Practices in Failure Analysis

Introduction

ANALYZING the inevitable failures that occur during testing, manufacturing, and service is an essential engineering process for continual improvement in product reliability. This article describes the general procedures, techniques, and precautions employed in the investigation and analysis of metallurgical failures that occur in service. The stages of investigation are discussed, and the various features of the more common causes of failure characteristics are described for fracture, corrosion, and wear failures. The mechanisms of fracture, corrosion, and wear failures are explained in more detail in other Sections of this Volume.

Practices in Failure Analysis

Stages of a Failure Analysis

Although the sequence is subject to variation depending on the nature of the failure and the availability of physical evidence or background information, there are stages that are common to all successful failure analyses. The combination of these stages comprises the total investigation and analysis. The following list includes many of the commonly used stages. The sequence in which these stages are used is not necessarily critical, and not all of the stages will or can be used in every failure analysis. However, a key principle is to not unnecessarily disrupt conditions that may require closer examination at a later date. Moreover, an additional constraint is to follow the Federal Rules of Evidence during investigation of a failure that may be destined for court.

The stages discussed in this article begin first with the preliminary steps of information gathering such as:

- Collection of background data and selection of samples
- Preliminary examination of the failed part (visual examination and record keeping)
- Nondestructive testing

These preliminary steps may then be followed by assessment of the damage and conditions leading to failure. These stages may differ depending on whether fracture, corrosion, and/or wear conditions are being investigated. In an analysis of a fracture, the following steps are described:

- Selection, identification, preservation, and/or cleaning of critical specimens
- Macroscopic examination and analysis (fracture surfaces, secondary cracks, and other surface phenomena)
- Microscopic examination and analysis of fracture surfaces
- Stress analysis to determine the actual stress state of the failed component
- Fracture mechanics
- Determination of the fracture mode

Following these topics on the analysis of fractures, separate sections also briefly describe factors and methods in the analysis of corrosion and wear failures.

In addition, investigations of a failure may utilize various techniques to characterize the condition of material. These include:

- Metallography or microstructural analysis
- Mechanical testing
Finally, the investigation concludes with a synthesis and interpretation of results. This step may actually require reiteration of previous steps or the introduction of new steps. Similar to design, failure analysis can be an iterative process of discovery and reexamination. Failure analysis can also be a multidisciplinary process and that may require consulting with experts in other disciplines throughout the investigation. Once all information has been assembled, then the final step is to synthesize all the evidence and formulate conclusions. This requires writing a report with follow-up recommendations on preventing future failures. The goal of every failure analyst is to determine not only the failure mechanism but also the root cause, which may be related to misuse, poor maintenance practices, or improper application, or related to the material properties, design, or manufacture of the product.

In cases that involve personal injury or will most likely involve legal pursuit of compensation from another company, care must be taken in preserving the scene and physical evidence. Accidental or deliberate destruction of evidence can result in diverting the legal liability of a failure to the person or company destroying the evidence, even though they may not have caused the original failure.

**Practices in Failure Analysis**

---

**Collection of Background Data and Samples**

The failure investigation should include gaining an acquaintance with all pertinent details relating to the failure, collecting the available information regarding the design, manufacture, processing, and service histories of the failed component or structure, and reconstructing, insofar as possible, the sequence of events leading to the failure. Collection of background data on the manufacturing and fabricating history of a component should begin with obtaining specifications and drawings and should encompass all the design aspects of the failed part as well as all manufacturing and fabrication details—maching, welding, heat treating, coating, quality-control records, and pertinent purchase specifications. Additional information on upfront planning of investigations is also described in the article “Organization of a Failure Investigation” in this Volume.

**Collecting Data and Samples**

On-Site Investigation. In the investigation of failures, it is also often desirable for the analyst to visit the scene, but for the analysis of some components it may be impractical or impossible for the failure analyst to visit the failure site. Under these circumstances, data and samples may be collected at the site by field engineers or by other personnel under the direction of the failure analyst. A field failure report sheet or checklist can be used to ensure that all pertinent information regarding the failure is recorded.

There also are situations where it is essential to perform failure analyses on the site. While it is recommended that examination be done in a laboratory, the requirements for on-site testing may involve the use of portable laboratories with metallographic equipment for grinding, mechanical polishing, and etching. Small specimens can be cut from a part on the site for preparation, examination, and photography immediately or upon return to a fully equipped laboratory. Photography is, of course, essential; it should be performed by the analyst or perhaps a professional photographer in the case of a large-scale accident scene. Other considerations for on-site examination at an accident scene are also discussed in more detail in the article “Modeling and Accident Reconstruction” in this Volume.

It is also frequently desirable to make acetate tape replicas or room-temperature-vulcanized (RTV) rubber replicas of fracture surfaces or of wear patterns of large parts during an on-site failure analysis. Several replicas should be made of the fracture-origin region using acetate tape softened in acetone, dried, then carefully stripped from the fracture surface. Upon return to the laboratory, the replicas may be gold coated and examined with a scanning electron microscope (SEM). Foreign particles removed from the fracture surface also may be analyzed.

The RTV rubber replica can be applied over a rather large area with less chance of missing a critical spot. A combination of acetate tape and RTV rubber replicas can assure the investigator of better coverage of the area in question. Room-temperature-vulcanized rubber does not provide the sensitivity of an acetate replica, and a setup time of several hours is required. However, the added area can be very important in an investigation.
The major components of the portable laboratory may include:

- A custom-made machine, plus auxiliary materials, for grinding and polishing small, mounted or unmounted metal specimens
- A right-angle head, electric drill motor with attachments and materials for grinding and polishing selected spots on large parts or assemblies. It is also used for driving the grinding and polishing machine described in the previous item
- A portable microscope, with camera attachment and film for use in photographing metallographic specimens
- Equipment and materials for mounting and etching specimens
- A handheld single-lens reflex 35-mm camera, with macro lenses and film
- A pocket-size magnifier, and a ruler or scale
- A hacksaw and blades for cutting specimens
- Portable hardness tester
- Acetate tape, acetone, and containers
- RTV rubber for replicas

Service History. The availability of a complete service history depends on how detailed and thorough the record keeping was prior to the failure. A complete service record greatly simplifies the assignment of the failure analyst. In collecting service histories, special attention should be given to environmental details such as normal and abnormal loading, accidental overloads, cyclic loads, temperature variations, temperature gradients, and operation in a corrosive environment. In most instances, however, complete service records are not available, forcing the analyst to work from fragmentary service information. When service data are sparse, the analyst must, to the best of his or her ability, deduce the service conditions. Much depends on the analyst's skill and judgment, because a misleading deduction can be more harmful than the absence of information.

Photographic Records. Photographs of the failed component or structure are oftentimes critical to an accurate analysis. A detail that appears almost inconsequential in a preliminary investigation may later be found to have serious consequences; thus, a complete, detailed photographic record of the scene and failed component can be essential. Photographs should be of professional quality, but this is not always possible. For the analyst who does his own photography, a single-lens reflex 35-mm or larger camera with a macro lens, extension bellows, and battery-flash unit is capable of producing excellent results. It may be desirable to supplement the 35-mm equipment with an instant camera and close-up lenses. Techniques and lighting are discussed in more detail in the article “Photography in Failure Analysis” in this Volume.

When accurate color rendition is required, the subject should be photographed with a color chart, which should be sent to the photographic studio for use as a guide in developing and printing. Some indication of size, such as a scale, coin, hand, and so forth, should be included in the photograph.

Samples should be selected judiciously before starting the examination, especially if the investigation is to be lengthy or involved. As with photographs, the analyst is responsible for ensuring that the samples will be suitable for the intended purpose and that they adequately represent the characteristics of the failure. It is advisable to look for additional evidence of damage beyond that which is immediately apparent. For failures involving large structures or key machinery, there is often a financially urgent need to remove the damaged structure or repair the machine to return to production. This is a valid reason to move evidence, but a reasonable attempt must be made to allow other parties, who may become involved in a potential legal case, to inspect the site. All concerned parties then can agree on the critical samples and the best way to remove them. If all parties are not available, care must be taken not to damage or alter critical elements to avoid spoiling evidence. Guidelines governing sample collection are covered in ASTM E 620, E 860, E 1020, and especially E 678. It is also recommended that samples be taken from other parts of the failed equipment as they may display supportive damage.

It is often necessary to compare failed components with similar components that did not fail to determine whether the failure was brought about by service conditions or was the result of an error in manufacture. For example, if a boiler tube fails and overheating is suspected to be the cause, and if investigation reveals a spheroidized structure in the boiler tube at the failure site (which may be indicative of overheating in service), then comparison with an unexposed tube will determine if the tubes were supplied in the spheroidized condition. As another example, in the case of a bolt failure it is desirable to examine the nuts and other associated parts that may have contributed to the failures. Also, in failures involving corrosion, stress-corrosion, or corrosion fatigue, a sample of the fluid that has been in contact with the metal, or of any deposits that have formed, will often be required for analysis.

Abnormal Conditions and Wreckage Analysis. In addition to developing a history of the failed part it is also advisable to determine if any abnormal conditions prevailed. Determine also whether events—such as an accident—occurred in
service that may have initiated the failure, or if any recent repairs or overhauls had been carried out and why. In addition, it is also necessary to inquire whether or not the failure was an isolated example or if others have occurred, either in the component under consideration or in another of a similar design. In the routine examination of a brittle fracture, it is important to know if, at the time of the accident or failure, the prevailing temperature was low, and/or if some measure of shock loading was involved. When dealing with failures of crankshafts or other shafts, it is generally desirable to ascertain the conditions of the bearings and whether any misalignment existed, either within the machine concerned or between the driving and driven components. In an analysis where multiple components and structures are involved, it is essential that the position of each piece be documented before any of the pieces are touched or moved. Such recording usually requires extensive photography, the preparation of suitable sketches, and the taking and tabulation of appropriate measurements of the pieces. Next, it may be necessary to take an inventory to determine if all of the pieces or fragments are present at the site of the accident. For example, an investigation of an aircraft accident involves the development of a considerable inventory, including listing the number of engines, flaps, landing gear, and the various parts of the fuselage and wings. It is essential to establish whether all the primary parts of the aircraft were aboard at the time that it crashed. Providing an inventory, although painstaking, is often invaluable. An experienced investigator determined the cause of a complex aircraft accident when he observed that a portion of one wing tip was missing from the main impact site. This fragment was subsequently located several miles back on the flight path of the aircraft. The fragment provided evidence of a fatigue failure and was the first component separated from the aircraft, thus accounting for the crash. The most common problem encountered in examining wreckage involves the establishment of the sequence of fractures to determine the origin of the initial failure. Usually, the direction of crack growth can be detected from marks on a fracture surface, such as V-shaped chevron marks. The typical sequence of fractures is shown in Fig. 1(b), where A and B represent fractures that intersect at about 90°. Here the sequence of fractures is clearly discernible from crack branching. Obviously, fracture A must have occurred prior to fracture B because the presence of fracture A served to arrest cracking at fracture B. This method of sequencing is called the *T-junction procedure* and is an important technique in wreckage analysis.

![Diagram of crack propagation and fracture origin](image)

**Fig. 1** General features to locate origin from crack paths (a) branching and (b) sequencing of cracking by the T-junction procedure, where fracture A precedes and arrests fracture B

Provided the fragments are not permitted to contact each other, it is also helpful to carefully fit together the fragments of broken components which, when assembled and photographed, may indicate the sequence in which fractures occurred. Figure 2 shows a lug that was part of a pin-joint assembly; failure occurred when the pin broke out of the lug. With the broken pieces of the lug fitted together, it is apparent from the deformation that fracture A must have preceded fractures B.
and C. However, parts will not fit well together because of plastic deformation that occurred before or during the fracture process.

Fig. 2  Fractured lug, part of a pin-joint assembly, showing sequence of fracture. Fracture A preceded fractures B and C.

Preliminary Examination

The failed part, including all its fragments, should be subjected to a thorough visual examination before any cleaning is undertaken. Often, soils and debris found on the part provide useful evidence in establishing the cause of failure or in determining a sequence of events leading to the failure. For example, traces of paint or corrosion found on a portion of a fracture surface may provide evidence that the crack was present in the surface for some time before complete fracture occurred. Such evidence should be recorded photographically.

Visual Inspection. The preliminary examination should begin with unaided visual inspection. The unaided eye has exceptional depth of focus, the ability to examine large areas rapidly and to detect changes of color and texture. Some of these advantages are lost when any optical or electron-optical device is used. Particular attention should be given to the surfaces of fractures and to the paths of cracks. The significance of any indications of abnormal conditions or abuse in service should be observed and assessed, and a general assessment of the basic design and workmanship of the part should also be made. Each important feature, including dimensions, should be recorded, either in writing or by sketches or photographs.

It cannot be emphasized too strongly that the examination should be performed as carefully as possible, because clues to the cause of breakdown often are present, but may be missed if the observer is not vigilant. Inspection of the topographic features of the failed component should start with an unaided visual examination and proceed to higher and higher magnification. A magnifying glass followed by a low-power microscope is an invaluable aid in detection of small details of the failed part.

Examination and Photography of the Damaged/Failed Part or Sample. The next step should be preliminary examination and general photography of the entire part and damaged or failed regions. Where fractures are involved, the entire fractured part, including broken pieces, should be examined and photographed to record their size and condition and to show how the fracture is related to the components. This should be followed by careful examination of the fracture. The examination should begin with the use of direct lighting and proceed at various angles of oblique lighting to delineate and emphasize fracture characteristics. This should also assist in determining which areas of the fracture are of prime interest and which magnifications will be possible (for a given picture size) to bring out fine details. When this evaluation has been completed, it is appropriate to proceed with photography of the fracture, recording what each photograph shows, its magnification, and how it relates to the other photographs. For information on photographic equipment, materials, and techniques, see the article “Photography in Failure Analysis” in this Volume and the article entitled “Photography of Fractured Parts and Fracture Surfaces,” on pages 78 to 90 in Fractography, Volume 12 of the Metals Handbook, 9th edition (now Volume 12 of the ASM Handbook).

Nondestructive Inspection
Although often used as quality-control tools, several nondestructive tests are useful in failure investigation and analysis: magnetic-particle inspection of ferrous metals, liquid-penetrant inspection, ultrasonic inspection, and sometimes eddy-current inspection. All these tests are used to detect surface cracks and discontinuities. Radiography is used mainly for internal examination. A photographic record of the results of nondestructive inspection is a necessary part of record keeping in the investigation.

Magnetic-particle inspection utilizes magnetic fields to locate surface and subsurface discontinuities in ferromagnetic materials. When the material or part to be tested is magnetized, discontinuities that generally lie transverse to the direction of the magnetic field will cause a leakage field to be formed at and above the surface of the part. This leakage field, and therefore the presence of the discontinuity, is detected by means of fine ferromagnetic particles applied over the surface, some of which are gathered and held by the leakage field. The magnetically held collection of particles forms an outline of the discontinuity and indicates its size, shape, and extent. Frequently, a fluorescent material is combined with the particles so that discontinuities can be detected visually under ultraviolet light. This method reveals surface cracks that are not visible to the naked eye.

Liquid-penetrant inspection is used to detect surface flaws in materials. It is used mainly, but not exclusively, with nonmagnetic materials, on which magnetic-particle inspection cannot be used. This technique involves the spreading of a liquid penetrant on the sample. Liquid penetrants can seep into small cracks and flaws (as fine as 1 μm) in the surface of the sample by capillary action. The excess liquid is wiped from the surface, and a developer is applied that causes the liquid to be drawn from the cracks or flaws that are open at the surface. The liquid itself is usually a very bright color or contains fluorescent particles that, under ultraviolet light, cause discontinuities in the material to stand out.

The main advantages of the liquid-penetrant method are its ability to be used on nonmagnetic materials, its low cost, its portability, and the ease with which results can be interpreted. The principal limitations of the liquid-penetrant method are:

- Discontinuities must be open to the surface.
- Testpieces must be cleaned before and after testing because the liquid penetrant may corrode the metal.
- Surface films may prevent detection of discontinuities.
- Penetrant may be a source of contamination that masks results in subsequent chemical analysis of fracture surfaces.
- The process is generally not suited to inspection of low-density powder-metallurgy parts or other porous materials.

Ultrasonic inspection methods depend on sound waves of very high frequency being transmitted through metal and reflected at any boundary such as a metal/air boundary at the surface of the metal, or a metal/crack boundary at a discontinuity within the part or component. High-frequency sound waves can detect small irregularities, but they are easily absorbed, particularly by coarse-grained materials.

The application of ultrasonic testing is limited in failure analysis because accurate interpretations depend on reference standards to isolate the variables. In some instances, ultrasonic testing has proved to be a useful tool in failure analysis, particularly in the investigation of large castings and forgings. Cracks, laminations, shrinkage cavities, bursts, flakes, pores, disbonds, and other discontinuities that produce reflective interfaces can be easily detected. Inclusions and other inhomogeneities can also be detected by causing partial reflection or scattering of the ultrasonic waves or by producing some other detectable effect on the ultrasonic waves.

The disadvantages of ultrasonic inspection include:

- Manual operation requires careful attention by experienced technicians.
- Extensive technical knowledge is required for the development of inspection procedures.
- Parts that are rough, irregular in shape, very small or thin, or not homogeneous are difficult to inspect.
- Discontinuities that are present in a shallow layer immediately beneath the surface may not be detectable.
- Couplants are needed to provide effective transfer of ultrasonic wave energy between transducers and parts being inspected.
- Reference standards are needed, both for calibrating the equipment and for characterizing flaws.

Radiography uses x-rays or gamma rays, which are directed through the sample to a photographic film. After the film has been developed, it can be examined by placing it in front of a light source. The intensity of the light passing through the film will be proportional to the density of the sample and the path length of the radiation. Thus, lighter areas on the plate correspond to the denser areas of the sample, whereas darker areas indicate a crack or defect running in the direction of the incident beam.

The main advantages of radiography are its ability to detect internal discontinuities and to provide permanent photographic records. However, certain types of flaws are difficult to detect by radiography. Laminar defects, such as cracks, present problems unless they are essentially parallel to the radiation beam. Tight, meandering cracks in thick
sections usually cannot be detected even when properly oriented. Minute discontinuities, such as inclusions in wrought material, flakes, microporosity, and microfissures, cannot be detected unless they are sufficiently segregated to yield a detectable gross effect. Laminations normally are not detectable by radiography because of their unfavorable orientation, usually parallel to the surface. Laminations seldom yield differences in absorption that enable laminated areas to be distinguished from lamination-free areas.

Eddy-current inspection can be used on all materials that conduct electricity. If a coil conducting an alternating current is placed around or near the surface of the sample, it will set up eddy currents within the material by electromagnetic induction. These eddy currents affect the impedance in the exciting coil or any other pickup coil that is nearby. Cracks or flaws within the sample will cause distortions in the eddy current, which in turn cause distortion in the impedance of the coil. The resulting change in impedance can be detected by attaching the appropriate electrical circuits and a meter. Flaws or cracks will show up as some deflection or fluctuation on the meter.

The advantages of electromagnetic inspection are:

- Both surface and subsurface defects are detectable.
- No special operator skills are required.
- The process is adaptable to continuous monitoring.
- The process may be substantially automated and is capable of high speeds.
- No probe contact is needed.

Limitations of electromagnetic inspection include:

- Depth of penetration is shallow.
- Materials to be inspected must be electrically conductive.
- Indications are influenced by more than one variable.
- Reference standards are required.

Residual Stress Analysis. X-ray diffraction is the most common method for direct, nondestructive measurement of residual (internal) stresses in metals. Stresses are determined by measuring the submicroscopic distortion of crystalline lattice structures by tensile or compressive residual stresses. However, it should be pointed out that measurement of residual stresses near fractures or cracks may be erroneous because the residual stresses have already been relieved by the fracture and cracks. Testing of undamaged similar, or exemplar, parts is frequently used as the only alternative in order to understand the residual stress system in the failed part prior to failure. More information is in the article “X-Ray Diffraction Residual Stress Measurement in Failure Analysis” in this Volume.

Acoustic-emission inspection detects and analyzes minute acoustic-emission signals generated by discontinuities in materials under applied stress. Proper analysis of these signals can provide information concerning the location and structural significance of the detected discontinuities.

Some of the significant applications of acoustic-emission inspection are:

- Continuous surveillance of pressure vessels and nuclear primary-pressure boundaries for the detection and location of active flaws
- Detection of incipient fatigue fracture in aircraft structures
- Monitoring of both fusion and resistance weldments during welding and cooling
- Determination of the onset of stress-corrosion cracking (SCC) and hydrogen damage in susceptible structures
- Use as a study tool for the investigation of fracture mechanisms and of behavior of materials
- Periodic inspection of tanks and aerial-device booms made of composite materials

This type of data may be useful background information in a failure analysis, or the technique might be used in evaluation of stress effects. Sources of acoustic emission that generate stress waves in material include local dynamic movements, such as the initiation and propagation of cracks, twinning, slip, sudden reorientation of grain boundaries, and bubble formation during boiling. This energy may originate from stored elastic energy, as in crack propagation, or from stored chemical-free energy, as in phase transformation.

Experimental stress analysis can be done by several methods, all of which may be valuable in determining machine loads and component stresses that can cause failures. Stress coating can be used effectively for locating small areas of high strains, determining the directions of the principal strains, and measuring the approximate magnitude of tensile and compressive strains. Gages can then be placed at the high-strain areas and in the principal-strain directions to measure the strain accurately on gage lengths 0.5 to 150 mm (0.02 to 6 in.). Although there are many mechanical, optical, and electrical devices capable of accurate strain measurements, the bonded electrical-resistance strain gage has become the standard tool for general laboratory and field use.
Photoelastic coatings have also been used for laboratory stress measurements. For this technique, a birefringent coating of controlled thickness is bonded to the testpiece with a reflective cement. Optical analysis is similar to conventional analysis but requires special equipment. The analysis may be recorded on color film with single-frame or movie camera.

**Practices in Failure Analysis**

## Fractures

Establishing the origin of a fracture is essential in failure analysis, and the location of the origin determines which measures should be taken to prevent a repetition of the fracture. The fracture-surface characteristics that show the direction of crack propagation (and conversely, the direction toward the origin) include features such as chevron marks, crack branching, and river patterns. Features that help identify the crack origin include concentric fibrous marks, radial marks, and beach marks. By a study of these features, crack progress can be traced back to the point of origin, and then it can be ascertained whether the crack was initiated by an inclusion, a porous region, a segregated phase, a corrosion pit, a machined notch, a forging lap, a nick, a mar, or another type of discontinuity, or was simply the result of overloading.

However, time employed in ascertaining all the circumstances of a failure is extremely important. When a broken component is received for examination, the investigator is sometimes inclined to prepare specimens immediately without devising an investigative procedure. To proceed without forethought may destroy important evidence and waste time. Some of the questions that should be raised concerning the nature, history, functions, and properties of the fractured part, and the manner in which it interacts with other parts, are:

- **Loading.** Were the nature, rate, and magnitude of the applied load correctly anticipated in the design of the part? Were repeated or cyclic loadings involved? What was the direction of the principal stress relative to the shape of the part? Were residual stresses present to an undesirable degree?
- **Material.** Was the recommended alloy used? Were its mechanical properties at the level expected? Were surface or internal discontinuities present that could have contributed to failure? Did the microstructure conform to that prescribed?
- **Shape.** Did the part comply with all pertinent dimensional requirements of the specification? Did the part have sufficient section thickness to prevent local overloading? Were fillets formed with sufficiently large radii? Were there adequate clearances between interacting parts? Were any of the contours deformed during service? Was there evidence of mechanical surface damage?
- **Environment.** Was the part exposed to a corrosive environment or to excessively high or low temperatures? Was the surface of the part suitably protected? Were the properties of the part altered by the exposure? Was there interaction (for example, galvanic) between the material of the part and that of adjacent components?

Examination of a fracture begins with visual scrutiny, which establishes:

- Whether there is gross evidence of mechanical abuse
- Whether there are indications of excessive corrosion
- Whether the part is deformed
- Whether there are obvious secondary fractures
- Whether the origin of the crack can be readily identified
- Whether the direction of crack propagation can be easily recognized

Often it is helpful to have an undamaged part of the same design as the fractured part available during this portion of the examination. The findings of this scrutiny will permit many deductions to be drawn concerning the service conditions existing at and prior to the time of fracture. These findings can then be extended by an examination of the fracture surface at low magnification with a stereomicroscope and then at high magnification by electron microscopy, metallography (occasionally), or some combination of these examination techniques. A survey at low magnification is important for identification of those areas that need further inspection at high magnification. The salient features are recorded in fractographs of appropriate magnification for report purposes and for future reference should subsequent handling or sectioning destroy evidence needed for failure analysis.

Fracture of a part in service is often intimately associated with the type of environment to which the part was exposed. Active chemical environments include water, salt air, salt water, acid solutions, alkaline solutions, molten metal, and even solid metal. Thermal environments that affect metal properties and fracture include exposure to low (cryogenic, for example) and high temperatures.
Selection and Preservation of Fracture Surfaces

The proper selection, preservation, and cleaning of fracture surfaces is vital to prevent important evidence from being destroyed or obscured. Surfaces of fractures may suffer either mechanical or chemical damage. Mechanical damage may arise from several sources, including the striking of the surface of the fracture by other objects. This can occur during actual fracture in service or when removing or transporting a fractured part for analysis.

Usually, the surface of a fracture can be protected during shipment by a cloth or cotton covering, but this may remove some loosely adhering material that might contain the primary clue to the cause of the fracture. Touching or rubbing the surface of a fracture with the fingers should definitely be avoided. Also, no attempt should be made to fit together the sections of a fractured part by placing them in contact. This generally accomplishes nothing and almost always causes damage to the fracture surface. The use of corrosion-inhibiting paper to package samples should be considered.

Chemical (corrosion) damage to a fracture specimen can be prevented in several ways. For instance, because the identification of foreign material present on a fracture surface may be important in the overall determination of the cause of the fracture, many laboratories prefer not to use corrosion-preventive coatings on a fracture specimen. When possible, it is best to dry the fracture specimen, preferably using a jet of dry, compressed air (which will also blow extraneous foreign material from the surface), and then to place it in a dessicator or pack it with a suitable dessicant. However, clean, fresh fracture surfaces should be coated when they cannot be protected from the elements. After failure of large structures several days may be required to remove critical specimens, and so coating the fracture surfaces would be the prudent decision.

Cleaning of fractured surfaces should be avoided in general, but must be done for SEM examination and often to reveal macroscopic fractographic features. Cleaning should proceed in stages using the least aggressive procedure first, then proceeding to more aggressive procedures if needed. Washing the fracture surface with water should especially be avoided. However, specimens contaminated with seawater or with fire-extinguishing fluids require thorough washing, usually with water, followed by a rinse with acetone or alcohol before storage in a dessicator or coating with a dessicant. Sometimes cleaning may also be required for removal of obliterating debris and dirt, or to prepare the fracture surface for SEM examination. Other acceptable cleaning procedures include use of a dry-air blast or of a soft-hair artist's brush; treating with inorganic solvents, either by immersion or by jet; treating with mild acid or alkaline solutions (depending on the metal) that will attack deposits but to which the base metal is essentially inert; ultrasonic cleaning; and application and stripping of plastic replicas.

Cleaning with cellulose acetate tape is one of the most widely used methods, particularly when the surface of a fracture has been affected by corrosion. A strip of acetate about 0.1 mm (0.005 in.) thick and of suitable size is softened by immersion in acetone and placed on the fracture surface. The initial strip is backed by a piece of unsoftened acetate, and then the replica is pressed hard onto the surface of the fracture using a finger. The drying time will depend on the extent to which the replicating material was softened, and this in turn will be governed by the texture of the surface of the fracture. Drying times of not less than 15 to 30 min are recommended. The dry replica is lifted from the fracture, using a scalpel or tweezers. The replicating procedure must be repeated several times if the fracture is badly contaminated. When a clean and uncontaminated replica is obtained, the process is complete. An advantage of this method is that the debris removed from the fracture is preserved for any subsequent examination that may be necessary for identification by x-ray or electron diffraction techniques. To be complete, the analyst should filter solvents used for cleaning to recapture insoluble particulates.

Sectioning. Because examination tools, including hardness testers and optical and electron microscopes, are limited as to the size of specimen they can accept, it is often necessary to remove from a failed component a fracture-containing portion or section that is of a size convenient to handle and examine. This is a destructive process and may spoil evidence in potential litigation cases.

Before cutting or sectioning, the fracture area should be carefully protected. All cutting should be done so that surfaces of fractures and areas adjacent to them are not damaged or altered; this includes keeping the fracture surface dry, whenever possible. For large parts, the common method of removing specimens is by flame cutting. Cutting must be done at a sufficient distance from the fracture site so that the microstructure of the metal underlying the surface of the fracture is not altered by the heat of the flame, and so that none of the molten metal from flame cutting is deposited on the surface of the fracture.

Heat from any source can affect metal properties and microstructures during cutting. Therefore, dry abrasive cutoff wheels should never be used near critical surfaces that will be examined microscopically. Therefore, sectioning should be performed with jewelers' saws, precision diamond-edged, thin cutoff wheels, hacksaws, band saws, or soft abrasive cutoff wheels flooded with water-based soluble oil solution to keep metal surfaces cool and corrosion-free. Dry cutting with an air-driven abrasive disk may also be used with care to remove small specimens from large parts if kept cool, along with coating the fracture surface for protection.

Secondary Cracks. When the primary fracture has been damaged or corroded to such a degree that most of the information relevant to the cause of the failure is obliterated, it is desirable to open any secondary cracks to expose their fracture surfaces for examination and study. These cracks may provide more information than the primary fracture.
In opening cracks for examination, care must be exercised to prevent damage, primarily mechanical, to the surface of the fracture. This can usually be accomplished if opening is done in such a way that the two surfaces of the fracture are moved in opposite directions, normal to the fracture plane. Generally, a saw cut can be made from the back of the fractured part to a point near the tip of the crack, using extreme care to avoid actually reaching the tip of the crack. This saw cut will reduce the amount of solid metal that must be broken. The final breaking of the specimen can be done in several ways:

- By clamping the two sides of the fractured part in a tensile-testing machine, if the shape permits, and pulling
- By placing the specimen in a vise and bending one half away from the other by striking it with a hammer in a manner that will avoid damage to the surfaces of the crack
- By gripping the halves of the fracture in pliers or vise grips and bending or pulling them apart

Cooling the part with liquid nitrogen often reduces the force and plastic deformation necessary to fracture the part. Fortunately, there is little confusion during subsequent examination as to which part of the fracture surface was obtained in opening the crack.

It is recommended that both the crack separation and the visible crack length be measured prior to opening. The analyst may have to use dye penetrant or other nondestructive evaluation technique to actually see the length of a tightly closed crack. Often, the amount of strain that occurred in the specimen can be determined from a measurement of the separation between the adjacent halves of a fracture. This should be done before preparation for opening a secondary crack has begun. The lengths of cracks may also be important for analyses of fatigue fractures or for consideration for the application of fracture mechanics.

**Macroscopic Examination of Fracture Surfaces**

One very important part of any failure analysis is the macroscopic examination of fracture surfaces. Performed at magnifications from 1 to 50 or 100×, it may be conducted by the unaided eye, a hand lens or magnifier, a low-power stereoscopic microscope, or a SEM. Macroscopic photography of up to 20× magnification requires a high-quality camera and special lenses; alternatively, a large magnifying glass may be used to enlarge a specific area in the photo, such as a crack or other small detail. A metallographic microscope with macroobjectives and lights may be used for somewhat higher magnifications. However, depth of field becomes extremely limited with light optics. For much greater depth of field, a SEM may be used for low-magnification photography as well as higher-magnification work. Stereo or three-dimensional photographs may also be made to reveal the topographic features of a fracture or other surface.

Frequently, a specimen may be too large or too heavy for the stage of the metallograph or the chamber of a SEM, and cutting or sectioning the specimen may be difficult or not allowed because of legal limitations at the time. In these instances, excellent results can be achieved by examining and—where appropriate—photographing replicas made by the method for cleaning fractures (see discussion under “Cleaning”). These replicas can be coated with a thin layer (about 20 nm, or 2 × 10−8 m, thick) of vacuum-deposited gold or aluminum to improve their reflectivity, or they may be shadowed at an angle to increase the contrast of fine detail. The replicas may be examined by incident-light or transmitted-light microscopy. Because they are electrically conductive, the coated replicas may also be examined with a SEM.

The amount of information that can be obtained from examination of a fracture surface at low-power magnification is extensive. A careful scan of the exterior surface of the part in the area adjacent to the fracture is required to determine whether specific stress raisers are present of a type that could have initiated the fracture. If any marks possess sharp reentrant angles, they constitute sites of stress concentration, a frequent cause of crack initiation. In this situation, the obvious remedy is more careful handling procedures and better inspection. Tool marks are another source of stress concentration. A fillet that has too small a radius, even though the surface of the fillet may be an excellent example of high-grade machining, is a recognized initiation site for fatigue cracks. Sharp-bottomed tool marks can initiate fatigue fractures even though the general contour of the area has a generous radius.

The shape, size, and cross section of a specimen or structural component can have a large effect on both the macroscopic and the microscopic appearance of the fracture surface, especially when pronounced stress raisers are present. Holes, corners, notches, machining marks, and, most of all, preexisting cracks actively influence fracture appearance. Pronounced stress raisers are more likely to be contained in a large part than in a small part, because large parts have greater volume and surface area.

The orientation of the fracture surfaces must be consistent with the proposed mode of failure and the known loads on the failed part. Failure in monotonic tension produces a flat (square) fracture normal (perpendicular) to the maximum tensile stress and frequently a slant (shear) fracture at about 45°. This 45° slant fracture is often called a “shear lip.” Many fractures are flat at the center, but surrounded by a “picture frame” of slant fracture. An example of this behavior is to be found in the familiar cup-and-cone fracture of a round tensile test bar.
In thin sheets, tube walls, and small diameter rods, slant shear fracture may occur because through-the-thickness stresses are minimized; that is, even though there may be a plane-strain condition, there may be minimal triaxial (hydrostatic) tensile stresses.

Macroscopic examination can usually determine the direction of crack growth and hence the origin of failure. With brittle, flat fractures, determination depends largely on the fracture surface showing a radial fanlike pattern of the type shown in Fig. 3. Cracks propagate parallel to shear lips if they are present. Where fracture surfaces show both flat and slant surfaces, this can be the terminal end of a fast-moving brittle fracture where the crack speed has slowed significantly. Crack extension can relax the stress so that final fracture occurs by slant shear fracture. Conversely, if a fracture has begun at a free surface, the fracture-origin area is usually characterized by a total absence of slant fracture or shear lip.

![Fig. 3 Surface of a fatigue fracture in a 4330V steel part. Chevron marks point to origin of fatigue in lower left corner. Arrows identify shear rupture along the periphery.](image)

Low-power examinations of fracture surfaces often reveal regions having a texture different from the region of final fracture. Fatigue, stress-corrosion, and hydrogen embrittlement fractures may also show these differences because the final failure is due to overload after the cross section is reduced by one of the aforementioned crack-initiation modes. Figure 4(a) shows the fracture surface of a steel tube and is an excellent example of the type of information that can be obtained by macroscopic examination. The V-shaped chevron marks and fanlike marks clearly indicate that the fracture origin is at the point marked by the arrow. This region, unlike the rest of the fracture, has no shear lip. The flat fracture surface suggests that the stress causing the failure was tension parallel to the length of the tube. The origin of the fracture as seen at higher magnification in Fig. 4(b) shows several small fracture origins having a texture different from that of the remainder of the fracture surface.
Microscopic Examination of Fracture Surfaces

Microscopic examination of fracture surfaces is typically done with a SEM. A SEM has the advantage over light microscopy because of the large depth of field and very high magnifications attainable, typically 5000 to 10,000×. In addition, SEMs are often equipped with microanalytical capabilities, for example, energy-dispersive x-ray spectrosopes (EDS). Chemical analysis can be helpful in confirming the chemistry of microstructural features that may be confused with fracture features.

The primary limitation of SEM analysis is sample size. A SEM analysis must be conducted in a vacuum so the sample must be put into a chamber that typically holds a sample less than 20 cm (8 in.) in diameter. Although there are some fracture surface features that are commonly associated with particular failure modes, the novice failure analyst must be very careful in fractographic analyses. Some of the more classic examples of fracture surface topography that indicates a fracture mode are:

- Dimpled rupture typical of overstress failures of ductile metals and alloys
- Cleavage facets, typical of transgranular brittle fracture of body-centered cubic and hexagonal close-packed metals and alloys
- Brittle intergranular fracture typical of temper-embrittled steel, where fracture is due to segregation of an embrittling species to grain boundaries (such as oxygen in iron or nickel), due to intergranular stress-corrosion cracking, or due to hydrogen embrittlement
Stage II striations, typical of some (not all) fatigue failures

Stress Analysis

It is sometimes quite apparent that an excessively high load or stress level was the direct cause or contributed significantly to the failure. Even so, an accurate stress analysis of the magnitude and type (axial, torsion, bending) of stress is required to substantiate the role of stress. In other failure analyses, the analyst may have strong evidence that the cause of a failure is related to excessively high static stresses (or cyclic stresses in the case of fatigue). In these cases, an analysis of the stress during normal operation (or abnormal operation if identified) must be conducted. Analytical, closed-form calculations based on engineering mechanics are often used by designers to predict stress levels in the early design stages. This method of using known “machine design and structural” formulas to predict the stress under a given load is also helpful to the failure analyst, especially in cases where this step may not have been used in the original design of the part. It is not uncommon to find products that have no record of any stress analysis in the development of the product. Even if such calculations were made, the failure analyst may not have access to them. The analyst must answer the questions “Was the component sized properly by the design stress analysis? Did the material have the properties assumed in the design? Did the part fail in a manner consistent with that assumed in design, or did it fail in a way not anticipated in the original design?”

In cases where unusual or abnormal loading is suspected, direct calculation of stresses will fall short and predict incorrect stress levels. In these cases, experimental stress analysis is used for determining machine loads and component stresses. This technique normally involves attachment of strain gages to similar parts in critical areas or typical areas where the failure has occurred. The strain gages are connected to a monitoring device either directly with wires or indirectly by radio signals for monitoring moving or rotating parts. In this way, the actual dynamic stresses can be determined. For products with very complex shapes and high thermal gradients, a finite-element analysis (FEA) may be required to estimate the level of stress that most likely existed in the failed component. These analyses can stand alone or can be used to help select critical locations for strain-gage attachment. Finite-element analyses can be time consuming and expensive, but they are necessary for an accurate assessment of stress levels in areas of complex geometry of some components. This type of analysis is almost essential for determining stresses caused by thermal gradients such as those found in welding. Overload failures are often a result of improper design or improper operation. A design analysis is essential in determining which of these is the root cause. Sometimes improper design is a result of incorrect information passed to the designer. In these cases, a failure is the only indication that the wrong input was used for the design. This is also true for fatigue failures. For proper design of rotating or moving parts, a detailed stress analysis is essential. It is much more difficult to predict dynamic stresses than static stresses.

Fracture Modes

Because the initial steps in failure analysis of a fracture involve visual and macroscopic observation, the first impressions should be based on obvious visual evidence. The simplest and most important observations relate to deformation: Was the metal obviously deformed? If it was deformed prior to fracture, yielding and fracture have occurred due to one or more gross overloads. It is predominantly a ductile fracture or a very high-stress, low-cycle fatigue fracture, as can be demonstrated by repeated manual bending of a paper clip or wire coat hanger. The deformation is directly related to the type of stress causing fracture: tension (stretched), bending (bent), torsion (twisted), or compression (shortened or buckled), or a combination of these stress types. The absence of gross deformation of the failed part indicates that the fracture is predominantly brittle. A brittle fracture should not be confused with brittle material. The shape or geometry of the part made from a ductile metal can result in an overload failure with little overall shape change, or a failure mechanism can operate to start and grow a crack such as a fatigue crack or stress-corrosion crack. When such a crack grows to the point that the remaining cross-sectional area of the part is overloaded by the normal loads, the final overload failure has little macroscale deformation associated with it. Thus, on a macroscale, the failure of a ductile metal can appear brittle. Of course, overload failures of brittle material always appear brittle on a macroscale.

It is usually more difficult to analyze a brittle fracture because there are a large number of possible mechanisms that can cause fracture with little or no obvious deformation. For single overload fractures, these include such factors as stress concentrations, low temperatures, high rates of loading, high metal strength and hardness, SSC, hydrogen embrittlement, temper embrittlement, large section size, and others. For fatigue fractures, causative factors can include stress concentrations, tensile residual stresses, large stress amplitudes, large numbers of load applications, corrosive environments, high temperatures, low metal strength and hardness, wear, and others. From this discussion, it should become clear that proper failure analysis is not simple but can become exceedingly complex, requiring considerable thought, examination, questioning, and reference to other sources of information in the
literature. However, identifying the failure mode is the key step in a failure analysis, and it is the essential part of determining the root cause.

Ductile Fracture. Overload fractures of many metals and alloys occur by ductile fracture. Overloading in tension is perhaps the least complex of the overload fractures, although essentially the same processes operate in bending and torsion as well as under the complex states of stress that may have produced a given service failure.

The classic example of ductile failure is a tensile test. In this fracture process, considerable elongation, that is, deformation, takes place before the geometric instability, necking, begins. Even after the deformation is localized at the neck, significant deformation occurs at the neck before the fracture process begins. After the neck forms, the curvature of the neck creates a region of tensile hydrostatic (or “triaxial”) stress. This leads to initiation of an internal crack near the center of the necked region. In commercial grade alloys, discontinuities such as inclusions or second-phase particles are sources of early void formation by separation of the matrix and the particle. Some of these voids coalesce to develop a crack, which is perpendicular to the tensile axis. The crack spreads until the state of stress, ductility of the metal, and flow condition reach a condition that favors a shear displacement. The crack path then shifts to a maximum shear plane, which is at an angle to the tensile axis (close to 45° in cylindrical specimens). Sometimes this shear lip forms only on one side of the initial flat crack. When this occurs, the resulting fracture surface has a macroscopic appearance known as cup-and-cone. For brittle materials, the majority of the fracture surface is perpendicular to the tensile axis with little or no fracture surface lying on a plane of shear.

Ductile failures in biaxially loaded sheet and plate structures often consist entirely of a shear lip. Pipe and pressure vessels are examples of biaxially stressed components. Often, failures in these components may first appear brittle with limited ductility; however, close inspection usually reveals some general thickness reduction but no necking at the fracture surface.

High-magnification examination of ductile fracture surfaces usually reveals dimples, which tend to be equiaxed when fractures occur under tensile load. Slant fractures or ductile fracture on planes of high shear stress generate elongated dimples. Ductile fractures (i.e., those with macroscopic deformation) are usually transgranular.

Brittle Fracture. There are two general types of brittle fracture caused by a single overload: transgranular cleavage and intergranular separation. Each has distinct features that make identification relatively simple. Transgranular cleavage can occur in body-centered cubic metals and their alloys (for example, ferritic steels, iron, tungsten, molybdenum, and chromium) and some hexagonal close-packed metals (for example, zinc, magnesium, and beryllium). Face-centered cubic metals and alloys (such as aluminum and austenitic stainless steels) are usually regarded as immune from this fracture mechanism. (See the article “Mechanisms and Appearances of Ductile and Brittle Fracture”).

Iron and low-carbon steels show a ductile-to-brittle transition with decreasing temperature that arises from a strong dependence of the yield stress on temperature. Brittle fracture of normally ductile metals depends on several physical factors, including specimen shape and size, temperature, and strain rate. Thus, a component or structure that has given satisfactory service may fracture unexpectedly; the catastrophic brittle fracture of ships in heavy seas and the failure of bridges on unusually cold days are examples. Metallurgical changes, especially strain aging, may cause the brittle fracture of such items as crane hooks and chain links after long periods of satisfactory operation.

Cleavage fracture is not difficult to diagnose because the fracture path is by definition crystallographic. In polycrystalline specimens, this often produces a pattern of brightly reflecting crystal facets, and such fractures are often described as crystalline (perhaps improperly as patterns are by definition crystalline). The general plane of fracture is approximately normal (perpendicular) to the axis of maximum tensile stress, and a shear lip is often present as a “picture frame” around the fracture. The local absence of a shear lip or slant fracture suggests a possible location for fracture initiation, for shear lips form during the final stages of the fracture process.

The fractography of cleavage fracture in low-carbon steels, iron, and other single-phase, body-centered cubic metals and alloys is fairly well established. Polycrystalline specimens contain numerous fan-shaped cleavage plateaus. The most characteristic feature of these plateaus is the presence of a pattern of river marks, which consist of cleavage steps or tear ridges and indicate the local direction of crack growth. The rule is that, if the tributaries of the “river lines” are regarded as flowing into the main stream, then the direction of crack growth is downstream. This is in contrast to macroscopic chevron marks, where the direction of crack growth, using the river analogy, would be upstream (see the article “Fracture Appearance and Mechanisms of Deformation and Fracture”).

Other fractographic features that may be observed include the presence of cleavage on conjugate planes, tear ridges, ductile tears joining cleavage planes at different levels, and tongues, which result from fracture in mechanical twins formed ahead of the advancing crack. Cleavage fracture in pearlitic and martensitic steels is less easily interpreted because microstructure tends to modify the fracture surface. In fact, cleavage fracture surfaces of pearlitic steel have characteristics similar to fatigue striations, so one must be careful not to confuse the fracture mode.

Intergranular fracture can usually be recognized, but determining the primary cause of the fracture may be difficult. Fractographic and microscopic examination can readily identify the presence of second-phase particles at grain boundaries. Unfortunately, the segregation of a layer a few atoms thick of some element or compound that produces intergranular fracture often cannot be detected by fractography. Auger analysis and sometimes EDS are useful for very
thin layers. Some causes of intergranular brittle fracture are given below, but the list is not exhaustive. It does, however, indicate some of the possibilities that need to be considered, and either eliminated or confirmed, as contributing to the fracture:

- The presence at a grain boundary of a large area of second-phase particles (such as carbides in Fe-Ni-Cr alloys or MnS particles in an overheated steel)
- Segregation of a specific element or compound to a grain boundary where a layer a few atoms thick is sufficient to cause embrittlement (embrittlement caused by the presence of oxygen in high-purity iron, oxygen in nickel, or antimony in copper and temper embrittlement of certain steels are examples of intergranular embrittlement where detection of a second phase at grain boundaries is difficult)

The conditions under which a progressively growing crack may follow an intergranular path before final fracture occurs include SCC, embrittlement by liquid metals, hydrogen embrittlement, and creep and stress-rupture failures. These failure modes are discussed in more detail elsewhere in this Volume. In addition, the article “Intergranular Fracture” in this Volume discusses various causes.

Fracture Mechanics Applied to Failure Analysis

The application of fracture mechanics is often pertinent to the investigation of failures, as well as to the formulation of preventive measures. In general, there are two types of conditions that may lead to structural failure:

- Net-section instability where the overall structural cross section can no longer support the applied load
- The critical flaw size \( a_c \) is exceeded by some preexisting discontinuity or when subcritical cracking mechanisms (for example, fatigue, SCC, creep) reach the critical crack size

Failures due to net-section instability typically occur when a damage process such as corrosion or wear reduces the thickness of a structural section. This type of failure can be evaluated by traditional stress analysis or FEA, which are effective methods in evaluating the effects of loading and geometric conditions on the distribution of stress and strain in a body or structural system. However, stress analyses by traditional methods or FEA do not easily account for crack propagation from preexisting cracks or sharp discontinuities in the material. When a preexisting crack or discontinuity is present, the concentration of stresses at the crack tip becomes asymptotic (infinite) when using the conventional theory of elasticity. In this regard, fracture mechanics is a useful tool, because it is a method that quantifies stresses at a crack tip in terms of a stress-intensity parameter \( K \):

\[
K = Y\sigma\sqrt{\pi a}
\]

where \( Y \) is a geometric factor (typically on the order of about 1), \( \sigma \) is the gross stress across the fracture plane, and \( a \) is the crack length. The stress-intensity parameter \( K \) quantifies the stresses at a crack tip, and a critical stress-intensity value \( (K_c) \) thus can be defined as:

\[
K_c = Y\sigma_f\sqrt{\pi a_c}
\]

where \( \sigma_f \) is the fracture stress occurring with a critical crack size, \( a_c \). The critical stress intensity, also known as fracture toughness \( (K_c) \), is the value of stress intensity \( (K) \) that results in rapid, unstable fracture. Fracture toughness \( (K_c) \) depends on both the thickness of the section and the ductility of the material. For a given material, the fracture toughness (or critical stress intensity, \( K_c \)) decreases as section thickness is increased. The value of \( K_c \) decreases with increasing section thickness until a minimum value is reached. The toughness at this minimum, which is an inherent material property, is the plane-strain fracture toughness \( (K_{Ic}) \). Plane-strain fracture is a mode of brittle fracture without any appreciable macroscopic plastic deformation and is thus referred to as linear-elastic fracture mechanics (LEFM). The general conditions for LEFM analysis are expressed as:

\[
\text{thickness} \geq 2.5(K_{Ic}/\sigma_y)^2
\]

where \( \sigma_y \) is the yield strength.

Linear-elastic fracture mechanics is a useful tool in failure analysis as many (and perhaps most) structural failures occur by the combined processes of crack initiation followed by subcritical crack growth mechanism (for example, fatigue, stress corrosion, creep) until a critical crack \( (a_c) \) is reached. In this regard, fracture mechanics is an effective tool for evaluating critical flaw size \( (a_c) \) that leads to rapid unstable fracture and can help answer questions during a failure analysis, such as:

- Where should one look for the transition from subcritical crack growth to unstable rapid fracture?
• What was the load on the component at the time of failure?
• Was the correct material used and manufacturing/processing sound?
• Was the part designed properly?
• Did the environment influence the failure?

Of course, many situations may involve thin sections and/or very ductile materials, where the conditional constraint for LEFM may not apply. In this case, the measure of toughness is plane-stress fracture toughness ($K_I$) and requires the use of elastic-plastic fracture mechanics (EPFM), as the process of unstable fracture involves some plasticity. Plane-stress fracture toughness ($K_I$) is higher than plane-strain fracture toughness ($K_{IC}$), but when thinner sections and more ductile materials are involved net-section instability becomes a factor. See the appendix in the article “Fracture Appearances and Mechanisms of Deformation and Fracture” for more information on the use of fracture mechanics in failure analysis.

Metallographic Examination

Metallographic examination of polished and of polished-and-etched sections by optical microscopy and by electron-optical techniques is a vital part of failure investigation and should be carried out as a routine procedure when possible. Metallographic examination provides the investigator with a good indication of the class of material involved and its structure. If abnormalities are present, these may be associated with undesirable characteristics that predispose to early failure. It is sometimes possible to relate them to an unsuitable composition or to the effects of service, such as aging in low-carbon steel that has caused precipitation of iron nitride, or gassing in copper. Microstructural examination may also provide information as to the method of manufacture of the part under investigation. It can reveal the heat treatment and possible deficiencies in heat treatment such as decarburization at the surface. Microstructural inspection can also reveal possible overheating through coarsening of carbides of superalloys and solution and precipitation of manganese sulfide. Other service effects, such as corrosion, oxidation, and severe work hardening of surfaces, also are revealed, and their extent can be investigated. The topographical characteristics of any cracks, particularly their mode of propagation, can be determined, for example, transgranular or intergranular. This provides information that can be helpful in distinguishing between different modes of failure. For example, fatigue cracks always propagate perpendicular to the maximum cyclic tensile stress while stress-corrosion cracking may propagate along grain boundaries.

Only a few general directions can be given as to the best location from which to take specimens for microscopic examination, because almost every failure has individual features to be taken into account. In most examinations, however, it must be determined whether the structure of a specimen taken adjacent to a fracture surface or a region at which a service defect has developed is representative of the component as a whole. This can be done only by the examination of specimens taken from the failure region and specimens taken from other locations. For instance, in the case of ruptured or bulged boiler tubes in which failure is usually restricted to one portion only, it is desirable to examine specimens taken from both sides of the fracture, from a location opposite the affected zone, and also from an area as remote from the failure as the size of the sample permits in order to determine whether the failure has been due to a material defect or to overheating—and, if the latter, whether this was of a general or localized nature. In investigations involving general overheating, sometimes the original condition of the material can be ascertained only from a sample cut from a part of the tube many feet away from the affected zone.

Metallographic specimens should be taken perpendicular to the fracture surface, showing the fracture surface in edge view. In cases where metal cleanliness may be an issue, the specimen orientation must be selected properly to determine inclusion density and morphology. This type of examination must be performed on the unetched metallographic specimen.

In the investigation of fatigue cracks, it may be desirable to take a specimen from the region where the fracture originated to ascertain if the initial development was associated with an abnormality, such as a weld defect, a decarburized surface, a zone rich in inclusions, or, in castings, a zone containing severe porosity. Multiple fatigue-crack initiation is very typical of both fretting and corrosion fatigue and may form in areas where there is constant stress across a section. Similarly, with surface marks, where the origin cannot be identified from outward appearances, a microscopic examination will show whether they occurred in rolling or arose from ingot defects, such as scabs, laps, or seams. In brittle fractures, it is useful to examine a specimen cut from where the failure originated, if this can be located with certainty. Failures by brittle fracture may be associated with locally work-hardened surfaces, arc strikes, local untempered martensite, and so forth. For good edge retention when looking at a fracture surface, it is usually best to plate the surface of a specimen with a metal, such as nickel, prior to mounting and sectioning, so that the fracture edge is supported during grinding and polishing and can be included in the examination. Alternative means include hard metal or nonmetal particles embedded in the mount adjacent to the edges.

Analysis of Metallographic Sections. As with hardness testing and macroscopic examination, the examination of metallographic sections with a microscope is standard practice in most failure analyses, because of the outstanding capability of the microscope to reveal material imperfections caused during processing and of detecting the results of a variety of in-service operating conditions and environments that may have contributed to failure. Inclusions,
microstructural segregation, decarburization, carbon pickup, improper heat treatment, untempered white martensite, second phases such as γ' in nickel-base superalloys and intergranular corrosion are among the many metallurgical imperfections and undesirable conditions that can be detected and analyzed by microscopic examination of metallographic sections. Even in the absence of a specific metallurgical imperfection, examination of metallographic sections is invaluable to the investigator in the measurement of microstructural parameters such as case depth, grain size, thickness of plated coatings, and heat-affected zone size (HAZ)—all of which may have a bearing on the cause of failure.

Mechanical Testing

Hardness testing is the simplest of the mechanical tests and is often the most versatile tool available to the failure analyst. Among its many applications, hardness testing can be used to assist in evaluating heat treatment (comparing the hardness of the failed component with that prescribed by specification), to provide an estimate of the tensile strength of steel, to detect work hardening, or to detect softening or hardening caused by overheating, decarburization, or by carbon or nitrogen pickup. Hardness testing is also essentially nondestructive except when preparation of a special hardness test specimen is required, as in microhardness testing. Portable hardness testers are useful for field examination, but the type of hardness test must be appropriate for the sample. For example, Brinell is preferred over Rockwell for a gray cast iron part. One must ensure the proper load is used for the test specimen thickness.

Other mechanical tests are useful in confirming that the failed component conforms to specification or in evaluating the effects of surface conditions on mechanical properties. Where appropriate, tensile and impact tests should be carried out, provided sufficient material for the fabrication of test specimens is available. After all photography, fractography, and nondestructive testing have been performed, the tensile properties of the failed component(s) may be tested. (It may be necessary to obtain permission for this destructive testing if litigation is involved.) This involves sectioning the component and machining test specimens. (In all phases, from rough sectioning to the application of load and measurement of the final dimensions, it is advisable to photodocument the process.) It is especially important to understand that the failed component may have been exposed to environmental conditions not experienced by exemplar components. For example, if the component was involved in a crash, there may have been a fire that exposed the component to temperatures that would have altered the mechanical properties. Likewise, the forces that acted to produce an overload failure in one area of the component may have also plastically deformed the component in other areas. To determine the mechanical properties of the component material, a location for testing must be chosen that has not been exposed to detrimental conditions. Here again, exemplar testing is a valuable tool for the failure analyst and should be employed when data from a failed component is skewed by environmental conditions.

Exemplar components are components that match the failed component (i.e., the same part number and hopefully the same lot or batch). The use of exemplars can range from simply being a visual reference in a demonstration to being used for testing specimens to provide mechanical property data about a particular part or batch of parts. The closer the exemplars are to the failed component (geometrically and chronologically), the more reliable the comparison of test data. If a raw material discrepancy is suspected, it may be that other components of the same batch or lot have the same properties. It may also be useful to have exemplars from other batches for comparison testing. The failure analyst should exercise care in interpreting mechanical test results. If a material has a tensile strength 5 to 10% below the minimum specified, this does not mean that low hardness or strength is the cause of its failure in service. Also, it should be understood that laboratory tests on small specimens may not adequately represent the behavior of a much larger structure or component in service. For instance, it is possible for a brittle fracture of a large structure to occur at or near ambient temperature, while subsequent laboratory tests of Charpy or Izod specimens show a transition temperature well below -18 °C (0 °F). The effects of size in fatigue, stress-corrosion, and hydrogen-embrittlement testing are not well understood. However, on the basis of the limited evidence available, it appears that resistance to these failure processes decreases as specimen size increases. Several investigators have found correlation problems of transition-temperature-type impact tests with service performance.

Occasionally, the mechanical properties may be acceptable over most of the component, but may vary at a bend or other discontinuity. Castings can have significant variations in properties from one location to another depending on the solidification practice for the casting. Thus, the location of the test specimen within the component can also be significant. Mechanical property tests for cast components are frequently performed on coupons separately cast for this purpose. Therefore, results from samples from the casting itself may not be directly comparable.

Other factors may affect material properties results. Material may have been tested prior to forming or other deformation. Subsequent coatings or case hardening may have improved or degraded mechanical properties. Variations from one location to another due to local material processing variations may help explain differences in mechanical properties between the bulk material and the failure origin.

Tensile tests, in many failure analysis investigations, do not provide enough useful information because relatively few failures result from metal that is deficient in tensile strength. Furthermore, samples cut from components that have failed in a brittle manner generally show adequate ductility under the conditions imposed during a tensile test.
Sometimes, however, there is justification for tensile testing of failed components to eliminate poor-quality material as a possible cause of failure. Often, these tensile tests for determining material quality are carried out by manufacturers and suppliers when examining components that have been returned to them for analysis.

