.: Prognostic health management for avionics system power supplies. In: IEEE aerospace conference, Big Sky, Montana 2005, pp. 1–7. IEEE
22. Veldman, J., Wortmann, H., Klingenberg, W.: Typology of condition based maintenance. J. Qual. Maint. Eng. **17**(2), 183–202 (2011)
23. Saranga, H.: Relevant condition-parameter strategy for an effective condition-based maintenance. J. Qual. Maint. Eng. **8**(1), 92–105 (2002)
24. Molent, L.: A unified approach to fatigue usage monitoring of fighter aircraft based on F/A-18 experience. In: ICAS98, Melbourne 1998, pp. 1–11. International Council of the Aeronautical Sciences
25. Hunt, S.R., Hebden, I.G.: Validation of the Eurofighter Typhoon structural health and usage monitoring system. Smart Mater. Struct. **10**, 497–503 (2001)
26. Fraser, K.F.: An overview of health and usage monitoring systems (HUMS) for military helicopters. In: Defence science and technology organisation, vol. DSTO-TR-0061. Melbourne, p. 24 (1994)
27. Heine, R., Barker, D.: Simplified terrain identification and component fatigue damage estimation model for use in a health and usage monitoring system. Microelectron. Reliab. **47**, 1882–1888 (2007)
28. Farrar, C.R., Lieven, N.A.J.: Damage prognosis: the future of structural health monitoring. Philos. Trans. Roy. Soc. A **365**, 623–632 (2006)
29. Vichare, N., Pecht, M.: Enabling electronic prognostics using thermal data. In: THERMINIC 2006, Nice, France 2006. TIMA

30. Kalgren, P.W., Baybutt, M., Ginart, A., Minella, C., Roemer, M.J., Dabney, T.: Application of prognostic health management in digital electronic systems. In: IEEE aerospace conference, Big Sky, Montana (2007). IEEE
31. Roemer, M.J., Byington, C.S., Kacprzynski, G.J., Vachtsevanos, G.: An overview of selected prognostic technologies with application to engine health management. In: ASME Turbo Expo, Barcelona, Spain 2006, pp. 1–9. ASME
32. Hess, A., Calvello, G., Frith, P., Engel, S.J., Hoitsma, D.: Challenges, issues, and lessons learned chasing the "big P": real predictive prognostics part 2. In: IEEE aerospace conference, Big Sky, Montana 2006, pp. 1–19. IEEE
33. Hess, A., Calvello, G., Frith, P.: Challenges, issues, and lessons learned chasing the "big P": real predictive prognostics part 1. In: IEEE Aerospace Conference, Big Sky, Montana 2005, pp. 1–10. IEEE
34. Brown, E.R., McCollom, N.N., Moore, E.E., Hess, A.: Prognostics and health management. A data-driven approach to supporting the F-35 Lightning II. In: IEEE Aerospace Conference, Big Sky, Montana 2007, pp. 1–12. IEEE
35. Lebold, M., Thurston, M.: Open standards for condition-based maintenance and prognostic systems. In: 5th annual maintenance and reliability conference (2001)
36. Dasgupta, A., Pecht, M.: Material failure mechanisms and damage models. IEEE Trans. Reliab. **40**(5), 531–536 (1991)
37. Tinga, T.: Stress intensity factors and crack propagation in a single crystal nickel based superalloy. Eng. Fract. Mech. **73**, 1679–1692 (2006)
38. Tinga, T., Brekelmans, W.A.M., Geers, M.G.D.: Time-incremental creep-fatigue damage rule for single crystal Ni-base super alloys. Mat. Sci. Eng. A **508**, 200–208 (2009)
39. Tinga, T., Brekelmans, W.A.M., Geers, M.G.D.: Incorporating strain-gradient effects in a multi-scale constitutive framework for nickel-base super alloys. Phil. Mag. **88**(30–32), 3793–3825 (2008)
40. Pecht, M., Ko, W.C.: A corrosion rate equation for micro-electronic die metallization. Int. J. Hybrid Microelectron. **13**, 41–51 (1990)
41. Homborg, A.M., Tinga, T., Zhang, X., van Westing, E.P.M., Oonincx, P.J., de Wit, J.H.W., Mol, J.M.C.: Time-frequency methods for trend removal in electrochemical noise data. Electrochim. Acta **70**, 199–209 (2012)
42. Engel, P.: Failure models for mechanical wear modes and mechanisms. IEEE Trans. Reliab. **42**(2), 262–267 (1993)
43. Zio, E.: Reliability engineering: old problems and new challenges. Rel. Eng. Syst. Saf. **94**, 125–141 (2009)
44. Uckun, S., Goebel, K., Lucas, P.J.F.: Standardizing research methods for prognostics. In: International conference on prognostics and health management, Denver, Colorado 2008, pp. 1–10. IEEE
45. Orsagh, R., Roemer, M.J., Sheldon, J., Klenke, C.J.: A comprehensive prognostic approach for predicting gas turbine engine bearing life. In: IGTI Turbo Expo, Vienna 2004, pp. 1–9. ASME
46. Orsagh, R., Sheldon, J., Roemer, M.J., Klenke, C.J.: Prognostics/diagnostics for gas turbine engine bearings. In: IEEE aerospace conference, Big Sky, Montana 2005, pp. 1–9. IEEE
47. Watson, M., Byington, C.S., Edwards, D., Amin, S.: Dynamic modelling and wear-based remaining useful life prediction of high power clutch systems. In: ASME/STLE international joint tribology conference, Long Beach, USA 2004, pp. 1–12. STLE
48. Jonge, J.B.D., Lummel, C.W.J.: Development of a crack severity index (CSI) for quantification of recorded stress spectra. In: TR 86049 L, vol. TR 86049 L. National Aerospace Laboratory, Amsterdam, (1986)
49. Adamides, E.D., Stamboulis, Y.A., Varelis, A.G.: Model-based assessment of military aircraft engine maintenance systems. J. Oper. Res. Soc. **55**(9), 957–967 (2004)
50. Vos, A., Andela, C., Kool, G., Silfhout, G.V., Habets, G., Koevoets, K., Mulder, M., Kempen, P.V.: Managing MRO—growing towards successful PBL contracting. World Class Maintenance, Breda (2011)
51. Wireman, T.: World class maintenance management. Industrial Press, New York (1990)

Further Reading

1. Crespo Marquez, A.: The maintenance management framework. Models and methods for complex maintenance. Springer Series in Reliability Engineering. Springer-Verlag, London (2007).
2. Kumar, U.D.: Reliability, Maintenance and Logistic Support; A Life Cycle Approach. Kluwer Academic Publishers, Norwell (2000)
3. Moubray, J.: Reliability-centered Maintenance. Industrial Press, New York (1997)

Chapter 6
Usage- and Condition-Based Maintenance

6.1 Introduction

In the previous chapter, an overview of maintenance policies was provided. It has been indicated that for any preventive policy, the key issue is the determination of the appropriate maintenance interval. In the present chapter, it will be demonstrated how knowledge on the physical failure mechanisms and monitoring of the system usage or condition can be combined to optimize the maintenance process.

To properly understand this challenge, in the next section, the origin of the uncertainty associated with the moment of failure will be discussed. Also, the benefits of reducing this uncertainty by applying physical failure models will be illustrated. Then, Sect. 6.3 will be devoted to model-based prognostics, one of the prognostic approaches that has been introduced in Sect. 5.5 and that can serve as a way of reducing uncertainty in maintenance interval prediction.

After this general treatment of prognostics, several maintenance concepts will be discussed that are based on the failure mechanisms and associated prognostic methods. These contain on the one hand the widely applied condition-based maintenance (CBM) policies, but on the other hand also the rather new usage (severity)-based and load-based maintenance (LBM) policies that were mentioned in the previous chapter. Section 6.4 will introduce the concepts of load- and usage-based maintenance. The development and application of these policies will be demonstrated using several case studies. Finally, Sect. 6.5 will treat condition monitoring and condition-based maintenance. The several available condition monitoring techniques will be presented, and the application of these techniques to enable CBM will be discussed. Also, the extension of diagnostic methods with a prognostic part will be treated, and the concept of structural health monitoring (SHM) will be discussed.

T. Tinga, *Principles of Loads and Failure Mechanisms*,
Springer Series in Reliability Engineering,
DOI: 10.1007/978-1-4471-4917-0_6, © Springer-Verlag London 2013

6.2 Uncertainty in Preventive Maintenance

Optimal application of a preventive maintenance policy requires determining the right moment to perform the maintenance tasks. Since the degradation rate of most systems depends on a lot of factors, it is seldom exactly known. Therefore, the optimal maintenance interval is associated with a certain amount of uncertainty. The exact origin of this uncertainty and they ways to cope with the uncertainty are the topic of this section [1], which extends the work by Hess, Engel and co-workers [2, 3] on this subject.

The procedure for the calculation of the service life is illustrated in Fig. 6.1, where the values of an arbitrary damage parameter D (e.g. a crack length, amount of wear) for a population of components at three moments in time (t_1, t_2 and t_3) are represented by three normal distributions. These distributions are typically obtained from a probabilistic analysis, where the variation in model parameters is incorporated in the life assessment of a part (see Sect. 7.6), which yields a service life distribution instead of one deterministic value for the service life of the part. The distribution at, for example, $t = t_1$ indicates that a considerable fraction of the population contains an amount of damage close to the average amount of damage at that time instance. However, some components have accumulated considerably more damage, while another set of components degraded much slower and thus has a D value that is much lower at $t = t_1$. The latter set of components apparently is loaded less severely than the other parts or has operated in a milder environment.

The uncertainty in the damage parameter, as represented by the width of the distributions in Fig. 6.1, is caused by the three following sources:

1. the actual usage of the individual components varies (e.g. hours per week, ratio of low power/high power operating hours)
2. uncertainty in the effect of usage on (internal) loads (for example, a certain rotational speed in a system not always causes exactly the same stress level in the rotating parts)

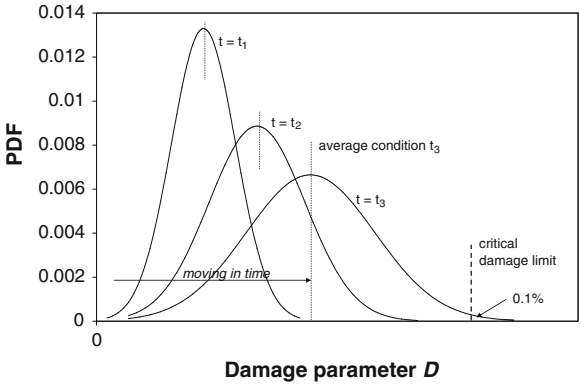

Fig. 6.1 Evolution of the damage parameter (D) distribution in time. The component must be replaced when a specified fraction (e.g. 0.1 %) of the distribution has reached the critical damage limit

3. variations in the life consumption for a given internal load, for example, caused by variations in material properties or dimensions.

The concepts of usage and loads will be explained in more detail in Sect. 6.3. Since the degradation proceeds in time, the damage gradually increases. In Fig. 6.1, this is represented by the movement of the distribution of D to the right. Simultaneously, the width of the distribution increases, since the *relative* uncertainties are assumed to be constant. The latter assumption implies that the damage rates for the individual components in a population are assumed to be constant, which means that the difference in damage between components with a high respectively low damage rate increases in time.

As the position of an individual component in the distribution of D is unknown, the replacement interval of the complete population of components is determined by the fastest degrading parts. Thus, the complete set of components must be replaced (or repaired) at the moment the tail of the distribution reaches a critical amount of damage. The criticality of the component determines what probability of failure is acceptable. Critical components will be replaced when only, for example, 0.1 % of the distribution has reached the damage limit, while non-critical components could be replaced when 50 % has reached the limit.

This figure shows that, due to the large uncertainty in the (predicted) condition, the component is already replaced at a time (t_3) at which the average component is far below the critical condition. However, since the usage is unknown, intervals must be calculated based on the most severe usage, which clearly results in conservative maintenance intervals. As a consequence, replacing the complete set of components at t_3 yields a considerable spill of remaining life in many components.

A similar approach was presented by Hess, Engel and co-workers [2, 3], who also studied the effect of uncertainty on the maintenance interval determination. For the point in time where a certain fraction of the population reaches the critical amount of damage, that is, the decision point (t_3 in Fig. 6.1), they introduce the term just-in-time-point (JITP). Further, the time period from the present time to the JITP is called the lead time interval (LTI). It is exactly the length of this LTI that determines whether a preventive maintenance concept is passive (limited or zero LTI) or proactive (sufficiently large LTI), that is, whether or not it contains a prognostic method (see Sects. 5.5 and 6.3). Also, in their work, they distinguish between a priori probability density functions (PDF) for the remaining life at time zero (both true and modelled) and a posteriori PDFs conditioned on observations during component use. Using the first type of PDFs, the service life is determined before the system enters service, while in the latter case the condition is assessed on a regular basis during service and the remaining useful life is determined. Therefore, these two approaches match with the preventive maintenance concepts 1 and 4b in Fig. 5.3, respectively. Moreover, these methods are fully experience based, since the probability density functions are based on historical measurements.

Hess and Engel [2, 3] also state that a balance must be found between preventing failures and avoiding over maintenance, where the JITP is the latest moment in time that satisfies the acceptable probability of failure. It is further

observed that the width of the PDF largely determines the amount of over maintenance, so a distribution as narrow as possible is preferred. It is shown in [2] that the distribution can be narrowed by assessing the system condition during service. Just by noting that the system is still operating at the moment of assessment, all cases in the original PDF with failure times smaller than the present time can be excluded. Then, the remaining subset can be normalized to maintain a cumulative probability of 1.0, which results in a narrower distribution and yields a smaller amount of conservatism (and over maintenance). In this way, a more accurate remaining life prediction is obtained as failure comes closer. This is completely opposite the case where no condition assessments are performed and the distribution is shown to become wider as time proceeds (see Fig. 6.1).

It should be noted that the reduction of uncertainty in the Hess and Engel approach is fully due to the condition assessments, which actually reduce the population of the original PDF. They state that the shape (and width) of the original PDF should be accepted as a given fact and cannot be controlled [2]. For an experience-based approach that is true, since the PDF is determined by the variations in usage and loading. However, in the next section, it will be shown that physical model-based approaches do enable a reduction of the PDF width. This ability to reduce uncertainty will appear to be the motivation for the usage- and load-based maintenance concepts discussed in Sect. 6.4.

Finally, uncertainties in the input of reliability analyses are often divided into two types [2, 4–6]. *Aleatory*, or irreducible, uncertainty due to physical variability and *epistemic*, or reducible, uncertainty arising from lack of knowledge on the system behaviour. The three sources of uncertainty in the damage parameter mentioned above can be classified as follows. The uncertainty in usage is a natural variation and thus aleatory. The uncertainty in the loads has both an aleatory and epistemic component. The model used to relate the usage to the loads provides an epistemic uncertainty, while the natural variation in dimensions (manufacturing process) and material properties gives an aleatory uncertainty. Similarly, the life prediction also contains an epistemic (lifing model) and aleatory (dimensions/properties) uncertainty. The next sections will demonstrate how monitoring usage, loads or condition can reduce both types of uncertainties. The use of physical models increases the knowledge of the system behaviour and thus reduces the epistemic uncertainty, while determination of the loads or condition eliminates the need for calculating these quantities, thereby reducing the associated aleatory uncertainty.

6.3 Model-Based Prognostics

In the previous section, it was demonstrated that the conservatism present in many maintenance intervals is largely due to uncertainty in the damage evolution in the components. However, if the relation between actual usage of individual components and their degradation can be quantified, the uncertainty can be reduced. In

other words, if an accurate model-based prognostic method is available, the optimal moment of replacement can be obtained for any component, based on the (monitored) usage or loading of that part [7].

6.3.1 Relation Between Usage, Loads and Degradation Rates

Normally, the usage of a system, in terms of operating hours, power settings, number of starts, etc., is known to the operator or can be monitored rather easily. However, the remaining life of the system determines when maintenance actions must be performed, whereas the relation between the usage and the remaining life is in many cases unclear. Insight into this relation can be obtained by zooming into the level of the material point, since that is the level at which the physical failure mechanisms are active. This requires translation of the usage (on the global level) to the local loads (e.g. stress, strain, temperature, electrical current) on the material level, as is illustrated in Fig. 6.2.

The loads are then related to the capacity of the material by some failure model (e.g. fracture, fatigue, creep, arc flash), which yields the damage accumulation, degradation rate or life consumption rate at the present load. Finally, assuming that the usage and/or loads can be estimated, a prognosis can be given for the remaining life of the system.

Two important relations in Fig. 6.2 are the usage-to-load relation and the load-to-life relation, denoted by the numbers 1 and 2, respectively. These relations can be assessed in a quantitative sense only when the physical background of the loading and the failure mechanism is understood. If accurate models are available

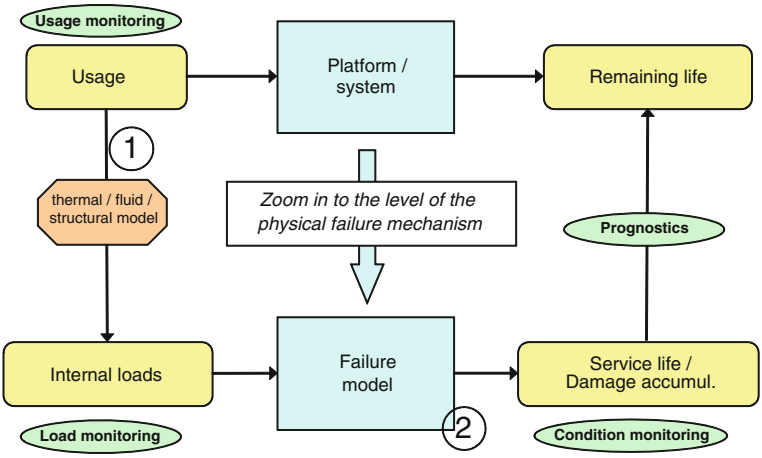

Fig. 6.2 Schematic representation of the relation between usage, loads, condition and life consumption. The most important relations are (*1*) the usage-to-load and (*2*) the load-to-life relations

for these relations, any usage history of the system can be translated into the associated damage accumulation or life consumption.

Note that the failure mechanism is modelled on the material point level in a component, whereas a system may contain numerous components with several failure mechanisms each. Therefore, before the method illustrated in Fig. 6.2 can be applied, a failure mode, effect and criticality analysis (FMECA) must be performed to determine which mechanism(s) in which component is/are critical to the service life of the complete system. This will be discussed in Chap. 8.

Monitoring of the usage or loading of the system is essential in a model-based approach. Figure 6.2 shows that monitoring can be performed at different levels. The lowest level of monitoring is usage monitoring, which implies registration of quantities like operating hours, rotational speeds or number of starts. Although usage monitoring is in most cases rather easy to perform, relating the data to degradation rates is generally not straightforward and requires quite some models and calculations. Load monitoring is one level higher, since it directly assesses the internal loading of components. This can be realized by applying sensors like thermocouples (to measure the temperature) or strain gauges (deformation). Monitoring at this level is generally somewhat more complex than usage monitoring, but the obtained information is related more directly to the component condition. The highest level of monitoring is condition monitoring, where the actual condition (i.e. the amount of degradation) is assessed directly and no calculations are required. However, this level of monitoring generally requires rather sophisticated sensors and is not always feasible, either technically (accessibility of the component) or economically.

During the design of a system, the original equipment manufacturer (OEM) must determine the maintenance intervals. However, the actual usage will be unknown at that stage. Therefore, assumptions are made regarding the number of operating hours per time period and the variation in the severity of usage (e.g. power setting). Then, based on this assumed usage, loads and lifetimes can be calculated, which ultimately leads to prescribed maintenance intervals. The uncertainty in these assumptions is generally covered by safety factors. As was discussed in Sect. 6.2, especially these safety factors cause the maintenance intervals to be (very) conservative.

In the next subsection, the model-based prognostic approach illustrated in Fig. 6.2 will be shown to reduce the uncertainty. The approach will therefore aid in reducing the spill of remaining life in prematurely replaced components and thus increase the maintenance efficiency.

6.3.2 Uncertainty Reduction

The conservatism encountered in the maintenance interval determination, as discussed in Sect. 6.3.1, can be reduced by decreasing the uncertainty in the predicted condition. This can be achieved by applying information about the actual usage for

the determination of the intervals. The result is a narrower distribution of the damage parameter, which leads to less conservative intervals, as is illustrated in Fig. 6.3. The narrower distribution can move further to the right before the tail reaches the critical damage limit. This yields a longer service life for the population of components. Also, the average amount of damage in de replaced components is higher, which reduces the spill of remaining lifetime.

Note that incorporating information about the usage or loading of the parts implies that the time to replacement is no longer given in calendar time (actual hours) but in equivalent hours. For example, for industrial machinery, the number of operating hours will probably be a better usage parameter than calendar time. For other applications, the effect of operating temperature may also have to be incorporated in the equivalent hours expression.

6.3.3 Applications

Model-based prognostic methods can be applied in several ways in the field of maintenance. The three most important applications are the following [7]:

1. Development and application of load-based maintenance (LBM) and usage (severity)-based maintenance (UBM/USBM) policies.
2. Development and application of advanced CBM approaches (prognostics + system development).
3. Understanding and reassessing failure data sets.

The load- and usage-based maintenance strategies will be introduced in Sect. 6.4. In Sect. 6.5 on health and condition monitoring, the role of model-based prognostics in CBM will be discussed, while its role in data analysis will be treated in Chap. 7. All these applications require (1) monitoring of the relevant quantities (usage, load, condition) and (2) a relation between the monitored quantity and the

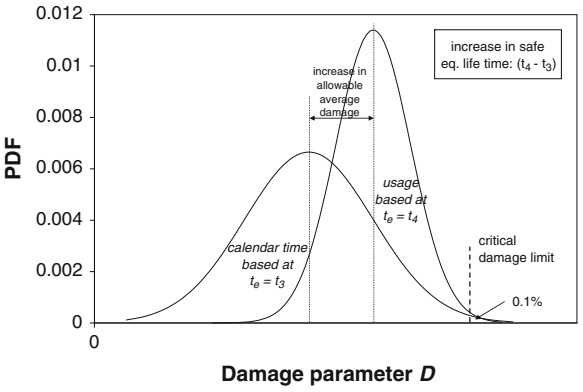

Fig. 6.3 Effect of reducing the width of the distribution: the service lifetime increases and the average amount of damage at replacement of the population is higher. Time is now given in equivalent hours, since the usage (severity) is incorporated

service life. For the latter relation, understanding of the physical mechanisms treated in the first part of this book is required.

As was already mentioned in Chap. 5, the physical model-based prognostic approaches have a number of clear benefits when compared to the experience-based prognostic approaches. Firstly, these methods do not rely on the collection of large sets of (failure) data. Once a physical model of the failure mechanism is available, knowledge of the material properties and local loads (as a function of usage) are sufficient to determine the component service life. And secondly, as opposed to the statistical and stochastic methods, the model-based methods are not based on historical data. Since the quantitative relation between usage and degradation is known, changes in usage can easily be incorporated in the prognostic analyses. This characteristic offers an opportunity for optimization of maintenance strategies that has not been fully utilized yet, as will be demonstrated in the remainder of this chapter.

6.4 Load- and Usage-based Maintenance

One of the major applications of the model-based prognostic approach presented in the previous section is the development of usage- and load-based maintenance strategies. These strategies rely on two basic requirements:

- The usage or loading of the system is monitored;
- A physical model-based prognostic method is available.

The optimal maintenance intervals can then be determined on the basis of the calculated (i.e. predicted) degradation of the system. Although the concept of including usage monitoring data in prognostic methods has been proposed by several researchers [8–11], practical applications to real systems are rather scarce. In this section, the functional and technical approaches to load- and usage-based maintenance will be discussed and several case studies will be described.

6.4.1 Functional versus Technical Approach

One of the challenges in load- and usage-based maintenance is to bridge the gap between the system level and the component level. Whereas the expected remaining life and associated maintenance intervals are required for the complete system, the physical models, that quantify the degradation for a certain usage pattern, are defined at the component or even material level.

Two approaches exist to bridge this gap: (1) the functional approach and (2) the technical or physical approach. In the functional approach, both life prediction and degradation are assessed on the system level, while in the technical approach both are determined on the component level, as is illustrated in Fig. 6.4. The figure

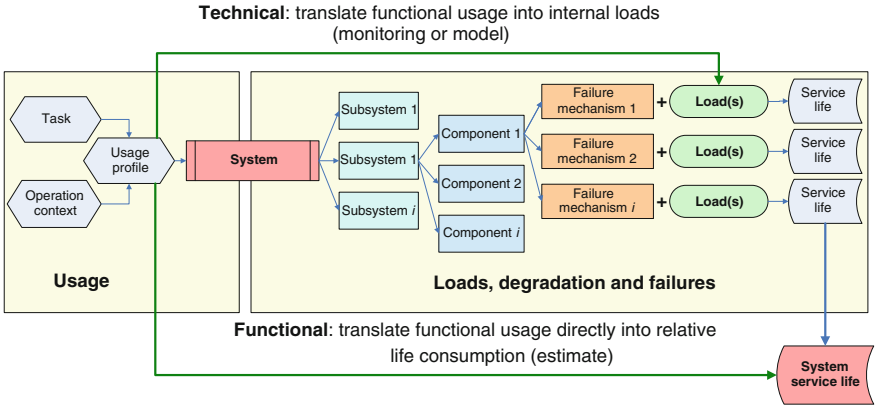

Fig. 6.4 Overview of usage- and load-based maintenance approaches

shows that the usage of a system can be defined in terms of the *task* (what the system is doing) and the *operational context* (in what conditions the system is operating). These two factors together determine the *usage profile* of the system [12]. A limited number of usage profiles typically suffice to characterize the (functional) usage of any system. The challenge is now to relate the usage profiles to the system degradation, in order to predict the remaining life and the optimal maintenance interval.

6.4.1.1 Technical Approach

In the technical approach, the system is decomposed into several subsystems, which are subsequently again decomposed into their individual components. For each component, the relevant failure mechanisms are determined and the loads that govern that failure mechanism are assessed. Moreover, the internal loads on the components are obtained from either a calculation (model) or a monitoring system (e.g. strain gauge, thermocouple). In that way, the usage of the system is translated into local loads, and their effect on the degradation of the individual components is assessed.

This approach requires a detailed (technical) analysis of the system and its components and for complex systems can be very time consuming. Therefore, in general, only the critical subsystems and components are included in the analysis. A useful approach to perform this selection process is the degrader analysis proposed by Banks et al. [13]. Firstly, the top 10 cost drivers and performance killers will be identified. Cost drivers are those subsystems or components that are responsible for a considerable fraction of the system maintenance costs. Subsystems or components that severely (and negatively) affect the system performance or availability are designated performance killers. Secondly, the failure modes and mechanisms of the cost driving and performance killing components will be

determined. Finally, relevant parameters to monitor these components will be determined and monitoring techniques will be selected.

The advantage of this technical approach is that quite accurate predictions of the remaining life can be obtained, but at the cost of dedicated sensors or data acquisition systems to gather the usage data and a considerable effort in developing suitable prognostic models.

6.4.1.2 Functional Approach

In the functional approach, the degradation is quantified on the system level, based on system level functional usage profiles. Contrary to the technical approach, the system is thus not decomposed into subsystems and components, but the effect of the usage variation on the complete system degradation is assessed. It is not possible to apply detailed physical models in this case and the relation between usage and degradation will be (at least partly) experience based. For example, two usage profiles can be defined for a car: (1) short trips within a city and (2) long trips on a highway. Based on experience and some knowledge on failure mechanism, it might be possible to define replacement intervals for the brake pads (which appear to be one of the critical components in a car) for both usage profiles. Monitoring the number of type 1 and type 2 trips then enables estimation of the next moment of brake pad replacement.

The advantage of this approach is that the functional usage profiles are generally rather easily described by an operator. Moreover, the system composition and development of physical models can be skipped, but the associated drawback is a more limited accuracy of the prediction.

The load- and usage-based strategies discussed in this section fill the gap between the rather simple experience-based prognostic strategies (i.e. trending, extrapolation) and CBM strategies. The former methods only provide an estimate of the present condition and remaining life. These are the methods commonly applied in health and usage monitoring systems (HUMS) commonly found in aircraft and helicopters. Their advantages are the relatively simple sensors and data acquisition techniques. On the other hand, the CBM approaches provide a much more accurate remaining life prediction, especially during the final phase of the service life. However, application of the appropriate sensors is not always feasible, either technically (availability of sensor, accessibility of location, etc.) or economically.

The proposed load- and usage-based strategies then provide a suitable alternative. Since no condition monitoring is required, relatively simple sensors can be applied to monitor the usage (e.g. operating hours, start–stops, rotational speeds) or loads (e.g. temperature, strain, electrical current). But at the same time, the results are much more accurate than those of the experience-based approaches, since a physical model is used to calculate the system degradation.

In the remainder of this section, two case studies will be presented. In the first case study, the technical approach is applied to a gas turbine, while the second case study demonstrates the functional approach on a military combat vehicle.