The role of directionality in tensile testing of wrought metals should also be considered. Specimens cut transversely to the longitudinal axis of a component (such as a shaft, plate, or sheet) usually give lower tensile and ductility values than those cut along the longitudinal axis. This is due to the marked directionality and the resulting anisotropy produced during rolling or forging. When sectioning tension-test specimens from the failed component, special attention should be paid to the orientation of the specimen. Some components may have quality-assurance notes on the drawings that indicate where tensile specimens are to be taken. This is especially true of critical aircraft components. Anisotropic materials have properties that vary with test specimen orientation. Tensile-strength and yield-strength specifications are usually given in the longitudinal and transverse directions. Typically, and unless otherwise specified, the tensile specimen should be taken with its major axis parallel to the direction of grain flow; however, test specimens are typically taken in two of the three directions: longitudinal and long transverse. (The short transverse direction is typically not tested since it is difficult to obtain specimens of sufficient length in that direction.) It may be necessary to lightly polish the component surface and use a macroetchant for the material being worked with to determine the grain flow. The specimens are removed from the parent component by first cutting a rough specimen shape, then final machining to the specified form. Residual stresses in the component may result in warped specimens or pinched cutting tools during rough machining. Care should be taken to document the residual stress observations, so that comparison to exemplar parts sectioned in the same manner can be made. Hopefully, final machining of the testpiece will correct the warped shape while still allowing the specimen to run parallel to the grain flow. Straightening of a warped tensile specimen during the test will result in a nonlinear indication on the initial portion of the stress-strain curve. This can be corrected to a line by using curve-fitting software or by drawing a line by eye back to the abscissa as shown in Fig. 5. There may also be nonuniformity in the specimen microstructure that affects the properties of the material, for example, a change in grain size due to cold work or changes in temper due to the HAZ of a weld.

**Fig. 5** Examples of stress-strain curves requiring foot correction. *D,* point where the extension of the straight (elastic) part diverges from the stress-strain curve. Source: Ref 1

It may be necessary to make some tests either at slightly elevated or at low temperatures to simulate service conditions. Also, it may be helpful to test specimens after they have been subjected to particular heat treatments simulating those of the failed component in service to determine how this treatment has modified mechanical properties. For example, treating a steel at a temperature in the embrittling range for about 1 h prior to impact testing will indicate any tendency to strain-age embrittlement. The determination of the ductile-to-brittle transition temperature may be useful in investigating brittle fracture of a low-carbon steel.

Component Proof Testing. Critical components are sometimes proof tested after manufacture. This is especially true of critical components that will be subjected to large loads. Proof testing involves loading a component past the recommended operating limits, and possibly slightly into the yield zone. This helps to ensure there are no manufacturing or materials defects that would cause premature failure. Proof loads are usually determined on a case-by-case basis and may be expressed in several forms, such as proof load to twice the operating load or proof load to 90% of the determined yield strength. Extreme care should be taken not to damage the component in the specification of a proof load. For example, stretching a component may alter a designed preload by relieving residual compressive stresses introduced during manufacture. There are usually documents with details of specific methods, tools, and loads used in proof testing and results of tests on critical components. If performing proof testing on exemplar parts from a given manufacturer, it will lend credibility to the results if the manufacturer's test procedures are followed. If the proof tests cite standard methods, such as ASTM, ISO, or other industry-accepted procedure, acquire a copy of the method and follow it closely. If results vary from those reported by the manufacturer, it may be necessary to prove the results are valid.
Conductivity of Aluminum Alloys. During a failure analysis investigation of a heat treated aluminum alloy, conductivity testing may be performed to evaluate proper heat treat condition, to assess areas for heat damage, or to estimate tensile strength. Conductivity is the reciprocal of electrical resistivity and is directly proportional to the mean free path of an electron in the crystal structure of the material. The mean free path is affected by the microstructure of the material, which is affected by heat treat condition.

Solution-annealed structures have soluble constituents that have precipitated out of solution, providing for a material with high conductivity. Aging causes fine precipitation of a second phase and thus decreases the conductivity of the material. Natural aging yields the lowest conductivity and intermediate strength, while artificial aging results in intermediate conductivity and highest hardness/strength. Overaging of the alloy causes the conductivity and hardness values to approach those of the solution-annealed condition. The manufacturer does not typically specify conductivity values of an aluminum alloy component. However, with proper design and material selection, the alloy and temper of the component are specified, and corresponding values of conductivity may be found in reference data such as specification SAE-MIL-H-6088. Conductivity measurements must be paired with hardness measurements to determine if the failed component meets alloy and heat treatment requirements. An aluminum alloy and its temper can be determined by measuring hardness and conductivity and verifying against reference data.

**Chemical Analysis**

In a failure investigation, routine analysis of the material is usually recommended. Often it is done last because an analysis usually involves destroying a certain amount of material. There are instances where the wrong material was used, under which conditions the material might be the major cause of failure. In many cases, however, the difficulties are caused by factors other than material composition.

In most instances, slight deviations from specified compositions are not likely to be of major importance in failure analysis. However, small deviations in aluminum content can lead to strain aging in steel, and small quantities of impurities can lead to temper embrittlement. In specific investigations, particularly where corrosion and stress corrosion are involved, chemical analysis of any deposit, scale, or corrosion product, or a substance with which the affected material has been in contact, is required to assist in establishing the primary cause of failure.

Where analysis shows that the content of a particular element is slightly greater than that required in the specifications, it should not be inferred that such deviation is responsible for the failure. Often, it is doubtful whether such a deviation has played even a contributory part in the failure. For example, sulfur and phosphorus in structural steels are limited to 0.04% in many specifications, but rarely can a failure in service be attributed to sulfur content slightly in excess of 0.04%. Within limits, the distribution of the microstructural constituents in a material is of more importance than their exact proportions. An analysis (except a spectrographic analysis restricted to a limited region of the surface) is usually made on drillings representing a considerable volume of material and therefore provides no indication of possible local deviation due to segregation and similar effects.

Also, certain gaseous elements, or interstitials, normally not reported in a chemical analysis, have profound effects on the mechanical properties of metals. In steel, for example, the effects of oxygen, nitrogen, and hydrogen are of major importance. Oxygen and nitrogen may give rise to strain aging and quench aging. Hydrogen may induce brittleness, particularly when absorbed during welding, cathodic cleaning, electroplating, or pickling. Hydrogen is also responsible for the characteristic halos or fisheyes on the fracture surfaces of welds in steels, in which instance the presence of hydrogen often is due to the use of damp electrodes. These halos are indications of local rupture that has taken place under the bursting microstresses induced by the molecular hydrogen, which diffuses through the metal in the atomic state and collects under pressure in pores and other discontinuities. Various effects due to gas absorption are found in other metals and alloys. For example, excessive levels of nitrogen in superalloys can lead to brittle nitride phases that cause failures of highly stressed parts.

Various analytical techniques can be used to determine elemental concentrations and to identify compounds in alloys, bulky deposits, and samples of environmental fluids, lubricants, and suspensions. Semiquantitative emission spectrography, spectrophotometry, and atomic-absorption spectroscopy can be used to determine dissolved metals (as in analysis of an alloy) with wet chemical methods used where greater accuracy is needed to determine the concentration of metals. Combustion methods ordinarily are used for determining the concentration of carbon, sulfur, nitrogen, hydrogen, and oxygen.

Wet chemical analysis methods are employed for determining the presence and concentration of anions such as Cl\textsuperscript{−}, NO\textsubscript{3}\textsuperscript{−}, and S\textsuperscript{2−}. These methods are very sensitive.

X-ray diffraction identifies crystalline compounds either on the metal surface or as a mass of particles and can be used to analyze corrosion products and other surface deposits. Minor and trace elements capable of being dissolved can be determined by atomic-absorption spectroscopy of the solution.

X-ray fluorescence spectrography can be used to analyze both crystalline and amorphous solids, as well as liquids and gases.
Infrared and ultraviolet spectroscopy are used in analyzing organic materials. When the organic materials are present in a complex mixture (such as, for example, solvents, oils, greases, rubber, and plastics), the mixture is first separated into its components by gas chromatography.

Analysis of Surfaces and Deposits. Wavelength-dispersive x-ray spectrometers (WDSs) and EDSs are frequently used for providing information regarding the chemical composition of surface constituents. They are employed as accessories for SEMs and permit simultaneous viewing and chemical analysis of a surface. The Auger electron spectrometer is useful for detecting the elements in extremely thin surface layers. The Auger electron spectrometer can provide semiquantitative determinations of elements with atomic numbers down to 3 (lithium). The size of the area examined varies greatly with the test conditions; it may be from 1 to 50 μm in diameter.

For chemical analysis of surface areas as small as 1 μm in diameter, the electron-microprobe analyzer is widely used. This instrument can determine the concentration of all but the low atomic number elements, with a limit of detection below 0.1%. The area examined with the ion-microprobe analyzer is a few microns in diameter larger than that examined with the electron-microprobe analyzer. The ion-microprobe analyzer has the advantage of being able to detect nearly all elements (including those of low atomic weights) in concentrations as low as 100 ppm. It is sometimes used to volatilize materials, which are then passed through a mass spectrometer.

Electron microprobes and other modern analytical instruments are described in greater detail in Materials Characterization, Volume 10 of the ASM Handbook.

The instruments discussed previously are used for direct analysis of surfaces; other techniques can be used for analyzing material that has been removed from the surface. For example, if material is removed in a replica (perhaps chemically extracted), it can be analyzed structurally by x-ray diffraction or electron diffraction. Also, depending on the quantity of material extracted, many of the routine chemical analysis techniques may be applicable.

Spot testing uses chemical tests to identify the metal, the alloying elements present, deposits, corrosion products, and soil. Spot tests can be performed both in the laboratory and in the field; they do not require extensive training in analytical chemistry. The only requirement is that the substance be dissolvable; hydrochloric acid or even aqua regia may be used to dissolve the material.

Spot tests for metallic elements such as chromium, nickel, cobalt, iron, and molybdenum are usually done by dissolving a small amount of the alloy in acid and mixing a drop of the resulting solution with a drop of a specific reagent on absorbent paper or a porcelain plate. Spot colorings produced in this way indicate the presence or absence of the metallic radical under test. Samples may be removed from gross surfaces by spotting the specimen with a suitable acid, allowing time for solution and collecting the acid spot with an eyedropper.

**Simulated-Service Testing**

During the concluding stages of an investigation, it may be necessary to conduct tests that simulate the conditions under which failure is believed to have occurred. Often, simulated-service testing is not practical because elaborate equipment is required, and even where practical it is possible that not all of the service conditions are fully known or understood. Corrosion failures, for example, are difficult to reproduce in a laboratory, and some attempts to reproduce them have given misleading results. Serious errors can arise when attempts are made to reduce the time required for a test by artificially increasing the severity of one of the factors—such as the corrosive medium or the operating temperature. Similar problems are encountered in wear testing.

On the other hand, when its limitations are clearly understood, the simulated testing and statistical experimental design analysis of the effects of certain selected variables encountered in service may be helpful in planning corrective action or, at least, may extend service life. The evaluation of the efficacy of special additives to lubricants is an example of the successful application of simulated-service testing. The aircraft industry has made successful use of devices such as the wind tunnel to simulate some of the conditions encountered in flight, and naval architects have employed tank tests to evaluate hull modifications, power requirements, steerage, and other variables that might forestall component failure or promote safety at sea.

Taken singly, most of the metallurgical phenomena involved in failures can be satisfactorily reproduced on a laboratory scale, and the information derived from such experiments can be helpful to the investigator, provided the limitations of the tests are fully recognized. However, many company managers prefer to conduct trials to verify improvements before major conversions are approved. This is a more conservative approach, but it only takes one improper recommendation that results in an adverse result to justify trials of major changes.

**Reference cited in this section**

1. P.M. Mumford, Test Methodology and Data Analysis, Tensile Testing, P. Han, Ed., ASM International, 1992, p 55
Corrosion Failures

Corrosion is traditionally defined as the destructive chemical or electrochemical reaction of a metal with its environment. A broader, more modern definition of corrosion is the deterioration of a material or its properties due to reaction with its environment. The latter definition makes three important points. It does not limit corrosion to chemical or electrochemical processes, because some forms of corrosion do not involve these processes. It also characterizes corrosion damage as the deterioration of the material or its properties, because some forms of corrosion weaken the material without visible changes in appearance or measurable weight loss. Finally, the modern definition of corrosion acknowledges that nonmetallic materials may corrode. Examples of corrosion of nonmetallic materials are the aging of rubber due to the effects of heat and/or oils; swelling, oxidation, stress cracking, and ultraviolet deterioration of plastics; destruction of the binder in concrete by environmental agents; and the biological attack or rotting of wood. This section focuses on analysis of metal corrosion.

There are many forms of metallic corrosion. These include: uniform corrosion, localized attack (e.g., crevice corrosion and pitting), galvanic attack, cracking phenomena (e.g., stress-corrosion cracking, hydrogen embrittlement, liquid metal embrittlement, and corrosion fatigue), velocity phenomena (e.g., erosion, cavitation, and impingement), fretting, intergranular attack, and dealloying or selective leaching. These forms are discussed in more detail in the article “Forms of Corrosion” in this Volume.

The type of corrosion, the corrosion rate, and the severity and extent of corrosion are influenced by the nature of the environment, the metal surface in contact with the environment, and the mechanical stresses (magnitude and direction). These factors do not necessarily remain constant as corrosion progresses. They are affected by externally imposed changes and those resulting from the corrosion process itself. Other factors that greatly affect corrosion processes include temperature and temperature gradients at the metal/environment interface, crevices in the metal part or assembly, relative motion between the environment and the metal part, and the presence of dissimilar metals in an electrically conductive environment. Processing and fabrication operations such as surface grinding, heat treating, welding, cold working, forming, drilling, and shearing produce local or general changes on metal parts that, to varying degrees, affect their susceptibility to corrosion.

When a corrosion failure has occurred, several means of preventing or minimizing future failures are available. Very often more than one method is used at the same time. The more important corrective and preventive measures are:

- Change in alloy, heat treatment, or product form (e.g., solution annealing of austenitic stainless steels minimizes the risk of intergranular attack and stress-corrosion cracking)
- Use of resin coatings (e.g., acrylcs, epoxies, phenolics, furanes and urethanes in the form of paints, potting compounds, adhesives, coatings, and linings)
- Use of inert lubricants (chemically inert resins such as silicones, esters, and fluorocarbons that sometimes can serve both as effective lubricants and as corrosion-resistant coatings and linings)
- Use of electrolytic and chemical coatings and surface treatments (e.g., anodizing of aluminum and aluminum alloys for protection in natural, “nonaggressive” environments)
- Use of metallic coatings (e.g., zinc-rich coatings)
- Use of galvanic protection (cathodic or anodic)
- Design changes for corrosion control
- Use of inhibitors
- Changes in pH and applied potential
- Continuous monitoring of variables.

Corrosion Failure Analysis

Complete investigations of corrosion failures can be very complex, although not all corrosion failures require a comprehensive, detailed failure analysis. Often the preliminary examination will determine the extent of investigation required. In general, the investigation should consider various possibilities without being unnecessarily costly or time consuming. Routine checks to determine that the specified material is used is important. Such checks have shown that “seamless” tubes that failed in service by developing longitudinal splits were actually welded tubes that corroded preferentially at the welds. Forgings that failed in service actually were castings in which failure was initiated by
corrosion at porous areas exposed at the surface of the metal. Also, Monel metal parts that corroded rapidly in an environment to which Monel is highly resistant actually were strongly magnetic and made of carbon steel. In another instance, failure of a braided copper wire was traced to the substitution of a carbon-black filler for the usual silica filler in a sheath covering the wire. Thus, galvanic action occurred between the carbon and the copper in the presence of moisture. While knowledge of the part and its application is desirable in any failure analysis, accurate history is especially important in the investigation of corrosion failures.

Information about the type of environment to which the failed part was exposed is essential. For instance, corrosion behavior in plants along rivers is affected by both local and upstream chemical composition in the system. Other determining factors include the temperature, whether exposure to the environment is continuous or intermittent and whether these and other factors varied during the service life of the part.

If available, engineering drawings and material and manufacturing specifications for the part should be examined. Particular attention should be given to any part changes that may have been made. Missing information should be obtained from operating and inspection personnel, if possible. At the same time, the accuracy of any relevant documentary information, such as daily log sheets or inspection reports, should be verified. The investigator should determine if any tests or changes affected the physical evidence of the failure.

On-Site Examination and Sampling

On-site examination is generally the same for corrosion failures as for other types of failures. The region of failure itself should be examined visually using hand magnifiers and any other suitable viewing equipment that is available. The areas immediately adjacent to and near the failure, as well as related components of the system, should be examined for possible effects on the failure. Remote, related equipment should be examined, especially in complex systems and where liquid or gases flow. Also, the possibility of introducing chemicals or other contaminants from upwind or upstream areas should be checked. In one case, for example, ammonia released to the atmosphere by a neighboring plant lead to the SCC failure of a carbon steel boiler.

If possible, the failed component and related components of the system should be photographed before samples are removed. Color photographs are highly desirable, especially when colored products of corrosion are present. A good practice is to calibrate the color at the beginning of every session by taking a picture of the film box, for example, using the same lighting conditions.

Accessories useful in on-site examinations include plastic bags for holding samples, a stainless steel spatula for digging out soft corrosion products, and a file capable of cutting through hard scales. A magnet can distinguish austenitic from martensitic and ferritic stainless steels, as well as steels from nonferrous alloys.

Sampling. When on-site sampling is done, the investigator should be guided by the information already obtained about the history of the failure. The bulk environment to which the failed part was exposed should be sampled, and suitable techniques should be used to obtain samples and make observations (such as pH, for instance) on the local environment at the point of failure.

In addition to taking samples from the failed area, samples from adjacent areas or apparently noncorroded regions should be obtained for comparison purposes. New or unused parts can provide evidence of the initial or unexposed condition of the part.

Removal of specimens and samples of corrosion products from the failed part or area requires the utmost caution. Care must be taken to avoid destroying valuable evidence or damaging the part and its related components. For example, torch cutting often is used for removal of specimens because it is fast and convenient. With torch cutting, cuts should be made away from the failure site to prevent alteration of the microstructure, thermal degradation of residues, and introduction of contaminants. If an abrasive cutoff wheel or a saw is used, the same precautions to avoid overheating apply. Also, coolants or lubricants that can contaminate or alter the part or any deposits present should not be used.

After the samples have been extracted, they must be suitably protected during transportation to the laboratory, such as with glass vials and polyethylene bags. One way to retain deposited material is to tape a covering of inert plastic over the critical area.

Preliminary Laboratory Examination

The procedures followed in the preliminary laboratory examination will vary, depending on whether an on-site examination has already been performed by the failure analyst and its completeness. An on-site examination by a well-equipped investigator will include much of the work that otherwise would have to be done in the preliminary laboratory examination.

When no on-site examination has been done, a failure analyst will be aided by records on the part and environment and the remainder of the failed part (or at the very minimum a good photographic record of it). Also helpful are undamaged or unused parts, related components, and samples of the environment.
Preservation of Evidence. Whether or not an on-site examination has been done, the samples should be handled so that maximum information can be gained before any sample is damaged, destroyed, or contaminated, preventing further tests. Also, a complete written and photographic record should be kept through all investigation stages.

Visual Examination and Cleaning. First, the sample is examined visually, most often with the aid of a low-power hand magnifier. At this stage, important features include the extent of damage, general appearance of the damage zone, and the color, texture, and quantity of surface residues. If substantial amounts of foreign matter are visible, cleaning is necessary before further examination. The residues can be removed in some areas, leaving portions of the failure region in the as-received condition to preserve evidence. When only small amounts of foreign matter are present, cleaning can be deferred so that the surface can be examined with a stereomicroscope before and after cleaning. Cleaning also can be deferred until necessary for surface examination at higher magnifications or for the preparation of metallographic specimens. Small amounts of residues can be removed using transparent tape or acetate replicas and retained for later analysis. Washing with water or solvent, with or without the aid of an ultrasonic bath, usually adequately removes soft residues that obscure the view. Inhibited pickling solutions will remove adherent rust or scale. Usually, the cleaning solutions should be saved for later analysis and identification of the substance removed.

Nondestructive Tests. For parts in which internal damage may have resulted from corrosion or from the combined effects of corrosion, stress, and imperfections, nondestructive testing is desirable. Radiography and ultrasonics can be used to locate internal discontinuities, and magnetic-particle and liquid-penetrant techniques are used to locate surface imperfections. Methods such as eddy current and holography are used less frequently, mainly because these methods require the use of standards to accurately interpret the data.

Microscopic Examination. Examination by both light microscopy and electron microscopy can be used to observe minute features on corroded surfaces, to evaluate microstructure of the metallic parts, and to observe the manner and extent to which the metal was attacked by the corrodent. Viewing the cleaned surface with a stereomicroscope clearly shows gross topographic features such as pitting, cracking, or surface patterns that can provide information about the failure mechanism. This information includes whether corrosion was the sole phenomenon involved, the type of corrosion, and whether other mechanisms, such as wear and fracture, also were operative.

If the features cannot be observed clearly using a stereomicroscope, instruments such as deep-field photographic microscopes or a SEM may be used. These instruments produce images with a greater depth of field and, therefore, can resolve the topographical features of very rough surfaces. Transmission electron microscopy, using replicas, can resolve extremely fine features.

Microscopic examination of polished or polished-and-etched specimens can reveal microstructural features as well as damage such as cracking. If the corrosion products possess sufficient coherence and hardness to be polished, they should be retained. One way to keep the surface material in place is to impregnate the sample with a casting-type resin and allow it to harden before cutting samples. Usually, it is helpful to vacuum impregnate the sample during the casting process to be sure that any surface-connected voids are filled with casting plastic. To secure maximum quality of retention, polishing on napless cloths with diamond abrasives is used.

Chemical Analysis of the Corrosion Products

The job of the failure analyst is to establish what role, if any, the corrosion products played in the failure and to identify and analyze the metal or metals of which the failed part was made, the environment to which the failed part was exposed, inhomogeneities in the part surface, and foreign matter and metal surfaces.

Both conventional techniques (such as wet chemical analysis, emission spectroscopy, x-ray diffraction, infrared spectrophotometry, gas chromatography, and x-ray fluorescence spectrogaphy) and special techniques (such as energy dispersive x-ray spectrometry, electron-microprobe analysis, ion-microprobe analysis, Auger-electron spectrometry, Mössbauer spectrometry, and electron diffraction) may be needed to define the composition and structure of various substances completely.

Identification and analysis of the metals of which the failed part was made usually is routine. Ordinarily, the purpose is not to look for minor deviations from the chemical composition specified for the part, but rather to check for possible major deviations in composition and to make sure that the correct alloy was used. However, for some austenitic stainless steels, the corrosion resistance and other properties of welded joints require close control over the composition of the stainless steel and the weld metal.

The bulk composition of the failure environment as well as the local composition of the environment at the metal interface are important in determining whether, or how, corrosion contributed to failure. Composition of the environment usually can be obtained chemically or spectroscopically. Serious corrosion damage can result from the presence of inhomogeneities in the surface of a metal part used in a corrosive environment. A classic example is severe local attack on a stainless steel because of embedded particles of “tramp” iron in the surface of the stainless steel. Establishing the nature of a layer of material on the metal surface may be more difficult than analyzing the material, especially when a thorough on-site investigation cannot be made. The surface
layer can be merely a trace of innocuous soil or a residue of corrodents or of corrosion products. It also can differ in composition from the bulk metal.

**Corrosion Testing**

Various types of corrosion-testing techniques are used to investigate corrosion failures and to evaluate the resistance to corrosion of metals and alloys for service in specific applications. They include accelerated tests, simulated-use tests, and electrochemical tests. Other techniques include monitoring performance in pilot-plant operations and in actual service. Some accelerated-test methods have been accepted as standard by both the military and industry. To shorten testing time, corrosion is accelerated in relation to naturally occurring corrosion, usually by increasing temperature, using a more aggressive environment, or increasing stress (e.g., SCC). Because various factors influence natural corrosion processes and differ widely in time dependence, the results of laboratory tests must be interpreted carefully. They can be related to expected actual service behavior only when close correlation with long-term service results has been established.

Simulated-use tests are used frequently to analyze corrosion behavior of metals and alloys in specific applications. In these tests, either actual parts or test specimens are exposed to a synthetic or natural service environment. Electrochemical tests provide data that can establish criteria for passivity or anodic protection against corrosion and determine critical breakdown or pitting potentials. The two general methods of electrochemical corrosion testing are controlled current and controlled potential. For either test method, ASTM G 3 provides useful guidance and standardization of the manner of recording and reporting electrochemical measurements. In the controlled-current method, the current (a measure of the corrosion rate) is controlled and the resulting corrosion potential is measured. Several instruments are available for such tests, in which either logarithmic or linear polarization curves are developed. Both galvanostatic and galvanodynamic polarization measurements are employed to plot anodic and cathodic polarization curves.

In the controlled-potential method of electrochemical testing, the corrosion potential (oxidizing power) is controlled and the resulting corrosion current is measured. Equipment is available for both constant-potential (potentiostatic) and varying-potential (potentiodynamic) testing to determine overall corrosion-rate profiles for metal-electrolyte systems over a range of potentials.

For the most part, electrochemical tests are more valuable for evaluating the corrosion resistance of materials per se rather than for direct use as a failure analysis tool.

**Tests for Corrective Action**

Once the failure cause and mechanism have been established, the conclusions and the effectiveness of recommended corrective actions can be confirmed through testing. These tests should simulate as closely as possible the environmental and mechanical conditions to which the failed part was subjected in service. Ideally, the part should be tested in service; however, such testing frequently is not feasible because of factors such as length of time to failure or extraneous damage caused by failure of the part. Thus, simulated-service tests may have to be performed in the laboratory. The following parameters should be carefully controlled:

**Environmental factors:**

- Temperature, which may be steady or fluctuating, and also may affect stress
- Single-phase or two-phase environment, which may involve alternate wetting and drying
- Environmental composition, including major and minor constituents, concentration and changes thereof, dissolved gases, and pH
- Electrochemical conditions, which may involve galvanic coupling or applied cathodic protection

**Mechanical factors:**

- Loading, which may be static or cyclic—if cyclic, mean stress may be zero, tensile, or compressive; also, stress-wave shape and period must be defined
- Surface damage, which may occur by fretting, abrasion, cavitation, or liquid-impingement corrosion

**Practices in Failure Analysis**
Wear Failures

Wear, friction, and lubrication are complex, interwoven subjects that may all affect the tendency of a part to fail or to cease being able to perform its intended function. While all three are very important, only wear receives emphasis in this section, while friction and lubrication are covered less extensively. The science of wear is called tribology, from the Greek word tribos, meaning wear, and it forms the basis for a systems approach to wear failure analysis. A systems approach to wear is necessary to understand the interactions among:

- The system material components—the bulk and surface properties of the wearing parts, the properties of the material causing the wear, and the interface medium
- The operating variables
- The interactions among the material components of the system
- The operating environment

It should already be clear that wear failures are not easy to analyze, describe, or prevent. There is no general agreement even as to the various types or forms of wear because they can be described in several different ways. One relatively simple way to classify the mechanical surface damage caused by wear is to describe eight major forms or mechanisms. This section discusses these eight form of wear, as follows:

- Abrasive wear
- Erosive wear
- Adhesive wear
- Fretting
- Cavitation
- Liquid-droplet impingement
- Rolling-contact fatigue
- Corrosive wear

About half of all wear failures are caused by abrasive wear, 15% by adhesive wear, and about 8% each by erosion and fretting. The other four mechanisms account for the remaining wear failures. The characteristics and mitigation of each are briefly described in the following section, along with lubricant failures.

Types of Wear

Abrasive Wear. The principal mechanism of abrasive wear is indicated by cutting. Abrasive wear is sometimes referred to as grinding wear. High-stress, low-speed particles or projections from hard materials cut and plow small grooves in softer materials. This type of wear is characteristic of damage to many ground contact tools, such as plows and cultivators, as well as other parts that dig into or rub against hard, abrasive materials. Again, it can lead to complete or partial failure and destruction of the part, particularly with respect to cutting tools, which can be badly dulled. Abrasive wear can be mitigated by changing material to:

- High-carbon steel
- 20 to 30% Cr white cast iron
- WC or TiC composites

and by increasing surface hardness using:

- Hard chrome plate
- Hardfacing alloy weld overlay
- Electroless nickel-phosphorus alloy plate
- High-velocity oxyfuel (HVOF) ceramic coatings (WC, CrC, aluminum oxide)

Lubrication is of marginal use in combating abrasive wear, unless a lubricant layer thicker than the particle size can be maintained. Further examples and details are given in the article “Abrasive Wear Failures” in this Volume.

Erosive wear is similar to abrasive wear, except that the force is provided by the kinetic energy of the particles as they are carried in a fluid. Low-stress, high-speed particles impinge upon the surface at some angle, tending to cut (ductile metal) or fracture (brittle material) very small wear chips, or particles from the surface. Erosion is a particular problem with various types of impellers, propellers, fans, and other parts where particles in a fluid, such as air or water, strike the
surface. The very large number of very small particles tends to cause channeling, forming grooves and removing soft surface layers (e.g., paint). Erosive wear also tends to round sharp edges of various types of parts, reducing efficiency and eventually leading to complete destruction.

In general, erosive wear resistance is increased with the hardness of the surface. Frequently, a relatively thick hardfacing or a relatively thin, hard, wear-resistant coating is used, because it may not be desirable to increase the general hardness of the metal. The increased hardness may lead to brittle fracture, which could be a worse problem than the erosive wear. Some specific steps that can be taken are to change flow conditions by:

- Reducing fluid velocity, but particle dropout must be avoided
- Eliminating turbulence at misalignments, diameter changes, gaps at joints, and so forth
- Avoiding sharp bends less than about 3.5 pipe diameters

and change material to:

- Harder material, but must be harder than particles, for example: high-chromium white cast iron or nitride case-hardened steel
- Hard coatings, for example: cement-lined pipe; tungsten carbide, hard chrome, electroless nickel; or cobalt-base alloy flame, arc, or HVOF deposits
- Elastomer or rubber lining in slurry piping

Adhesive wear occurs by dynamic metal-to-metal contact between two surfaces sliding relative to one another, when there is poor or nonexistent lubrication. The principal mechanism of adhesive wear is described by the key word “microwelding,” which is similar to friction welding. At low stress, bonding is on a microscale at high points, called asperities, at least initially. At later stages and higher stresses, bonding is more extensive, and severe damage occurs. Adhesive wear, then, involves microwelding between two metals that are mutually soluble; that is, they are inherently capable of being welded together. Since adhesive wear is often basically a lubrication problem, a lubricant is frequently involved and must be considered as part of the system. Discussion of lubricants and lubrication is contained in a later section of this article.

While the term dynamic adhesive wear is preferred, other terms are frequently used to describe varying degrees of damage. Some of these terms, in order of increasing severity, are:

- **Scuffing**: superficial scratches on the mating faces
- **Scoring**: grooves cut into the surface of one of the components
- **Galling**: severe tearing and deep grooving of one face and buildup on the mating surface
- **White layer**: in steels, the formation of a very hard (>800 HV) white etching phase by frictional heating
- **Seizure**: “friction welding” of the mating parts so they can no longer move

These terms are less accurate, and it is preferred to use the general term of adhesive wear. The “dynamic” aspect refers to the movement of one surface sliding past another surface, such as a shaft rotating in a sleeve bearing, making and breaking of threaded connections, or two gear teeth contacting under load. Adhesion is favored by chemically clean surfaces, nonoxidizing conditions, and by chemical and structural similarities between the sliding couple. If a poor lubricant, or no lubricant, is present in the interface, the adhesion between the two surfaces rapidly escalates and very large wear scars may occur, accompanied by gross overheating. The heat comes from the friction between the two surfaces (see the information on friction in a later section of this article). In some cases, complete destruction of surfaces may occur. In order to have destruction of the surface, frictional heating must bring the local temperature at the interface into the 870 to 1090 °C (1600 to 2000 °F) range, or higher, for steel. At this temperature, great changes occur in the microstructure of hardened steels, which are temperature-sensitive. Rehardening will occur, with the formation of a very hard, brittle, untempered martensite, “white layer,” at the surface, surrounded by softer, highly tempered martensite below the surface. The structure has a pattern that is quite similar to that of grinding burn.

To prevent or minimize adhesive wear, all facets of the situation must be considered; a system approach must be taken. The result may be that several courses of action will be taken. The considerations must include the nature of the wearing metals, the surfaces of these mating components, and the lubrication or other environment that may be present. Adhesive wear is the one wear mechanism that is most degraded by poor lubrication, or lack thereof, and the most benefited by good lubrication.

Lubrication to separate and cool the surfaces and remove wear debris can be used to minimize adhesive wear:

- Keep bulk lubricant cool to prevent overheating at the contact surfaces from frictional heat.
- Use lubricating oil with EP or other additives that form chemical films to prevent metal-to-metal contact.
- At higher temperatures, use greases or solid lubricants.
Materials with better inherent adhesive wear resistance can be selected:

- Martensitic and precipitation-hardened (PH) stainless steels have generally poor galling resistance.
- Austenitic stainless steels, with high work-hardening rate, are better.
- For carbon and alloy steels, higher hardness is beneficial.
- Stainless steel S21800 and nickel casting alloy CY5SnBiM are very resistant to adhesive wear.

Surface finish and coatings are important for minimizing adhesion:

- Relatively smooth surfaces do not have projections (asperities) that penetrate the lubricant film, which, in turn, should have as high of a viscosity as possible for the application.
- A surface that is too smooth (<10 μm.) will not carry the lubricant into the contact zone and will not tolerate debris.
- Hard coatings, plating, or overlays can be beneficial.
- Diffusion coatings, such as nitriding, tuftriding, and sulfidization, can be good.
- Use mutually insoluble metals with low shear strength to prevent adhesion. (Gold or silver plating is sometimes used on one of two mating high-speed gears, or other contacting parts.)

Fretting, sometimes referred to as static adhesive wear, is wear that occurs between two closely contacting surfaces having oscillatory relative motion of extremely small magnitude. The term fretting is derived from the French word “frottement,” which means rubbing. Fretting wear has also been called fretting corrosion, friction oxidation, chafing fatigue, and wear oxidation. Fretting is the cause of discoloration and damage in close-fitting joints or between any two closely fitted parts that are then subjected to vibration during service. Some examples are:

- Press fits on shafts at coupling hubs, gears, wheels, and bearing inner races
- Bolts, rivets, and pins under the heads and in their holes
- Rolling-element bearings while stationary during shipping
- Wire rope between the strands
- Turbine and compressor blade roots
- Heat exchanger tubes in tube sheets and baffles

Instead of having a dynamic sliding motion as discussed previously, fretting involves adhesive wear in a joint that is essentially stationary except for the small vibratory elastic motion that inevitably occurs during machine operation. In other words, the joint is supposed to be stationary, but actually it is not, due to a minute metal-to-metal rubbing motion at the interface. Again, microwelding occurs, tearing tiny fragments from one or both surfaces. This wear debris remains between the surfaces and oxidizes to form a reddish-brown “cocoa” (mostly αFe₂O₃ iron oxide) on iron and steel, or a black powder (aluminum oxide) on aluminum. It is the wear debris that oxidizes when exposed to air that causes fretting to also be called fretting corrosion, although the “corrosion” seems to be incidental to the wear resulting from the adhesion.

One of the major problems with fretting is not the destruction of the joint itself, but the fact that it frequently leads to fatigue cracks and fracture originating in the interface. Fretting can cause the endurance limit of a steel to be reduced from, say, 515 to about 375 MPa (75 to about 55 ksi), often enough to cause a service failure. In many cases, the site where the fretting fatigue originates is different from the expected location of fracture, particularly if the expected failure location has been strengthened by heat treating or mechanical prestressing of the surface.

The major variables in fretting are the tightness of the interface and the magnitude of the vibration that is causing relative motion. The probability of the fretting leading to fatigue failure depends on the magnitude of the cyclic stresses in the fretted part and the reduction in the endurance or fatigue limit. Fretting is not easy to prevent or eliminate, although it can be minimized by one or more of the following procedures:

- Increase the interference of press-fit components, so as to increase the clamping force and prevent relative motion. Recommended values are given in design handbooks.
- Use plastic or rubber pads, gaskets, or bushings to absorb some or all of the vibration.
- Increase the hardness of one or both surfaces to increase the resistance to microfracture.
- Roughen the surfaces in an attempt to “lock” them together, although this probably will not work unless the pressure or interference fit is also increased between the two surfaces.
- Chromium plate one surface to reduce the coefficient of friction and try to prevent microwelding to the other surface.
- Heat treat with a salt bath or gas atmosphere diffusion treatment to form an εFeN compound in the surface that resists fretting of ferrous metals.
Elimination of the vibration causing the fretting is often impossible, because it is a normal aspect of the operation of the equipment or because it is so small as to be beyond one’s ability to correct (e.g., by better balancing of rotating components).

Cavitation is the removal of material from a surface by the formation and rapid collapse of bubbles of gas or vapor in an adjacent liquid. Sometimes called cavitation corrosion or cavitation erosion, this type of wear is somewhat unique in that only one metal surface is involved. The other material is a liquid, frequently water or water/antifreeze. The basic mechanism is true pitting fatigue, resulting in pits in the surfaces of parts such as diesel engine wet cylinder sleeves, hydraulic turbines, nozzles, impellers, and propellers, or any other part that moves rapidly through liquid. It can also occur in pressure let-down valves in rapidly flowing liquid piping systems. High-speed flow of liquid in these devices causes local hydrodynamic pressures to vary widely and rapidly. The resulting pits can be quite small or very large and deep, depending on the particular circumstances.

Cavitation pitting is caused by rapid, repetitive, relative movement between the metal and the liquid. When the local pressure is reduced, as when the metal and liquid momentarily move in opposite directions, small cavities are formed in the liquid at the interface with the metal in the low-pressure regions. These gas or vapor cavities form when the local pressure drops below the vapor pressure of the liquid or dissolved gas. Then, when the metal and the liquid move toward each other with reversal of movement, the cavities at the interface implode, or collapse, on the metal surface, causing very high local-contact stresses. If vibration or other relative movement is continued for a large number of load applications at these local points on the metal, the implosion forces cause fatigue fracture or pits to form in the metal surface. Continued service of this type can cause complete penetration of the metal and/or large pits in the metal. One of the characteristics of cavitation pitting is that the pitting occurs primarily in a pattern on the surface, with most of the pits concentrated in specific areas, depending on the local vibration or flow characteristics. Another characteristic of cavitation damage is that the pitting is very rough and jagged. For example, if cavitation is occurring in a centrifugal pump, the damage will generally be on the back side of each vane of the impeller at the same location near the inlet and will be jagged. Erosion damage, on the other hand, will be on the front or pressure side of the vanes and will generally be grooved.

Cavitation pitting of diesel engine wet cylinder sleeves can result as the sleeve wall moves in and out due to the internal pressure cycles. It can cause complete destruction of an engine if the coolant gets into the combustion chamber or can cause loss of efficiency of an impeller or propeller. Any corrosion is incidental to the pitting fatigue that causes the pits, although corrosion can easily occur on fresh, unprotected metal fracture surfaces.

Some of the variables involved in cavitation are:

- Liquid composition, density, surface tension, and viscosity
- Dissolved gas content
- Pressure relative to the vapor pressure of the liquid at the system temperature
- Temperature relative to the boiling point at the local pressure
- Vibration levels
- Flow velocity—there is a power-law relationship with an exponent between 3 and 7
- Material

Cavitation pitting is not easily prevented, because the vibration and liquid flow characteristics are not easily controlled. Some possible ways to avoid it are to streamline fluid flow by design changes to avoid sudden pressure drops, stiffen the part to change vibration characteristics, increase the vapor pressure of the liquid or the total pressure in closed systems, and reduce surface roughness. Pumps should be operated at or very near their design flow rate and head pressure; cavitation occurs in off-design operation. Material changes, such as using a metal with a higher fatigue strength, can slow but not eliminate damage. For example, a Stellite overlay can increase resistance by 10 times or more over bare type 304 stainless steel, but will not stop the cavitation. If none of these possible solutions work, it may be necessary to simply replace the parts during scheduled maintenance.

Liquid-droplet impingement is, in some ways, the inverse of cavitation, but it produces damage that is very similar to cavitation damage in appearance. Liquid-droplet impingement is metal loss from the repeated impact of liquid droplets that are being carried along in a gas or vapor stream. The damage is generated by the initial high-pressure pulse in the area of impact causing cracks and radially displaced metal, followed by breaking off of the asperities thus formed by subsequent rapid radial flow from subsequent droplet impacts. The droplets usually form by condensation from a vapor stream, as the temperature, the pressure, or both are reduced. The most common places for this damage to occur are the blades in the late stages of steam turbines and downstream of elbows in piping systems. Similar to cavitation, the area where metal has been removed by liquid-droplet impingement will be very rough and jagged.

The significant variables in liquid-droplet impingement damage include:
• Material resistance
• Flow velocity
• Pressure and temperature drops
• Droplet quantity and size
• Liquid specific gravity

The primary direction to take in combating this form of wear damage is to modify the flowing gas properties. Design or process changes can be used to eliminate, or at least reduce, the size and quantity of the droplets. Material changes to minimize damage could include:

• Increasing the local hardness in the damage zone, such as by flame or induction hardening
• Using weld overlay, laser clad, or brazed-in inserts of a material such as Stellite
• Applying very hard coatings, such as HVOF tungsten carbide

Rolling-contact fatigue is surface damage that results by fatigue from repeated rolling, or rolling and sliding contact between curved metal surfaces. The key words descriptive of this type of wear are “pitting fatigue” with origins on or near contacting surfaces. Although the cause of the fatigue and the locations of the origins are somewhat different in each subtype, these are true fatigue fractures that are caused by shear-stress origins due to contact stresses between two metal surfaces. Three mechanisms have been identified and are described: subsurface-origin pitting, surface-origin pitting, and subcase-origin spalling. This form of wear damage is most commonly seen in rolling-element (ball and roller) bearings, gears, and rolling-mill rolls.

The common thread linking the three types of rolling-contact fatigue is that each represents the limit of load-carrying capacity; improvements are made only with great difficulty, if at all, assuming that the materials are proper and the geometries are accurate. Where metal-to-metal contact is concerned, such as in gears or rolling-contact bearings, the materials are usually hardened steels, with hardness values near 60 HRC. More details on rolling-contact fatigue are given in the article “Fatigue Failures” in this Volume.

Corrosive wear really is not a separate type of wear, but is simply a combination of corrosion and wear occurring either alternatively or simultaneously. Corrosive wear can result in more rapid degradation of material than would be observed by either the corrosion or the wear acting individually. Two conditions can occur. In one case, corrosion deposits that form are more easily removed by one of the wear mechanisms, such as erosion. In the other case, the wear process can remove protective deposits or passive layers, exposing bare metal to the corrosive environment.

Wear of parts operating in a corrosive environment is particularly difficult to prevent because of the difficulties in specifying and testing alternate metals. Frequently, various types of surface coatings are of great help in solving the problems of corrosive wear. In some cases, the use of lubricants containing corrosion inhibitors can be beneficial.

**Lubricants**

Almost any surface film acts as a lubricant, preventing cold welding of asperities on opposing surfaces or allowing opposing surfaces to slide across one another at a lower frictional force than would prevail if the film was not present. Lubricants may be either gas, liquid, or solid. The main function of a lubricant is to maintain separation of the surfaces by pressure within the film, which opposes the applied contact force. Another function of a lubricant is to carry away heat generated by two surfaces sliding under contact pressure. Liquid lubricants dissipate heat better than solid or semifluid lubricants, but the shear properties of all types of lubricant are critical to performance. A third function of liquid lubricants is to flush dirt and wear debris out of the contact zone, so they can be removed subsequently by filtration.

Lubricating oils are relatively free-flowing organic substances that lower the coefficient of friction in mechanical devices. They are available in a broad range of viscosities, and many are blended or contain additives to make them suitable for specific uses. In general, lubricating substances that are fluid at 20 °C (68 °F) are termed *oils*; lubricating substances that are solid or semifluid at 20 °C (68 °F) are termed *greases*.

Oils are derived from petroleum (mineral oils), plants, or animals (fixed oils). Commercial mineral, oil-base products consist mainly of saturated hydrocarbons (even though naphthene-base crudes are predominantly unsaturated) in the form of chain or ring molecules that are chemically inactive and do not have polar heads. These commercial products may or may not contain waxes, volatile compounds, “fixed” oils, and special-purpose additives.

Silicone-based oils have been developed that can be used for somewhat higher temperatures, where mineral oils or fixed oils would break down and degrade.
Fixed oils are usually considered to have greater “oiliness” than mineral oils. Oiliness is a term that describes the relative ability of any lubricant to act as a boundary lubricant. Electron-diffraction experiments have shown that molecules of an effective lubricating agent—a long-chain fatty acid of high molecular weight, such as stearic acid or oleic acid—are attached to a metallic surface by polar bonding. They stand up much like individual strands in a pile carpet. The result is a surface layer with high adhesion and high resistance to contact stress and low resistance to lateral shear along the surface. Lubricating grease is a semifluid product consisting of a dispersion of a thickening agent in a liquid lubricant. In more practical terms, most greases are stabilized mixtures of mineral oil and metallic soap. The soap is usually a calcium, sodium, or lithium compound and is present in the form of fibers whose size and configuration are characteristic of the metallic radical in the soap compound. Some greases may contain additives to make them more suitable for specific applications, such as metalworking or threads on fasteners. Some of these materials include graphite, molybdenum disulfide, polytetrafluoroethylene (PTFE or Teflon), copper, nickel, and zinc powders.

Solid lubricants provide certain advantages in high-vacuum, aerospace, or cryogenic applications, where liquids would evaporate or congeal, and at somewhat higher temperatures, where liquids could decompose. Numerous solid inorganic and organic compounds as well as certain composite materials may be classified as solid lubricants. Notable examples are graphite, molybdenum disulfide, and PTFE. Dozens of proprietary compounds may also be classified as solid lubricants. Although solid lubricants may be applied to achieve design simplification or weight reduction, they usually are adopted because of their good stability, retaining their lubricity at elevated temperatures, in chemically active environments, and under exposure to nuclear radiation. However, there has been some evidence in the literature that molybdenum disulfide can break down in hot, moist environments to release hydrogen sulfide, which can cause hydrogen embrittlement of some steels.

Lubrication Modes. Several basic lubrication modes should be understood in order to analyze and correct wear failures. In all modes, contact surfaces are separated by a lubricating medium, which may be a solid, a semisolid, or a pressurized liquid or gaseous film. In hydrodynamic lubrication, the shape and relative motion of the sliding surfaces cause the formation of a fluid film having sufficient pressure to separate the surfaces. In hydrostatic lubrication, the lubricant is supplied under sufficient external pressure to separate the opposing surfaces by a fluid film. Elastohydrodynamic lubrication is a system in which the friction and film thickness between the two bodies in relative motion are determined by the elastic properties of the bodies, in combination with the viscous properties of the lubricant at the prevailing pressure, temperature, and shear rate. In dry-film (solid-film) lubrication, a coating of solid lubricant separates the opposing surfaces and the lubricant itself wears away. Boundary lubrication and thin-film lubrication are two modes in which friction and wear are affected by properties of the lubricant. In boundary lubrication, each surface is covered by a chemically bonded fluid or semisolid film, which may or may not serve to separate opposing surfaces, and viscosity of the lubricant is not a factor affecting friction and wear. In thin-film lubrication, the lubricant usually is not bonded to the surfaces; it separates opposing surfaces, and lubricant viscosity affects friction and wear.

Lubricant Failures. Devices depending on lubricants to minimize friction and wear can fail when the lubricant fails. Most lubricant failures occur by chemical decomposition, contamination, changes in properties from exposure to elevated temperature, or loss of film. Lubricating oils and greases can fail by any one of the aforementioned processes alone; however, in most situations, chemical decomposition, contamination, and temperature are all involved and are interrelated. In addition, pressure-lubricating systems are susceptible to certain types of failure that result in the inability of the system to provide the required oil flow rate where it is most needed. Thus, failure by lubricant starvation often results. Decomposition. When oil is heated in the presence of air, oxidation occurs. Oxidation increases the viscosity and organic-acid concentration of mineral oils, causing varnish and lacquer deposits to form on hot metal surfaces. Under severe conditions, the deposits may be converted to hard, carbonaceous substances. Fixed oils absorb oxygen more readily than do mineral oils, and some may dry, thicken, and form elastic solids. Certain fixed oils (castor, olive, sperm whale, and lard oils) oxidize more slowly than others. These fixed oils are more widely used in blended oils because of their nondrying characteristics. Temperature also affects oxidation rates. In mineral and blended oils, for example, the rate doubles with each 10 °C (18 °F) rise in temperature. Oxidation rates also are higher when the oil is agitated or when catalysts such as copper or acids are present.

In general, solid-film lubricants fail by mechanical removal of microscopically thin layers. Wear debris, which consists primarily of lubricant particles, is generated by the sliding action of a sharp edge against the bonded film on a contact surface. The sharp edge shears a layer of the film (and sometimes the entire film) from the substrate. Contact between a rolling element and a sharp ridge can chip the film, often initiating more extensive failure. This process eventually causes a lack of dynamic stability (as would result from excessive clearance in a bearing) or adhesive wear of metallic contact surfaces.

Many bonded, solid lubricants derive their adhesion from binders, which are incorporated into the film in quantities up to about 20% by volume. Metals, oxides, silicates, or other ceramics are the most common binders. When wear debris contains binder particles, it abrades the remaining film more rapidly than when it consists solely of particles of the lubricating substance.
Contamination of the lubricant with water or chemicals can lead to lubricant decomposition, corrosion of contact surfaces, or both. Contamination with abrasive substances or debris can cause abrasive wear, especially when the size of the contaminant particles is as thick or thicker than the lubricating film.

In internal-combustion engines, water, halide acids, sulfur acids, and products of partial combustion of fuel hydrocarbons are picked up by the lubricating oil. These contaminants can cause various undesirable chemical reactions with oil and metallic surfaces, resulting in the formation of varnish deposits, sludges, viscous emulsions in the oil, or in corrosive wear of engine components.

Because of their chemical nature, fixed oils are particularly susceptible to chemical alteration by alkalis. Alkalis cause saponification (formation of soap) by direct chemical reaction with the fatty acids in fixed oils. This reaction alters the nature of the lubricant and, consequently, its lubricating properties.

The viscosity of virtually all oils and greases is affected by temperature and pressure. Almost any change in temperature significantly affects viscosity: the higher the temperature, the lower the viscosity. Increase of pressure increases viscosity, but the effect is seldom significant except at very high pressures.

A decrease in temperature of mineral oil from 100 to 0 °C (212 to 32 °F) can increase viscosity by 100 times or more. The viscosity of fixed oils is also affected by temperature, but to a lesser extent than for mineral oils. For the above decrease in temperature range, lard oil will increase its viscosity by approximately 30 times.

Oils containing substantial quantities of volatile compounds may lose these components by evaporation when operating temperatures are too high. This process not only alters the viscosity, but also upsets the chemical nature of the oil, thus changing other properties. In some instances, volatilization can take place within the lubricating film, as when frictional heat is ineffectively removed by the circulating fluid. Bubble formation within the film reduces the load-carrying capacity of the film, leading to adhesive or abrasive wear or, in severe cases, cavitation pitting of opposing surfaces.

Loss of Film. When boundary lubrication is provided by soft metallic soaps (iron stearate, for example), a rise in surface temperature can result in a marked increase in the coefficient of friction and the wear rate suddenly changes from mild to severe. The temperature at which this change occurs is the transition temperature. Transition temperatures generally are within the range of 120 to 205 °C (250 to 400 °F), depending on the lubricant and the chemical composition of the metal substrate. Extreme-pressure lubricants function by reaction with the metal surface rather than by adsorption of components in the lubricant. They often are used as substitutes for soap-type boundary lubricants when operating temperatures exceed the transition temperature.

Prevention of Lubricant Failures. Often, lubricant failure can be traced to the selection of an inappropriate lubricant. Petroleum lubricating oils are available in various formulations, and they have various special properties. When these properties cannot be obtained by conventional refining techniques or are obtainable only at a very high cost by refining, they are imparted to the lubricant by additives. Often, certain oil additives impart entirely new performance characteristics to the base oil.

In many instances, lubricant failures (and subsequent component failures) can be prevented or corrected by changes in design of the device or in design of the lubrication system. Starvation of a bearing caused by inadequate lubricant flow or by clogging of oil passages sometimes can be corrected by increasing the size of the passages. Often, an increase or decrease in clearance between sliding surfaces will enable the lubricant to function more effectively. Shields, covers, and seals sometimes can prevent lubricant contamination from external sources. In other instances, filtration or absorption devices can be incorporated into the system to remove unwanted contaminants.

Analysis of Wear Failures

Analysis of wear failures depends greatly on knowledge of the service conditions under which the wear occurred. However, many determining factors are involved, requiring careful examination, both macroscopic and microscopic. Wear failures generally result from relatively long-time exposure, yet certain information obtained at the time the failure is discovered is useful in establishing the cause. For example, analyzing samples of the environment (especially the lubricant), sludge from the lubricating system, or a used oil filter can reveal the nature and amount of wear debris or abrasive in the system. Therefore, analyses of wear failures involve some techniques in other types of failure. As previously mentioned, it is necessary to consider the complete system involved in the failure. Examination of the worn parts alone can lead to erroneous conclusions and inadequate recommendations.

Background Information

- Determine the nature of the application and the specified manufacturing methods of the parts, including surface treatments or coatings.
- Establish the design and the actual operating conditions: loads, speeds, temperatures, and so forth.
• Find out what lubricant, if any, was specified and how was it applied.
• Establish the ambient environment and service conditions.
• Determine whether the observed wear is normal and typical for the particular application, or if it is abnormal and unique.

**Examination**

• Define the surface configuration of both the worn surface and the original surface.
• Define the relative motions in the system, including direction and velocity.
• Define the force or pressure between the mating surfaces, or between the worn surface and the wear environment on both the macroscopic and microscopic scales.
• Identify the actual materials in the worn part, the environment, the abrasive, the wear debris, and the lubricant.
• Define the type and effectiveness of any lubricant: oil, grease, surface film, naturally occurring oxide layer, adsorbed film, other, or none.
• Determine the wear rate and coefficient of friction.
• Reach a conclusion as to the wear mechanism, or the combination of mechanisms: abrasive, adhesive, rolling-contact fatigue, and so forth.
• Devise a solution and formulate recommendations.

The following sections discuss some of the details and special methods that are involved in the analysis of wear failures. Service Conditions. The first step in wear failure analysis is the initial identification of the type of wear or, if more than one type can be recognized, evaluation of the relative importance of each type as quantitatively as possible. This identification of the type or types of wear requires a detailed description of the service conditions based on close observation and adequate experience. A casual and superficial description of service conditions will be of little value. Descriptions of service conditions often are incomplete, thus imposing a serious handicap on the failure analyst, especially if working in a laboratory remote from the service site. For instance, assume that an analyst must study the problem of a badly seized engine cylinder. Obviously, this is an instance of adhesive metal-to-metal wear or lubricated wear, because use of a suitable engine oil is implied. Furthermore, assume that during an oil change the system had been flushed with a solvent such as kerosene to rinse out the old oil and had been left filled with solvent instead of new oil inadvertently. Also assume that a slow leak, resulting in loss of the solvent, was not detected during the operating period immediately preceding seizure. The analyst probably would receive the damaged parts (cylinder block and pistons) after they had been removed from the engine, cleaned, and packed. If evidence of the substitute “lubricant” could not be established, clear determination of the cause of failure would be extremely difficult or perhaps impossible. Another example could involve a rolling-element bearing that was running hot because of the lack of an adequate oil supply. The first thing maintenance people will often do when they find a hot bearing is add oil. So, when the bearing fails and is sent to the laboratory, it arrives showing clear signs of overheating, but is covered with oil.

Similarly, incomplete descriptions of service conditions can be misleading in analysis of abrasive wear. For example, in describing the source of abrasion that produces wear of mining and ore-handling equipment, generalized references to the ore, such as copper ore, are common. Such descriptions are too vague to be meaningful; the mineral being extracted usually has little effect on the abrasiveness of the mixture, whereas the bulk rock, or gangue, is the principal source of abrasive particles. Unless the gangue minerals are studied both qualitatively and quantitatively, a valid assessment of wear, whether normal or abnormal, is impossible.

In analysis of conventional lubricated wear, detailed description of the lubricant is essential and often must be supplemented by data regarding pressures applied to mating surfaces, operating temperatures, and surface conditions. When corrosion is a factor in lubricated wear, determination of the temperature, degree of aeration, hydrogen-ion concentration, velocity of the lubricant, and the composition and concentration of the corrodent in the lubricant may be difficult. Other analysis complications include the presence of substances that inhibit or accelerate corrosion. In some pumps and compressors, the fluid being processed may leak into the lubrication system, contaminating the oil and causing it to become corrosive. So, it is important in such cases to know the composition of the process stream and to submit a sample of the lubricant for analysis.

Initial Examination and Measurements. Examination of a worn part generally begins with visual observation and measurement of dimensions, usually requiring micrometers, calipers, and standard or special gages. Observations of the amount and character of surface damage often must be made on a microscopic scale. An optical comparator, toolmaker's microscope, recording profilometer, or other fine-scale measuring equipment may be required to assess adequately the amount of damage that has occurred.

Weighing a worn component or assembly and comparing its weight with that of an unused part can help define the amount of material lost. The difference in weight is a measure of the severity of the wear. This material is lost; it is transferred to an opposing surface by abrasive wear or eliminated by abrasive wear, erosion, cavitation, or combinations
Examination of the worn surface can provide much information, including:

- The amount of material removed
- The type of damage (scratching, gouging, plowing, adhesion, pitting, corrosion, spalling, or simple penetration)
- The existence and character of surface films
- Whether certain constituents are being attacked preferentially
- The direction of relative motion between a worn surface and abrading particles
- Whether abrading particles have become embedded in the surface

Because wear is a surface phenomenon, the original surface configuration of the components in contact influences wear by influencing resistance to relative motion. Microscopy may be useful to study features of the worn surface including the configuration, distribution, direction of scratches or gouges, and indications of the preferential removal of specific constituents of the microstructure. Abrasive particles or wear debris should be studied under the microscope to observe their shape and the configuration of their edges (sharp or rounded) and to determine if they have fractured during the wear process. These particles can also be chemically analyzed in an SEM using EDS to determine their nature and indicate their source.

Examination of the worn surface by light microscopy at magnifications up to about 100 diameters usually is required to detect uneven or abnormal wear patterns and to reveal the direction of relative movement between the worn surface and the opposing surface or abrasive. Sometimes, higher magnifications are required, and a SEM may be necessary to study areas of slight wear. These techniques are described further in Friction, Lubrication, and Wear Technology, Volume 18 of the ASM Handbook.

Direct observation at magnifications greater than about 50 diameters can be difficult if the part does not fit in the stage of a metallurgical microscope. Sectioning to remove a portion of the worn surface for direct observation precludes repair and reuse of the part. Some stereobinocular microscopes are capable of such magnifications. Replication is another technique that can be used for light-microscope or SEM observations of worn surfaces of large parts. Replication using plastic films or harder cast materials offers the additional advantage that a reproduction of the surface can be obtained at a remote site and carried back to the laboratory for detailed study.

Changes in surface configuration that occur during the wear process affect subsequent stages of wear. As mentioned earlier, each manufacturing method produces surfaces that are characteristic of that process, whether it is grinding, machining, burnishing, or something else. The process of “wearing in,” which involves progressive reduction in surface roughness by adhesive or abrasive wear of opposing surfaces, generally is followed by a period of relatively little wear. The initial smoothing out of asperities, particularly in lubricated systems that operate under boundary lubrication, reduces the microscopic hills and valleys on the surface to a height about the same as the thickness of the lubricant film. The surfaces then ride on each other with no interference between peaks on the opposing surfaces, and wear essentially ceases. In normal service, then, there will be a characteristic pattern of this break-in wear that will define the expected contact between mating surfaces.

In other instances, particularly if the initial surfaces are somewhat rougher or if boundary lubrication is ineffective, adhesive wear may result in progressive surface roughening and eventual failure. If this process releases wear debris into the joint and the particle size of this debris exceeds the thickness of the lubricating film, combined adhesive and abrasive wear between the opposing surfaces and the wear debris can result in rapid deterioration.

Direction of Relative Motion. When only unidirectional sliding is involved, scratches or gouges produced on the worn surface are aligned with the direction of relative motion. In a sleeve bearing, for example, the scratches should run circumferentially on the inner surface of the bearing and on the mating shaft. Scratches resulting from wear that are oriented in other directions indicate misalignment, vibration, or looseness. These factors can contribute to the severity of the wear.

In devices that undergo combined rolling and sliding, knowledge of the relative velocities and directions of rolling and sliding is necessary for definition of the wear mechanism. The direction of rolling is defined as the direction in which the point of contact moves. The direction of rolling is always opposite to the direction of rotation of a rolling element. On a given surface, a condition of positive sliding exists if the direction of sliding is the same as the direction of rolling.
Negative sliding occurs on the mating surface, where the directions of rolling and sliding are opposite. Most surface-origin, rolling-contact fatigue failures originate in regions of negative sliding, because the shear stresses are usually more severe there than in regions of positive sliding. Negative sliding occurs on the dedenda of gear teeth, on the cam follower riding on a cam, and in other devices on the part that have the lower surface velocity in a rolling-sliding system.

Metallography. Once the worn surface has been characterized, it is appropriate to prepare metallographic sections through the wear damage and through comparable unworn areas. Because wear is a surface phenomenon, it is almost always necessary to use special techniques to avoid rounding of the edge of the metallographic specimen. One technique is nickel plating on the specimen before it is sectioned and mounted. Additions of powdered glass, hard steel shot, or pelletized alumina, to name a few examples, may be used in the mounting material to improve edge support. The mounted specimen must be polished with care using polishing cloths with little nap, such as silk or nylon, so that the edge is not rounded. For valid analyses of very thin surface layers, techniques such as taper sectioning are needed to allow metallographic observations and microhardness measurements. Taper sectioning involves sectioning the specimen at a shallow angle, and it produces a geometric magnification in the direction of sectioning of surface layers and coatings.

Etchants, in addition to preparing a specimen for the examination of microstructure, also reveal characteristics of the worn surface. Two features that can be revealed by etching a worn surface are phase transformations and metal transfer caused by localized adhesion to an opposing surface and the results of overheating caused by excessive friction. An example of a structure produced by overheating is the “white layer” (untempered martensite) that sometimes develops on steel or cast iron under conditions of heavy sliding contact. Etching a worn surface also can help in detecting the selective removal of specific constituents of the microstructure. It is very important that the worn surface be examined under the microscope and photographed before it is etched, because some topographic features may be easier to observe on the unetched surface.

Metallography will determine whether or not the initial microstructure of the worn part met specification, both in the base material and in any hardened case or surface coatings that may have been applied. It also will reveal existence of localized phase transformation, sheared or cold-worked surface layers, or, as described in the case study on chromium-plated cylinders, the presence of embedded abrasive particles.

Macroscopic and microscopic hardness testing indicate the resistance of a material to several wear mechanisms. Because harder materials are likely to cut or scratch softer materials, comparative hardness of two sliding surfaces may be important. Microhardness measurements on martensitic steels may indicate that frictional heat has overtempered the steel and, when used in conjunction with a tempering curve (a plot of hardness versus tempering temperature), can allow a rough estimate of surface temperature. In some failures, such as in case crushing, a microhardness traverse to verify the depth of the hard case will be needed. Bulk hardness measurements also can indicate whether or not a worn part was heat treated correctly.

Effect of Microstructure and Hardness on Wear. The microstructural heterogeneity of a wear surface influences the wear process, because constituents such as carbides, inclusions, intermetallic compounds, and dispersed phases have properties different from those of the matrix. Hard microconstituents such as carbides make a metal extremely resistant to abrasive wear if they are closely spaced in a relatively hard matrix.

Matrix hardness is important to wear resistance. If hard microconstituents are dispersed widely in a matrix that is not hard enough to have good wear resistance of its own, the matrix may wear away rapidly, leaving the hard particles projecting from the surface, where they can cut into a mating surface. For this reason, under dry-sliding conditions, fine pearlite exhibits considerably better wear resistance than does coarse pearlite or a mixture of ferrite and pearlite.

X-ray and electron diffraction analyses disclose the structure of a crystalline solid. These techniques are particularly valuable for analyzing abrasives, wear debris, or surface films because they can identify compounds, not merely elements. Modern diffraction units, with computerized data collection and analysis using search-match techniques, can help by identifying metals, corrosion products, abrasive minerals, and wear debris. Microstructural features such as retained austenite cannot always be seen by microscopic examination of an etched specimen; quantitative diffraction analysis can reveal the relative amounts of such unresolved constituents in the microstructure.

Chemical Analysis. One or more of the various techniques of chemical analysis—wet chemical analysis, spectroscopy, colorimetry, x-ray fluorescence, atomic absorption, EDS, or electron-beam microprobe analysis—usually is needed for properly analyzing wear failures. The actual compositions of the worn material, the wear debris, the abrasive, and the surface film must be known in order to devise solutions to most wear problems.

Although not so familiar to most failure analysts, a similar number of analytical methods are available for evaluating lubricants. If an ample quantity of lubricant is available, some of the bulk properties, such as viscosity, can be measured. Identification of the oil, its additives, and contaminants can be done by Fourier transform infrared (FTIR) analysis and inductively coupled photospectroscopy (ICP). An analysis of the lubricant can determine if the proper base stock and additives were present, and if any contamination or degradation has occurred in service.

Chemical analysis may be needed to establish or confirm the wear mechanism. The following example describes an instance in which adhesive wear caused by solid-phase welding between mating surfaces was verified by electron-beam microprobe analysis.
Geological studies of soils or similar abrasive mixtures of minerals are mandatory in analyzing wear of tillage tools, earthmoving equipment, and ore-handling devices. The abrasive characteristics of sandy soils are different from those of coarse, rocky soils, clay, or fine silt. Particle hardness and shape (jagged or rounded) are important. The degree of compaction determines the amount of pressure that forces abrasive soil particles against a surface that is being dragged through the mixture. In addition, moisture content is important because it determines, in part, the temperature, lubricity, and cohesiveness of the mixture. Many ores, slags, and similar bulk materials are extremely abrasive to equipment for handling and moving materials; geological analysis can define the abrasive character of these aggregates.

Practices in Failure Analysis

**Formulating Conclusions and Report Writing**

At a certain stage in every investigation, the evidence revealed by the examinations and tests is analyzed and premises are formulated. Obviously, many investigations will not become a series of clear-cut stages. If the probable cause of failure is apparent early in the examination, the pattern and extent of subsequent investigation will be directed toward confirmation of the probable cause and the elimination of other possibilities. Other investigations will follow a logical series of stages, as outlined in this article, and the findings at each stage will determine the manner in which the investigation proceeds. As new facts modify first impressions, different hypotheses of failure will develop and will be retained or abandoned as dictated by the findings. Where extensive laboratory facilities are available to the investigator, maximum effort will be devoted to amassing the results of mechanical tests, chemical analysis, fractography, and microscopy before the formulation of preliminary conclusions is attempted. Finally, in those investigations in which the cause of failure is particularly elusive, a search through reports of similar instances may be required to suggest possible clues. Some of the work performed during the course of an investigation may be thought to be unnecessary. It is important, however, to distinguish between work that is unnecessary and that which does not bear fruitful results. During an examination, it is to be expected that some of the work done will not assist directly in determining the cause of failure; nevertheless, negative evidence may be helpful in excluding some potential causes of failure from consideration. Sometimes the evidence that excludes possible failure causes may be stronger than the available evidence supporting the probable cause for a failure.

On the other hand, any tendency to curtail work essential to an investigation should be guarded against. In some instances, it is possible to form an opinion regarding the cause of failure from a single aspect of the investigation, such as visual examination of a fracture surface or examination of a single metallographic specimen. However, before final conclusions are reached, supplementary data confirming the original opinion, if available, should be sought. Total dependence on the conclusions that can be drawn from a single specimen, such as a metallographic section, may be readily challenged unless a history of similar failures can be drawn upon. Yet, circumstances sometimes dictate that a conclusion be formulated on limited data. This is especially true in forensic investigations because of the usually limited extent of the problem and the need to preserve as much evidence as possible. Also, of course, economic limitations are always present. It is rarely justified for the failure analyst to conduct every type of test and investigation possible.

The following checklist, which is in the form of a series of questions, has been proposed as an aid in analyzing the evidence derived from examinations and tests and in formulating conclusions. The questions are also helpful in calling attention to details of the overall investigation that may have been overlooked:

- Has the failure sequence been established?
- If the failure involved cracking or fracture, have the initiation sites been determined?
- Did cracks initiate at the surface or below the surface?
- Was cracking associated with a stress concentration?
- How long was the crack present?
- What was the level or magnitude of the loads?
- What was the type of loading: static, cyclic, or intermittent?
- Is the fracture surface consistent with the type of loading assumed in the design?
- What was the failure mechanism?
• What was the approximate service temperature at the time of failure?
• Did high or low temperature contribute to failure? Were there temperature excursions?
• Did wear contribute to failure?
• Did corrosion contribute to failure? What was the type of corrosion?
• Was the proper material used? Is a better material required?
• Was the cross section adequate for the type of service? Were the stresses too high?
• Was the quality of the material acceptable in accordance with specifications?
• Were the mechanical properties of the material acceptable in accordance with specifications?
• Was the component that failed properly heat treated?
• Was the component that failed properly fabricated?
• Was the component properly assembled or installed?
• Was the component repaired during service and, if so, was the repair correctly performed?
• Was the component properly “run in”?
• Was the component properly maintained and properly lubricated?
• Was failure related to abuse in service?
• Can the design of the component be improved to prevent similar failures?
• Was the failure primary, or was it damaged by failure of another part?
• Are failures likely to occur in similar components now in service, and what can be done to prevent their failure?

In general, the answers to these questions will be derived from a combination of records and the examinations and tests previously outlined in this article. However, the cause or causes of failure cannot always be determined with certainty. In this instance, the investigation should determine the most probable cause or causes of failure, distinguishing findings based on demonstrated facts from conclusions based on conjecture. In most cases, the conclusion should be on the first page, closely followed by recommendations.

Failure Analysis Report Preparation and Writing. A failure analysis report is the summation of all of the individual tests and analyses performed during the course of an investigation. The report coalesces the various results and presents them in a concise and logical format. It should be written to describe the particulars of the subject being studied, convey pertinent information gathered from testing and analysis, and logically and accurately interpret and explain data and results. A report allows the reader to follow the investigator's thought processes during the analysis and documents and explains the course the investigation followed. The author may be the sole investigator, a coinvestigator, a contributor, or other person designated as the primary point of contact. Basically, the body of the report tells the story of the investigation, explaining why certain tasks were performed, in what order, what the results were and their interpretations, and answers the questions originally posed by the customer at the onset of the study. Brevity is important, but not at the expense of completeness and clarity. Reports should be written with the target audience in mind. For example, are the intended readers technically oriented? A report intended for engineers and scientists would likely differ from a report submitted to marketing or finance personnel. Frequently, readers will scan only one or two sections of the report, usually either the abstract, summary, or conclusions section, prior to deciding whether to read the entire report. Consequently, these sections must adequately convey the scope and findings of the study.

Adequate preparation is required prior to actually writing a report. An author must compile all of the information generated during an investigation and review the data. Many investigators find it is helpful to assemble all pertinent data, such as photographs, spectra, and mechanical and physical test results, and lay the information out in front of them sequentially, corresponding to the order the tests were actually performed. Glean the useful data from the redundant or superfluous. Next, review some basic questions to decide exactly what information the report should convey. What did the customer request? Frequently customers ask for something, but it is the responsibility of the investigator to discern what they really want. Which tests and analyses were performed? Constraints may be placed upon the investigation—generally, time or cost. If an investigation is requested, but limited by deadlines or funding, the analyst must decide which tests and procedures will furnish the most useful information in the most timely fashion without exceeding any preestablished budget. If certain tests or analyses were excluded, whether for time, budgetary, or other reasons, it is often beneficial to explain why they were not performed. Certain failures require minimal amounts of testing and analysis; others involve multiple disciplines and extensive investigations. A series of questions,
posed prior to writing the final report, help to ensure that the analysis adequately addresses the most important considerations, regardless of whether the investigation is large scale or routine (Ref 2):

- Has all relevant background information been collected?
- What was the most likely sequence of failure?
- For failures involving cracking or fracture, was there a single initiation site, or multiple sites?
- Was crack initiation surface or subsurface?
- Was a stress concentrator a factor in cracking?
- Was the cracking fresh? If not, how long was the crack extant.
- What was the load intensity?
- Did cyclic, intermittent, or static loading contribute?
- What was the stress orientation?
- Was the failure temperature related? If so, what temperatures did the component encounter?
- Was the failure wear related?
- Did corrosion contribute to the failure? If so, what type of corrosion?
- Was an appropriate material used? Should a more suitable material be used? Was it within specification? It is worth remembering that a specified material may still be inadequate.
- Were the mechanical and physical properties of the material adequate? Within specification?
- Was the component properly heat treated? fabricated? assembled and/or installed? run/broken in? lubricated? maintained?
- Was the component ever repaired? If so, was the repair adequate?
- Did in-service abuse contribute to the failure?
- Would improving component design preclude similar failures?
- How can future failures of comparable components be avoided?

It is preferable for the report to list work chronologically, as the tasks were performed. The flow of the report should be from nondestructive to destructive. The results must satisfy the customer's original requirements. Were standard procedures followed and documented? Individual companies or organizations usually have guidelines concerning the format and content of technical reports, required signatures, distribution, and so forth. Likewise, legal considerations, such as what constitutes a business record and the length of time records and supporting data are retained, should adhere to company policies.

Technical reports are not prose, but the author can still imbue singularity into the work, while maintaining professionalism and cogency. Although there is no single universally accepted report format, most share common features and are tailored to satisfy the needs of the customer and employing organization. For instance, technical reports have historically been written in the past tense from a third-person perspective. However, more authors are now using first person to describe their analyses. They believe first person instills the reader with a better understanding of the flow of the investigation and helps them understand the thought processes behind the analyses performed. Many experienced analysts prefer to put their conclusions at the beginning of a report, so that the reader understands the pertinence of the information reported. Photographs, figures, charts, and so forth should augment and enhance the textual narrative of the report, not replace it; many experts feel that imbedding such graphics in the body of the report, rather than at the end, tends to help the flow of the report. Company procedures and practices may address such issues as voice and format. If no formal company policy exists, there are many organizations that provide standard formats and guidelines in a variety of media. The American National Standards Institute (ANSI) can provide hardcopy references and standards for guidance (Ref 3). Likewise, the Internet World Wide Web can provide excellent on-line reference sites (Ref 4, 5).

Elements of the Report. Traditionally, failure analysis reports contain, as a minimum:

- An introduction that includes background information on the subject of the report and restates the work requested and may serve as a written record for verbal work requests
- A section detailing investigative procedures and results
- Conclusions

Lengthier reports can comprise:
Most reports begin with a **heading**. The heading may include the company name, department, author, date, report number, and title of the report. The title should be in title case and be as descriptive as possible without being too verbose.

An **abstract** is useful to succinctly describe the scope and objectives of the investigation and the findings of the analysis. All information, descriptions, and conclusions stated in an abstract should be expounded upon in the body of the paper. An abstract should be brief, generally between 150 and 500 words. The abstract should not be a complete summary of the paper; instead, it should convey only the salient points of the report (Ref 3).

A **table of contents** should be included if the report contains many pages, typically ten or more. A table of contents sequentially lists the sections of the report and their respective page numbers.

The **background**, or **foreword**, section of the report supplies the reader with pertinent information and helps serve as an introduction. This section of the report should contain as much useful information as possible, including:

- Nomenclature of the parts involved
- Description of the work requested
- Descriptions of the events surrounding the subject failure
- Related documentation such as specifications, standards, drawings
- Manufacturing and service histories
- Applicable part numbers, materials, loads, and so forth

An **investigative procedures** section details the tests and methods utilized during an analysis. The progression of tests should follow the investigator's test plan. Typically, testing is performed on components sequentially, from nondestructive to destructive. Destructive testing, by its very nature, destroys potential evidence. If litigation is involved, or potentially involved, it is necessary to obtain agreement between all parties as to the scope of the work, and also then to obtain a court order to allow the destructive testing to be performed. This should be addressed and documented. A detailed as-received description is beneficial, complemented with photodocumentation. Photos, figures, charts, and similar visuals should be captioned with an explanatory statement, including applicable scales, units, and so forth. Any cleaning and preparation should be noted. The macroscopic features can be described, followed by the microscopic features. Microscopy generally progresses from optical to electronic. Metallographic and chemical analysis methods used during the investigation should be detailed. Equipment utilized to perform the various tests can be stated in this section. Physical and mechanical testing performed should be documented. Mechanical testing can include hardness and tensile testing. Physical testing can include density and electrical conductivity.

The **results** section of the paper reports the pertinent data and information obtained from the various tests performed during the investigation. The data are subsequently used in the discussion and conclusions sections; therefore, extemporaneous data should not be included. The results should be presented using descriptions, tables, figures, charts, or other graphics. Mathematical and computational procedures and formulaic manipulations and equations should be described textually. If multiple equations are used, each should be numbered sequentially for reference. Testing and data accuracy should be addressed. The appropriate appendix should be referenced concerning error analysis and computational and analytical details.

The **discussion** section of the report describes the interpretation and analysis of the findings obtained from the various tests. The most important results should be recapped. Relationships between various properties, facts, and findings should be explained. The characteristics of the subject specimens can be compared to those of laboratory-manufactured exemplars; this helps gage the accuracy of the tests and provides valuable reference.
information. Plausible explanations relating to failure modes and causes should be addressed. The discussion should proceed from specific to general. Any variations in data should be explained. Possibility versus probability of various scenarios should be considered and discussed.

The conclusions section of a technical report should summarize the most significant results and findings of the investigation and present the most cogent and plausible explanations. It may be necessary to reiterate information discussed in prior sections to bolster any conclusions or theories presented. Conclusions are generally based on facts, but often informed speculation and opinion are required to satisfy customer requirements. State if a conclusion is based on supposition, and describe supporting factual data. As stated previously, frequently the abstract and conclusions sections are the only portions of a report that are read, so they need to be especially concise, accurate, and well written.

The three main outcomes of a failure analysis are: determining how the subject component failed (e.g., overload), establishing why the component failed (e.g., overtorque), and making recommendations to prevent future failures, when applicable and appropriate. Recommendations should address, and rectify, the deleterious factors that contributed to the failure. It is worth noting that sometimes recommendations may be costly to implement or unpopular with others in an organization. However, it is essential that the investigator preclude failures, thereby increasing safety, reducing the likelihood of injury, and increasing service life of the given component.

A signature block is normally required for technical reports. Local company policies generally identify necessary signatories. Signatories’ titles and contact information should be included, immediately below the appropriate signature. As a minimum, the primary investigator/author should sign the report. Any concurrence or editorial signatures should follow the primary signature.

An acknowledgments, or contributors, section is frequently included in technical papers. Coinvestigators, colleagues, technicians, analysts, and so forth who made contributions during an investigation can be acknowledged for their efforts.

Any work from other authors cited in a paper should be referenced. There are two major formats for citations, or referencing work: the number-identification system and the author-date system. The former ascribes specific numbers to the references, corresponding to entries in a successively numbered list of references. The latter parenthetically cites an author's name and publication dates in the body of the text and corresponds to an entry in an alphabetically arranged reference list. Generally, the order of information is: the author's name, the article title, and publication information. Specific formats and examples are discussed at length in various sources, such as the Modern Language Association (MLA) Handbook, 5th edition (1999) and Chicago Manual of Style, 14th edition (1993). Since the advent of the Internet, on-line sources are becoming more common. Although no standards have as yet been adopted, several present similar formats (Ref 6). Most writing style guides and standards organizations generally agree that an Internet reference should include the authors, a descriptive document title, date, and uniform resource locators (URLs).

Appendices should contain computational, factual, and analytical details referenced in the body of the report. Error calculations should be placed in the appendix. Multiple appendices should be designated by sequential letters (Appendix A, Appendix B, etc.). The appendices should contain sufficient information so that the reader can understand and evaluate the results without needing to refer to previous publications.

A glossary of terms used in the report can follow the appendix. Terms and their respective definitions should be in alphabetical order, each on a separate line.

A final review and edit of the report is always beneficial. Useful considerations when reviewing a report include:

- Alternative viewpoints and ideas can be addressed if additional people review a report. This must be balanced against the time involved during the editing cycle.
- Always perform spelling and grammar checks.
- Is all the information contained in the report pertinent to the subject investigation?
- Are the results and conclusions logical and accurate?
- Have the failure mode and cause been ascertained?
- Does the report have a logical flow and continuity of results?
- Are the recommendations plausible?
- Were all the original questions answered?
- Were there any contradictory results? If so, retest or explain.
 Were proper procedures followed and documented?
 Were the test plan followed as designed, or was divergence necessary?
 Were all graphics properly captioned? Were scales included with all photodocumentation?
 Were units used consistently?
 Was the equipment used in the investigation documented and in calibration?
 Is all the information in the report suitable for distribution? Were concerns regarding classified, restricted, or proprietary information addressed?

The Audience. Obviously, not every report will require coverage under each of these sections. Lengthy reports should begin with an abstract or executive summary. Because readers of failure analysis reports are often people in purchasing, operations, accounting, management, and even legal staff, the avoidance of technical jargon wherever possible is highly desirable. A glossary of terms may also be helpful. The use of appendices, containing detailed calculations, equations, and tables of chemical and metallurgical data can serve to keep the body of the report clear and uncluttered. The use of literature or book references is recommended along with references to pertinent standards such as ASTM, and so forth.

Customer needs vary. Some may request a formal report, others an abbreviated report or executive summary. Not all reports require each section previously mentioned. In certain instances, sections of the report can be combined for better flow and continuity. For example, the results and discussion sections, or the procedures and results portions, can be combined into one section. An internal company report may not require the detail that a potentially litigious investigation might merit. It is worth remembering, however, that even the most innocuous report may potentially be used in litigation. Therefore, semantics and nomenclature should be chosen carefully. Potentially inflammatory or insensitive wording should be avoided. As an example, the use of the word “flaw” versus “defect.” Not every flaw is a defect, but all defects are flaws (Ref 7).

Consult Other Disciplines. When available, take advantage of the resources of other knowledge and experience available to you. This can be someone in your company or even a vendor. An interdisciplinary approach to complex failure analyses of large structures or machines is often warranted. Activities such as nondestructive testing and FEA are usually subject areas in which the failure analyst is not proficient. Working with engineers and specialists in other disciplines can be required to reach the root cause. Most specialty support such as nondestructive testing or FEA analysis can be purchased as a service. However, simple collection of results without a meaningful dialogue of the engineering and metallurgical variables at play in a particular failure may lead to poor results and improper interpretation. Developing a team to solve complex problems can ensure the best possible outcome when the failure requires a broad analysis.

Follow-Up on Recommendations. The purpose of most industrial failure analysis is to determine the basic or root cause of failure of metal parts. Failures are, at best, a source of irritation and, at worst, a safety hazard, but they are always an economic loss. Therefore, study of the failure should result in carefully formulated recommendations in the report that are aimed at reducing or eliminating similar failures in the future. Such recommendations may involve adding warning labels, changes in design, metallurgy, manufacture, quality control, maintenance, repair practices, and anticipated usage of the product. If recommendations are made in the report, they should be reviewed with appropriate personnel after the report has been issued so that the recommendations are not overlooked. There may be valid reasons for not implementing the recommendations, but at least those performing the failure analysis and writing the report will have fulfilled their responsibility.

References cited in this section


The following list summarizes key guidelines for conducting a failure analysis investigation:

- There are 14 possible steps to follow in conducting a failure analysis—from collecting and examining samples to testing, determining the failure mode, writing the report, and developing follow-up recommendations.
- In working to produce economical, reliable products, it is vital for manufacturers to recognize the importance of analyzing failures that occur during product development testing or in service.
- Of primary importance in conducting a thorough failure analysis is the compilation of information regarding the manufacturing, processing, and service history of the failed part. This includes obtaining original specifications and drawings, if available.
- It is important for failure analysts to keep written and visual records of their investigations.
- An essential first step in any failure analysis is a thorough visual examination including consideration of fluids, soil, debris, paint, or corrosion found on the surface of the failed part.
- Nondestructive testing methods (magnetic-particle, liquid-penetrant, ultrasonic, radiography, and eddy-current testing) as well as mechanical tests (hardness, tensile, and impact, can be helpful in determining the cause of a failure.
- The influence of the chemical and thermal environment must be kept in mind during any failure analysis. Corrosion and property changes may be critical in an analysis.
- Macroscopic examination of the fracture surfaces can be considered the most important part of a failure analysis. It is typically performed at magnifications of 1 to 100× using the unaided eye, a hand lens, a low-power stereoscopic microscope, or SEM.
- Performing both macro- and microscale examination of unetched and etched specimens provides important evidence about the type and structure of the material involved in the failure.
- Brittle fractures are typically more difficult to analyze because of the large number of possible fracture mechanisms and the lack of obvious deformation.
- In any failure analysis, all avenues must be explored to ensure that the investigator's analysis is complete and that the final conclusions are well documented and accurate.

Acknowledgments

This article is based on contents from the following lessons in the course Principles of Failure Analysis (Materials Engineering Institute, ASM International):

- D. McGarry, “General Procedures for Failure Analysis”
- R. Parrington, “Corrosion Failures”
Contents on reporting writing were contributed by Steve McDaniels, National Aeronautics and Space Administration. Contents on mechanical testing were provided by Daryll McKinley, McSwain Engineering.

Practices in Failure Analysis

References


Photography in Failure Analysis

Richard H. McSwain, McSwain Engineering, Inc.

Introduction

FAILURE ANALYSIS is an investigative process in which visual observations are made and conclusions are drawn based on physical evidence and features present on a failed component. Recording the features observed in a failure analysis is critical both to document the component condition and to communicate the results of the failure analysis to others.

The primary means of documenting the conditions and features observed during a failure analysis investigation is photography. Failure analysis photography is a combination of science and art. Proper imaging techniques are required to produce an acceptable fracture surface photograph. Experience is required to produce a visually acceptable fracture surface photograph. The use of proper equipment, proper technique, and the study of the resulting photographs can lead to exceptional results. This article reviews photographic principles as applied to failure analysis and specific techniques employed in both the field and in the laboratory.
Visual Examination

Visual Observations. Failure analysis is a process centered in visual examination and analysis. The details observed on a fracture surface are based on differences in fracture surface topography, reflectivity, and color characteristics. Observation of these factors requires the failure analyst to use proper illumination techniques and suitable magnification to resolve and distinguish the fracture features present. During the failure analysis process, features on a microscopic scale must be imaged and interpreted; this interpretation needs to be considered in conjunction with macroscopic features, the gross condition of a component, and the environment in which the component was operated to determine the mode and cause of failure.

Fracture Feature Notation. Prior to beginning photography of a failed component, the component should be examined thoroughly to determine its condition; to evaluate which features are important and which features are extraneous, that is, postfracture mechanical damage; and to determine if special treatment or cleaning of the failed component is required. This process typically involves examination with the unaided eye and proceeds to higher magnification, usually by stereomicroscopic examination. A handheld magnifier may also be of assistance at this stage of the investigation. Illumination at varying degrees of obliqueness and orientation can be helpful in evaluating fracture conditions present.

The failed component and any fractured and separated pieces should then be photographed. At this stage, it is important to document how the fracture is related to the remainder of the component as well as to document the relative size of the component, using suitable reference scales in the photograph.

Photographic Documentation Strategy. The failure analysis process, dependent on visual examination, is also dependent on the recording of features observed at each step of the investigation. The failure analysis process may change the appearance of the fractured component by cleaning or sectioning for scanning electron microscopy examination and/or metallography. It is important to photograph the component at each stage of the process to retain the visual information available and the component condition at each step.

Field Photographic Documentation

Composition. Photographic documentation of failed components in the field requires photography of the failed component, the assembly, and the entire structure. Documentation should include the orientation and connection of various components and should include multiple views of each component. Field photographic documentation is used in the failure analysis laboratory for component and fracture orientation, feature reference, and condition assessment. Compositions should avoid extraneous objects in the image that would excessively complicate the photograph. Where possible, the use of multiple photographs taken at closing distances is beneficial in referencing component placement and orientation within an assembly during later parts of the investigation. Photographic lighting in the field should avoid harsh shadows that result in excessive contrast. Harsh shadows can be minimized by use of fill-flash.

Component Identification. It is important during the field photo documentation process to make sketches and record the name and description of components to assist in the laboratory failure analysis process as well as in report preparation and other communications. It is helpful to have component manuals present during the field documentation process, so that proper terminology can be used for the components being examined.
Techniques of instant photography may be beneficial in documenting components in the field and in recording component nomenclature on the photograph.

Field Photographic Documentation Specifications. The American Society for Testing and Materials (ASTM) provides guidelines for photographic documentation of components involved in a failure analysis that may become part of litigation. ASTM E 1188, “Standard Practice for Collection and Preservation of Information and Physical Items by a Technical Investigator,” states that photographic documentation should begin as soon as possible after the incident. If items involved in the incident are disassembled or subjected to destructive testing, each step of the disassembly or testing shall be documented by contemporaneous photographs or videotaping. The failure analyst should constantly keep in mind the fact that the purpose of photography is to record the condition of a failed component at any given period of time for transfer of information in the future. The specific photographic documentation may be the only record of the condition of the component available subsequent to the field examination.

Other ASTM standards providing guidance for failure analysis work and failure analysis documentation include:

- ASTM E 860, “Standard Practice for Examining and Testing Items That Are or May Become Involved in Litigation”
- ASTM E 1020, “Standard Practice for Reporting Incidents”

Numerous photographic texts also are available to assist the failure analysis photographer. Some of these are cited in the Selected References listed in this article.

Photography in Failure Analysis

Richard H. McSwain, McSwain Engineering, Inc.

Laboratory Photographic Documentation

As-Received Condition. It is important to document the condition of the failed component on receipt in the failure analysis laboratory. This documentation is consistent with the guidelines given in ASTM E 860, “Examining and Testing Items That Are or May Become Involved in Litigation,” as well as proper failure analysis technique. The as-received photographic documentation should accurately depict the condition of the failed component, including the fracture surface condition as well as the overall condition of the component.

Accurate color in laboratory photographic documentation is critical, and periodic use of photographic color standards assists in this task.

The as-received photographic process should begin with direct lighting and a scale or ruler in the photograph for reference. After direct lighting, varying degrees of oblique lighting can be used to highlight the topographic features present on the fracture surfaces of failed components. To the extent possible, confusing backgrounds should be eliminated through the use of background paper or material. Stereophotography is extremely useful at this stage of the failure analysis process to record component configuration and spatial characteristics.

The overall component photodocumentation is followed by macrophotography of the component and fracture details, stereomicroscopic photography of the component and fracture details, and scanning electron microscopy of the component and fracture details.

Further analysis may require component sectioning and metallography. Optical metallographic photography is used to document the microstructure and internal structural condition of a failed component.

Component failure simulation testing, if performed, should be photographically documented in a manner similar to the documentation of the failed component condition, using still photography and videodocumentation. In some cases, high-speed videodocumentation may be used in the failure simulation testing.

Laboratory Component Lighting. Most failure analysis laboratory photographic documentation is performed using flash photography. Flash photography ensures that the component is adequately lighted and that the
colors of the component, including any degradation, are accurately recorded. Color-balanced flood-lights can also be used for laboratory photodocumentation. Color photography reveals heating effects, corrosion effects, mechanical damage, coating condition, and other types of component degradation not readily distinguishable using black-and-white photography. Floodlighting may also be used in laboratory photographic documentation to ensure that fracture features are properly highlighted and that the fracture condition is accurately recorded. Side lighting of fracture surface features is especially useful in revealing fracture surface texture that contains information related to the failure mode. In those cases, side lighting of the fracture enhances the topographical features present on the fracture surface and provides a means of communicating the fracture features to others. For relatively small components, the photographic copystand provides an excellent means for documenting component and fracture surface condition by varying the angle and direction of the incident lighting. The copystand should be rigid to avoid vibration when long exposures are used for close-up photography. The copystand allows lighting from single or dual sides of the component. Orientation of the component with respect to the incident lighting highlights different characteristic features on the fracture surface. Highly reflective components can create glare or reflected images of the camera or lights. Special techniques can be employed to deal with highly reflective conditions and are described later in this article.

Magnification Determination. The relative size of a failed component and the magnification of the photograph are important to incorporate in the laboratory failure analysis process. This is most easily accomplished by placing a scale adjacent to the failed component and imaging the scale with the failed component, while keeping the lens-to-film plane constant. The photographic magnification then becomes the ratio of the actual scale size to the size of the scale in the printed image. Printing of the failure analysis photograph to any enlargement size allows the magnification of the enlargement to be determined by measuring the ratio of the actual scale size to the enlarged scale size.

---

**Photography in Failure Analysis**

Richard H. McSwain, McSwain Engineering, Inc.

---

**Photographic Equipment**

Camera Bodies. The most common camera body used in failure analysis photography is the single-lens reflex 35 mm camera. The 35 mm failure analysis camera may have either manual focus and manual exposure or automatic focus and automatic exposure control. The automatic camera body that allows both aperture and shutter priority control is preferable for field and laboratory failure analysis work. The camera body that allows interchangeable lenses is most suitable for failure analysis work, because of the need to perform close-up photography as well as overall, wide-angle, and telephoto-range photography.

Camera Lenses. The most versatile lens used for failure analysis photography is the zoom lens, with a focal-length range of approximately 28 to 120 mm. A typical zoom lens useful for failure analysis photography is shown in Fig. 1. Such a range allows the formatting of the image, so that the primary object in the field of view is the failed component or assembly. Extraneous features in the photograph can be detracting and add difficulty in communicating failure analysis information. The other lens that finds particular use in failure analysis photography is the macro- or close-up lens. A typical macro lens is shown in Fig. 2. Close-up photography is often required in failure analysis to document the detailed features present on a fracture surface. In order to accomplish this, either a macro lens in the 50 or 100 mm focal-length range or a zoom lens with a macrocapability is useful. Adapter rings and extension tubes can also be used to adjust the focal length and provide an increase in magnification. The higher focal lengths have the disadvantage of reduction of depth of field. Either type of macro lens may use manual or automatic focus. In many conditions, the automatic lens is switched to the manual mode in order to achieve focus at very close focal distances.
Camera Flashes. The primary flash used for failure analysis photography must be capable of being mounted on the camera as a “hot shoe” flash or on an extension cord for side lighting. A typical flash arrangement with an off-camera extension cord is shown in Fig. 3. It is preferable that flash photography be performed in a through-the-lens metering mode to achieve optimal exposure. For most failure analyses, use of the off-camera flash in a side-lighting condition provides beneficial shadowing that brings out fracture detail and makes interpretation of fracture features easier. Another flash useful for failure analysis photography is the ring flash, which provides for optimal flash lighting and exposure in macrophotography. A typical ring flash arrangement is shown in Fig. 4. The ring flash also allows selection of flash from a single side of the lens or from above or below the lens. Ring flash lighting is especially useful in recording fracture surface topographic features and surface markings, such as part numbers.
Camera Accessories. The primary accessory important to the failure analyst is a suitable carrying case for camera equipment. A typical case for field use is shown in Fig. 5. Such a case should provide protection for the camera equipment being used in both field and laboratory conditions. Many times, the failure analyst travels to other laboratories or the field, and the camera case must adequately protect the camera equipment in those conditions. Camera cases may be soft-sided, plastic, or metal. The case should be equipped with adequate padding to separate camera components and to ensure cameras are free of damage during transit.
Film Photography

Film Imaging. The primary photographic recording media used in failure analysis is 35 mm film. The 35 mm format, coupled with high-quality film, produces excellent enlargements to poster-sized prints. Film is available as color negatives, color transparencies, and as black-and-white negatives. The most common film used in failure analysis is the color negative, which results in a color print. Fine-grain film is preferable in failure analysis because of the ability to make enlargements. Color prints are able to precisely record the conditions observed by the failure analyst and to provide a means of communicating those observations to others. Film speed is important to the failure analyst because of the nature of the film itself. High-speed films typically have coarser grain and are less capable of producing high-resolution enlargements. Color print films with speeds in the range of 100 to 200 American Standards Association (ASA) are optimal for the failure analysis application. They offer both reasonable shutter speeds and have sufficiently fine grain to produce enlargements up to 75 cm (30 in.) per side.

Film Emulsion. Certain films also incorporate an extra layer of emulsion for recording light-colored detail, such as in portrait photography. These films are especially useful in failure analysis because of their ability to capture detail in light-colored, highly reflective areas often encountered in fracture surface photography.

Digital Photography

Digital Imaging. Scientific digital photography is similar in many ways to color film photography. In film photography, the image is recorded on silver halide crystals in the film emulsion. In digital photography, the image is recorded on a microelectronic sensor made up of an array of light-sensitive elements similar to the silver halide crystals in a film emulsion. The captured image is stored in a removable media that is subsequently used to transfer the image file to a computer or printer for printing.
Digital photography provides immediate availability of images viewable on the camera imaging screen or on a computer screen. Inadequacies in illumination or exposure can be quickly detected and corrected. It is especially useful in documenting stages of a failure analysis where cleaning or sectioning is involved. Each step of the process can be documented with digital photography, and the image captured at each stage can be verified.

Image Size/Resolution. The resolution in a digital image is determined by the number of pixels in the imaging sensor; the higher the number of pixels, the greater the image resolution. High-resolution photographs are attainable with multiple megapixel sensor arrays. Image file sizes in the multiple megabyte range produce excellent photographs with enlargement capability.

File Formats. The two primary digital camera file formats used in failure analysis photography are the TIFF and JPEG file structures. The TIFF and bitmap file formats are uncompressed and provide maximum resolution. The TIFF format is the most commonly used uncompressed format for scientific photography. The JPEG is a compressed file storage format with compression that typically results in a reduced file size with no discernable difference in image characteristics or quality. Care should be taken to avoid multiple JPEG compression steps or high-compression formats that can result in loss of image quality.

Digital Cameras and Devices. Digital cameras have features ideal for failure analysis work. Those include a liquid crystal diode display screen for display of the image recorded, removable storage media, zoom lenses, and macro-focus capability. Such a camera offers all of the standard features found on a professional 35 mm single-lens reflex film camera. It has the same flexibility of interchangeable lenses and off-camera flash. The digital camera has the added advantage of the immediate viewing of the captured image and the rapid production of high-quality prints.

Lenses for Digital Cameras. The professional digital camera body accepts the same lenses as used on professional single-lens reflex film cameras. Typically, the zoom lens is beneficial for recording the condition of the component, and a macro-lens is useful for recording up-close detail.

Digital Storage Media. The digital storage media used in professional digital cameras has several formats, depending on the media selected by the camera manufacturer. Digital media has sufficient storage capability to record the equivalent of multiple rolls of photographic film images. However, care must be taken to not bring the digital media into close proximity to any magnetic device, and it must be protected from physical damage, especially to the electrical contacts.

Digital scanners are useful in failure analysis for introduction of photographs into reports and for electronic transmission of photographs obtained during failure analysis. Professional scanners are preferred because of their increased resolution and image quality. Image resolution should be carefully selected to avoid excessive file sizes that are unreasonable for storage, reproduction, and electronic transmission. An image resolution of 200 to 400 dots/in. is suitable for failure analysis photograph scanning, and both film and digital photographs can be scanned with excellent results.

**Photography in Failure Analysis**

Richard H. McSwain, McSwain Engineering, Inc.

**Basics of Photography**

Aperture selection in failure analysis photography is one of the most critical aspects of the photographic process, because the aperture setting directly controls the depth of field as well as the exposure time. A large aperture opening (small aperture number) results in a shallow depth of field. In those conditions, it is possible that the entirety of the fracture surface may not be in focus. A small aperture opening (high aperture number) results in a large depth of field, which is beneficial in documenting the overall failed component condition. However, the small aperture has limited light transmission and may result in very long exposure times. Long exposure times can lead to loss of sharpness if the camera or support has any vibration present. The typical aperture setting for most failure analysis photography is F8 to F11. Higher aperture settings may be used in conditions where the camera can be mounted either on a copystand or tripod and floodlighting can be used.
those situations, the aperture setting can go to F32 and higher without loss of image sharpness. For automatic cameras, the aperture priority setting is preferred in failure analysis, because of the need to ensure that the entirety of the fracture surface is in focus in the photograph.

Shutter Speed Selection. The shutter speed is critical in creating sharp photographs. The highest shutter speed possible is preferred, because it minimizes the possibility of image blur from camera movement during exposure. A minimum shutter speed of \( \frac{1}{100} \) s produces adequately sharp photographs. If a suitable shutter speed cannot be attained by hand-holding the camera, a flash should be used. Most flash units synchronize to the camera shutter speed of \( \frac{1}{60} \) s, but because of the short flash duration, flash photography does not suffer from loss of sharpness due to camera movement. Floodlight photography can suffer from shutter-speed problems if an excessively large aperture number is selected, which results in very long exposure times. With a camera mounted on a tripod or on a copystand, exposure times as long as 2 s can still yield sharp photographic images.

Feature Exposure Selection. For most failure analysis photography, the correct exposure is one that accurately records the condition of the component. Under certain conditions, multiple exposures are taken to highlight certain features of the component for specific purposes. One of those conditions would be the lighting of the inside surface of a hollow component. In that situation, the outside of the component may be underexposed or overexposed, but the internal surface of the hollow component would be properly illuminated. The general guideline is that exposure should document the overall condition of the component, followed by adjustment, in order to accurately record the visual characteristics of the part of the failed component under investigation.

Filter Selection. The main filter used in failure analysis photography is the ultraviolet haze or UV filter. The UV filter protects the lens from scratching and from accumulation of deposits and dirt. The UV filter also aides in sharpening the photograph under hazy conditions. Polarizing filters may be used in highly reflective conditions. If tungsten lighting is used, then a blue filter, such as an 80B, is commonly used to correct the color temperature shift caused by the tungsten lighting.

---

**Photography in Failure Analysis**

Richard H. McSwain, McSwain Engineering, Inc.

---

**Photographic Lighting**

Natural lighting is typically used in photographic field documentation of failed components and structures. It should be remembered that the color temperature of the lighting varies through the day, and color temperature change affects the appearance of the photograph. Early-morning photographs tend to have a blue tint, and late-afternoon photographs have a golden tint. Use of color standard cards can assist in ensuring that the color recorded is the “true” color of the component. Natural lighting also suffers from variations in image contrast due to sunlight conditions, changing from bright conditions to overcast conditions. Bright sunlight conditions create harsh shadows that may prevent the proper printing of the photograph to show all features present. Severely overcast conditions may result in long exposure times for the required aperture setting and may lead to a loss of sharpness in the image. Fill-flash is often useful outdoors for both the bright sunlight conditions, to fill in heavy shadows, as well as in overcast conditions, to produce adequate overall lighting.

Backlighting is often used for emphasizing a component contour or shape. Such lighting can be accomplished using a lighted panel viewer or other technique to project light from behind a component. It should be kept in mind that the type of lighting used must be compensated for color temperature in order to achieve a proper color photograph. Use of a clear plastic sheet, supported to provide separation from the table top and lit from below, is also useful in producing photography without shadows.

Flash lighting is the most commonly used lighting technique in failure analysis photography. It is well suited to failure analysis because it is directional, and the shadows produced emphasize the fracture features present on a failed component, where the flash is used off-camera. It is critical in using an off-camera flash to properly direct the flash to the component being photographed. This can be accomplished by use of camera flash.
brackets or by carefully holding the flash unit during photography. Care must be taken, as well, to not overexpose components with the flash. Flash overexposure with digital photography is especially critical because of the limited exposure latitude of a digital image compared to a film image.

Copystand Lighting. The two side lights on a copystand provide uniform lighting, with excellent contrast for recording the features of a failed component. Copystand lighting also provides the option of lighting from only one side of the failed component, to better reveal fracture surface details.

Fiber Optic Lighting. Fiber-optic techniques are useful in directing light to a certain location, including cavities and hollow areas, as well as providing a means to emphasize fracture surface topography. The fiber-optic lighting color temperature should be compensated for by using the appropriate filter on the camera lens. Fiber-optic lighting using a borescope can be especially useful in highlighting cavities that are difficult to light by normal means.

Photography in Failure Analysis

Richard H. McSwain, McSwain Engineering, Inc.

Fracture Surface Photography

Fracture Surface Side Lighting. The topography of a fracture surface is one of the most important features present in a failed component. It is the means by which the failure analyst determines the mode of failure. Surface topography is best revealed through proper lighting techniques. For general fracture surface documentation, dual incandescent side lighting with the copystand provides an excellent lighting condition to document the overall condition of a component. High-intensity moveable lighting and other light sources may be required for specific component conditions.

A typical arrangement of a dual-side incandescent lighting copystand is shown in Fig. 6, where lights “A” and “B” both illuminate the specimen. The resulting photograph with dual-side illumination is shown in Fig. 7(a). Dual-side lighting is limited in revealing surface topography, because topographic feature shadows are filled by the opposing light source.

![Fig. 6 Copystand for dual-side incandescent lighting](image)
Fig. 7 Fracture-surface side lighting. (a) Photograph of a fatigue fracture resulting from dual-side incandescent lighting with lights A and B (Fig. 6). (b) Photograph of a fatigue fracture resulting from single-side incandescent lighting from light A (Fig. 6).

The arrangement for a single-side incandescent lighting copystand is to illuminate the specimen with just one of the lights. With illumination from just light “A” in Fig. 6, the resulting photograph clearly reveals the fractographic features present (Fig. 7b). When using single-side lighting, the specimen should be rotated under the lens, and the effects of the lighting should be evaluated to reveal the most detail in the fracture surface.

In general, the light should flow across the grain of the fracture and not parallel to it. Figure 8(a) shows the effect of sample side lighting of the fatigue fracture with the light parallel to the fracture features, and Fig. 8(b) shows the same fracture with the side lighting perpendicular to the fatigue fracture feature. The side lighting perpendicular to the fracture features reveals much more of the fracture detail typical of fatigue. A white card can be used in the position of the opposing light to reflect light into harsh shadowed areas, as shown in Fig. 9(a) and (b).

Fig. 8 Photographs from single-side incandescent light of a fatigue fracture with (a) parallel lighting and (b) perpendicular lighting.
Fracture Surface Ring Lighting. Ring lighting produces excellent uniform light coverage of a fracture surface. A typical ring light attached to a camera lens and the resulting fracture photograph are shown in Fig. 10(a) and (b). The ring light is less able to highlight fracture features but is beneficial in photo-documentation at closer focal lengths.

Fracture Surface Fluorescent Lighting. Fluorescent lighting produces a more diffuse lighting when photographing highly reflective fracture surfaces and provides excellent overall documentation of the component conditions. A typical arrangement of a single-side fluorescent lighting copystand is shown in Fig. 11(a). The resulting fractograph is shown in Fig. 11(b). A typical fluorescent ring light attached to a camera lens and the resulting fracture photograph are shown in Fig. 12(a) and (b). The fluorescent light is limited in showing detail from shadowing of fracture surface features but is especially useful when photographing highly reflective fracture surfaces, such as fatigue in aluminum.
Fracture Surface Fluorescent Ring Lighting. In some cases, the side-lighting effect may need emphasis, as in the case of a fatigue fracture origin. An incandescent fiber-optic light source is useful in such conditions. A typical fiber-optic side-lighting arrangement and resulting fracture photograph are shown in Fig. 13(a) and (b). The resulting fractograph is high in contrast, and the fracture surface topography features are maximized.

Fracture Surface Fiber Optic Lighting. Highly reflective components, such as bearings and polished metallographic specimens, are difficult to photograph because of unwanted reflections. The use of diffuse incandescent illumination is beneficial when photographing highly reflective surfaces. The light diffusion is accomplished using a white paper “tent.” A typical tent lighting arrangement and the resulting photograph are shown in Fig. 14(a) and (b). The outcome is a reduction in unwanted reflections and a resulting acceptable photograph.
Fracture Surface UV Lighting. There are certain requirements where UV lighting may be necessary to document the condition of a component. One such requirement is instrument panel photo documentation, where fluorescent materials are used in the manufacturing process. Another requirement is in the documentation of fluorescent penetrant nondestructive inspection indications. A typical UV lighting arrangement and resulting photograph are shown in Fig. 15(a) and (b). Film photography often has difficulty in capturing the UV fluorescent indications, and long exposure times are often required. Digital photography is more useful in capturing fluorescent indications under UV light, especially with cameras in which the media speed setting can be maximized.

Macrophotography

Macrolenses. Macrophotography in failure analysis work is performed using a zoom lens with macrocapability or a macrolens at either the 50 or 100 mm range. Typically, the maximum magnification achievable with the macrophotography lens is 1 to 1 magnification. A typical macrophotography arrangement and resulting fracture photograph are shown in Fig. 16(a) and (b).
Macrophotography lighting is critical, because the camera is in close proximity to the component being photographed, and normal lighting techniques may not adequately reach the fracture surface or the component. Because of those requirements, either a macroflash is used or some type of side lighting that allows the proper lighting of the fracture surface without shadowing from the lens.

Depth of field in macrophotography is critical. At very close focal distances, the depth of field is critical. Typically, higher f-stop numbers are selected to ensure that the entire fracture surface is in focus.

Reference Locator Photographs. In macrophotography, it is useful to have available a reference locator photograph that shows the overall appearance of a component undergoing a failure analysis. As close macrophotographs are taken, the reference overall photograph can be marked to record the location and condition of each of the macrophotographs taken.

---

**Microscopic Photography**

Optical stereomicroscopy is routinely used in failure analysis photography to examine and characterize fracture surface features. Images observed in the stereomicroscope must be recorded for documentation of the features present and to transfer that information to others. A typical stereomicroscope photographic arrangement is shown in Fig. 17, and the resulting photograph is shown in Fig. 18. The primary techniques for stereomicroscopy photography is film photography, instant film photography, digital photography, and, in some applications, videography. All of the techniques except film photography provide immediate feedback as to image quality. Stereomicroscopic photography of a fracture surface requires adjustment of the microscope to give adequate depth of field. This is accomplished by adjusting the field aperture. All of the techniques of fracture surface lighting equally apply to stereomicroscopic photography, and similar results are obtainable.
Optical metallography is used in failure analysis photography to characterize the macro- and microstructure of a failed component. As in optical stereomicroscopy, the features of the microstructure observed with optical metallography must be recorded and then printed. A typical optical metallograph photographic arrangement and resulting photograph are shown in Fig. 19(a) and (b). The photographic techniques employed in optical metallography include digital photography, film photography, and instant film photography. Optical metallography has historically used monochromatic lighting and black-and-white photography. However, color photography is advantageous for showing heating effects, corrosion
product, and microstructural features with color. Additional and expanded information on optical metallography is in *ASM Handbook*, Volume 9.

![Optical metallograph](image)

**Fig. 19** Optical metallograph. (a) Typical optical metallograph used in failure analysis photography. (b) Optical metallograph photograph of a microstructure

Scanning electron microscopy is used in failure analysis photography for the examination of surfaces, including fracture surfaces, and the recording of fracture features. Scanning electron microscopy is used in failure analysis photography for the examination of fracture surfaces and other surface features and the recording of features present, as well as high-magnification recording of the microstructure. A typical scanning electron microscope and resulting scanning electron microscope photograph of a fracture surface are shown in Fig. 20(a) and (b). The scanning electron microscope is a higher-magnification extension of the optical stereomicroscope as well as the optical metallograph. Scanning electron microscope images are monochromatic and, as such, do not benefit by color photography. The primary techniques of photography associated with scanning electron microscopy are digital photography and instant film photography. X-ray imaging spectroscopy associated with scanning electron microscopy typically uses digital photography for the recording of x-ray images and maps. Videographic cameras may also be used to record entire scanning electron microscopy examination sessions.
Fig. 20 Scanning electron microscopy. (a) Typical scanning electron microscope used in failure analysis photography. (b) Scanning electron microscope photograph of a fatigue fracture

Photography in Failure Analysis
Richard H. McSwain, McSwain Engineering, Inc.

Special Methods

Stereophotography can be used in failure analysis work for documenting component condition and recording information in three dimensions. Stereophotography consists of the capturing of two images, either simultaneously with a stereocamera or sequentially by the repositioning of a single camera. The resulting image pairs are viewed with a stereoviewer, enabling the observation of three-dimensional features and characteristics. Similar techniques are used with the scanning electron microscope by the capture of sequential images, with the second photograph taken after tilting the sample. A professional stereocamera is shown in Fig. 21. Stereophotographs may also be accomplished by using a single camera and a slide bar, as shown in the resulting stereopair image in Fig. 22.
Videography is used in failure analysis photography to capture dynamic events as well as to document rapidly changing component conditions. Current videography cameras are available with either analog or digital technology. Digital videocameras have the advantage of the input of video-images into a computer for frame-by-frame analysis. Videography is also useful in failure analysis photography for low-light-level documentation, including UV lighting conditions as used in nondestructive inspection and infrared imaging in very low light levels. Videography is especially useful in failure analysis photography to document the disassembly of components, sectioning of components, and cleaning of components and fractures. Figure 23 is a digital image resulting from a typical digital videocamera.
Fig. 23 Videocamera photograph of a fatigue fracture

High-speed videography has use in failure analysis photography for the recording of high-speed events during exemplar component testing, impact testing, and fracture testing. High-speed cameras employ either high-speed film recording or high-speed digital technology. High-speed digital cameras have recording speeds of 1000 to 10,000 frames/s and higher recording times of 1 to 4 s. The recording technique is similar to digital video, but, instead of recording image data onto tape, the images are stored in random access memory. The images are typically transferred to a computer by either cable or digital storage media. The lighting requirements for high-speed video are much greater than for normal photography or videography because of the short duration exposure times. A typical digital high-speed video-camera being used in a dynamic testing application is shown in Fig. 24(a), and a resulting high-speed video-photograph is shown in Fig. 24(b).
Fig. 24 High-speed videography. (a) Typical high-speed digital videocamera used in failure analysis photography. (b) High-speed videophotograph of an impact test

Photography in Failure Analysis

Richard H. McSwain, McSwain Engineering, Inc.

Selected References


Chemical Analysis of Metals in Failure Analysis

Debbie Aliya, Aliya Analytical
Chemical Analysis of Metals in Failure Analysis
Debbie Aliya, Aliya Analytical

Bulk Composition Verification

Before the details of the analytical method are discussed, it should be noted that there are few analytical methods that can be properly performed without knowledge of the intended composition. Most of the analytical methods in use today are instrumental methods whose accuracy depends on calibrations based on alloys of similar composition. The response of the instrument to the alloys present is rarely linear with respect to composition. Thus, a general idea of the expected amount of any element is required in order to determine the composition in an accurate and precise manner. Of course, if enough material is provided, the chemist can repeat the test several times, approaching the proper calibration setup each time. However, this is costly in time, and there is still the risk of missing an element that was specified. If the analyst does not trust the chemist to actually perform the test and provide the results obtained, then another chemist should be found. Do not expect a competent analyst to work “blindly” out of a misguided belief that this will produce the “most objective and correct” results.

If the analyst does not know what the exact specified material is, knowing that it is a carbon steel, for example, versus an alloy steel or a stainless steel will be helpful. Obviously, the experienced metals chemist will be able to tell that something is a piece of aluminum or a copper alloy or a piece of steel. More unusual alloys may cause problems. However, for many methods, where individual elements must be run separately, it could be very expensive to test for every element that might be present in any specific alloy within the family of alloys. If the levels of trace elements that are not generally specified or reported on the certification documentation of the producing mill are desired, the analyst must specify what amount is being sought. For example, embrittlement from amounts of tin less than 0.001% can cause a problem, if the tin is in the “wrong places” in the microstructure. Many of the standard or economy methods simply will not detect these small amounts.

Likewise, very high concentrations of elements can cause false readings with many of the common methods. If the chemist has not calibrated with a similar specimen with a higher content than the specimen in question, an analytical error could easily go undetected. Many composition analysis methods must be calibrated with a material not only having a similar amount of the alloying element in question, but also having the same matrix element and a similar concentration...
of alloying element. For example, it may not be a good idea to use a nickel-base superalloy with 15% chromium as a calibration check for an iron-base stainless steel with 15% chromium.

Another important characteristic that the chemist should be made aware of is whether the material has been wrought or is in an as-cast condition. As-cast products usually have more segregation than wrought products, so the sampling needs to be done from more layers of the material than may be acceptable for wrought product.