6.4.2 Case Study Technical Approach: Gas Turbine Blade

The methods discussed in the previous section are applied to a case study to demonstrate the benefits of using information on the specific usage or loading of the system, thereby applying the knowledge on the associated physical failure mechanism. A gas turbine blade has been selected here as case study, since the service life can rather easily be related to one single failure mechanism (creep). Moreover, the usage of gas turbines is generally well known, since most OEMs provide at least usage monitoring systems (and sometimes condition monitoring systems) with their machines. A typical low pressure turbine blade is shown in Fig. 6.5.

The (assumed) usage history of the gas turbine is shown in Figs. 6.6 and 6.7. Figure 6.6 shows the variation of the number of operating hours per year, whereas Fig. 6.7 shows the variation in the power setting of the machine (fractions of time at high, middle and low power). The average usage per year, in this example, is 2,130 h with a standard deviation of 967 h.

6.4.2.1 Physical Models

Physical models are required for both the usage-to-loads and load-to-life relations as indicated by the numbers 1 and 2 in Fig. 6.2. In the following, the loads and damage accumulation are related to the usage of the gas turbine presented in Figs. 6.6 and 6.7.

The components rotate at speeds in the order of 10.000 revolutions per minute (rpm) and operate in a hot gas stream with temperatures in the range from 500 to 1,000 °C. Due to the rotation, a centrifugal force (F) acts on the turbine blades,

Fig. 6.5 Low pressure turbine blade

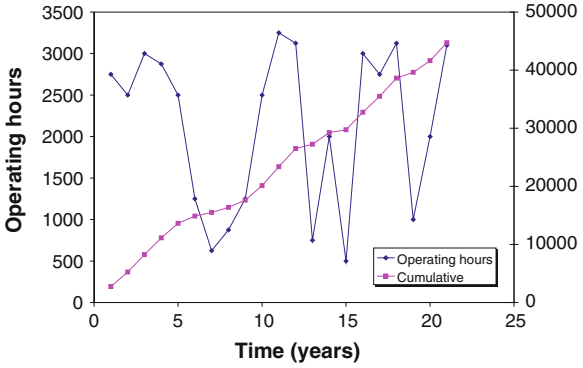

Fig. 6.6 Usage history of the gas turbine in terms of operating hours per year

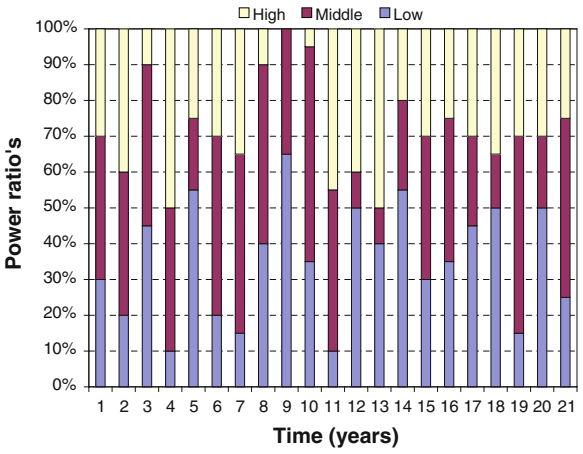

Fig. 6.7 Variation of gas turbine usage in terms of power settings

causing radially directed normal stresses in the root of the blade. The magnitude of this stress (σ) depends on the mass (m) of the blade, the rotational speed (ω), the distance from the blade to the engine centre axis (r) and the area of the blade cross section (A) in the following way

$$\sigma = \frac{F}{A} = \frac{m\omega^2 r}{A} \tag{6.1}$$

The rotational speed and gas temperature depend on the power setting of the gas turbine. The values for the three power settings shown in Fig. 6.7 are given in Table 6.1. Using the values $m = 0.2$ kg, $r = 0.2$ m and $A = 1.4 \times 10^4$ m^2, the corresponding stress values can be calculated. Since the turbine blades are solid and uncooled, the blade temperature will equal the gas temperature in a steady-state situation.

Table 6.1 Turbine blade loads at different power settings

Power setting	Rotational speed [rad/s]	Stress [MPa]	Gas temperature [°C]	Blade temperature [°C]
Low	597	102	500	500
Middle	733	154	750	750
High	1,047	313	900	900

6.4.2.2 Damage Accumulation

For the present component, creep is the life limiting failure mechanism. Creep is a high temperature deformation process that depends on both the stress and temperature levels. It results in inelastic deformation of the blade, which could lead to failure in two ways: either the elongation of the blade causes it to touch the casing or a locally high creep strain initiates a crack, which leads to fracture of the blade. Modelling of the creep behaviour can be performed at several levels, ranging from detailed multiscale models [14, 15] to rather simple power law relations. The latter approach will be followed here to demonstrate that simple methods can already provide a large improvement of the efficiency. The creep strain rate for the present material is described by a Norton creep law and depends on the temperature (T) and stress (σ) as follows

$$\dot{\varepsilon}_{cr} = \frac{\mathrm{d}\varepsilon_{cr}}{\mathrm{d}t} = AT^4\sigma \tag{6.2}$$

where the value of the constant A equals 2×10^{-20} $(\mathrm{MPa})^{-2}$ $(^{\circ}\mathrm{C})^{-4}$ $(\mathrm{h})^{-1}$. A creep strain of 1.0 % is defined as the critical amount of creep deformation ($\varepsilon_{\mathrm{crit}}$), leading to an unacceptable elongation of the blade. Note that it is assumed that the creep deformation up till 1 % creep strain is mostly in the secondary creep regime, which means that the creep strain rate is constant. The time to failure (t_{f}) can then be calculated as

$$t_f = \frac{\varepsilon_{\mathrm{crit}}}{\dot{\varepsilon}_{cr}} \tag{6.3}$$

The accumulated amount of damage (D) can be obtained using Robinson's damage rule [16]:

$$D = \sum_i \frac{\Delta t_i}{t_{f,i}} \tag{6.4}$$

where Δt_{i} is the time period spent at some conditions (stress and temperature) and $t_{\mathrm{f,i}}$ the failure time at those conditions. Failure will occur when the damage parameter D attains the value 1.

Using these equations, the damage accumulation for the three power settings can be calculated. The results are shown in Table 6.2.

Table 6.2 Creep deformation and damage accumulation rates at different power settings

Power setting	Creep strain rate [h^{-1}]	Failure time [h]	Damage accumulation rate [h^{-1}]
Low	1.27×10^{-7}	78,587	1.27×10^{-5}
Middle	9.72×10^{-7}	10,293	9.72×10^{-5}
High	4.11×10^{-6}	2,432	4.11×10^{-4}

6.4.2.3 Comparison of Methods

Five preventive maintenance policies will be compared here in terms of effectiveness. For each method, a criterion is derived to replace the turbine blades. Then, applying the usage history from Fig. 6.6, the moments of replacement are determined and the total number of replacements during the 21 years history are counted and compared.

It is assumed that the real behaviour of the blades as a function of the usage is described exactly by the physical models presented in the previous subsection. This means that, during the simulations, the elongation of the blades at any moment is fully known and consequently the most optimal replacement intervals (when the damage parameter $D = 1$) are also known. Normally, this real behaviour is unknown to the operator, unless it is possible to continuously and accurately monitor the blade elongation.

In practice, a physical model will only approximate the real behaviour, and a certain amount of (epistemic) inaccuracy will be present in the model predictions. Therefore, using the model in Eq. (6.1) to calculate the loads from the usage, an uncertainty of 5 % (standard deviation) is assumed. So when the rotational speed is known, relation (6.1) provides the average blade stress with a standard deviation of 5 %. For the load-to-life model in (6.2), a standard deviation of 10 % is assumed. Further, in Sect. 6.2, it was illustrated that a component has to be replaced when the tail of the distribution (0.1 %, see Fig. 6.1) reaches the critical amount of damage. In a normal distribution, this 0.1 % probability corresponds to the $\mu + 3\sigma$ value (three times the standard deviation added to the mean value). Therefore, in the analyses in the following subsections, the $\mu + 3\sigma$ values of the calculated creep strains are compared to the critical creep strains to determine the failure time. The procedure followed is schematically shown in Fig. 6.8.

Next, the simulation of the five preventive maintenance policies will be discussed.

6.4.2.4 Calendar Time Based

If an OEM, during the design phase, would have to provide a fixed calendar time for replacement of the blades, assumptions must be made for the expected number of operating hours per time period and the expected ratios of power settings. Based on the history shown in Fig. 6.6, with an average usage of 2,130 h per year, a safe

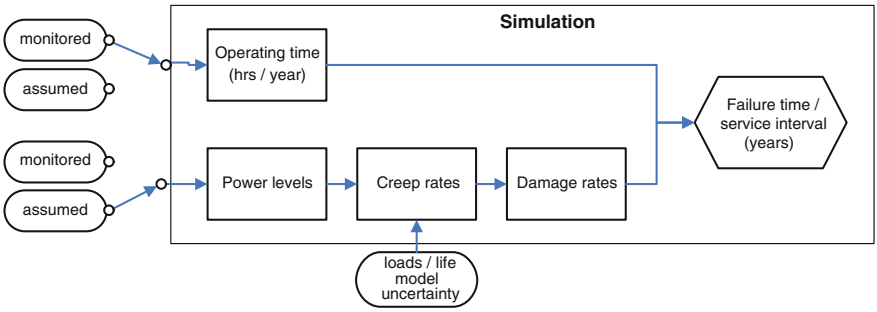

Fig. 6.8 Schematic representation of the simulation. Depending on the maintenance concept, either a monitored or assumed quantity is used as input and a specific model uncertainty is applicable

estimate of the expected usage can be obtained by taking the average number of hours increased by the standard deviation (967 h). This yields an estimated usage of 3,096 h per year. The ratio between the different power settings is hard to estimate a priori. Since it is possible that an operator runs the machine at the high power setting constantly, a safe assumption would be to use only the high power setting in the life assessment.

At this high power setting, the average creep rate according to the creep model equals $4.11 \times 10^{-6}\ \mathrm{h}^{-1}$ and the associated failure time appears to be 2,432 h (Table 6.2). But since the load and failure model contain a total uncertainty of 15 % (standard deviation), the upper limit of the creep rate (associated with 0.1 % probability of failure) is obtained by adding three times the 15 % standard deviation, which yields a creep rate equal to $5.96 \times 10^{-6}\ \mathrm{h}^{-1}$ (see Fig. 6.9). The associated failure time is obtained by relating this creep rate to the critical creep strain, see Eq. (6.3), yielding a failure time of 1,677 h. Therefore, the proposed replacement interval would in this case be obtained by relating the failure time to the estimated yearly operating time: 1,667/3,096 = 0.54 year ($\sim$6 months).

6.4.2.5 Usage Based

By taking into account the actual number of operating hours per year, a rather simple variant of UBM is obtained. No assumptions are required for the operating hours, but the ratios of power settings are still unknown, so the conservative assumption that all hours are run at the high power setting still must be made. As was shown for the previous method, the lower limit for the failure time is then 1,677 h. In this case, the blades are only replaced when the actual number of operating hours exceeds the calculated failure time. This means that the intervals are in most cases considerably longer than the 0.54 year obtained from the CTBM method, as the 3,096 h per year was a rather conservative assumption.

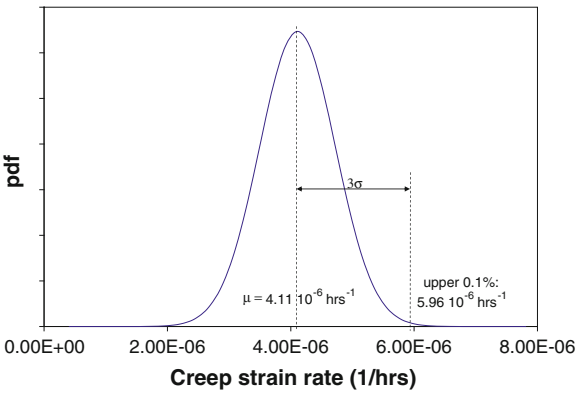

Fig. 6.9 Creep strain rate distribution, showing the difference between the average value and the upper 0.1 % value

6.4.2.6 Usage Severity Based

In this policy, not only the usage of the system in terms of operating hours is taken into account, but also the severity of the usage. Therefore, the power settings are monitored in addition to the operating hours. This means that during the operating hours at low or middle power, much less damage is accumulated than during the hours at high power, as can be observed in Table 6.2. The resulting lower limits for the failure time, considering the total uncertainty of 15 %, are now 54,198, 7,099 and 1,677 h for the three power settings. Although still the uncertainties of both the loads and life model must be taken into account, this policy again increases the replacement intervals considerably, since only a fraction of the operating hours are performed at the high power setting.

6.4.2.7 Load Based

In a load-based maintenance policy, the internal loads are monitored, which in this case means that the stress in the blades is measured. Practically, this is not very easy to achieve in a gas turbine, due to the high temperature and limited accessibility. However, for many other systems that is much easier and it is assumed here that it is technically possible to determine the stress. Consequently, no physical model is required to calculate the loads, but only a model for the loads-to-life calculation is needed. As a result, the uncertainty associated with the loads calculation (5 %) has no effect in the method anymore. The only remaining uncertainty is the 10 % inaccuracy in the creep model. The calculated upper limit of the creep strain is again compared to the allowable creep strain to determine the moments that the damage parameter attains the value 1 and the blades must be replaced.

6.4.2.8 Condition Based

The final step is to directly monitor the condition of the components, in this case the elongation of the turbine blades. In practice, the blades can be measured during periodic inspections, but it is here assumed that the elongation can be monitored continuously. For this policy, no physical models are required anymore, so the associated uncertainties do no longer play a role. The moment of replacement can be determined very accurately on the basis of the monitored elongation. Note that the blade is only replaced at the moment the monitored damage attains the value 1, so there is no lead time between the detection of incipient end-of-life and the actual moment of replacement. In practice, a certain amount of prognostics would be applied to create this lead time, but at the cost of an increase in the uncertainty.

6.4.2.9 Results

The five policies described above have been simulated using the history shown in Fig. 6.6. The evolution of the damage parameter D is calculated using equation (6.4) and the lower limits of the failure time as defined before. The resulting evolution of the damage parameter is shown in Fig. 6.10. The decrease in the uncertainty in the usage, loads and condition of the blade yields the observed decrease in calculated damage.

Replacements of the blade are triggered by the damage parameter reaching a multiple of 1.0. The resulting moments of replacement are shown in Fig. 6.11, whereas the total number of replacements over a period of 21 years is shown in Fig. 6.12.

For the calendar time-based method (CTBM), one or two replacements are necessary for each year. Taking into account the usage in terms of operating hours reduces the number of replacements considerably, especially in the years with low numbers of operating hours (e.g. year 7 and 8). The usage severity-based method (USBM), which also takes into account the variations in power settings, is again

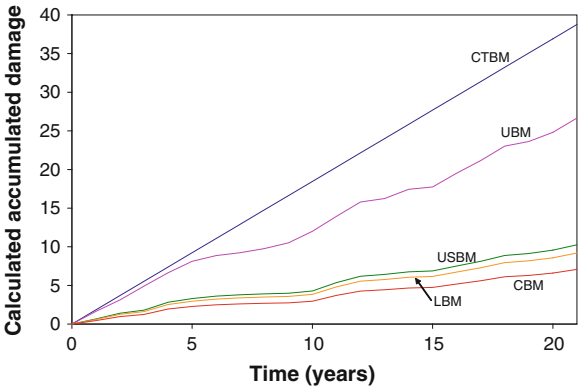

Fig. 6.10 Calculated evolution of the damage parameter using different assumptions about the usage and different levels of uncertainty in the models

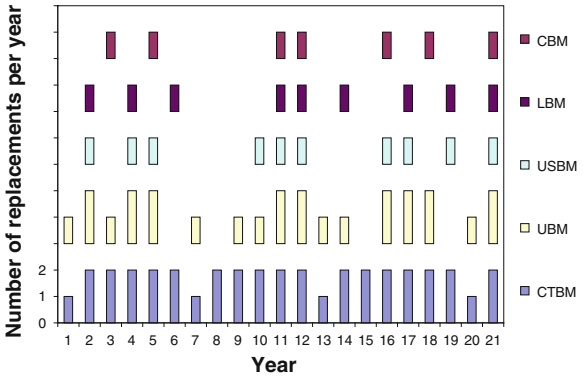

Fig. 6.11 Overview of blade replacement moments for five different maintenance strategies

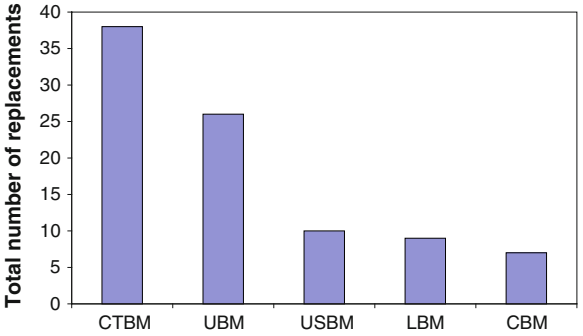

Fig. 6.12 Comparison of the total number of replacements during a period of 21 years for five different maintenance strategies

more efficient. By using the measured loads (LBM) instead of the calculated loads, the uncertainty in the life prediction is reduced, which yields another decrease of the number of replacements. Finally, CBM provides the most efficient replacement intervals, since in that method also the uncertainty associated with the creep calculation is absent.

Figure 6.12 illustrates that application of information about the usage, loads or condition yields a significant improvement of the efficiency of the maintenance process, while the effectiveness (probability of failure) remains the same. CBM is the most efficient method, but for many systems no suitable sensors are available or the accessibility of the system is too limited. In that case, usage- or load-based methods are interesting alternatives, since loads and especially usage are much easier to monitor. However, these methods require understanding of the component failure mechanism(s) and (the development of) physical models to calculate the loads and lifetime. In this case study, the largest improvement is obtained at the transition from usage based to USBM. Applying LBM and CBM does not yield a significant additional benefit. However, the relative efficiency gains depend on the uncertainties in the different models.

6.4.3 Case Study Functional Approach: Military Combat Vehicle

In this second case study, the functional approach of UBM will be demonstrated. The military combat vehicle CV90, see Fig. 6.13, will be used as the case object [17].

The approach followed consists of the following steps:

1. Determine the critical (sub)systems and their failure modes,
2. Define a limited number of usage profiles for the system,
3. Quantify the relation between the usage profiles and the degradation rates for the identified failure modes.

When these three steps have been completed, a methodology for USBM is in place for the system. Monitoring the usage profiles then enables the assessment of the optimal maintenance intervals. The three steps in the method development will be discussed separately below.

6.4.3.1 Critical Subsystems and Failure Modes

To determine which subsystems of the CV90 are critical, the degrader analysis proposed by Banks et al. [13] is applied. This analysis starts with assessing the top 10 cost drivers and the top 10 availability killers. The cost drivers are those subsystems or components that are responsible for a considerable fraction of the system maintenance costs. Subsystems or components that severely (and negatively) affect the system performance or availability are designated performance killers. Based on data obtained from information systems and expert opinion, a list of the critical subsystems/components is obtained for the CV90, see Table 6.3. Three of the subsystems in this table are both cost drivers and availability killers, while the other subsystems are either cost or availability critical.

Fig. 6.13 Combat vehicle CV90 (*source* Wikipedia)

After completing the list of critical subsystems, the failure modes of these subsystems must be determined. Maintenance information systems and expert experience could be used to determine the most common failure modes for four subsystems, as is indicated in the final column of Table 6.3. For the remaining subsystems, the failure modes could not easily be assessed.

6.4.3.2 Definition of Usage Profiles

The second step in this functional UBM approach is the definition of a limited number of usage profiles that enable the description of the system usage. As was explained before, a usage profile is a combination of the mission or task of the system and the operational context (e.g. environment). For the combat vehicle, the mission and operational context have been defined in terms of 14 parameters, which have been grouped in 7 categories, as indicated in Table 6.4.

Note that some of the parameters are binary (yes or no), for example, water crossing and combat loaded. Other parameters like kilometres, speed and hours are continuous parameters, and some parameters have a discrete number of options. For example, the terrain surface type can be selected from paved road, unpaved road, light terrain, medium terrain or heavy terrain. Similarly, the roughness of the terrain is either flat, hilly or mountainous.

The functional missions performed by the vehicle can now all be defined in terms of the parameters in Table 6.4. For example, if the vehicle is used to set up a road block, the parameter values shown in Table 6.5 can be defined. For the basic mission parameters like kilometres and hours, an average value for this type of mission can be estimated. In a similar way, the operational context parameters can be determined.

Table 6.3 Overview of 10 most critical subsystems and their associated failure modes

	Subsystem/component	Availability killer or cost driver	Failure mode
1	Track	Cost, availability	Track link: wear, fracture Track pad: wear
2	Engine	Cost, availability	
3	Seats	Availability	Seat cover wear, safety belt/head rest fracture
4	AC/heating assembly	Cost, availability	
5	Seat comm/Gunn assy	Availability	
6	Final drive	Availability	Final drive fracture
7	Idler wheel	Cost	Crankshaft bearing seizure, idler wheel fracture
8	Power unit	Availability	
9	Gun mount assembly	Availability	
10	Thermal camera	Cost	

Table 6.4 Parameters adopted to define the usage profile in terms of mission and context

Mission		Operational context	
Manoeuvrability		Location and climate area	
	Idling		Climate area
Driving			Humidity of area
	Driving behaviour		
	Speed	Terrain	
	Water crossing		Type of surface
Load			Roughness of terrain
	Combat loaded	Deployment	
Basic parameters			Spectrum of force
	Kilometres		
	Hours		
	Gun rounds		
	Number of starts/trips		

Table 6.5 Mission parameter values for the mission 'set-up road block'

Mission	
Manoeuvrability	
	Idling = Yes
Driving	
	Driving behaviour = calm
	Speed = 0 km/h
	Water crossing = No
Load	
	Combat loaded = Yes

In this way, the functional usage of the vehicle can completely be described in terms of the usage parameters. A certain mix of different usage profiles, that can easily be indicated by the operators, can then be translated into the usage parameters, which govern the degradation of the subsystems.

6.4.3.3 Quantifying Relation between Usage Profiles and Degradation Rates

The third step in the analysis is the most important and generally the most difficult step. The relation between the different usage profiles or usage parameters and the degradation rates of the critical systems must be quantified. The procedure will be demonstrated here for the most critical subsystem, the track (pad) of the combat vehicle.

The dominant failure mode for the track pads was identified to be wear of the pads. This means that especially the surface type and roughness of the terrain will have a large influence on the wear of the track pads. To quantify the relative

Table 6.6 Relative severity of wear for different surface types and terrain roughness

Surface type		Roughness of terrain	
Paved road	2.00	Flat	1.00
Unpaved road	1.00	Hilly	1.25
Light terrain	1.00	Mountainous	1.50
Medium terrain	1.25		
Heavy terrain	1.50		

severity of the surface types and terrain roughness, several experts have assessed the estimated effect of these parameters on track pad wear. The results are shown in Table 6.6. This indicates, for example, that driving on a paved road causes the track pads to wear a factor two faster than for driving on an unpaved road. Note that the assessment of the usage severity is performed here in a rather subjective manner (expert opinion). More objective measures could be obtained by collecting failure data for a longer period or by performing wear tests with track pads at different surfaces.

To apply this information in a USBM policy, the usage profiles of the combat vehicle must be specified. Table 6.7 shows the distribution of driving distance over the different combinations of surface type and terrain roughness for the usage profile 'training'. In this profile, more than one-third (35 %) of the kilometres is driven in heavy terrain at medium roughness.

From the relative usage severity numbers in Table 6.6, the severity numbers for the surface type—roughness combinations in Table 6.7 can be calculated, as is shown in Table 6.8.

Finally, the relative severity of the 'training' usage profile can be assessed by combining Tables 6.7 and 6.8. For a given total driving distance of 800 km in the 'training' usage profile, the equivalent distance for each combination of surface

Table 6.7 Distribution of driving distance over surface type and terrain roughness for usage profile 'training'

Roughness	Surface type		
	Paved (%)	Unpaved (%)	Heavy terrain (%)
Light	2	4	14
Medium	5	10	35
Heavy	4.5	15	10.5

Table 6.8 Relative severity of wear for different combinations of surface type and roughness

Roughness	Surface type		
	Paved	Unpaved	Heavy terrain
Light	2.00	1.00	1.50
Medium	2.50	1.25	1.88
Heavy	3.00	1.50	2.25

type and terrain roughness can be calculated. For example, driving on paved road at medium roughness represents 5 % of the total driving distance, which is 40 km. Moreover, the relative severity of this usage is 2.50. Multiplication of these two numbers yields an effective distance of $2.5 \times 40 = 100$ km. This result is shown in Table 6.9, together with the effective distances for all other conditions.

Summation of all nine contributions yields a total effective distance of 1,432 km. This number implies that a nominal distance of 800 km at this usage profile causes damage to the track pads which is equivalent to driving 1,432 km at unpaved roads and light terrain roughness (which is the reference situation with relative severity equal to 1.0). The average usage severity of this 'training' usage profile is therefore $1{,}432/800 = 1.79$.

As the 'training' usage profile is executed a lot in practice, collected failure data on the track pads will generally be related to this usage profile. From the collected failure data, the mean time to failure (MTTF) for the pads appears to be 1,883 km.

Combination of this MTTF and the calculated severity (1.79) of the 'training' usage profile then enables the prediction of the MTTF for any other usage profile. For a usage profile only consisting of driving on unpaved road at light roughness (severity = 1.0), the expected MTTF will be $1.79/1.0 \times 1{,}883 = 3{,}370$ km. However, for a more severe usage profile, for example, severity = 2.7, the expected MTTF will be even shorter, that is, $1.79/2.7 \times 1{,}883 = 1{,}248$ km.

6.4.3.4 Summary

This case study of the functional UBM approach demonstrates that the effect of variations in usage of the system can be incorporated in the maintenance interval determination, without the development of complex physical models and detailed monitoring of loads or usage. By estimating the quantitative effect of different usage profiles on the system degradation, just specifying the functional usage (mission + context) enables the application of a much more dynamic maintenance policy. The accuracy of this approach obviously largely depends on the accuracy of the usage–degradation relation. However, even when initially an accurate estimate is difficult to make, the collection of additional failure data can be used to improve the method continuously.

Table 6.9 Effective driving distance (km) for different combinations of surface types and roughness for the 'training' usage profile (800 km)

Roughness	Surface type			
	Paved	Unpaved	Heavy terrain	
Light	32	32	168	232
Medium	100	100	525	725
Heavy	106	180	189	475
Total	238	312	882	1,432 km

6.5 Health and Condition Monitoring

In Sect. 6.3, it has been indicated that monitoring can be performed at different levels. Monitoring the usage or loading of a system or component is generally not very complex, but also delivers a limited amount of direct information about the degree of degradation of the system. In the previous section, it has been explained how physical models can aid in obtaining the required information on the system condition from the monitored usage or loads. On the other hand, monitoring the health or condition of a system with dedicated sensors directly yields this information without the need of complex models and calculations. Only when a prediction of the remaining useful life is required, (physical) models are again required.

Condition monitoring techniques have already been applied in industry for many decades. A number of methods are therefore well matured and are now available as commercial systems. These systems will be discussed in the first subsection. At the same time, the last decade has shown a significant development of new condition monitoring techniques. This is on the one hand due to the fact that a large variety of new (reliable) sensors have become available, enabling the monitoring of a wide range of load and condition parameters. Sensors detecting fatigue cracks, disbonded joints, erosion, impact damage, composite delamination and corrosion are now available and enable SHM of complex systems like aircraft or wind turbines. On the other hand, the increased computational power of modern computers has made the analysis of all collected data feasible.

In addition to the sensors that directly monitor the condition of the system, also sensors are available that monitor performance parameters, like temperature or pressure, that are related to the system condition. These sensors enable an indirect type of condition monitoring that is frequently applied in the process (e.g. refineries) and power industry (e.g. gas turbines). Large numbers of sensors installed in complex systems enable characterization of the (thermodynamic) process and the corresponding condition of the system.

Due to the boost in performance and availability of sensors and other hardware, condition monitoring systems are offered by OEMs in several industries as *the* way to increase maintenance efficiency. However, many operators, and even some manufacturers, do not realize that CBM can only be performed in an efficient way when both a condition monitoring system (the sensors) and a prognostic method are available. The sensor data only provide information about the current state of the component. Just waiting till the moment that a monitored condition parameter exceeds a critical value means that immediate action is required, which is difficult to plan (e.g. personnel, spare parts) and may have serious consequences for the system availability. Therefore, a prognostic method is required to determine when future maintenance activities are necessary. This issue is discussed in Sect. 6.5.3.

6.5.1 Condition Monitoring Techniques

The most widely used condition monitoring techniques nowadays are vibration monitoring and oil analysis, but also acoustic methods and thermography are commonly applied. In addition to these methods focusing on mechanical failures, corrosion monitoring (by periodic inspections) is also a widely applied way of condition monitoring. Many books are available on the details of these techniques [18–20]. In this subsection, the essential aspects of the commonly applied techniques will be provided, while the next subsection will shortly discuss the more recent developments in SHM.

6.5.1.1 Vibration Analysis

In vibration analysis, the dynamic characteristics of a mechanical system in motion are used to identify its operating condition. The technique can be applied to rotating equipment (centrifugal pumps, compressors, gear boxes), reciprocating machines (engines, cylinders) and linear motion systems. Each system has a certain 'signature' in terms of the frequencies and amplitudes of its vibrations. In a healthy condition, the frequencies are related to the operating condition, for example, the rotational frequency, of the system and the design (e.g. number of balls in a bearing). Upon degradation of the system, the amplitude of the vibrations may increase, frequencies may shift to other values, or new frequencies may appear. By monitoring the vibrations, either continuously or periodically, changes in the signature can be detected and appropriate actions can be undertaken.

The actual monitoring of the vibrations is performed by transducers. The most commonly applied sensor is an accelerometer that measures accelerations of the system, but also sensors to measure displacements or velocities are applied. The sensors may be mounted to the machinery at various locations to perform continuous monitoring, or a portable device (see Fig. 6.14) is used to perform periodic measurements.

Fig. 6.14 Portable vibration analysis device

The obtained vibration data are then analysed, and in most cases a fast Fourier transform is applied to convert the time domain data to a frequency spectrum. In such a plot in the frequency domain, the shift or amplitude increase of certain vibration frequencies can easily be detected. Basically, two approaches can be followed to analyse the collected data. In the first approach, only the total amount of energy in the spectrum is determined by calculating the power spectral density. The increase of this quantity for a system indicates that the system is degrading and the overall vibration level increases. Although the analysis is rather straightforward and can be performed quickly, the information obtained is rather limited. Only a warning that something is happening is provided, but determination of the cause of the spectrum change is not possible with this analysis.

In the second and more detailed approach, separate frequencies are analysed instead of the overall spectrum; especially when certain frequencies are known to be related to specific parts of mechanisms in the system, for example, certain bearings or gear sets, observed changes in those frequencies can aid in assessing the root cause of a system failure. Typical faults that can be detected from this type of vibration analysis are as follows [19]:

- Imbalance: increase in the peak at the fundamental running speed (1X) of the shaft;
- Misalignment: increase of the fundamental (1X) and 2nd harmonic (2X) frequency;
- Bearing damage or wear: frequency depends on running speed (1X), but also on number of rolling elements (NX) and ratio of ball and pitch diameter;
- Gear damage: changes in the gear mesh (=set of frequencies depending on the number of gear teeth and running speed);

To be able to detect these types of failures, a good understanding of the machine and its parts is required and a basic spectrum for the healthy machine must be used as a reference.

6.5.1.2 Ultrasonics

Also based on the analysis of vibrations is the ultrasonics monitoring technique. While regular vibration monitoring frequencies range from 1 Hz to around 30 kHz, ultrasonics monitors frequencies higher than 30 kHz. This technique is used for very specific applications, where leak detection is the most common. Leaks in pressure or vacuum vessels generally create high frequency noise caused by the expansion of air, gasses or liquids flowing through a small orifice. Note that this technique is primarily used to diagnose the system, that is, detect the occurrence of leaks, but is not very suitable for monitoring the gradual degradation of a system.

6.5.1.3 Oil Analysis

By analysing the lubrication oil from mechanical or electrical systems, information can be obtained on the condition of both the oil and the systems themselves. The

former can be applied to determine the optimal moment for lubricant replacement. The analysis is based on a chemical analysis of the oil, where typically viscosity, pollution and water content are determined.

More interesting from a condition monitoring point of view, however, is the fact that the lubricant also contains wear particles from the system it is lubricating. The concentration of these wear particles is a good indication for the wear rate in the system and thus provides information on the present condition. Moreover, the shape and chemical composition of the wear particles provide insight in which part inside the machine is wearing and what the dominant wear mechanism is. These analyses are typically performed with optical imaging systems combined with analysis and pattern recognition algorithms that provide both a morphological analysis and the particle size distribution.

Condition monitoring based on oil analysis is in most cases performed by periodically sampling the lubricant and analysing the samples in a laboratory. However, also some sensors are available that enable continuous monitoring. A well-known example is the magnetic plug that is present in many aerospace applications. The plug is positioned in the lubricant flow and collects ferritic particles. Once a certain amount of particles is present, an electrical signal provides a warning that is used to trigger the replacement of the lubricant or the inspection of the associated mechanical systems. Also online sensors to measure the water content of lubricants are available nowadays. These sensors are mainly used in maritime applications.

6.5.1.4 Thermography

This technique monitors the emission of infrared energy to assess the operating temperatures of systems. The working principle is based on the radiation law discussed in Sect. 2.3.3, stating that all bodies with an absolute temperature $T > 0$ K radiate energy, where the amount of energy is proportional to T^4. Initially, thermographic measurements were done with infrared thermometers or line scanners, but nowadays infrared cameras are available to visualize the radiation from a system. However, interpretation of the results is not always easy, since not only the system itself produces radiation, but also all other systems in its vicinity. Moreover, part of the radiation is also reflected by walls, ceilings and other systems, which makes that a thermographic analysis in most cases cannot be used to determine absolute values of system temperatures. Nevertheless, a comparative analysis providing information on which parts of the system are at elevated temperature in many cases yields valuable insight into the condition of a system. Finally, although the method is often applied to electrical systems, it can also detect anomalies in mechanical systems, for example, a temperature increase in a seizing bearing.

6.5.1.5 Corrosion Monitoring

Monitoring of corrosion processes is widely applied to static equipment in the oil and gas sector (e.g. pipelines, pressure vessels) and in the offshore and maritime industry. Although in recent years online corrosion sensors are being developed that enable continuous monitoring of structures, most of the monitoring in this field is achieved by performing periodic inspections. By periodically assessing the damage caused by corrosion, that is, determining the weight loss or reduction of wall thickness (see also Sect. 4.10.1), trends in degradation rates can be obtained and decisions on the timing of maintenance or replacement of structures can be taken.

Since many structures operating in corrosive environments are protected by coatings, monitoring the condition of the coating is also important. This is again achieved by performing periodic inspections, which can either be visual or be performed using analysis techniques like electrochemical impedance spectroscopy (EIS) or electrochemical noise measurements (ENMs) [21]. The latter technique has considerable potential for application in (wireless) corrosion sensors due to its passive nature (i.e. application of a current to the structure is not required to perform the measurement).

6.5.2 Structural Health Monitoring

A research field closely related to condition monitoring is the field of SHM. Whereas condition monitoring mainly focuses on rotating and reciprocating equipment in plants and process industry, SHM aims to assess the condition of (mostly large) structures in civil, mechanical and aerospace engineering. Typical structures for which SHM systems have been developed are bridges, offshore rigs, wind turbines, helicopters, aircraft and spacecraft. Moreover, as was discussed in the previous subsection, condition monitoring comprises several techniques like vibration monitoring, thermography and oil analysis. SHM mainly utilizes the changes in the dynamic response of the structure to assess the present condition.

The basic premise of the SHM methods is that damage will alter the stiffness, mass or energy dissipation properties of a system, which in turn alter the measured dynamic response of the system [22]. The SHM process therefore involves the observation of a system over time using periodically sampled dynamic response measurements from an array of sensors, the extraction of damage-sensitive features from these measurements and the analysis of these features to determine the current state of the system's health. As was discussed for failure prognostics in Sect. 5.5.3, also in this field both experience-based and model-based approaches can be adopted to relate the observed changes in dynamic response to the system deterioration.

In the experience-based approach, data is collected over a prolonged period of time and statistical pattern recognition techniques are applied to identify the occurrence of damage in the structure. The drawback of this approach is that only

after a certain period of time, sufficient experience is gathered to make reliable decisions. This drawback can be circumvented by applying the model-based approach. If a reliable model of the system is available, the response of the system to a certain amount of damage at a specific location can be calculated and can be compared to the data obtained in practice. In that way, damage can be located and quantified without a large amount of experience. Since modelling the dynamic response of a large structure is quite complex and requires a large computational effort, this approach is presently only applied on a component or subsystem level. For example, the change in dynamic response of a composite T-beam can be used to detect composite delamination [23].

Some important challenges in SHM are the local character of damage, the influence of (operating) conditions on the dynamic response and the difficulty of a priori identification of the effects of damage [22]. Since damage generally occurs very locally and initially is quite limited in size, the effect on the response of a complete structure may be limited and hard to detect. This requires the development of accurate feature recognition methods, but also procedures to select the optimal number and locations of the sensors to maximize the collection of useful data.

Another challenge is the influence of operating conditions on the dynamic response of a structure. The effects of changes in these conditions may easily exceed the effects of a local damage, which might then be impossible to detect. It is therefore important to filter out these effects, which might be easier when proper models of the system and registration of the loads and conditions are available. The final challenge in this field is that damage detection must be performed in an unsupervised learning mode. This means that the effect of a certain damage type on the response of a large structure cannot easily be determined experimentally and experience must be built to recognize certain damages. For simple structures, this process can be accelerated using physical models, but for large and complex structures that is not feasible yet.

Finally a technical development that could be an enabler for more sophisticated SHM systems will be discussed shortly: the fibre optic sensor.

6.5.2.1 Fibre Optics

A quite recent development in this field is the application of fibre optic sensors for load monitoring and damage detection. The basis of this technique is an optical fibre in which a fibre Bragg grating (FBG) has been constructed, see Fig. 6.15. If a light pulse is sent into the optical fibre, the grating with a certain spacing d will transmit most of the wavelengths in the spectrum, but light with the Bragg wavelength λ_B will be reflected by the grating. This means that both the transmitted and reflected spectral response will show a clear identification of the Bragg wavelength.

Since the Bragg wavelength depends on the spacing of the grating, straining of the optical fibre will change the spacing and thus the reflected wavelength. Therefore, the optical fibre can conveniently be applied as a strain sensor. And due to its small dimensions, it can rather easily be embedded in several structures, for

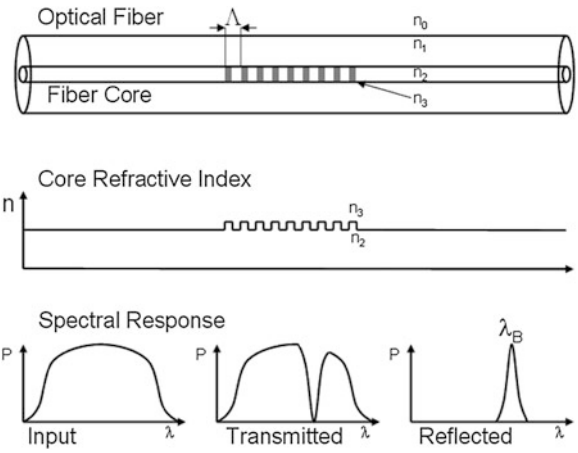

Fig. 6.15 Working principle of a fibre Bragg grating

example, concrete civil structures or composite aerospace structures. Moreover, by applying temperature sensitive or humidity sensitive coatings to the outer surface of the optical fibre, the strain sensor can be transformed into a temperature or humidity sensor.

6.5.3 Condition-Based Maintenance

Having a condition monitoring system (CMS) or structural health monitoring system (SHMS) in place does not automatically imply that performing CBM is trivial. The CMS or SHMS provides information about the present status of the system, that is, diagnostic information. However, based on this diagnosis, decisions must be taken on the timing and scheduling of maintenance tasks like repair or replacement. Only if a solid basis for these decisions, and the tools and methods to support them, are available, an effective and efficient CBM policy can be applied. The requirements for the useful application of condition monitoring data will be discussed in the present section.

To illustrate the challenges on this issue, the concepts of the *P–F* interval or delay time will be treated first. Most systems will start to degrade soon after they have been taken into operation. This degradation generally starts at the material level, for example, by accumulating fatigue or corrosion damage, but will initially not affect the performance of the system. Also a condition monitoring system will at this stage not be able to detect any significant changes in the condition of the system. However, at some stages, the CMS will be able to detect a certain anomaly and an indication (warning/alarm) is obtained that the system is prone to fail. It then takes another period of time before the system actually (functionally) fails. In Fig. 6.16, this process is illustrated by a plot of the system condition versus (operating) time. The point in time where a potential failure can be detected is

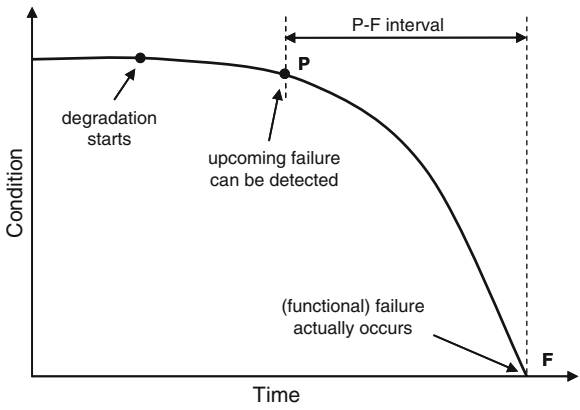

Fig. 6.16 Relation between the observance of a potential failure (*P*) and the actual failure (*F*)

called *P*, while the point of actual failure is called *F*. Therefore, the time period between these two points, that is, the *P–F* interval, is the time available to repair or replace the system to prevent the failure to occur [24].

This interval is also called the delay time [25], defined as the time elapsed from when a future failure could first be noticed until the time when its repair can be delayed no longer because of its consequences. This opportunity window therefore is the period within which the defect should be corrected before it leads to a failure.

It may be clear that the length of the *P–F* interval or delay time is decisive for the success of a CBM policy. If only limited time is available after the detection of an imminent failure, only a very flexible and reactive maintenance organization will be able to prevent failures. On the other hand, a long delay time provides a large opportunity window in which the optimal moment of maintenance activities can be chosen, maximally making use of the benefits of clustering maintenance of similar systems. The length of the interval can be increased by applying high quality sensors, performing appropriate data analysis, but most of all by identifying the proper failure mechanism. If, for example, a system is monitored with a corrosion sensor, but fatigue is the life limiting failure mechanism, the CMS is not expected to give a timely warning for the imminent fatigue failure. Therefore, understanding the failure mechanisms of a system and the associated loads is essential to properly apply condition monitoring, as will be discussed further below.

In addition to the length of the *P–F* interval, also the uncertainty in this interval is an important factor. At the moment of detection (*P*), an estimate must be made of the remaining time to failure, that is, a prognosis of the remaining useful life of the system. If a large amount of uncertainty is present in that estimate, it will be hard to make proper decisions. This is the same problem as the uncertainty in maintenance interval determination that has been discussed in detail in Sect. 6.2. And again, making use of physical failure models will be shown to provide the solution.

In traditional condition monitoring systems, trending methods and growth models are used to extrapolate trends in monitored condition parameters (e.g. vibration levels) to determine component replacement or repair intervals. But similar to the (experience-based) statistical and stochastic methods discussed before, these trending methods lack a physical basis and assume some model form like linear, exponential, etc. (this is often called the black box model). Moreover, the trends are based on historical data, which enables only accurate predictions when the current and future usage of the system is similar to the past usage. For systems that are operated in a variable way, reliable predictions are rather difficult to make in this way.

Knowledge of the physical failure mechanisms can assist in improving the maintenance efficiency. By applying physical model-based prognostic methods, the effects of changes in usage can be taken into account (which turns the black box model into a grey box model). Moreover, understanding the failure mechanism provides valuable information when a new condition monitoring system is developed. These two aspects will be discussed next.

6.5.3.1 Physical Model-based Prognostics

Knowledge of the governing failure mechanism and having access to a quantitative relation between loads and degradation rate (see Fig. 6.2) enables accurate component prognostics for CBM. Changes in usage or loading of the component can easily be incorporated in the prognostic method, and the prediction of the maintenance activities no longer depends on past usage-based trends only.

The large difference with the prognostic methods used for fixed maintenance intervals (Sect. 6.3) is the availability of the monitoring data. This data specifies the actual condition of the component at any moment, which provides two important advantages. Firstly, the prediction only covers a fraction of the total component service life, starting from a known condition. This considerably increases the accuracy of the prognosis, especially when approaching the end-of-life, since the uncertainties in, for example, usage and material properties are reduced. This can be illustrated using the following example. A crack length sensor has been applied to a certain structure. When the crack has propagated to half the critical crack length, a crack growth model can be used to predict (with some uncertainty) the number of additional cycles required to reach the critical crack length, based on the assumed magnitude of the additional cycles. If, on the other hand, no sensor was available, the complete crack propagation process would have to be predicted, starting from the situation with no crack in the structure. Not only the assumption on the expected magnitude of the cyclic load would be more uncertain in this case, but also the moment the crack would initiate and start to grow is uncertain. Therefore, the predicted service life for this structure would be much more uncertain without the sensor data available.

Secondly, the monitoring data may be used to validate and update the prognostic model. Provided that the relevant loads on the component are also monitored (or can be calculated from any other monitored parameter), the present

monitored condition can be compared to the present calculated condition. Any deviation between the measured and calculated condition can be used to update the prognostic model, which reduces the uncertainty in all predictions to come. In this way, an accurate model is obtained quite rapidly. The only requirements are a representative physical model and the registration of the governing loads. If the physical model does not represent the failure mechanism correctly, the calculations will always deviate from the measurements and updating the model parameters is not possible. Monitoring the loads requires some additional effort, but ultimately yields a significant benefit. For the example of the crack sensor system discussed before, this means that the cyclic load on the component has to be monitored (e.g. by applying a strain gauge). Using that data, the crack propagation can be calculated by applying a crack growth law, for example, the Paris' law (Sect. 4.4.6), and can be compared to the measured crack length. Deviations between the calculated and measured crack length can be used to update the values of the model constants (C and n).

To summarize, physical model-based methods improve the accuracy of the prognostics in a CBM concept, since variations in usage can easily be incorporated. At the same time, the monitoring data can be applied to validate and update the used model.

6.5.3.2 Condition Monitoring System Development

Another aspect of CBM that can benefit from knowledge on the physical failure mechanisms is the development of new condition monitoring systems. One of the key issues of developing a new system is the selection of the parameters to be monitored. This selection should be based on a thorough understanding of the component loading and the corresponding failure mechanisms. Only monitoring a condition parameter that is representative for the critical failure mechanism yields the appropriate data that can be used to perform component prognostics. In practice, however, the availability of certain sensors often guides the selection of these parameters, leaving the operator with a huge amount of sensor data that cannot (easily) be translated into useful maintenance information.

In a recent research project [26], a set of guidelines have been developed for setting up a condition monitoring system. These guidelines assist a manufacturer or operator in developing the system in a structured way using the decision tree shown in Fig. 6.17. Typical questions to be considered are as follows:

- is it possible to identify the quantities that govern the maintenance needs of the asset (i.e. critical failure mechanism and associated condition parameter)?
- can these quantities be measured?
- can the (values or trends in) measured quantities be used to predict failures or be translated into maintenance intervals (i.e. prognostics)?
- does application of CBM yield any financial advantage (lower maintenance costs, higher availability) or does it increase the safety level?

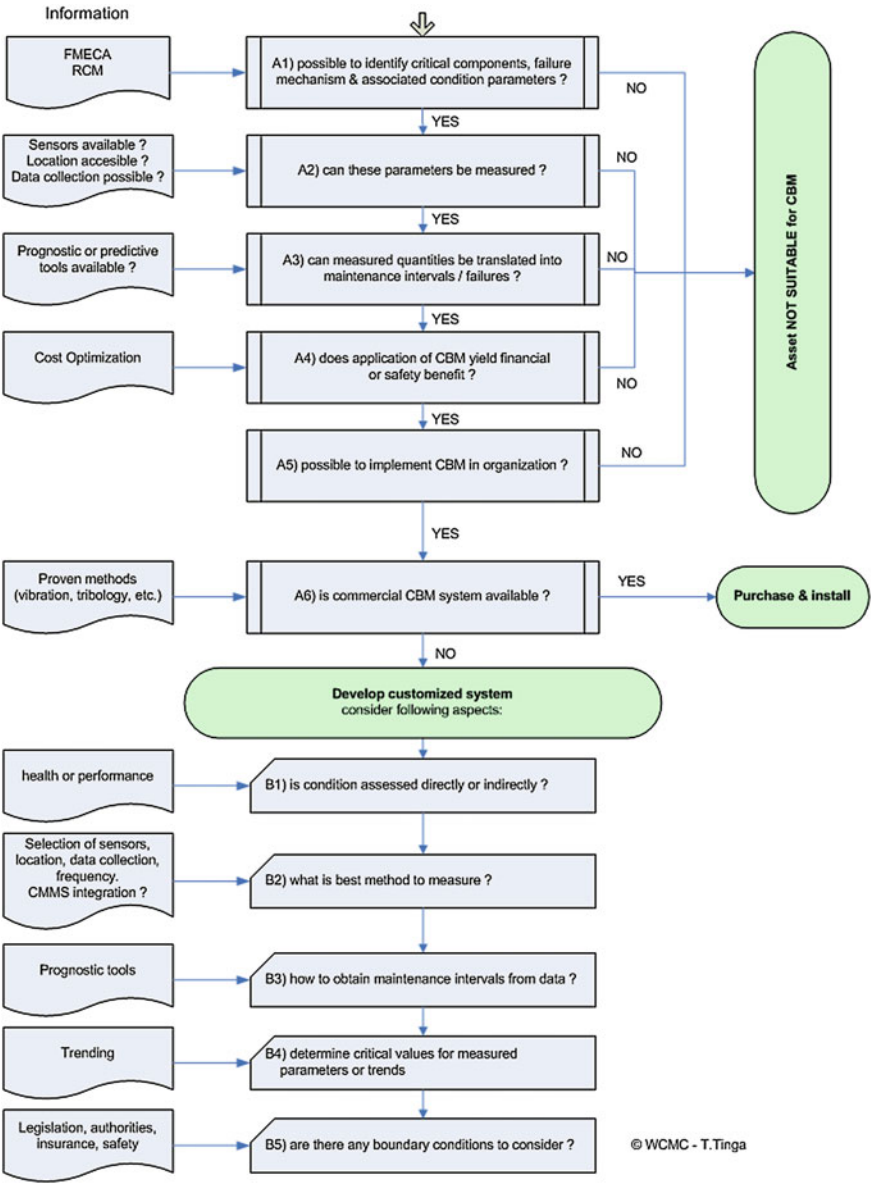

Fig. 6.17 Decision tree for condition-based maintenance system development

When all these questions can be answered positively for a certain asset, it is suitable to be maintained on a condition basis. The next step is then to decide how the CBM system can be realized. That requires detailed answering of the following (technical) questions:

- is the condition assessed directly (vibrations) or indirectly (performance)?
- what is the best method to measure the required quantities?
  - is a suitable sensor available?
  - is the location accessible?
  - is data collection possible (local, remote/online)?
  - what will be the sample frequency (real time or regular inspections)?
- how can the (values or trends in) measured quantities be translated into maintenance intervals (prognostics)?

Most of these questions can only be answered when the physical mechanisms leading to failure of the component are understood. Only then it is possible to select the appropriate quantities to be monitored and to define the prognostic method to be used.

To conclude this section on CBM, it can be stated that the main challenges are the increase of the *P–F* interval and the decrease in the uncertainty in this interval. This section has demonstrated that both challenges can largely benefit from the knowledge of failure mechanisms and their associated loads.

6.6 Summary

In this chapter, the application of knowledge on physical failure mechanisms and their associated loads to develop improved maintenance concepts has been discussed. The chapter started with explaining the role of uncertainty in maintenance interval determination. The more uncertainty is present in the prognostic method, the more conservative the maintenance intervals will have to be, which sincerely affects the efficiency. Applying model-based prognostic methods and adopting usage- and usage severity-based maintenance policies are shown to reduce the uncertainty and thus to increase the maintenance efficiency. For CBM, the same methods are shown to be required for the increase of the delay time (or *P–F* interval) and the reduction of their uncertainty.

References

1. Tinga, T.: Application of physical failure models to enable usage and load based maintenance. Rel. Eng. Syst. Saf. **95**(10), 1061–1075 (2010)
2. Engel, S.J., Gilmartin, B.J., Bongort, K., Hess, A.: Prognostics, the real issue involved with predicting life remaining. In: IEEE Aerospace Conference, pp. 457–469. Big Sky, Montana (2000)
3. Hess, A., Calvello, G., Frith, P., Engel, S.J., Hoitsma, D.: Challenges, issues, and lessons learned chasing the "big P": real predictive prognostics part 2. In: IEEE Aerospace Conference, pp. 1–19, Big Sky, Montana (2006)

4. Hall, P.L., Strutt, J.E.: Probabilistic physics-of-failure models for component reliabilities using Monte Carlo simulation and Weibull analysis: a parametric study. Rel. Eng. Syst. Saf. **80**, 233–242 (2003)
5. Hora, S.C.: Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. Rel. Eng. Syst. Saf. **54**, 217–223 (1996)
6. Helton, J.C., Burmaster, D.E.: Guest editorial: treatment of aleatory and epistemic uncertainty in performance assessment for complex systems. Rel. Eng. Syst. Saf. **54**, 91–94 (1996)
7. Tinga, T.: Physical model based component prognostics. In: Andrews, J., Bérenguer, C., Jackson, L. (eds.) Maintenance Modelling and Applications, pp. 166–184. DNV, Hovik (2011)
8. Farrar, C.R., Lieven, N.A.J.: Damage prognosis: the future of structural health monitoring. Philos. Trans R. Soc. A **365**, 623–632 (2006)
9. Romero, R., Summers, H., Cronkhite, J.: Feasibility study of a rotorcraft health and usage monitoring system (HUMS): Results of operator's evaluation. NASA Lewis Research Center, Cleveland (1996)
10. Lebold, M., Thurston, M.: Open standards for condition-based maintenance and prognostic systems. In: 5th Annual Maintenance and Reliability Conference, TN (2001)
11. Heine, R., Barker, D.: Simplified terrain identification and component fatigue damage estimation model for use in a health and usage monitoring system. Microelectron. Reliab. **47**, 1882–1888 (2007)
12. Wubben, J.P.C.: Usage Profiles and their Effect on Maintenance Intervals for Military Systems. Netherlands Defence Academy, Breda (2010)
13. Banks, J.C., Reichard, K.M., Hines, J.A., Brought, M.S.: Platform degrader analysis for the design and development of vehicle health management systems. In: International Conference on Prognostics and Health Management (2008)
14. Tinga, T., Brekelmans, W.A.M., Geers, M.G.D.: Incorporating strain-gradient effects in a multi-scale constitutive framework for nickel-base superalloys. Phil. Mag. **88**(30–32), 3793–3825 (2008)
15. Tinga, T., Brekelmans, W.A.M., Geers, M.G.D.: Time-incremental creep-fatigue damage rule for single crystal Ni-base superalloys. Mat. Sci. Eng. A **508**, 200–208 (2009)
16. Robinson, E.L.: Effect of temperature variation on the long-time rupture strength of steels. Trans. ASME **74**(5), 777–781 (1952)
17. Tiddens, W.W.: Towards Improving Maintenance Performance in a Military Context. University of Groningen, Groningen (2011)
18. Rao, B.K.N.: Handbook of Condition Monitoring. Elsevier Science Ltd, Oxford (1996)
19. Mobley, R.K.: An Introduction to Preventive Maintenance, 2nd edn. Butterworth-Heinemann, Oxford (2002)
20. Vachtsevanos, G., Lewis, F.L., Roemer, M.J., Hess, A., Wu, B.: Intelligent Fault Diagnosis and Prognosis for Engineering Systems. Wiley, Hoboken (2006)
21. Homborg, A.M., Tinga, T., Zhang, X., van Westing, E.P.M., Oonincx, P.J., de Wit, J.H.W., Mol, J.M.C.: Time-frequency methods for trend removal in electrochemical noise data. Electrochim. Acta **70**, 199–209 (2012)
22. Sohn, H., Farrar, C.R., Hemez, F.M., Shunk, D.D.: A review of structural health monitoring literature: 1996–2001. In: vol. LA-13976-MS. Los Alamos National Laboratory, Los Angeles, (2004)
23. Loendersloot, R., Ooijevaar, T.H., Warnet, L., de Boer, A., Akkerman, R.: Vibration based Structural Health Monitoring and the modal strain energy damage index algorithm applied to a composite T–beam. In: Vibration and Structural Acoustics Analysis: Current Research and Related Technologies. Springer, London (2012)
24. Moubray, J.: Reliability-Centered Maintenance. Industrial Press, New York (1997)
25. Christer, A.H., Waller, W.M.: Delay time models of industrial inspection maintenance problems. J. Oper. Res. Soc. **35**(5), 401–406 (1984)
26. Tinga, T., Soute, D., Roeterink, H.J.H.: Guidelines for Condition Based Maintenance. World Class Maintenance Consortium, Breda (2009)

Further Reading

1. Moubray, J.: Reliability-Centered Maintenance. Industrial Press, New York (1997)
2. Mobley, R.K.: An Introduction to Preventive Maintenance, 2nd edn. Butterworth-Heinemann, Oxford (2002)
3. Vachtsevanos, G., Lewis, F.L., Roemer, M.J., Hess, A., Wu, B.: Intelligent Fault Diagnosis and Prognosis for Engineering Systems. Wiley, Hoboken (2006)

Chapter 7
Reliability Engineering

7.1 Introduction

In the discussion on different maintenance concepts in Chap. 5, a division was made between experience-based and model-based approaches. In the experience-based approach, data on usage or failures of a system are collected, and based on this historic data, predictions of future failures are made. This is extensively done in the field of reliability engineering, where generally a numerical relationship describing the data is adopted. Parametric families of distributions, like exponential and Weibull, have been used as failure functions for decades now. The accuracy of this approach heavily depends on the amount, quality and relevance of the collected data. One of the important issues here is the definition of a functional failure, as was also discussed in Sect. 1.1, for example, in terms of exceeding a certain norm value. Moreover, if the data set is limited or the usage profile of the system has changed considerably, the distribution function based on the data is not representative and accurate prognostics are not possible.