If the material has been subjected to a surface treatment such as carburization of steel, the chemist should be told, so that an erroneous carbon content based on analysis of layers including the enriched surface layers is not reported. If the material has been subject to a surface treatment such as carburization of steel, the chemist should be told, so that an erroneous carbon content based on analysis of layers including the enriched surface layers is not reported.

When a component or assembly is subject to damage that prevents its continued functioning, those involved often want to know whether the correct material was used to manufacture the subject items. There is a long history of use of verification of composition as “acquitall” of a component that has stopped functioning. This practice is in general based on ignorance of the principles of engineering, which require that a specific loss of functionality be linked to a specific deficiency in the component. On the other hand, lack of conformance to a particular requirement in no way automatically implies a cause of failure. Also, conformance to all specified requirements does not automatically imply that the product was designed and manufactured properly. Despite the fact that determining the bulk composition is not often a key piece of information in a failure analysis, it is often a useful piece of information, in particular, when the analyst is not very familiar with the alloy system in question.

For example, the potential consequences of exceeding impurity limits for zinc and aluminum casting alloys are discussed in various articles in Properties and Selection: Nonferrous Alloys and Special-Purpose Materials, Volume 2 of ASM Handbook. If some of the consequences appear to have happened, it might be useful to perform a composition analysis. However, the analyst needs to be aware that the composition can be altered by service conditions. Particularly where corrosion occurs, some elements are generally removed in greater amounts than others. An example is “dezincification” of brass. Thus, the fact that the composition does not meet the specification today does not prove that it did not meet the specification at the time of manufacture. Great care is required in interpreting results of bulk-composition test results performed on actual products, especially when they have been in service.

Even when no corrosion is involved, the analyst needs to be aware that most listed steel specifications, for example, are based on ladle analysis. The ladle analysis is more representative of the entire heat than the relatively tiny specimen usually used in product composition testing. Most steel is sold based on this ladle analysis. However, it is entirely possible, especially for steel produced by the older rimming or semikilled ingot processes, for the composition to change prior to complete solidification. Even in steel that is produced by a “killed” or “deoxidized” process, chemical reactions can continue in the ladle as the metal is being cast. Although the mills often test multiple specimens from a single ladle, they rarely report more than one of the analysis results on the certification. As long as all the elements remain “in spec,” the entire heat will be sold with the same “ladle analysis.” For carbon and alloy steels and zinc casting alloys in product form, a broader range of allowed concentrations of intentional and residual elements is published by the Society of Automotive Engineers.

Also notable is that many product form metal specifications have other requirements than the composition. Whether a component met intended mechanical property requirements may be much more significant to the failure investigation than the composition, unless the failure was a corrosion failure of a supposedly corrosion-resistant alloy. The complete applicable product specification must be carefully read and interpreted before an out-of-specification composition is “blamed” for the failure.

In addition to corrosion reactions, another source of composition change that happens after the certifying test is performed occurs when intentional carburization is carried out to improve wear resistance. Many small parts, such as self-tapping screws, are virtually impossible to analyze for original base-metal carbon content, since the carbon added during carburizing often penetrates through much of the material. This example illustrates the importance of knowing what the specified processing sequence was for any component subjected to a bulk composition analysis. A microstructure analysis is often more informative.

There are many difficulties in analyzing the base metal composition of manufactured components. Often the component has been painted or plated. The plating or paint must be removed without changing the composition of the underlying metal. While this may be done relatively easily with paint on a heavy steel component by burning off the paint, this burning may alter the composition of a very light gage material. A solvent paint removal process may be required in that situation. Please note that only an experienced metals chemist should recommend proper specimen preparation techniques, as some might be quite dangerous if not performed with adequate knowledge of proper safety precautions. Another example of a difficult plating removal process would be a copper/nickel/chromium plating on a zinc die casting. Since the zinc is more soluble in acid than any of the plating layers, another approach is obviously required, such as sanding or drilling. Obviously, the odd shape of many products makes sampling difficult at best. Small specimens that have had part of their composition altered by carbon or nitrogen diffusion or that have been plated or painted tend to be more problematic than larger pieces of structural alloys.

Some analysts remelt and chill cast their metals specimens in order to get a disk large enough to analyze by the faster and thus more economical test methods, such as optical emission spectroscopy (OES) or glow discharge spectroscopy. As long as the chips or bits that are remelted are truly representative of the base metal (i.e., plating or other surface coating
has been removed, no carburized or decarburized material is unintentionally included, no chunks that have been altered by corrosion processes have been included, etc.) and the process is performed properly in a protective atmosphere, this can be a legitimate technique. However, many things can go wrong in this process, especially if the chemist is not familiar with the details of metal processing methods and service conditions that may have altered the composition. In fact, any tool that is used to machine chips that are analyzed can actually contaminate the specimen. Extreme care by an experienced metals chemist is the minimum requirement for important analysis of metals composition in failure analysis work. If the milling to obtain chips for analysis is done in a machine shop where other machining is done, it may be necessary to set up a protective “tent” (for example, out of aluminum foil) around the chips being collected, to prevent contamination by airborne chips from other machines.

Economy Methods. The economy methods of metals composition analysis are usually methods that can be preformed on a “small chunk” of the metal in question. The economic benefits of these types of methods include the fact that no time is spent milling, drilling, weighing, and dissolving the specimen. As long as a clean flat area is available on the surface of the part, analysis may proceed after a quick grinding operation on a special sanding belt. The need for the special sanding belt is because the abrasive grit must not contain any of the elements that are going to be analyzed. For example, a zirconium oxide abrasive is often used to prepare steels for this type of analysis, because zirconium is rarely specified in steels. Thus, if a small piece of zirconium oxide does get embedded in the metal, it will not affect the analysis. The actual size of the “small chunk” required varies with the particular test method. A common requirement for an optical emission spectrophotometer (OES) is a 6 mm (\(\frac{1}{4}\) in.) diameter, flat-ground spot on a specimen at least 1.6 mm (\(\frac{1}{16}\) in.) thick. Thinner material may burn through during the test, and a smaller flat area will not properly seal the vacuum chamber where the emitted light waves are analyzed. A common requirement for a glow discharge spectrophotometer (GDS) is a 13 mm (\(\frac{1}{2}\) in.) diameter, flat-ground spot on a specimen at least 1.6 mm (\(\frac{1}{16}\) in.) thick. These analytical methods are quick to perform, especially once the instrument has been calibrated for the type of alloy in question. The main drawback to these methods is that most metals do not have a completely uniform composition from surface to core. Most metals are subject to some degree of segregation during solidification. Thus, sulfur is more likely to be concentrated in the core of a piece of steel, while many high-temperature-forming processes deplete carbon at the surface of steels. While the differences may not be great in modern wrought product, cast products can be very heavily segregated. Extreme care is required when interpreting such results.

Referee Methods. Most referee methods are either “wet chemical” or “combustion” methods. In these methods, the referee designation is usually due to the fact that the sampling process allows a more representative specimen to be analyzed. The accuracy of the “economy methods” described previously is very good on a properly maintained and calibrated piece of equipment. However, the methods performed on solid specimens can only measure the layers available. A specimen properly sampled for wet chemical or combustion analysis will contain all the layers of the material, from surface to core, except when the surface layers are specifically intended to be discarded. This might be the case for a heavy casting or forging.

The failure analyst should remember that the composition at the surface may be of more interest than that of the bulk. For example, if a leaf spring has fractured in fatigue, it may be valuable to analyze the outer 0.1 mm (0.004 in.) for carbon separately by combustion, even if the bulk composition is analyzed with an economy method. Again, even a properly performed standard-certification verification can be very misleading to an inexperienced failure analyst. Common wet chemical analysis methods for metals include atomic absorption, where a beam of light is directed through a solution of known concentration prepared by dissolving metal chips in the proper acid mixture. The light is absorbed preferentially at certain wavelengths by certain elements. Thus, the more of a certain element present, the less light will get through to the detector. Again, proper calibration is extremely important. Another method of wet chemical analysis is the inductively coupled plasma spectrometer. Again, a solution of dissolved metal is fed into the instrument. The solution is dispersed and heated to plasma temperatures. The metal atoms, excited by the heat, emit light waves that have known correspondences to particular elements.

A drawback of all wet methods is that if there are microconstituents that are not easily soluble in the commonly available acids, erroneous results can easily occur. It is not the instrument that is at fault, but the fact that the method used to dissolve the metal was inadequate. These types of analysis errors may be very difficult to see. Many tool steels and other highly alloyed corrosion resistant materials are difficult to analyze for this reason. Again, if correct results are important, the analyst should choose a chemist who has proven competency in analyzing the particular alloy in question. One way of demonstrating competency is participation in the “round robin” programs administered by the National Institute of Standards and Technology or the Collaborative Testing Service.

Please note that the competent chemist can take advantage of this usually problematic differential solubility to gain additional information about some materials. For example, aluminum is often added to steel to create conditions conducive to maintaining a fine grain size, even in the heat-affected zones of welds. This effect is created because aluminum nitrides keep the grain growth to a low level, even during exposure to the relatively high temperatures encountered during welding. However, aluminum is very highly oxidizable at molten steel temperatures, so it often
happens that some or much of the aluminum added to the steel ends up as an oxide rather than a nitride. The aluminum oxides are useless as grain-size controls. Aluminum oxides are not easily soluble in acids, while the aluminum nitrides are easily soluble. Thus, running the test for aluminum twice, once using an “economy” method that measures all of the aluminum (if the aluminum is not macrosegregated) and once using a wet method on a solution that has been strained to remove the insoluble aluminum oxides, can provide information on why an apparently “conforming” aluminum-killed steel has a coarsened grain structure.

Another referee composition method, used particularly often with ferrous alloys, is one of the various combustion methods. Here, a known quantity of solid chips of the metal is heated in a crucible while an oxygen-containing gas is passed over the crucible. The oxygen combines with carbon and sulfur present in the metal, forming carbon dioxide, carbon monoxide, sulfur dioxide, and so on. Some sort of detection device is used to, in effect, “count” the molecules of compound gases. These numbers are (generally, with modern instruments) automatically converted to a percent of carbon and/or sulfur in the original metal specimen.

Alternatives to Economy and Referee Methods. All of the methods described so far are in some manner destructive. When a less destructive method is required or when a failure investigation of an in-process incident reveals that the material in house has been “mixed” with an unspecified material, spot-check test kits may be useful. Most of these tests have limited ability to distinguish certain grades from other somewhat closely related grades. For example, a drop of potassium thiocyanate/stannous chloride solution will turn pink if placed on 316 stainless steel, which contains molybdenum, while it will not have any appearance change if placed on 304 stainless, which has no molybdenum. Other examples are also shown in Table 1 and in Ref 1 and 2. Kits containing premixed solutions are available from specialty companies.

<table>
<thead>
<tr>
<th>Element tested for</th>
<th>Reagent</th>
<th>Positive result</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromium</td>
<td>Diphenylcarbazide</td>
<td>Violet color</td>
<td>Identify chromium-free steel</td>
</tr>
<tr>
<td>Cobalt</td>
<td>Ammonium thiocyanate/acetone</td>
<td>Blue color</td>
<td>Identify cobalt-base alloys</td>
</tr>
<tr>
<td>Copper</td>
<td>Dithizone</td>
<td>Purple color</td>
<td>Sort copper-bearing stainless steels</td>
</tr>
<tr>
<td>Iron</td>
<td>Potassium ferricyanide</td>
<td>Blue precipitate</td>
<td>Sort low-iron high-temperature alloys</td>
</tr>
<tr>
<td>Lead</td>
<td>Sulfuric acid</td>
<td>White precipitate</td>
<td>Sort leaded bronze</td>
</tr>
<tr>
<td>Molybdenum</td>
<td>Potassium thiocyanate/stannous chloride</td>
<td>Pink color</td>
<td>Sort molybdenum-bearing stainless steels</td>
</tr>
<tr>
<td>Nickel</td>
<td>Dimethylglyoxime</td>
<td>Red precipitate</td>
<td>Sort 300 series from 400 series stainless steel</td>
</tr>
</tbody>
</table>

Hydrogen Embrittlement. On the subject of bulk chemical analysis of metals, hydrogen analysis in steels can be important in parts suspected to have broken or cracked because of hydrogen embrittlement. While reliable methods are becoming more available, the analysis of hydrogen in high-strength steel in the extremely low concentrations often associated with hydrogen embrittlement is still very costly. In addition, the amount of hydrogen in the steel is usually not of any particular significance. It is generally the atomic (“nascent”) hydrogen that is present in the grain boundaries that causes the biggest problems. The hydrogen analysis methods available cannot distinguish between hydrogen atoms and hydrogen molecules and cannot distinguish between hydrogen within the grains and hydrogen at the grain boundaries. In addition to this, the hydrogen that has actually weakened the grain boundaries of a broken part has probably dissipated into the atmosphere by the time it gets to the chemistry lab. Thus, for routine investigations, hydrogen analysis is unlikely to provide much useful information. Hydrogen embrittlement diagnosis is better confirmed by other methods, including microstructure analysis to preclude other more observable causes of brittle intergranular cracking and confirmation that the cracking was delayed with respect to the application of stress or was caused by a sustained stress. Auger analysis may be used in bigger investigations of more costly failures, to preclude embrittlement of grain boundaries by other low melting point elements, such as sulfur, antimony, tin, and phosphorus. Any of these tests would probably shed more light on the causes of fracture than hydrogen analysis of a broken steel object.

Procuring a Chemical Analysis. If a piece of metal requires analysis and it is important to determine whether the potential chemist has the expertise needed, it may be necessary to ask some questions before selecting the chemist. The key items that should be discussed are the specimen size required, turnaround time, the cost, the accuracy and precision of the analysis and whether the chemist has a proven “digestion” method for any wet analysis being contemplated. The OES and GDS methods generally require a 6 to 13 mm (1/4 to 1/2 in.) diameter flat surface. The common wet analysis methods generally require about 0.5 to 2.5 g (0.02 to 0.09 oz) of material. If the instrument to be used for the analysis is not calibrated with a specimen of the same matrix as the sample to be analyzed, great errors can occur. This is of particular
concern for wet chemical analysis, where the specimens are put into solution before being analyzed. For example, a 15% solution of chromium in water may give a different peak height at the emission wavelength than a 15% solution of chromium with 84% iron would. This has obvious applications for analysis of stainless steels. This type of difficulty, as well as difficulties in getting certain types of second phase particles into solution, are probably two of the biggest concerns that the failure analyst should have when procuring bulk chemical analysis, especially of the less common alloys.

If services are procured from an outside company, the analyst also needs to determine how long the test will take and how much it will cost. For example, consider a hypothetical failure of a heavy duty industrial fan that is supposed to have been fabricated from 316 stainless steel. The analyst calls a laboratory and asks if the lab can verify the composition. The first question may be how large a piece to send. What are the dimensions of the specimens? The second question is why is this analysis requested. Does the analyst just want to know if it meets the specification for 316 stainless? Or does the analyst need to know an actual, very precise average value for some of the elements? In this regard, several possible alternatives might be considered. For example, one alternative is that the purpose is just to know if it is a 316 stainless steel. In this case, the lab representative may suggest analysis by GDS with results available in two days, once the specimen is received.

However, if the analyst is suspicious about the strain-hardening behavior and wants to know the representative analysis on the nickel content, then additional testing may be desirable. In this case, the lab may suggest GDS testing and also inductively coupled plasma (ICP) testing to double-check the nickel content. This additional analysis, of course, will cost more and may take more time. Additional cost and time may also be required if ICP testing is performed for additional elements in addition to nickel. For example, if alloying content requires closer scrutiny, then the lab may suggest an ICP test with carbon and sulfur tested with a combustion method. This may take five days to get results, once the specimen is received.

The preceding example is not intended to show how much a particular cost or turn around time should be for a particular test. The intent is to indicate the trend of increasing cost and time for the more definitive methods and to show what subjects should be covered if both parties are to have a good chance of being satisfied by the transaction. Laboratories vary widely in how much they charge and how long they take to analyze a component. Some labs may run some alloys only once a week, for example, because of the time it takes to calibrate the instrument. They prefer to calibrate for fifteen similar alloy specimens than to calibrate five times for three similar alloy specimens.

Format of a Bulk Metal Composition Analysis Report. The following table is an example of a suggested format of a bulk metal composition analysis report. Note that the chemist has flagged (with an asterisk) the element (in this case, Mn) that does not meet the given specification:

<table>
<thead>
<tr>
<th>Element (symbol)</th>
<th>Actual % (by weight)</th>
<th>Specified range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbon (C)</td>
<td>0.43</td>
<td>0.37–0.43</td>
</tr>
<tr>
<td>Manganese (Mn)</td>
<td>0.57*</td>
<td>0.35–0.55</td>
</tr>
<tr>
<td>Phosphorus (P)</td>
<td>0.007</td>
<td>0.010 max</td>
</tr>
<tr>
<td>Sulfur (S)</td>
<td>0.014</td>
<td>0.025 max</td>
</tr>
<tr>
<td>Silicon (Si)</td>
<td>0.17</td>
<td>0.15–0.30</td>
</tr>
<tr>
<td>Aluminum (Al)</td>
<td>0.0015</td>
<td>Not specified</td>
</tr>
<tr>
<td>Chromium (Cr)</td>
<td>1.32</td>
<td>1.15–1.40</td>
</tr>
</tbody>
</table>

Many labs will double-check the concentration of an element that falls outside the specified range before reporting that the material does not meet the specification. Some laboratory accreditation boards require that the element be checked with a different test method the second time, if one is available. Note also that the elements are listed in the order that they would be listed in the applicable specification. This makes it easy to follow. Many chemical analysis laboratories follow the less-easy-to-interpret practice of listing the elements in the order in which they were tested or in an apparently random order.

When evaluating composition, it should be noted that standard specifications permit a composition range, or in some instances a maximum value for a given element. Slightly exceeding nominal limits for a given element may not be as critical a potential root cause as to exceeding the limit of a maximum value element, for example, phosphorus.

References cited in this section


Microchemical Analysis in Failure Analysis

Often failure analysis work requires microchemical analysis to find evidence of contamination, to evaluate the composition of microphases revealed by a metallographic specimen or at a grain boundary exposed by a fracture surface, or to evaluate corrosion products. Important information on specimen handling is provided at the end of this article. Failure to properly protect the specimens at all phases of the failure analysis can make correct interpretation of the results of this type of test impossible.

Backscattered Electron Imaging. Most scanning electron microscopes have a useful feature called backscatter electron mode. While the secondary electrons, the normal imaging mode, provide topological information about the specimen, the backscatter mode gives information about topology as well as composition. But the user should not switch over entirely to backscattered images. The images are less sharp than those that can be obtained by the secondary imaging. It is useful to use the two imaging modes to show differences between unknown materials. Figure 1 shows some flaky metallic contaminant that was of an unknown composition. Particle brightness appears similar in the image from the secondary electron imaging (SEI) mode of Fig. 1(a), while there are clearly two categories of particles (brighter and darker) in the image from the backscatter mode in Fig. 1(b). The electrons impinging on the specimen interact with the specimen in several ways. Higher-atomic-number elements “reflect” or backscatter more electrons than lower-atomic-number elements. This makes higher-atomic-number–rich areas of a specimen appear brighter.

![SEI and Backscatter Images](image.png)

**Fig. 1** SEM image of flaky metallic contaminants determined to be a mix of steel and aluminum particles by backscatter mode image. (a) Secondary electron imaging mode where particles are not distinguished by composition. (b) Image from backscatter mode, which shows the lower atomic number particles (in this case aluminum 1, 2, 3, 4, 5, 6, and 8) as darker than the iron particles (3A and 7). Note also that the fibers seen in (a) are invisible in this backscatter imaging mode.

The SEM backscatter mode is often useful for rough composition analysis. The backscattered image in Fig. 1(b) provides a fast and economical method of getting a general idea of how many particle types are present. The actual elemental composition was determined with the EDS microchemical analysis feature for several particles independently. The subsequent EDS analysis showed that the fibers (which are invisible in the backscatter image) were mainly composed of carbon and oxygen, with significantly lower atomic numbers than the aluminum (dark flakes) or the steel (bright flakes). The particles and fibers in both photos are held in place by a special double stick tape whose only EDS detectible element is carbon.

Energy-Dispersive Spectrometry Microchemical Analysis. The EDS feature of a scanning electron microscope is an even more powerful tool for microchemical analysis work. The EDS method works when electrons from the SEM analyzing...
beam knock electrons out of their orbits around the atom. Electrons that are in higher-energy orbitals have a tendency to fill in the empty orbitals. The excess energy that cannot be carried into the closer orbital is released from the atom by means of a photon. Many different elements emit photons of energy levels detectable by the EDS detector.

Some elements emit photons of several particular energy levels. Figure 2 is a typical scan of a piece of some organic material. The vertical scale represents the number of photons detected during the analysis time. This scan was obtained using an EDS unit with a “low element” detector. The scan has also been corrected to remove the background. Many older EDS systems leave a differently shaped curve, even if no detectable elements are present. The use of the low element detector allows the carbon and oxygen to be detected. Zinc and sodium are indistinguishable in trace amounts with this test method, so the analyst has labeled the peak in question with both possibilities. The palladium (Pd) was used to sputter coat the specimen prior to putting it in the SEM to make it conductive and suitable for analysis. Other people use graphite or gold. Graphite has the obvious limitation that if you are looking for carbon, it confuses the issue. Gold also has many peaks that are at the same positions as other elements that are often of interest to the analyst. The unlabeled peak at the far left is the internal calibration peak for the EDS unit.

Fig. 2 Typical EDS scan of a polymeric hydrocarbon material

Most EDS units now available have computers to help the analyst determine the elements present, based on the energy levels of the emitted photons. Examination of the scan illustrates a significant point. Energy-dispersive spectrometry analysis only tells what elements are present. It gives no definitive information about molecule types. Hydrogen is not detectable at all. This means that polyethylene, polypropylene, and polystyrene will all look the same as paraffin and many liquid oils, because only the carbon will show up.
If photons of more than one energy level are emitted by a given type of atom, then the ratio of the number of photons at each energy level is constant, as long as the test conditions (in particular, accelerating voltage) are constant. Figures 3(a) and (b) show the scans for the bright and dark flakes shown in Fig. 1(b). The scans match to iron and aluminum, respectively. In both plots, the vertical scale is exaggerated to show the trace elements more clearly, and thus the tops of the iron peak (Fig. 3a) and aluminum peak (Fig. 3b) are cut off. The darker flakes have peak emission rate at an energy level that may indicate the presence of bromine or aluminum. Since bromine is not usually solid, aluminum is the likely contributor. In addition, bromine has a second peak at another energy level, and because no second peak was found, even when analyzed at high accelerating voltages, the flakes appear to be aluminum.

![Fig. 3 EDS scans from particles detected in Fig. 1 showing composition differences. (a) Portion of EDS scan obtained from one of the (bright) iron based particles in Fig. 1(b). Some traces of carbon are also seen here, probably an oily residue on the flake. The silicon may be due to a dirt contaminant. (b) Portion of the scan obtained from one of the (dark) aluminum based particles. The unlabeled peak at the far left of both plots is the internal calibration peak for the EDS unit.](image)

If small peaks are being sought while additional elements are present, especially if corrosion products are involved, it may be difficult to decide whether the peak is due to aluminum or a trace of bromine that may have contaminated the specimen. This is because the smaller peaks at the higher energy levels may not protrude above the baseline electronic noise. This indicates another important point about EDS analysis. It is not foolproof. The computer will give you an answer to every question that you ask it. However, you must have some understanding of chemistry to be able to interpret the results. “Murphy's Rules of Microchemical Analysis,” a humorous attempt to prevent the beginner from making some of the more tempting errors of inexperience, states, “The probability of detecting argon decreases with the experience of the analyst.” Since the traditional SEM and EDS units operate in a relatively high vacuum, it is indeed unlikely that argon will be present. Yet the computers list its emission energy at 2.957 keV.

Figure 4(a) shows a typical scan obtained from a piece of steel. Note that the main element present is iron with three peaks. There is also a trace of carbon detected, probably from oil on the surface of the part. There is no nitrogen present at a detectable level, and the location where it would have been found was labeled to emphasize this in Fig. 4(a). A small trace of manganese is sometimes detectable with the EDS but was not detected in this case. While most steels have specified levels of carbon, sulfur, phosphorus, silicon, and other elements, the EDS cannot usually detect these elements, as they are too dilute.
Fig. 4  Plot from EDS scan of low carbon steel sheet at (a) 15 keV and (b) 5 keV. The high energy iron peaks (above 5 keV) are missing in the spectrum in (b) produced from 5 keV electrons. The carbon peak is also higher in Fig. 4(b), suggesting a trace of carbon, probably from oil, on the surface of the part. The unlabeled peak at the far left is the internal calibration peak for the EDS unit.

Figure 4(b) shows another scan obtained from the same piece of steel but with a lower accelerating voltage for the electrons in the SEM. Note that the height ratios of the iron peaks have changed. This illustrates another important feature of EDS analysis. It is necessary to know what the analysis conditions are. In this case, the difference between the two scans is accelerating voltage of the impinging electrons. A voltage of 15 keV is a common test mode for high quality imaging on electrically conductive materials. The scan in Fig. 4(b) was produced with an accelerating voltage of only 5 keV. Note that the two larger energy peaks at approximately 6.3 keV are absent. A 5 keV electron can never allow a 6.3 keV photon to be emitted. Only the lower energy photons at 0.70 keV can escape from the specimen.

The lower accelerating voltage also means that the electrons will not penetrate as deeply. So if the layers nearest the surface are of interest, lower accelerating voltages may be helpful. For example, very thin oxides may be undetectable at 15 keV, because most of the signal will come from the subsurface layers. If the oxide layer is only 0.1 μm thick, it will be much easier to detect when the beam penetrates to approximately 0.4 μm than when it penetrates to 1 μm.

It is also important to point out that EDS analysis is considered to be a semiquantitative method. Part of the reason for this is the fact that the light elements are either not detectable at all or not detectable unless they are present in very concentrated amounts. Most of the atom types under atomic number 22 (titanium) need to have at least a few percent in the area being penetrated by the analyzing beam in order to produce a detectable signal. Heavier atoms may be detectable in lower concentrations, depending on what other elements are present. So if the specimen is oil with some fine bits of steel along with some suspended silica particles, carbon, iron, silicon, and oxygen may be detected if a modern “low-element” detector is used. If a less modern detector is used, only iron and silicon will be detected.

In any case, even the semiquantitative analysis is only valid for the heavy elements. In this case, the analysis might give iron 85% and silicon 15%. The detector will force the total to add to 100%. Even this ratio can be useful in some cases. However, the bigger problem with interpreting semiquantitative-analysis results from an EDS is that the specimen size is unknown. It is not a trivial matter to know exactly how deeply the electrons that are causing photon emission are penetrating. The shape of the volume from which the photons are being emitted is not easy to determine either. And, of course, unobserved subsurface second-phase particles may also contribute to the EDS signal.

Microscopically flat and homogeneous specimens can give quite good quantitative results with EDS. Such specimens are rare in failure analysis work. However, analysts working with a limited budget and who are doing SEM work for other reasons may find it informative to run a semiquantitative analysis on an available smooth and clean area of the component adjacent to the fracture. This information could provide a clue, for example, to distinguish 304 stainless from 316 stainless steel, without the added expense of a proper bulk chemical analysis. Especially in routine investigations where the requesting party has documentation showing that the proper material was used, this semiquantitative “microchemical” method can substitute for an economy bulk chemical analysis.

For this reason, analysis of suspected contaminants contained within a crumbly corrosion product on a fracture surface should generally be confined to the strictly qualitative mode. Interpreting the quantitative results from such a test would be very difficult at best.
Wavelength-Dispersive Spectrometry Microchemical Analysis. In modern x-ray microanalysis, the analyst actually has a second (less common in commercial laboratories) choice for the x-ray spectrometer: the wavelength-dispersive spectrometer, or WDS. The main difference between the EDS and WDS microprobe techniques is the method of detecting the x-rays. The original design by Castaing utilized a wavelength-dispersive spectrometer and was the model for the first commercial systems (see the article “Electron Probe X-Ray Microanalysis” in Materials Characterization Volume 10 of ASM Handbook. These systems consisted of an x-ray detector and single-crystal diffractometers that moved synchronously around the sample to detect the various wavelength x-rays (Fig. 5). This configuration is excellent for the quantitative analysis of flat samples but is awkward and time-consuming to set up and is difficult to use on irregular topography, such as a fracture surface. With the advent of semiconductor devices, a new type of detector was developed that, with no moving parts, could simultaneously count and determine the energy of the x-rays (Fig. 6). These detectors are the ones used for EDS modules on SEM devices.

Fig. 5 Schematic diagram of the components of a wavelength-dispersive x-ray spectrometer
Even though the EDS method is more common, the wavelength dispersive spectroscopy (WDS) method has utility in certain applications because of its ability to distinguish certain elements that are difficult in some cases to distinguish with the EDS method. In particular, sulfur and molybdenum are difficult to distinguish with EDS, as the positions of their peaks are so close to each other. With WDS, the peaks are quite discernible. Differences between EDS and WDS capabilities are described in more detail in the article “Electron Probe X-Ray Microanalysis” in Materials Characterization Volume 10 of ASM Handbook.

Specimen Handling Information. The methods discussed in this section on microchemical analysis are often used as surface chemical analysis methods. This is because the same instrumental techniques often deal with microscopic features, trace chemical contamination, and/or thin surface films. Therefore, the analyst must always be concerned with sample handling to be sure that fragile features are not altered before critical information about the surface is obtained.

Specimen collection and storage methods must be planned to prevent the possibility of physical damage or chemical contamination and to ensure that sample integrity is preserved. If specimens for analysis are removed from a larger object, the method for removal should not allow excessive heating or contamination of the specimen area. Cutting without coolants at a distance that does not heat the sample is critical for surface analysis samples, as well as for those for metallurgical analysis. If this is not possible or convenient, acetate tape softened with acetone may be used to remove partially adherent contaminant prior to sectioning in a standard wet metallurgical (abrasive) cutoff saw.

Specimens should be stored in clean containers. This seems an obvious caveat, but even new containers can contain residual chemical contaminant that could be a confusing factor in the failure analysis. For example, mold release agents in new plastic bags can be transferred to a sample and detected later by chemical surface analysis. In addition, specimens should be stored in a cool dry environment to prevent superficial corrosion or other surface contamination prior to analysis.

Even the handling practices for routine fracture analysis work must be considered sources of possible damage to microscopic fracture features or surface chemistry. Poking or probing of fracture surfaces with fingers or tools should be avoided to prevent damage to microscopic features. Furthermore, although it seems to be human nature to reassemble the
pieces created by a fractured component, analysts (and others) must resist the urge to do so. Critical microscopic fracture features can be completely obliterated by this simple act, even on hardened alloy steel components.

Poor sample handling practices are often a problem in surface chemical analysis. Surface sensitive chemical analysis techniques can analyze thin layers only a few atoms thick. Chemical treatments commonly applied by manufacturers to the surface of metal components, such as those used for “corrosion protection,” will prevent surface analysis methods from being an effective tool. Oils and perspiration from touching the surface with bare hands (Fig. 7) or spittle (Fig. 8) contaminate surfaces such that accurate analysis for failure evaluation cannot be performed. Residue from adhesive tapes also adds substantial artifact contamination to a surface, which can confound the failure analysis when the chemical nature of the surface is important.

![Fig. 7](a) SEM image and (b) EDS spectrum (10 keV) for fingerprint residue on a clean steel surface. Sometimes other elements, from perspiration, such as chlorine, or potassium and sodium may also appear in fingerprints.

![Fig. 8](EDS spectrum (10 keV) for analysis of residue from dried human spittle)
It is essential to plan for surface evaluation early in the failure analysis and consider the effect of prior specimen preparation or analyses on possible later tests. Some surface evaluation techniques can cause damage to a surface that will affect the results of subsequent analyses. For example, heat from high intensity lights used for visual examination, photography, or light microscopy can drive off volatile chemical residues from the surface or distort a polymer component. The concentrated energy from an electron beam can distort or melt polymeric materials during SEM examination.

A less obvious example is that examination of a sample by SEM can result in a thin film of hydrocarbon contamination (Fig. 9), as well as a shallow layer of radiation damage that will break chemical bonds in the material struck by the electron beam. These two factors will prevent accurate identification of chemical compounds on the surface by subsequent surface chemical analyses. Therefore, consideration of potential subsequent analyses of a surface should be considered before routinely jumping directly to SEM examination.

Fig. 9 SEM image showing carbon residue buildup from the electron beam scan area in the SEM

Footnote

* *Larry Hanke, Materials Evaluation and Engineering, Inc., provided this section on specimen handling.

Chemical Analysis of Metals in Failure Analysis

Debbie Aliya, Aliya Analytical

References


Characterization of Plastics in Failure Analysis

Jeffrey A. Jansen, Stork Technimet, Inc.

Introduction
THE ULTIMATE OBJECTIVE of a failure analysis is to ascertain the mode and the cause of the failure, regardless of the material from which the part was fabricated. The investigation is performed in generally the same manner, whether the failed component was produced from metal or plastic or a combination of these materials. Thus, the general steps required to conduct a comprehensive failure investigation are the same, and these are outlined in Fig. 1. In general, the failure analysis process is analogous to putting together a jigsaw puzzle. A failure analysis requires assembling bits of information into a coherent and accurate portrayal of how and why the part failed. Reaching the objectives of the plastic failure analysis, namely, the determination of the mode and cause of the failure, or expressed alternatively, evaluating how the part failed and why it failed, requires a scientific approach and a broad knowledge of polymeric materials.
Plastic components can fail via many different modes, including catastrophic mechanisms, such as brittle fracture, ductile overload, creep rupture, environmental stress cracking, molecular degradation, and fatigue. In the case of failure involving fracture, the determination of the failure mode involves identifying how the crack initiated and how it subsequently extended. This is usually ascertained using a number of visual-based techniques, such as stereomicroscopy, scanning electron microscopy (SEM), and the preparation of mounted cross sections. Noncatastrophic failure modes are also relevant, and these include discoloration, distortion, and contamination. Assessing the mode of the failure is often not as difficult as establishing why the part failed. Evaluating why the part failed usually requires analytical testing beyond the visual-based techniques. In many cases, a single cause cannot be identified, because multiple integrated factors may have contributed to the failure. All of the factors that affect the performance of a plastic component can be classified into one of four categories: material, design, processing, and service conditions (Ref 1). These factors do not act independently on the component but instead act in concert to determine the performance properties of a plastic component. This is represented graphically in Fig. 2 (Ref 1).

The principal differences between how failure analyses are performed on metal and plastic materials center on the techniques used to evaluate the composition and structure of the material. Unlike metals, polymers have a molecular structure that includes characteristics such as molecular weight, crystallinity, and orientation, and this has a significant impact on the properties of the molded article. Additionally, plastic resins usually contain additives, such as reinforcing fillers, plasticizers, colorants, antidegradants, and process aids. It is this combination of molecular structure and complex formulation that requires specialized testing (Ref 2). While the chemical composition of a failed metal component can often be evaluated using a single spectroscopic technique, a similar determination requires multiple analytical approaches for a part produced from a plastic resin.

This article reviews those analytical techniques most commonly used in plastic component failure analysis. The description of the techniques is not designed to be a comprehensive review and tutorial but instead is intended to make the reader familiar with the general principles and benefits of the methodologies. The descriptions of the analytical techniques are supplemented by a series of case studies. The technique descriptions refer to the case studies, and the two are written in a complimentary manner to illustrate the significance of the method. The case studies also include pertinent visual examination results and the corresponding images that aided in the characterization of the failures.
Characterization of Plastics in Failure Analysis

Jeffrey A. Jansen, Stork Technimet, Inc.

Fourier Transform Infrared Spectroscopy

Fourier transform infrared spectroscopy (FTIR) is a nondestructive microanalytical spectroscopic technique that involves the study of molecular vibrations (Ref 2). The analysis results provide principally qualitative, but also limited quantitative, information regarding the composition and state of the material evaluated. Fourier transform infrared spectroscopy uses infrared energy to produce vibrations within the molecular bonds that constitute the material evaluated. Vibrational states of varying energy levels exist in molecules. Transition from one vibrational state to another is related to absorption or emission of electromagnetic radiation (Ref 3). These vibrations occur at characteristic frequencies, revealing the structure of the sample. Fourier transform infrared spectroscopy produces a unique spectrum, which is comparable to the fingerprint of the material. It is the principle analytical technique used to qualitatively identify polymeric materials.

Method

Several different sampling techniques, all involving either transmission or reflection of the infrared energy, can be used to analyze the sample material. This allows the evaluation of materials in all forms, including hard solids, powders, liquids, and gases. Depending on the spectrometer and the corresponding accessories, most samples can be analyzed without significant preparation or alteration. In the analysis of polymeric materials, transmittance, reflectance, and attenuated total reflectance (ATR) are the most common sampling techniques. Additionally, a microscope can be interfaced with the spectrometer to focus the infrared beam and allow the analysis of samples down to 10 μm. Regardless of the sampling technique, the beam of infrared energy is passed through or reflected off of the sample and directed to a detector. The obtained spectrum shows those frequencies that the material has absorbed and those that have been transmitted, as illustrated in Fig. 3. The spectrum can be interpreted manually, or more commonly, compared with voluminous library references with the aid of a computer.
The results generated through FTIR analysis are referred to as an infrared spectrum. The spectrum graphically illustrates the relative intensity of the energy absorbed on the y-axis versus the frequency of the energy on the x-axis. The frequency of the energy can be represented directly in microns (μm) or, more popularly, as reciprocal centimeters (cm\(^{-1}\)) referred to as wavenumbers. The discrete spectral features present with a FTIR spectrum are known as absorption bands.

Uses of FTIR in Failure Analysis

Material Identification. Possibly the most important use of FTIR in failure analysis is the identification of the base polymer used to produce the sample. The determination of the composition of the failed component is an essential part of the investigation. Because different polymers have a wide variation in their physical, mechanical, chemical resistance, and aging properties, the use of the wrong resin can yield detrimental results in many applications. Fourier transform infrared spectroscopy is well suited for the identification of polymers having different molecular structures, and this is illustrated in Fig. 4. Confirming that the failed article was produced from the specified material is the primary consideration of the failure analyst in assessing the cause of the failure. Thus, FTIR is often the first analytical test performed during a plastic failure analysis. The use of FTIR in characterizing the composition of the plastic-resin base polymer is illustrated in examples 1, 4, 7, and 9 in this article. One area where FTIR is inadequate is in differentiating between polymers having similar molecular structures, such as the members of the nylon family, and poly(ethylene terephthalate) and poly(butylene terephthalate). In these cases, other techniques, such as differential scanning calorimetry, must be used to augment the FTIR results.
Fig. 4 Fourier transform infrared spectral comparison showing distinct differences between the results obtained on various plastic materials

Aside from the determination of the base polymer, FTIR is used to characterize other formulation constituents. Fourier transform infrared spectroscopic analysis can provide information regarding the presence of additives and filler materials. Due to the nonlinearity of infrared absorptivity of different molecular bonds, it is not possible to accurately state minimum concentration detection limits. However, it is generally considered that materials present within a compounded plastic resin at concentrations below 1% may be below the detection limits of the spectrometer. Given this restriction, it is likely that most major formulation additives, such as plasticizers, can be characterized, while low-level additives, including antioxidants, may go undetected. Given that FTIR is principally used for the analysis of organic materials, its use in the evaluation of inorganic filler materials is somewhat limited. However, some commonly used fillers, such as calcium carbonate, barium sulfate, and talc, produce unique, identifiable absorption spectra. Example 6 in this article shows the analysis of plastic-resin formulation constituents.

Contamination. Similar to its ability to identify the plastic formulation constituents, FTIR is extremely useful in the determination of contaminant materials within the failed part material. While contamination is never an intended part of a plastic compound, its presence certainly can have a tremendous impact on the properties of the molded component. Through the electronic manipulation of the obtained FTIR results, including spectral subtraction, extraneous absorption bands not attributed to the base resin can be used to characterize contaminant materials. Fourier transform infrared spectroscopy is useful in the identification of contaminant material, whether it is mixed homogeneously into the resin or present as a discrete inclusion. The role of FTIR in the identification of contaminants is discussed in examples 3 and 8 in this article.

Degradation. Fourier transform infrared spectroscopy is a valuable tool in assessing a failed component material for degradation, such as oxidation and hydrolysis. Molecular degradation, often involving molecular weight reduction, has a significant detrimental impact on the mechanical and physical properties of a plastic material. This degradation can result from several stages in the product life, including resin compounding, molding, and service. As a polymeric material is degraded on a molecular level, the bonds comprising the material are altered. Fourier transform infrared spectroscopy detects these changes in the molecular structure. While FTIR cannot readily quantify the level of degradation, it is useful in assessing whether the material has been degraded and determining the mechanism of the degradation. Specifically, several spectral bands and the corresponding molecular structure can be ascertained, including carbonyl band formation.
representing oxidation, vinylidene group formation as an indication of thermal oxidation, vinyl, vinylene functionality for photooxidation, and hydroxyl group formation indicating hydrolysis (Ref 4). Case studies showing the effectiveness of FTIR in assessing molecular degradation are presented in examples 1, 13, and 15 in this article.

Chemical Contact. Parallel to the application of FTIR in addressing polymeric degradation, the technique is also useful in evaluating the failed sample material for chemical contact. Plastic materials can be affected in several ways through contact with chemical agents. Depending on the polymer/chemical combination, solvation, plasticization, chemical attack, or environmental stress cracking can occur. In the case of property alteration through solvation or plasticization, FTIR can be helpful in identifying the absorbed chemicals. Because these chemicals are present within the failed plastic material, the likelihood of distinguishing the agents is high. Based on the observed spectral changes, mechanisms and chemical agents producing chemical attack, including nitration, sulfonation, hydrolysis, and aminolysis, can be detected (Ref 4). Environmental stress cracking, the synergistic effect of tensile stress while in contact with a chemical agent, is one of the leading causes of plastic failure. The chemical agent responsible for the cracking may be identified using FTIR. However, given that such materials are often volatile organic solvents, the chemical may not be present within the sample at the time of the analysis. Examples 2, 9, and 14 in this article illustrate the identification of chemicals that had been in contact with a failed plastic component.

References cited in this section


Characterization of Plastics in Failure Analysis

Jeffrey A. Jansen, Stork Technimet, Inc.

Differential Scanning Calorimetry

Differential scanning calorimetry (DSC) is a thermoanalytical technique in which heat flow is measured as a function of temperature and/or time. The obtained measurements provide quantitative and qualitative information regarding physical and chemical changes involving exothermic and endothermic processes, or changes in the heat capacity in the sample material (Ref 2). Differential scanning calorimetry monitors the difference in heat flow between a sample and a reference as the material is heated or cooled (Ref 5). The technique is used to evaluate thermal transitions within a material. Such transitions include melting, evaporation, crystallization, solidification, cross linking, chemical reactions, and decomposition. A typical DSC result is presented in Fig. 5. Differential scanning calorimetry uses the temperature difference between a sample material and a reference as the raw data. In the application, the instrumentation converts the temperature difference into a measurement of the energy per unit mass associated with the transition that caused the temperature change. Any transition in a material that involves a change in the heat content of the material can be evaluated by DSC (Ref 5), the limitation being that commercially available equipment may not be able to detect transitions within materials that are present at concentrations below 5% (Ref 4).
Fig. 5 Differential scanning calorimetry thermogram showing various transitions associated with polymeric materials. The (I) indicates that the numerical temperature was determined as the inflection point on the curve.

Method

Sample preparation for DSC analysis includes placing the specimen within a metal pan. The pan can be open, crimped, or sealed hermetically, depending on the experiment. A reference, either in the form of an empty pan of the same type or an inert material having the same weight as the sample, is used. The most commonly used metal pan material is aluminum; however, pans made of copper and gold are used for special applications. The sample and reference pans rest on thermoelectric disc platforms, with thermocouples used to measure the differential heat flow (Ref 5). Specimen size typically ranges between 1 and 10 mg, although this can vary depending on the nature of the sample and the experiment. The normal operating temperature range for DSC testing is -180 to 700 °C (-290 to 1290 °F), with a standard heating rate of 10 °C/min (18 °F/min). A dynamic purge gas is used to flush the sample chamber. Nitrogen is the most commonly used purge gas, but helium, argon, air, and oxygen can also be used for specific purposes. Often, two consecutive heating runs are performed to evaluate a sample. A controlled cooling run is performed after the initial analysis in order to eliminate the heat history of the sample. The first heating run assesses the sample in the as-molded condition, while the second run evaluates the inherent properties of the material.

Results

The plotted results obtained during a DSC analysis are referred to as a thermogram. The thermogram shows the heat flow in energy units or energy per mass units on the y-axis as a function of either temperature or time on the x-axis. The transitions that the sample material undergoes appear as exothermic and endothermic changes in the heat flow. Endothermic transitions require heat to proceed, while exothermic transitions give off heat.

Uses of DSC in Failure Analysis
Melting Point and Crystallinity. The primary use of DSC in polymer analysis is the detection and quantification of the crystalline melting process. Because the crystalline state of a polymeric material is greatly affected by properties including stereoregularity of the chain and the molecular weight distribution as well as by processing and subsequent environmental exposure, this property is of considerable importance (Ref 5). The melting point \( (T_m) \) of a semicrystalline polymer is measured as the peak of the melting endotherm. A composite thermogram showing the melting transitions of several common plastic materials is presented in Fig. 6. The \( T_m \) is used as a means of identification, particularly when other techniques, such as FTIR, cannot distinguish between materials having similar structures. This can be useful in identifying both the main resin and any contaminant materials. The material identification aspects of DSC are illustrated in examples 4, 5, 7, 8, 10, 11, 12, and 15 in this article.

![Fig. 6 Differential scanning calorimetry used to identify polymeric materials by determination of their melting point.](image)

The heat of fusion represents the energy required to melt the material and is calculated as the area under the melting endotherm. The level of crystallinity is determined by comparing the actual as-molded heat of fusion with that of a 100% crystalline sample. The level of crystallinity that a material has reached during the molding process can be practically assessed by comparing the heat of fusion obtained during an initial analysis of the sample with the results generated during the second run, after slow cooling. The level of crystallinity is important, because it impacts the mechanical, physical, and chemical resistance properties of the molded article. In general, rapid or quench cooling results in a lower crystalline state. This is the result of the formation of frozen-in amorphous regions within the preferentially crystalline structure. Examples 11 and 12 in this article show applications involving DSC as a means of assessing crystallinity. Recrystallization, or the solidification of the polymer, is represented by the corresponding exothermic transition as the sample cools. The recrystallization temperature \( (T_c) \) is taken as the peak of the exotherm, and the heat of recrystallization is the area under the exotherm. Some slow-crystallizing materials, such as poly(ethylene terephthalate) and polyphthalamide, undergo low-temperature crystallization, representing the spontaneous rearrangement of amorphous segments within the polymer structure into a more orderly crystalline structure. Such exothermic transitions indicate that the as-molded material had been cooled relatively rapidly. Example 9 in this article shows how low-temperature crystallization was detected via DSC.

Glass Transition in Amorphous Plastics. Polymers that do not crystallize and semicrystalline materials having a significant level of amorphous segments undergo a phase change referred to as a glass transition. The glass transition
represents the reversible change from/to a viscous or rubbery condition to/from a hard and relatively brittle one (Ref 6). The glass transition is observed as a change in the heat capacity of the material. The glass transition temperature ($T_g$) can be defined in several ways but is most often taken as the inflection point of the step transition. A composite thermogram showing the glass transitions of several common plastic materials is presented in Fig. 7. The $T_g$ of an amorphous resin has an important impact on the mechanical properties of the molded article, because it represents softening of the material to the point that it loses load-bearing capabilities.

![Graph showing glass transitions of Polycarbonate, Polysulfone, and Polystyrene with temperatures marked as 148.87°C and 103.64°C (Polycarbonate), 187.79°C (Polysulfone), and 103.64°C (Polystyrene).](Image)

**Fig. 7** Differential scanning calorimetry used to detect glass transitions within amorphous thermoplastic resins. The (I) indicates that the numerical temperature was determined as the inflection point on the curve.

Aging, Degradation, and Thermal History. As noted by Sepe (Ref 5), “DSC techniques can be useful in detecting the chemical and morphological changes that accompany aging and degradation.” Semicrystalline polymers may exhibit solid-state crystallization associated with aging that takes place at elevated temperatures. In some polymers, this may be evident as a second $T_m$ at a reduced temperature. This second $T_m$ represents the approximate temperature of the aging exposure. Other semicrystalline materials may show an increase in the heat of fusion and an increase in the $T_m$. The thermal aging of both the resin and the failed molded part is illustrated in example 10 in this article. Amorphous resins exhibit changes in the glass transition as a result of aging. In particular, physical aging, which occurs through the progression toward thermodynamic equilibrium below the $T_g$, produces an apparent endothermic transition on completion of the glass transition.

Degradation and other nonreversible changes to the molecular structure of semicrystalline polymers can be detected as reduced values for the $T_m$, $T_c$, or heat of fusion. Instances of degradation detected by DSC are presented in examples 7, 14, and 15 in this article. Similarly, degradation in amorphous resins can be observed as a reduction in the $T_g$ or in the magnitude of the corresponding change in heat capacity. Further, the resistance of a polymer to oxidation can be evaluated via DSC by standard methods or experiments involving high-pressure oxygen or air exposure. Such evaluations usually measure the oxidative induction time or the temperature at which oxidation initiates under the experimental conditions. This can be used to compare two similar materials or to determine whether a plastic resin has undergone partial oxidation.
Thermogravimetric Analysis

Thermogravimetric analysis (TGA) is a thermal analysis technique that measures the amount and rate of change in the weight of a material as a function of temperature or time in a controlled atmosphere. The weight of the evaluated material can decrease due to volatilization or decomposition or increase because of gas absorption or chemical reaction. Thermogravimetric analysis can provide valuable information regarding the composition and thermal stability of polymeric materials. The obtained data can include the volatiles content, inorganic filler content, carbon black content, the onset of thermal decomposition, and the volatility of additives such as antioxidants (Ref 4).

Method

Thermogravimetric analysis instruments consist of two primary components: a microbalance and a furnace. The sample is suspended from the balance while heated in conjunction with a thermal program. A ceramic or, more often, a platinum sample boat is used for the evaluation. As part of the TGA evaluation, the sample is usually heated from ambient room temperature to 1000 °C (1830 °F) in a dynamic gas purge of nitrogen, air, or a consecutive switch program. The composition of the purge gas can have a significant effect on the TGA results and, as such, must be properly controlled. The size of the sample evaluated usually ranges between 5 and 100 mg, with samples as large as 1000 mg possible. Minimal sample preparation is required for TGA experiments.

Results

The results obtained as part of TGA evaluation are known as a thermogram. The TGA thermogram illustrates the sample weight, usually in percent of original weight, on the y-axis as a function of time or, more commonly, temperature on the x-axis. The weight-change transitions are often highlighted by plotting the corresponding derivative on an alternate y-axis.

Uses of TGA in Failure Analysis

Composition. Thermogravimetric analysis is a key analytical technique used in the assessment of the composition of polymeric-based materials. The quantitative results obtained during a TGA evaluation directly complement the qualitative information produced by FTIR analysis. The relative loadings of various constituents within a plastic material, including polymers, plasticizers, additives, carbon black, mineral fillers, and glass reinforcement, can be assessed. The assessment of a plastic resin composition is illustrated in Fig. 8. This data is important as part of a failure analysis in order to determine if the component was produced from the correct material. The weight-loss profile of the material is evaluated, and, ideally, the TGA results obtained on the material exhibit distinct, separate weight-loss steps. These steps are measured and associated with transitions within the evaluated material. A thorough knowledge of the decomposition and chemical reactions is required to properly interpret the obtained results. In most situations, however, distinct weight-loss steps are not obtained, and in these cases, the results are complemented by the corresponding derivative curve. Noncombustible material remaining at the conclusion of the TGA evaluation is often associated with inorganic fillers. Such residue is often further analyzed using energy-dispersive x-ray spectroscopy (EDS) in order to evaluate its
composition. The use of TGA in characterizing plastic composition is presented in examples 8, 10, 12, and 15 in this article. Additionally, example 11 illustrates the quantification of an absorbed chemical within a failed plastic component.

![Thermogravimetric analysis thermogram](image)

**Fig. 8** Thermogravimetric analysis thermogram showing the weight-loss profile for a typical plastic resin

Thermal Stability. Thermogravimetric analysis data can also be used to compare the thermal and oxidative stability of polymeric materials. The relative stability of polymeric materials can be evaluated by assessing the onset temperature of decomposition of the polymer. Quantitatively, these onset temperatures are not useful for comparing the long-term stability of fabricated products, because the materials are generally molten at the beginning of decomposition (Ref 5). However, a comparison of the obtained TGA thermograms can provide insight into possible degradation of the failed component material. Example 6 in this article illustrates a comparison of the thermal stability of two polymeric materials, while example 13 shows the effects of molecular degradation. Degradation experiments involving polymeric materials can also provide information regarding the kinetics of decomposition. Such studies provide information regarding the projected lifetime of the material. Such measurements, however, provide little information pertinent to a failure analysis.

Evolved Gas Analysis. Thermogravimetric analysis evaluations can also be performed whereby the evolved gaseous constituents are further analyzed using a hyphenated technique, such as FTIR or mass spectrometry (MS). Such TGA-FTIR or TGA-MS experiments are referred to as evolved gas analysis.

**References cited in this section**


Thermomechanical Analysis

Thermomechanical analysis (TMA) is a thermal analysis method in which linear or volumetric dimensional changes are measured as a function of temperature, time, or force (Ref 2). Thermomechanical analysis is used to study the structure of a polymeric material by evaluating the implications of the material dimensional changes.

Method

Standard solid samples evaluated via TMA should be of regular shape, having two flat, parallel sides. Additionally, fiber and film samples can also be tested with minimal preparation. Experiments conducted to evaluate expansion and contraction of solid materials are performed on a quartz stage. The sample is placed on the stage, with a quartz probe resting on the opposing end. Thermomechanical analysis data can be acquired in compression modes, including expansion, penetration, dilatometry, rheometry, and flexure or tension mode (Ref 2). The analysis of film and fiber samples requires special fixturing, similar in principle to a universal mechanical tester. For all analysis configurations, the stage assembly is surrounded by a furnace and a cooling device. The normal operating range for TMA experiments is -180 to 1000 °C (-290 to 1830 °F), with a 5 °C/min (9 °F/min) heating rate commonly used. A compressive force is normally applied to the probe configuration throughout the evaluation for purposes of preload and stability.

Results

Plotted results generated through a TMA analysis, similar to the printed data obtained from all of the thermal analysis techniques, are referred to as a thermogram. The thermogram presents the sample dimension, either as length or a percentage of original length, on the y-axis, and as a function of temperature, time, or force on the x-axis. Temperature is the standard independent variable. Changes in the sample are presented as expansion or contraction.

Uses of TMA in Failure Analysis

Coefficient of Thermal Expansion. The coefficient of thermal expansion (CTE) is the change in the length of a material as a response to a change in temperature. The derivative of the slope of the line showing the dimensional changes with respect to temperature represents the CTE. This is a significant property when plastic materials are used under highly constrained conditions. This is commonly the case when plastic parts are used in conjunction with components produced from other materials, such as metals and ceramics. In general, the CTEs of polymeric materials are substantially greater than those of metals and ceramics. Thus, comparative testing of mating materials can produce data used to illustrate and even calculate the potential interference stresses on the materials in a multimaterial design. The evaluation of the CTEs of mating plastic and metal components is illustrated in examples 10 and 14 in this article.

Material Transitions. According to Sepe (Ref 5), “The CTE is an important property in itself; however, it is of particular value in polymers, because sudden changes in CTE can signal important transitions in the material structure.” Within semicrystalline polymers, the \( T_g \), signaling the conversion from a hard, brittle material to a rubbery condition, is accompanied by an increase in the CTE. A thermogram representing a typical semicrystalline resin is shown in Fig. 9. The physical properties of the material can be expected to be significantly different across this transition. Amorphous resins soften at the \( T_g \), and because of this, samples undergo compression under the inherent load of the testing conditions. A thermogram showing the glass transition in an amorphous resin is shown in Fig. 10. The evaluation of the glass transition is presented in example 14 in this article. Thermomechanical analysis is generally accepted as a more accurate method for assessing the \( T_g \) of polymeric materials, relative to DSC. “By using the prescribed attachments and the appropriate force, TMA can be used to determine two commonly measured properties of plastic materials: the heat-deflection temperature and the Vicat softening temperature.” (Ref 5).
Fig. 9 Thermogravimetric analysis thermogram representing a typical semicrystalline plastic resin.
Fig. 10  Thermogravimetric analysis thermogram representing a typical amorphous plastic resin

Molded-In Stress. Internal molded-in stress is an important source of the total stress on a plastic component and is often sufficient to result in the failure of plastic materials. Such stresses are particularly important in amorphous resins, which are prone to environmental stress cracking. Molded-in stresses are commonly imparted through the forming process, especially injection molding and thermoforming. Molded-in stress is observed in amorphous resins as a marked expansion in the sample dimension at temperatures approaching the $T_g$, as illustrated in Fig. 11. This expansion is associated with rapid expansion as the internal stresses are relieved. This stress relief is due to molecular reorientation on attaining sufficient thermal freedom. In the absence of molded-in stress, the sample would compress due to the loss of load-bearing capabilities as the material undergoes glass transition.

Fig. 11  Thermogravimetric analysis thermogram showing a high level of residual stress in an amorphous plastic resin

Chemical Compatibility. The chemical compatibility of a plastic material with a particular chemical agent can be assessed using TMA. In particular, the volume swell of a polymeric material by a chemical can be tested. The sample material is constrained in a quartz vessel, and the chemical agent is added. Dilatometry is used to measure the volume expansion of the material over time.

References cited in this section


Dynamic Mechanical Analysis

Dynamic mechanical analysis (DMA) is a thermoanalytical technique that assesses the viscoelastic properties of materials. Dynamical mechanical analysis evaluates the stiffness, as measured by modulus, as a function of temperature or time. Polymeric materials display both elastic and viscous behavior simultaneously, and the balance between the elastic recovery and viscous flow changes with temperature and time (Ref 5). Measurements can be made in several modes, including tension, shear, compression, torsion, and flexure. The results obtained as part of a DMA experiment provide the storage modulus, loss modulus, and the tangent of the phase-angle delta (tan delta). Dynamic mechanical analysis is not routinely used as a failure analysis technique, but it can provide valuable material information.

Method

Dynamic mechanical analysis experiments can be performed using one of several configurations. The analysis can be conducted to apply stress in tension, flexure, compression, shear, or torsion. The mode of the analysis determines which type of modulus is evaluated. The measurement of modulus across a temperature range is referred to as temperature sweep. Dynamic mechanical analysis offers an advantage over traditional tensile or flexural testing in that the obtained modulus is continuous over the temperature range of interest. In addition, special DMAs can also be conducted to evaluate creep through the application of constant stress or stress relaxation by using a constant strain. Dynamic mechanical analysis studies can be performed from -150 to 600 °C (-240 to 1110 °F), usually employing a 2 °C/min (4 °F/min) heating rate.

Results

The results obtained as part of a DMA evaluation are plotted to illustrate the elastic or storage modulus ($E'$) and the viscous or loss modulus ($E''$) on the y-axes and as a function of temperature on the x-axis. Less frequently, time is used, depending on the type of experiment. Additionally, the tangent of the phase-angle delta ($E''/E'$) is also calculated. A typical DMA thermogram is presented in Fig. 12.
Fig. 12 Dynamic mechanical analysis thermogram showing the results obtained on a typical plastic resin. Tan delta is ratio of the loss modulus to the storage modulus.

**Uses of DMA in Failure Analysis**

Temperature-Dependent Behavior. The temperature-dependent behavior of polymeric materials is one of the most important applications of DMA. In a standard temperature-sweep evaluation, the results show the storage modulus, loss modulus, and the tan delta as a function of temperature. The storage modulus indicates the ability of the material to accommodate stress over a temperature range. The loss modulus and tan delta provide data on temperatures where molecular changes produce property changes, such as the glass transition and other secondary transitions not detectable by other thermal analysis techniques. The superiority of DMA over DSC and TMA for assessing the glass transition is well documented (Ref 5). Secondary transitions of lesser magnitude are also important, because they can relate to material properties such as impact resistance. The ability of a plastic molded component to retain its properties over the service temperature range is essential and is well predicted by DMA.

Aging and Degradation. Changes in the mechanical properties of plastic resins that arise from molecular degradation or aging can be evaluated via DMA. Such changes can significantly alter the ability of the plastic material to withstand service stresses. While the cause and type of degradation cannot be determined, DMA can assess the magnitude of the changes. This can provide insight into potential failure causes.

Solid and Liquid Interactions. Sepe (Ref 5) notes that “DMA is sensitive to structural changes that can arise when a solid polymer absorbs a liquid material.” This effect can arise from the absorption of water or organic-based solvents. Dynamic mechanical analysis experiments can assess changes in the physical properties of a plastic material that can result from such absorption, including loss of strength and stiffness. Example 11 in this article shows the changes in mechanical properties of a plastic resin associated with chemical absorption. The experiments can also evaluate the recovery after the removal or evaporation of the absorbed liquid.
Methods for Molecular Weight Assessment

The aspect of molecular structure, and specifically molecular weight, makes polymeric materials unique among materials commonly used in engineering applications, including metals and ceramics. Molecular weight and molecular weight distribution are probably the most important properties for characterizing plastics (Ref 4). These parameters have a significant impact on the entirety of characteristics of a plastic resin, including mechanical, physical, and chemical resistance properties. Molecular weight assessment can be used to evaluate the characteristics of a base resin or to assess the effects of compounding, molding, or service on the material. Changes in molecular weight can occur throughout the material life cycle and can significantly impact the performance of the molded part. Changes can result in molecular weight decreases through such mechanisms as chain scission, oxidation, and hydrolysis, or as increases through destructive cross linking. Because of this, the characterization of molecular weight is an important aspect of a thorough failure analysis.

Gel permeation chromatography (GPC), which is also referred to as size exclusion chromatography, is an analytical method used to characterize the molecular weight distribution of a polymeric material. Similar to all chromatographic techniques, GPC uses a packed column to segregate various constituents. One or multiple columns used in conjunction are used to separate the polymeric and oligomeric materials within the plastic resin. The polymer is further separated by molecular weight, producing essentially a histogram representing the molecular weight distribution of the material. From these results, a numerical average molecular weight can be calculated. Detectors, based on refractive index or ultraviolet detection, are used to identify the changes in molecular weight.

Gel permeation chromatography offers the advantage, unlike melt viscosity and solution viscosity techniques, of producing results that directly represent the actual molecular weight and molecular weight distribution of the plastic resin. Another advantage is that GPC requires a relatively small sample size, 30 to 120 μg, for a complete evaluation (Ref 7). The technique, however, is often complicated to perform, using sophisticated instrumentation, and difficult to interpret. Example 10 in this article reviews the use of GPC in a failure investigation.

Melt Flow Index. The melt flow index or melt flow rate (MFR) describes the viscosity of a plastic material in the molten state. The sample material is heated through the melting or softening point and extruded through a die having a standard-sized orifice. Different materials use various test conditions, including temperature and load. The method for determining the MFR is described in ASTM D 1238. Melt flow rate is the simplest technique for assessing the molecular weight of a plastic material and is inversely proportional to the molecular weight of the polymer (Ref 4). Melt flow rate is widely used to describe the molecular weight of a plastic resin and is commonly cited by suppliers on a material data sheet. The units used to indicate MFR are grams per 10 min. Examples 7, 11, 12, and 14 in this article describe the use of melt flow in assessing molecular weight in a failure analysis.

While MFR is relatively easily determined and is commonly used to describe molecular weight, the technique has several negative aspects. Melt flow rate does not measure the molecular weight distribution of the analyzed material and represents only the average molecular weight of the material. Because of this, the blending of polymers having different molecular weight distributions and average molecular weights can result in equal determinations between very different materials having distinct properties.

Solution Viscosity. The traditional approach for determining only the molecular weight of a resin, but not the molecular weight distribution, involves dissolving the polymer in a suitable solvent. However, the more structurally complicated macromolecules require the use of hostile solvents, tedious sample preparations, and
costly time delays to obtain limited, single datapoint values. For example, the solution viscosity determination of polyvinyl chloride (PVC), according to ASTM D 1243, requires either a 1 or 4% concentration in cyclohexanone or dinitrobenzene, while polyamides, or nyons, require formic acid. Other engineering polymers might require tetrahydrofuran, dimethylformamide, dimethylsulfoxide, or other equally hostile solvents (Ref 8). The obtained solution viscosity values are only indications of molecular weight and do not reflect the absolute weight values (Ref 8). Example 9 in this article illustrates the use of solution viscosity in a failure investigation.

References cited in this section


Characterization of Plastics in Failure Analysis
Jeffrey A. Jansen, Stork Technimet, Inc.

Mechanical Testing

Because a wide range of mechanical tests are available to evaluate plastics and polymers, they initially do not seem to constitute a rational set. The totality of mechanical tests can be partitioned into logical groups in several distinct ways (Ref 9). One very useful way to classify the various mechanical test methods is to distinguish between tests that evaluate long-term properties, as opposed to those that evaluate short-term properties. Short-term tests include those that assess what are generally considered to be material properties. These include tensile tests, flexural tests, and the evaluation of impact resistance. Short-term tests, while generally easy to conduct and interpret, lack the ability to predict or assess the long-range performance properties of a material. As such, short-term tests are frequently listed on material data sheets.

Tests for Short-Term Properties. The most commonly performed mechanical test used to evaluate plastic material properties is the tensile test. This testing is performed on a dumbbell-shaped specimen and is outlined in ASTM D 638. Tensile testing provides data regarding the yield point in the form of yield strength and elongation at yield, the break properties as tensile strength at break and elongation at break, and the stiffness of the material as elastic modulus. Additionally, the tensile test generates information regarding the proportional limit.

A second short-term mechanical method that is used to evaluate plastic materials is flexural testing. Flexural testing simulates bending of the test sample. The test specimen is evaluated on a universal mechanical tester, and the tests can be performed using a three- or four-point bend configuration. Flexural testing provides two pieces of data: flexural modulus and break strength. This testing is performed in accordance with ASTM D 790. Several different types of tests are used to evaluate the impact properties of a plastic material. These include pendulum-based tests, such as Izod and Charpy tests, and falling weight tests, such as the dart penetration configuration. Unlike tensile and flexural testing, the results obtained from impact testing do not provide fundamental material properties. Instead, impact testing results are more performance-based. Given these different methodologies of assessing the impact properties of a plastic material, the falling weight or dart impact tests are generally considered to be superior to the pendulum configurations. Falling weight tests
evaluate the sample material in two dimensions and not one, because the specimen is a plate rather than a beam. The data obtained during an instrumented falling weight impact test include the energy to maximum load, representing the energy required to initiate cracking, and the total energy to failure. The ratio of these two is an indication of the ductility of the material. Additionally, an examination of the test specimens is used to classify the failure mode from brittle to ductile. Falling weight impact testing is described as part of ASTM D 3029.

Tests for Long-Term Properties. Fatigue testing of plastic materials exposes the samples to cyclic stresses in an attempt to evaluate the samples in a manner that would produce fatigue failure while in service. Testing procedures are used to simulate flexural fatigue and tensile fatigue. The analyses are normally conducted in a way that does not excessively heat the specimen, thus altering the failure mode. The results of a fatigue test are shown in the form of a stress-number of cycles curve.

A second long-term test methodology assesses the creep resistance of the material. Creep testing exposes the sample to a constant stress over a prolonged period of time. This is done to simulate the effects of static stresses on the performance of a material in service. The extension or strain of the sample over time is measured. Traditional creep testing can take an extended period of time. Similar results can, however, be obtained through a DMA creep study, which can be performed in the course of a few days.

Mechanical Testing as Part of a Failure Analysis. The use of mechanical testing in a failure analysis is limited. The preparation of specimens from the failed component may not be possible. Further, published standard mechanical data, including yield strength, elastic modulus, and flexural modulus, are very dependent on the specimen configuration and testing conditions. Given that most published data are generated on specially molded test specimens, the testing of samples excised from molded articles may not provide an adequate comparison. In some cases, it is not apparent whether observed differences are the result of material deficiencies or variations in test specimen configuration. Instead, mechanical testing is most useful in comparing a known good or control sample with a failed part. Many times, this is best accomplished through some sort of proof load testing. Proof load testing involves measuring the strength and dimensional changes as a function of an applied load. In most cases, this testing involves producing a catastrophic failure within the test sample. The use of proof load testing as part of a failure analysis is illustrated in example 12 in this article.

Reference cited in this section


Characterization of Plastics in Failure Analysis

Jeffrey A. Jansen, Stork Technimet, Inc.

Considerations in the Selection and Use of Test Methods

Through the application of analytical testing and a systematic engineering approach, it is possible to successfully ascertain the nature and cause of a plastic component failure. The testing, however, must be performed in a sound manner, with the obtained data only being as good as the analysis method. Further, the data presented by the analytical methods are often complicated and, in many cases, require an experienced analyst to be properly interpreted.

The aforementioned analytical tests are not meant to be an all-encompassing list of the methods used to evaluate failed plastic components. Certainly, there are numerous testing methodologies that provide data pertinent to a plastic component failure analysis. Other analysis techniques, including EDS and SEM, are important tools in a plastic component failure analysis. Scanning electron microscopy and EDS methods are covered elsewhere in this Volume. More specialized chromatographic methods, including gas chromatography and gas chromatography-mass spectroscopy, are extremely useful in assessing low-concentration additives within a plastic resin. Nuclear magnetic resonance spectroscopy is useful in polymeric analysis, providing information related to composition beyond FTIR. Nuclear magnetic resonance can provide data regarding stereoregularity, carbon content, chemical composition, and copolymer structure (Ref 10). Additionally, surface analysis spectroscopic techniques, such as secondary ion mass spectroscopy, x-ray photoelectron spectroscopy, and electron spectroscopy for chemical analysis, are specifically used to characterize very shallow surface layers. These techniques can be used to analyze material composition but are particularly suited for the analysis of surface contaminants (Ref 10). While these analytical techniques can provide valuable data as part of a plastics failure analysis,
the tests described in this article are considered to be the most important in the majority of cases. Given the charge that this article be treated in a practical manner, test methods used less often were omitted. A summary showing both the treated and omitted analysis methods and the corresponding information gained is included in Table 1 (Ref 3).

Table 1  Practical information derived from polymer analysis methods

<table>
<thead>
<tr>
<th>Test method</th>
<th>Properties measured</th>
<th>Use in failure analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fourier transform infrared spectroscopy (FTIR)</td>
<td>Molecular bond structure</td>
<td>Material identification, contamination, degradation, chemical contact</td>
</tr>
<tr>
<td>Differential scanning calorimetry (DSC)</td>
<td>Heat of fusion, melting point, glass transition temperature, heat capacity</td>
<td>Material identification, level of crystallinity, aging/degradation, thermal history</td>
</tr>
<tr>
<td>Thermogravimetric analysis (TGA)</td>
<td>Weight loss over temperature or time</td>
<td>Composition, thermal stability, evolved gas analysis</td>
</tr>
<tr>
<td>Thermomechanical analysis (TMA)</td>
<td>Dimensional changes over temperature</td>
<td>Coefficient of thermal expansion, material transitions, molded-in stress, chemical compatibility</td>
</tr>
<tr>
<td>Dynamic mechanical analysis (DMA)</td>
<td>Elastic modulus, viscous modulus, tan delta</td>
<td>Temperature-dependent behavior, aging/degradation, solid-liquid interactions</td>
</tr>
<tr>
<td>Gel permeation chromatography (GPC)</td>
<td>Weight-average molecular weight, molecular weight distribution</td>
<td>Degradation, suitability of material for use</td>
</tr>
<tr>
<td>Melt flow rate (MFR)</td>
<td>Melt viscosity</td>
<td>Degradation, compliance with material specification</td>
</tr>
<tr>
<td>Solution viscosity</td>
<td>Intrinsic viscosity</td>
<td>Degradation</td>
</tr>
<tr>
<td>Mechanical testing</td>
<td>Strength and elongation properties, modulus</td>
<td>Compliance with material specification, mechanical properties</td>
</tr>
<tr>
<td>Scanning electron microscopy (SEM)</td>
<td>Surface and particle morphology</td>
<td>Fracture mode</td>
</tr>
<tr>
<td>Energy dispersive x-ray spectroscopy (EDS)</td>
<td>Elemental concentrations</td>
<td>Material composition, fillers, additives</td>
</tr>
<tr>
<td>Nuclear magnetic resonance (NMR)</td>
<td>Molecular bond structure</td>
<td>Material identification</td>
</tr>
<tr>
<td>Mass spectroscopy (MS)</td>
<td>Molecular structure</td>
<td>Material identification, additives</td>
</tr>
<tr>
<td>X-ray photoelectron spectroscopy (XPS)</td>
<td>Elemental concentrations</td>
<td>Chemical composition of surfaces</td>
</tr>
<tr>
<td>Auger electron spectroscopy (AES)</td>
<td>Elemental concentrations</td>
<td>Chemical composition of surfaces</td>
</tr>
</tbody>
</table>

Source: Ref 3

References cited in this section


Characterization of Plastics in Failure Analysis

Jeffrey A. Jansen, Stork Technimet, Inc.

Case Studies
Example 1: Embrittlement of a Polycarbonate Bracket. A plastic bracket exhibited relatively brittle material properties, which ultimately led to catastrophic failure. The part had been injection molded from a medium-viscosity polycarbonate resin and had been in service for a short duration prior to the failure.

Tests and Results. A visual examination of the bracket revealed a series of surface anomalies, and it was suspected that the presence of the defects was related to the premature failure. The component base material was analyzed using micro-FTIR in the attenuated total reflectance (ATR) mode. The obtained spectrum exhibited absorption bands characteristic of polycarbonate, as shown in Fig. 13. No evidence of material contamination was found. A similar analysis was performed on the part surface in an area that showed the anomalous surface condition. The spectrum representing the surface was generally similar to the results obtained on the base material. However, the surface spectrum showed a relative increase in the intensity of a spectral band between 3600 and 3350 cm$^{-1}$, indicative of hydroxyl functionality. Additionally, the spectrum also showed changes in the relative intensities of several bands, as compared to the results representing the base material. A spectral subtraction was performed, and the results produced a good match with a library reference of diphenyl carbonate. This is illustrated in the spectral comparison presented in Fig. 14. Diphenyl carbonate is a common breakdown product produced during the decomposition of polycarbonate.

![Fig. 13 The FTIR spectrum obtained on the bracket base material, exhibiting absorption bands characteristic of polycarbonate](image-url)
Conclusions. Overall, the obtained results suggested that the anomalous surface condition observed on the bracket represented molecular degradation of the polycarbonate. This is consistent with the brittle properties exhibited by the component. The most likely cause of the molecular degradation was improper drying and/or exposure to excessive heat during the injection molding process.

Example 2: Chemical Attack of Acrylonitrile-Butadiene-Styrene Grips. A set of plastic grips from an electric consumer product failed while in service. The grips had been injection molded from a general-purpose grade of an acrylonitrile-butadiene-styrene (ABS) resin. The parts had cracked while in use after apparent embrittlement of the material.

Tests and Results. An examination of the grips confirmed a severe level of cracking, covering the majority of the grip surface. Handling of the parts revealed that the grip material exhibited very little integrity, unlike the usual ductility associated with ABS resins. A white discoloration was also observed on the otherwise red grips. The surface of the grips was evaluated using SEM, revealing isolated areas that showed significant degradation in the form of material loss, as shown in Fig. 15. The observed morphology suggested selective degradation of the polybutadiene domains within the ABS resin.
Fig. 15  Scanning electron image showing isolated degradation of the grip material. 30×
Micro-FTIR in the ATR mode was used to analyze the base material and the surfaces of the grips. The results obtained on the base material were characteristic of an ABS resin. Analysis of the surface of the part produced a somewhat different result. The spectrum representing the grip surface contained absorption bands associated with ABS; however, the results contained additional bands of significant intensity. A spectral subtraction was performed, thereby removing the bands associated with the ABS resin. The obtained subtraction spectrum produced a very good match with glyceride derivatives of fats and oils. This identification is illustrated in Fig. 16.

![FTIR Spectrum](image)

Fig. 16 The FTIR spectrum obtained on the grip surface. The spectrum contains absorption bands indicative of glyceride derivatives of fats and oils in addition to bands associated with the base ABS resin.
Conclusions. It was the conclusion of the analysis that the grips failed via brittle fracture associated with severe chemical attack of the ABS resin. A significant level of glyceride derivatives of fatty acids, known to degrade ABS resins, was found on the part surface. The glyceride derivatives selectively attacked the polybutadiene domains within the molded ABS part, leading to apparent embrittlement and subsequent failure.

Example 3: Inclusion within an ABS Handle. The handle from a consumer product exhibited an apparent surface defect. The handle had been injection molded from a medium-viscosity-grade ABS resin. The anomalous appearance was objectionable to the assembler of the final product and resulted in a production lot being placed on quality-control hold.

Tests and Results. The surface of the part was examined using an optical stereomicroscope. The defect appeared as a localized area of lightened color, and the zone immediately surrounding the anomaly was slightly recessed. A mounted and polished cross section was prepared through the part, revealing distinct included material within the base molding resin. The inclusion, as shown in Fig. 17, did not appear to contain a significant level of the blue pigment, as present in the base material.

![Fig. 17 Micrograph showing the included material within the handle. 24×](image)

The preparation of the cross section not only allowed a thorough inspection of the defect but also served to facilitate further analysis of the material. The sample was initially analyzed using EDS. The results obtained on the included material showed exclusively carbon and oxygen, precluding an inorganic contaminant. The base resin and the included material were further analyzed using FTIR in the reflectance mode. The spectrum representing the base material contained absorption bands indicative of an ABS resin. Analysis of the included material produced distinctly different results. The spectrum obtained on the included material was characteristic of polybutadiene, the rubber-modifying agent present in ABS. This identification is presented in Fig. 18.
Conclusions. It was the conclusion of the evaluation that the handle contained an inclusion, which produced the apparent surface anomaly. The included material was identified as polybutadiene. The most likely source of the included polybutadiene was an undispersed gel particle formed during the production of the molding resin.

Example 4: Relaxation of Nylon Wire Clips. A production lot of plastic wire clips was failing after limited service. The failures were characterized by excessive relaxation of the clips, such that the corresponding wires were no longer adequately secured in the parts. No catastrophic failures had been encountered. Parts representing an older lot, which exhibited satisfactory performance properties, were also available for reference purposes. The clips were specified to be injection molded from an impact-modified grade of nylon 6/6. However, the part drawing did not indicate a specific resin.

Tests and Results. A visual examination of the clips showed that the failed parts were off-white in color, while the control parts had a pure white appearance. An analysis of both sets of parts was performed using micro-FTIR in the ATR mode. A direct comparison of the results produced a good match, with both sets of spectra exhibiting absorption bands that were characteristic of a nylon resin. The comparison, however, revealed subtle differences between the two sets of clips. The spectrum representing the reference parts showed a relatively higher level of a hydrocarbon-based impact modifier, while the results obtained on the failed parts showed the presence of an acrylic-based modifier. The differences in the spectra suggested that the two sets of clips were produced from resins having different formulations, particularly regarding the impact modifier.

The clip materials were further analyzed using DSC. The thermogram representing the reference part material, as shown in Fig. 19, exhibited an endothermic transition at 264 °C (507 °F), characteristic of the melting point of a nylon 6/6 resin. Additionally, the results contained a second melting point, of lesser magnitude, at 95 °C (203 °F). This transition was indicative of a hydrocarbon-based impact modifier, as indicated by the FTIR results. The thermogram obtained on the failed clip material also showed a melting point characteristic of a nylon 6/6 resin. However, no evidence was found to indicate a transition corresponding to the hydrocarbon-based modifier found in the control clip material.
Fig. 19 The DSC thermogram representing the reference clip material, exhibiting an endothermic transition characteristic of the melting of a nylon 6/6 resin. The results also showed a second melting transition attributed to a hydrocarbon-based impact modifier.

Conclusions. It was the conclusion of the analysis that the control and failed clips had been produced from two distinctly different resins. While both materials satisfied the requirements of an impact-modified nylon 6/6 resin, differences in the impact modifiers resulted in the observed performance variation. From the results and the observed performance, it appeared that the material used to produce the failed clips had different viscoelastic properties, which produced a greater predisposition for stress relaxation.

Example 5: Embrittlement of Nylon Couplings. Molded plastic couplings used in an industrial application exhibited abnormally brittle properties, as compared to previously produced components. The couplings were specified to be molded from a custom-compounded glass-filled nylon 6/12 resin. An inspection of the molding resin used to produce the discrepant parts revealed differences in the material appearance, relative to a retained resin lot. Specifically, physical sorting resulted in two distinct sets of molding resin pellets from the lot that had generated the brittle parts. Both of these sets of pellets had a coloration that varied from that of the retained reference resin pellets. A sample of retained molding resin, which had produced parts exhibiting satisfactory performance, was available for comparative analysis.

Tests and Results. Micro-FTIR in the ATR mode was used to analyze the molding resin samples. The results obtained on the three molding resin samples were generally similar, and all of the spectra exhibited absorption bands characteristic of a nylon resin. Further analysis of the resin samples using DSC indicated that the control material results exhibited a single endothermic transition at 218 °C (424 °F), consistent with the melting point of a nylon 6/12 resin, as specified. The DSC thermograms obtained on the two resin samples that produced brittle parts also exhibited melting point transitions associated with nylon 6/12. However, additional transitions were also apparent in the results, indicative of the presence of contaminant materials. The results obtained on one of the resin samples, as presented in Fig. 20, showed a secondary melting point at 165 °C (330 °F), indicative of polypropylene. The thermogram representing the second resin sample, as included in Fig. 21, displayed a second melting transition at 260 °C (500 °F), characteristic of a nylon 6/6 resin.
Fig. 20 The DSC thermogram representing a molding resin pellet that had produced brittle parts. The thermogram shows a major melting transition associated with nylon 6/12 and a weaker transition attributed to polypropylene.
Conclusions. It was the conclusion of the analysis that the molding resin used to produce the brittle couplings contained a significant level of contamination, which compromised the mechanical properties of the molded components. Two distinct contaminants were found mixed into the molding pellets. The contaminant materials were identified as polypropylene and nylon 6/6. The source of the polypropylene was likely the purging compound used to clean the compounding extruder. The origin of the nylon 6/6 resin was unknown but may represent a previously compounded resin.

Example 6: Failure of Plasticized Poly(vinyl chloride) Tubing. A section of clear polymeric tubing failed while in service. The failed sample had been used in a chemical transport application. The tubing had also been exposed to periods of elevated temperature as part of the operation. The tubing was specified to be a poly(vinyl chloride) (PVC) resin plasticized with trioctyl trimellitate (TOTM). A reference sample of the tubing, which had performed well in service, was also available for testing.

Tests and Results. The failed and reference tubing samples were analyzed using micro-FTIR in the ATR mode, and the results representing the reference tubing material were consistent with the stated description: a PVC resin containing a trimellitate-based plasticizer. However, the spectrum representing the failed tubing material was noticeably different. While the obtained spectrum contained absorption bands characteristic of PVC, the results indicated that the material had been plasticized with an adipate-based material, such as dioctyl adipate. This identification is shown in Fig. 22.
In order to assess their relative thermal stability, the two tubing materials were analyzed via thermogravimetric analysis (TGA). Both sets of results were consistent with those expected for plasticized PVC resins. The thermograms representing the reference and failed sample materials showed comparable plasticizer contents of 28 and 25%, respectively. The results also showed that the reference material, containing the trimellitate-based plasticizer, exhibited superior thermal resistance relative to the failed material, containing the adipate-based material. This was indicated by the elevated temperature of weight-loss onset exhibited by the reference tubing material.

Conclusions. It was the conclusion of the evaluation that the failed tubing had been produced from a formulation that did not comply with the specified material. The failed tubing was identified as a PVC resin with an adipate-based plasticizer, not TOTM. The obtained TGA results confirmed that the failed tubing material was not as thermally stable as the reference material because of this formulation difference, and that this was responsible for the observed failure.

Example 7: Cracking of Poly(butylene terephthalate) Automotive Sleeves. A number of plastic sleeves used in an automotive application cracked after assembly but prior to installation into the mating components. The sleeves were specified to be injection molded from a 20% glass-fiber-reinforced poly(butylene terephthalate) (PBT) resin. After molding, electronic components are inserted into the sleeves, and the assembly is filled with a potting compound. A retained lot of parts, which had not cracked, were available for reference purposes.

Tests and Results. The reference and failed parts were analyzed using micro-FTIR in the ATR mode. The spectra obtained on both sets of parts contained absorption bands characteristic of a thermoplastic polyester, such as PBT or poly(ethylene terephthalate) (PET). Different types of polyester resins cannot be distinguished spectrally, because of the similar nature of their structures. However, subtle but distinct differences were apparent in the results, suggestive of degradation of the failed part material.

Differential scanning calorimetry was performed on the sleeve materials using a heat/cool/heat methodology. Testing of the reference material produced initial heating results indicative of a PBT resin, as illustrated by the melting point at 224 °C (435 °F). Analysis of the failed sleeve samples produced a melting transition at a significantly reduced temperature, 219 °C (426 °F). Additionally, the failed material transition was broader, and overall, the results suggested molecular degradation of the failed sleeve material. A comparison of the initial heating thermograms is presented in Fig. 23.
Fig. 23 A comparison of the initial heating run results, suggesting degradation of the failed sleeve material

The identification of degradation was supported by the second heating DSC results, obtained after slow cooling. The second heating thermogram representing the failed sleeve material showed additional differences relative to the results obtained on the reference material. The failed material did not produce the bimodal melting endothermic transition normally associated with PBT after slow cooling. This was thought to be the result of molecular degradation, which produced shorter polymer chain lengths, therefore reducing steric hindrance. A comparison of the second heating thermograms is included in Fig. 24.
Fig. 24 A comparison of the second heating run results, further suggesting degradation of the failed sleeve material

The sleeve materials were further analyzed using TGA. The thermograms obtained on the reference and failed samples were generally consistent, including equivalent glass contents. Additionally, the results were in agreement with those expected for a PBT resin.

The melt flow rates (MFRs) of the reference and failed sleeve materials were determined. Because no molding resin was available for comparison purposes, the nominal range from the specification sheet, 14 to 34 g/10 min, was used. The testing showed that the failed sleeve material had been severely degraded, producing a MFR of 128 g/10 min. This was in agreement with the DSC data and indicated severe molecular degradation of the PBT resin. A review of the results generated by the reference parts also showed significant molecular degradation. While the extent of the degradation was less, the obtained melt flow rate, 50 g/10 min, still demonstrated a substantial reduction in the average molecular weight.

Conclusions. It was the conclusion of the evaluation that the failed sleeves had cracked due to embrittlement associated with severe degradation and the corresponding molecular weight reduction. The degradation was clearly illustrated by the reduced melting point and uncharacteristic nature of the associated endothermic melting transitions as well as the substantial increase in the MFR of the molded parts. The reduction in molecular weight significantly reduced the mechanical properties of the sleeves. The cause of the degradation was not evident, but the likely source appears to be the molding operation and exposure to elevated temperature for an extended period of time. It is significant to note that the reference parts also showed a moderate level of molecular degradation, rendering them susceptible to failure over a longer duration.

Example 8: Cracking of ABS Protective Covers. Numerous protective covers, used in conjunction with an electrical appliance, failed during assembly with the mating components. The failures were traced to a particular production lot of the covers and occurred during insertion of the screws into the corresponding bosses. The parts had been injection molded from an ABS resin to which regrind was routinely added. Retained parts, which exhibited normal behavior during assembly, were available for comparative analysis.

Tests and Results. A visual examination of the failed parts revealed relatively brittle fracture features, without significant ductility, as would be apparent as stress whitening or permanent deformation. Core material taken from the reference and failed parts was analyzed using micro-FTIR in the ATR mode. Both obtained spectra exhibited absorption bands associated with an ABS resin. However, the spectrum representing the failed part showed additional absorption bands.
Spectral subtraction was performed, thereby removing the absorbances attributed to the ABS resin from the spectrum obtained on the failed part. The spectral subtraction results were consistent with a thermoplastic polyester, such as PET or PBT. However, these two materials cannot be distinguished spectrally, because of similarities in their structures. As such, the melting point is usually used to differentiate between these materials. The FTIR results indicated the presence of contaminant material exclusively within the ABS resin used to mold the failed covers.

In order to further identify the contaminant material, a sample taken from the failed part was analyzed via DSC. The obtained DSC thermogram, as presented in Fig. 25, showed a glass transition at approximately 101 °C (214 °F), consistent with the expected results for an ABS resin. The results also showed an additional endothermic transition at 222 °C (432 °F), indicative of a PBT resin. The failed cover material was also analyzed using TGA in order to assess the level of the contamination. The TGA analysis was performed using high-resolution temperature programming, and the results revealed adequate separation of the ABS and PBT resins. Based on the results, the contamination was estimated to account for approximately 23% of the failed cover material.

![DSC Thermogram](image)

**Fig. 25** The DSC thermogram obtained on the failed cover material. The thermogram shows an endothermic transition associated with PBT. The (I) indicates that the numerical temperature was determined as the inflection point on the curve.

**Conclusions.** It was the conclusion of the evaluation that the appliance covers failed via brittle fracture associated with stress overload. The failures, which occurred under normal assembly conditions, were attributed to embrittlement of the molded parts, due to contamination of the ABS resin with a high level of PBT. The source of the PBT resin was not positively identified, but a likely source appeared to be the use of improper regrind.

**Example 9: Failure of Polycarbonate/PET Appliance Housings.** Housings from an electrical appliance failed during an engineering evaluation. The housings had been injection molded from a commercial polycarbonate/PET (PC/PET) blend. The parts were being tested as part of a material conversion. Parts produced from the previous material, a nylon 6/6 resin, had consistently passed the testing regimen. The housing assembly included a spring clip, which applied a static force on a molded-in boss extending from the main body of the housing. Grease was applied liberally within the housing assembly during production.

**Tests and Results.** A visual inspection of the tested parts showed catastrophic failure within the molded-in boss. The failures were consistent across all of the parts and were located at an area where the spring clip contacted the housing.
boss. While the final fracture zone exhibited limited features associated with ductility in the form of stress whitening, no such characteristics were apparent at the locations corresponding to the crack origins. The fracture surfaces were further examined via SEM. The SEM inspection showed the presence of multiple crack initiation sites along the side of the boss that had mated with the spring clip. No evidence of significant ductility was found with the crack initiation locations, as represented in Fig. 26. The overall features observed on the fracture surface were indicative of environmental stress cracking.

![Image](image_url)

Fig. 26 Scanning electron image showing brittle fracture features at the crack initiation site, characteristic of environmental stress cracking. 24×

Micro-FTIR in the ATR mode was performed on the housing material, and the resulting spectrum was in agreement with the stated resin description, a blend of PC and polyester. No signs of material contamination were found. The housing material was further evaluated using DSC. The thermogram obtained during the initial heating run, as shown in Fig. 27, exhibited an endothermic transition at 253 °C (487 °F), characteristic of the melting point of a PET resin. The initial heating run results also showed a low-temperature exothermic transition associated with the crystallization of the PET resin. These results indicated that the material had not been fully crystallized during the molding process. The results generated during the second heating run, after slow cooling, did not show the low-temperature crystallization. The glass transition associated with the PC resin was observed in the second heating run.
Fig. 27 The initial heating DSC thermogram, exhibiting a melting transition consistent with a PET resin. A low-temperature crystallization exothermic transition was also apparent. The (I) indicates that the numerical temperature was determined as the inflection point on the curve.

In order to assess the molecular weight of the housing material, the intrinsic viscosity of the resin was measured. A comparison of the results with historical data revealed a substantial reduction in the viscosity of the failed part material. This indicated that the housing material had undergone significant molecular degradation during the injection molding process.

The grease present within the housing assembly was analyzed using micro-FTIR. The FTIR test results indicated that the grease was composed of a relatively complex mixture. The lubricant contained a hydrocarbon-based oil, a phthalate-based oil, lithium stearate, and an amide-based additive. These results were significant, because phthalate esters are known to be incompatible with PC resins.

Conclusions. It was the conclusion of the analysis that the appliance housings failed through environmental stress cracking. The required chemical agent was identified as a phthalate-based oil present within the grease used to lubricate the assembly. Specifically, the phthalate oil was not compatible with the PC portion of the resin blend. The source of the stress responsible for the cracking appears to be the interference related to the spring clip. While the previous parts, produced from the nylon 6/6 resin, were also under similar stresses, this resin was not prone to stress cracking in conjunction with the lubricant. Thus, the resin conversion was the root cause of the failures. Additionally, the test results also showed that the injection molding process left the material susceptible to failure. Specifically, the molded parts had been undercrystallized, reducing the mechanical strength of the molded articles, and, more importantly, the resin had been degraded, producing a reduction in the molecular weight and reducing both the mechanical integrity and chemical-resistance properties of the parts.

Example 10: Failure of PET Assemblies. Several assemblies used in a transportation application failed during an engineering testing regimen. The testing involved cyclic thermal shock, immediately after which cracking was observed on the parts. The cracking occurred within the plastic jacket, which had been injection molded from an impact-modified, 15% glass-fiber-reinforced PET resin. The plastic jacket had been molded over an underlying metal coil component. Additionally, a metal sleeve was used to house the entire assembly. Prior to molding, the resin had reportedly been dried
at 135 °C (275 °F). The drying process usually lasted 6 h, but occasionally, the material was dried overnight. The thermal shock testing included exposing the parts to alternating temperatures of -40 and 180 °C (-40 and 360 °F). The failures were apparent after 100 cycles. Molding resin and non-failed parts were also available for analysis.

Tests and Results. The failed assemblies were visually and microscopically examined. The inspection showed several different areas within the overmolded jacket that exhibited cracking. The cracked areas were located immediately adjacent to both the underlying metal coil and the outer metal housing. The appearance of the cracks was consistent with brittle fracture, without significant signs of ductility. The examination also revealed design features, including relatively sharp corners and nonuniform wall thicknesses, that appeared to have likely induced molded-in stress within the plastic jacket.

The fracture surfaces were further inspected using SEM, and the examination revealed features generally associated with brittle fracture, as shown in Fig. 28. No evidence of microductility, such as stretched fibrils, was found. The fracture surface features indicated that the cracking had initiated along the outer jacket wall and subsequently extended through the wall and circumferentially around the wall. Throughout the examination, no indication of postmolding molecular degradation was found.

![Fig. 28 Scanning electron image showing brittle fracture features on the failed jacket crack surface. 20×](image)

Micro-FTIR was performed in the ATR mode on a core specimen of the jacket material. The resulting spectrum was consistent with a thermoplastic polyester resin. Such materials, including PET and PBT, cannot be distinguished spectrally, and a melting point determination is usually used to distinguish these materials. The failed jacket material and reference molding resin were analyzed using TGA, and the results obtained on the two samples were generally consistent. This included relatively comparable levels of volatiles, polymer, carbon black, and glass reinforcement. Further, the results were in excellent agreement with those expected for the stated PET material.

The failed jacket and reference materials were evaluated via DSC. Analysis of the failed jacket material produced results that indicated a melting transition at 251 °C (484 °F), consistent with a PET resin. However, a second endothermic transition was also present. This transition, at 215 °C (420 °F), suggested the melting of annealed crystals, indicating that the part had been exposed to a temperature approaching 215 °C (420 °F). The thermal shock testing appeared to be the only possible source of this thermal exposure.

Analysis of the molding resin also produced results consistent with a PET resin. The results also exhibited a second melting endotherm at 174 °C (345 °F). Again, this transition was associated with melting of annealed crystals for material exposed to this temperature. The apparent source of the exposure was the drying process. This was well in excess of the stated drying temperature.

Further analysis of the assembly materials using thermomechanical analysis (TMA) produced significantly different results for the PET jacket and the steel housing material. Determination of the coefficients of thermal expansion (CTEs) showed approximately an order of magnitude difference between the two mating materials.

An assessment of the molecular weight of the failed jacket samples as well as a non-failed part and the molding resin samples was performed using several techniques. A combination of MFR, intrinsic viscosity, and finally, gel permeation chromatography (GPC) was used, because of conflicting results. The MFR determinations showed that the drying process produced a considerable increase in the MFR of the resin, corresponding to molecular degradation in the form of chain scission. This was contrasted by the results generated by the intrinsic viscosity testing. These results showed an increase in the viscosity of the dried resin relative to the virgin resin. This increase was suggestive of an increase in molecular weight, possibly through partial cross linking. Testing of the resin samples and the molded parts via GPC produced results that reconciled the discrepancy. The GPC results showed that the drying process produced competing reactions of chain...
scission and cross linking. The net result was severe degradation of the dried resin, which predisposed the molding material to produce jackets having poor mechanical properties. The GPC testing showed that the molded jackets were further degraded during the injection molding process.

Conclusions. It was the conclusion of the investigation that the assemblies failed via brittle fracture associated with the exertion of stresses that exceeded the strength of the resin as-molded. The stresses were induced by the thermal cycling and the dimensional interference caused by the disparity in the CTEs of the PET jacket and the mating steel sleeve. However, several factors were significant in the failures. It was determined that the resin drying process had exposed the resin to relatively high temperatures, which caused substantial molecular degradation. The drying temperature was found to be approximately 173 °C (344 °F), well in excess of the recommendation for the PET resin. Further degradation was attributed to the molding process itself, leaving the molded jacket in a severely degraded state. This degradation limited the ability of the part to withstand the applied stresses. Additionally, the testing itself exposed the parts to temperatures above the recognized limits for PET, and this may have significantly lowered the mechanical properties of the part.

Example 11: Cracking of a Polyethylene Chemical Storage Vessel. A chemical storage vessel failed while in service. The failure occurred as cracking through the vessel wall, resulting in leakage of the fluid. The tank had been molded from a high-density polyethylene (HDPE) resin. The material held within the vessel was an aromatic hydrocarbon-based solvent.

Tests and Results. A stereomicroscopic examination of the failed vessel revealed brittle fracture surface features. This was indicated by the lack of stress whitening and permanent deformation. Limited ductility, in the form of stretching, was found exclusively within the final fracture zones. On cutting the vessel, significant stress relief, in the form of distortion, was evident. This indicated a high level of molded-in stress within the part. The fracture surface was further inspected using SEM. The observed features included a relatively smooth morphology within the crack origin location, which was indicative of slow crack initiation. This area is shown in Fig. 29. Features associated with more rapid crack extension, including hackle marks and river markings, were found at the midfracture and final fracture areas, as represented in Fig. 30. The entirety of the fracture surface features indicated that the cracking had initiated along the exterior wall of the vessel. The cracking extended transversely through the wall initially, and subsequently, circumferentially around the wall. Throughout the examination, no signs of postmolding molecular degradation or chemical attack were found.

---

**Fig. 29** Scanning electron image showing features associated with brittle fracture and slow crack growth within the crack initiation site. 100×
The failed vessel material was analyzed using micro-FTIR in the ATR mode. The obtained spectrum exhibited absorption bands characteristic of a polyethylene resin. No evidence was found to indicate contamination or degradation of the material.

Material excised from the failed vessel was analyzed using DSC. The results showed a single endothermic transition associated with the melting point of the material at 133 °C (271 °F). The results were consistent with those expected for a HDPE resin. The results also showed that the HDPE resin had a relatively high level of crystallinity, as indicated by the elevated heat of fusion.

Thermogravimetric analysis was performed to further evaluate the failed vessel material. The TGA testing showed that the HDPE absorbed approximately 6.3% of its weight in the aromatic hydrocarbon-based solvent. Overall, the TGA results were consistent with those expected for a HDPE resin.

The MFR of the vessel material was evaluated, and the testing produced an average result of 3.8 g/10 min. This is excellent agreement with the nominal value indicated on the material data sheet, 4.0 g/10 min. As such, it was apparent that the vessel material had not undergone molecular degradation. The specific gravity of the resin was measured. The material produced a result of 0.965. This indicated that the material had a relatively high level of crystallinity, as suggested by the DSC results.

In order to assess the effects of the hydrocarbon-based solvent on the HDPE vessel, the material was evaluated using dynamic mechanical analysis (DMA). The vessel material was analyzed in two conditions. Material samples representing the vessel material in the as-molded condition as well as material from the failed vessel were evaluated. A comparison of the DMA results showed that in the saturated, equilibrium state, the HDPE resin lost over 60% of its elastic modulus at room temperature, because of the plasticizing effects of the solvent. A comparison of the DMA results, indicating the reduction in mechanical properties, is shown in Fig. 31.

---

**Fig. 30** Scanning electron image showing features indicative of rapid crack extension within the final fracture zone. 20×
Fig. 31 A comparison of the DMA results, showing a loss of over 60% in the elastic modulus as a result of the effects of the solvent

Conclusions. It was the conclusion of the investigation that the chemical storage vessel failed via a creep mechanism associated with the exertion of relatively low stresses. Given the lack of apparent ductility, the stresses responsible for the failure appear to have been below the yield strength of the material. The source of the stress was thought to be molded-in residual stresses associated with uneven shrinkage. This was suggested by the obvious distortion evident on cutting the vessel. The relatively high specific gravity and the elevated heat of fusion are indicative that the material has a high level of crystallinity. In general, increased levels of crystallinity result in higher levels of molded-in stress and the corresponding warpage. The significant reduction in the modulus of the HDPE material, which accompanied the saturation of the resin with the aromatic hydrocarbon-based solvent, substantially decreased the creep resistance of the material and accelerated the failure. The dramatic effects of the solvent had not been anticipated prior to use.

Example 12: Failure of Polyacetal Latch Assemblies. Components of a latch assembly used in a consumer device exhibited a relatively high failure rate. The latches are used as a safety restraint, and failure in the field could result in severe injury. The failures were occurring after installation but prior to actual field use. Specifically, the failures occurred as cracking within retaining tabs used to secure a metal slide. The cracking was limited to an older design, with newer components showing no signs of failure. The latch assembly components were injection molded from an unfilled commercial grade of a polyacetal copolymer. As part of the evaluation, both failed parts representative of the older design and newer components were available for testing.

Tests and Results. A visual examination of the failed parts confirmed cracking within the retaining tab adjacent to the metal slide. The failures were present at consistent locations on all of the parts. The crack surfaces showed evidence of macroductility in the form of stress whitening within the final fracture zone exclusively. Throughout the visual examination, it was also apparent that the parts exhibited a very sharp corner formed by the retaining tab and the main body of the latch assembly body. Sharp corners are considered a poor design feature in plastic components, because they can result in severe stress concentration and can produce areas of localized poor fusion.

The fracture surface was further evaluated using SEM. The SEM examination showed a clear crack origin at the corner formed by the retaining tab. The crack origin areas exhibited brittle fracture features without signs of significant microductility. Secondary crazing was also apparent at the crack origin location. A typical crack initiation site is shown in Fig. 32. The overall features were suggestive of cracking caused by a relatively high strain rate event and/or very high stress concentration. The midfracture surface showed an increase in the apparent ductility, as evidenced by an overlapping morphological structure. The final fracture zone showed significant deformation and stretching, indicative of ductile
overload. A laboratory failure was created by overloading the tab from a nonfailed part in a manner consistent with the insertion of the corresponding metal slide. The laboratory fractures exhibited surface features that were in excellent agreement with those exhibited by the failed parts, as apparent in Fig. 32.

![Fig. 32 Scanning electron images showing excellent agreement between the features present within the crack initiation sites of (a) the failed latch assembly and (b) the laboratory fracture. Both surfaces showed relatively brittle fracture features. 59×](image)

The failed latch assembly material was analyzed using micro-FTIR in the ATR mode, and the obtained spectrum exhibited absorption bands characteristic of a polyacetal resin. It is significant to note that polyacetal copolymers and homopolymers cannot be differentiated spectrally, and a melting point determination is often used to distinguish between these materials.

Differential scanning calorimetry was used to analyze the latch material. The obtained results showed that the material underwent a single endothermic transition at approximately 165 °C (330 °F), characteristic of the melting point of a polyacetal copolymer. The results also showed that the part was somewhat undercrystallized. This was evident through a significant increase in the heat of fusion between the initial heating run and the second heating run, after slow cooling. Undercrystallization can reduce the mechanical strength of the molded article and is usually the result of molding in a relatively cold tool. The level of undercrystallization found in the failed parts, however, was moderate in nature and not thought to be a major factor in the failures. Thermogravimetric analysis was also performed on the latch material, and the obtained results were consistent with those expected for an unfilled polyacetal copolymer.
The latch material was also analyzed to determine its MFR. Parts representing the older, failed components and the newer, current design were evaluated. Both sets of molded parts produced results ranging from 10.7 to 11.0 g/10 min. This was in good agreement with the nominal MFR for the molding resin, 9.0 g/10 min. Throughout the analytical testing of the failed latch material, no evidence was found to indicate contamination or degradation of the molded parts. Mechanical testing was performed in order to assess the effect of the recent design change. Because of the configuration of the parts, standard mechanical testing could not be performed. Instead, a proof load test was devised to directly assess the stress required to produce failure within the tab, at an area consistent with the failure latch assembly. A direct comparison was made between the two sets of parts. The parts representing the older design, with the sharp corner at the retaining tab, produced an average value of 78.7 N (17.7 lbf) at failure. More importantly, the parts within this group produced an average tab extension of 0.76 mm (0.03 in.) at failure. The evaluation of the part representing the new design generated significantly different results. Specifically, the failures occurred at a higher load, 92.1 N (20.7 lbf), and a greater tab extension, 2.5 mm (0.10 in.). A comparison of the mechanical test results is shown in Fig. 33. This mechanical evaluation clearly illustrated the advantage afforded by the design change, effectively increasing the tab radius.

![Fig. 33](image.png)

**Fig. 33** A comparison of the mechanical test results, showing a significant improvement in the parts produced from the new design

Conclusions. It was the conclusion of the evaluation of the failed latch assemblies that the parts failed via brittle fracture associated with stress overload. The stress overload was accompanied by severe apparent embrittlement resulting from a relatively high strain rate event and/or significant stress concentration. The relatively sharp corner formed by the retaining tab was shown to be a primary cause of the failures, with the newer, redesigned parts producing superior mechanical test results.

Example 13: Failure of a Nylon Filtration Unit. A component of a water filtration unit failed while being used in service for approximately eight months. The filter system had been installed in a commercial laboratory, where it was stated to have been used exclusively in conjunction with deionized water. The failed part had been injection molded from a 30% glass-fiber- and mineral-reinforced nylon 12 resin.

Tests and Results. A visual examination of the filter component revealed significant cracking on the inner surface. The cracking ran along the longitudinal axis of the part and exhibited an irregular pattern. The surfaces of the part presented a flaky texture, without substantial integrity, and displayed significant discoloration. The irregular crack pattern, flaky texture, and discoloration were apparent on all surfaces of the part that had been in contact with the fluid passing through the component. Several of the crack surfaces were further examined using SEM. The fracture surface exhibited a coarse morphology, as illustrated in Fig. 34. The reinforcing glass fibers protruded unbounded from the surrounding polymeric matrix. The fracture surface also showed a network of secondary cracking. Overall, the observations made during the visual and SEM inspections were consistent with molecular degradation associated with chemical attack of the filter component material.
To allow further assessment of the failure, a mounted cross section was prepared through one of the cracks. The cross section, as presented in Fig. 35, showed a clear zone of degradation along the surface of the part that had contacted the fluid passing through the filter. The degradation zone extended into the cracks, which indicated massive chemical attack. The prepared cross section was analyzed using energy-dispersive x-ray spectroscopy, and the results obtained on the base material showed relatively high concentrations of silicon, calcium, and aluminum, with lesser amounts of sulfur and sodium in addition to carbon and oxygen. The results were consistent with a mineral- and glass-filled nylon resin. Analysis of the surface material, which exhibited obvious degradation, showed a generally similar elemental profile. However, significant levels of silver and chlorine were also found. This was important, because aqueous solutions of metallic chlorides are known to cause cracking and degradation within nylon resins.

The filter component material was further analyzed using micro-FTIR in the ATR mode. Analysis of the base material produced results characteristic of a glass- and mineral-filled nylon resin. However, analysis of the surface material showed additional absorption bands characteristic of substantial oxidation and hydrolysis of the nylon. A spectral comparison showing this is presented in Fig. 36. The presence of these bands is consistent with the high level of molecular degradation noted during the visual and SEM examinations.
Fig. 36 Fourier transform infrared spectral comparison showing absorption bands associated with hydrolysis at 3350 cm$^{-1}$ and oxidation at 1720 cm$^{-1}$ in the results obtained on the discolored surface.

Comparative TGA of the base material and the surface material also showed a significant difference. In particular, the results obtained on the surface material showed a lower temperature corresponding to the onset of polymer decomposition. This is illustrated in Fig. 37.
Conclusions. It was the conclusion of the evaluation that the filter component failed as a result of molecular degradation caused by the service conditions. Specifically, the part material had undergone severe chemical attack, including oxidation and hydrolysis, through contact with silver chloride. The source of the silver chloride was not established, but one potential source was photographic silver recovery.

Example 14: Failure of a PC Switch Housing. A housing used in conjunction with an electrical switch failed shortly after being placed into service. A relatively high failure rate had been encountered, corresponding to a recent production lot of the housings, and the failed part was representative of the problem. The housing had been injection molded from a commercially available, medium-viscosity grade of PC, formulated with an ultraviolet stabilizer. In addition to the PC housing, the design of the switch included an external protective zinc component installed with a snap-fit and two retained copper press-fit contact inserts. Control parts representing an earlier production lot were available for reference purposes.

Tests and Results. A visual examination of the submitted housing revealed massive cracking within the base of the part, including the retaining tabs securing the contacts. The fractures were primarily located adjacent to the copper contacts. Gray streaks, commonly referred to as splay, were also apparent on the PC housing, as shown in Fig. 38. Splay is often associated with molecular degradation, caused by insufficient drying or exposure to excessive heat, from the molding process. The visual examination also revealed that the contacts corresponding to the failed housing retaining tabs extended significantly, relative to contacts in nonfailed areas. This suggested a high level of interference stress between the contact and the tab. The fracture surface was further inspected using an optical stereomicroscope. The fracture surface showed no evidence of ductility, as would be evident in the form of stress whitening or permanent deformation. An oily residue was evident covering the crack surface.
Fig. 38  A view of the housing showing gray streaks characteristic of splay

The crack surface was further examined via SEM. The fracture surface exhibited multiple apparent crack origins and classic brittle fracture features, including hackle marks, river markings, and Wallner lines. A representative area on the fracture surface is shown in Fig. 39. No evidence of ductility, which would be apparent as stretched fibrils, was found. Overall, the observed features were indicative of brittle fracture associated with the exertion of stresses below the yield point of the material, over an extended period of time, by a creep mechanism.
Fig. 39 Scanning electron image showing characteristic brittle fracture features on the housing crack surface. 100×

The housing base material was analyzed using micro-FTIR in the ATR mode, and the resulting spectrum contained absorption bands characteristic of PC. The results produced an excellent match with a spectrum obtained on a reference part, without evidence of contamination. The oily residue found on the part, including the fracture surface, was also analyzed. The obtained spectrum was characteristic of an aliphatic hydrocarbon-based oil, with no signs of aromatic hydrocarbons or other chemicals known to produce stress cracking in PC resins.

The housing material was also analyzed using DSC. The DSC thermogram showed a single transition at 141 °C (286 °F), associated with the glass transition temperature ($T_g$) of the material. This temperature was somewhat lower than expected for a PC resin, which usually undergoes this transition closer to 150 °C (300 °F). This difference was thought to be an indication of potential molecular degradation.

Thermomechanical analysis (TMA) was used to evaluate the failed retaining tab material, using an expansion probe. The TMA results confirmed the relatively low $T_g$, in particular, with a comparison to a reference part. A comparison showing this is presented in Fig. 40. No evidence was found in the results to indicate molded-in residual stress.
Fig. 40 The TMA results obtained on the failed and reference parts. The results exhibit differences corresponding to a reduction in the glass transition of the failed material.

Melt flow testing of the housing samples showed the submitted reference part to have a MFR of 39.7 g/10 min, compared with 78.1 g/10 min for the failed components. The nominal value for the resin used to produce the housing was 9 to 12 g/10 min. This indicated not only severe molecular degradation within the failed housing material but also within the reference parts. The most likely source of the degradation was the molding process. This degradation was consistent with the presence of splay observed on the part as well as the reduced $T_g$.

Conclusions. It was the conclusion of the evaluation that the switch housings failed via brittle fracture, likely through a creep mechanism. The failure was caused by severe embrittlement of the housing resin associated with massive molecular degradation produced during the molding process. A potential contributing factor was the design of the part, which produced significant interference stresses between the contact and the mating retaining tab.

Example 15: Failure of Nylon Hinges. A production lot of mechanical hinges had failed during incoming quality-control testing. The hinges were used in an automotive application and had cracked during routine actuation testing. Similar parts had been through complete prototype evaluations without failure. However, a change in part supplier had taken place between the approval of the prototype parts and the receipt of the first lot of production parts. The mechanical hinges were specified to be injection molded from an impact-modified, 13% glass-fiber-reinforced nylon 6/6 resin. A resin substitution was suspected, corresponding to the supplier change. Samples representing the failed components and the original prototype parts were available for the failure investigation.

Tests and Results. A visual examination of the failed parts confirmed catastrophic cracking within the mechanical hinge in an area that would be under the highest level of stress during actuation. The failures did not show signs of macroductility, which would be apparent in the form of stress whitening and permanent deformation. The fracture surfaces of the failed parts were further inspected via SEM. While the presence of glass-reinforcing fibers can render a plastic resin inherently more brittle, a certain level of ductility is still expected at the 13% glass level. This ductility is often only apparent at high magnification and only between the individual glass fibers. However, the failed hinge components did not exhibit any signs of ductility even at high magnification, with the fracture surface showing only brittle features. A laboratory failure was created on one of the prototype parts by overloading the component. Examination of the fracture surface using SEM showed the normally anticipated level of ductility, as indicated by the overlapping, rose-petal morphology. The crack surfaces of both the failed part and the laboratory fracture are shown in Fig. 41.
Fig. 41  Scanning electron images showing (a) brittle fracture features on the failed hinge and (b) ductile fracture features on the laboratory fracture. 118x.

Analysis of the failed components and the corresponding molding resin via micro-FTIR produced results characteristic of a nylon resin. The molding resin and failed parts generated generally similar results. However, a distinct difference was apparent in that the spectra obtained on the failed parts showed an additional absorption band at approximately 1740 cm\(^{-1}\), indicative of partial oxidative degradation of the resin. A spectral comparison illustrating this is presented in Fig. 42. Because the parts had not yet been in service, this degradation was thought to have occurred during the molding process.
Fig. 42 Fourier transform infrared spectral comparison showing absorption bands at 1740 cm\(^{-1}\), characteristic of oxidation within the results obtained on the failed parts.

The failed parts were further tested using DSC. The obtained DSC results showed a melting point of 263 °C (505 °F), consistent with a nylon 6/6 resin. The molding resin was also analyzed via DSC, and a comparison of the results further indicated degradation of the molded nylon resin. This was apparent by a noted reduction in the heat of fusion in the results representing the failed parts.

The failed parts and the prototype parts were also analyzed using conventional thermogravimetric analysis (TGA), and both analyses produced results indicative of a nylon resin containing approximately 13% glass-fiber reinforcement. Further testing was performed using TGA in the high-resolution mode. This analysis was conducted in order to assess the level of impact-modifying rubber resin. The weight loss associated with the rubber was observed as a shoulder on the high-temperature side of the weight loss representing the nylon resin. This weight loss was particularly evident in the derivative curve. Because the weight losses could not be totally resolved, an absolute level of rubber could not be determined. However, a comparison of the results allowed a determination of the relative level of the impact modifiers in the two materials. This comparison showed a distinctly higher level of impact modifier in the prototype part material, relative to the failed part material.

Conclusions. It was the conclusion of this evaluation that the hinge assemblies failed through brittle fracture associated with stress overload during the actuation of the parts. The failed part material was found to be degraded, as indicated by both the FTIR and DSC analysis results. This degradation likely occurred either during the compounding of the resin or during the actual molding of the parts. A significant factor in the hinge failures is the conversion to a different grade of resin to produce the failed production parts as compared to the prototype parts. While both resins produced results characteristic of a 13% glass-fiber-reinforced, impact-modified nylon 6/6, the failed part material contained a significantly lower level of rubber. This decrease in rubber content rendered the parts less impact resistant and subsequently lowered the ductility of the molded hinge assemblies.
References


Characterization of Plastics in Failure Analysis

Jeffrey A. Jansen, Stork Technimet, Inc.