This limitation can partly be solved by following a model-based approach, in which physical models are adopted to calculate the loading and degradation of systems and components. A limited amount of 'experience' or a changing usage profile in this case does not impede the prediction of future failures, since the models quantify the relationship between usage and degradation in a more generic way. This approach has been discussed extensively in Chap. 6.

However, application of the model-based approach may not always be feasible. The development of physical models in some cases takes a considerable effort, and knowledge on the failure mechanisms may not always be available. The experience-based approach is in that sense more easily accessible, since no detailed knowledge of the system and its failure behaviour is required. A collected set of data can rather easily be processed with a wide variety of reliability engineering methods that are readily available. Therefore, if the accessibility of the reliability engineering methods could be combined with the accuracy and robustness of the

T. Tinga, *Principles of Loads and Failure Mechanisms*,
Springer Series in Reliability Engineering,
DOI: 10.1007/978-1-4471-4917-0_7, © Springer-Verlag London 2013

model-based approaches, an optimal methodology would be obtained in many occasions.

The present chapter aims to provide that integration between reliability engineering and physical failure models. The knowledge on the failure mechanisms will be utilized to select the appropriate parameters for a data analysis and to optimally interpret the results. In the next section, the basic concepts of reliability engineering will be introduced briefly. Then, in Sect. 7.3, the challenge of selecting a relevant failure parameter (RFP) is discussed. Sect. 7.4 treats the application of the life exchange rate matrix (LERM) in reliability engineering and discusses its limitations. Then, in Sect. 7.5, the analysis of real failure data is discussed, and finally, Sect. 7.6 treats the concept of stochastic service life analyses.

7.2 Reliability Engineering Basics

The basic principles of reliability engineering will be introduced in this section for further reference in the remainder of this chapter. For a more detailed treatment of the topics in this field, the reader is referred to one of the many specialized books on reliability engineering [1–4].

The time to failure (ttf) of a system or component is generally considered to be a random variable. Its exact value depends on a number of factors, like the usage, loading, environment and strength, which makes it hard to predict exactly. For analysis purposes, it is convenient to find a probability density function (e.g. exponential, normal, Weibull) that correctly represents the distribution of the random values x and to identify the parameters of that function that best fit the values. Such a distribution function is denoted $f(x)$, which means that the probability of the random variable X (e.g. ttf) to be in the interval $[a, b]$ equals

$$P(a \leq X \leq b) = \int_a^b f(x)\mathrm{d}x \tag{7.1}$$

The definition of the cumulative distribution function $F(y)$ is

$$F(y) = P(X \leq y) = \int_{-\infty}^{y} f(x)\mathrm{d}x \tag{7.2}$$

such that the derivative of $F(y)$ is again the probability density function $f(x)$:

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x} \tag{7.3}$$

If the random variable is the ttf, the cumulative distribution function $F(t)$ is also called the failure probability function, since it represents the probability that the ttf is less than a certain lifetime t. Related to this failure function is the survival or reliability function. The reliability $R(t)$ is defined as the probability that a system will still function after a specified period of time t:

$$R(t) = P(\text{ttf} > t) = \int_t^\infty f(x)\mathrm{d}x \tag{7.4}$$

which is closely related to the cumulative density function $F(t)$:

$$R(t) = 1 - F(t) = 1 - \int_0^t f(x)\mathrm{d}x \tag{7.5}$$

Another important quantity in reliability engineering is the hazard rate or failure rate function $\lambda(t)$, representing the instantaneous failure rate. $\lambda(t)$ is a conditional probability, referring to the probability that the system will fail at time t, but under the condition that it is still functioning just before t:

$$\lambda(t)\Delta t = P(t \leq \text{ttf} \leq t + \Delta t | \text{ttf} > t) \tag{7.6}$$

The probability density function $f(t)$ also represents a failure velocity, but in that case, it is unconditional: it refers to the complete population, while $\lambda(t)$ only considers the surviving fraction of the population. The relationship between the hazard rate and reliability can be derived from the previous expression:

$$\lambda(t) = \frac{R(t) - R(t + \Delta t)}{R(t)\Delta t} = -\frac{1}{R(t)}\frac{\mathrm{d}R(t)}{\mathrm{d}t} = \frac{f(t)}{R(t)} \tag{7.7}$$

The evolution of the failure rate during the life cycle of a system is often used to characterize the failure behaviour. If a system shows completely random failure behaviour, as is often experienced for electronic systems and components, the failure rate λ will be constant in time. The age of the system thus does not affect the probability of failure: a new component is as likely to fail as an old component. Contrary to this random failure behaviour, components that have a clear degradation mechanism, like wear, corrosion, creep or fatigue, will exhibit an increasing failure rate. As the components are getting older or have been used more, failure will become more likely.

Finally, also a decreasing failure rate is observed in practice. It is generally associated with infant mortality or running in: if a system is just employed, some unexpected problems will appear, but after a certain period of operation, these failures have been solved and will not occur anymore. For some systems, this behaviour is also observed after a maintenance period, implying that faults are introduced in the system during maintenance. In that case, maintenance raises instead of lowers the failure probability and should be postponed as long as

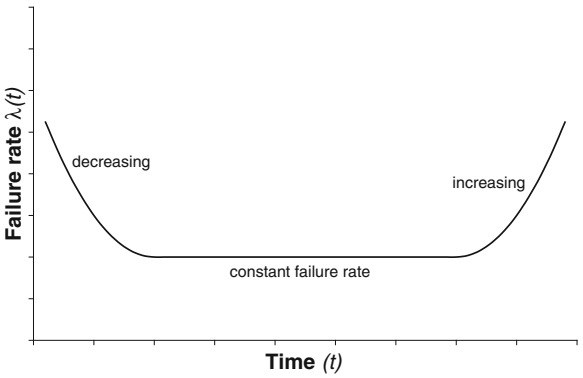

Fig. 7.1 Bathtub curve showing regimes of decreasing, constant and increasing failure rates

possible. Combining the three failure rate regimes, that is, decreasing, constant and increasing, yields the well-known bathtub curve, see Fig. 7.1. This curve typically represents the (ideal) failure behaviour of a system during its life cycle.

Note, however, that in practice, many systems do not behave like this at all. As the different parts of the bathtub curve are associated with different failure mechanisms, the curve shown in Fig. 7.1 is only observed when these mechanisms become active in the considered system at the right moment in time. As maintenance is generally aimed at one specific (the most critical) failure mechanism, considering the failure rate to be either constant, increasing or decreasing is in most cases the best option.

7.2.1 Parametric Probability Density Functions

A number of parametric functions are commonly used in reliability engineering to represent the stochastic behaviour of failure events and repair processes. The three distribution functions shown in Figs. 7.2 (PDF) and 7.3 (CDF), the exponential,

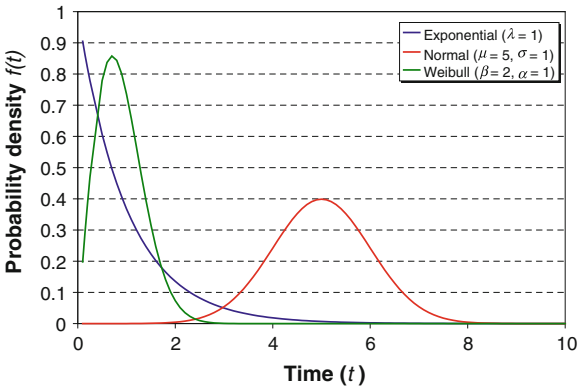

Fig. 7.2 Probability density functions for exponential, normal and Weibull distribution

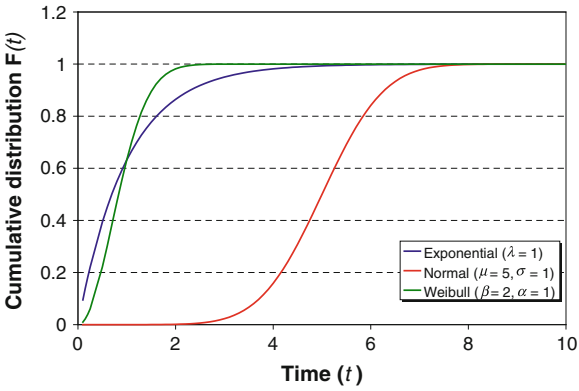

Fig. 7.3 Cumulative distribution functions for exponential, normal and Weibull distribution

normal and Weibull distributions, will be shortly discussed here. All these functions have one or two parameters, whose value must be determined from the considered data set using statistical methods.

The probability density function $f(t)$ and cumulative distribution function $F(t)$ for the exponential distribution with parameter λ are given by

$$\begin{aligned} f(t) &= \lambda e^{-\lambda t} \\ F(t) &= 1 - e^{-\lambda t} \end{aligned} \tag{7.8}$$

For the exponential distribution, the failure rate $\lambda(t)$ is constant and equal to the parameter λ. Due to its simple form and convenient properties, the exponential distribution is widely used in maintenance modelling, although perfectly random failure behaviour is not representative for many components in the real world.

The normal distribution is defined as follows:

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) \tag{7.9}$$

where μ represents the mean value and σ the standard deviation. The functions $f(t)$ and $F(t)$ for the normal distribution are plotted in Figs. 7.2 and 7.3.

Finally, the probability density function for the Weibull distribution is given by

$$f(t) = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} \exp\left[-\left(\frac{t}{\alpha}\right)^{\beta}\right] \tag{7.10}$$

where α is the scale parameter and β is the shape parameter. If the Weibull distribution is used to describe the ttf of a set of components, the scale parameter represents the characteristic life of the population and the shape parameter indicates the failure behaviour. For $\beta = 1$, the distribution reduces to an exponential distribution, and consequently, the failure rate is constant in time. For $\beta < 1$, the failure rate is decreasing (infant mortality), and for $\beta > 1$, the failure rate increases. In failure data analysis, typically a distinction is made between

$1 < \beta < 4$ and $\beta > 4$. In the former case, the system is gradually degrading and an optimal preventive replacement interval can be determined. In the latter case, the system is rapidly degrading, which typically only occurs near the end of the service life. Performing inspections to detect the onset of this stage of the life cycle is the most suitable policy in such a case.

7.2.2 Mean Time to Failure and Mean Time to Repair

A quantity that is often used in practice to represent the reliability of a system is the mean time to failure (MTTF). The MTTF is the average value of the ttf and can be obtained either from a set of failure data (see Sect. 7.2.3) or from the probability density function as

$$\mathrm{MTTF} = \int_0^\infty t f(t)\mathrm{d}t = -\int_0^\infty t \frac{\mathrm{d}R(t)}{\mathrm{d}t}\mathrm{d}t \tag{7.11}$$

Using the integration by parts technique,

$$\mathrm{MTTF} = |-tR(t)|_0^\infty + \int_0^\infty R(t)\mathrm{d}t = \int_0^\infty R(t)\mathrm{d}t \tag{7.12}$$

since[1]

$$\lim_{t\to\infty}[tR(t)] = \lim_{t\to\infty}\left[t\exp\left(-\int_0^t \lambda(x)\mathrm{d}x\right)\right] = 0 \tag{7.13}$$

In the special case of a constant hazard rate (exponential distribution), determination of the MTTF is straightforward:

$$\mathrm{MTTF} = \int_0^\infty R(t)\mathrm{d}t = \int_0^\infty e^{-\lambda t}\mathrm{d}t = \left|-\frac{1}{\lambda}e^{-\lambda t}\right|_0^\infty = \frac{1}{\lambda} \tag{7.14}$$

Instead of the MTTF, also the mean time between failures (MTBF) is often used in practice. The difference between MTBF and MTTF is the repair period following a failure, as is indicated in Fig. 7.4. The typical length of the repair period is denoted as the mean time to repair (MTTR). In a similar way, as for the MTTF, also the MTTR can be obtained from a probability density function. This function,

[1] As $\int_0^t \lambda(t)\mathrm{d}t = \int_{R(0)=1}^{R(t)} (-\mathrm{d}R(t)/R(t))$, it is derived that $R(t) = \exp\left(-\int_0^t \lambda(x)\mathrm{d}x\right)$

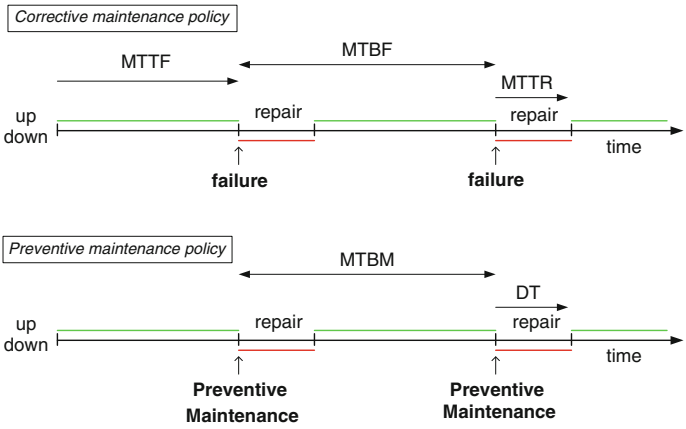

Fig. 7.4 Definitions of mean failure and repair times

which is denoted $g(t)$, defines the distribution of the time to repair (ttr). Analogous to the failure probability function $F(t)$, the maintainability $M(t)$, being the probability that a component will be repaired at the moment t, is defined as

$$M(t) = \int_0^t g(x)\mathrm{d}x \tag{7.15}$$

The associated conditional repair rate function $\mu(t)$ is very similar to the failure rate $\lambda(t)$. As the ttr is commonly also assumed to be exponentially distributed with parameter μ, the MTTR in that case is given by

$$\mathrm{MTTR} = \frac{1}{\mu} \tag{7.16}$$

In Sect. 5.6, on maintenance performance, the term mean time between maintenance (MTBM) was introduced. This term is very similar to the term MTBF, but it is applicable to a preventive maintenance policy setting, where actual failures generally do not occur. As is indicated in Fig. 7.4, the repair process in that case is a replacement process represented by the associated downtime (DT).

7.2.3 Non-parametric Reliability Evaluation

In the previous part of this section, the failure probability and reliability of a system have been described by parametric functions. However, in real applications, a set of failure data are obtained, which generally do not exactly match a parametric distribution. In that case, analysing the data based on the assumed parametric function (and its fitted parameters) may lead to inaccurate results.

Therefore, in this subsection, some evaluation methods based on real data will be discussed.

Several methods are available to analyse a set of failure data. In the following, it is assumed that an ordered set of n failure times t_1, t_2, t_3,…, t_n with $t_i \leq t_{i+1}$ are available. A reliability plot of this data can then be constructed using the direct method. If i is the number of failures within the set at time t_i, an estimation of the reliability function $R(t)$ at time t_i can be made by calculating the fraction of the components that survived up to t_i

$$\hat{R}(t_i) = \frac{n-i}{n} = 1 - \frac{i}{n} \tag{7.17}$$

From this expression, also the estimates for the functions $F(t)$, $f(t)$ and $\lambda(t)$ can be obtained rather easily.

The direct method estimates the reliability to be zero from the moment of the latest failure in the data set. However, since the data set only contains random samples, it is not very likely that t_n is actually the longest possible survival time. Therefore, an improved direct method was proposed, estimating $R(t)$ at t_i as

$$\hat{R}(t_i) = 1 - \hat{F}(t_i) = 1 - \frac{i}{n+1} = \frac{n+1-i}{n+1} \tag{7.18}$$

The MTTF can also be estimated from the data set, according to

$$\text{MTTF} = \sum_{i=1}^{n} \frac{t_i}{n} \tag{7.19}$$

These methods can be applied to sets of failure data representing actually failed components. However, in practice, components are often preventively replaced before failure occurs. The time of replacement is often included in the data set, but does not represent a ttf. Nevertheless, this type of data can be incorporated in the analysis, since it provides useful information (i.e. that the component has survived a certain time period).

A data set containing such incomplete data is called a censored data set. Three different types of censoring exist [1]:

1. *right-censored data*: the part has not failed yet;
2. *left-censored data*: the part has failed, but the time it entered service is unknown;
3. *interval-censored data*: exact failure times are unknown, only the number of failures in a certain period is recorded.

To include the censored data in the reliability analysis, the methods discussed above must be modified. Assume that in a set of n parts, only $r < n$ failed and the remaining $n–r$ parts have been replaced preventively. This means that the data set is right-censored. If a part has failed, a failure time t_i is available; otherwise, the

replacement time t_i^+ is used. Applying the improved direct method discussed above, it follows that

$$\frac{\hat{R}(t_i)}{\hat{R}(t_{i-1})} = \frac{\left(\frac{n+1-i}{n+1}\right)}{\left(\frac{n+2-i}{n+1}\right)} = \frac{n+1-i}{n+2-i} \tag{7.20}$$

and then the recursive relation

$$\hat{R}(t_i) = \frac{n+1-i}{n+2-i}\hat{R}(t_{i-1}) \tag{7.21}$$

is obtained. This expression can now be used to step-wise analyse the data set. If a real failure is encountered at t_i, the reliability is estimated using Eq. (7.21). For a censored time t_i^+, the reliability is estimated by the reliability at t_{i-1}. The initial reliability is set to $R(0) = 1$.

A similar method is the Kaplan–Meier approach, which adopts a slightly different recursive equation:

$$\hat{R}(t_i) = \left(1 - \frac{1}{n+1-i}\right)\hat{R}(t_{i-1}) \tag{7.22}$$

An application of these nonparametric methods will be shown in Sect. 7.5.

7.3 Relevant Failure Parameter

As was discussed in Sect. 6.2, the key issue of preventive maintenance concepts is the determination of the length of maintenance intervals. To guarantee the effectiveness of a maintenance concept (quantified by a certain allowable probability of failure), it has been shown in Sect. 6.2 that more conservatism is required when the uncertainty about the degradation process increases [5]. The total uncertainty in a failure distribution is the combined effect of a range of uncertainties in the underlying processes, like usage, loads and material properties. However, if a number of these uncertainties can be quantified, the total uncertainty can be reduced considerably. The challenge is therefore to find the parameter that provides the best correlation with the ttf. This parameter will be denoted here as the relevant failure parameter (RFP).

Traditionally, failure distributions are plotted in terms of ttf, where generally the calendar time is considered. The average time between the moment of installing a new component and the moment of failure, the MTTF, is used to describe the components' failure behaviour. However, when the usage of the component is variable in time, this is not the most suitable parameter. In line with the concepts of model-based prognostics (see Fig. 6.2) discussed in the previous chapter, this section will demonstrate that there are three steps to improve the RFP in describing the component service life:

1. Incorporate information on the usage,
2. Incorporate information on the severity of usage or loading,
3. Incorporate information on the condition.

Note that these steps fully correspond to, respectively, the UBM, USBM/LBM and CBM maintenance policies introduced and discussed in Chaps. 5 and 6.

These policies as applied in the gas turbine blade case study in Sect. 6.4.2 are shortly summarized here again:

- *Calendar time–based maintenance* (*CTBM*): fixed periods of calendar time are used, for example every month or every year. This means that no information about the usage or loads is considered. It is even unknown whether or not the system is actually operated in a certain period. This is the traditionally used parameter in failure distributions.
- *Usage-based maintenance* (*UBM*): the usage is incorporated in the determination of the maintenance intervals, for example by using the number of operating hours or cycles.
- *Usage severity–based maintenance* (*USBM*): the usage severity during operation of a system is taken into account by monitoring the variation in usage (e.g. power setting, rotational speed). This requires knowledge of the relationship between usage severity and life consumption, which is provided by suitable physical models.
- *Load-based maintenance* (*LBM*): the usage severity during operation of a system is taken into account by monitoring the actual relevant internal loads in the component (e.g. temperature, strain). A suitable physical failure model then yields the life consumption associated with the measured loads.
- *Condition-based maintenance* (*CBM*): the condition of the system is assessed in a (almost) direct way, by using appropriate sensors (e.g. crack length sensors). The amount of uncertainty is limited, and maintenance can be performed just-in-time.

Now, the failure distributions for a population of similar components are plotted as they are assumed when these five different maintenance policies are applied. The result is shown in Fig. 7.5. The large difference in distribution width observed for the five different policies is caused by the fact that five different RFPs are used on the horizontal axis. For the gas turbine blade, the service life is creep dominated. Therefore, the RFPs used for the five maintenance policies in Fig. 7.5 are the following:

• CTBM:	Calendar time (days)
• UBM:	Operating time (hours)
• USBM:	Equivalent operating time, that is, weighted sum of operating time at low, middle and high power setting (equivalent hours)
• LBM:	Equivalent operating time, that is, weighted sum of operating time at three stress and temperature levels (equivalent hours)
• CBM:	Condition parameter, that is, blade elongation (mm)

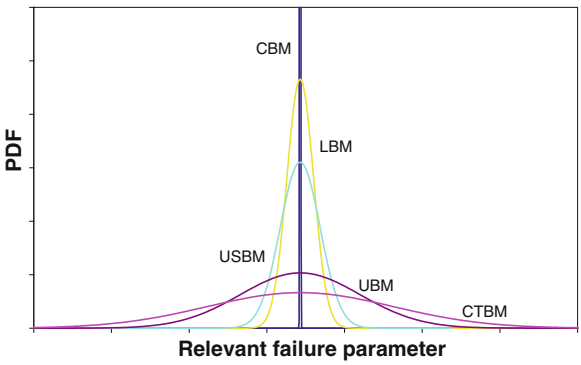

Fig. 7.5 Failure distributions for a population of components using five different relevant failure parameters (*RFPs*). The RFPs depend on the applied maintenance policy

The definition of these parameters and their motivation will be explained in the next subsections. Further, note that (by appropriately scaling the horizontal axis) the five distributions have been plotted such that the mean values of the distributions coincide.

The results in Fig. 7.5 can be explained as follows. With the CTBM policy, the variation in failure time (in terms of calendar time) appears to be very large. The reason is that some gas turbines in the population are operated intensively, while others have much less operating time or are used less severely. Therefore, the average service life of the blades is around 10 months, but after 16 months, at least 5 % of the blades are still expected to be in service. The variation in expected service life is thus extremely large, and an efficient maintenance policy is hard to establish.

For the CBM policy, it is completely different, since not the calendar time is used as RFP, but the directly monitored blade elongation. Because the failure criterion in terms of elongation is accurately known, that is, 3 mm, all blades are expected to fail very close to this limit, as is illustrated by the very narrow distribution in Fig. 7.5. Based on the (periodic or continuous) blade length measurements, the optimal moment for replacement can thus be determined accurately and maintenance can be performed just-in-time.

The failure distributions $f(t)$ in Fig. 7.5 can also be plotted as reliability curves, as is shown in Fig. 7.6. The reliability $R(t)$ was defined in Eq. (7.4).

The reliability curves can conveniently be used to determine the moment in time where the reliability drops below a critical level, for example 95 % as indicated by the dashed line. At this value of the RFP, action should be taken to prevent regular failures. The figure shows that an increasing uncertainty requires action at an earlier point in time and thus yields a less efficient maintenance process.

In the next three subsections, the three steps in defining more sophisticated and better correlating RFPs will be discussed.

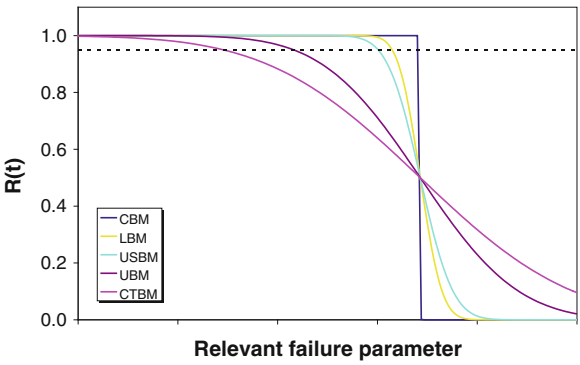

Fig. 7.6 Reliability curves for a population of components using five different relevant failure parameters (*RFPs*). The RFPs depend on the applied maintenance policy

7.3.1 Incorporating the Usage

As was mentioned before, the essence of the RFP is that for every failure (mechanism), the parameter must be found that quantifies the service life of the component the best. The first step in this search is looking for the relevant usage parameter. For example, for a structure that is continuously exposed to a corrosive environment, elapsed calendar time since the start of the exposure is probably a good quantity to describe the service life. However, as was shown in Chap. 4, the fatigue life of a component is better expressed in the number of stress cycles, while adhesive wear is best quantified by the sliding distance. Note that in literature, other terms have been proposed to incorporate the usage of the system, for example 'maintenance relevant quantity' used by Smit [6].

To demonstrate the concept once more, an example of an airline with a fleet of aircraft will be used. Many of the components in the aeroengines have a fatigue-dominated service life. For these components, the number of operating hours (UBM) generally does not correlate very well with the service life, while the correlation with the number of stress cycles (LBM) is much better. This means that number of cycles should be selected as a usage parameter rather than operating hours in this case.

It is assumed that each flight provides one major fatigue cycle and that failure occurs for the considered part after 10,000 cycles. For regional flights, with a typical duration of 2 h, on average, one half cycle per flight hour is accumulated. This means that the ttf is around 20,000 flight-hours. Intercontinental flights of 5 h, on average, only face 0.2 cycle per flight hour, implying that 50,000 flight-hours can be accumulated before the part fails. This is visualized in Fig. 7.7, where in the left-hand plot, the failure distribution in terms of operating hours is shown. The large width of this distribution is due to the variation in flight durations: short flights are associated with relatively short times to failure, and performing long flights increases the average lifetime. However, if the failure distribution is plotted as a function of number of cycles to failure, as is shown in the right-hand-side plot

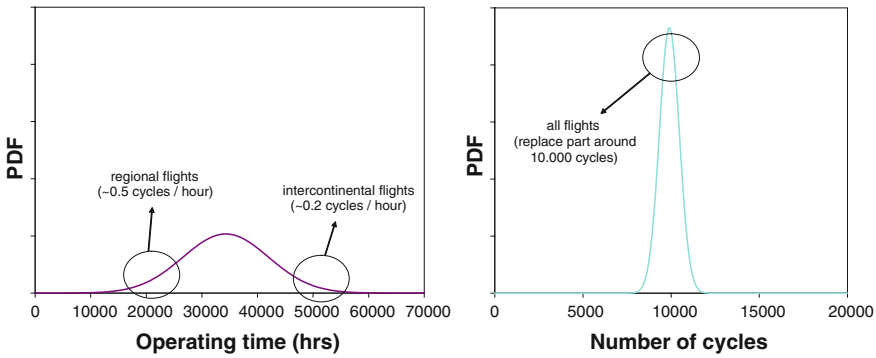

Fig. 7.7 Failure distribution functions in terms of operating hours and number of cycles. The various types of flights have significantly different times to failure, but similar number of cycles to failure

of Fig. 7.7, the type of flight has no effect anymore, since all parts are expected to fail around 10,000 cycles.

The selection of the appropriate usage parameter is only possible when the physical failure mechanisms are understood and the internal loads that govern this mechanism are known. In the first part of this book, the failure mechanisms and their governing loads have been treated extensively. In summary, the appropriate usage parameters that should be adopted as RFP for the most common mechanisms are as follows:

- Fatigue → number and magnitude of stress or strain cycles,
- Creep → temperature, stress level and time,
- Wear → normal load and sliding distance,
- Corrosion → exposure time and environmental conditions.

This knowledge enables the first improvement in the maintenance process efficiency, since replacement intervals of components can be calculated in terms of the variations in the dominating load. However, in this way, only the usage is taken into account, whereas the usage severity (how damaging is each cycle or operating hour?) is not considered. To quantify the differences in usage severity, a physical failure model can be used to define a more sophisticated relative failure parameter. This will be discussed in the next subsection.

7.3.2 Incorporating the Usage Severity or Loads

Whereas the selection of an appropriate usage parameter is acknowledged by many to improve the maintenance efficiency and is also commonly applied now, the additional potential of incorporating usage severity has not been widely recognized yet. This concept will be demonstrated in the present subsection by extending the

gas turbine blade case study introduced in Sect. 6.4.2. The basic idea is that the number of operating hours still provides a limited correlation with the component service life, as the operating conditions and associated damage rate largely vary across the operating hours. To be more specific, an hour operated at high power yields much more damage in the blade than an hour at a low power setting. By applying physical models to quantify the damage rates, this difference can be incorporated in a RFP.

In the previous chapter, physical models have been selected to calculate the internal loads and the resulting creep strain accumulation in the low-pressure turbine blade. For convenience, the equations will be repeated here again. The blades rotate at speeds in the order of 10,000 revolutions per minute (rpm) and operate in a hot gas stream with temperatures in the range of 500 to 1,000 °C. Due to the rotation, a centrifugal force acts on the turbine blades, causing radially directed normal stresses in the root of the blade. The magnitude of these stresses depends on the mass (m) of the blade, the rotational speed (ω), the distance from the blade to the engine centre axis (r) and the area of the blade cross section (A) in the following way:

$$\sigma = \frac{F}{A} = \frac{m\omega^2 r}{A} \tag{7.23}$$

The rotational speed and gas temperature depend on the power setting of the gas turbine. Since the turbine blades are solid and uncooled, the blade temperature will equal the gas temperature in a steady-state situation.

For the present component, creep is the life-limiting failure mechanism. Creep is a high-temperature deformation process that depends on both the stress and temperature levels. The creep strain rate for the present material is described by the Norton's creep law and depends on the temperature (T) and stress (σ) as follows:

$$\dot{\varepsilon}_{\mathrm{cr}} = \frac{\mathrm{d}\varepsilon_{\mathrm{cr}}}{\mathrm{d}t} = AT^4\sigma \tag{7.24}$$

where the value of the constant A equals $2{\cdot}10^{-20}$ $(\mathrm{MPa})^{-2}$ $(^{\mathrm{o}}\mathrm{C})^{-4}$ $(\mathrm{hour})^{-1}$. A creep strain of 1.0 % is defined as the critical amount of creep deformation ($\varepsilon_{\mathrm{crit}}$), leading to an unacceptable elongation of the blade. Note that it is assumed that the creep deformation up until 1 % creep strain is mostly in the secondary creep regime, which means that the creep strain rate is constant. The ttf can then be calculated as

$$t_f = \frac{\varepsilon_{\mathrm{crit}}}{\dot{\varepsilon}_{\mathrm{cr}}} \tag{7.25}$$

The accumulated amount of damage (D) can be obtained using Robinson's damage rule [7]:

$$D = \sum_i \frac{\Delta t_i}{t_{f,i}} \tag{7.26}$$

where Δt_i is the time period spent at some condition (stress and temperature) and $t_{f,i}$ the failure time at those conditions. Failure will occur when the damage parameter D attains the value 1.

The usage of the gas turbine is defined in terms of operating hours per year and a usage severity based on the fractions of time spent at low, middle and high power settings, respectively (see Figs. 6.6 and 6.7). The induced blade loads and resulting creep damage accumulation are shown in Table 7.1.

The creep strain rates are a good indication of the usage severity, so these quantities are used here to define a RFP in terms of equivalent operating hours:

$$\Delta t_{\text{eq}} = \frac{\dot{\varepsilon}^{\text{cr}}_{\text{low}}}{\dot{\varepsilon}^{\text{cr}}_{\text{high}}} \Delta t_{\text{low}} + \frac{\dot{\varepsilon}^{\text{cr}}_{\text{mid}}}{\dot{\varepsilon}^{\text{cr}}_{\text{high}}} \Delta t_{\text{mid}} + \frac{\dot{\varepsilon}^{\text{cr}}_{\text{high}}}{\dot{\varepsilon}^{\text{cr}}_{\text{high}}} \Delta t_{\text{high}} \tag{7.27}$$

where Δt_{low}, Δt_{mid} and Δt_{high} are the numbers of hours spent at the three power settings and $\dot{\varepsilon}^{\text{cr}}_i$ are the associated creep strain rates. Using the numbers from Table 7.1, the following expression for the equivalent operating hours is obtained:

$$\Delta t_{\text{eq}} = 0.03\ \Delta t_{\text{low}} + 0.24\ \Delta t_{\text{mid}} + 1.00\ \Delta t_{\text{high}} \tag{7.28}$$

If all operating hours would be run at high power, Δt_{eq} equals the number of operating hours. If the machine is operated at lower power settings for a certain period of the time, Δt_{eq} will be lower than the number of actual operating hours. From Table 7.1, it appears that at the high power setting, the failure time is 2,432 h. Since the usage parameter in (7.27) is normalized to this high power setting, the failure time in terms of equivalent hours is also 2,432 h.

Now the effect of the usage on the lifetime of the blade is known, a sensitivity study is performed with different usage profiles. It is assumed that always 20 % of the operating hours are run at middle power. Then, the fraction of time operating at low power is varied from 0 to 80 % (which means that the remaining hours are run at high power). Using the equivalent time parameter as defined in (7.27), the resulting service life of the gas turbine is calculated, as is shown in Fig. 7.8. It can be observed that for low fractions of time at low power, the service life approaches the 2,434-h service life associated with the high power setting. For an increasing fraction of time at low power, the service life considerably increases.

Table 7.1 Turbine blade loads and creep damage accumulation at different power settings

Power setting	Rotational speed (rad/s)	Stress (MPa)	Blade temperature (°C)	Creep strain rate (hour^{-1})	Failure time (hours)
Low	597	102	500	$1.27 \cdot 10^{-7}$	78,587
Middle	733	154	750	$9.72 \cdot 10^{-7}$	10,293
High	1,047	313	900	$4.11 \cdot 10^{-6}$	2,432

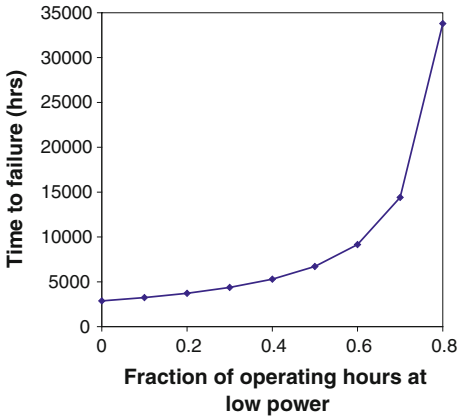

Fig. 7.8 Effect of usage severity on gas turbine blade creep life

In this sensitivity study, it is assumed that the power ratios are fixed during the complete service life. In practice, this will generally vary in the course of time, which means that monitoring the actual ratios provides the information to assess the damage accumulation in the blades. It should be stressed that the proposed RFP can only be applied when the hours at the three power settings are actually monitored.

Further, the RFP presented here takes into account the severity of the usage to a certain extent. The power settings are divided into three classes only, which for some applications may be too course. However, it may be relatively easy to monitor these three classes in a USBM maintenance strategy. If the internal loads can be monitored directly (LBM), a much finer classification of the usage variations is possible. Also, a usage parameter now has been presented for a creep failure-dominated component, but similar parameters can be defined for fatigue (equivalent cycles) or wear (equivalent sliding distance).

To conclude this subsection, it can be stated that adopting an expression like (7.27) as the RFP incorporates the usage severity in the parameter, and enables to differentiate between operating hours at different conditions and severity. Therefore, the accuracy and efficiency of a maintenance policy can be increased by applying sophisticated RFPs, but the additional effort in monitoring the relevant usage parameters is a strict requirement.

7.3.3 Incorporating the Condition

If it is possible to monitor the condition of the considered part, that would enable the final step in improving the RFP. As a periodic or continuous condition assessment provides the most direct information on the damage accumulation process within the part, the uncertain process of predicting the deterioration based

on the usage is not required anymore. As is shown in Fig. 7.5, this yields a failure distribution with a very limited amount of uncertainty.

However, as was discussed in Sect. 6.5, condition monitoring is not always feasible (either technically or economically). In that case, the RFPs discussed in the previous two subsections constitute useful methods to predict future failures and to determine the optimal maintenance intervals.

7.4 Life Exchange Rates

For a system that contains several critical components, the reliability of the complete system depends on the reliabilities of the individual components, as is generally calculated using reliability block diagrams (RBDs). For three components that all are critical to the functionality of the system, the RBD consists of three reliabilities $R_i(t)$ in series. The system reliability $R_s(t)$ is then obtained by

$$R_s(t) = \prod_{i=1}^{3} R_i(t) \tag{7.29}$$

This equation assumes that all individual component reliabilities are given as a function of the same variable t (generally representing operating time). However, in the previous section, it has been stated that the selection of the appropriate RFP is very important. Therefore, it is quite probable that the reliabilities of the three components are given as a function of three different variables, for example hours, cycles and kilometres. In that case, Eq. (7.29) cannot be used to calculate $R_s(t)$ and a life exchange rate matrix (LERM) is often used [2].

The LERM is a matrix containing factors to exchange two parameters, for example hours and cycles. The entries $r_{i,\,j}$ have the following meaning: one life unit of parameter i (t_i) equals $r_{i,\,j}$ times a life unit of parameter j (t_j). Therefore, Eq. (7.29) can be rewritten as

$$R_s(t_i) = \prod_{j=1}^{3} R_j(t_j) = \prod_{j=1}^{3} R_j(r_{j,i}t_i) \tag{7.30}$$

Apart from the application of the LERM in RBDs, this theory can also be used to transform a single reliability (or failure distribution) from one to another usage parameter. For example, the component reliability in terms of cycles can be transformed into an hour-based reliability if the number of cycles per hour is known. However, this cannot be done unconditionally, as will be discussed next.

In Sect. 7.3, it was argued that always the most specific RFP should be used to minimize the uncertainty in the failure distribution and reliability functions. The actual shape of the distribution can then be obtained by either calculating (using the physical models) or measuring the failure times. However, in the latter case, the required narrow distribution is only obtained when the actual usage (severity)

is monitored, thus enabling the assessment of the selected RFP. Eq. (7.30) suggests that a distribution in terms of a less specific usage parameter (e.g. operating hours instead of cycles for a fatigue case) can be transformed into the required reliability function, but in that case, the required narrow distribution is not obtained. Only, a distribution equivalent to the wider original distribution is obtained, since the actual severity of the usage is still not specified.

This can be illustrated using the example of the airline from Sect. 7.3.1 and Fig. 7.7. Having collected the failure data in terms of operating hours, as represented by the distribution in the left-hand plot of Fig. 7.7, an average number of cycles per flight hour could be used to transform the distribution to cycles. An exchange rate between flight-hours and cycles of 0.33 cycles/hour produces the following LERM:

$$\text{LERM} = \begin{bmatrix} 1 & 0.33 \\ 3 & 1 \end{bmatrix} \tag{7.31}$$

Applying this LERM to the failure distribution in Fig. 7.7 provides the same distribution function in terms of cycles, with only the x-axis scaled with a factor 3, see Fig. 7.9. It is clear that this distribution is much wider than the failure distribution in terms of cycles in the right-hand plot of Fig. 7.7.

The reason for the large width of the distribution in Fig. 7.9 is the use of a fixed exchange rate of 0.33 cycles per hour. As was discussed in Sect. 7.3.1, the number of cycles per flight hour varies significantly, especially between regional and intercontinental flights. Only when the actual ratio for each individual flight is used in the transformation of the distribution function, the narrow distribution in the right-hand side of Fig. 7.7 is obtained. This demonstrates that a distribution function in terms of a less specific failure parameter cannot be transformed into a distribution of a more specific parameter without monitoring the variations in that more specific parameter.

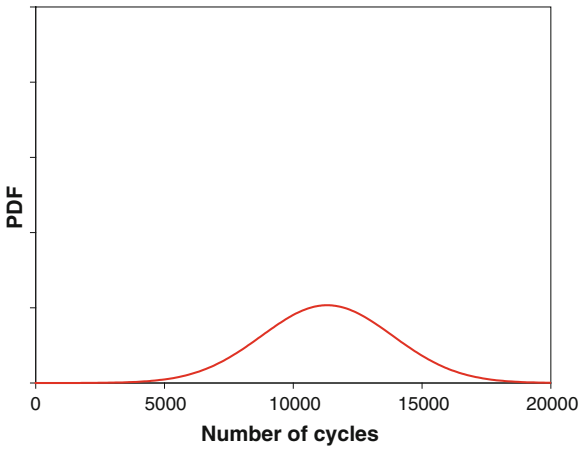

Fig. 7.9 Failure distribution function in terms of number of cycles, as obtained from transforming the operating hours distribution function using a fixed exchange rate

On the other hand, using the LERM to transform a distribution in terms of a very specific usage parameter into a more general parameter (e.g. from cycles to hours for the fatigue case) yields a quite narrow distribution based on the specific value of the exchange rate used. However, this is quite dangerous. If for some reason, the exchange rate changes (e.g. more cycles per hour), the original distribution in terms of cycles remains the same, but the transformed distribution in terms of operating hours would have to shift and is thus not representative anymore. Failure will occur at a lower number of operating hours, since the critical number of cycles will have accumulated earlier. If the assumption of the fixed exchange rate is unknown, application of the hour-based distribution may lead to unconservative maintenance planning and associated failures.

For example, if for a gas turbine, the average cycle to hour ratio has been determined by monitoring the fleet of a specific operator, the cycle-based reliability can be transformed into an hour-based reliability using that average exchange rate. But if one of the gas turbines is operated such that much more start–stops are made than on average, the time-based reliability is not representative for this machine anymore. Also, another operator probably uses the same gas turbines quite differently, which also means that the hour-based reliability cannot be used to perform an accurate analysis.

The conclusion of this subsection is that the use of exchange rates seems to be a convenient way to transform distributions, but care should be taken when applying only a fixed (e.g. average) exchange rate. On the one hand, the exchange rates cannot be used to transform distributions in terms of general usage parameters into the more precise distributions in terms of a specific usage parameter. And on the other hand, transforming back specific distributions into more general distributions is dangerous, since the transformation is based on one specific exchange rate, which is not representative for many other situations.

7.5 Interpretation of Failure Data

If a sufficiently large set of failure data are available for a system or component, a reliability analysis can be performed, providing insight into, for example, the MTBF and the change in system reliability as time proceeds. However, care should be taken in interpreting both the failure data and the results of the reliability analysis. The failure data set is just a set of numbers (e.g. times to failure for a population of components), but the data in the set may not always be homogeneous. For example, although it is generally assumed that replacing a component makes the system ‘as good as new’, in practice that might not always be the case. Also, replacing components preventively, so before actual failure of the component occurs, affects the data set, since failure data and replacement data are mixed.

In this section, a case study on fuel injectors of a diesel generator will be used to demonstrate that analysing a set of failure data appropriately and interpreting the

results of the associated reliability analysis require insight into the failure behaviour and the maintenance policy of the system.

7.5.1 Case Study Description

On a fleet of 16 diesel generators used to generate electric power, the fuel injectors are observed to fail regularly. Normal practice for these machines is that failures are registered in the computerized maintenance management system (CMMS), where only the calendar date is recorded. The number of operating hours at each failure cannot be retrieved from the system. Also, only failure information on the system level (the generator) is stored, so no data on failures of individual fuel injectors is available. As can be learnt from Sect. 7.3 on the RFP, analysing this data set will yield results with a large amount of uncertainty and detailed insights are not expected to be obtained.

To be able to solve the problem, the operator started to collect more detailed failure data. The moment of replacement of each injector was registered, including the number of operating hours of the engine at that moment and the position of the injector (each generator contains 12 identical fuel injectors) on the generator. After a certain period, this resulted in a data set of 315 injector replacements.

The maintenance policy applied to the generators is such that after 4,000 operating hours, all 12 injectors are replaced. However, it occasionally occurs, for example, due to exceptional operating conditions, preventive replacement takes place before the limiting number of operating hours is reached. Moreover, if an individual injector is detected to malfunction, it is also replaced (i.e. corrective maintenance). The removed injectors (defective or preventively replaced) are sent to a workshop, where they are repaired and sent back to the stock point.

Using the complete set of replacement times, a reliability plot for the injectors can be constructed, as is shown in Fig. 7.10. The plot shows that the reliability decreases almost linearly, reaching $R = 0$ at a maximum number of 5,000 operating hours. This means that replacements are performed at a wide range of ages and certainly not only at the prescribed interval of 4,000 h, although a relatively large drop in the reliability is observed between 4,000 and 4,400 operating hours. The number of operating hours at a reliability of 50 % can be considered as the typical service life of the injectors. In this case, that value equals 2,373 operating hours.

As the plot in this form does not provide much insight into the actual reliability of the injectors, a more detailed analysis of the data is required. Two aspects will be investigated more in-depth: (1) the difference between actual failures and preventive replacements and (2) the assumption that replaced injectors are as good as new.

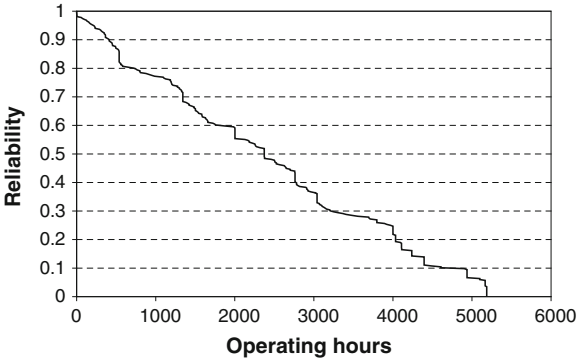

Fig. 7.10 Reliability plot of fuel injectors, based on full set of replacement data

7.5.2 Actual Failures versus Preventive Replacements

In the registration of the fuel injector replacements, the reason for the replacement, being either a malfunctioning of the injector or a (periodic) preventive replacement, could be retrieved. As a result, the complete data set can be split into a set of actual failures and an additional set of replaced (but not failed) injectors. The reliability plot for the actually failed injectors is shown in Fig. 7.11.

This curve initially decreases rather fast up until 1,500 operating hours, then becomes less steep, but shows another faster decrease around 3,000 h. This suggests that not all injectors behave identically, but two groups of injectors could be identified. This will be investigated further in the next subsection. Further, the typical service life (at $R = 50$ %) in this case equals only 2,000 h, which means that removing the data on the non-failed injectors has reduced the average service life of the injectors.

Now, it will be checked what the additional value of the data of preventively replaced injectors is. Combining the two subsets again into one complete set yields a censored data set (Sect. 7.2.3), in which the unfailed injectors provide additional information on the reliability of the components. The Kaplan–Meier method is

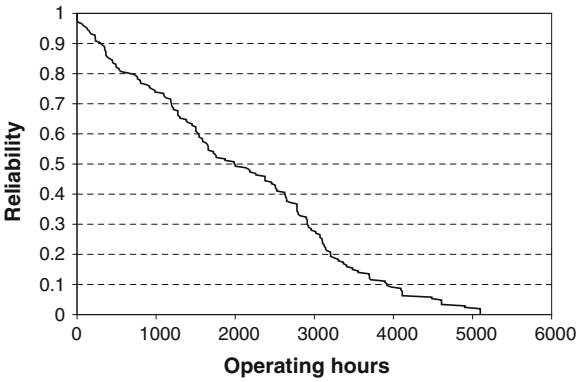

Fig. 7.11 Reliability plot of fuel injectors, based on replacement data of failed injectors

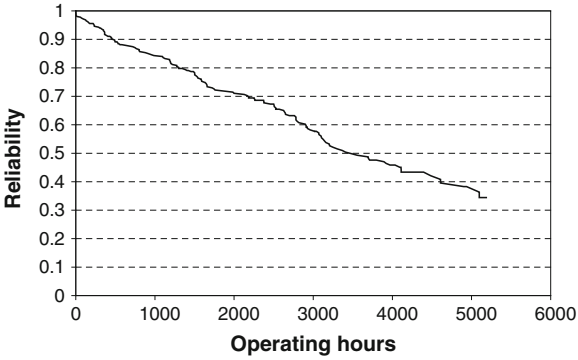

Fig. 7.12 Reliability plot of fuel injectors, based on all replacement data and using Kaplan–Meier approach to include the effect of censoring

applied here to include the censored data in the analysis. The resulting reliability plot is shown in Fig. 7.12.

From this plot, it can be observed that the surviving injectors considerably increase the average reliability of the injectors. The typical service life (at $R = 50$ %) now equals 3,407 h. Only using the actual failure data, as is shown in Fig. 7.11, thus yields a very conservative estimate of the injector service life. Note that in the plot, the reliability does not decrease until 0 %, since the data points with the largest number of operating hours are non-failures. The fact that they survived more than 5,000 h suggests that the reliability indeed is not expected to be close to zero at that age.

7.5.3 As Good as New Assumption

In the previous subsection, the analysis (Fig. 7.11) suggested that the population of fuel injectors consists of two sets with somewhat different behaviour. To investigate this in more detail, a histogram of the failure times is constructed, see Fig. 7.13. This plot shows actually three ranges of operating hours where failure is more likely to occur. As it is suspected that replacing the injectors possibly affects their service life during the subsequent operating period, the assumption that replacing the injector makes it as good as new will be checked.

This is done by separating the failures of the original injectors from the failures of replacing injectors. The reliability curves for both subsets are shown in Fig. 7.14.

The plot shows that a clear difference in reliability exists between the original and replacing injectors. For the original fuel injectors, the typical service life (at $R = 50$ %) equals 2,797 operating hours, while for the replacing injectors, this value is only 1,587 h, almost a factor of two lower.

After obtaining these results, the operator started an investigation to find the cause of the reduced lifetime of replacing injectors. Soon it was discovered that the cleaning of the injector seats during replacement, as prescribed by the OEM,

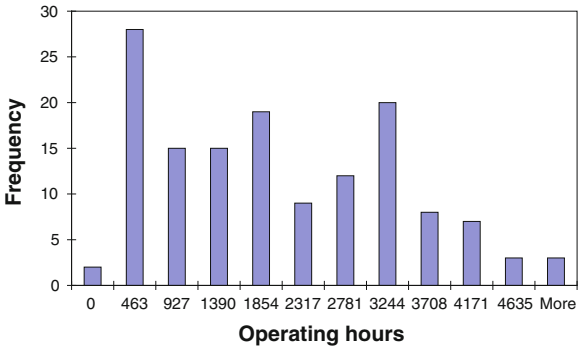

Fig. 7.13 Histogram of fuel injector failure times

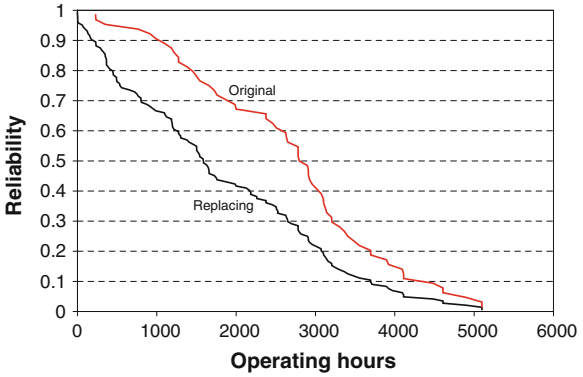

Fig. 7.14 Reliability plot comparing original and replacement fuel injector failure times

caused scratches in the seats. As a result, during operation of the generator, hot combustion gas could leak around the fuel injectors and enhanced the degradation. Therefore, the injectors failed much earlier than expected. By changing the replacement and cleaning procedure, the problem was solved and the service life of the injector could be increased again.

7.6 Stochastic Life Assessment

In the engineering practice, many problems are analysed in a deterministic way. This means that each quantity is assumed to have a single value, and the variation in that value is neglected. For example, the elastic modulus of a set of mechanical test bars is assumed to be exactly equal to the value supplied by the material supplier, and the width of the bar is assumed to match the nominal value exactly. Also, the methods treated in part I of this book to calculate loads and ttf are completely deterministic.

However, in practice, there will always be a slight variation in any quantity across a set of samples. For material properties, like the elastic modulus or the

coefficient of thermal expansion, this is caused by slight variations in the chemical composition or the microstructure of the material. The dimensions of parts may vary due to slightly different manufacturing processes, where the magnitude of the variations depends on the (required) accuracy of the applied manufacturing technique. The variations in material properties or dimensions also lead to variations in the properties that depend on these basic characteristics. For example, the mentioned variation in elastic modulus in the test bars leads to a variation in the measured elongations for a given load. Moreover, in Chap. 6, these variations in dimensions and material properties were indicated to constitute one of the sources for the uncertainty in a service life prediction.

In numerical analyses, the effect of this variation is normally covered by safety factors. By using a lower limit of the material properties, a conservative result is obtained from the analysis. The lower limits are identified by their 95 % confidence level on the first (also called *A*-value) and tenth percentiles (also called *B*-value). Similarly, in the determination of maintenance intervals, the uncertainty in the expected remaining useful life of components was shown to be covered by applying conservative (shorter) intervals. However, the amount of conservatism in the final result in this case cannot be quantified. For that reason, probabilistic or stochastic analyses can be performed [8], in which the variation in the input parameters is incorporated. This provides two important benefits. Firstly, based on the known variation in the input quantities, the uncertainty in the output (e.g. the reliability) can be quantified. And secondly, it provides insight into the causes of the scatter in the output (sensitivity analysis).

7.6.1 Stochastic Analysis

Scatter or variation in one of the model parameters is incorporated by treating the parameter as a stochastic or random variable. Such a variable is described mathematically by a distribution function, as was discussed in Sect. 7.2.1.

Performing a stochastic analysis provides benefits, as mentioned in the introduction, but it also has two drawbacks. Firstly, the analyses are computationally more demanding, and secondly, the required data are not always available. The first issue is not very critical with modern computers, but the second issue is often more difficult to address. When insufficient data are available, the distributions of input parameters cannot be determined accurately. However, in some cases, engineering judgement and past experience can be used to find an acceptable approximation of the distribution functions.

Two basic types of stochastic analyses exist [8]. The first type is a sensitivity analysis, which is used to determine how sensitive the design or service life is for variations in the model parameters (e.g. dimensions, material properties). The results of this sensitivity analysis can be applied to determine whether a certain parameter should be treated as stochastic variable or can be assigned a constant deterministic value. The second type of analysis is the probabilistic life assessment

(or reliability analysis), which is used to determine the service life distribution or the reliability in terms of the probability of failure for a certain load sequence.

In general, a complete stochastic analysis consists of the following steps:

1. Creation of an accurate deterministic model, for example, a model to predict the creep life for a component subjected to a certain load sequence (stress and temperature history).
2. Selection of the random variables and their distribution functions. The results of a sensitivity analysis can assist in making this selection, whereas a sufficient amount of data is required to determine the appropriate distribution functions.
3. Definition of the failure function, specifying when failure occurs.
4. Solution of the stochastic problem by application of a suitable stochastic (sampling) method. The different methods will be discussed in the next subsection.
5. Interpretation of the results, especially the determination of the allowable probability of failure.

It is very important that a reliable deterministic model is created, since that model serves as a basis for the stochastic analysis. Any inaccuracy in the model of the failure mechanism will lead to much larger inaccuracies in the results of the stochastic analysis.

7.6.2 Failure Function

One of the essential steps in performing a stochastic analysis is the definition of a failure function (G). This means that a function should be defined that attains a value smaller than zero for all combinations of the random variables that lead to failure. A value larger than zero indicates a safe condition, and $G = 0$ is called the limit state.

For example, an analysis can be performed on a cantilevered beam, for which the deflection δ depends on the magnitude of both the applied force F (= load) and the elastic modulus E (= capacity). To investigate the variation in the expected deflection, F and E are considered to be stochastic variables, and the resulting distribution of the deflection δ (F, E) will be a function of these two variables. Further, the maximum allowable deflection of the beam is 4 mm, which means that deflections exceeding this value are considered to be a failure. For this example, the failure function G can then be defined as

$$G(F, E) = 4 - \delta(F, E) \tag{7.32}$$

For deflections less than 4 mm, the failure function will have a positive value, while it becomes negative for deflections larger than 4 mm.

Using this kind of failure function, a probabilistic problem can always be formulated as

$$p_f = P(G(\vec{x}) \leq 0) = \iint\limits_{G(\vec{x}) \leq 0} f(\vec{x}) \mathrm{d}\vec{x} \tag{7.33}$$

where $\vec{x}$ is the vector containing the random variables and $f(\vec{x})$ is the joint probability density function (JPDF) of these variables. Solving the stochastic problem therefore requires solving the integral equation.

Since the JPDF is generally unknown and also the limit state $G = 0$ is normally not available in explicit form, the only way to solve the equation is to perform multiple evaluations of the failure function. In its simplest form, this means that a sampling method is used to randomly take values for the stochastic variables from their distributions. For each combination of random variables, the failure function is evaluated and the relative occurrence of failed states determines the probability of failure. A number of available sampling methods will be discussed in the next subsection.

7.6.3 Sampling Methods

The most extensively applied sampling method is the Monte Carlo method. This method randomly selects values for all stochastic variables, taking into account the specified distribution functions, and evaluates the failure function G. If a negative value is obtained, failure has occurred. An approximation of the integral equation in (7.33), representing the probability of failure of the system, is then obtained by

$$p_f = P(G(\underline{x}) \leq 0) = \frac{\text{number of simulations with } G \leq 0}{\text{total number of simulations } N} \tag{7.34}$$

This method is very robust and always converges to the exact solution. However, a large number of failure function evaluations are required to obtain accurate results, which makes the method computationally demanding. This is especially the case when a G-evaluation requires a complex analysis (e.g. finite element analysis) instead of the evaluation of a simple explicit equation.

In engineering applications, the allowed probabilities of failure are often (very) small ($<10^{-3}$), which means that a failure state ($G < 0$) is only found occasionally during the sampling process. The number of simulations required to obtain an accurate solution depends on the probability of failure in the following way:

$$N \approx 270 \frac{1 - p_f}{p_f} \tag{7.35}$$

For $p_f = 10^{-3}$, the required number of simulations is already 270,000. However, the required number of simulations is not dependent on the number of stochastic variables, as is the case for other methods. Therefore, the Monte Carlo method is relatively efficient for problems with many random variables.

Whereas in a Monte Carlo simulation the values of the stochastic variables are taken completely at random, the efficiency of the method can be improved by a more dedicated selection of the variables. A method that applies this principle is the Latin hypercube sampling (LHS) method, where for an analysis with N simulations, the probability distribution is split into N intervals with equal probability. This means that in the tails of the distribution, the intervals cover relatively large ranges of the random variable, while in the regions with a high probability density, the intervals are much narrower. During the simulation, only one sample is taken from each interval, which guarantees a uniform sampling of the distribution. Compared with the Monte Carlo method, convergence to the exact answer is faster for LHS. The method is also very robust, but is still very inefficient for small probabilities of failure, since still a large number of simulations are required then.