Selected References

• D. Wright, *Failure of Plastics and Rubber Products*, RAPRA Technology, Shawbury, U.K., 2001
Stress Analysis and Fracture Mechanics

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

Introduction

THE STRESSES acting on a component may cause unacceptable deformation, cracking, or fracture. By definition, after a stress related failure has occurred, it is a given that the stresses reached a critical level, at least at one critical location. Or it may be that fracture did not occur. It should be remembered that some parts are designed to fracture under certain circumstances. Not doing so may constitute the failure. Shear pins, for example, are commonly used to protect valuable equipment by acting as mechanical fuses. Stress directors in helicopter gearing are also designed so that if gear tooth fracture occurs, only half the tooth breaks off. This leaves the other half of the gear tooth to carry the load, simultaneously creating sufficient vibration to warn the pilot of imminent failure but allowing him sufficient flying time to land.

In any case, the goal of failure analysis is to understand a situation sufficiently well to determine if changes in design, manufacture, or other areas are needed. When confronted with a cracked, fractured, or deformed component, the failure analyst will usually seek to answer some basic questions:

- Were the loads and stresses encountered by the part at the level anticipated during design? Or did some unexpected condition(s) contribute to the failure?
- Was the material in the area of the cracking or deformation capable of meeting the conditions anticipated during design? Was there some deficiency or discontinuity that contributed to the failure, or was there a local stress raiser at the critical location? Was this taken into account by the designer?

Of course, the root causes of unacceptable stress concentrators are varied. They may occur from design (e.g., changes in section), manufacturing (e.g., surface condition deleterious inclusions or unacceptable porosity), and/or service exposure (e.g., corrosion and wear). The critical point is that if the causes (i.e., stresses and associated loads) contributing to the failure are not sufficiently understood, then it may be most difficult to prevent it from happening again. For example, consider a part with a fillet radius that is somewhat sharper than indicated on a drawing, and that the part fails at the fillet radius. Did the part fail because the fillet radius was out of design tolerance, or did it fail because of some other application or manufacturing factor(s)? A safety factor was probably used when the drawing was created with a maximum assumed load, but if the part was loaded well above the original design load, it could have failed whether the fillet radius was correct or not.

There are many reasons why a failure may occur, and a complete picture is required to gain insights on a situation that is potentially complex (and apparently unanticipated as well because, by definition, failures are unanticipated). The root-cause circumstances of a failure are therefore potentially subtle, complex, and/or synergistic. In this context, the analyst must be aware of not only the details of the component and the stress analysis, but also the “big picture.” It is common for the materials failure analyst to identify the fracture origin and any material discontinuity at the origin. However, did the welds meet the conditions for “fitness-of-purpose” requirements? Were the weld codes and standards applicable for the actual conditions?

A part is generally expected to fracture at the maximum stress location unless there is a material “deficiency” at some other location. However, if the deficiency is an expected material deviation that is properly anticipated during design, then the material condition may not be the root cause. In this case, a design review may be initiated with a more extensive analysis of loads and stresses. This design review also may find that none of the expected stresses are sufficient to explain the condition of the component (i.e., the designer correctly accommodated the anticipated stresses). In this case, the source of the additional stress may be unanticipated loads and possibly, inappropriate use. In this regard, the possibilities are many, such as operator overloading, unusual resonance vibration, or unexpected stress introduced as a result of assembly. The critical point is that
the analyst must look at the overall situation, including the loads and the distribution of stresses within the configuration. Is there an additional load/stress that was not included in the analysis, which changes the stress distribution? Or is the stress analysis incorrect?

In many cases, however, stress analysis alone is not sufficient to explain how a fracture occurred. The failure analyst must also apply the principles of fracture mechanics to understand how a crack developed and grew. The issue in this portion of the failure analysis is to determine if the expected stresses would have been sufficient to initiate, propagate, and grow a crack at the specific location and under the conditions for which the crack actually developed. The most insidious failures are those in which one would not expect fracture from consideration of the stress level alone, but cracking and ultimate fracture nevertheless do occur, often due to a combination of localized stress concentration, fluctuating loads, and/or aggressive environment.

This article describes concepts and tools that can be used by the failure analyst to address the above questions and additional queries that naturally follow. Only an overview and review of some of the basics is possible in an article such as this. The reader is referred to some of the many excellent texts (see the “Selected References” for this article) for further study. The perspective of this article is that of the failure analyst who is confronted with a problem; that is, a failed part, and asked to understand how it happened and how to keep it from happening again.

Throughout the material herein, it also should be kept in mind that the application of stress analysis and fracture mechanics tools is somewhat different than it is in design. The goal of the designer is to prevent failures from occurring by assuring that stresses in service do not reach a critical level. Uncertainty and unknowns in analysis, manufacture, service, and material conditions are typically addressed by applying so-called “factors of safety.”

**Stress Analysis and Fracture Mechanics**

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

**Analysis of Applied Stresses**

Stress analysis provides a road map for the failure analyst. If fracture initiates in a low stress area, several working hypotheses are suggested. The part may not have been used as intended. It may have suffered mechanical or environmental damage prior to use or the presence of a large material flaw may have predisposed it to failure by creating a point or volume in which the actual stress exceeds the strength of the material. Additionally, the microstructure may vary from location to location. On the other hand, if failure occurs in an area of high stress, overload may also be suspected. Of course the possibility of improper design cannot be discounted. Knowledge of the stress distribution can point the failure analyst to the crack origin, especially if the crack surface is nearly featureless. It can also help check for consistency with the fracture features.

In the conventional approach to design and failure analysis, an elastic stress analysis of the component is necessary. Stress analysis is initially performed in the design stage of a development project. Frequently, principal stresses and maximum shear stresses are important to the designer because both are used in common failure expressions to calculate maximum load capability (e.g. multiaxial yield criteria). One failure mode theory is the Rankine (or maximum normal stress) criterion, where the maximum principal stress is used. The other is the Tresca theory based on the maximum shear stress. The Rankine criterion holds that inelastic deformation at a point begins when the maximum normal stress reaches the yield strength of the material. Tresca’s theory states that yield at a point begins when the maximum shear stress equals the shear stress in a tension specimen when it begins to yield. Another common failure model is the von Mises criterion, which is based on a determination of distortion energy at yield and recognizes that in a constant-volume isotropic material, hydrostatic stresses cannot cause failure by yield.

An octahedral shear stress criterion provides a mathematically identical expression to the von Mises criterion. Several other criteria have been proposed (e.g., one based on maximum strain). Additionally, some criteria take into account polar behavior (unequal tensile and compressive failure stresses), anisotropic behavior and nonconstancy of volume during deformation. The historically early models (Rankine, Tresca, von Mises, and octahedral shear) were phenomenological and as proposed could be applied to either failure by yield or failure by fracture. They have also been applied to cyclic loading. In practice, the Tresca and von Mises criteria are most often used today to predict yield in isotropic constant-volume materials. The Rankine model is not often used because of its inability to predict behavior when one principal stress is negative and another positive. When used to predict failure of materials that are not very ductile, none of the
The file is downloaded from www.bzfxw.com

Stress Analysis and Fracture Mechanics

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

Fundamentals of Stress Analysis

Most of the conventional failure criteria are stress based. As such, the stress analysis of a geometrical component is important for design, reliability analysis, and failure analysis. This stress analysis is usually based on the mechanics of materials approach but could also use the theory of elasticity approach in some cases. There are many references in the literature that describe the state of stress for common geometries (Ref 1, 2, 3, 4), so in many instances it is not necessary to do original work but rather to know how to apply the equations that describe the state of stress.

In this section, issues related to the determination and use of stress are detailed. Stress is defined, and a procedure to deal with stress by determining maximum values through stress transformation is described. The stress analysis equations of typical component geometries are given, and some of the implications of the stress analysis relative to failure in components are also discussed. The component stresses determined by analysis usually lie in directions defined by the shape of the component geometry, for example, along major axes or perpendicular to major axes. These stresses may not be the maximum ones, so they must then be examined to find the actual maxima to be used in a failure criterion. These maxima are determined by taking stress transformations to find maximum normal stresses (principal stresses) and maximum shear stresses as well as maximum shear stresses. A procedure for calculating the planes on which the principal stresses and maximum shear stresses act is described in the section “Transformation of Stresses” in this article.

Definitions
Stress Definition. A simple definition of stress is as a force per unit area. In a geometry that has a uniform stress distribution, such as a bar in simple tension, a force, \( F \), applied along the axis of the bar that has area, \( A \), has uniform stress, \( \sigma \), where:

\[
\sigma = \frac{F}{A} \quad \text{(Eq 1)}
\]

See Fig. 1(a). If the force is perpendicular to the area as shown, the stress, \( \sigma \), is defined as an average normal stress. In a similar way a body may have a force, \( V \), that is parallel to the area, \( A \). The stress that is generated by this force is defined as a shear stress and usually given the symbol, \( \tau \). Then:

\[
\tau = \frac{V}{A} \quad \text{(Eq 2)}
\]

is the average shear stress (see Fig. 1b).

For most geometrical configurations, the stress is not uniform but varies as a function of position in the body. A more general definition of the stress is then given as follows. An arbitrary body having forces \( F_i \) applied (such that the body is in equilibrium) is cut, revealing an internal surface (Fig. 2). The internal surface has a distribution of force that is needed for equilibrium of the cut section. Over an increment of area, \( \Delta A \), an increment of force, \( \Delta F \), is required. Then the stress vector, \( S \), is defined as:

\[
S = \lim_{\Delta A \to 0} \left( \frac{\Delta F}{\Delta A} \right) \quad \text{(Eq 3)}
\]
If the force $\Delta F$ is resolved into mutually perpendicular components such that $\Delta N$ is a component normal to the area, $\Delta A$, and $\Delta V$ is parallel to $\Delta A$, then the normal component of stress, $\sigma$, and shear component of stress, $\tau$, are defined as:

$$\sigma = \lim_{\Delta A \to 0} \left( \frac{\Delta N}{\Delta A} \right)$$

(Eq 4)

and

$$\tau = \lim_{\Delta A \to 0} \left( \frac{\Delta V}{\Delta A} \right)$$

(Eq 5)
The force $\Delta F$ and its components are functions of their position on the cut surface. The possibility that the stresses $\sigma$ and $\tau$ are functions of position on the surface must also be permitted (as in the stress field around a stress concentrator or a stress state in bending or torsion). In addition, if the surface is cut in a different orientation, the forces are different and the stresses are different as well. Therefore the stresses, described with reference to a coordinate system are functions of position in a body as well as the orientation of the area over which the stress is defined and the coordinate system used to determine stresses.

The stress definitions given by Eq 4 and 5 is generated as a result of forces applied to a surface of the body. These stresses derived from surface forces are often labeled as primary stresses. Stresses can also be generated as a result of body forces and strain mismatches. Stresses coming from strain mismatch are often labeled secondary stresses. Common secondary stresses that are encountered are residual stresses and thermal stresses.

Residual stresses are important in failure analysis. They are different from the applied stresses in that when the forces are released, the stresses created by applied loads go to zero, but the residual stresses remain locked in the body. The way to treat residual stresses has been investigated and is still a matter of concern. Some methods to deal with it are discussed later.

Stress Components. The description of stress at a point in three dimensions requires the labeling of stresses that act on three mutually perpendicular planes and in three mutually perpendicular directions. Most generally, there are nine components of stress at a point. The labels can be given for the type of coordinate axes chosen, for example, Cartesian, polar, or spherical. For defining stress components, the labels may be given in Cartesian coordinates where the mutually perpendicular axes are labeled $x$, $y$, and $z$. Keeping the stresses defined the same as previously, $\sigma$ for normal stress and $\tau$ for shear stress, the nine stress components are defined with the subscripts referring to the coordinate directions. For example, $\sigma_{xx}$ is the normal stress acting on a plane perpendicular to the $x$ coordinate direction and acting in the $x$ direction. Similarly, $\tau_{xy}$ is the stress acting on a plane normal to $x$ and acting in the $y$ direction. Using this nomenclature, nine stress components can be defined by:

$$
\begin{bmatrix}
\sigma_{xx} & \tau_{xy} & \tau_{yx} \\
\tau_{yx} & \sigma_{yy} & \tau_{yz} \\
\tau_{zx} & \tau_{zy} & \sigma_{zz}
\end{bmatrix}
$$

These nine components are said to comprise the stress tensor. They are shown in Fig. 3 on a stress element.

![Fig. 3 Three-dimensional stress components](image)
Stress Transformations

Internal forces as well as those on internal areas are functions of the orientation of the forces and areas. The redefining of stress components from one orientation to another is accomplished by an operation referred to as stress transformation. In a component design, reliability analysis, or failure analysis, the use of the stress transformation is essential so that the highest and most damaging stresses can be determined and applied in the analysis. In most stress analyses, the formulas that are developed from mechanics of materials principles or from theory of elasticity are in directions that correspond to the geometrical construction of the body. For example, Cartesian coordinates can be taken along an axis of a bar and perpendicular to the axis. For round shapes such as pipes or pressure vessels, the radial, circumferential, and axial directions can be chosen. These are taken for convenience in the analysis of the stresses in these components. However, finding the stresses in the directions corresponding to the body dimensions does not guarantee that they are the maximum and most damaging stresses. The ability to transfer stresses to another orientation (reference frame) is essential in determining these maximum stresses. The methods of stress transformations are discussed first, and then the determination of stress maxima is discussed.

To define stress transformation, it is easier to start with a two-dimensional case. One common way to define a two-dimensional (2-D) stress state is with a plane stress assumption. Plane stress assumes that all stresses in the third direction, \( z \), are zero so that:

\[
\sigma_z = 0 \quad \tau_{xz} = 0 \quad \tau_{yz} = 0
\]

(Eq 6)

Also, the plane stress assumption requires the nonzero stresses to be only functions of \( x \) and \( y \).

The stress transformation technique can be illustrated easily in two dimensions by using the principles of statics. In two dimensions, only three of the six stress components are used: \( \sigma_x, \sigma_y, \tau_{xy} \). These can be shown on a 2D stress element, Fig. 4a, where \( x \) is a horizontal component and \( y \) a vertical component. The element is cut at an angle, \( \theta \), and new axes defined as \( x' \) and \( y' \) are in the direction defined by a rotation of the axes through the angle \( \theta \) (Fig. 4b). A new stress element defined by the cut surface takes the shape of a triangle defined by one face perpendicular to \( x \), a second face perpendicular to \( y \), and the third perpendicular to the \( x' \) direction. The element is taken to have a unit thickness such that the area of each face can be defined by \( \theta \) and the area of either the \( x \) or \( y \) face as shown in Fig. 4(c). The stress components on the face perpendicular to the \( x' \) direction are labeled \( \sigma_{x'} \) and \( \tau_{x'y'} \).

![Stress Transformation Diagram](image)

**Fig. 4** Stress transformation in two dimensions. (a) Two dimensional stress at a point with a cut angle \( \theta \). (b) Stresses along rotated axes, \( x', y' \). (c) Area values on cut element

Static equilibrium can be applied to forces, not stresses. Therefore, the stresses have to be converted to force by multiplying by an appropriate area. In Fig. 4(c) the areas of the primed faces are related to the original faces by the angle \( \theta \). If \( A \) is the area normal to \( x' \), then \( A \cdot \cos \theta \) is the area normal to the \( x \) axis and \( A \cdot \sin \theta \) is the area normal to the \( y \) axis.
The force in the \( x' \) direction is \( \sigma x' A \). This can be used to define equilibrium in the \( x' \) and \( y' \) directions by using the statics equilibrium equations:

\[
\sum F_{x'} = 0 \quad \sum F_{y'} = 0 \quad (\text{Eq 7})
\]

Equilibrium in the \( x' \) direction gives:

\[
\sum F_{x'} = 0 = \sigma x(A) - \sigma x(A \cos \theta) \cos \theta - \sigma y(A \sin \theta) \sin \theta - \tau_{xy}(A \cos \theta) \cos \theta
\]

(Eq 8)

Solving for \( \sigma x' \) gives:

\[
\sigma x' = \sigma x \cos^2 \theta - \sigma y \sin^2 \theta - 2 \tau_{xy} \cos \theta \sin \theta
\]

(Eq 9)

These are usually transformed to double angle format by:

\[
\cos 2\theta = (1 + \cos 2\theta)/2 \\
\sin 2\theta = (1 - \cos 2\theta)/2
\]

(Eq 10)

\[
2 \tau_{xy} \cos \theta \sin \theta = \sin 2\theta
\]

(Eq 11)

Writing the sum of the forces in the \( y' \) direction, an expression can be developed for \( \tau_{x'y'} \):

(Eq 12)

For a cut face with perpendicular direction, \( y' \), the same equilibrium can be analyzed or \( \theta \) simply can be replaced by \( 90^\circ + \theta \) in Eq 11 to get \( \sigma y' \).

(Eq 13)

Equations 11 to 13 are the 2D stress transformations. In a more complex way, a three-dimensional (3D) stress transformation equation can be defined for an arbitrary direction \( x' \) in 3D space. If the axis \( x' \) makes angles with the original \( x, y, \) and \( z \) axes defined by direction cosines \( l, m, \) and \( n \) where \( l \) is the cosine of the angle between \( x \) and \( x' \), \( m \) is the cosine of the angle between \( y \) and \( x' \), and \( n \) is the cosine of the angle between \( z \) and \( x' \), then the stresses on the plane normal to \( x' \) would be:

\[
\sigma x' = \frac{\sigma x + \sigma y}{2} - \frac{\sigma x - \sigma y}{2} \cos (2\theta + 180^\circ)
\]

\[
+ \tau_{xy} \sin (2\theta + 180^\circ)
\]

\[
\sigma y' = \frac{\sigma x + \sigma y}{2} - \frac{\sigma x - \sigma y}{2} \cos 2\theta - \tau_{xy} \sin 2\theta
\]

(Eq 14)

The 3D stress transformation is complex as are most stress analyses in three dimensions. Therefore, in most analyses an attempt is made to define conditions to achieve a 2D stress state such as plane stress. Therefore, most of the discussion will concentrate on 2D stress operations. However, it is essential to recognize those cases for which a 2D solution does not obtain the maximum shear stress in the body or the plane on which it acts (see the section “Three-Dimensional Stress Analysis” in this article).

One of the simple tools that is used to perform 2D stress transformation is the construction of Mohr’s circle. With the use of Mohr's circle, a stress transformation or a determination of maximum stresses can be made with relative ease and without the memorization of all the stress transformation equations. Squaring and adding the 2D transformation equations, Eq 11 and 12, gives:

\[
\left( \frac{\sigma x' - \sigma x + \sigma y}{2} \right)^2 + \tau_{x'y}^2 = \left( \frac{\sigma x + \sigma y}{2} \right)^2 + \tau_{xy}^2 \quad (\text{Eq 15})
\]

It was noted that Eq 15 is the equation of a circle with axes \( \sigma \) and \( \tau \) and centered on the \( \sigma \) axis with:

\[
\text{center} = \left( \frac{\sigma x + \sigma y}{2} \right)
\]

(Eq 16)

and radius, \( R \), given by:
The result is that stress transformations can be performed by using the geometric principles of a circle. For example, if stresses are known at a point, they are plotted on a figure that has horizontal axis $\sigma$ and vertical axis $\tau$ so that $(\sigma_x, \tau_{xy})$ is the coordinate point at horizontal position $\sigma_x$ and vertical position $\tau_{xy}$ and so that $(\sigma_y, \tau_{xy})$ is the coordinate that corresponds to $\sigma_y$ and $\tau_{xy}$. Since the center of the circle is on the $\sigma$ axis, this can be easily found by Eq 16. The radius of the circle is given by Eq 17. From this, the entire circle can be drawn. This is illustrated by the schematic Mohr’s circle in Fig. 5.

\[ R^2 = \left( \frac{\sigma_x + \sigma_y}{2} \right)^2 + \tau_{xy}^2 \]  

(Eq 17)

Fig. 5  Mohr's circle for two-dimensional stress transformation

Then for a rotation of axes by $\theta$ in the stress element, the position on the Mohr’s circle must go through a rotation of $2\theta$ in the same direction but around the circumference of the circle. Every point on the circle corresponds to a possible stress pair. Note that a rotation of 180° of the circle corresponds to a rotation of the stress element by 90° so that $\sigma_i$ is transformed to $\sigma_j$ as shown in Fig. 5.

Principal Stresses. The purpose of a stress transformation is mainly to find the stresses that can be used in a failure criterion. These would be either the largest magnitude stresses in any direction or the magnitude of stresses on a weak plane. For the former, the extreme stresses can be found by taking Eq 11 and applying the calculus principle $d\sigma_x/d\theta = 0$. That is, principal stresses are normal (perpendicular) stresses on planes for which the shear stresses are zero. The values of principal stresses are given by:
where the positive radical gives a maximum stress labeled $\sigma_1$ and the negative radical a minimum stress $\sigma_2$. Notice that the positions of the extreme stresses are the points on the Mohr's circle where the circle crosses the $\sigma$ axis (horizontal axis) (Fig. 5). The right side of the circle corresponds to a maximum normal stress and the left side to a minimum stress. These stresses as called principal stresses and are often labeled $\sigma_1$ and $\sigma_2$. Notice that the extremal stress given by Eq 18 is equivalent to taking the center $\pm$ the radius of the Mohr’s circle. In this way the extreme values of stresses can be easily found by simply using the Mohr’s circle.

The maximum shear stress, $\tau_{\text{max}}$, can also be found by taking $\tau_{xy}'$ from Eq 12 and using the calculus principal $d\tau_{xy}/d\theta = 0$. This results in:

$$\tau_{\text{max}} = \pm \sqrt{\left(\frac{\sigma_x + \sigma_y}{2}\right)^2 + \tau_{xy}^2} \tag{Eq 19}$$

The value of $\tau_{\text{max}}$ is the radius of the Mohr’s circle, and the corresponding points on the Mohr’s circle are at the top and bottom of the circle as shown in Fig. 5. The maximum shear stress could also be written as:

$$\tau_{\text{max}} = \frac{\sigma_1 - \sigma_2}{2} \tag{Eq 20}$$

Note that the maximum normal stresses are defined at a position where the shear stresses are zero. A plane with zero shear stress is defined as a principal stress plane, that is, a plane having only normal stresses. This can be extended to three dimensions. Given a symmetric stress tensor as defined previously, a transformation is desired to a plane that results in the stress tensor having only normal stress components; all shear components are zero. This would be given by:

$$\begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \\ 0 & 0 & \sigma_3 \end{pmatrix}$$

The transformation results in a cubic equation in $\sigma$ given by:

$$\sigma^3 - I_1\sigma^2 + I_2\sigma - I_3 = 0 \tag{Eq 21}$$

where

$$I_1 = \sigma_x + \sigma_y + \sigma_z$$
$$I_2 = \sigma_x\sigma_y + \sigma_y\sigma_z + \sigma_z\sigma_x - \sigma_{xy}^2 - \sigma_{xz}^2 - \sigma_{yz}^2$$
$$I_3 = \sigma_x\sigma_y\sigma_z + 2\sigma_{xy}\sigma_{xz}\sigma_{yz} - \sigma_x^2\sigma_{yz} - \sigma_y^2\sigma_{xz} - \sigma_z^2\sigma_{xy} \tag{Eq 22}$$

The values of the coefficients, $I_1$, $I_2$, and $I_3$ are called the stress invariants and are independent of the coordinate system used to describe the stress state. However, here they are used to solve the stress cubic given in Eq 21 so that three principal stresses result: $\sigma_1$, $\sigma_2$, and $\sigma_3$. In three dimensions, the maximum shear stresses can be found in the same way to the 2D stresses, that is as one-half the difference of any two principal stresses. If stresses $\sigma_1$, $\sigma_2$, and $\sigma_3$ are ordered from maximum to minimum, then the absolute maximum shear stress in three dimensions is:

$$\tau_{\text{max(abs)}} = \frac{\sigma_1 - \sigma_3}{2} \tag{Eq 23}$$

The transformation of stress in three dimensions can be made with equations like Eq 14. For this transformation, a circle like the Mohr's circle cannot be defined. There is thus, no simple graphical way to transform the stresses or determine $\sigma_1$, $\sigma_2$, $\sigma_3$, and maximum shear stresses for a general 3D system of stresses. However, once $\sigma_1$, $\sigma_2$, and $\sigma_3$ are known from the solution of the cubic equation (Eq 21), it is possible to draw three circles (Fig. 6) having centers, $C$, at

$$C_1 = \frac{\sigma_1 + \sigma_2}{2}; \quad C_2 = \frac{\sigma_2 + \sigma_3}{2}; \quad C_3 = \frac{\sigma_1 + \sigma_3}{2} \tag{Eq 24}$$

and radii
Fig. 6 Special three-dimensional Mohr's circle defined with principal stresses

These circles show the three principal stresses and the corresponding maximum shear stresses. Using them makes it easy to see that the absolute maximum shear stress is given by Eq 23. They show that all of the principal stresses occur on planes that have zero shear stress. These three-dimensional (3D) Mohr's circles cannot be used for a general stress transformation but can be used for a particular transformation. If in one direction the principal plane is held fixed, then the circle that does not contain that point is the one for the other directions. A rotation of this circle gives all of the stresses that can be obtained from a rotation of the 3D element, holding the axes fixed in the perpendicular plane. For example, if the z axis is chosen for \( \sigma_3 \) direction, then a rotation of points on the circle containing \( \sigma_1 \) and \( \sigma_2 \) would define all of the stress states in the planes perpendicular to z that are the x-y plane for Cartesian coordinates.

There are cases in which a Mohr's circle construction can be used for a case of 3D stresses. Those cases are ones in which only one shear stress acts in the body. That is, if the stress tensor has the general form:

\[
\begin{bmatrix}
\sigma_{xx} & \sigma_{xy} & 0 \\
\sigma_{yx} & \sigma_{yy} & 0 \\
0 & 0 & \sigma_{zz}
\end{bmatrix}
\]

In this case, \( \sigma_{zz} \) is a principal stress. Because the directions of the principal stresses are mutually orthogonal, the other two principal stresses must lie in the xy plane and can therefore be determined from the two-dimensional case considering \( \sigma_{xx}, \sigma_{yy}, \) and \( \tau_{xy} \).

This case clearly illustrates the dangers of considering only the applied loads in determining the maximum shear stress in the body. Consider a case of balanced biaxial tension (\( \sigma_{xx} = \sigma_{yy} \)) and consider only the two-dimensional (2D) case. It is incorrectly concluded that there is no shear stress in the body because the radius of the Mohr's circle in the xy plane has been reduced to a point. However, when the 3D case is considered, \( \sigma_{zz} \) is located on the \( \sigma-\tau \) axes, and now there are two circles having equal radius given by \( \frac{1}{2} (\sigma_{xx} - \sigma_{zz}) \) and \( \frac{1}{2} (\sigma_{yy} - \sigma_{zz}) \) and a third circle of zero radius given by \( \frac{1}{2} (\sigma_{xx} - \sigma_{yy}) \) so that there are now four planes on which the same maximum shear stress exists.

One plane of interest for defining stresses is called the octahedral plane (or \( \{1, 1, 1\} \) plane in crystallographic terminology). It is defined as the plane with normal \( I = M = N \); that is, the normal to the plane makes equal angles to the x, y, and z axes. On this plane, there is a normal stress given by:

\[
\sigma_{net} = \frac{\sigma_x + \sigma_y + \sigma_z}{3} = \frac{\sigma_1 + \sigma_2 + \sigma_3}{3}
\]

(Eq 26)

and a shear stress given by:

\[
R_1 = \frac{\sigma_1 - \sigma_2}{2}, \quad R_2 = \frac{\sigma_2 - \sigma_3}{2}, \quad R_3 = \frac{\sigma_3 - \sigma_1}{2}
\]
The shear stress on the octahedral plane is often given in the form of an equivalent or effective stress, $\sigma_e$, given by:

$$\sigma_e = \frac{1}{\sqrt{2}} \left( \sqrt{(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2} - \frac{1}{3} \sqrt{\left(\sigma_1 - \sigma_2\right)^2 + \left(\sigma_2 - \sigma_3\right)^2 + \left(\sigma_3 - \sigma_1\right)^2} \right)$$  (Eq 27)

The shear stress on the octahedral plane is often given in the form of an equivalent or effective stress, $\sigma_e$, given by:

$$\sigma_e = \frac{1}{\sqrt{2}} \sqrt{(\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 + (\sigma_3 - \sigma_1)^2}$$  (Eq 28)

This effective stress and the von Mises stress have equal magnitude in this case, and along with the maximum shear stress is used to describe the yield condition in ductile materials. Alternatively, the effective stress could be described in terms of the Tresca criterion:

$$\sigma_e = \frac{\sigma_1 - \sigma_3}{2}$$

Other effective stresses can be defined for polar and/or anisotropic materials.

Three-Dimensional Stress Analysis. Although many practical stress states can be analyzed in two dimensions, the more complex 3D case is now examined. Complicated loading and/or geometry may induce a stress state in which none of the stress components is zero. For example, normal and tangential surface tractions as occur in combined rolling and sliding problems often generate a local 3D stress state, and failures are often associated with the attendant sliding friction. Notches, which are frequent crack initiation sites, when subjected to external loads often generate triaxial stress states and locally embrittle the material, predisposing it to failure. Thus, manipulation of the entire stress tensor is necessary for analysis.

As noted previously, Mohr’s circle is a graphical representation of the 2D stress transformation equations. The cases described previously are convenient to show how Mohr’s circle can be used to obtain the principal stresses and the planes of maximum shear stress. Three pairs of principal stresses define six planes of relative maximum shear stress (the sign of a shear stress does not matter). For purposes of failure analysis, it is important to know that the largest of these stresses has been obtained. A simple plane stress example illustrates the point. Consider a case of in-plane balanced biaxial tension of magnitude $S$; how does one determine the magnitude of the maximum shear stress and the plane(s) on which it exists? Treating this as a 2D problem would be incorrect, since every planar stress transformation leads to the same stress tensor, namely biaxial tension. Mohr’s circle is a point. By improperly neglecting the third principal stress, even though it is zero, one would conclude that no shear stresses are present, when, in fact, $\tau_{\text{max}}$ is $S/2$ and acts on two perpendicular planes.

Consider next the case of a torsional stress, $\tau$, combined with an axial stress, $\sigma_1$. The 2D stress tensor is:

$$\begin{bmatrix} \sigma_1 & \tau \\ \tau & 0 \end{bmatrix}$$

Mohr’s circle correctly predicts the maximum shear stresses to be:

$$\tau_{\text{max}} = \pm \sqrt{\left(\frac{\sigma_1}{2}\right)^2 + \tau^2}$$

However, if a second direct stress, $\sigma_2$, is added such that the 2D stress tensor becomes:

$$\begin{bmatrix} \sigma_1 & \tau \\ \tau & \sigma_2 \end{bmatrix}$$

then Mohr’s circle may or may not give the correct maximum shear stress. It can be easily shown that when $\sigma_1$ and $\sigma_2$ are both positive or both negative, the incorrect maximum shear stress is incorrectly calculated assuming a two dimensional problem. This is because the third normal principal stress (which happens to have zero magnitude) has been neglected. Next consider plane strain. If $\varepsilon_z = 0$, then plane strain conditions exist wherever $\gamma_{xz} = \gamma_{yz} = 0$. Plane deformation, in which $u_z = 0$ (where $u_z$ is displacement in $z$-direction) and both $u_x$ and $u_y$ are independent of $z$, also leads to plane-strain conditions. It is approximated in the case of a long prismatic cylinder whose lateral-surface (parallel to the $z$-axis) boundary conditions are uniform (hence independent of $z$) along the length ($z$-axis) and are applied parallel to the $x$-$y$ plane. Appropriate end conditions are $u_z = 0$, $\varepsilon_{xz} = 0$ and $\tau_{xz} = 0$. 
Suppose that a linearly elastic material has Poisson's ratio, $\nu$, of 0.5, and that $\sigma_x = \sigma_y$. Then by Hooke's law $\sigma_z = \sigma_y = \sigma_x$, that is:

$$\varepsilon_z = \frac{1}{E} \left[ \sigma_z - \nu (\sigma_x + \sigma_y) \right]$$

Therefore, the hydrostatic stress ($\sigma_h$) is given by:

$$\sigma_h = \frac{\sigma_x + \sigma_y + \sigma_z}{3} = \frac{3\sigma}{3} = \sigma$$

This describes a hydrostatic stress state, that is, the stress tensor, $T$, is defined as:

$$T = \begin{bmatrix} \sigma_h & 0 & 0 \\ 0 & \sigma_h & 0 \\ 0 & 0 & \sigma_h \end{bmatrix}$$

Even if Poisson's ratio is unequal to 0.5, a hydrostatic stress whose magnitude is:

$$\sigma_h = \frac{(1+\nu)(\sigma_x + \sigma_y)}{3}$$

still exists. The effect of the hydrostatic stress is to raise the stress level at which yield occurs, causing it to approach the fracture stress. In a constant-volume material (metals, Ceramics), the hydrostatic stress cannot cause yield, but a hydrostatic stress does increase the magnitude of the stresses created by the applied loads at which yield occurs. In a nonconstant volume material (polymers), the hydrostatic stress affects yield behavior, so that a von Mises, octahedral shear, or Tresca-type criterion is inaccurate.

In application, consideration of the plane strain state of stress is significant in understanding macroscale brittle fracture surface orientation in tensile loading and the difference in fracture toughness for plane stress and plane strain loading (see the section “Fracture Mechanics” in later paragraphs). The difference in plane-stress and plane-strain loading is also important in understanding differences in fracture surface orientation and neck shape during macroscale ductile fracture of plate versus cylindrical tensile components (see the article “Mechanisms and Appearances of Ductile and Brittle Fracture in Metals” in this Volume).

Loading relatively thick components in tension or bending and which contain notches, cracks, or similar stress concentrators parallel to the thickness direction, taken as the $z$ direction to be consistent with the notation above, results in plane-strain loading conditions. The material at the tip of the notch or crack (area of maximum stress concentration) is constrained by the adjacent material, which is not as highly stressed. Thus, $\varepsilon_z$ is constrained to be (close to) zero. As discussed in other sections of this article and other articles in this Volume, the fracture toughness observed under plane-strain conditions is a minimum and is considered a material property, $K_{IC}$. The underlying explanation for this observation follows from the preceding paragraph. Plane strain conditions increase the hydrostatic stress and inhibit yielding.

Under purely hydrostatic loading, no shear stresses exist, as is easily shown by examination of the three Mohr's circles. The circles reduce to zero radius. Hydrostatic stress is sometimes referred to as spherical stress. (A stress state in which a hydrostatic stress exists is sometimes incorrectly referred to as a case of triaxial stress. Triaxial loading indicates only that the stresses act in three directions in the body.) Spherical stress is perhaps a better term since it conveys the omnidirectional nature of the loading and suggests the absence of shear stresses. The three 3D Mohr's circles reduce to circles of zero radius; that is, they become a point in a manner analogous to the case of biaxial tension analyzed in two dimensions. The difference, however, is that in a spherical stress state, no shear stresses are present, while under biaxial tension or compression, the 2D analysis leads to the incorrect conclusion that shear stresses are absent. This only highlights the need for proper stress analysis.

### Stress-Strain Relationships

Stress and strain in an isotropic body for elastic deformation are related by linear expressions that result from Hooke's law. These are given by:

$$\varepsilon_x = \frac{\sigma_x}{E} - \frac{\nu}{E} \left( \sigma_y + \sigma_z \right)$$  \hspace{1cm} (Eq 29)

$$\varepsilon_y = \frac{\sigma_y}{E} - \frac{\nu}{E} \left( \sigma_x + \sigma_z \right)$$  \hspace{1cm} (Eq 30)
\[ \varepsilon_x = \frac{\sigma_x}{E} - \frac{v}{E} \left( \sigma_x + \sigma_y \right) \]  
\(\text{(Eq 31)}\)

\[ \gamma_{xy} = \frac{\tau_{xy}}{G}; \quad \gamma_{xz} = \frac{\tau_{xz}}{G}; \quad \gamma_{yz} = \frac{\tau_{yz}}{G} \]  
\(\text{(Eq 32)}\)

where \(E\) is the elastic modulus (often called Young’s modulus), \(v\) is Poisson’s ratio, and \(G\) is the shear modulus (sometimes called modulus of rigidity). Equations 29 to 32 are the 3D forms of the stress-strain equations for materials that are homogeneous and isotropic. The analysis of the stress-strain relationships shows that there are only two independent constants. Then \(E\), \(G\), and \(v\) are related by:

\[ E = \frac{G}{2(1+v)} \]  
\(\text{(Eq 33)}\)

The volume change in a material is given by:

\[ \Delta V = \left( \varepsilon_x + \varepsilon_y + \varepsilon_z \right) V = \left[ 1 - \frac{2v}{E} \left( \sigma_x + \sigma_y + \sigma_z \right) \right] V \]  
\(\text{(Eq 34)}\)

where \(V\) is the original volume and \(\Delta V\) is the volume change.

For typical materials the Poisson’s ratio, \(v\), varies from 0 to 0.5. For \(v = 0\), there is no lateral strain with axial strain. From Eq 34, it can be seen that the volume change gets smaller as \(v\) increases from zero to 0.5. Experimental evidence indicates that metallic materials when plastically strained have a Poisson ratio of 0.5 and therefore deform under constant-volume conditions. An equivalent statement is that they are incompressible. Alternatively, polymeric materials have higher compressibility and in many applications should not be considered constant volume materials. As previously indicated, constant volume corresponds to \(v = 0.5\). The most usual values of \(v\) for elastic metals are between 0.25 and 0.35. If exact data are not available, a value of \(v = 0.3\) is typically used.

Two-dimensional forms can also be given for special stress or strain cases. For example, only the stresses in the \(x-y\) plane may be nonzero for an in-plane stress. The corresponding stress-strain relationships become:

\[ \varepsilon_x = \frac{\sigma_x}{E} - \frac{v}{E} \sigma_y \]  
\(\text{(Eq 35)}\)

\[ \varepsilon_y = \frac{\sigma_y}{E} - \frac{v}{E} \sigma_x \]  
\(\text{(Eq 36)}\)

\[ \gamma_{xy} = \frac{\tau_{xy}}{G} \]  
\(\text{(Eq 37)}\)

These can be inverted to solve the stresses as a function of strain:

\[ \sigma_x = \frac{E}{1-v^2} \left( \varepsilon_x + v \varepsilon_y \right) \]  
\(\text{(Eq 38)}\)

\[ \sigma_y = \frac{E}{1-v^2} \left( \varepsilon_y + v \varepsilon_x \right) \]  
\(\text{(Eq 39)}\)

\[ \tau_{xy} = G \gamma_{xy} \]  
\(\text{(Eq 40)}\)

The 2D equations in Eq 35 to 40 are the representation of Hooke’s law for a homogeneous and isotropic material. A homogeneous material does not vary in properties from point to point; that is, for a translation of the coordinate axes there is no difference in material properties. Isotropic materials do not vary for a rotation of the axes. That is, the elastic properties do not vary for a change in orientation. No polycrystalline material satisfies these conditions exactly; however, the added benefit that comes from an analysis that uses elastic relationships for nonhomogeneous or anisotropic materials often does not justify the added difficulty introduced into the problem. For example, a completely anisotropic material can have 21 elastic constants. Even an orthotropic material can have nine elastic constants. Trying to work with these relationships presents a problem of such difficulty that it can usually only be handled numerically. However, it should be pointed out that anisotropy is taken into account (as for example with an anisotropic yield criterion) and controlled when large-strain plastic forming operations are carried out with metallic sheet.

Most metallic materials conform to these linear stress-strain relationships during the initial part of their loading. Most stress analyses are based on this linear-elastic relationship. The requirement for elastic behavior is that the body returns to its original shape on unloading. If metals are loaded beyond a point called the elastic limit, the body does not return to its original shape with unloading. Metals are said to yield and the relationship between stress and strain typically may
become nonlinear. (Note: the elastic limit and the yield point or yield stress are not equivalent. A material may be, and almost always is, nonlinearly elastic.) Above the yield stress, material behavior is nonlinear, and plastic strain results from increased stress. The volume of the material is assumed to be constant and Eq 34 shows that \( v \) approaches 0.5 for a constant-volume material. The relationship between stress and strain in the plasticity region often takes a nonlinear form, often parabolic in shape (i.e., \( y = ax^n \)), expressed as: \( \sigma = K\varepsilon^n \), where \( K \) is the constant strength coefficient and \( n \) is the constant strain hardening exponent. It is important to note that this equation is of fundamental importance in discussion of plastic forming behavior and that the stress, and strain are an instantaneous values, not nominal values. Finally, it is worthwhile to reiterate the previous discussion of constraints on deformation along the \( z \) axis. Equations 35 to 40 are 2D equations, but the deformation along the \( z \) axis depends on the constraint imposed by the thickness of the material in the \( z \)-direction. When the part is thick, a lateral constraint is imposed on deformation along the \( z \) axis. When plane-strain conditions occur (i.e., \( \varepsilon_z = 0 \)), a stress component develops along the \( z \) axis:

\[
\sigma_z = v(\sigma_x + \sigma_y) \quad \text{(plane strain)}
\]  

(Eq 41a)

In contrast, if the section is thin along the \( z \)-axis, then the lateral constraint is reduced and deformation occurs along all major axes. In this case, the condition is plane stress, where (as the name suggests), stress is zero in one direction:

\[
\sigma_z = 0 \quad \text{(plane stress)}
\]

(Eq 41b)

Other Components of Stress

Effect of Residual Stresses. Residual stresses are locked into the body and are not released when external forces are released. In that way they are always present. However, in the absence of external forces, the body must be in equilibrium and hence have a zero net force. This gives the residual stress pattern a character that is called self-equilibrating. That is, for a set of residual stresses in tension, there must also be a similar set of residual stress in compression so that the net force caused by them is zero. For this reason the effect of the residual stresses cannot be accounted for correctly by simply adding the maximum value of residual stress to the primary stresses, those that come from the applied forces. Just as applied forces create complex internal states of stress, locked-in residual stresses may also be complex, and 3D stresses are vector quantities. Residual stresses must be added vectorially to applied stress. The ways for handling residual stresses are varied but not always well understood. In the most elementary and conservative manner, it is sometimes assumed that the residual stresses are at a value of the yield strength and that they add algebraically (in a given direction) to the stresses from applied loads. In that way the combined applied and residual stresses are very severe, and the analysis of them is very conservative. Other ways for handling residual stresses are to measure the residual stress pattern and to add this measured pattern to the existing stress pattern that comes from applied forces. There is not always a good way to measure this residual stress pattern. Among the ways to measure residual stresses is hole drilling with strain gages, a destructive procedure. X-ray diffraction, a surface measurement, and neutron diffraction are not always readily available. Because residual stresses come from strain mismatch, they are easily disturbed by such things as plastic deformation or geometrical change such as the introduction of a cracklike defect. So even with a measured residual stress pattern, it is not clear what will happen to the residual stress pattern during the life of a component.

Contact (Hertzian) Stresses. Sometimes neglected by designers, contact stresses can initiate failure, especially in the presence of friction. Many finite-element analysis (FEA) practitioners assume that component failures occur at locations far removed from the loading and, for the sake of convenience, apply either point or line loads to their models. This generally yields acceptable remote stress fields, but stresses local to the load may be seriously in error if that load is applied to a single node or is improperly distributed over an assumed area. The theoretical area of contact between two nonconforming surfaces is frequently quite small. The interaction between these surfaces is often described as either “point” or “line” contact. Common examples of “point” contact are mating helical gears, cams and crowned followers, ball bearings and their races, and train wheels and rails. If the mating parts can be considered semi-infinite and if material behavior is linearly elastic, then the local stress state can be described by Hertzian theory as long as the contacting surfaces can be modeled as quadratic functions of those spatial coordinates defining the surfaces. Even in the absence of friction, the resulting 3D stress state is quite complex, and although the local stress state is compressive, that is, the principal stresses beneath the load are negative, large subsurface shear stresses, which can serve as crack initiation sites, exist beneath the load. A thorough analysis should be performed, especially if the contact stresses are the relevant stresses.

On the other hand, “line” contact, which occurs in cylindrical rollers and similar apparatuses, frequently can be considered to be a case of plane strain. Errors of FEA, such as load application over zero area, that is, applying a line load, may again yield acceptable far-field results, but local stresses will be overestimated. Uniform distribution of surface pressure will underestimate them. Again, if the contact stresses are responsible for failure, for example, pitting or spalling, proper load distribution is important to the failure analysis.
Consider radial loading of a cylindrical pin in a hole. First, look at the case in which significant clearance exists between the pin and the hole. In that event, a near-Hertzian stress state may exist beneath the pin, given sufficient material between the hole and an edge. However, if the hole is located near an edge and the loading is perpendicular (or nearly so) to that edge, then a large tensile stress exists at that edge beneath the hole. The magnitude of this tensile stress is strongly influenced by the stress state below the contact. Now examine the other extreme: little or no clearance between the pin and hole. In this event, Hertzian theory cannot be employed to develop the near-field subsurface stress state. Due to better load distribution, the edge tensile stress now will be significantly lower, as will the maximum subsurface shear stress. These two cases define the envelope in which pin/hole contact occurs. Differences between the two edge tensile stresses could approach 100%. Again this serves to highlight the need for proper stress analysis.

For additional information on complex stresses in components, see the article “Overview of Mechanical Properties and Testing for Design” in Mechanical Testing and Evaluation, Volume 8 of ASM Handbook.

References cited in this section


Stress Analysis and Fracture Mechanics

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

Stress Analysis Of Common Geometries

The stress analysis of common geometries can be made with either a mechanics of material approach or a theory of elasticity approach. The mechanics of materials approach follows four steps to arrive at a solution; these steps are taken in a prescribed sequence. The mechanics of materials always proceeds as follows:

1. Define the geometry of deformation.
2. Define strain from the geometry of deformation.
3. Define stress from this strain through the stress-strain relationships.
4. Relate the stresses to the boundary loadings through equilibrium.

Following these four steps always leads to a solution. In the procedure, the first step is the most difficult, defining the geometry of deformation. This has been done in the past for most common geometries by applied mechanics theoreticians who studied deformation models. In developing the mechanics of materials solutions here, the author uses these classical deformation patterns invents rather than new ones.

The theory of elasticity method requires that a number of conditions be satisfied simultaneously, namely, the body must satisfy internally a set of equilibrium equations and a set of compatibility equations. The equilibrium equations ensure that all parts of the body are in equilibrium under the stress field, and the compatibility equations ensure that the displacements are single-valued and possible. The compatibility equations derive from strain-displacement equations. The theory of elasticity always gives an exact set of equations for the solution to a problem. However, the equations are often a partial differential equation that cannot be easily solved and/or which must be solved numerically.

Therefore, the common elements that are presented in the following sections for stress analysis follow the mechanics of materials method for solution. Examples presented include the tensile geometry, torsion geometry, bending and beam shear, pin shear, thin-walled pressure vessels, curved beams, and springs.

Simple Shaft or Bar
A summary in Fig. 7 shows simplified, 2D free-body diagrams illustrating the orientations of the normal-stress and shear-stress systems at any internal point in a shaft loaded in pure tension, torsion, and compression. The single-overload fracture behavior of both ductile and brittle materials also is illustrated for each type of load.

**Fig. 7** Free-body diagrams showing orientation of normal stresses and shear stresses in a shaft and the single-overload fracture behavior of ductile and brittle materials. (a) Under simple tension. (b) Under torsion. (c) Under compression loading.

Tensile Bars. The tensile bar has an axis, \(x\), that goes through the centroid of a cross-sectional area. The area is often constant along the \(x\) axis but need not be. Tensile bars also are often round but need not be. Any shape of a tensile bar is permissible, but it must have a cross-sectional area with a centroid that lies on the load line. Assume that the load, \(P\), is directed along the axis, that is, through the centroid of the area. Off-center loads, those not along a centroid, result in more complex loading and would not be analyzed as simple axial loading.

Step 1 in the mechanics of materials assumes that the deformation mode of the bar is uniform elongation, that is, the strain is averaged over the entire length \(L\) of the body. Step 2 gives the strain as the stretch, \(\delta\), divided by the total length of the bar, \(L\). The only nonzero stress is \(\sigma_x\), which lies along the \(x\) axis so that step 3 gives the stress from the stress-strain equation as \(\sigma_x = E\varepsilon\). The equilibrium condition from step 4 gives the relationship between the stress, \(\sigma\), and the external loading, \(P\), as load equals the stress times area (A). These steps result in following equations:

\[
\sigma = \frac{P}{A} \quad \varepsilon = \frac{\delta}{L} \\
\delta = \int \frac{Pdx}{AE} = \frac{PL}{AE}
\]

(Eq 42)

The strain in the above equation should be indicated as the strain in the direction of load (e.g., \(\varepsilon_y\)).

The second result for the deflection, \(\delta\), is correct when the load, \(P\); area, \(A\); and modulus, \(E\) do not vary along the \(x\)-axis. For simple geometries like the tension bar, there is only one stress component. As discussed in preceding sections, there can be transverse strains even though no stress is applied in these directions.

Under tension loading, the tensile stresses (\(\sigma_1\) in Fig. 7) are longitudinal, whereas the compressive stress components (\(\sigma_3\) in Fig. 7) are transverse. The stresses created by tensile loading of cylindrical or approximately square specimens cause considerable deformation (elongation and necking) before fracture, which originates near the center of the shaft within a
region perpendicular to the load and propagates toward the surface, ending with a shear lip. However, in a brittle material, the fracture surface from a tensile overload is roughly perpendicular to the direction of tensile stress and involves little or no macroscale visible permanent deformation. The elastic stress distribution in pure tension loading, in the absence of a stress concentration, is uniform across the section. Thus, fracture in this case can originate at any point within the highly stressed volume. However, once the specimen necks, the stress distribution is no longer constant across the cross section, and fracture is predicted to initiate either along the centerline or at the surface, depending on specimen geometry and material properties (strain handling and strain-rate hardening). See the article “Mechanisms and Appearances of Ductile and Brittle Fracture in Metals.”

Torsion Bars. A bar is loaded with a torsional load if the loading is a couple, called torque, $T$, directed along the axis of the bar. The torsional loading of a round bar can be analyzed easily by the mechanics of materials method. It requires the same four steps previously listed in the same sequence required for all mechanics of materials solutions. The torsion of a round bar results in a state of stress described as pure shear. For the deformation mode, the assumption is made that the bar twists with no planar distortion and for the round bar, that the plane sections do not warp; that is, the plane sections remain plane. If the torsional bar is viewed as a roll of round coins, then the bar twists with one coin rotating slightly relative to the one next to it. As each coin rotates, the entire roll has a uniform shear strain that is linear in radius $r$. The coins do not distort but retain their original shape as they rotate about the axis. Therefore the strain is zero at the axis of the bar and reaches a maximum at the external surface Fig. 8.

![Fig. 8 Round bar in torsion](image)

Because the strain is linear with $r$ and the stress strain equation gives only a shear stress, the stress is also linear in $r$, (Eq 37). The evaluation of step 4 gives the relationship between the stress and the applied load, which here is a torque $T$. This results in a shear stress that is related to the torque by:

$$\tau = \frac{Tr}{J}$$

(Eq 43)

where $T$ is the applied torque and $J$ is the polar moment of inertia. $J = (\pi/2)r^4$ for a round bar. The round bar does not need to be constant in size along its length. For this stress the strain is:

$$\gamma = \frac{\tau}{G} = \frac{Tr}{JG}$$

(Eq 44)

The rate of twist is:

$$\frac{d\theta}{dx} = \frac{\gamma}{r} = \frac{T}{rJG}$$

(Eq 45)

so the total twist, $\theta$, can be found by:

$$\theta = \int_0^L \frac{Tdx}{JG} = \frac{TL}{JG}$$

(Eq 46)

where the second part of the equation is correct when all of the quantities are constant along the bar, $T$, $G$ and $J$ are not functions of $x$.

If the bar is not round, there is warping; that is, the bar displaces out of the plane of the cross-section in the $x$ direction. The simple mechanics-of-materials method used for the round bar does not work, and some more advanced method like theory of elasticity can be used. For a common shape like a bar with a rectangular cross section (Fig. 9), the membrane analogy gives a method from which the torsional stress can be calculated. Since this stress is not a simple function of the applied torque, the solution is often given in parametric format. For example with a torsion bar that has a rectangular cross section $a \times b$, the maximum stress occurs at the middle of the long side and is given as:

$$\tau = \beta \frac{T}{ab^2}$$

(Eq 47)

where $\beta$ is a parameter that depends on the ratio of $ab$. For $a = b$, a square cross section, $\beta = 4.81$. For $a = 10b$, a long thin rectangular cross section, $\beta = 3.21$. 
In a ductile material loaded to failure in torsion, shear stresses cause considerable deformation before fracture. However, this deformation may not be usually obvious, because the shape of the shaft has not been changed. The distortion will be obvious if there were axial grooves or lines on the shaft before twisting or if the metal is hot etched to reveal grain-flow twisting or if a mounted microstructural section is examined. If a shaft loaded in torsion is assumed to consist of an infinite number of infinitely thin disks that slip slightly with respect to each other under the torsional stress, visualization of deformation is simplified. Torsional single-overload fracture of a ductile material usually occurs on the transverse plane, perpendicular to the axis of the shaft. In pure torsion, the final-fracture region is at the center of the shaft. Slight bending will cause it to be off-center. Additionally, final off-axis fracture is expected if the crack does not grow radially in a uniform manner with respect to the circumference. Since the latter is relatively uncommon, off-axis final fracture locations are common.

A brittle material in pure torsion will again fracture perpendicular to the tensile-stress component, which is now 45° to the shaft axis. The resulting fracture surfaces usually have the shape of a spiral. The elastic stress distribution in pure torsion is maximum at the surface and zero at the center of the shaft. Thus, in pure torsion, fracture normally originates at the surface (in non-surface-treated components), which is the region of highest stress. When surface-treated components are loaded in torsion, the material strength profile may cause subsurface crack initiation.

Compression. When a cylindrical shaft is loaded in axial compression (Fig. 7c), the stress in the axial direction is compressive (σ in Fig. 7c) and tensile in the transverse direction (σ in Fig. 7c). The maximum shear stresses are 45° to the shaft axis, as they are during axial tension loading.

In a ductile material overloaded in compression, shear stresses cause considerable deformation and possibly the creation of some cracking but usually do not result in total fracture. The shaft is shortened and bulges laterally. A brittle material loaded in pure compression, if it does not buckle, again will fracture perpendicular to the maximum tensile-stress component. Because the tensile stress is transverse, the direction of brittle fracture is parallel to the shaft axis. If fracture occurs, it will likely be in the longitudinal direction, because compression loading increases the shaft diameter and stretches the metal at the circumference.

**Beam Bending**

Beam Analysis: Internal Forces. A member is usually treated with a beam analysis if the loads or applied couples are transverse to the beam length. The tensile bar has forces along its axis, and the torsion bar has a couple along its axis. However, a beam has forces and couples that are perpendicular to its axis. The tension and torsion bars are most often loaded with constant forces or ones that change in discrete steps. The beam has internal forces that usually vary as a function of position along the axis of the beam. An internal force balance indicates that the beam is subjected to a shear force, \( V \), an axial force and internal and moment. These generate two stress components: a normal stress, \( \sigma \), acting along the length from the moment (\( M \)), and a shear stress, \( \tau \), which acts on both longitudinal and transverse planes from the shear force.

Critical points in the beam correspond to the points where the shear force and bending moment are highest in magnitude, positive or negative. The sign convention for the internal loading in a beam is illustrated in Fig. 10. The positive configuration for \( V \) has it in the following directions, up on the left and down on the right. The positive configuration for \( M \) has it pointing up on both sides. The opposite directions correspond to negative values.
Fig. 10  Sign convention for shear \((V)\) and bending moment \((M)\)

The sign convention for the beam can be used to keep track of the stress directions; however, it is easier for the simple beam to use a physical sense to tell the direction of the stress. The beam can be loaded with a distributed load that can be a function of position, discrete forces that are transverse to the beam axis, and moments that are transverse to the beam axis. Typical beam loads are shown in Fig. 11.

For the beam, the horizontal and vertical axes \(x\) and \(y\) can be taken from the left end of the beam. The loads shown in Fig. 11 are \(w\), a distributed load that may be a function of position, \(w = w(x)\); \(P\), a discrete force; and \(M_o\), a discrete moment. The relationships between these applied loads and internal forces, \(V\) and \(M\) of Fig. 10, are determined from equilibrium as:

\[
\begin{align*}
    w &= -\frac{dV}{dx} \\
    V &= \frac{dM}{dx}
\end{align*}
\]  

(Eq 48)

or,

\[
\begin{align*}
    V_2 - V_1 &= -\int_{x_1}^{x_2} wx \, dx \\
    M_2 - M_1 &= \int_{x_1}^{x_2} V \, dx
\end{align*}
\]  

(Eq 49)

These relationships can be used to analytically determine the shear force and moment in a beam or to make a graphical determination of them using the shear and bending moment diagram approach. For the latter, the value of the shear force, \(V\), is plotted as a function of position on the beam (using the area under the \(w\) curve as given by Eq 49) to accumulate continuous change in \(V\). The applied load, \(P\), causes a sudden, discontinuous jump change in \(V\). The \(M\) diagram uses the area under the \(V\) diagram to give the continuous change in \(M\) and a discrete moment to cause a jump change in \(M\). The advantage of the diagram approach is that it is easy to determine the maximum magnitudes of the values of \(V\) and \(M\). These values are used in the stress equations to determine maximum values of \(\sigma\) and \(\tau\).

Beam Analysis: Stresses. The stress equations are developed from the mechanics of material methods. For this method to work, the loads must be applied in a plane of symmetry of the beam. This works for many geometries. For example, rectangular and circular cross sections as well as U-shaped and T-shaped beams have a plane of symmetry and can be analyzed by the simple mechanics of materials method. L-shaped or C-shaped beams do not fulfill this symmetry requirement and must be analyzed by other means.

The bending moment, \(M\), gives a normal stress \(\sigma_z\); the shear force, \(V\), gives a shear stress \(\tau_{yz} = \tau_{zy}\). If all applied loads are parallel to each other, the subscripts sometimes are dropped. (Note, however, that retention of subscripts minimizes errors in determining the net stress acting in a given direction as the combination of axial stress due to pressure and bending stress in a pressure vessel, or in a beam subjected to a torque.) The normal stress, \(\sigma\), is analyzed by developing a geometry of deformation that assumes that there is a neutral axis in the beam over which there is a reversal of the stress and strain signs. On one side of the neutral axis the stress is positive and tension exists; on the other side, stress is negative and compression exists. The sections that were originally plane remain plane and rotate about this neutral axis without distortion of the cross section. This geometry of deformation is used to define a strain given by:

\[
\varepsilon = -\frac{y}{\rho}
\]  

(Eq 50)

where \(\rho\) is a radius of the beam curvature. This makes the strain linear in \(y\), the distance from the neutral axis. Here the strain is negative for a positive value of \(y\) and positive for a negative value of \(y\). The stress is then also linear in \(y\). Writing the equilibrium equations to relate the internal stress to the boundary load gives the position of the neutral axis and the relationship between stress and applied loading. Taking \(\Sigma F = 0\) shows that the neutral axis coincides with the centroidal axis of the cross-sectional area of the beam; that is, the centroidal axis of the cross section that is perpendicular
to the beam axis, the one with a normal orientation in the x direction in Fig. 11. Taking $\sum M_x = M$, the moment that comes from the applied loading gives the relationship between the stress, $\sigma$, and $M$. It is:

$$\sigma = \frac{My}{I}$$  \hspace{1cm} (Eq 51)

where $M$ is the bending moment at the point, usually $M = M(x)$, $I$ is the moment of inertia, and $y$ is a position in the cross section taken in a perpendicular direction from the neutral axis. Here the stress is negative or in compression for a positive value of $M$ and $y$. Since $M$ is a function of $x$ and $\sigma$ a function of $y$, the stress, $\sigma$, varies from point to point in the beam. It is often desirable to find maximum values of this stress. To find the maximum value of stress, a maximum $M$ must be obtained as well as a maximum value of $y$. The maximum $M$ comes from the beam analysis, for example, from a moment diagram, and the maximum $y$ is at the outer edge of the beam, a distance farthest from the neutral axis. This is given the label $c$. Then:

$$\sigma_{\text{max}} = \frac{M_{\text{max}}c}{I}$$ \hspace{1cm} (Eq 52)

If the beam has two planes of symmetry, then the maximum stress occurs at $y = \pm c$, one sign giving maximum tension and the other maximum compression. This stress is sometimes called the outer fiber stress; it is the one used in determining design limits. For some beams, a quantity called section modulus, $Z$, is calculated where $Z = I/c$ so that at the outer edges:

$$\sigma = \frac{M}{Z}$$  \hspace{1cm} (Eq 53)

The beam shear stress is calculated from the shear force, $V$, using:

$$\tau = \frac{VQ}{lt}$$ \hspace{1cm} (Eq 54)

When $\tau$ is determined at some point $y$ from the neutral axis, then $Q$ is the first moment of the area from a free surface, usually the top or bottom of the beam to the point $y$ and taken about the neutral axis, $I$ is the moment of inertia of the entire cross section, and $t$ is the beam thickness at the point $y$. The thickness has to take all of the length that a cut through the section would encounter at $y$. For example, if the beam has two legs, as in a U beam, then $t$ is the combined thickness of both sides.

As an example of using Eq 53 and 54 to calculate $\sigma$ and $\tau$ in a beam that has a rectangular section $b \times h$, where $b$ is the thickness of the section and $h$ is the height, the stresses:

where the normal stresses, $\sigma$, are taken at the outer edges and the shear stresses, $\tau$, are at the neutral axis. (For beam stresses, $\tau$ are usually very small compared with $\sigma$.) The above $\tau$ is the maximum shear stress from the beam shear load and for a $x$-$y$ oriented element. However, for a failure criterion based on shear stress, the maximum $\tau$ does not occur in the $x$-$y$ orientation. At the top or bottom of the beam, the large $\sigma$ usually dominates the stress magnitudes in the beam, and this stress is in the $x$ direction. At $45^\circ$ to the $x$ direction, a plane of maximum shear stress occurs and the shear stress on that plane is $\sigma/2$. This can be found by applying a stress transformation as would be done in a Mohr's circle analysis like that illustrated in Fig. 5. For most beam geometries, $\sigma$ from bending is at least an order of magnitude greater than $\tau$ from beam shear. Thus the shear stress associated with bending stresses, that is, the stress found by rotating the stress element, is the one that is used for developing safety against failure by yield.

Curved Beams. The above stress analysis of a beam is for a beam that is essentially straight, that is, the axis of the beam, $x$, must be straight. Some elements are beamlike in loading but have a curved axis. Examples might be chain links, loading hooks, and machine frames. For the curved beam to be analyzed by the mechanics of materials method, the two curved surfaces of the beam must be concentric circles. The beam is then specified by the circle dimensions, an inner radius, $r_i$, and an outer radius, $r_o$. The curved beam does not have its neutral axis on the centroid as a straight beam does. Therefore, both the centroidal and neutral axes must be calculated. The centroidal axis is given the symbol $R$ and the neutral axis the symbol $r_n$. The main parameters of the curved beam are shown in Fig. 12.
To find $r_n$, the following expression can be evaluated:

$$r_n = \frac{A}{\int_A \frac{dA}{r}}$$

(Eq 56)

although for most common cross-sectional shapes the value of $r_n$ is tabulated. An important parameter is the difference between the centroidal radius and the neutral axis radius. This is called an eccentricity, $e$, where $e = R - r_n$. The stress, $\sigma$, for a curved beam loaded by end moments, $M$, is given by:

$$\sigma = \frac{My}{Aer}$$

(Eq 57)

The distance $y$ is taken from the neutral axis. If the moment is positive then $y$ is positive going toward the center of the circles. The stress becomes maximum at the inner radius, $r_i$. For example, at the inner radius $y = r_n - r_i$ and at the outside radius $y = r_n - r_o$. The latter is a negative number and reflects a negative bending stress at the outer radius. The stress in Eq 57 is for a pure moment, $M$, applied at the end of the curved beam. Often a force is applied. Fig. 13 shows the two possible cases. For an applied end force the stress must include the axial component of stress as well so that:

$$\sigma = \frac{My}{Aer} + \frac{F}{A}$$

(Eq 58)

where $F$ is the applied force. In the case of Fig. 13, the axial force adds to the tensile bending stress at the inside radius and subtracts from the magnitude of the bending compression stress at the outside radius. The moment of an applied force is taken relative to the centroidal axis of the beam and not the neutral axis.

Other Geometries

Thin-Walled Pressure Vessels. A pressure vessel with a nominal radius, $R$, that is at least 10 times the thickness, $t$, is considered to be a thin-walled pressure vessel. The vessel is subjected to only internal pressure, $p$. The internal pressure is the gage pressure and is not absolute. The effect of the wall thickness is to control the magnitude of the radial stress. In thin-walled vessels, this radial stress is assumed to be zero, and the hoop stress is assumed to be constant across the cross section. The shear stress $\sigma_{\theta r}$ is zero due to symmetry. Then, the only stresses present in the section due to an internal pressure are the hoop stress, $\sigma_{\theta\theta}$ and an axial stress, $\sigma_{zz}$. 

![Fig. 12 Definitions of distances for a curved beam](image)

![Fig. 13 Different ways of loading a curved beam](image)
The magnitudes of the axial and hoop stress are obtained by statics principles from equating the force of the internal pressure with the force of the stresses in the walls of the thin vessel. The pressure vessels can be cylindrical or spherical in shape. For the cylindrical vessel these stresses are:

\[ \sigma_\theta = \frac{pr}{t}, \quad \sigma_z = \frac{pr}{2t} \]  

(Eq 59)

and the axial stress depends on having a capped end for the pressure vessel. A thin-walled pipe can also be treated in the same way but would not have a capped end. In that case the axial stress is zero. If the vessel is spherical, the stresses are equal in every direction

\[ \sigma_{th} = \frac{pr}{2t} \]  

(Eq 60)

The largest normal stress for a thin-walled pressure vessel is the hoop stress on a cylindrical vessel. The axial stress is half of the hoop stress. All of the stresses in a pressure vessel are principal stresses. Mohr's circle can be sketched (or transformation equations applied) to show that the maximum shear stress occurs on a plane rotated 45° around the axial direction and has a magnitude:

\[ \tau_{\text{max}} = \frac{\sigma_\theta}{2} \]  

(Eq 61)

Therefore, at the macroscale, a fracture on a hoop plane is considered to be brittle fracture, and fracture on an inclined plane through the wall is considered to be ductile fracture. Macroscale brittle fracture (and therefore with no transverse thinning) is sometimes described as “thick-lipped” fracture. (An example is in the article “Fracture Appearance and Mechanisms of Deformation and Fracture).

Pin-Loaded Clevis. One geometry that is often analyzed is the pin-loaded clevis or pin-loaded plate. An example of a pin and clevis is shown in Fig. 14. This could be used for loading two bodies that must be coupled. The analysis can be applied to the pin as well as the clevis. For the pin, the load is a shear load.
The pin often is in double shear as it is in this case. That means that it is stressed on both sides of the center. Then a load $P$ is applied to the middle of the pin, and the internal shear force $V = P/2$ is resisted on either side of the central load. Given a pin diameter, $d$, and the pin shear stress, which is usually taken as the average stress given by:

$$\tau = \frac{V}{A} = \frac{P/2}{(\pi/4)d^2} = \frac{2P}{\pi d^2} \tag{Eq 62}$$

The pin could be analyzed as a beam to get a maximum shear stress at the neutral axis but that is not commonly done. The clevis or plate with a pinhole can also be analyzed. The stress from the pinhole loading is not large enough to cause a problem unless the pinhole is too close to an edge, in which case it could be critical. The pinhole resists a shear load over the area that goes from the middle of the pinhole to the plate edge along a distance $d$ (Fig. 15). That distance could go from the middle of the pinhole to the surface of the clevis as shown on the right; however, it is often advised to take $d$ from the top of the pinhole to the edge of the clevis as shown on the left, here labeled $d_{\text{min}}$. That provides a conservative calculation of the stress. If the $d_{\text{min}}$ dimension is about the same as the pin diameter, the design is usually okay. The stress then is:

$$\sigma = \frac{P}{2td_{\text{min}}} \tag{Eq 63}$$

where $t$ is the thickness of the plate. The stress concentration associated with the hole is not taken to influence the area defined by $d_{\text{min}}$. Here $P$ is the total load applied to the one side of the clevis, as shown in Fig. 15, and not the load applied to the pin, as shown in Fig. 14.

Spring Stresses. Spring stress analysis can be done for two types of springs, the linear helical spring and the helical torsion spring. The linear helical spring has a coil of wire and is wound in a round helical shape. It is loaded along the axis of the spring (Fig. 16). In this case the spring wire is round as well. In Fig. 16, $D$ is the diameter of the coil taken at midwire, $d$ is the wire diameter for a round wire spring, and $C$ is the spring index ($D/d$).
The only stress present in the linear spring is a shear stress, created primarily by the torsional load on the spring wire. Given force $F$ applied to the spring along the axis and in the center of the coil diameter, $D$, the stress is given as:

$$\tau = K_s \frac{8FD}{\pi d^3}$$  \hspace{1cm} (Eq 64)

where $K_s = 1 + (0.5/C)$.

A helical torsion spring (rattrap spring) has a helical coil but it is loaded in a direction normal to the spring axis. This type of loading causes normal stresses resulting from a bending moment. The load is usually applied on an extension wire so that the applied moment is $M = FL$, where $F$ is the force and $L$ is the distance from the force application point, on the extension, to the center of the spring. The analysis is like that of a curved beam in bending. The regular bending formula of Eq 55 is used to calculate the bending stress, but a factor $K_b$ is applied to the stress to account for the curvature. The resulting stress equation for the torsional spring is:

$$\sigma = K_b \frac{32M}{\pi d^3}$$  \hspace{1cm} (Eq 65)

where $K_b$ takes different values for the inside of the beam versus the outside. On the inside:

$$K_{bi} = \frac{4C^2 - C - 1}{4C(C - 1)}$$  \hspace{1cm} (Eq 66)

and on the outside

$$K_{bo} = \frac{4C^2 + C - 1}{4C(C + 1)}$$  \hspace{1cm} (Eq 67)

and again, $C = D/d$.

Stress Concentration. In many structural shapes there are holes, corners and other geometric changes that alter the stress distribution. The result of such disturbances is a localized stress concentration. These usually result in a stress, $\sigma_{\text{max}}$, which is greater than the nominal stress, $\sigma_{\text{nom}}$, where $\sigma_{\text{nom}}$ is calculated from the application of the stress analysis formulas. This maximum stress, $\sigma_{\text{max}}$, is determined from the nominal stress, $\sigma_{\text{nom}}$, by multiplying it with a stress concentration factor, $k_c$.

Therefore:
\[ \sigma_{\text{max}} = k_t \sigma_{\text{nom}} \quad \text{(Eq 68)} \]

Qualitatively, stress concentration factors increase as the difference in section at the discontinuity increases and as the fillet radius between the two sections decreases. It is also important to recognize that stress concentration is created not only by axial loading but also by bending and twisting loads and that the magnitude of the stress concentration factor depends on the type of loading.

The nominal stress from the stress formulas may be calculated based on either the net section or the reduced section and is indicated on each stress concentration factor chart; that is, for a plate with a hole, either the nominal area discounting the presence of the hole or the net section after the hole area is removed is used.

Standard references for stress concentration factors (Ref 2, 3) typically only provide values for the maximum stress created by the discontinuity, usually in a graphical format. Fig. 17 illustrates how a hole in a plate causes a stress concentration locally at the edge of the hole.

![Stress concentration at eyebolt](image)

**Fig. 17 Stress concentration at eyebolt.** \( k_t \), stress concentration factor; \( \sigma_a \), nominal (average) stress

A complete stress analysis often indicates that other stresses may be present and that the magnitude of those stresses depends on location and distance around the discontinuity. In some instances these stresses may be compressive. An example is the stress state associated with a circular hole in a wide plate subjected to an axial load. The stress concentration factor for such a geometry is taken as 3 (infinitely wide plate) or the actual correction for a finite width plate. However, it is important to recognize that the complete description of the stress state around the hole shows that a compressive stress is present. A stress concentration of 3 exists on a plane perpendicular to the applied load, but this factor changes to -1 on a plane parallel to the applied load at the top and bottom of the hole. The plate is not expected to fail at this location for uniaxial loading, and therefore the compressive stress is not considered in the design. However, if the plate is loaded biaxially, the situation changes. If the plate is loaded biaxially with balanced forces, the net stress concentration factor, obtained by superposition, is 2, not 3, and this stress concentration exists at four locations rather than two.

**References cited in this section**


**Stress Analysis and Fracture Mechanics**

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

**Application of Stress Analysis**
As noted at the beginning of this section, stress analysis provides a road map for the failure analyst. Stress analysis is an important and usually necessary part of the procedure to obtain a root cause for multiple reasons, including determining:

- Safe operating loads on a component in the original design
- Whether accurate and sufficiently precise stress analysis was done in the initial design in the event of failure
- From the macroscale examination of the fracture surface of a fractured component, whether the location of crack initiation is consistent with that expected based on proper fabrication/assembly and use of the component. If it is not, there are suggestions that the part may have been used for something other than its intended purpose or, alternatively, a metallurgical or geometric imperfection may be present in the material that caused the crack initiation site to be moved from the predicted location based on the nominal stress analysis.

Stress Analysis in Failure

Failure Criteria. Using a stress analyses for failure analysis requires a definition of the material behavior that constitutes failure. There can be many definitions for failure; yielding, fracture and buckling to name a few. A conventional design usually picks a yield criterion or a fracture criterion based on stress. For the former, the conventional yield criteria for metallic materials are based on the maximum shear stress approach (Tresca) or the maximum distortion energy approach (von Mises). For the former, Tresca, the maximum shear stress as given by Eq 23 is compared with the maximum shear stress in a tensile test. The failure stress for ductile materials is usually taken as the 0.2% offset yield strength, \( \sigma_{0.2} \), but in some cases (including brittle materials) it could be ultimate tensile strength, \( \sigma_{ut} \). For design, a safety factor is usually chosen to determine a design stress from the failure stress. As an example, the yield strength with a safety factor of 2 or an ultimate tensile strength with a safety factor of 3 could be chosen. The maximum shear stress in a tensile test is \( \sigma/2 \) at 45° to the loading direction. Therefore having the maximum shear stress in a component equal half of the yield strength would be an example of a failure criterion based on the Tresca criterion.

The maximum distortion theory (von Mises) uses the effective stress of Eq 28. For a tensile test \( \sigma_0 \) equals the yield strength. Therefore a general failure criterion, \( \sigma_{f} = \sigma_{y} \), is usually the von Mises failure criterion. The Tresca and von Mises criteria do not give the same result except for a few special loading conditions, pure tension being one of them. Nevertheless, they are always within 15% of giving the same result and the Tresca criterion is more conservative. Both the isotropic von Mises and Tresca criteria are used for design purposes for ductile materials. (If the material is highly anisotropic, the common criterion is due to Hill, not discussed here). For the more brittle materials, the simplest criterion is a maximum normal stress and is often used but often does not accurately predict failure when uniaxial loading is compressive or for multi-axial loading when principal stresses are of opposite sign. This would be based on the principal stress, \( \sigma_{p} \), reaching a limit, for example, a value of failure stress. For some brittle metallic materials and for most polymeric materials, a better representation of experimental data is obtained by using a criterion that includes both normal and shear stresses (that is, the normal stress on the plane of maximum shear stress affects the permissible shear stress for no yield). For materials that behave in a brittle manner at the macroscale (due to constraint created by geometry or for an inherently brittle material), it is good to consider the effect of defects in the material. In that case, the fracture mechanics approach presented in the next section is usually a better way to access failure potential than a maximum principal stress criterion. For the preceding criteria it is assumed that the material behavior is even, that is, the failure occurs at the same load magnitude in tension loading as it does in compression loading. This assumption is acceptable for many metals. For some other materials it is not a good criterion. (A discussion of failure in uneven materials is given in many textbooks as the Coulomb-Mohr criterion. It is not discussed here; however, this criterion is important in reference to polymeric materials.)

Fracture Orientation. The orientation of the fracture surface also depends on the type of behavior. For a material that behaves in a macroscale brittle manner, the failure is perpendicular to the maximum normal direction. A macroscale ductile fracture is assumed to occur on a plane of maximum shear stress. The maximum shear stress often occurs on a 45° plane between that of the maximum principal stress and the minimum principal stress. Since the minimum principal stress is often the zero stress of a free surface, the fracture plane is often oriented at 45° to a free surface. (However, for plane strain deformation of plate geometries, this angle is changed to approximately 55° for uniaxial loading of isotropic material.) Mixed-mode fractures result when constraint is incomplete so that fracture occurs on both planes of maximum normal stress and maximum shear stress. This causes shear lips in plate-like structures and the cup and cone fracture of a tensile specimen. Any fracture that causes a 45° ridge to a free surface is usually caused by a ductile fracture mode. On the other hand, a fracture surface that is flat is not always an indication of a brittle fracture. A microscale ductile fracture may occur with relatively flat fracture. That is, the macroscale appearance is brittle (with little transverse contraction) but the fracture mechanism (microscale) is ductile (i.e., microvoid coalescence).

Defects and Imperfections. An imperfection in a body may cause failure at a relatively low nominal stress, and if this stress is below the design stress, the imperfection can be considered a defect. Failures predicted by some of the preceding criteria even when stress concentration is taken into account, are often nonconservative. The fracture mechanics approach was developed to account for the effect of defects on failure for a body containing crack-like imperfections. This approach
is discussed in the next section and not described here. However, the question of whether an imperfection in the body or material constitutes a cracklike defect that should be analyzed by the fracture mechanics approach is one that needs some discussion. This is extremely important because a traditional approach to increase safety of a design is to increase the section to decrease the nominal stress. However, if the material contains a cracklike imperfection and the section thickness is increased, the design becomes less safe, not more safe.

It is not uncommon that many of the nondestructive methods used to find defects are unable to precisely distinguish the type of defect and how damaging it would be. Assuming the imperfection is cracklike and applying the fracture mechanics approach is the most conservative way to do a safety analysis, but it is necessary if cracklike imperfections are present.

In any stress analysis used in design, there are two inputs: material properties and loading conditions. Failure occurs when the loading conditions cause material strength, ductility, or toughness properties to be exceeded. The question then is how to handle the presence of imperfections. Two types of imperfections can exist in materials: geometric imperfections created by manufacturing operations or abuse, and material imperfections. Geometric imperfections include nicks and gouges, and in some cases, surface roughness created by machining. Material imperfections include deleterious microstructures (inherent brittleness due to mechanical/thermal processing errors, too large a grain size, precipitate-free zones, micro-and macroscale segregation, too high an inclusion content, shrinkage porosity in castings, etc.). Fracture mechanics analysis is a macroscale analysis and therefore properly used to analyze the effect of macroscale imperfections and thus primarily geometric imperfections. One example where a material imperfection might be used in a fracture mechanics analysis would be a macro-scale region of porosity (although this is not standard procedure in a fracture mechanics based analysis). Alternatively, material (microscale) imperfections cause a change in apparent mechanical properties and therefore are not used in determining the effect of loading conditions (and therefore as stress intensity input in a fracture mechanics analysis) but rather as input of the strength, toughness, or ductility of the material.

Critical Locations. Finding critical locations for possible failure and applying the analysis to these areas is an important consideration in the application of the stress-based failure analysis. To apply the ductile failure criteria, the regions with the highest stresses must be found. These are not always the largest tensile stress as in the first principal stress but the stresses that are used in the ductile failure criteria, such as the largest shear stress or the largest value of effective stress. With an analysis that involves the mechanics of materials presented previously, the critical locations can be determined by considering the nature and stress distribution that are found from each loading:

- Tensile loading is assumed to create a uniform distribution of stress.
- Torsion-bar loading has the highest stress in outer regions.
- Bending beam has the highest stress where the moment is highest and at a maximum distance from the neutral axis.

These are important and should be part of finding the most critical stress region. If failure is determined not to initiate at these locations, the implication is that a material imperfection (geometric or material) has moved the initiation site. In addition, the highest stresses that result for the mechanics of materials formulas may not be the highest stresses in the body. These are only stresses in the major directions of the body. A stress transformation to find principal stresses and maximum shear stresses should be done. For ductile materials, shear stress or difference between stress components may be more important than the absolute maximum normal stress value.

An example of finding critical locations is shown in Fig. 18. A round beam has an applied torque, axial load, and transverse load. The stresses that would result from these loads are axial normal stress, bending normal stress, torsional shear stress, and beam shear stress. The location of maximum normal stress and maximum shear stress is then found by considering the location where the net stresses are maximum. The axial stress is independent of location, and the stress tensor associated with this stress is given by:

\[
\sigma_{\text{axial}} = \begin{bmatrix}
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & \tau_{\text{ax,ax}}
\end{bmatrix}
\]
The bending stresses vary with position and are given by:

\[
\sigma_{\text{bend}} = \begin{bmatrix}
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & \sigma_{z,z,b}
\end{bmatrix}
\]

\[
\tau_{\text{bend}} = \begin{bmatrix}
0 & 0 & 0 \\
0 & 0 & \tau_{b} \\
0 & \tau_{b} & 0
\end{bmatrix}
\]

The torque applied to the section does not vary with position and is given by:

\[
\tau_{\text{torque}} = \begin{bmatrix}
0 & 0 & 0 \\
0 & 0 & \tau_{\text{torque}} \\
0 & \tau_{\text{torque}} & 0
\end{bmatrix}
\]

Therefore, examination of these stresses indicates that:

- The axial stress will add to the variable bending stress.
- The shear stress created by the torque will add to the variable shear stress created by the bending.

Clearly, then, the maximum net axial stress will occur on the bottom of the beam where the beam bending stress is tensile and will have a magnitude of the sum of these two stresses. Also, the maximum shear stress will occur along the neutral axis where the maximum shear stress from bending exists and will have a magnitude of the sum of the shear stress from the torque plus that from the bending. (In most cases, the shear stress from the bending is small in relation to that from the torque and is often neglected in an analysis. For solid prismatic sections, it is \(\frac{4}{3} \frac{V}{A}\) and for solid circular sections it is \(\frac{3}{2} \frac{V}{A}\), where \(V\) is the nominal shear force in the beam. However, for hollow tubes, the maximum shear at the neutral axis increases to \(2 \frac{V}{A}\), and therefore should not be so quickly neglected.

Examination of the aforementioned stress tensors then indicate that the critical locations to examine are (a) where the net shear stress is a maximum and (b) where the net axial stress is at a maximum. Location (a) is along the neutral axis and (b) is at the location of the maximum bending moment. Ductile fracture for (a) on the plane of maximum shear stress will occur on either a longitudinal or transverse plane, while brittle fracture at this location will occur on a plane at 45° to the maximum shear stress. Brittle fracture due to the net axial stress will occur on a plane normal to the axis, and ductile fracture will occur on a plane at 45° to the axis.

If the lateral load were not at the midlength or if several sets of lateral loading were involved, a shear and moment diagram construction could identify the critical locations for maxima. Then again, the top and bottom of the beam should be analyzed for the flexural stress and the midplane for shear. Other components can be analyzed in a similar way to find critical points. When the geometry or loading pattern becomes so complex that a simple application of the mechanics of
materials formulas cannot be made or does not reveal the critical locations, then a numerical analysis can be made and critical stress locations found.

Limit Analysis. An alternate method for stress analysis is a limit analysis. Here the approach is to analyze a structure with the assumption that is has reached a plastic limit. The calculations for this can often be much easier than than the elastic stress analysis described in the section “Stress Analysis of Common Geometries” in this article.

The limit analysis is done with two limit theorems, an upper bound theorem and a lower bound theorem. The upper bound theorem chooses a plastic collapse mechanism and balances external work on the structure against the internal energy dissipated by the collapse mechanism. Different collapse mechanisms can be assumed, and the one giving the lowest collapse load would be closest to being correct. The lower bound theorem uses an equilibrium analysis while requiring that no stress in the structure is above a limit stress like yield strength. The lower bound analysis can use different stress distributions, especially regarding the points that are at a limit stress. The analysis that gives the highest limit load would be the closest to being correct. If the lower and upper bound approaches give the same result, an exact limit solution is reached. If they differ, they bracket the exact solution, which may not be exactly determined.

The basic assumption on most limit analyses is that the body is undergoing ductile failure. In those cases the residual stresses do not have much influence. Under a large strain field as that imposed in a limit analysis, the strain mismatches that caused residual stresses tend to be washed out and the residual stresses disappear. Therefore in limit analysis, secondary stresses like limit residual stresses are often ignored.

Strain rate effects in safety analysis can have a confusing set of effects. For a ductile material, a fast loading rate would raise the strength of the material, possibly resulting in a higher failure load. However, if a brittle mechanism of failure may occur, a faster loading rate may lower the failure load (e.g., a higher DBTT with a higher strain rate). This is especially true for a structure with a notch or cracklike defect. In addition, the faster loading rate may change the mechanism of failure from a ductile one to a brittle one, and a failure prediction based on ductile behavior may not be correct. Therefore, it is important to consider the effect of a fast loading rate on the failure of a structure on a case by case basis and not use general principles to try to predict the effect of the loading rate. The evaluation requires knowledge of the failure mechanisms, for example, fracture or yield, material property data under rapid loading, and the loading rate (see ASM Handbook, Volume 8 for the regimes of strain rates). A fast loading rate may alter the stress analysis by introducing inertial effects that influence the stress distributions. A loading to failure that is in the millisecond range usually does not result in inertial effects but rates an order of magnitude faster than that which could introduce inertial effects.

Examples

The failure analyst also should realize that many low- and medium- technology products are designed and manufactured without ever having an engineer perform any sort of stress analysis. This obviously makes the failure analysis more challenging. Fortunately, there is usually someone who is able to help determine where the maximum stress should have been. In some cases, this is quite helpful and all that is necessary for the scope of the investigation (as a properly performed stress analysis may be out of the range of the time and budgetary constraints for the failure analysis project). In addition, if analysts gain an understanding of the basic types of loading (tension, compression, bending, and torsion), and study the expected crack appearances for each of these four loading types in cylindrical and prismatic geometries, they will be able to characterize the stresses of fracture regardless of the presence or absence of a correct or incorrect stress analysis.

It also is not always necessary for the failure analyst to personally become involved in a complex stress analysis. Those services may require extensive training, and the failure analyst should be careful not to perform work outside of his or her area of expertise. However, the failure analyst must be able to appreciate the implications of the analysis. It is more important for the failure analyst to be able to examine a failed part and to know whether the failure location and, in the case of fracture, the fracture surface orientation and initiation site are consistent with those expected in proper use of the component. Such stress analysis is qualitative or semiquantitative and may be key to identifying cause for failure. Two examples are cited, one simpler than the other, to illustrate this point.

Example 1: Stresses in a Screwdriver Used as a Pry-Bar. Normal use of the screwdriver imposes a torsional state of stress on the shank of the screwdriver. If the screwdriver is used as a pry-bar, the state of stress is that created by bending. Identification of root cause of a broken screwdriver clearly resides in recognizing indications on the fracture surface that tell whether the screwdriver failed during expected usage (torsional stresses), or while being used for other than its intended purpose (bending). The identification can be made because of difference in orientation of the fracture surface relative to the shaft. The screwdriver could still fail when serving its intended and proper use, but the implications of the failure are totally different in that case.

Example 2: Stresses in a Pressurized Cylindrical Section. As a second example, consider an internally pressurized cylindrical section (see the section “Thin-Walled Pressure Vessels” in this article for a more complete discussion of the stress state in such a section). The stresses present in this (assumed to be thin-walled) section are a tensile stress acting along the axis of the cylinder and a tensile circumferential stress. The magnitude of the circumferential stress is twice as
large as the axial stress. Therefore, failure, if it occurs, is assumed to be caused by the circumferential stress. However, the vessel must be supported in some way, so it is likely that bending stresses are also present in the vessel. The longitudinal stress created by the bending adds to the longitudinal stress created by the pressure and must exceed the longitudinal stress created by the pressure before it can change the plane on which the maximum stress acts. The latter could occur for example by the presence of a nick or gouge in the material in a location of high bending stress. There are then four possible ways in which fracture could occur: ductile or brittle fracture due to the internal pressure; or ductile or brittle fracture due to the internal pressure plus the bending stress. Identification of root cause clearly rests in part on being able to examine the fracture surface to determine whether bending loads were involved in the failure and whether the fracture was macroscale ductile or brittle.

**Stress Analysis and Fracture Mechanics**

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

---

**Fracture Mechanics**

In general, there are two types of conditions that may lead to structural failure:

- Net-section instability where the overall structural cross section can no longer support the applied load
- The critical flaw size \(a_c\) is exceeded by some preexisting discontinuity or when subcritical cracking mechanisms (e.g., fatigue, stress-corrosion cracking, or creep) cause the crack to reach critical size.

Failures due to net-section instability typically occur when a damage process such as corrosion or wear reduces the thickness of a structural section. This type of failure can be evaluated by traditional stress analysis or finite element analyses. These methods can also help identify the likely areas of stress concentration where cracking may have initiated. However, stress analyses by traditional methods do not easily account for crack propagation from preexisting cracks or sharp discontinuities in the material. Basic elasticity calculations show that both stress and strain become astronomical at a discontinuity such as a crack, far exceeding any recognized property levels that might offer some sort of limitation. Two situations are then possible:

- The crack reaches a critical length so that instability occurs and cataclysmic fracture occurs. The fracture is macroscale brittle (microscale ductile or brittle).
- The crack blunts, redistributing the stress state, with continued loading creating a tear zone (and sharpened crack-tip radius) in front of the crack. In steels, this tear zone can then cause the critical crack length to be exceeded such that unstable cleavage fracture occurs or unstable microscale ductile fracture is induced. Which event occurs depends on the temperature and the loading rate, but in either event, crack propagation is unstable (i.e., does not require an increasing load after creation of the tear zone). See the discussion below and in the article “Mechanisms and Appearances of Ductile and Brittle Fracture in Metals.”

**Stress Analysis and Fracture Mechanics**

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

---

**Fracture Mechanics Concepts**

The analysis of macroscale brittle fracture induced because of the presence of cracklike imperfections in a material is the basis of fracture mechanics. In a simplistic overview, fracture mechanics says that there is a triangle of variables that must be considered in the analysis of fracture. This is in contrast to traditional design in which there are only two components of the design criteria—material properties and applied loading conditions. For linear elastic fracture mechanics (LEFM) analysis, the three parameters involved are the stress created by the applied load \(\sigma\), the stress
intensity \((K)\), and the crack and component geometry. The \(K\) created by the applied loads is then related to the stress and crack length via:

\[
K = \sigma (\sqrt{\pi a}) \left[ f \left( \frac{a}{W} \right) \right]
\]

where \(a\) is the crack length and \(f(a/W)\) is a geometry correction factor tabulated in handbooks, which applies to a given component and crack geometry. Common examples include an edge-crack in a plate, a center crack, a double edge crack, and so forth. Fracture is assumed to occur when the \(K_{\text{applied}}\) created by the applied loads exceeds a material property \(K_c\). It is important to recognize that the geometry correction factor is not a stress concentration factor.

Fracture can be macroscale brittle, macroscale ductile, or macroscale mixed ductile and brittle. The attraction of fracture mechanics for fracture analysis is that so long as the volume of material described by the stress intensity factor is small with respect to the dimensions of the component and the size of the plastic zone at the crack tip is small with respect to the \(K\)-zone, the description of the stress state in the vicinity of the defect can be assumed to be elastic (LEFM analysis) and the resulting stress state and stress intensity factor is the same, no matter what the component geometry. Hence, the “one parameter” \((K)\) description of the fracture process of fracture exists. Results from simpler laboratory test specimens can be transferred one-to-one to a full scale engineering component of different geometry. As shown below, there is a plastic zone at the tip of the crack. If the size of this zone is large with respect to the \(K\)-field, linear elastic analysis may not be appropriate. In this case, the more recently developed field of elastic-plastic fracture mechanics (EPFM) must be utilized in which some other material property (such as the J-integral) replaces the stress intensity factor to describe the “toughness” of the material. Analysis by LEFM is easier to apply (fracture control is described by a one-parameter analysis, the critical value of \(K\)). Additionally, from a practical standpoint, more material data is available in the literature for LEFM fracture than for EPFM fracture. If EPFM analysis is used, a two-parameter approach is necessary. The following section focuses on LEFM analysis, with some mention of EPFM analysis. In conclusion, it is important to recognize that in the application of fracture mechanics analysis to a failure analysis, the material is treated as a black box. Analysis is based only on macroscale parameters of nominal stress, crack length, and component geometry. (This should not be interpreted to imply that microstructural parameters do not affect the value of \(K\) at fracture. Clearly, fracture toughness depends strongly on the specific microstructure present in the material and also on environmental conditions.)

**Stress Analysis and Fracture Mechanics**

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

### An Introduction to LEFM

Assume the presence of a cracklike imperfection in a material as shown in Fig. 19. Irwin showed that the stress state around the imperfection could be described two dimensionally, where:

- On the plane of the crack, the only existent stress is \(\sigma_{yy}\).
- There is a singularity in the stress field as \(r\) approaches zero; that is, the stresses approach infinity.
On the plane of the crack ($\theta = 0$ in Fig. 19), the only existent stress is $\sigma_{yy}$:

$$\sigma_{yy} = \frac{K}{\sqrt{2\pi r}} \left[ f'(\theta) \right]$$

where $a$ = crack length and $W$ is the width of the plate containing the cracklike imperfection.

If the material is not inherently brittle, the crack tip stress $\sigma_{yy}$ reaches the yield stress (YS) at some distance in front of the crack tip, causing a plastic zone to be created. There are various ways to estimate the size and shape of the plastic zone (i.e., based on use of the yield equations in the previous section). However, the simplest approach is to assume that yield occurs at a distance in front of the crack tip:

$$r = \left( \frac{1}{2\pi} \right) \left( \frac{K_c}{YS} \right)^2$$

The Irwin analysis was 2D. If the body is sufficiently thick (see subsequent paragraphs), the state of stress becomes 3D by the addition of a transverse stress, $\sigma_{zz}$, that acts in the thickness direction of the body and that has a magnitude given by:

$$\sigma_{zz} = \sigma_{xx} + \sigma_{yy}$$

Stated differently, the state of stress at the crack tip changes from plane stress to plane strain, with the transverse stress $\sigma_{zz}$ preventing flow in the transverse direction. Then, application of yield equations shows that the plastic zone is reduced in size, being approximately one third as large as for the plane stress case. Irwin also observed that the critical value of $K$ at fracture then becomes independent of section thickness and that the lateral contraction (due to $\sigma_{yy}$ at the crack tip) reaches a minimum value (i.e., the strain is 2D, plane strain). Since the value of $K$ at fracture no longer depends on section thickness, it is considered to be a material property and described not as $K_c$, but as the plane-strain fracture toughness $K_{IC}$.

Determination of material property toughness is covered by several ASTM specifications, E-399 specifying the determination of $K_{IC}$.

For quantitative evaluation in either design or in failure analysis, the material toughness at fracture must be known. Tabulated values in the literature are for plane strain conditions. If the section thickness to be used in the design or that of a failed component does not create sufficient constraint to develop plane strain conditions, the fracture toughness at fracture is unknown but higher than the tabulated plane strain value. For purposes of design, the required section thickness to obtain plane strain conditions can be estimated (for a through-thickness crack loaded in opening mode) according to:
where $B =$ section thickness and $S =$ design stress, typically the yield stress decreased by a factor of safety. This criterion is conservative in that plane strain values will be obtained at this section thickness but may also be obtained for smaller section thicknesses. For application to failure analysis of a fractured component, examination of the fracture surface will indicate whether all of the fracture surface is flat and normal to the applied loads or whether the fracture is mixed mode and contains shear lips. If shear lips are present, constraint was incomplete so that the value of $K$ at the onset of fracture is greater than $K_{lc}$.

This equation can also be used for purposes of inspection and determination of suitability for service. Assuming that $K_{lc}$ and the yield strength of the material is known, the equation can also be used where the crack length, $a$, replaces the section thickness $B$, allowing the estimation of critical crack lengths. Such estimates predict critical crack lengths at room temperature of a few thousands of an inch for a temper-embrittled steel to several inches for hot rolled and or properly quenched and tempered steels. Similar examination of the toughness of high-strength aluminum alloys indicate that unless these materials are specially tempered to maximize fracture toughness (e.g., the T73 temper), critical section thicknesses may be less than $\frac{1}{4}$ of an inch.

The general trend in all cases (of properly processed material is that the fracture toughness tends to decrease as the strength level is increased for materials typically used as structural or machine components. In steels, the fracture toughness does increase initially with carbon content for low (lower than that of the structural grades) carbon contents. However, the specific variation is quite sensitive to the microstructure present. Additionally, fracture toughness can be highly anisotropic in fabricated components. In rolled plate, fracture toughness is highest when cracks propagate in the long transverse direction and lowest when cracks propagate parallel to the rolling direction. Similar results are obtained for other deformation patterns with low fracture toughness associated with the largest principal fabrication strain in the body. This anisotropy is due to both banding (microscale solute segregation, which may result in microstructure constituent banding as for pearlite distribution in hot rolled steel) and fibering created by inclusion shape, distribution, and morphology. Some success in reducing fracture toughness anisotropy due to fibering and raising the overall toughness level has been achieved in the steel industry via ladle calcium treatments.

The analysis therefore must be modified to include a small zone of plasticity at the crack tip. The modified stress distribution is shown schematically in Fig. 20. The small plastic enclave is called a plastic zone, and its extent directly ahead of the crack is denoted by $R_p$.

**Fig. 20 Stress distribution ahead of crack in which small-scale plasticity is included**

The stress field at the crack tip can be characterized in terms of the stress intensity ($K$), where the plastic zone size depends on both $K$ and $\sigma_y$:

$$R_p = C \left( \frac{K}{\sigma_y} \right)^2$$
The constant, C, depends on various factors including thickness and deformation characteristics of the material. It usually has a value between $1/\pi$ and $1/3\pi$. This equation is only valid for linear elastic bodies; that is, when $R_p$ is much smaller than the crack size (i.e., $R_p < 10\%$ of the crack size) and other physical dimensions of the body (Ref 5). As $K$ increases (e.g., by increasing the load) $R_p$ also increases, and the material in the plastic enclave becomes more severely strained. This plastic enclave is very important in metals and polymers since it is the region in which energy is absorbed and is responsible for the relatively high fracture toughness of metals compared to all other materials.

As a result of the direct relationship between $K$ and the strains at the crack tip, a critical value of $K$ can be defined for when fracture takes place. When the region of plastic deformation around a crack is small compared to the size of the crack (as is often true for large structures and high-strength materials), the magnitude of the stress field (in tension) around the crack is thus related to the stress-intensity factor, $K$, as follows:

$$K = Y\sigma\sqrt{\pi a}$$

where $Y$ is a geometric factor, $\sigma$ is the gross stress across the fracture plane, and $a$ is the crack length. In this way, linear elastic analysis of small-scale yielding can be used to define a unique factor, $K$, that is proportional to the local crack-tip stress field outside the small crack-tip plastic zone. The stress-intensity factor can have a simple relation to applied stress and crack length, or the relation can involve complex geometry factors for complex loading, various configurations of real structural components, or variations in crack shapes. For example, the geometric factor, $Y$, is typically on the order of 1 for a tension-loaded plate with a through-thickness crack (Fig. 21). Geometric factors (expressed as a flaw shape parameter, $Q$) for surface cracks and embedded cracks are also shown in Fig. 21 (Ref 6). Specific formulas for $K$ depend on the load/crack geometry, and numerous solutions for $K$ can be found in Ref 7 and in the article “Summary of Stress-Intensity Factors” in Fatigue and Fracture, Volume 19 of ASM Handbook (Ref 8).
These concepts provide a basis for defining a critical stress-intensity factor ($K_c$) defined as:

$$K_c = Y\sigma_f\sqrt{a_c}$$

where $\sigma_f$ is the fracture stress occurring with a critical crack size, $a_c$. The critical stress intensity, also known as fracture toughness ($K_c$), is the value of $K$ that results in rapid, unstable fracture. The value of $K_c$ depends not only on the material being considered but also on the thickness of the material in which the crack is found. Plastic deformation is the most important source of toughness for most structural materials, and anything that impedes plastic deformation will reduce the toughness. As previously noted, toughness is manifested primarily through plastic deformation at the crack tip; anything that impedes crack-tip plastic deformation will reduce the toughness. The plastic deformation might be visible in a test specimen as a non-linear displacement curve with rising load (although this could correspond to crack opening due to stable ductile tearing.) However, the deformation may not be visible on the specimen surface, but it can be positively identified fractographically on a microscale.

The macroscopic deformation of the material also influences toughness. If the material is ductile enough or thin, it tends to contract (neck) parallel to the crack plane. This is indicative of plane strain conditions (Fig. 22b); that is, the stress along the crack plane is zero (Eq 41b). However, if the section size is sufficiently thick (or if the material is more brittle), the lateral contraction is opposed by the bulk of the thick material, and an additional stress component develops parallel to the crack plane (i.e., plane strain conditions, Eq 41a). Under plane-strain conditions, plastic deformation at the crack tip is severely limited. Lateral contraction is constrained along the crack plane, and the fracture surface is very flat with little or no macroscopic deformation perpendicular to the crack plane (Fig. 22a).

The value of $K_c$ thus depends not only on the ductility of the material but also on the state of stress. If the component is thin or sufficiently ductile, there is contraction along the crack plane, and the work of lateral deformation contributes to toughness. As the thickness of the section increases, the critical stress intensity is reduced until it reaches a minimum value for plane-strain fracture toughness ($K_{IC}$) (Fig. 23). In the plane-strain state, a material is at its lowest point of resistance to unstable fracture, and the mode of fracture is brittle without any appreciable macroscopic plastic deformation. The plane-strain fracture toughness ($K_{IC}$) is a minimum value and an inherent material property. Valid linear-elastics conditions for determination of $K_{IC}$ require a thickness $\geq 2.5(K_{IC}/\sigma_y)^2$.
In general, when the specimen thickness and the in-plane dimensions near the crack are large enough relative to the size of the plastic zone, then the value of $K$ at which growth begins is a constant and the plane-strain fracture toughness can be determined. If not, elastic-plastic fracture mechanics (EPFM) is required.

Linear elastic fracture mechanics (LEFM) is a useful tool in failure analysis, because many structural failures occur by subcritical crack growth mechanism (e.g., fatigue, stress corrosion, and creep) until a critical crack ($a_c$) is reached. In this regard, fracture mechanics is an effective tool for evaluating critical flaw size ($a_c$) that leads to rapid unstable fracture and can help answer questions during a failure analysis such as (Ref 6):

- Where should one look for the transition from subcritical crack growth to unstable rapid fracture?
- What was the load on the component at the time of failure?
- Was the correct material used and manufacturing/processing sound?
- Was the part designed properly?
- Did the environment influence the failure?

Linear elastic fracture mechanics can be applied to interpretation of subcritical crack growth, as briefly described in the next section. Care must be taken in applying LEFM to the onset of unstable (rapid) fracture of ductile materials, because the plastic component of fracture resistance may be significant. When plastic deformation is more excessive, the errors due to the assumption of elastic behavior become too large to be acceptable or realistic. In this case, elastic-plastic fracture mechanics (EPFM) is required.

When the crack plane is subjected to mixed modes (tension, shear, torsion) of loading, linear superposition is used to determine crack-tip stress fields (see Ref 9 for examples). However, failure criterion is written in terms of an effective $K$, which for modes I and II would be: $K_{\text{eff}}^2 = K_1^2 + K_II^2$. Many applications are for crack growth by opening (Mode I or Mode 1). When stress is applied in the crack plane, stress-intensity values are termed as $K_{1c}$ (shear or sliding), or as $K_{3c}$ (tearing or torsion along the crack plane).

References cited in this section


Subcritical Fracture Mechanics

Fracture mechanics analysis dealing with cracking (as opposed to fracture) is sometimes referred to as subcritical fracture mechanics. The final fracture is by overload. In subcritical cracking, the amount of plastic deformation during cracking is very small, and so LEFM concepts are typically valid for crack-growth analysis. An exception is creep cracking, for which EPFM concepts must be used. Creep crack growth is discussed in Ref 10.

Fatigue. In the case of cyclic loading, subcritical crack growth from fatigue loading, \( \frac{da}{dN} \), is characterized by the stress-intensity range, \( \Delta K = K_{\text{max}} - K_{\text{min}} \). During a load cycle, \( K \) varies from a minimum and maximum:

\[
K_{\text{min}} = Y\sigma_{\text{min}}\sqrt{\pi a} \\
K_{\text{max}} = Y\sigma_{\text{max}}\sqrt{\pi a}
\]

where the stress ratio is \( R = K_{\text{min}}/K_{\text{max}} \). The crack growth rate, \( \frac{da}{dN} \), depends on \( \Delta K \), and \( R \) and must be provided by a test of the material. Crack growth rates versus \( \Delta K \) typically have three distinct regions of behavior, as discussed in more detail in the article “Fatigue Failures” in this Volume.

Region 1 is the stage of early crack growth with a threshold value \( (\Delta K_{\text{th}}) \) below which crack rates become diminishingly small. If values of \( \Delta K \) are above \( \Delta K_{\text{th}} \), the crack size grows and thus causes the stress intensity to increase at the crack tip. The crack growth rates accelerate during stage 1 until they reach stage 2, where rates are described by the Paris equation:

\[
\frac{da}{dN} = C(\Delta K)^n
\]

where \( C \) and \( n \) are empirical constants, which depend on the stress ratio. In practice, fracture mechanics testing is typically done at a small positive \( R \) value (\( n, R \sim 0.1 \)), while service conditions range from -1 to +1 for \( R \). Because the Paris equation covers only one \( R \) value, an alternate form is:

\[
\frac{da}{dN} = \frac{C}{(1-R)^m} \Delta K^n = C \frac{\Delta K^m}{(1-R)^m} \Delta K^{n-m}
\]

This is known as the Walker equation. Many other equations are used, but it should be pointed out that the equations are phenomenological curve-fitting equations. Conditions, especially environments, can alter the shape of the curves. For example, Fig. 24 shows three cases of environmental effects:

- **True corrosion fatigue** (TCF): Fig. 24(a). Material-environmental combination where the threshold value is lowered and the chemical reaction is enhanced by the plastic behavior at the crack tip under fatigue loading. The chemical reaction is not influenced by the static stress at the crack tip.
- **Stress corrosion fatigue** (SCF): Fig. 24(b). Material-environmental combination where the chemical reaction is strongly influenced by the static stress at the crack tip (i.e., stress corrosion is observed) but is not influenced by the plastic behavior at the crack tip as a result of fatigue loading.
- **TCF combined with SCF**: Fig. 24(c)

The latter is the most common.
Fig. 24 Cases of environmental effects on fatigue crack growth. (a) True corrosion fatigue. (b) Stress-corrosion fatigue. (c) Stress-corrosion fatigue on true corrosion fatigue

Additionally, it is necessary to recognize that the Paris constants are in general anisotropic and for the same reasons that fracture toughness is anisotropic; that is, they depend on the direction of crack growth in mechanical/thermally processed material because of banding and segregation. Many references to $\Delta K - \frac{da}{dN}$ curves have superimposed comments about sensitivity of crack growth rates to loading condition variables (i.e., the $R$ value), to the environment and to the microstructure. These references indicate high sensitivity of the microstructure in regions I and III but low sensitivity in region II (the Paris law region). This must be carefully interpreted. It is true that alloy compositions (e.g., the range of austenitic stainless steels) all have about the same Paris constants and that all steels, pearlitic and quench and tempered, have approximately the same Paris constants. This is of considerable benefit when back calculations are done using experimental determination of critical overload crack lengths and fatigue striation spacing, since constants may not be available for the specific alloy. However, as just noted, the Paris constants can be very sensitive to solute and microstructure banding as well as inclusion morphology and distribution. Thus, in a sense, there is also strong microstructure sensitivity in the Paris law region of the curve.

Stress-Corrosion Cracking. In the case of stress-corrosion cracking, the rate of growth will depend on $K$ (similar to fatigue), but the rate of growth depends directly on time ($\frac{da}{dt}$) under a static stress. The curve of $\frac{da}{dt}$ versus $K$ typically exhibits three distinct regions (Fig. 25) in a fashion similar to fatigue. Stage 1 is the threshold region, where $\frac{da}{dt}$ depends strongly on $K$. Stage 2 is the steady-state region where crack growth is only slightly dependent on $K$. Stage 3 is the $K$-dependent region of fast fracture. The threshold value is best determined by performing tests on cracked specimens and by measuring the time to fracture. A plot of the stress intensity applied as a function of time to failure will show an asymptote (Fig. 26). For combined stress-corrosion and fatigue cracking (Fig. 24c), one obtains $\frac{da}{dN}$ from a cyclic-loading test in the appropriate environment.
Fig. 25 Stages of stress-corrosion crack growth for a component under static loading in an aggressive environment

Fig. 26 Effect of environment (3.5% salt solution) on time at sustained load on fracture toughness for titanium alloys and steel. Source: Ref 10

Reference cited in this section

Elastic-Plastic Fracture Mechanics

Different approaches are used for evaluating materials that fracture (or behave) more plastically than they behave in LEFM. This requires the application of EPFM or nonlinear elastic fracture mechanics. Elastic-plastic fracture mechanics is still based on elasticity concepts, but it must also account for the nonlinear portion of the stress-strain curve in the region of plastic deformation. Hence, the term nonlinear fracture mechanics is used. It is described only briefly here with more details contained in Ref 11.

The J-integral and the crack-tip opening displacement (CTOD) methods are most often used to characterize fracture at the crack tip in EPFM. As plastic deformation is involved, fracture resistance in EPFM can be characterized logically and expressed in units of energy per unit area. For example, a general (and simplistic) form of the J-integral may be given as follows (Ref 12):

\[ J = J_{el} + J_{pl} \]

where the elastic portion of the fracture-resistance energy \( J_{el} \) is:

\[ J_{el} = \frac{K_c^2(1 - \nu^2)}{E} \]

and the plastic portion of the fracture-resistance energy \( J_{pl} \) is:

\[ J_{pl} = \frac{\eta A_{pl}}{B_N B_0} \]

where \( \eta \) is factor based on the type of specimen, \( A_{pl} \) is the area under the load against displacement plot, \( B_N \) is the specimen thickness, and \( B_0 \) is the uncracked portion of the specimen (ligament). These equations apply to specific specimen geometries and load-crack situations as described in the ASTM test standards for fracture toughness testing (Ref 12, 13, 14). These expressions are given only as an example to show the link between \( J \) and \( K \) and to further show how the J-integral is expressed in two parts: the elastic and the plastic components. If the LEFM toughness is expressed in terms of an energy criterion, it equals \( \frac{K_c^2}{E} \) so that the dimension will be energy per unit area.

Due to the manner in which \( J \) is measured, the fracture energy often appears to increase when the fracture is in progress. For low-toughness materials, this increase is generally small and can then be ignored. For high-toughness material, the increase can be considerable. In this case, the fracture may first be stable, and a further increase in load is necessary to cause an uncontrollable (unstable) fracture. Because of historical development, \( J \) is usually referred to as the J-integral, and it can be expressed in the form of an integral. The J-integral is useful for computation of geometry factors, but for fracture mechanic applications, \( J \) can be derived without the integral.

References cited in this section


**Stress Analysis and Fracture Mechanics**

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

**Applications**

The strength of a structural component decreases with crack size, and calculations of residual strength (rather than the original design strength as a function of crack size) can be done using LEFM or EPFM. In most cases, the residual strength analysis is performed with LEFM and is based on the stress-intensity factor, $K$, and the toughness (the plane-strain fracture toughness, $K_I$, or the plane-stress fracture toughness, $K_c$), as discussed in more detail in the article “The Practice of Damage Tolerance Analysis” in Fatigue and Fracture, Volume 19 of ASM Handbook (Ref 15). Residual strength analysis on the basis of EPFM and plastic fracture mechanics is discussed in the article “Residual Strength of Metal Structures” in Fatigue and Fracture, Volume 19 of ASM Handbook (Ref 16). One method is the failure assessment diagram, which represents the whole gamut of fractures from brittle to fully plastic. The failure assessment diagram and the application of $J$-integral and CTOD data are discussed in more detail in the article “Failure Assessment Diagrams” in this Volume.

Subcritical fracture mechanics also permits calculation of the crack growth rates. Fracture control can be exercised, but only if both the residual-strength diagram and the crack-growth curve are available. If either is missing, nothing worthwhile can be said about the fracture safety of the structure. For example, if a crack is detected, one may be able to show that it can be sustained under the prevailing service loading. However, because the crack may be much longer the next day, it is not possible to keep the structure in operation unless it is known how long it will take for the crack to grow to a size that cannot be sustained. The crack-growth curve must be known as well.

Use of fracture mechanics in failure analysis can help in the evaluation of conditions leading to fracture and in fact may be critical to the determination of whether an imperfection is a defect (i.e., causative of failure). From the amount of crack growth (crack size at fracture), known stresses, and growth-rate properties, a reasonable insight can always be obtained regarding the question of misuse; for example, continuous overloading. The time to failure and final crack size are determined using fracture mechanics, as discussed in the following paragraphs. When results are not in accordance with observations, the analysis can be repeated to determine how much higher (or lower) the stresses would have had to have been to produce the cracking time and extent of cracking observed.

Fractography generally provides the crack size at which final fracture took place. Because fracture occurs in accordance with either LEFM or EPFM conditions, equations are available for using the fractographic results in various ways. If the operation stresses are known, equations can be used to calculate the toughness, which, if compared to the anticipated toughness, will show whether or not the material had a toughness higher or lower than expected. If the material specification indicated a toughness value, the calculation will show whether or not the material was in accordance with the specification. For more discussion, see the appendix in “Fracture Appearance and Mechanisms of Deformation and Fracture.”

Conversely, if the fracture stress is unknown, the fracture mechanics equations can be used to determine the stress at the time of fracture. In this case, a reasonably accurate (sometimes estimated) toughness value must be used. With the stress at fracture known, it can be determined whether the fracture took place due to an excessively high load or whether in general the operating stress levels were too high.

This information can often be corroborated by measuring the size of the stretched zone from the fractographs, provided stereomicroscopy is used (Fig. 27) (Ref 17). The size of the stretched zone is a measure of the CTOD at the time of fracture. It can be shown that the CTOD at fracture is related to the toughness as $\text{CTOD} = \frac{K^2}{\partial E\sigma}$ or $\text{CTOD} = J/\sigma$.
Fig. 27 Measurement of the stretched zone in a fractured specimen. (a) Crack-tip blunting and the stretched zone. (b) Topography of stretched zone. Source: Ref 17

Therefore, the toughness can be calculated from the measured stretched-zone size. Figure 28 (Ref 18) shows that careful measurement can give reliable toughness values. With the toughness known, fracture stresses can be calculated as previously discussed.
Crack Growth. Fatigue-striation counts (if striations are present and if they can be accurately measured) generally provide a reasonable account of the crack-growth rates and crack-growth curve. If the crack-growth-rate behavior of the material and the stresses are known, the stress intensity can be calculated and a comparison can be made between actual and
anticipated properties for a conclusion about the adequacy of the material. Conversely, if the stresses are not known, the measured rates and rate properties can be used to estimate what the acting stresses were. This procedure at least provides the magnitude of the stresses.

Any change of loading or change in environment during the cracking process is likely to leave its mark. Such changes produce a change in crack-growth rate, which is usually associated with a change in topography (macro-scale surface roughness and at micro-scale a discontinuous change in striation spacing). Because the roughness affects the reflection of light, a change in roughness will appear as a line (beach mark). Any beach mark is an indication of a change in circumstances during cracking. Beach marks are also created by corrosion at the crack tip when the material is in the unloaded condition for a period of time. Therefore, the presence of a beach mark may only indicate that the material was unloaded for some period of time, not that the stress intensity range changed. If inclusions or second phase particles are present in the microstructure, the striation spacing is also changed as the crack front interacts with these particles. The striation spacing thus is locally affected by the particle, not by a change in stress intensity range. The beach marks on fatigue-crack surfaces are well known. Similar marks may occur on a stress-corrosion crack surface due to changes in loading or environment.