Importance sampling

To increase the efficiency for problems with low probabilities of failure, importance sampling methods have been developed. These methods artificially raise the probability of sampling from the ranges within the input distributions that cause the extreme values of interest (i.e. failures) in the output. The use of this biased distribution will lead to a biased result of the simulation. However, the simulation outputs are weighted to correct for the use of the biased distribution, which ensures that the final result is unbiased. The challenge of these methods is to find an appropriate biased distribution, or, in other words, determine the restricted sampling domain.

A specific type of importance sampling is the adaptive radial-based importance sampling (ARBIS) method. In this method, a circular area with radius β_0 is defined in the plane of the random variables (or space for >2 variables) inside which no samples are taken. The radius β_0 is unknown, but must be selected as large as possible, without intersecting the limit state(s). This is done in an adaptive way. Since only samples are taken outside the circular area defined by the radius β_0, which contains most of the samples in a Monte Carlo simulation, many samples are in the important region near the limit state. This makes the method much more efficient than the Monte Carlo method.

Response surface methods

In a limited number of cases, an analytical solution for the integral equation in (7.33) exists. One of these cases is the situation where (1) all stochastic variables follow a standard normal distribution; (2) all stochastic variables are statistically independent and (3) the limit state ($G = 0$) is a hyperplane defined by the stochastic variables. In two dimensions, this is a straight line and in 3 dimensions a plane, etc. If these three conditions are satisfied, the following analytical solution can be used:

$$P(G(\vec{x}) \leq 0) = \iint\limits_{G(\vec{x}) \leq 0} f(\vec{x}) \mathrm{d}\vec{x} = \Phi(-\beta) \tag{7.36}$$

where Φ is the standard normal cumulative distribution function and β the smallest distance from the origin (in standard normal space) to the limit state. This represents the point with the highest probability, which is therefore called the most probable point (MPP). Solving the integral equation has now reduced to an optimization problem: finding the smallest distance β.

In practice, the three requirements are never fully met. The first two requirements can be satisfied by applying variable transformations, but the required straight line for the limit state is often not realistic. In the first-order reliability method (FORM), the limit state is approximated by a straight line through the MPP. The inaccuracy of the method is then determined by the deviation of the real limit state from the straight line. The optimization process used to find the MPP requires a certain number of failure function evaluations (the actual number depending on the number of random variables), but in general, the method is considerably more efficient than the Monte Carlo methods. However, the required variable transformations and variable independency and the definition of an efficient optimization scheme make this method rather complex.

A further improvement in the accuracy can be achieved by applying a second-order reliability method (SORM), which approximates the limit state by a second-order Taylor polynomial. To make this approximation, somewhat more failure function evaluations are required, but the accuracy of the method is much better. For these response surface methods, the probability of failure does not affect the required number of simulations, as is the case for the Monte Carlo methods. That makes these methods suitable for engineering applications. On the other hand, the number of random variables does affect the required number of simulations, which means that for problems with very large numbers of stochastic variables, the efficiency benefit compared with the Monte Carlo method will be lost. Another disadvantage is that the method easily fails (not robust) for more complicated problems and is not capable to handle more complex shapes of the limit state.

Comparison of methods

The sampling methods mentioned in this section have been described in more detail by Grooteman [8], who also compared the performance of the methods using a benchmark problem. The results of the comparison are shown in Table 7.2, indicating the number of simulations required for the convergence of the results and the calculated probability of failure. Deviation from the actual value of $7.1{\cdot}10^{-5}$ indicates the inaccuracy of the methods.

Table 7.2 Comparison of different sampling methods

Method	# Simulations	Calculated p_f
Monte Carlo/LHS	790.000	7.1×10^{-5}
Importance sampling (ARBIS)	475	7.4×10^{-5}
FORM	25	4.9×10^{-5}
SORM	35	7.0×10^{-5}

This shows that the more sophisticated methods are much more efficient than the Monte Carlo method, with acceptable accuracies. The only exception is the FORM method, which is efficient, but less accurate.

7.6.4 Application

The methods discussed above can be applied to life assessments in two slightly different ways. The most frequently applied method is the calculation of the probability of failure, as was discussed before in this section. This means that a failure state is defined a priori (e.g. allowable deflection of the beam, allowable amount of creep deformation or fatigue crack length) and that the risk of failure is calculated using the assumed variations in material properties or loads. These risk assessments are frequently performed during the design of aircraft, in the nuclear power generation sector and in civil engineering (bridges, dams), where failures have serious consequences. The acceptable risk level is defined, and all designs will have to meet this requirement (see also Sect. 9.5 on probabilistic design).

However, stochastic analyses can also be used to create the complete distribution of lifetimes for a certain failure mechanism. This is often done for crack propagation analyses [8], where the material properties characterizing the crack growth behaviour (e.g. the parameters C and m in the Paris law, see Sect. 4.4.6) belong to a certain distribution. Simulations are performed for a large number of samples, which yields a large collection of crack growth curves. Based on the calculated number of cycles to failure (corresponding to a certain critical crack length), a distribution function for the service life can be constructed. Of course, from this distribution, the probability of failure corresponding to a certain number of cycles can be derived, so the two methods are strongly related. However, the latter method provides much more insight into the sensitivity of the resulting distribution for variations in the input distributions.

7.7 Summary

In this chapter, the potential applications of the knowledge on loads and failure mechanism in reliability engineering have been discussed. After a brief introduction of the basic principles of reliability engineering, the challenge of selecting a RFP has been treated. Understanding the failure mechanism and recognizing the governing load were shown to be the key factors in selecting the most appropriate failure parameter and minimizing the uncertainty in the analysis. This has also been demonstrated in a case study on the failure of diesel generator fuel injectors. The application of the life exchange rate matrix in reliability engineering methods has been discussed, and the limitations of this approach have been indicated. Finally, the concept of stochastic service life analyses is described, and it has been

shown that how these analyses aid in quantifying the uncertainty in the service life of a component.

References

1. Manzini, R., Regattieri, A., Pham, H., Ferrari, E.: Maintenance for Industrial Systems. Springer Series in Reliability EngineeringSpringer, London (2010)
2. Kumar, U.D.: Reliability, Maintenance and Logistic Support; A Life Cycle Approach. Kluwer Academic Publishers, Norwell (2000)
3. Ebeling, C.E.: An Introduction to Reliability and Maintainability Engineering. McGraw-Hill Companies, Inc., Boston (1997)
4. Pham, H.: Handbook of Reliability Engineering. Springer, London (2003)
5. Tinga, T.: Application of physical failure models to enable usage and load based maintenance. Rel. Eng. Syst. Saf. **95**(10), 1061–1075 (2010)
6. Smit, K.: Onderhoudskunde. VSSD, Delft (2010)
7. Robinson, E.L.: Effect of temperature variation on the long-time rupture strength of steels. Trans. ASME **74**(5), 777–781 (1952)
8. Grooteman, F.: A stochastic approach to determine lifetimes and inspection schemes for aircraft components. Int. J. Fat. **30**, 138–149 (2008)

Further Reading

1. Manzini, R., Regattieri, A., Pham, H., Ferrari, E.: Maintenance for Industrial Systems. Springer Series in Reliability EngineeringSpringer, London (2010)
2. Pham, H.: Handbook of Reliability Engineering. Springer, London (2003)
3. Ebeling, C.E.: An introduction to reliability and maintainability engineering. McGraw-Hill Companies, Inc., Boston (1997)

Chapter 8
Failure Analysis

8.1 Introduction

Despite all maintenance activities performed in industry, unexpected failures will always keep occurring in practice. However, if a failure has serious consequences in terms of costs, safety, environmental effects or consequential damage, there will generally be a large drive to prevent such a failure to occur again in future. Also, less critical failures may become very awkward if they occur regularly. In those cases, it is essential to find the root cause of the failure, since knowing that cause enables to find a solution for the problem, either by reducing the loads on the system or by increasing the load-carrying capacity.

In the present chapter, firstly several existing methods to analyse failures, their effects and their causes will be discussed. Both methods to assess possible future failures and methods analysing failures that already occurred will be treated. These methods will guide the failure analysis and ensure that a structured approach is followed. Then, in Sect. 8.3, a procedure is proposed for a mechanism-based failure analysis, which follows the introduced structured approach, but at the same time optimally utilizes the knowledge on loads and failure mechanisms introduced in the first part of this book. In Sect. 8.4, the method is applied in a case study from industry. After this discussion of analysis methods, the remainder of the chapter is devoted to the determination of the failure mechanism. Once the precise failure mechanisms are known, finding a way to prevent such a failure is generally rather straightforward. However, performing a solid failure analysis generally requires quite some knowledge and experience. Therefore, the final two subsections aim to provide background information and procedures that assist less-experienced people in assessing the failure mechanism of a failure at hand.

T. Tinga, *Principles of Loads and Failure Mechanisms*,
Springer Series in Reliability Engineering,
DOI: 10.1007/978-1-4471-4917-0_8, © Springer-Verlag London 2013

8.2 Methods

Several methods are available that assist in performing a structured analysis of failures. The major goal of all these methods is the prevention of failures, especially those failures that have large consequences. The methods can be divided in two classes. The first class of methods is applied in the design phase of the system, before it has entered service and before any failure has occurred. These methods, like the failure mode, effects and criticality analysis (FMECA) and the fault tree analysis (FTA), aim to identify possible future failure modes. If the risks associated with some failure modes are perceived to be too high, a modification of the design can be considered or appropriate maintenance tasks can be defined (e.g. periodic inspections).

The second class of methods is applied after a failure has occurred. These methods focus on finding a way to prevent additional failures to occur, either by looking for the root cause of the failure (e.g. root cause analysis—RCA) or by selecting the failures with the highest priority (e.g. Pareto and degrader analysis). The four mentioned analysis methods will be discussed in the next subsections.

8.2.1 Failure Mode, Effects and Criticality Analysis

In the failure mode and effects analysis (FMEA), all possible failures for a certain system are identified, but also the effects of these failures are described in terms of financial, safety and functional consequences. A FMEA is an inductive or bottom-up method, since the analysis starts with the possible failures of components and derives what the consequences (on the higher system level) are. A FMEA is generally executed by a group of people with different backgrounds. By including experience from design, operation, maintenance and finance in the team, it is more probable that all possible failures are identified and that their effects are properly estimated.

Whereas the FMEA is a purely qualitative analysis, only describing what the possible failure modes and their effects are, the analysis can be made more quantitative by adding a criticality analysis. The method is then called FMECA. For each failure mode i, the criticality is quantified by calculating the risk priority number (RPN) defined as

$$\mathrm{RPN}_i = S_i O_i D_i \tag{8.1}$$

The RPN is a product of severity (S), occurrence (O) and detection (D). The severity of a failure mode quantifies how large the consequences of that failure mode are. Values are typically obtained from predefined tables, indicating on a scale from 1 to 10 or from 1 to 5, the different grades of severity. Further, the occurrence parameter quantifies its likelihood of occurrence, for example, ranging from extremely unlikely to frequent, and the detection parameter specifies the

probability that a failure is *not* detected when it occurs. Also, the values of these two parameters are selected from predefined tables. By multiplying the three quantities, the RPN properly expresses that a failure mode is associated with a higher risk when it occurs more often (O), and its consequences are more severe (S), or when the probability that the failure is not detected (D) is higher.

Although the RPN is obtained from an objective multiplication of the parameters S, O and D, the definition of the tables for these parameters and the selection of the values is still rather subjective. Therefore, the scoring of the failure modes should preferably be performed independently by several people from the FMEA team to obtain a more objective result. Moreover, the obtained RPNs are not risk numbers in an absolute sense, since they depend on the chosen tables. This means that the boundary between acceptable and unacceptable risks (i.e. the RPN threshold value) should be determined for each analysis separately. Finally, the three quantities S, O and D quantify rather different aspects of risk. It should be realized that an increase in occurrence (O) not always represents the same increase in risk as an equally large increase in severity (S). Therefore, the obtained RPN values should be interpreted with care.

The results of a FMECA analysis are generally collected in a large table, which is called the FMECA form. An example of such a form will be shown below for a simple case (Tables 8.1 and 8.3). By completing all columns of the form, the analysis is performed in a structured and complete manner. Note further that the FMECA is closely related to the reliability-centred maintenance (RCM) strategy [1] (see Sect. 5.3.1). In fact, the first five steps in this approach constitute the FMECA. Moreover, ideally the FMECA is a dynamic process, which means that failure data obtained during operation should be utilized to

Table 8.1 First part of FMECA form for bicycle case study

Function	Failure mode	Effect	Causes	Detection
Rolling	Flat tire	Driving impossible	Nails, wear	Trivial, inspection
	Damaged wheel bearing	High driving resistance	Wear, bad lubrication	Inspection
Steering	Stem seizure	Driving impossible	Corrosion	Trivial
	Broken handlebar	Driving impossible	Overload, fatigue	Trivial
Carrying	Broken frame	Driving impossible	Overload, corrosion	Trivial
	Broken seat post	Driving almost impossible	Overload	Trivial
Driving	Broken chain	Driving impossible	Overload, corrosion, wear	Trivial
	Damaged bearing	High driving resistance	Wear, bad lubrication	Inspection
Illumination	Failed front light	No driving in dark	Collision, degradation	Testing
	Electrical wires disconnected	No driving in dark	Overload	Inspection

Fig. 8.1 Bicycle used as case study in failure mode, effects and criticality analysis

update the FMECA. However, in practice, this is not very common. The analysis is generally performed before the system enters service and is not updated thereafter [2].

Several standards are available for FMEA, where the MIL-STD-1629A, British Standard BS 5760 and the J1739 from the Society of Automotive Engineers are the most important ones. The general procedure to perform a FMEA is as follows [3]:

1. FMEA group formation.
2. System analysis.
3. FMEA execution.
4. Risk evaluation (FMECA).
5. Corrective action planning.

The steps 2–5 of this procedure will be demonstrated below for a rather simple system, a bicycle (see Fig. 8.1).

8.2.1.1 System Analysis

In this first step of the actual analysis, the product must be analysed to define the system structure. Either a functional breakdown is performed or the system hierarchy in terms of subsystems and components is identified. This structured breakdown is used as a reference for the further analysis and defines the first column of the FMECA form, as is shown in Table 8.1. Also, the system boundary must be set, to decide which parts of the system are considered in the analysis.

For the present bicycle case study, a functional breakdown is performed, identifying the function of the bike (personal transport) and the different functions of its subsystems: driving, illumination, carrying and rolling. The results are shown in Table 8.1.

8.2.1.2 Failure Mode and Effects Analysis

In this second step, the actual FMEA is performed, so for every function identified in the previous step, the possible failure modes and their effects are defined. Also, the possible cause is indicated in the table, but this is not an essential part of the FMEA. Determination of the cause of failures can be done by other methods, like root cause analysis that will be discussed below. Finally, also the detection method for each failure is indicated.

For the bicycle case study, the results are shown in the FMECA form in Table 8.1. For each function, several failure modes are listed and their effects are specified. Many of the failure modes are easy to detect, that is, breaking of the chain will immediately be observed, but for some failure modes (e.g. bearing degradation), an inspection may be required to asses the damage.

8.2.1.3 Risk Evaluation

In the third step of the analysis, the risk of the different failure modes is quantified by performing the criticality analysis. This step transforms the analysis from a purely qualitative into a more quantitative analysis. For the three parameters, occurrence, severity and detection rating scales must be selected from standards or must be defined specifically for the analysed system. Then, for each failure mode, values are assigned to the parameters and multiplication yields the RPN value.

For the bicycle case, the rating scales used are shown in Table 8.2. Then, values are assigned to O, S and D for each failure mode, and the RPN values are calculated, as is shown in Table 8.3.

8.2.1.4 Interpretation of Results: Corrective Action Planning

The final step in the analysis is then the interpretation of the results. The failure modes with the highest RPN represent the highest risks, and corrective actions should therefore be planned for these failure modes. Reducing the risk can be achieved by reducing any of the three parameters. Application of higher-quality parts possibly reduces the failure frequency and thus decreases the value of O. The consequences of a failure can be diminished by changing the structure of the

Table 8.2 Rating scales for O, S and D for the bicycle case study

Rate	Occurrence	Rate	Severity	Rate	Detection
1	Once per lifetime	1	Slight disfunctioning	1	Trivial
2	Once per year	2	Driving in certain conditions not possible	2	
3	Once per quarter	3	Driving becomes uncomfortable	3	Inspection
4	Once per week	4	Severe effect on driving	4	
5	Once per day	5	Driving impossible	5	Difficult

Table 8.3 Second part of FMECA form for bicycle case study

Function	Failure mode	*S*	*O*	*D*	*RPN*	Action
Rolling	Flat tire	5	4	1	20	Apply better tires
	Damaged wheel bearing	3	2	3	18	Preventive maintenance (lubrication)
Steering	Stem seizure	5	2	1	10	Inside parking
	Broken handlebar	5	1	1	5	Reduce loading
Carrying	Broken frame	5	1	1	5	Reduce loading
	Broken seat post	4	1	1	4	Reduce loading
Driving	Broken chain	5	3	1	15	Preventive maintenance (lubrication)
	Damaged bearing	3	2	3	18	Preventive maintenance (lubrication)
Illumination	Failed front light	2	3	2	12	Apply higher-quality lights
	Electrical wires disconnected	2	4	2	16	Better protection of wires

system, for example by incorporating redundancy. Failure of a single subsystem or part will then not automatically result in failure of the complete system, and the severity (*S*) will be reduced. Finally, setting up an inspection program might increase the detectability of certain hidden or hard-to-detect failures, which decreases the value of *D*.

For the bicycle case study, the results in Table 8.3 demonstrate that the RPN values for the different failure modes range from 4 to 20. This means that a clear priority to solve the failure modes emerges, suggesting to start with high RPNs, for example those failures with RPN > 16. It is also observed that the highest RPNs are associated with failures that either occur frequently (flat tire, disconnected electrical wires) or that are hard to detect (bearing damage). The first type of failures can be solved by applying higher-quality parts to reduce the failure frequency, while the second type of failures benefits from periodic inspections and preventive maintenance.

8.2.2 Fault Tree Analysis

Another method to asses the possible failure modes and mechanisms of a system is the fault tree analysis (FTA). Contrary to the FMECA, the fault tree analysis is a deductive or top-down method. Starting from a system failure, which is called the top event, all possible underlying events and failures of subsystems or components are identified. In the end, a series of basic events are obtained that may be responsible for the occurrence of the top event. The analysis is presented in the graphical form, where a 'tree' of events is constructed. An example of a fault tree with top event *T* and basic events *A, B* and *C* is shown in Fig. 8.2.

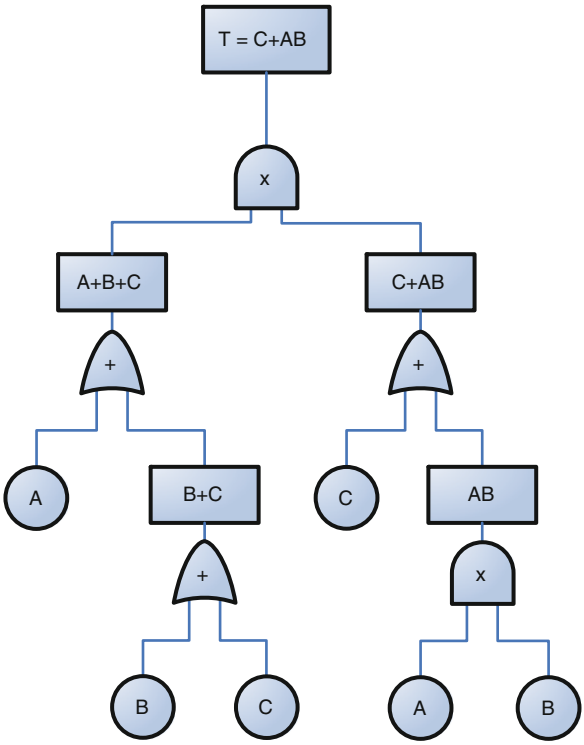

Fig. 8.2 Example of a fault tree with top event T and basic events A, B and C

The events causing a higher-level event are connected through different types of gates. An OR gate (+) indicates that the higher-level fault occurs when at least one of the input faults occurs. For an AND gate ($\times$), all of the input faults must occur before the higher-level fault occurs. In this way, the dependence of the system on its subsystems and components can be analysed accurately.

The fault tree can be translated into expressions using Boolean algebra. For example, the lower right-hand part of the fault tree in Fig. 8.2 shows an AND gate connecting the basic events A and B. The higher-level fault can then be expressed as AB. Both A and B can only attain the value 0 (false, i.e. no failure) or 1 (true, failure occurs), so only if both faults occur simultaneously, AB becomes true and failure occurs. Similarly, the OR gate one level higher connects AB and C, which yields an expression for the higher-level fault $C + AB$. This means that this event will occur when either C or AB attains the value 1. Using this procedure, the total fault tree can be analysed, which eventually yields the top-event expression $T = (A + B + C)(C + AB)$.

The obtained Boolean expression representing the fault tree can be transformed into an equivalent fault tree (EFT), which is a two-level fault tree that represents the essential behaviour of the complete fault tree. The constituents of the EFT are so-called minimal cut sets (MCS), which are combinations of basic events that are essential for the top event to occur: if a single failure in the cut set does not occur,

the top event will not occur. The benefit of having an equivalent fault tree for a system is that the dependence of the system on its critical faults is directly visible. The general form of an EFT for a top event T is as follows:

$$T = \sum_{i=1}^{n} \mathrm{MCS}_i = \sum_{i=1}^{n} \left(\prod_{j=1}^{m_i} C_{ij} \right) \tag{8.2}$$

where n minimal cut sets MCS_i exist, each consisting of m_i primary events C_{ij}.

To derive the EFT from a general fault tree expression, the rules of Boolean algebra must be applied (see e.g. [3] for an overview). The expression obtained for the fault tree in Fig. 8.2 can be reduced as follows:

$$\begin{aligned} T &= (A + B + C)(AB + C) \\ &= AAB + ABB + ABC + CA + CB + CC \end{aligned} \tag{8.3}$$

Applying the idempotent law, stating that $X_A \cdot X_A = X_A$ and $X_A + X_A = X_A$, this expression can be reduced further to

$$\begin{aligned} T &= AB + AB + ABC + CA + CB + C \\ &= AB + ABC + CA + CB + C \end{aligned} \tag{8.4}$$

Finally, using the law of absorption, $X_A + (X_A X_B) = X_A$, the expression for the EFT is obtained to be

$$T = AB + C \tag{8.5}$$

This simple expression containing two MCSs, AB and C, immediately shows that the system only fails when either component A and B both fail or C fails. This could not be observed easily from the original expression.

Finally, once the minimal cut sets for the system have been determined, the probability of failure for a system can easily be calculated from the fault tree, provided that the failure probabilities of the basic faults or events are known. The basic probability rules can be applied to the Boolean expressions:

$$P(A + B) = P(A) + P(B) - P(A)P(B) \tag{8.6}$$

$$P(AB) = P(A)P(B) \tag{8.7}$$

where $P(event)$ represents the probability that *event* occurs. In more general terms, this is expressed as

$$\begin{aligned} P(X_1 + X_2 + \ldots + X_i) &= \coprod_i P(X_i) \\ &= 1 - (1 - P(X_1)) \cdots (1 - P(X_i)) \end{aligned} \tag{8.8}$$

$$P(X_1 X_2 \ldots X_i) = \prod_i P(X_i) = P(X_1)P(X_2)\ldots P(X_i) \tag{8.9}$$

Applying these rules to the fault tree in Fig. 8.2 and given that over a certain operating period the failure probabilities for the components are $P(A) = 0.01$, $P(B) = 0.1$ and $P(C) = 0.05$, the probability that the system fails is $P(T) = 0.051$.

8.2.3 Pareto and Degrader Analysis

The two methods discussed in the previous subsections are generally applied before the system enters service and thus before any failure has occurred. They are used to get insight into the possible risks and aim to govern the system design process. For systems that are already operative, the data collected on failures during service provide very useful additional information that can be utilized to further improve the system or its operation.

The Pareto analysis can be used to prioritize such improvement efforts for complex systems. In general, complex systems show many different failures, but not all failures are equally harming the operation of the system. The Pareto analysis provides a structured methodology to filter out the most important failures. It is based on the observation that 20 % of the failures are responsible for 80 % of the maintenance costs, or 80 % of the total downtime of the system. This top 20 % of the failures should therefore be aimed at in improving the system, since solving them provides a major reduction in costs or a significant improvement in the uptime. At the same time, putting effort in one of the failures outside the top 20 % will yield very limited benefits.

The tool generally used to perform a Pareto analysis is the Pareto chart, in which the data, for example failures of a system, are divided into a number of classes and then plotted in a bar diagram. However, before plotting the data, the classes are sorted such that the first bar represents the largest number of failures. In addition to the bars, also a line plot of the relative cumulative count of failures is often presented. An example for gas turbine failures in different modules is shown in Fig. 8.3.

The Pareto chart directly visualizes the class containing the largest number of failures, indicating that improving the performance of this class will provide the largest improvement in the complete system.

Instead of plotting the number of failures, in maintenance applications, often the costs of maintenance (repair, replacement, labour) or the effect on system availability is plotted in a Pareto chart. This yields an overview of the top cost drivers and performance killers, which generally drive the system or maintenance process improvement.

Identifying the cost drivers and performance killers from a Pareto analysis is also the basis of the platform degrader analysis proposed by Banks et al. [4], which is partly based on the reliability-centred maintenance [1] approach (see also Sect. 5.3.1). The degrader analysis aims to determine which components and subsystems contribute the most to the loss of system operational availability. It then identifies

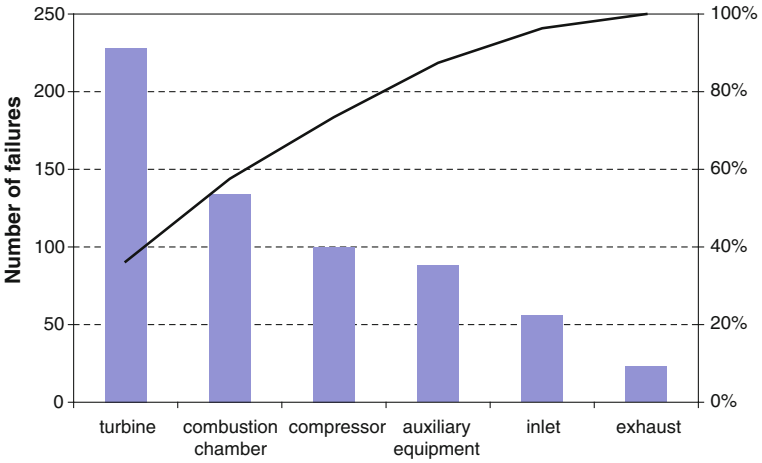

Fig. 8.3 Pareto chart of gas turbine failures per module

diagnostic, predictive and prognostic technologies that are mature and appropriate to apply to these specific components and subsystems. As the method focuses on the top candidates for health monitoring, rather than conducting full FMECA on each platform, an in-depth focus on relevant components is achieved, rather then providing superficial recommendations for many components.

The platform degrader analysis that already has been applied in a case study in Sect. 6.4.3, which consists of three steps:

1. Identify components which have the lowest reliability and greatest number of maintainability issues (i.e. Pareto analysis).
2. Evaluate how these components fail and determine their dominant and critical failure modes (applying FMECA on only the top degrader components).
3. Identify appropriate solutions for monitoring each dominant and critical failure mode, capable of providing a diagnostic or prognostic assessment.

Finally, it should be noted that the data collected by companies in their computerized maintenance management systems (CMMS) is not always suitable to be applied in more quantitative analyses. Firstly, the accuracy and completeness of the collected (failure) data are in most cases not very high level. But more importantly, the background of the data is often unclear. In addition to regular failures, also failures due to human errors or wrong procedures are included in the data, without the possibility to separate these types of failures. Also, information on the level of the failure mechanisms is generally not available. However, despite these limitations, analysing the data in many cases yields very useful insights into the failure behaviour of systems.

8.2.4 Root Cause Analysis

The final method discussed in this section is the RCA, which is applied to thoroughly research the cause of a failure or accident. Formally, RCA is not a well-defined method, but actually covers all structured approaches aiming to find the deepest cause of a failure, accident or event. One of the first publications on RCA is the famous book by Keppner and Tregoe [5]. The essence of the RCA is that only addressing the *root* cause of a problem, as opposed to merely address the symptoms of the problem, will ensure that it will not happen again. This generally requires 'out of the box' thinking.

The method is often applied in investigations of major accidents (e.g. aircraft crashes, nuclear power plant accidents), where the main purpose is to discover the chain of events leading to the accident and learn the lessons from that insight to prevent similar accidents to occur again. However, also liability often plays an important role, since knowing the root cause often enables to put the (financial) responsibility at some party.

During the course of a root cause analysis, some of the previously discussed methods may be applied. For example, a fault or event tree analysis generally is a suitable methodology to structure the problem. Moreover, the analysis can be performed both in a deductive (top-down) and in an inductive (bottom-up) manner. In a RCA, different types of causes can be found. In general, the following three classes of causes are recognized [6]: technical or physical causes, human errors and latent causes. The latter class is associated with errors in legislation or regulations and with common practices in a certain companies that may lead to failures.

The essence of a RCA is that it is executed until a sufficiently deep level is reached. Otherwise, the real root cause is not discovered, but only intermediate failures (e.g. of subsystems) are obtained, and only symptoms or consequences of the problem will be acted on. To ensure that a sufficiently deep level is reached, a technique called 'five whys' is sometimes applied. It is based on the heuristic that the actual root cause is found only at the fifth level, so asking 'why did it fail?' for five consecutive times will probably yield the root cause.

In a maintenance context, the consequences of many failures are not very severe, which makes that a RCA is not performed very often. If a component fails, it will be merely replaced without investigating what the cause of the failure has been. However, if the root cause of the problem has not been addressed, it is quite probable that the failure will occur again rather soon. Therefore, applying RCA to maintenance problems can yield significant benefits in terms of cost or downtime reduction.

But then again, the RCA can only be successful if it is executed to a sufficiently deep level. For maintenance problems, this means that not only the failing part and failure mode must be determined, as is generally done, but also the failure mechanism and the associated loads should be identified. If this information is available, it will be rather simple to decide whether the loading of the part was too high, or the load-carrying capacity was too low. The benefit of knowing whether

the loading or the capacity of the system is the root cause is that finding a solution is rather straightforward. If the loads appear to be too high, the usage of the system (which causes the loads) must be altered. On the other hand, when the capacity appears to be insufficient, a redesign must be considered, where other materials could be applied or the dimensions of the part could be changed.

The balance between loads and capacity has been introduced in Sect. 1.2 and is the basis of the complete first part of this book. The basic principles of the treated loads and failure mechanisms can therefore assist in determining the root cause of many failures. Another benefit of executing the RCA until the level of the failure mechanism is that the number of possible causes becomes quite limited. Whereas infinite numbers of failure modes exist, the number of failure mechanisms on the material level is only in the range of 15 to 20 (see part I). On that level, it does not matter whether a fatigue failure occurs in a helicopter part, a structural member of a bridge or a production machine component. The only challenge is to determine the internal load on the material level from the specific dimensions, materials and usage of the system under consideration.

8.3 Mechanism-Based Failure Analysis

As was indicated in the previous section, executing a sound failure analysis in a maintenance context requires that the following conditions are met:

1. a concise and structured approach is followed,
2. a proper selection is made of relevant failures to be investigated,
3. the analysis is executed to a sufficiently deep level.

In this section, a procedure is proposed that meets these requirements and at the same time optimally utilizes the knowledge on loads and failure mechanisms introduced in the first part of this book.

The procedure, which is called the mechanism-based failure analysis (MBFA), combines the four methods introduced in the previous section. The generic process of MBFA and the role of FTA, RCA, FMECA and Pareto are schematically shown in Fig. 8.4. A step-wise guideline for performing a failure analysis is shown in Fig. 8.5. Starting from a failed asset or system to be analysed, firstly a fault tree analysis is performed to identify all possible failure modes that could lead to the system (functional) failure.

After the completion of this overview, it is required to identify the most critical failure modes, since generally solving all possible failure modes is not feasible. Note that also completing a full FTA might not always be feasible. Especially for large and complex systems, the fault tree will be quite extensive. In those cases, steps 2 and 3 from Fig. 8.5 can be executed simultaneously, focusing the FTA on the most critical failures.

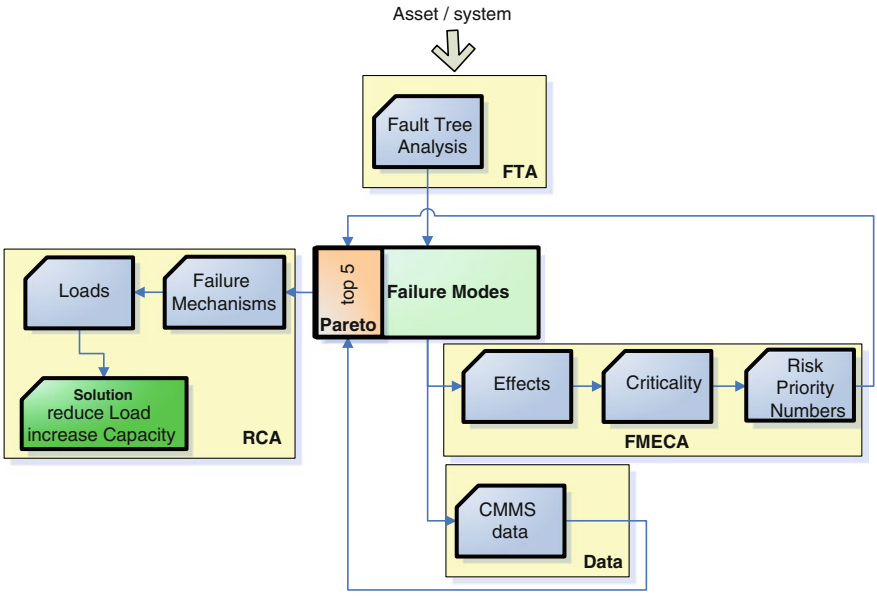

Fig. 8.4 Failure analysis process diagram showing the role of *FTA*, *FMECA*, *Pareto* and *RCA* in identifying critical failure modes and their root causes

To be able to perform the Pareto analysis, which determines the top 5 or 10 most critical failure modes, data must be generated on which the sorting process in the Pareto analysis can be based. Two options are available to generate this quantitative data: (1) collect failure data from the CMMS or (2) perform a FMECA analysis and calculate RPNs for all failure modes. Based on either the CMMS data (e.g. costs, MTBF) or the RPN values, the Pareto analysis will yield the top 5 cost drivers or performance killers.

Then, for the critical failure modes, a root cause analysis must be performed. It is essential that the level of detail of this RCA is such that the failure mechanisms and the internal loads for each failure mode can be assessed. Moreover, the relationship between the governing loads and the usage of the system must be identified, which implies that any excessive load can be linked to a certain usage condition. Monitoring data on loads and usage could be very useful in this assessment. These steps (4 and 5) in the process can be aided by the list of common failures in Sect. 8.5 and the decision scheme in Sect. 8.6, but also by the basic principles on loads and failure mechanisms in part I of the present book.

Note that an original equipment manufacturer (OEM) in general already has detailed knowledge on the possible failure modes and mechanisms of its system, because this type of knowledge is essential in the process of designing and developing the system. Therefore, an OEM can generally skip the first four steps in the procedure in Fig. 8.5. However, linking the loads on the system to the usage

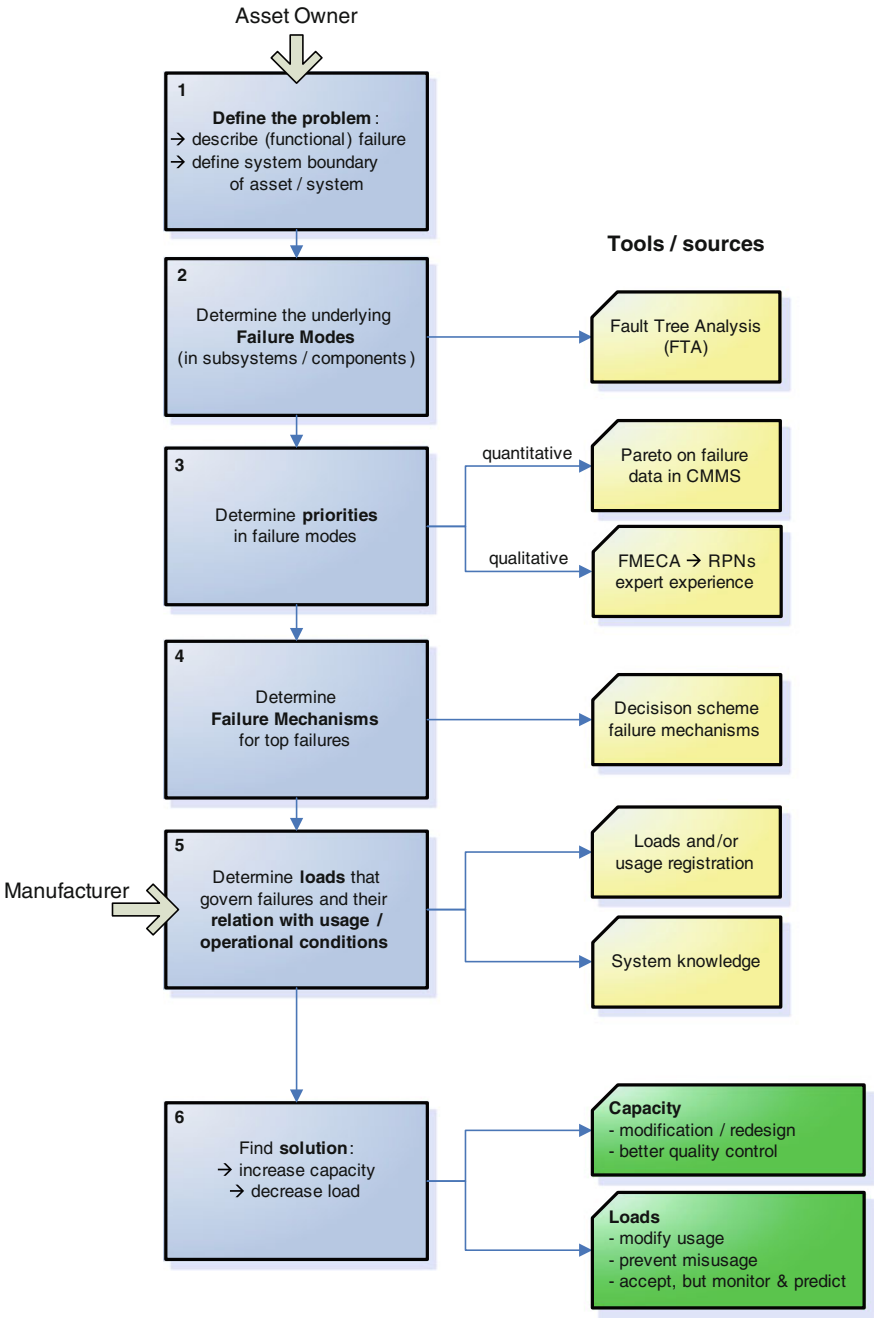

Fig. 8.5 Failure analysis process guidelines showing the different steps in the analysis and indicating the tools and sources for each step

profiles of different operators is not always trivial for an OEM. Very often the manufacturer has no access to usage data of the operators, while this information is essential in remedying the majority of the critical failures. The biggest challenge for an OEM in this procedure is therefore getting insight into the usage profiles.

Finally, the solution for the problem, that is, prevention of similar failures in the future, must be found. Since the failure mechanisms and governing loads have been determined, it is generally rather easy to decide whether the loads or the capacity of the system constitutes the root cause. For capacity problems, a modification of the system should be considered, while for loading problems, the usage and associated loading of the system should be reduced. If changing the usage profile of the system is not feasible, the (frequent) failures must be accepted, but setting up a monitoring program for the usage, loads or condition (see Sects. 6.4 and 6.5) may aid to make the failures predictable.

In the next section, the relevance of case studies on real systems will be discussed and the approach proposed in the present section will be demonstrated on a fire extinguishing pump case study.

8.4 Case Studies

The number of different systems operated nowadays in industry, transportation and infrastructure is enormous. Moreover, the systems consist of numerous subsystems and components, and all systems are operated differently in terms of loads, operational and environmental conditions. This makes that every failure is almost unique, and the number of failure modes is rather extensive. This is one of the aspects that make performing a failure analysis a complex process. The method proposed in the previous section can assist in following a structured approach, but experience is also a very important prerequisite.

To make the experience of investigators on failure analyses accessible to others, many books and reports have been published, containing case studies [7–9] on a wide range of failures. Also, several journals (e.g. Journal of Failure Analysis and Prevention, Journal of Engineering Failure Analysis) exist in this field. During the process of analysing a failure, studying the description and results of analyses on similar failures might provide parts of the solution of the problem under consideration.

In the present section, one additional case study of an engineering failure will be presented. Since this failure is also a very specific problem which might only be interesting for those directly involved in similar failures, the focus in this case study will be on the application of the approach presented in Sect. 8.3.

8.4.1 Centrifugal Pump

On board of ships, the fire extinguishing system is a very important and critical system. In case of fire, the system must be operative immediately, so high requirements are set to the reliability and availability of the system. To meet these requirements, the system consists of several centrifugal pumps that, on the one hand, ensure that the pressure of the system is maintained and, on the other hand, provide the required flow of sea water when the system is used.

Regarding this high criticality, a failure analysis is executed on the fire extinguishing pumps, in order to increase the system availability and at the same time decrease the maintenance costs. The procedure shown in Fig. 8.5 is followed, and the consecutive steps will be discussed next.

8.4.1.1 Problem Definition

The first step in the analysis is the problem definition, starting with the specification of a failure. For this analysis, failure is defined as 'the centrifugal pump is not functioning correctly'. This implies that several situations are regarded as a pump failure: the pump does not produce any flow (no yield), the pump produces insufficient flow (low yield) or the pump shows unusual behaviour (e.g. vibrations, heating up). Although this definition is not very precise, it does incorporate all possible failures in the analysis. Since in this case, the operator's insight into the failure behaviour is limited, the chosen general definition of failure ensures that no important failure modes are excluded. Another part of the problem description is the definition of the system boundary. The considered centrifugal pump, as shown in Fig. 8.6, is driven by an electric motor. It is decided that this motor is outside the considered system, and the boundary is at the shaft in between pump and motor.

8.4.1.2 Fault Tree Analysis

The next step in the failure analysis procedure is the execution of the fault tree analysis to identify all failure modes that could possibly lead to the pump failure as defined before. The resulting fault tree is shown in Fig. 8.7. Basically, three branches appear in this fault tree, which are associated with the three indicated failure conditions: no yield, low yield and unusual behaviour. For each condition, several lower-level failures have been identified and, ultimately, a number of basic failures are obtained, as represented by the coloured circles at the lower end of the fault tree. The meaning of the different colours will be explained later on.

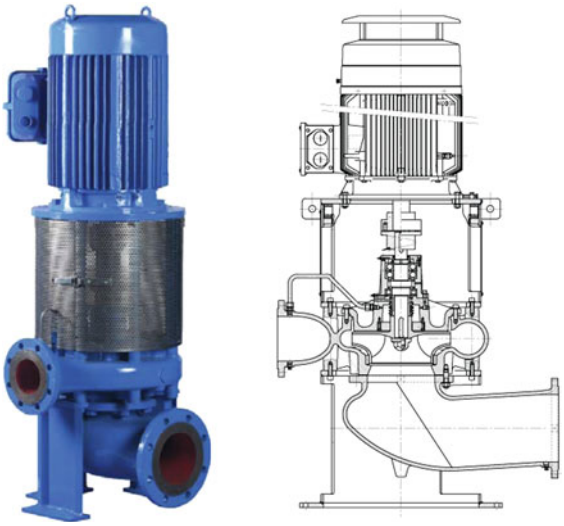

Fig. 8.6 Centrifugal pump driven by an electric motor (Published with kind permission of SPX Flow Technology Assen B.V.)

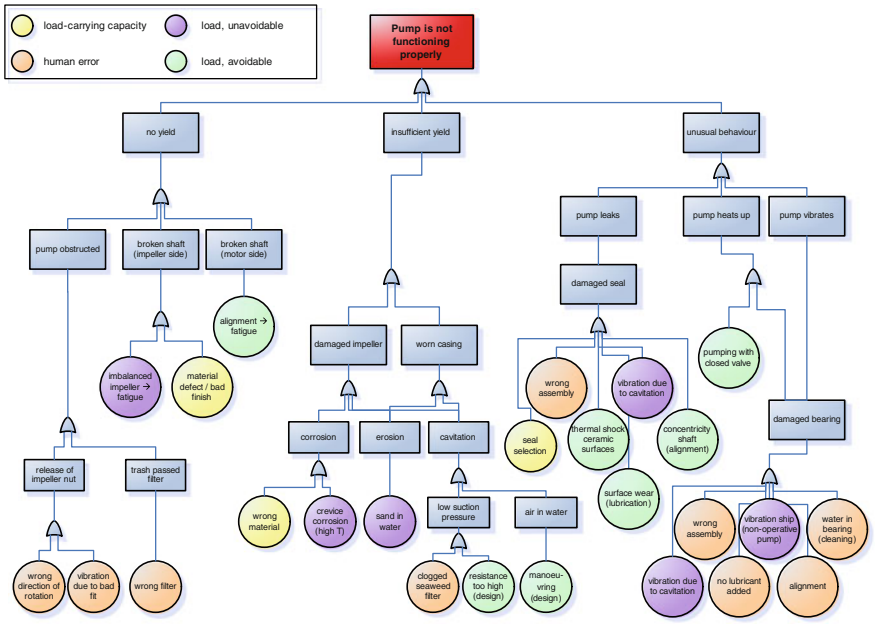

Fig. 8.7 Fault tree analysis of centrifugal pump. The *colours* of the basic failures indicate the type of failure cause

8.4.1.3 Determine Priorities in Failure Modes

Since the number of failure modes is considerable, a selection of the most critical modes must be made. The selection can be based either on costs or on failure frequency (which is related to availability). Since in this company, the maintenance costs for individual systems could not easily be retrieved, the focus has been on the failure frequencies. Therefore, all records in the CMMS associated with this type of pumps, installed across the fleet of ships, have been collected to get insight into the most frequent failures.

This analysis proved to be difficult, since the level of detail of the failure registration was limited. Although it was possible to get information on pump failures, the failure mode causing that failure was in most cases not available. Therefore, the experience of a group of operators and maintenance staff was utilized to prioritize the failure modes. The combined results of the CMMS data and expert experience yielded the following top-priority failure modes:

- Seal leakage
- No yield from impeller
- Bearing replacement (showing excessive vibrations)
- Shaft fracture

8.4.1.4 Determine Failure Mechanisms

The next step in the failure analysis is the assessment of the failure mechanisms causing the various failure modes. As was discussed before, this final deepening step of a root cause analysis is essential, since it provides valuable insight into the possible solutions for the problem. In this case study, the identification of the failure mechanisms is performed at two different levels. For each basic failure (i.e. failure mode) in the fault tree analysis, the failure cause has been selected from four possible types of causes:

1. Insufficient load-carrying capacity of the system or part (yellow): This is often caused by applying parts that do not comply with specifications. *Example* using an impeller manufactured from normal steel instead of stainless steel, resulting in corrosion problems.
2. Human error (orange): Often caused by disregarding regulations, by the absence of clear regulations or by inadequate training. *Example* cleaning the pumps with a high-pressure water jet removes the lubricant from the bearings, resulting in bearing failures.
3. Excessive load on the system due to avoidable (mis)use (green): The usage of the system deviates from the design specification, but can rather easily be changed to comply with the specifications. *Example* ceramic seals heat up when running the pump with no flow. A sudden opening of the valve produces a very high cooling rate, and the resulting thermal shock causes the seals to fracture.

Slowly opening the valve would allow the seals to cool down at a moderate rate.

4. Excessive load on the system due to unavoidable (mis)use (purple): The usage of the system deviates from the design specification, but adaptation of the usage is unacceptable or impossible. *Example* when the ship is in shallow water, the sea water ingested by the pump generally contains sand particles that cause erosion of the impeller.

In the fault tree in Fig. 8.7, the different types of causes have been indicated with the colour that has been mentioned for the four types above. After this general analysis, the four identified critical failure modes are analysed in more detail and the failure mechanisms are determined.

- Seal leakage: several failure mechanisms can cause this type of failure:
  - seal surfaces wear when no water is present in the pump during operation,
  - thermal shock due to sudden cooling of heated seals causes fracture,
  - vibrations of the pump (e.g. by cavitation in the impeller, misalignment or wrong assembly) yield high loads on the seals, producing overload damage.
- Insufficient yield: the impeller gets damaged due to several mechanisms:
  - corrosion due to prolonged exposure to sea water,
  - crevice corrosion due to elevated temperature in pump,
  - erosion due to (sand) particles in sea water flow,
  - fatigue damage due to cavitation in impeller.
- Bearing damage: this damage is caused by the following mechanisms:
  - local fatigue and wear damage due to vibration of non-operative pumps (caused by vibration of other machines in same room),
  - wear damage due to misalignment (high loads), bad lubrication or bad assembly of the bearings.
- Shaft fracture: these failures are caused by misalignment of the pump and motor. The shaft is then loaded asymmetrically, which yields a fatigue fracture after a certain number of revolutions. The alignment is performed periodically by maintenance staff, but apparently the procedure is not adequate.

8.4.1.5 Determine Loads and Their Relation with Usage

As was mentioned in the previous step, the division into four different types of causes actually combines the failure mechanism assessment and the loads to usage linkage. Especially in the case of overloading the system (either avoidable or unavoidable), the relationship of the observed overload with the usage could be established in most cases.

For example, one of the failures is overheating of the pump, caused by operating the pump, while the valve in the output circuit is still closed. The absence of any water flow inside the pump implies that no cooling is realized, and the pump starts to heat up. It is clear that the thermal loading of the pump in this case is directly related to the closed valve. Moreover, such a failure can be prevented by prescribing that the pump is not allowed to be operated, while the valve is closed.

8.4.1.6 Find Solution

The final step of the failure analysis procedure is to find solutions for the problem, that is, ways to prevent similar failures in the future. Since the type of cause has been identified for all failure modes in one of the previous steps, finding solutions is generally rather straightforward. Each of the four types of causes has a clear solution direction:

1. Insufficient load-carrying capacity: modification or redesign of the system to increase the capacity. In case of non-compliant parts: better quality control of spare parts.
2. Human error: better regulations, better training.
3. Avoidable (mis)use: change the usage profile to bring it back within specifications.
4. Unavoidable (mis)use: accept failures to occur, but try to make them predictable by usage or condition monitoring.

It can be observed in Fig. 8.7 that for the present case study, human errors (orange circles) occur relatively often and that also quite a number of avoidable overloading (green) takes place. As was indicated above, these failures can be prevented rather easily by changing the way the system is operated and by improved training of personnel. The four critical failure modes can be solved in the following way:

- Seal leakage: all failure mechanisms mentioned in the previous step are caused by operating the pump incorrectly. Therefore, the failures can be prevented by better instruction and training of the operators. Especially, preventing that the pump is operated without flow is important. Also ensuring that assembly and alignment are executed in the right manner can prevent many failures. The excessive wear of the seals when running dry might also be reduced by applying an oil-lubricated seal.
- Insufficient yield: most of the impeller failures are unavoidable, since they are due to regular usage of the pumps. Making the failures predictable, for example, by monitoring the number of operating hours while in shallow water, might reduce this problem. The damage due to cavitation might be prevented, since the occurrence of cavitation is related to the way of operating the pump.

- Bearing damage: the local damage in non-operative (vibrating) pumps can be reduced by periodically running the pumps, ensuring that other locations in the bearings are loaded. All other failures can be prevented by improving procedures for alignment and assembly of new bearings.
- Shaft fracture: improve the alignment procedure and training to ensure proper alignment of the system. Additionally, the system can be made more robust by replacing the fixed coupling between pump and motor by a magnetic coupling. The latter is much less sensitive for misalignment.

8.4.2 Conclusions from Case Studies

The case study in this section demonstrates that the procedure proposed in Sect. 8.3 can be applied to real-world problems. Moreover, it demonstrates that analysing the failure mechanisms and associated loads provides the leads for solving the problems in a rather straightforward way. This case study was part of a recent research project [9], in which more case studies from industry have been analysed using this approach. The general conclusions from those case studies can be summarized as follows:

- Obtaining details on failure mechanisms from information systems like CMMS appears to be cumbersome. In most systems, such detailed information is not collected, or the quality of the information is rather low due to the lack of knowledge and the lack of envisaged benefit of the data collection at the people entering the data.
- If the data on failure mechanisms or failure frequencies cannot be obtained from information systems, interviews with experts (operators, maintainers) are found to provide much of the required information. Also involving the OEM of the system in the analysis is experienced to be very useful. As the designer of the system, the OEM has detailed insight into the failure modes and loading of the system.
- A significant fraction of the failures is due to human errors or abuse of the system. Therefore, before considering modifications of the system, considerable improvements in system availability can be realized by better instruction and training of operators. Creating the awareness that certain ways of operating the system lead to many failures is essential, but often requires a detailed analysis of the frequent failures to be able to find root causes and set priorities.

8.5 Lists of Common Failures

To aid the root cause analysis of common failures, this section provides an overview of typical failures in industrial equipment and systems. For each failure mode, the possible failure mechanisms and the associated loads are described. During the root cause analysis, this overview may provide the possibly active failure mechanisms, but more importantly, also directly indicates which load governs that mechanism. The latter relationship is crucial to find a sound solution for the observed problem, as was demonstrated in the previous sections. The detailed information on the various loads and failure mechanisms in part I of this book may then help to make a detailed quantitative analysis.

As was mentioned earlier, a very extensive number of failure modes exist. Each system or piece of equipment can functionally fail in several ways. It is therefore impossible to make a complete list. However, finding failures in the lists provided in this section that are similar to a failure to be analysed might assist in determining the precise failure mechanism and associated loads. Therefore, this section provides overview of dominant failure modes in a variety of system/component types. These are shown in Tables 8.4, 8.5, 8.6, 8.7, 8.8 and 8.9 for rotating equipment, static equipment, standard components and electronic components and devices, respectively.

8.6 Decision Scheme for Failure Mechanisms Determination

The present book has stressed numerous times that knowing and understanding the failure mechanism is an important requirement to improve the reliability of a system and set up effective maintenance programs. However, assessing the mechanism when a failure has occurred requires a certain level of knowledge and mostly also a certain amount of experience. It is the aim of the present book to provide the knowledge, but experience cannot be obtained from books.

Nevertheless, in the present section, a decision support scheme will be presented that aims to assist in determining the failure mechanism for a certain failure. The scheme is based on the working approach of experienced failure analysts. Moreover, it combines the three major sources that can evidence the occurrence of a certain failure mechanism: fracture surface analysis, load history analysis and material analysis.

Fracture surface analysis in many cases provides valuable information on the failure process [7, 10]. By inspecting the fracture surface by optical or electron microscopy, specific characteristics may be observed that can be linked to a specific failure mechanism. Some examples of fracture surfaces and their specific characteristics for various failure mechanisms are shown in Chap. 4.

Table 8.4 Failure modes, mechanisms and associated loads in rotating equipment

System	Component	Failure mode	Failure mechanisms	Load
Gas turbine*	Turbine blade	Fracture	Overload, fatigue (LCF), creep	Centrifugal + thermal stress, temperature
		Elongation/deformation	Creep, plastic deformation	Centrifugal stress, temperature
		Surface damage	Oxidation, erosion	Temperature, particle size + speed
	Compressor blade	Fracture	Overload, fatigue (LCF + HCF)	Centrifugal stress, vibrations
		Surface damage	Oxidation, erosion	Temperature, particle size + speed
	Combustor	Surface damage	Oxidation	Temperature
		Cracking	Thermal fatigue, creep	Thermal stress, temperature
		Deformation	Creep	Thermal stress, temperature
Diesel engine	Piston	Dimensional change (clearance)	Wear along cylinder wall	Normal load, speed, friction
		Seizure	Thermal expansion	Temperature (overheating)
	Fuel injector	Fractured spring	Fatigue	Mechanical stress in spring
	Valve	Fracture	Fatigue	Mechanical stress
		Leakage	Erosion	Gas velocity and composition
	Cylinder	Cracking	Fatigue	Thermal + mechanical stress
		Dimensional change (clearance)	Wear (piston)	Normal load, speed, friction
Generator	Rotor	Winding short circuit	Intrinsic breakdown	Electric field
	Stator	Coil winding failure	Intrinsic breakdown	Electric field, mechanical stress, temperature
Pump (centrifugal)	Impeller	Dimensional change	Erosion, corrosion	Fluid composition, temperature and velocity
	Casing	Surface damage, leakage	Erosion, corrosion	

* *Note* General components present in all rotating equipment, e.g. bearings and shafts, are shown in Table 8.6

Table 8.5 Failure modes, mechanisms and associated loads in static equipment

System	Component	Failure mode	Failure mechanisms	Load
	Pipe	Reduced wall thickness, leakage	Corrosion	Chemical, temperature
			Erosion	Mechanical stress due to erosive Fluid/gas
		Burst	Overload, fatigue, creep	Mechanical stress (pressure), temperature
	Valve	Non-functioning (open or close)	Corrosion, wear (high friction)	Chemical, temperature, mechanical stress + motion
		leakage	Erosion, corrosion	Mechanical stress due to erosive fluid/gas, chemical, temperature
	Tank	Leakage	Corrosion, fatigue, creep	Chemical, mechanical stress, temperature
		Burst	Overload, fatigue, creep	Mechanical stress (pressure), temperature
Heat exchanger		Fouling	No physical mechanical, process-related	-
		Leakage	Corrosion, erosion	Chemical, temperature
		Fracture of pipes	Stress corrosion cracking, fatigue, overload	Mechanical stress, chemical, temperature

The load history of a failed component may also yield important information on the failure mechanism. On the one hand, the type of loading, for example, mechanical (either static or cyclic), thermal or electric, determines which class of failure modes may have been active, while on the other hand, the magnitude of the load and variation in the load can provide insight into the expected time to failure for a specific failure mechanism. The estimated service life can then be compared with the observed time to failure to decide whether the considered mechanism could be responsible for the failure.

Finally, also analysis of the material is often required to determine the failure mechanism. The first reason is that some materials are hardly affected by certain failure mechanisms. For example, stainless steel applied in ambient conditions is not expected to corrode extensively, so failure of such a component due to corrosion is not very likely. Similarly, many steel alloys have a clear fatigue limit, which means that those materials are not expected to suffer from fatigue at low applied stress levels. A second reason to analyse the material is to asses the load-carrying capacity of the material, since that quantity represents the boundary between failure and no failure. The capacity, for example, the tensile strength of a metal, can often be obtained from material handbooks. However, these nominal values may not always be valid. The material microstructure may have degraded, for example, due to prolonged exposure to elevated temperatures or due to a welding process, which also may have reduced its capacity. Also, the chemical composition of the material may deviate from normal specifications, which

Table 8.6 Failure modes, mechanisms and associated loads in standard components

Component	Failure mode	Failure mechanisms	Load
Bearing			
Rolling element	Seizure	Wear (surface) fatigue of races	Mechanical stress (shaft), # revolutions
		Cage or elements	
		plastic deformation	mechanical stress (shaft)
	Vibrations/ noise	wear, surface fatigue	mechanical stress (shaft), # revolutions
Shaft	Fracture	Overload, fatigue, creep	Mechanical stress, temperature
	Deformation	Plastic deformation, creep	Mechanical stress, temperature
Seal	Leakage		
Metal		Wear, deformation, corrosion	Mechanical stress, motion, chemical
Ceramic		Fracture, wear	Mechanical + thermal stress, motion
Synthetic		Deformation, degradation	Mechanical stress, temperature
Gear set	Teeth damage	Fatigue, wear	Mechanical stress, # revolutions, friction
	Teeth fracture	Overload, plastic deformation	Mechanical stress, temperature

generally also changes the mechanical properties. Although microstructural and chemical analysis requires quite some knowledge, experience and facilities, the results may provide a direct indication for the failure mechanism.

The challenge in determining the failure mechanism that has caused a certain failure in practice is combining the information obtained from the failure, that is, the visible features, with knowledge on the load history and material. This reasoning process is quite similar to the process of diagnosing a patient in a medical setting. From an initial set of possible causes, certain possibilities can be excluded based on available information (e.g. obtained from interviews or simple tests). Ideally, in the end, only one possible cause remains.

A similar procedure is proposed here for the failure mechanism determination process. Based on the visible features of the failure, an initial set of possible failure mechanisms are constructed. Then, by going through a set of questions, all except one mechanism are tried to be excluded. The three mentioned sources to find evidence for a certain failure mechanism are the ingredients for the questions. With a focus on mechanical failures, the following three starting points, that is, failure features, for a failure analysis have been defined:

1. A crack is visible or the part is completely fractured.
2. The part is deformed.
3. Surface damage is observed at the investigated part.

Table 8.7 Failure modes, mechanisms and associated loads in electronic components

Component	Failure mode	Failure mechanisms	Load
Resistor	Fail open	Cracking	Mechanical stress
		Melting	Electric current
Adjustable resistor potentiometer	Bad contact/changed resistance in wiper	Wear	Mechanical stress, Sliding distance
		Deformation	Mechanical stress
		Corrosion	Humidity
		Surface contamination	
Capacitor	Shorting or partial loss of dielectric strength	Breakdown, often due to degrading dielectric	Electric field, temperature
	Electrolyte dry-out, contamination or gas creation	Leakage (mechanical)	Stress, temperature
Cables	Fracture	Chaffing—wear	Movement/vibration
		Static overload	Mechanical stress
Connectors/leads	Failing of solder (especially surface mounted)	Static overload	Temperature (change) mechanical stress
	Failing of solder due to brittle intermetallic layers	Static overload + thermal degradation	Temperature (change) mechanical stress
	Fracture of leads	Static overload, fatigue	Temperature (change) mechanical stress
	Non-conductive lead due to oxide formation	Corrosion	Humidity, temperature

Table 8.8 Failure modes, mechanisms and associated loads in semiconductor components

Component	Failure mode	Failure mechanisms	Load
Diode	Exceeding reverse breakdown voltage	Breakdown	Electric field
Transistor	Property change	Electromigration	Electric field
		Melting/vaporizing of metallization layers	Electric current

For these three starting points, the potential failure mechanisms and a set of questions have been defined in Tables 8.10, 8.11 and 8.12, respectively. The procedure to determine the failure mechanism is then as follows:

1. Select the appropriate starting point based on the features of the failure.
2. Write down the proposed potential failure mechanisms (FM).
3. Process all questions in the appropriate table column by column. The answers to the questions (yes or no) determine the action to be taken:

Table 8.9 Failure modes, mechanisms and associated loads in electronic circuits

Component	Failure mode	Failure mechanisms	Load
Printed circuit boards (PCB)	Cracking of traces	Fatigue	Mechanical stress, temperature (thermal stress)
		Static overload	
	Corrosion of traces	Corrosion	Humidity, temperature
	Short circuit between traces due to growth of conducting filaments	Corrosion	Temperature, humidity
		Diffusion	Temperature
		Breakdown	Electric field
Integrated circuits (IC)	Failure due to electrostatic discharge	Melting, evaporation	Electric current
		Breakdown	Electric field
	Fracture	Overload	Mechanical/thermal stress
	Functional failure	(inter)diffusion	Temperature
	Increased resistance	Corrosion	Temperature, humidity

a. for questions labelled with ***rq***: this condition is *required* for the associated FM to occur. Therefore:

(i) negative answer: the mechanism cannot be active → remove this FM from the list;
(ii) positive answer: this FM is still possible → leave the FM on the list;

b. for questions labelled with ***c***: this condition *confirms* the presence of the associated FM. Therefore:

(i) negative answer: this FM is still possible → leave the FM on the list;
(ii) positive answer: this FM is evidenced to be active → remove the other FMs from the list;

c. for questions labelled with ***x***: this condition *excludes* the presence of the associated FM. Therefore:

(i) negative answer: this FM is still possible → leave the FM on the list;
(ii) positive answer: this FM is excluded → remove the FM from the list;

4. After processing all questions in the appropriate table, in most cases, only one potential failure mechanism remains, which must be the mechanism that caused the failure. If still more than one FM remains, additional information or testing must be arranged to differentiate between the remaining FMs.

Table 8.10 Decision support scheme for failure mechanism determination I: crack or fracture

Observed feature: visible crack or fractured part

Possible FM	Load		Material		Fracture surface		Other questions	
Overload	Has a load in the order of the material strength be applied?	**rq**	Is the material strength possibly reduced by bad production (microstructure, voids) or processing (e.g. welding)?	**rq**	Can dimples be observed?	**c**		
Fatigue	Is a cyclic load (mechanical or thermal) present?	**rq**	Is the cyclic load exceeding the material fatigue limit (if present)?	**rq**	Can striations be observed?	**c**		
Creep	Is a combination of elevated temperature (> 400 °C) and a mechanical load present?	**rq**	Does the applied combination of stress and temperature lead to finite creep rupture times in the present material?	**rq**	Can facets be observed?	**c**		
Wear (surface fatigue)	Is there a contact situation with repetitive loading (e.g. rolling contact)?	**rq**	Is the combination of contact stress level and number of repetitions sufficient to cause fatigue cracking in the present material?	**rq**			Completely fractured or through crack (not just surface crack)?	**x**
Stress corrosion cracking (SCC)	Is a combination of a corrosive environment and a static mechanical load present?	**rq**	Is the material susceptible to stress corrosion cracking?	**rq**	Is any evidence of corrosion visible?	**rq**		

Table 8.11 Decision support scheme for failure mechanism determination II: deformed

Observed feature: deformed part							
Possible FM	Load		Material		Fracture surface	Other questions	
Plastic deformation	Has a load exceeding the material yield strength been applied (either mechanical or thermal)?	**rq**	Is the material strength possibly reduced by bad production (microstructure, voids) or processing (e.g. welding)?	**rq**			
Creep	Is a combination of elevated temperature (>400 °C) and a mechanical load present?	**rq**	Does the applied combination of stress and temperature lead to a creep strain rate in the present material that might produce the observed deformation?	**rq**			
Wear (adhesive, abrasive)	Is there a contact situation between two parts that move relative to each other?	**rq**	Are the friction and normal load in the contact sufficiently high (e.g. unlubricated) to cause a dimensional change due to wear?	**rq**		Are wear tracks visible on the surface?	**c**
Erosion	Is there a contact situation between a medium (gas, fluid, possibly containing particles) and a part?	**rq**	Are the friction and normal load in the contact sufficiently high to cause a dimensional change due to wear in this material?	**rq**		Are erosion tracks visible on the surface?	**c**
Melting	Has a high temperature close to the material melting temperature been present?	**rq**	Is the nominal material melting temperature possibly reduced by bad production (chemical composition)?	**rq**		Are any signs of flowing material (in fluid condition) visible?	**c**

Table 8.12 Decision support scheme for failure mechanism determination III: surface damage

Observed feature: surface damage							
Possible FM	Load		Material		Fracture surface	Other questions	
Wear	Is there a contact situation between two parts that move relative to each other?	**rq**	Is the friction and normal load between the two moving parts in contact sufficiently high (e.g. unlubricated) to cause a dimensional change due to wear in this material?	**rq**		Are wear tracks visible on the surface?	**c**
Erosion	Is there a contact situation between a medium (gas, fluid, possibly containing particles) and a part?	**rq**	Is the friction and normal load between the part and medium/particles in contact sufficiently high to cause a dimensional change due to wear in this material?	**rq**		Are erosion tracks visible on the surface?	**c**
Melting	Has a high temperature close to the material melting temperature been present?	**rq**	Is the nominal material melting temperature possibly reduced by bad production (chemical composition)?	**rq**		Are any signs of flowing material (in fluid condition) visible?	**c**
Corrosion	Is a corrosive environment present?	**rq**	Is the material susceptible for corrosion?	**rq**		Are any signs of corrosions (e.g. oxide layer, blisters, pits) visible?	**c**

8.7 Summary

In this chapter, four existing methods to analyse failures, their effects and their causes have been discussed. Both methods to assess possible future failures, that is, FMECA and FTA, and methods analysing failures that already occurred, RCA and Pareto, have been treated. After that, the mechanism based failure analysis procedure has been proposed. This method follows the structured approach of existing methods but at the same time optimally utilizes the knowledge on loads and failure mechanisms introduced in the first part of this book. The procedure has also been applied to a real system in practice, a centrifugal pump.

After this discussion of analysis methods, the remainder of the chapter was devoted to the determination of the failure mechanism. Once the precise failure mechanisms are known, finding a way to prevent such a failure has been shown to be rather straightforward. Therefore, the final two subsections provided background information and procedures that assist less-experienced people in assessing the failure mechanism of a failure at hand.

References

1. Moubray, J.: Reliability-Centered Maintenance. Industrial Press, New York (1997)
2. Braaksma, A.J.J., Klingenberg, W., Veldman, J.: Failure mode and effect analysis in asset maintenance: a multiple case study in the process industry. Int. J. Prod. Res. 1–20 (2012) (in press)
3. Manzini, R., Regattieri, A., Pham, H., Ferrari, E.: Maintenance for Industrial Systems. Springer Series in Reliability Engineering. Springer, London (2010)
4. Banks, J.C., Reichard, K.M., Hines, J.A., Brought, M.S.: Platform degrader analysis for the design and development of vehicle health management systems. In: International Conference on Prognostics and Health Management (2008)
5. Keppner, C.H., Tregoe, B.B.: The Rational Manager: A Systematic Approach to Problem Solving and Decision-Making. McGraw Hill, New York (1965)
6. Marquez, A.C.: The Maintenance Management Framework: Models and Methods for Complex Maintenance. Springer Series in Reliability Engineering. Springer, London (2007)
7. Nishida, S.: Failure Analysis in Engineering Applications. Butterworth-Heinemann, Oxford (1992)
8. Jones, D.R.H. (ed.): Failure analysis case studies II. Elsevier Science, Oxford (2001)
9. Tinga, T.: Mechanism based failure analysis. Final report of World Class Maintenance Innovation Project—WP4. World Class Maintenance, Breda (2012)
10. Wulpi, D.J.: Understanding how components fail, 2nd edn. ASM International, Materials Park (1999)

Further Reading

1. Manzini, R., Regattieri, A., Pham, H., Ferrari, E.: Maintenance for Industrial Systems. Springer Series in Reliability Engineering. Springer, London (2010)
2. Nishida, S.: Failure Analysis in Engineering Applications. Butterworth-Heinemann, Oxford (1992)

Chapter 9
Design

9.1 Introduction

The foregoing chapters in part II of this book focused on application of the knowledge on loads and failure mechanisms in maintenance and reliability. This means that in most cases, systems are concerned that are already in the operational phase of their life cycle. Failures may already have occurred or are expected to occur, and methodologies have been presented to prevent future failures and optimize the system availability accordingly.

However, also during the design phase of the system life cycle, understanding the system failure behaviour and estimating the expected loads on its subsystems and components is important. In the design phase, decisions are made on material selection, dimensioning of the parts and assembly of the system. These decisions have a major impact on the reliability of the system and thus affect the maintenance effort required to assure a certain system availability.

In this final chapter, the benefit of understanding the failure behaviour on the design of systems will be discussed. The next subsection explains the concept of life cycle management. Then, the important aspects of design for maintenance will be treated in Sect. 9.3, and a number of design philosophies will be presented in 9.4. Finally, the concept of probabilistic design will be explained in Sect. 9.5.

9.2 Life Cycle Management

The term life cycle management (LCM) is used to indicate all the activities that are required to exploit a capital asset in the most effective and efficient manner during its complete life cycle. The life cycle of an asset constitutes all phases from the initial idea to the final disposal. In systems engineering, the following phases are recognized [1]: conceptual design, preliminary system design, detailed design and

T. Tinga, *Principles of Loads and Failure Mechanisms*,
Springer Series in Reliability Engineering,
DOI: 10.1007/978-1-4471-4917-0_9, © Springer-Verlag London 2013

development, production and/or construction, distribution, operation, maintenance and support, retirement, phase-out and disposal.

Closely related to LCM are the concepts of life cycle costing (LCC) and total cost of ownership (TCO). A classical investment decision was primarily based on the initial investment costs. However, it is now realized more and more that the additional costs during the operational phase, for example, for maintenance and operational supplies, can reach the same order of magnitude as the initial investment. And for certain very specific types of assets with typically a long service life ($\sim$25–30 years), for example, military aircraft, the operational costs can be a factor 3–5 times the initial investment costs.

This observation makes that deciding on initial investment is being replaced by taking decisions based on the life cycle costs: All the costs during the complete life cycle, including maintenance, supplies and disposal, are taken into account. The sum of all these costs, both the direct life cycle costs and the associated fixed costs (e.g. required facilities) constitute the total cost of ownership. For example, if a new system is to be acquired and two OEMs can deliver a system that meets the functional requirements, the manufacturer offering the system for the lowest price is not automatically the best option. If the other OEM delivers the system with a condition monitoring system, that slightly raises the investment costs, selecting this system may yield considerably lower operational costs, since less unnecessary maintenance is executed. The life cycle costs will therefore be significantly lower, which more than compensates for the higher initial investment.

Another important issue is that the majority of the life cycle costs are already fixed during the design phase. The decisions taken there greatly determine the amount of required maintenance and the need for certain types of spare parts, the level of mechanics and the required facilities, etc. Therefore, considerably reducing the life cycle costs is only possible by making changes during the design phase.

It is concluded that, although it is difficult to estimate all future costs accurately, taking into account the operational and disposal costs in an investment decision and during the design process to some extent generally leads to a lower total cost of ownership.

9.3 Design for Maintenance

Most design processes focus on realizing a system that best meets the functional requirements, that is, delivers the best performance. Since maintainability and supportability of the system are generally not included in the functional requirements, these aspects can severely hamper the performance of the system during operation. A well maintainable and supportable system can be realized in several ways [2]:

- Producing a reliable system, with low numbers of failures, reducing the absolute amount of maintenance required;

- Making the system easy to maintain, thereby reducing the average maintenance time;
- Ensuring that the system is easy to support, reducing the costs of supplies and facilities.

Each of these three aspects will be treated separately in the next subsections.

9.3.1 System Reliability

The best way to design a system that has a high availability during service is to ensure a high reliability of the system. If the system rarely fails, and thus only needs a limited amount of maintenance, the complexity of the maintenance tasks or the supportability of the system do hardly affect the availability.

A reliable system can be realized in several ways. The first way is to ensure a high load-carrying capacity of the components. This can be achieved during the design process by selecting high quality materials, as is for example done in aerospace applications. Another way to create a high capacity is over-dimensioning of parts. By explicitly increasing wall thicknesses or the dimensions of critical cross-sections, the internal loads on the parts can be reduced, as was explained in Chap. 3. This approach is often followed in (infra)structural assets like bridges, tunnels and dams, where the additional weight of the structures is no problem.

A second way to design highly reliable systems is to include redundancy in the system. This means that several subsystems or components with similar tasks are incorporated in the system, ensuring that failure of one of those subsystems not immediately leads to failure of the complete system. If the failed subsystem is repaired or replaced before the remaining subsystems have failed, the system will not face any downtime.

Other ways to increases the system reliability [2] are the application of components and materials with a verified reliability (only use 'proven technology') or the minimization of the number of parts, and especially the number of moving parts. The latter is motivated by the fact that moving parts are generally more prone to failure than static parts.

It is clear that incorporating these approaches in the design process to increase the system reliability require insight in the failure behaviour of the system; especially the balance between the loading of the system and its capacity must be clear.

9.3.2 Maintainability

Although the aim of the designer must be to design a reliable system, it is in most cases impossible to prevent all failures. In that case, the designer should also consider how difficult the required maintenance will be, since that might severely affect the system availability.

The two most important ways to increase the maintainability of the system are improving accessibility and designing modular systems. When the subsystems or components that are expected to be replaced regularly (either preventively or correctively) can be accessed easily, the associated maintenance tasks can be completed in a short period of time. If, however, many other components will have to be disassembled before the specific component can be reached, the downtime of the system will be considerably longer.

The second way to improve maintainability is the concept of modularity. This means that in case of failures, complete modules or subsystems can easily be removed and replaced. The removed module is then repaired offline in a workshop, while the system downtime is only limited.

Finally, also application of standard components is the system generally increases the maintainability [2], since a technician recognizes the parts and knows how to (correctly) maintain them.

9.3.3 Supportability

The supportability of the system is determined by the supplies and facilities that are required to maintain the system. An important factor in this aspect is the availability of spare parts. Parts with a long lead time yield a low supportability of the system. It is therefore advisable to apply standard components and subsystems that can be applied across different systems within the fleet and can be obtained from several suppliers. Application of the spares across the fleet enables the support from other local stock points (which keep the same parts in stock) in case of a stock-out. When parts can be obtained from several suppliers, the supplier with the shortest lead time can be selected.

Another aspect that affects the system supportability is the use of specific equipment or facilities. If these are not available at the location where maintenance must be executed, the downtime of the system will increase.

Finally, incorporating suitable prognostic methods (see Chaps. 6 and 7) may considerably improve the logistic forecast, which means that the supply chain can timely anticipate to future maintenance tasks.

9.4 Design Philosophies

A system and all its components are designed for a certain service life; especially in critical applications, as are found in sectors like aerospace and nuclear power generation, it is important that components do not fail during the expected service life. As was illustrated in part I of this book, the basic manner to achieve this is to design the components such that the expected loads will not cause significant damage during this service life.

However, several design philosophies can be adopted by a manufacturer to realize this. The most important approaches that are commonly used for aerospace components are the safe-life and damage tolerance philosophies. The choice for a certain philosophy also brings about implications for the maintenance strategy or service life management, that is, the activities that must be employed to ensure safe operation of the component or system, like inspections, replacements and repairs. Therefore, the selection of a certain philosophy is only possible when the associated maintenance policy is feasible, for example, some policies require that the system is inspectable.

The two design philosophies, which are used for both airframe and aero engine design, will be described generally in the next subsections [3]. After that the more specific application details will be illustrated.

9.4.1 Safe-life

The first design philosophy that was developed in the 1950s is the safe-life (SL) approach. In the safe-life approach, a component is designed for a finite service life during which significant (fatigue, creep) damage will not occur. Basic to this approach is that either the structure is not inspectable or that no inspections are planned during the service life. Service life management of safe-life components appears to be simple: No inspections are planned, and the components must be retired at the certified lifetimes. In practice, this gives some complications. Components may be found prematurely damaged, requiring repair, replacement or redesign and replacement. On the other hand, many components reach their certified lifetimes with little or no indications of damage (see also Sects. 6.2 and 6.3), and there is an understandable wish to extend their lives. Service life extension for safe-life components means an increasing risk of failure that is poorly quantifiable. For non-critical components, this situation may be acceptable, but it is not for critical components like gas turbine blades, discs and shafts. In other words, service life extension for critical safe-life components is difficult or impossible even when they show no evidence of damage. This is one of the main reasons why alternative philosophies have been developed, as will be discussed in the next subsection.

Although the safe-life philosophy originated in the aerospace world, it is also widely applied to other applications. In fact, in all applications where weight is not a decisive design criterion, it is possible to include a certain safety factor in the design of components. For example, by increasing the area of critical cross-sections to reduce the internal load, the service life of a component can be increased beyond the system service life, which actually implies that a safe-life approach is followed: The part is not expected to be damaged or fail during its service life. In other words, the design does not allow any failures [4].

9.4.2 Damage Tolerance

An alternative to the safe-life design philosophy has been developed since 1970 by the United States Air Force (USAF). This philosophy is called the damage tolerance (DT) approach and differs from the original safe-life approach in two major respects:

1. The possibility of cracks or flaws being present in new structures must be explicitly accounted for;
2. Structures are considered either to inspectable or to be non-inspectable in service.

The key aspect of the damage tolerance approach is the first point, that is, the presence of initial damage in a part cannot be avoided in all cases and thus must be accounted for. And since the design allows failures, the safety of the system or structure must be ensured by maintenance [4]: (1) Regular inspection must be performed to assess the size and growth rate of the damage; (2) components must be replaced at specified times to prevent the initial damage to reach a critical size. Note that, although the philosophy may be generalized to all slowly developing degradation mechanisms, it is mostly applied to structural parts in which fatigue is the dominant failure mechanism.

Components designed according to the safe-life approach are not inspected during service since no damage is expected to occur. However, in non-inspectable damage tolerance designs the damage that is expected to be present cannot be assessed, which provides a safety risk. Therefore, to ensure that the damage will not reach a critical size, these designs should be qualified as either slow crack growth structures or fail-safe systems. For slow crack growth structures, initial damage must grow slowly and not reach a size large enough to cause failure before the end of the service life. Fail-safe systems are designed such that, in the event of a failure, they respond in a way that will cause no harm. In a structural part, this can be achieved by creating multiple load paths. If one load path fails, the remaining paths can carry the load, thus preventing complete failure of the part. Another example of a fail-safe system is the air brakes on railway trains and trucks. The brakes are held in the off position by air pressure created in the brake system. In case of a brake line fracture, the air pressure will be lost and the brakes applied.

In the damage tolerance approach, it is thus recognized that, specifically for fatigue critical components, it can contain a manufacturing defect or material discontinuity which could act as a crack starter, that is, a site where crack growth commences. This can, in the first instance, be interpreted conservatively as follows:

1. The crack starter sizes are at the detection limits of pre-service non-destructive inspection (NDI) techniques;
2. Crack growth starts as soon as the components enter service.

In practice, one of these two interpretations is often relaxed, since otherwise the calculated service lives would be very limited. Therefore, the first assumption is often relaxed by reducing the initial crack sizes to 'equivalent initial flaw sizes' (EIFS). These are obtained by crack growth calculations that retrace crack growth from the final crack sizes and fatigue (LCF) lives of tested components. The second assumption is relaxed by adding crack initiation lives to crack growth lives calculated using fracture mechanics (see also Sect. 4.4.6). It is thus assumed that it takes a certain period of time (or actually number of cycles) before a microscopic material defect has evolved to a macroscopic crack, which thus extends the component's service life.

The maintenance policy or service life management for the damage tolerance approach implies that in-service inspections are planned to assess the damage evolution at regular instances. Moreover, components must be retired at a certified lifetime (=life limit), which is calculated using a crack propagation analysis starting from the assumed initial crack size. The replacement intervals of the components are thus not affected by the inspection results: not finding a crack does not imply that the interval can be extended. The reason is that inspections are not perfect and a crack can always be missed. Only in case of an unexpectedly large crack, the component will be replaced before the end of the certified interval. The advantage of using damage tolerance concepts instead of the traditional safe-life approach lies not so much in obtaining longer lives, but in making safety more quantifiable.

9.4.3 Application

The vast majority of aircraft structural parts and gas turbine components are treated according to the safe-life approach. Only for a small but increasing number of components, the damage tolerance approach is used. Application to real parts is performed as follows.

In the conventional safe-life approach, the life limit for a fatigue-dominated component is calculated with an *S*–*N* curve (see Sect. 4.4.4). This life limit is associated with the time required to initiate a 1/32-inch-long surface crack in a part with no pre-existing defect. The value is determined from a large amount of test data, which gives a distribution of crack initiation lives. For the LCF life limit, the B.1 value (=the time where in 1 out of 1,000, or 0.1 %, a crack has initiated) is used.

In the damage tolerance approach, the initial life limit estimate is calculated following the safe-life approach, but in addition to this, the time for a 1/32 inch crack to grow to a critical size is calculated. The life limit is then calculated as follows:

$$\text{Life Limit} = \text{B.1 Initiation Life} + \text{B.1 Propagation Life} \tag{9.1}$$

The B.1 value of the initiation life is the same value as used in the safe-life approach. The B.1 propagation life is obtained in a similar way. Again a lot of tests are done, and the time where 1 out of 1,000 has reached the critical crack length is the B.1 propagation life or safety limit. Note that the assumed initial damage consists of manufacturing defects. Inspections are planned during service, and if a crack is detected, the component is retired. However, the chance that this happens is very small, because both for the initiation life as for the propagation life, the B.1 value is used, so the total risk is only 10^{-6} (assuming that the initiation and propagation lives are independent). If a component has reached the life limit, it is also retired, whether or not it contains a crack. The drawback of this damage tolerance approach is therefore that still most components are retired prematurely, that is, in the uncracked state.

The inspection intervals, which are used for DT components, are determined from the safety limit (the crack propagation life). The first inspection is recommended to occur between 0.5 and 1.0 times the calculated safety limit. Subsequent inspections are recommended at between 0.5 and 1.0 of the time required for the maximum undetectable flaw to grow to the critical size. Note that for the calculation of the first inspection interval (from the safety limit), the maximum probable flaw size is used as initial crack size, whereas for subsequent inspections, the inspection limit of in-service NDI is used.

9.5 Probabilistic Design

As was discussed in Chaps. 6 and 7 , many engineering calculations are deterministic analyses, while considerable uncertainty exists in many of the input variables (e.g. usage, material properties). The uncertainty can be accounted for by applying large safety factors or by using conservative estimates of the input variables. However, as was demonstrated in Chap. 6, this leads to very conservative maintenance intervals and service lives.

Section 7.6 illustrated that performing a stochastic life assessment can aid in quantifying the uncertainty and associated probability of failure. The same principle can be applied in design calculations, which is then called probabilistic design. By quantifying or estimating the variations in future loading of the system and defining acceptable risk levels, the system can be designed to meet these requirements. This approach is commonly applied in civil engineering for large infrastructural assets (bridges, dams) and for nuclear power facilities. In both cases, the risks of failure are enormous (i.e. flooding, nuclear radiation damage) and the design requirements can only be set when the probabilities of failure can be compared to certain risk levels. These analyses are also known as probabilistic risk assessments (PRAs).

In these analyses, the probabilities of failure can only be assessed in a reliable way when the potential failure mechanisms and the (variations in) loads are known. The methods as discussed in Sect. 8.2, that is, FMECA and FTA, are

therefore generally the starting points of these analyses, but the failure models discussed in Chap. 4 are required to perform the actual quantitative analyses.

9.6 Summary

In this chapter, the design process of complex assets has been discussed, with a special focus on how knowledge on failure mechanisms can be utilized to design better performing and more reliable systems. The concepts of life cycle management and life cycle costs have been briefly explained. Then, the different aspects of design for maintenance were discussed, demonstrating in which ways the design can be modified to assure a sufficiently high reliability, maintainability and supportability of the system. To put the design process in a somewhat broader perspective, two important design philosophies are compared and finally the concept of probabilistic design was briefly discussed.

References

1. Blanchard, B.S., Fabrycky, W.J.: Systems Engineering and Analysis, 3rd edn. Prentice Hall International, London (1998)
2. Mulder, W., Basten, R.J.I., Dongen, L.A.M.v.: Set of Design Rules for Design for Maintenance. University of Twente, Enschede (2012)
3. Tinga, T., Wanhill, R.J.H., Hoeve, H.J.t.: Aerospace Lifing Methods, Specifically for Gas Turbines. National Aerospace Laboratory, Amsterdam (1998)
4. Nishida, S.: Failure Analysis in Engineering Applications. Butterworth-Heinemann, Oxford (1992)

Further Reading

1. Dhillon, B. S.: Life Cycle Costing for Engineers. CRC Press, Boca Raton (2009)

Index

T. Tinga, *Principles of Loads and Failure Mechanisms*,
Springer Series in Reliability Engineering,
DOI: 10.1007/978-1-4471-4917-0, © Springer-Verlag London 2013