Initial Defects. If fractography shows the presence of initial defects (mechanical or metallurgical), crack-growth analysis can be used, starting from the initial defect. Again, depending on the objective, information can be obtained on stresses (known material properties) or material properties (known stresses). One can thus obtain information on whether the design stresses were too high or whether the material was inadequate.

Residual, Contact, and Thermal Stresses. In general, the stress intensities due to different stress systems are additive as long as all stress systems are of the same mode, for example, mode I. The crack tip stress field due to combined bending and tension is the sum of those due to tension and bending. However, the effective stress intensity factor is then given by:

$$K_{\text{eff}}^2 = K_I^2 + K_{II}^2$$

Residual stresses due to welding, cold deformation, surface treatments, or heat treatments give rise to a stress intensity. If this stress intensity is of the same mode as that due to external loads, the two are additive. Although this is very useful in principle, the stress intensity due to residual stresses usually cannot be determined simply because the residual stresses are unknown.

For a fracture problem, it may be assumed that local yielding at the crack tip is so extensive that residual stresses will be annihilated and can be ignored for the fracture analysis. In the case of fatigue-crack growth, the residual stress is static as opposed to the cyclic applied stress. As such, the effect of the residual stress may be interpreted as a simple change in roughness and at micro-scale a discontinuous change in striation spacing. The beach marks on fatigue-crack surfaces are well known. Similar marks may occur on a stress-corrosion crack surface due to changes in loading or environment.

Residual stresses occur only when plastic deformation occurs very locally. On release of elastic strains of the surrounding material, the plastically (permanently) deformed region does not fit in the surroundings. The surrounding elastic material will force it to fit and exert stresses to do so. These stresses are called residual stresses because they occur without the presence of an external load. The surrounding elastic material will essentially return to zero strain, which forces the permanently deformed material nearly back to zero strain. Just as the original plastic strain required yielding, the return to zero strain will require reverse yielding. This leads to the conclusion that residual stresses always equal the yield stress. However, if the material is at elevated temperature where the yield is lower, the residual stress may be equal to the lower yield stress only.

These guidelines can be used to evaluate the effect of residual stress on $K$. It should be noted, however, that $K$ is an elastic parameter, and if the residual stress equals the yield, the applicability of $K$ is doubtful. Accounting for residual stresses in EPFM is equally debatable. However, application of this rule can at least lead to an estimate of the effect. Cautious engineering judgment is certainly required. The same arguments hold more or less for thermal stresses and contact stresses. Stress intensities are additive, but the problem is in knowing what to add.

Example 3: The Application of Fracture Mechanics to Fractography (Ref 6). In this example, the material is quenched-and-tempered 4340 steel with a yield strength ($\sigma_y$) of 160 ksi and a fracture toughness ($K_{IC}$) of 80 ksi $\sqrt{in}$. On a failed component made from this material, fractography is used to determine values for critical flaw size ($a_c$) and $2c$ (as defined in Fig. 21a) for surface cracks. For a surface flaw, the fracture mechanics relationships are:

$$K_{IC} = 1.1\sigma_y\sqrt{\pi a_c}/Q$$

$$\sigma_y = K_{IC}/1.1\sqrt{\pi a_c}/Q$$

If $a_c = 0.35$ in., $a/2c = 0.35/0.7 = 0.5$, and $Q = 2.2$, then:

$$\sigma_y = 80/1.1\sqrt{\pi 0.35/2.2}$$

$$\sigma_y = 103 \text{ ksi}$$

In this case, fracture stress is as would be expected for a correctly designed structure (that is, the safety factor (SF) is 1.5). Failure occurred because of extensive service-induced subcritical crack growth.
If $a_c = 0.17$ in., $a/2c = 0.17/0.34 = 0.5$, and $Q = 2.2$, then $\sigma_f = 148$ ksi. In this case, fracture stress is well over the expected design stress—thus, the component was overstressed. Failure occurred either because of a design error or an operating error.

If $a_c = 0.17$ in., $a/2c = 0.17/0.34 = 0.5$; $Q = 2.2$; and one knows that $\sigma_f$ can be no greater than 100 ksi, then $K_{Ic} = 54$ ksi $\sqrt{\text{in}}$. In this case, the fracture toughness of the material is well below the expected value of $80$ ksi $\sqrt{\text{in}}$

Failure occurred because the wrong material was used or the material was processed incorrectly.

If $a_c = 0.17$ in.; $a/2c = 0.17/1.7 = 0.1$; and $Q = 1.05$, then $\sigma_f = 102$ ksi. In this case, fracture stress is well below the expected design stress. The configuration of the crack (large width, $2c$, to its depth, $a$) suggests a failure due to a machining mark, surface scratch, or impact ding.

Note that these sample calculations are simple and are intended to show several different situations. However, the material is steel, and temperature is another important factor. There is a possibility that temperature at the time of failure was less than room temperature, and high strain rates could shift the DBTT. These calculations assume room temperature conditions and no strain rate effects on DBTT.

References cited in this section


Stress Analysis and Fracture Mechanics

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

Acknowledgments

Portions of this article were also adapted from Ref 5, 6, and 17.

References cited in this section


Appendix: Stress Tensor Decomposition and Transformation of Stress

Stress Tensor Decomposition

Any stress state can be decomposed into spherical and deviatoric tensors. In many engineering materials, while the spherical stress tensor is associated with volumetric change only, the deviatoric stress tensor (sometimes called the stress deviator) is associated with a change in shape. In these materials, plastic deformation is associated with a change of shape. Yield, therefore, occurs incompressibly, implying that Poisson’s ratio during plastic deformation, \( \nu_p \), is very nearly 0.5.

Decomposition of the stress tensor is straightforward and uncomplicated. If \( \mathbf{T} \) represents the stress state at a given point, then \( \mathbf{T} = \mathbf{T}_h + \mathbf{T}_d \), where \( \mathbf{T}_h \) and \( \mathbf{T}_d \) are the hydrostatic and deviatoric stress tensors, respectively. Mathematically:

\[
\mathbf{T} = \begin{bmatrix}
\sigma_h & 0 & 0 \\
0 & \sigma_h & 0 \\
0 & 0 & \sigma_h
\end{bmatrix} + \begin{bmatrix}
\sigma_x - \sigma_h & \tau_{xy} & \tau_{xz} \\
\tau_{xy} & \sigma_y - \sigma_h & \tau_{yz} \\
\tau_{xz} & \tau_{yz} & \sigma_z - \sigma_h
\end{bmatrix}
\]

in which

\[
\sigma_h = \frac{\sigma_x + \sigma_y + \sigma_z}{3}
\]

A yield criterion for ductile metals whose failure envelope is quite close to Tresca's is the distortion energy theory. Also known as the von Mises-Hencky theory, it can be shown that the von Mises effective stress, \( \sigma_e \), is a function only of the stress deviator. Thus, spherical stresses play nearly no role in plastic deformation of these materials. A consequence of this is that if the normal stresses, \( \sigma_x \), \( \sigma_y \), and \( \sigma_z \) are equal or nearly so, the spherical stress is relatively large and the effective stress is relatively small, reducing the tendency to yield.

Transformation of Stress

Ordered principal stresses, denoted as \( \sigma_1 \), \( \sigma_2 \), and \( \sigma_3 \), are the eigenvalues of the stress tensor, while the associated eigenvectors, \( \hat{\mathbf{n}}_1, \hat{\mathbf{n}}_2, \hat{\mathbf{n}}_3 \), are unit normals to the planes on which the principal stresses act. Several commercially available software packages such as MathCAD have routines that easily extract both eigenvalues and eigenvectors. Other embedded capabilities include both numerical and symbolic manipulation of mathematical objects such as matrices, integrals, and differentials in addition to nonlinear equation solvers.

Maximum shear stresses (both positive and negative) and the planes on which they act can be easily determined once the principal stresses and directions are known. Because maximum shear stresses are oriented on planes rotated 45° with respect to the planes on which both \( \sigma_1 \) and \( \sigma_3 \) act, unit normals to these maximum shear stress planes \( \hat{\mathbf{n}}_1 \) and \( \hat{\mathbf{n}}_2 \) are found to be:
As many fractures originate on these planes this information has both predictive and diagnostic value. In the design stage possible failure sites and orientations can be predicted. In the event of failure, stress analysis can point to improper use or overload as contributing factors.

As a numerical example to determine principal stresses, maximum shear stresses, and the planes on which they act, consider the following stress tensor, $T$:

$$
T = \begin{bmatrix}
10 & 4 & 8 \\
4 & 20 & 12 \\
8 & 12 & -5
\end{bmatrix} \text{ksi}
$$

The principal stresses and associated directions are found, using MathCAD, to be:

$$
\sigma_1 = 27.531 \text{ ksi} \quad n_1 = \begin{pmatrix}
-0.3737 \\
-0.8366 \\
-0.4005
\end{pmatrix}
$$

$$
\sigma_2 = 9.026 \text{ ksi} \quad \hat{n}_2 = \begin{pmatrix}
0.8844 \\
-0.4515 \\
0.1181
\end{pmatrix}
$$

$$
\sigma_3 = -11.557 \text{ ksi} \quad \hat{n}_3 = \begin{pmatrix}
0.2797 \\
0.3101 \\
-0.9086
\end{pmatrix}
$$

Note that to maintain a right-handed coordinate system, $\hat{n}_1 \times \hat{n}_2 = \hat{n}_3$. Hence, normals to the planes of maximum shear stress are given by:

$$
\hat{n}_1^{(1)} = \begin{pmatrix}
-0.066 \\
-0.372 \\
-0.926
\end{pmatrix} \quad \text{and} \quad \hat{n}_2^{(2)} = \begin{pmatrix}
-0.462 \\
-0.811 \\
0.359
\end{pmatrix}
$$

Transformation of a three-dimensional (3D) stress state from one Cartesian coordinate system to another Cartesian coordinate system that has been rotated with respect to the first is governed by the following:

$$
T' = A T A^T
$$

in which $T$ represents the original stress tensor and $T'$ represents the stress state in the primed system. $A$ is the matrix of direction cosines, and $A^T$ is its transpose. Each element of $A$, denoted $a_{ij}$, is the cosine of the angle between the $i$th axis of the primed system and the $j$th axis of the original, or unprimed system, $i, j = x, y, z$. Thus, if stresses are wanted in a coordinate system rotated 30° counterclockwise about the $y$-axis of a right-handed coordinate system, the matrix of direction cosines becomes:

$$
A = \begin{bmatrix}
0.866 & 0 & -0.500 \\
0 & 1 & 0 \\
0.500 & 0 & 0.866
\end{bmatrix}
$$

The corresponding stress state, $T'$, is then:
If another coordinate transformation (into the double primed system) is necessary, then a second matrix of direction cosines, \( B \), is required. If \( B \) represents the orientation between the double primed system and the primed system, then the transformation equation becomes:

\[
T'' = B A T A^T B^T
\]

The transformations \( A \) and \( B \) can be combined by defining \( C = B A \). In that case

\[
T'' = C T C^T
\]

which is in the same form as the single transformation.

---

**Stress Analysis and Fracture Mechanics**

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

---

**References**


Stress Analysis and Fracture Mechanics

John D. Landes, University of Tennessee; W.T. Becker, University of Tennessee, Emeritus; Roch S. Shipley, Packer Engineering; Julian Raphael, Columbus McKinnon Corporation

Selected References

**Mechanics of materials**


**Elasticity**


**Fracture mechanics and fatigue**

X-Ray Diffraction Residual Stress Measurement in Failure Analysis

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

Introduction

X-RAY DIFFRACTION (XRD) residual stress analysis is an essential tool for failure analysis. X-ray diffraction residual stress measurement methods are commonly applied to the successful completion of failure analyses and, in many cases, are the only viable ones for the acquisition of the required data. X-ray diffraction residual stress analysis methods can be used in (but are not necessarily limited to) situations where failures result from overloading, stress corrosion, fatigue, stress concentrations, or inappropriate manufacturing processes. X-ray diffraction residual stress analysis methods can also be applied to the evaluation of corrective measures and the optimization of production parameters for extending the service life of components. This article focuses primarily on what the analyst should know about applying XRD residual stress measurement techniques to failure analysis. Discussions are extended to the description of ways in which XRD can be applied to the characterization of residual stresses in a component or assembly, and to the subsequent evaluation of corrective actions that alter the residual stress state of a component for the purposes of preventing, minimizing, or eradicating the contribution of residual stress to premature failures.

The article presents a practical approach to sample selection and specimen preparation, measurement location selection, and measurement depth selection; measurement validation is outlined as well. A number of case studies and examples are cited. Failures with root causes other than residual stress are beyond the scope of this article and are not discussed. The article also briefly summarizes the theory of XRD analysis and describes recent advances in equipment capability. Comprehensive, detailed treatments of XRD residual-stress measurement theory can be found in Ref 1, 2, 3, 4, 5, 6, 7, 8.

X-ray diffraction techniques have been applied as early as 1925 to the measurement of residual stress on a variety of materials (Ref 9). Considerable advances have since been made in detector speed and resolution (Ref 10, 11), data analysis and handling (due primarily to the advent of the personal computer), and equipment portability (Ref 12). The importance of the XRD method resides in its ability to measure residual and applied stress with high spatial resolution, speed, and excellent accuracy, and, in many cases, measurements can be performed nondestructively (Ref 13). The measurement of residual stress via XRD is generally limited to polycrystalline materials (Ref 6), that is, in materials with a grain structure (long-range ordering) as normally found in metals and their alloys as well as in polycrystalline ceramics (i.e., oxides, carbides, nitrides, etc.). The size of specimens that can be evaluated using XRD can vary widely from as small as the head of a pin to as large as a ship (Ref 12, 14), a building/structure (Ref 15, 16), or a bridge (Ref 17, 18). Line of sight for both the incident beam and the detector to the measurement location is required, and the incident x-ray beam must irradiate the volume of interest (Ref 8). The precision and accuracy of results obtained is generally a function of the instrument used, the material condition, the measurement technique used, and subsequent data analysis. Unless otherwise specified, data analysis is based on the assumption that a near-random grain orientation distribution is sampled in a homogeneous and isotropic material (Ref 3). Although a detailed discussion is
beyond the scope of this article, XRD techniques can also be used for the measurement of percent retained austenite (Ref 19), and other phases, pole figures (Ref 7), and other analyses too numerous to mention.

References cited in this section


Residual Stress in Failures and XRD Analysis

The residual stresses present in a component can arise in almost every step of processing (Ref 1) and, in many cases, can play a key role in determining the effective service life of a component. When the failure analyst suspects residual stress as a contributing factor to a premature failure, such suspicions may be validated by experiment and by measurement. X-ray diffraction residual stress measurement analysis should be specified when damaged components:

- Experience static in service loads that are a significant fraction of the maximum allowable equivalent stress (stress concentrations, overloaded)
- Distort or form cracks without applied loads
- Are placed in corrosive environments (stress corrosion)
- Experience cyclic loading in service (mechanical or thermal fatigue)
- May have been subject to improper processing on manufacture (shot peening, grinding, milling, etc.)
- May have been subject to inappropriate heat treatment (stress relief, induction hardening, service temperature, thermal strains, etc.)

Reference cited in this section


X-Ray Diffraction Theory and Residual Stress Measurement

The XRD technique uses the distance between crystallographic planes, that is, \( d \)-spacing, as a strain gage. This method can only be applied to crystalline, polycrystalline, and semicrystalline materials (Ref 1, 20), where thousands of grains are sampled in a typical measurement. When the material is in tension, the \( d \)-spacing increases in the direction of stress, and when the material is in compression, the \( d \)-spacing decreases. The presence of residual stresses in the material produces a shift in the XRD peak angular position that is directly measured by the detector (Ref 6). For a known x-ray wavelength, \( \lambda \), and \( n \) equal to unity, the diffraction angle, \( 2\theta \), is measured experimentally, and the \( d \)-spacing is then calculated using Bragg's law:

\[
\sin \theta = \frac{\lambda}{2d}
\]
Once the \( d \)-spacing is measured for unstressed (\( d_0 \)) and stressed (\( d \)) conditions, the strain is calculated using the following relationship:

\[
\varepsilon = \frac{(d - d_0)}{d_0}
\]

(Eq 2)

When applying a plane-stress model, the unstressed lattice spacing, \( d_0 \), can be substituted with the \( d \)-spacing measured for the specimen of interest at \( \psi \) (psi) = 0; that is, the unstressed lattice spacing need not be known precisely for the material in question (Ref 3). To evaluate stresses in the direction normal to the specimen surface \( \sigma_{33} \), the unstressed lattice spacing for the material in question must be known with an accuracy of better than 0.01%. An error in the estimation of the unstressed lattice spacing, \( d_0 \), as small as 0.1% can lead to large errors in the stress measurement results (Ref 1, 21).

For the \( \sin^2 \psi \) method, where a number of \( d \)-spacings are measured, stresses are calculated from an equation derived from Hooke’s law for isotropic, homogeneous, fine-grain materials:

\[
\varepsilon_{\psi\psi} = \left( S_2 \left( \sigma_{11} \cos^2 \phi + \sigma_{12} \sin 2\phi + \sigma_{22} \sin^2 \phi - \sigma_{33} \right) \right) \sin^2 \psi + \frac{1}{2} S_1 \left( \sigma_{11} + \sigma_{22} + \sigma_{33} \right) \sin 2\psi + \frac{1}{2} S_2 \left( \sigma_{13} \cos \phi + \sigma_{23} \sin \phi \right) \sin 2\psi
\]

(Eq 3)

where \( \frac{1}{2} S_2 \) and \( S_1 \) are the x-ray elastic constants of the material, variations of \( \sigma_{ij} \) are the stress-tensor components, \( \phi \) is the angle of the direction of strain measurement with respect to the frame of reference, \( \psi \) is the angle subtended by the bisector of the incident and diffracted beam, with the specimen normal, and \( \varepsilon_{\psi\psi} \) is the strain for a given \( \phi\psi \) orientation. Figure 1 is a schematic showing the reference axes and direction of measurement.

Fig. 1 Definition of the reference axes and the direction of measurement in XRD residual-stress analysis

Evaluation of the stress-tensor components, \( \sigma_{ij} \), is generally straightforward and is normally carried out by plotting the measured \( d \)-spacing versus \( \sin^2 \psi \), with careful selection of the measurement directions, \( \psi \) and \( \phi \). A variety of mathematical models and measurement approaches have been proposed to evaluate the stress-tensor components of interest (Ref 1, 2, 3, 4, 5, 6, 7, 8).

References cited in this section


2. V. Hauk, Structural and Residual Stress Analysis by Nondestructive Methods, Elsevier, 1997


X-Ray Diffraction Residual Stress Measurement in Failure Analysis

J.A. Pineault, M. Belassell, and M.E. Brauss, Proto Manufacturing Ltd.

Analysis of XRD Data

Residual-stress measurement \( d \) versus \( \sin^2 \psi \) data can normally be categorized as follows: linear, elliptical with \( \psi \)-splitting, and nonlinear with oscillatory behavior (Fig. 2). Data obtained from near-isotropic and homogeneous samples with no shear stress are linear and resemble that shown in Fig. 2(a). Elliptical data, shown in Fig. 2(b) and commonly defined as \( \psi \)-splitting in \( d \)-spacing versus \( \sin^2 \psi \) plots obtained via XRD, are generally evidence of shear stress and/or instrument misalignment (Ref 1). Data similar to that found in Fig. 2(c) are most often caused by a nonrandom crystallographic orientation of the coherently diffracting domains in the volume sampled by the incident x-ray beam, that is, material condition.

**Fig. 2** Common types of \( d \)-spacing versus \( \sin^2 \psi \) plots. (a) Linear: exhibiting no shear stress. (b) Elliptical: exhibiting \( \psi \)-splitting due to shear stress. (c) Nonlinear: oscillatory behavior due to preferred crystallographic orientation. Source: Ref 1

Mechanical surface treatments with forces out of plane with the surface create shear stresses (and thus, \( \psi \)-splitting). Shear stresses can typically be found in samples after grinding, turning, rolling, inclined shot peening, and honing processes. Shear stresses can also be created due to material heterogeneity with nonuniform microstrains. This is often seen in multiphase materials (Ref 2). When \( \psi \)-splitting is observed, a nonlinear regression should be employed to fit the data points in \( d \)-spacing versus \( \sin^2 \psi \) plots (Eq. 3). A triaxial analysis can also be employed (Ref 1, 2, 5). These analyses generally require the collection of many data points at both positive and negative \( \psi \) angles and in many \( \phi \) directions.
The examples shown in Fig. 3 demonstrate that an elliptical fit gives good results (Fig. 3a), whereas a linear fit using either the $\psi > 0$ or the $\psi < 0$ branch leads to incorrect results (Fig. 3b). In this case, the results from the two opening branches of the curve differ by approximately 54 ksi.

![Graph](image)

**Fig. 3** $\psi$-splitting on steel using (a) elliptical fit and (b) linear fit for $\psi > 0$ and $\psi < 0$

With recent advances in equipment speed, the full XRD peak profile can be obtained quickly and easily. When using instrumentation having single-channel detectors, the full diffraction profile should always be collected. Equipment that allows the use of only two $\psi$ tilts (typically, $\psi = 0°$ and $\psi = 45°$) has limited applications. The use of only a few $\psi$ tilts can lead to erroneous results; thus, the double-exposure method with $\psi > 0$ or $\psi < 0$ is only viable in special cases and when it is certain that the $d$-spacing versus $\sin^2\psi$ plot is linear. This means that, to ensure accurate results on a wide variety of samples and measurement locations, many $\psi$ tilts using both $\psi > 0$ and $\psi < 0$ and elliptical fitting methods should be applied. The use of true $\psi$ or $\chi$ (chi) geometry, sometimes referred to as $\Omega$ (omega) and $\psi$, respectively, is recommended. The use of $\chi$ goniometer geometry has advantages, primarily a slightly increased tilt range where the $\psi$ tilt range is limited due to sample geometry constraints. However, when using the modified $\chi$ (or side-inclination) geometry, errors may be introduced due to the $\psi$ angle offset and therefore, should only be used when the specimen geometry eliminates the preferred methods (i.e., true $\psi$ or $\chi$). True-$\chi$ and modified-$\chi$ geometry are more sensitive to goniometer alignment, particularly with small goniometer radii; thus, care must be taken in sample and goniometer alignment.

**References cited in this section**

Instrument Calibration and Validation of Stress Measurements

A comprehensive approach to the measurement of residual stress using XRD should address instrument calibration and the validation of results. The five steps required to calibrate instrumentation and to validate stress measurement results are:

1. Alignment of XRD instrumentation
2. X-ray elastic constant determination
3. Surface condition evaluation
4. Collection parameter selection
5. Repeatability and reproducibility determination

Once all of the previous five steps have been completed successfully, residual-stress measurements via XRD techniques can be applied with a high level of confidence.

Alignment of XRD instrumentation is performed according to the ASTM E 915 standard (Ref 22). This procedure is also performed periodically to verify instrument conformity to specifications. The maximum acceptable stress is ±2 ksi (for steel), tensile or compressive. The verification of instrument alignment is usually performed on a stress-free powder of the material in which residual stress is to be measured. Ideally, the shear stress should also be on the same order of magnitude as the normal stress, that is ±2 ksi, although technically, this would exceed the ASTM E 915 specification. Excellent instrumental alignment can easily be achieved for a goniometer with a focal radius as small as 30 mm (1.2 in.) (Fig. 4).

Fig. 4 Elliptical analysis on stress-free iron powder with a 30 mm (1.2 in.) focal radius goniometer

X-Ray Elastic Constant Determination (ASTM E 1426). For stress measurements using XRD techniques, one needs to determine the x-ray elastic constants (XECs). They are a function of the mechanical properties of the material and the crystallographic plane selected to perform the measurements. Once known, they can be used in subsequent measurements on the same alloy but can be affected by preferred orientation. In general, the XECs are determined experimentally or by calculation using theoretical methods. Theoretical models, such as those in Ref 23 and 24, are generally used when a suitable coupon cannot be obtained. The experimental method is usually the preferred method, particularly when the physical parameters of the material structure are not well known. The experimental method uses either a four-point bend or uniaxial tensile specimen, and ASTM E 1426 (Ref 25) is usually applied. In the case of four-point bend testing, a strain gage is attached to the opposite side and/or near the measurement location on the specimen. The applied strains are
recorded via a strain gage, and the stresses are measured by XRD at different load increments. The maximum applied stress is approximately 70% of the yield stress. For example, the slope of the stress measured via XRD versus the applied stress for a ground steel coupon (Fig. 5) was used to calculate the $\frac{1}{2}S_2$ parameter in Eq 3. $S_1$ is calculated for triaxial methods, and its determination is not described by ASTM E 1426 (Ref 25).

Fig. 5 Results obtained from a steel four-point bend specimen in round robin tests. The XEC $\frac{1}{2}S_2$ was determined to be $3.687 \times 10^{-5}$ ksi$^{-1}$

Surface Condition Evaluation. The effect of factors such as surface condition (surface roughness, oxide layers, etc.) on XRD residual stress measurement results should be addressed prior to surface-stress measurements on real components. For example (Ref 26), a steel specimen was machined in three regions, each with different machine tools, to introduce variations in surface roughness on the same specimen. The stress was then measured in the three regions via XRD techniques under independently monitored tensile loads. X-ray diffraction results were plotted versus the applied stress for locations where the average roughness ($R_a$) was 1, 3, and 6 μm (1 μm = 39.4 μin.). The linear least squares regression for each data set yields slopes of 1.01, 0.96, and 1.03 for 1, 3, and 6 μm $R_a$, respectively (Fig. 6a). All of these results are well within the experimental error (less than 5%), thus indicating that an $R_a$ of up to 6 μm for these experimental conditions has no measurable effect on residual-stress results obtained via XRD.

Fig. 6 X-ray diffraction stress versus applied stress for varying average roughness ($R_a$). (a) Samples with $R_a$ of 1, 3, and 6 μm. (b) Samples with $R_a$ of 1, 40, and 56 μm. (Source: Ref 26)

Another sample was prepared with regions machined to an $R_a$ of 1, 40, and 56 μm. It was loaded in tension, and stresses were measured via XRD. In this case, the slopes were 0.9, 0.15, and 0.18 for 1, 40, and 56 μm $R_a$, respectively (Fig. 6b).
This demonstrates that the surface-stress measurements do not respond as well to the applied stress when the roughness is increased.

The response of stress measurements via XRD to the applied load for an $R_a$ of up to 56 $\mu$m is plotted in Fig. 7. Here, it can be seen that the sensitivity to the applied load is reduced when the $R_a$ is greater than 10 $\mu$m (0.0004 in.), in this particular case.

![Fig. 7 Ratio of measured stress and applied stress for varying $R_a$. Source: Ref 26](image)

When the surface $R_a$ is less than the penetration depth of x-rays in the material, the measured stress results more accurately reflect the applied load (Fig. 8a). However, when the surface $R_a$ is greater than the penetration depth of the x-rays, the measured stress reflects the applied load to a lesser degree (Fig. 8b).

![Fig. 8 Effect of surface $R_a$ on XRD stress measurements. (a) X-ray penetration depth is greater than $R_a$. (b) X-ray penetration depth is less than $R_a$. Source: Ref 26](image)

The depth of penetration of x-rays in a material is dependent on the wavelength, $\lambda$, (i.e., the energy) of the incident radiation and the mass absorption coefficient, $\mu$, of the material (Ref 6). The mean depth of penetration can vary but is typically 10 to 20 $\mu$m for most metals.

Effect of Corrosion. The effect of surface corrosion on measurement response can be demonstrated by an example. A strain gage was attached to the surface of a steel wire on the inner side adjacent to the XRD stress-measurement location of interest. The XRD measurement location was visibly corroded. The wire was then incrementally loaded in tension. The stress was measured using XRD techniques, and the applied stress was monitored via the strain gage. The slope of the XRD stress versus applied stress was 0.889. The same steel wire was electropolished, and the experiment was repeated. The slope of the XRD stress versus applied stress was 0.997.

The results plotted in Fig. 9 indicate that, in this particular case, the electropolished surface (Fig. 9b) responds better than the corroded surface (Fig. 9a). Because the near-surface layers are often of interest to the failure analyst, electropolishing techniques and near-surface residual-stress measurements may be required.
Collection Parameter Selection. A number of factors can contribute to the introduction of random and/or systematic errors into residual-stress measurements performed via XRD techniques; thus, the data collection parameters should be carefully selected. Components can exhibit different microstructures that require the use of different collection parameters. In general, these collection parameters can be summarized as follows:

- Collection time
- Number of $\psi$ tilts used for $d$-spacing versus $\sin^2\psi$ plots
- X-ray diffraction peak position determination
- Effects of microstructure
- Effects of surface curvature and beam size

Optimization of the previous collection parameters should be addressed prior to proceeding with residual-stress measurements.

Collection Time. Low collection times can introduce random errors due to insufficient x-ray counting statistics (Ref 1, 7, 27). Random errors become apparent during repeatability tests; therefore, x-ray data collection times should be increased sufficiently until the desired repeatability is achieved. This technique works well for all detector types, particularly multichannel solid-state detectors that do not resolve individual counts. When using position-sensitive proportional counters, the number of counts can be used as a measure of counting statistics.

Number of $\psi$ Tilts Used for $d$-Spacing Versus $\sin^2\psi$ Plots. Improper selection of the number of $\psi$ angles used can introduce systematic errors (see the section “Analysis of XRD Data” in this article). Systematic errors become apparent during reproducibility tests. In general, a minimum of five $\psi$ tilts are required for a reasonable assessment of the shape of the $d$-spacing versus $\sin^2\psi$ curve; however, it is recommended that more than five tilts be used as a general practice.

X-Ray Diffraction Peak Position Determination. Improper selection of the peak location method can also introduce systematic errors in XRD residual stress measurement results. Depending on the specimen under examination, the XRD $K\alpha_1$-$K\alpha_2$ doublet may or may not be resolved. A number of factors can affect how well the doublet is resolved, including material effects such as the dislocation density, the crystallite size in the sample, and instrumental effects such as slit size and defocusing (Ref 1). In practice, that is, in most cases, XRD patterns obtained from real-life components exhibit broadening due to cold-working effects such as machining, grinding, shot peening, and so on. If the doublet is not well resolved, the peak position can be determined by fitting the blended doublet as a whole. The position of the blended doublet can thus be determined via analysis of the full profile (e.g., gravity center methods), and the splitting of $K\alpha_1$ and $K\alpha_2$ lines is quasi-automatically averaged. The application of peak fitting methods based on the analysis of selected parts of the profile should only be used exceptionally when peak asymmetry exists (Ref 2).

If the $K\alpha_1$-$K\alpha_2$ doublet is reasonably well resolved, it is recommended that the Rachinger correction be applied (Ref 7). This method consists of eliminating $K\alpha_2$, assuming that it has the same shape as $K\alpha_1$ and its intensity is half. Once the $K\alpha_2$ portion of the doublet has been eliminated, symmetric functions such as Voigt, Gaussian, Pearson VII, and full-profile methods work well (Ref 2). In contrast to the center-of-gravity method, the Gaussian offers the advantage that it may easily be extended to two or more profiles with their $K\alpha_2$ peaks removed, using the Rachinger correction. Because the intensity ratio and the difference in wavelength, $\Delta\lambda$, of the $K\alpha_1$ and $K\alpha_2$ lines is known for all x-ray tube targets, two functions (such as Pearson VII, Gaussian, modified Lorentzian, etc.) with the appropriate angular separation and relative intensity can be used to fit the doublet. In special material conditions or in the case of high-resolution experiments, it is recommended to use Pearson VII or the more physically based Voigt functions (Ref 2). The cross-correlation method is another method for peak location determination; however, this method is not applicable for textured materials (Ref 1).
Effects of Microstructure. All materials, at the single-crystal level, are anisotropic to some degree. If a nonrandom grain orientation distribution is sampled, the effects become apparent in either the shape and/or relative intensity of the diffraction peaks and in the $d$-spacing versus $\sin^2 \psi$ plots. Typically, large grain size or preferred orientation effects result in a nonlinear or oscillatory $d$-spacing versus $\sin^2 \psi$ plots (Fig. 2c). In the case of large grain-size effects, the peak shape generally changes drastically from $\psi$-tilt to another $\psi$-tilt, and corrective actions may include increasing the aperture size and/or oscillation of the $\psi$ and/or $\phi$ axes. In the case of preferred orientation, the peak shape generally remains constant with progressive broadening, but the peak intensity varies widely from $\psi$-tilt to another $\psi$-tilt. As a rule of thumb, an intensity ratio greater than 2.0 indicates there is a potential material condition problem. Corrective actions may include linearization (Ref 2) over the highest $\psi$-tilt range possible and others (Ref 1, 28, 29).

Effects of Surface Curvature and Beam Size. For specimen curvature effects, the following rule of thumb should be adhered to: for cylindrical specimens of radius $R$, the maximum incident x-ray spot size to use is $R/6$ for 5% accuracy and $R/4$ for 10% accuracy in the hoop direction, and $R/2.5$ and $R/2$ for 5% and 10% accuracy, respectively, in the axial direction. In cases where the beam size cannot be made sufficiently small, corrections can be applied (Ref 5).

Repeatability and Reproducibility. To assess the effects of random and/or systematic errors in residual stress measurement results, the repeatability and reproducibility of such measurements may be determined. Repeatability is a measure of the consistency in measurement results collected sequentially without a change in experimental setup. Reproducibility is a measure of the consistency in measurement results when the experiment is setup again at a different time, with a different instrument and/or with a different operator. The residual-stress measurement data found in Table 1 were collected on a laboratory standard. The repeatability test result average in Table 1 has a standard deviation of 0.7 ksi, and the reproducibility test result average has a standard deviation of 1.1 ksi. The slightly higher standard deviation in the reproducibility test versus the repeatability test is due to:

- A slightly different population of grains sampled
- Small variations in sample positioning and alignment
- The nonuniformity in surface residual stress of the specimen

Table 1  **Repeatability and reproducibility of stress results**

<table>
<thead>
<tr>
<th>Measurement No.</th>
<th>Stress, ksi</th>
<th>Repeatability</th>
<th>Reproducibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-62.8</td>
<td>-62.9</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-62.9</td>
<td>-61.1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>-63.2</td>
<td>-63.0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>-62.6</td>
<td>-63.7</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>-61.4</td>
<td>-63.8</td>
<td></td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td>-62.6</td>
<td>-62.9</td>
<td></td>
</tr>
<tr>
<td><strong>Standard deviation</strong></td>
<td>0.7</td>
<td>1.1</td>
<td></td>
</tr>
</tbody>
</table>

A similar repeatability and reproducibility test should be performed when a new material, material condition, or type of sample is to be analyzed using XRD.

**References cited in this section**

Sample Selection

In general, a statistically representative population of specimens should be chosen for analysis, so as to accurately sample the potential scatter within a population. In practice, the number of specimens that are representative of a population is related to how well the manufacturing processes are controlled. Such scatter (or lack of process control) in and of itself can be a contributing factor to a wide variability in the effective service-life-to-failure period for components manufactured with uncontrolled or poorly controlled processes. This can result in a drastically reduced predicted life for a given component, simply to accommodate the shortcomings of a small percentile of the component population.

If a normal distribution is used to predict the residual-stress distribution at a given location among a population of specimens, then a minimum number of five specimens should be chosen to adequately sample the potential distribution within the population. For processes where the standard deviation in the residual stress is less than the experimental error, it makes economic sense to reduce the number of specimens to represent the population. In the authors' experience, budget constraints often limit the sample population to two or three specimens; however, if variations in results at equivalent locations on equivalent specimens are larger than the experimental error, more specimens should be obtained for measurement.

Separate specimen populations should be selected to represent the manufacturing stage(s) of interest, so as to evaluate the effect of a given process on the stress state of the component. This can help identify the manufacturing process or production line that may be introducing an undesirable stress state in the component. Knowledge of the component service history can also help identify how undesirable stress states were introduced in service or how a failure may have occurred. Ideally, samples should be selected from real production populations to more accurately represent the true variance that can be expected within a population. Characterizing a sample population of only one using XRD residual-stress analysis can be risky, because it may inadequately represent the statistical distribution of residual stresses in the component population. Components that have already failed may be of limited use, because residual stresses may have changed or relaxed considerably upon failure. Ideally, nonfractured components with a similar history should be considered for analysis as well.
Measurement Location Selection and Location Access

When a failure occurs, the failure analyst must define the damage type and the origin of failure. This may be relatively straightforward or difficult, depending on the type of component and/or structure in question. Generally, the point of origin of a crack can be easily located on a given component with standard nondestructive evaluation techniques (i.e., visual inspection, eddy current, fluorescent dye penetrant, ultrasonic, etc.). In more complicated structures, the failure analyst must identify the components and locations most vulnerable to cracking.

Empirical Observation Method for Measurement Location Selection. An empirical guideline for determining the locations where cracks are most likely to occur and thus, where residual-stress measurements should be concentrated, can be summarized as follows:

- The highest-stressed areas are the areas with the smallest thickness or cross section for a given applied load.
- Stress concentration geometry: undercuts, necks, radii, weld toes, fillets, etc.
- Locations of dynamic contact (fatigue contact)
- Locations exposed and susceptible to corrosion

Once a crack has already initiated or failure has already occurred, the location of interest for XRD measurement is in that area. It should be noted that in the vicinity of the crack, the residual stress is in part or completely relieved in the direction normal to the crack line, hence the importance of evaluating parts prior to crack initiation. Triaxial XRD stress measurements are preferable to determine the orientation and the magnitude of the stress field, when it is possible or practical. In cases where the accessibility around the location of interest is limited to one or two directions (either due to sample geometry, economic, or time constraints), the priority is given to stress measurements in the direction perpendicular to the crack line.

Finite-Element Method for Measurement Location Selection Finite-element methods are very useful to predict the stress distribution in a component under applied loads. This method requires:

- Building a two- or three-dimensional (2-D, 3-D) map of the component or structure
- Determining the boundary conditions
- Performing calculations in 2-D or 3-D, depending on the complexity of the loading
- Printing the 2-D or 3-D stress distribution results (components in different directions, plastic strains, equivalent strains and stresses, etc.)

The results identify high- and low-stress areas and stress gradients. The locations with high priority for residual-stress measurements are the highest-tensile-stressed areas, that is, the areas that are most susceptible to localized yielding or cracking. At the boundaries where the stress gradient changes sign, shear stresses may be present; thus, triaxial stress measurements at these locations are recommended. For components of complex geometry where finite-element methods are not economically practical, the empirical observation method is preferable.

Stress-Mapping Method for Measurement Location Selection The introduction of stress-mapping techniques has allowed the rapid and precise characterization of entire component surfaces, including areas of interest such as steep stress gradients (as found in welds) and their associated tensile residual-stress maxima. In cases where the component geometry is such that stress measurements can be performed in the areas of interest (near the failure), stress mapping can be performed in both loaded and unloaded conditions. Stress-mapping techniques are routinely applied to welds, ground areas, shot-peened areas, rolled areas, and locally heat treated areas with temperature gradients. Subsequent measurements should be concentrated in areas identified by stress maps as the most tensile with higher resolution and/or with depth.

The stress-map display makes available to the failure analyst a complete and accurate visual analysis of the magnitude and distribution of residual stresses in components. The plots shown in Fig. 10, and 11 are examples of stress maps collected on components with surface-stress gradients. In Fig. 10, the tensile stresses detected in the center of the map (in the weld metal and adjacent heat-affected zone) may limit the service life of the saw blade (Ref 30). Similarly, in Fig. 11, the unpeened tensile stress region of a 316L stainless steel butt welded plate may be the location most susceptible to failure (Ref 31).
Accessing Measurement Locations of Interest. Using XRD, stresses in areas as small as a fraction of a millimeter can be measured. Accessing measurement locations on flat surfaces is straightforward but impossible at locations such as the inside diameter of a 2 mm (0.08 in.) hole. X-ray diffraction residual-stress measurements at locations where access is a problem normally require sectioning of the component to access the location of interest and, in general, require the evaluation of applied stresses caused by the relaxation due to sectioning. The most common method to monitor the strain relaxation during sectioning is with the use of electrical resistance strain gages. The placement and orientation of strain-gage elements prior to sectioning requires careful consideration.

For example, when residual-stress measurements are required on the inner diameter of a small ring, and the diameter of the ring is too small for the goniometer to obtain line of sight to the inner-diameter measurement locations, strain gages should thus be placed on both the outer and inner surfaces, so as to measure strains in the direction(s) of interest. Both outer and inner gages should be as close to the stress measurement location as possible. If other directions are to be monitored, dual-element strain gages or rosettes can be used for this purpose. The relaxation is generally assumed to be elastic; however, in highly stressed materials, it can also be elasto-plastic. The strain monitoring should therefore be performed in real time, that is, continuous mode. Care must be taken to ensure that the chosen cutting method does not cold work or overheat the XRD stress measurement location of interest.

The following summarizes the steps for sectioning samples prior to XRD analysis:

- Orient strain-gage element(s) so as to measure strains in the direction parallel to XRD stress measurement direction(s).
- Use coolant with the cutting tool (saw blade, cutoff wheel, EDM wire, etc.) to keep the sample temperature low, thus minimizing localized thermal effects and preventing stress relaxation due to heat dissipation generated during sectioning.
- Always protect the gages from coolant contact (particularly if the coolant is conductive) with epoxy, lacquer, or equivalent coatings (see strain gage manufacturer's recommendations).
• Monitor (in real-time) the variation of the strains during sectioning.
• Look for strain peaks during sectioning, and compare the value to the yield stress of the material, because plastic deformation may occur.
• Use the final strains to calculate the stress relaxation correction due to sectioning. In general, the calculated stress values are subtracted (added with opposite sign).

When complicated sectioning is required, more advanced models may be required and applied (see Ref 8 for more details). Finite-element methods can also be useful in such cases.

References cited in this section


X-Ray Diffraction Residual Stress Measurement in Failure Analysis

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

Selecting Measurement Directions and Depths

Selecting Measurement Directions. X-ray diffraction residual stress measurement directions can be selected as follows:

- For characterization of the complete stress tensor (triaxial analysis), the measurement of stresses in six directions is recommended (a minimum of three are required).
- Initial stresses generated during processing can be a good indication; for example, in the case of rolling, grinding, or turning processes, the directions parallel and perpendicular to the rolling, grinding, or turning direction are of interest.
- In specimens where induced stresses are omnidirectional, one direction or two orthogonal directions may be sufficient.
- If a crack has initiated in any direction, the highest priority measurements are in the direction perpendicular to the crack.
- When measurement directions are constrained by sample geometry, measurements at 0°, 45°, and 90° for any reference frame may be used to calculate principle and shear stresses using Mohr’s circle for the biaxial case. Similar methods can be applied for the triaxial case as well.

Selecting Measurement Depths. A comprehensive residual-stress investigation using XRD is seldom limited to the surface, and thus, subsurface measurements are generally required. If subsurface locations must be measured nondestructively, neutron diffraction is recommended (Ref 5, 8). When surface and subsurface measurements are performed, stress gradients normal to the specimen surface can be characterized, thus exposing potentially beneficial or harmful subsurface residual stresses in the material. Using XRD, the residual stresses should be corrected for stress relaxation using the Moore-Evans (Ref 32) method, where the material removed is over the whole surface. The stress correction with this technique requires measurements in one direction for flat surfaces and two directions for cylindrical surfaces (hoop and axial). For more complicated specimen geometry, finite-element techniques can be used to correct for stress relaxation due to material...
removal (Ref 3, 33). In general, the correction is within the experimental error of the residual stress measurement (i.e., is negligible) if the deepest layer removed is a small fraction of the total thickness of the component. When steep stress gradients normal to the surface exist in the component, the stress gradient correction should also be applied to collected data (Ref 3). The number of subsurface residual stress measurements performed is not limited, and measurement depths are generally selected (as required) to view the full shape of the stress versus depth profile to depths where stress gradients tend to level off. The actual depths required to characterize a subsurface stress gradient are very process- and sample-dependent and should be determined on a case-by-case basis.

References cited in this section


X-Ray Diffraction Residual Stress Measurement in Failure Analysis

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

Specimen Preparation

Most mechanical and structural components are protected when used in aggressive or harsh environments to prevent or minimize material degradation mechanisms such as oxidation, wear, erosion, corrosion, and so on. Structures exposed to the elements are often coated with zinc or painted, whereas mechanical components used in a dynamic mode are often lubricated. Protective coatings can partially or completely attenuate the incident x-ray beam and must be removed prior to residual-stress measurement. Removal of coatings must be performed without modifying the surface or subsurface residual-stress state. Mechanical polishing or grinding modifies the residual stress at the surface and the subsurface and is not recommended. If mechanical surface preparation is required for other nondestructive testing techniques (such as eddy current or ultrasonic) to be used in parallel with XRD, the XRD analysis should be performed before the surface is additionally disturbed or cold worked. It is important that the chemicals used to remove coatings never etch the material. If etching occurs, subsequent electropolishing below the etched layer may be required. To evaluate the appropriateness of the surface-preparation techniques applied and the surface condition of the component to XRD residual-stress measurements, a complete surface-condition evaluation (four-point bend test) must be performed on a coupon taken from the component in the as-prepared condition.
Residual-Stress Effects on Components under Quasi-Static Loading (Ref 2)

The basis for a quantitative assessment of statically loaded components is dependent on the determination of the maximum allowable equivalent stress, $\sigma_e$, which is determined by the underlying effect chosen (i.e., the resulting normal stresses, shear stresses, deformations, etc.), its location, and the defined failure criterion (i.e., fracture, plastic deformation, etc.). Thus:

$$\sigma_e \leq \frac{R}{s}$$

(Eq 4)

where $R$ is the resistance to failure (and depends on the failure criterion), $s$ is the safety factor, and the allowable equivalent stress, $\sigma_e$, is a function of the loading stresses and residual stresses. The sign and magnitude of the residual stress may increase or diminish the equivalent stress, and subsequently, surface and subsurface residual-stress gradients can modify their contributions locally. In cases where plastic deformation occurs in ductile materials, their effect may diminish, because residual stresses can change or relax substantially prior to and on failure and thus may or may not play a significant role in the failure (depending on the failure criterion). The effects of residual stresses are generally more dominant in brittle failures or when plastic deformation is the failure criterion (Ref 2).

Tensile residual stresses can significantly reduce fracture loads (Fig. 12), whereas compressive residual stresses generally act to increase the crack opening thresholds for given loading stresses (Fig. 13).

![Fig. 12 Effect of tensile residual stress (RS) on fracture loads as a function of test temperature. Source: Ref 34](image-url)
When applying linear elastic fracture mechanics, the residual-stress contributions to the stress-intensity factor can be calculated and used to predict whether crack growth and/or arrest occurs. It should be noted that the effect of residual stresses on crack growth rates tends to diminish with increasing fracture toughness. Residual stresses can also be taken into account when elastic-plastic fracture mechanics are used (Ref 2).

References cited in this section


---

X-Ray Diffraction Residual Stress Measurement in Failure Analysis

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

Stress-Corrosion Cracking and Corrosion Fatigue (Ref 31)
Environmentally assisted cracking, also known as stress-corrosion cracking (SCC), is a major source of potential failures in the process industries, in pulp mills, in storage vessels, and even in aircraft. Quite often, SCC occurs in the heat-affected zone (HAZ) immediately adjacent to a weld, simply because the HAZ is left in a state of very high residual tensile stress as a result of the shrinkage and differential cooling occurring in most welds. Tensile stress (resulting from the superposition of residual and applied stresses) (Ref 61) is the main component of the SCC triangle; the other two are a susceptible metal and an environment that often needs to be only slightly corrosive to that metal. For instance, grade 316 stainless steel is essentially inert to the corrosive effect of common salt unless tensile stresses are present, when it becomes very sensitive to chloride-induced SCC.

There are a number of possible solutions to the SCC problem. The obvious one is to change the environment, but that is rarely possible. The next is to change the metal, but usually that is expensive and, if the component in question is already built, impractical. Thermal stress relieving is a partial solution at best, because, to completely relieve all the tensile stresses in a component, it is necessary for the heating to reach the annealing temperature that may change the material properties. In addition, annealing cannot be used to overcome any subsequently applied tensile service loads. Corrosion engineers have long recognized that an effective solution for the retardation or even prevention of SCC is the introduction of compressive stresses.

For example, it can be seen in Fig. 14 that the shot peening technique used had a significant effect on the stress state of the weld and parent material, as seen by the “step” or drop in residual stress near the center of the stress map. On the left-hand side, a typical weld stress map is observed, with tensile residual stresses in the weld and in the HAZ, then dropping off in the parent material. The right side of this map was the peened portion. Here, the characteristic profile is much more compressive (or less tensile) and smooth; however, tensile residual stresses still exist. This indicates that the peening process had the effect of reducing the tensile residual-stress field in the weld and HAZ and introducing a much more uniform compressive residual-stress level in the parent material. However, it was not sufficient to make the surface stresses in the HAZ entirely compressive. This would suggest that the postweld treatment could be changed or augmented to increase the compressive residual stress imparted on the weld and HAZ.

Fig. 14  X-ray diffraction residual-stress map showing the introduction of compressive surface residual stresses in the parent material and the reduction, but not elimination, of tensile residual stresses in the weld metal on the unmasked side from shot peening a nickel alloy weldment. Source: Ref 31

X-ray diffraction techniques can thus be used to characterize the stress state of components that may be susceptible to SCC either prior to or after they have been put into service. In the case of corrosion fatigue, compressive surface residual stress has a beneficial effect on lifetime and strength. In the case of data shown in Fig. 15 (Ref 36), it can be seen that XRD residual stress measurements can be used to compare the effects of grinding and shot peening on the surface and subsurface residual stress state.
The strong impact that surface and near-surface residual stresses have on the fatigue life of components underlies the importance of studying the effect of surface treatments and manufacturing processes (Ref 2, 37, 38). It is now known that if reliable fatigue-life estimates are to be made, it is necessary to characterize the residual-stress fields in test specimens and engineering components (Ref 39).
When a component undergoes cyclic loading, it can be susceptible to fatigue. The fatigue life for a given component is often characterized with a Wöhler diagram or $S$-$N$ curve, where $S$ is defined as the total stress range the material experiences during cyclic loading, and $N$ is defined as the number of cycles to failure. Consider the simple case of constant amplitude cyclic loading (Fig. 16) (Ref 40). The peak-to-peak difference in the maximum and minimum stress is defined as the stress range. However, in the case of Fig. 16, the mean stress cannot be simply considered as the mean applied stress but must be considered as the superposition of the mean applied stress and the residual stress. Thus, a nonzero residual stress offsets the mean stress about which the stress amplitude cycles. This principle applies to variable amplitude cycling as well. Changes in the residual stress for a component in a given application (and thus, the mean total stress) can have the effect of displacing the $S$-$N$ curve and subsequently changing the fatigue life.

![Fig. 16 Typical constant amplitude cyclic loading spectrum. Source: Ref 40](image1)

$S$-$N$ curves are plotted in Fig. 17 for two identical gears of identical hardness, except one was double shot peened. The stress range of the fatigue limit of gear A is 1256 MPa, whereas an increase of 38% to 1710 MPa is observed for the double shot peened gear B (Ref 41).

![Fig. 17 S-N curves for as-hardened (gear A) and as-hardened plus double shot peened (gear B) gears. Source: Ref 41](image2)

Plastic strain, which is generally more significant when operating in cyclic fatigue, has a predominant influence in the number of cycles to failure (Ref 42). The relation between the plastic and the elastic strain can be summarized as follows: in high-life ranges, the plastic strain rapidly diminishes to negligible values, and the elastic strain range dominates. In the low-life range, the plastic component of strain dominates the material behavior. The predominant driving force in fatigue (elastic or plastic) is thus determined by the in-service strain/stress range. In the case of short fatigue lives, plastic strain is more dominant than the elastic strain, so that low-cycle fatigue (LCF) life is controlled by the ductility of the material. At longer fatigue lives, as in the case of high-cycle fatigue (HCF), the elastic strain is more significant than the plastic strain, and the fatigue life is determined by fracture strength. Another difference between LCF and HCF failures is that LCF is characterized by multiple cracking in highly stressed areas, whereas HCF failures generally initiate at precise stress concentration sites, with cracks propagating from a single initiation (Ref 43).
In cases of LCF failures, the in-service stress exceeds the elastic limit of the material. Low-cycle fatigue is usually used to define the reverse elastoplastic loading mode (Ref 44, 45). Low-cycle fatigue is characterized by cumulative fatigue damage associated with cycles to failure of up to $10^4$ to $10^5$ cycles. Experience has also demonstrated that LCF life is sensitive to changes in stress (or strain). A 5 to 10% difference in stress can result in a 50% difference in life; thus, errors in the representation of total stress have a profound effect on life, if residual stress is not in the equation (Ref 46).

Example 1: Use of XRD to Assess Residual Stresses in Steel Springs. A steel spring used in an automotive application suddenly began to fail in the field. It was understood that “nothing had changed” in the fabrication process of these springs, yet the incidence of field failures suddenly increased dramatically. Fatigue tests, using springs fabricated prior to field failures, lasted 500,000 cycles to failure, whereas fatigue tests performed on springs fabricated after field failures lasted only 50,000 cycles to failure. It was discovered that the percent coverage of shot peening prior to and subsequent to the increase in failure incidence was much less than 100%, with a shot peening time of 12 min. Subsequently, a potential corrective action was introduced by the engineers: an increase in the time the spring was shot peened from 12 to 60 min. The residual-stress state of “as fabricated” springs in three conditions were thus evaluated: springs manufactured prior to failure incidence increase, 12 min peen; springs manufactured following failure incidence increase, 12 min peen; and 60 min peen.

The residual-stress measurement results, as seen in Fig. 18 indicate that something had indeed changed, namely the effective depth and magnitude of compressive residual stresses in the shot-peened springs when comparing before and after the increase in failure incidence. The cause of this change may have been either material or process related. Additionally, increasing the peening time from 12 to 60 min significantly increased the compressive residual-stress levels in the springs. In fact, the springs that were shot peened for 60 min were found to fail at much more than 500,000 cycles in fatigue tests. Thus, in this example, the suspicion that inadequate compressive residual stresses were imparted by the peening process was validated, and the validity and effect of the proposed corrective action was substantiated. In this case, an increased peening time resulted in an increased percent coverage and effective compressive stress level. This example demonstrates that once the source of the failure is understood and validated by experiment, corrective measures can be applied, verified, and subsequently monitored with confidence. This example also demonstrates that XRD can be used to explore quality-related issues. The residual-stress benchmarks established in this case study can be referenced by the manufacturer or the customer in future residual-stress measurement audits and/or incorporated into the blueprints for this component.

![Stress versus depth profiles for different steel coil springs](image)

**Fig. 18 Stress versus depth profiles for different steel coil springs**

Example 2: Use of XRD to Assess the Effect of Mechanical Loading on Stress Relaxation of Machined and Shot-Peened Nickel-Base Alloy (Waspaloy). Experiments were conducted on rotating beam specimens at different loading levels. The specimen conditions are defined as shot peened and not shot peened/polished. Initial surface residual stresses were
determined using XRD techniques, and the specimens were thus divided into three groups: first level, second level, and third level (shot peened), where the third level was the most compressive. Prior to failure, multiple cracks were observed at 500× magnification adjacent to the macroscopically visible main crack, indicating that damage was not localized in the middle of the gage section, that is, evidence of LCF (Fig. 19).

![Fig. 19 Observation of failed nickel-base alloy (Waspaloy) specimen after rotating bend fatigue. (a) Macro view. (b) Micrograph. Source: Ref 43](image)

$S$-$N$ curves for all three stress levels can be seen in Fig. 20. Level 1 and level 2 curves are very similar; thus, the difference is negligible, according to the physical mechanism of failure. The number of cycles to failure for shot-peened specimens is longer at low applied bending moment and shorter at higher applied bending moment, as for non-shot-peened specimens. This can be explained by the effect of two parameters: the level of the compressive residual stress and the surface roughness. The roughness of the surface due to peening can act as a stress raiser and may diminish the fatigue life, but the high compressive stresses can counterbalance the effect of the roughness (Ref 47, 48). In addition, the benefit of shot peening at higher applied fatigue loads may be diminished due to the surface roughness and the potentially steeper evolution (i.e., lack of stability) of residual stresses when the yield strength of ductile materials, such as Waspaloy, are exceeded.

![Fig. 20 Applied bending load and number of cycles to failure, $R = -1$, for Waspaloy specimen. Source: Ref 43](image)

Residual stress relaxation is a consequence of micro- and/or macroplastic deformations and is important, whether the superposition of residual, loading, or mean stresses exceeds the monotomic or cyclic yield strength. The fatigue limit of high-strength materials is generally lower than the cyclic yield strength, whereas in low-strength materials, these values may nearly coincide. This explains why residual stresses do not relax in high-strength materials subjected to load amplitudes in the range of the fatigue limit (Ref 2).

In the case of materials such as Waspaloy, the evolution of residual stresses with LCF can be used as a mechanism for tracking the useful fatigue life remaining in the component. It has been shown, in some cases, that a linear relationship holds between the residual-stress relaxation rate and the residual stress amount of the starting condition (Ref 43, 45, 46, 49, 50).
Example 3: Using XRD to Track Residual-Stress Levels at Critical Locations on Turbine Engine Disks as They Accumulate Engine Cycles. The results shown in Fig. 21 indicate that there is a continual degradation of residual compressive stress with an increase in operational engine cycles. Based on the sensitivity of fatigue life with stress, the degradation of compressive residual stress from cold-working processes is an indication of increased “active” stress and could, therefore, be used as a measure of remaining fatigue life. For this to be realized, a trend must exist whereby the compressive residual stress, from shot peening or other cold-working processes, relaxes to a value below which the disk is at increased risk to crack. In other words, it is assumed that residual-stress degradation is a precursor to fatigue crack initiation in LCF mode where loads exceed the yield strength of the material (Ref 45). (The implications of the comparison in Fig. 21 of the “large spread—process not controlled” versus the “controlled process reduces spread” are discussed in the section “Sample Selection” in this article.)

![Residual stress data plots](image)

**Fig. 21** Theoretical model versus XRD-generated experimental data plots of residual stress versus number of cycles. Source: Ref 43

Residual Stress and HCF. High-cycle fatigue may occur when the cyclic load range does not exceed the cyclic yield strength of the material but exceeds the fatigue limit (endurance limit) of the material (although many materials, such as aluminum, do not have well-defined fatigue limits). In components where no initial stresses are present, stresses may develop, and, if the stresses that develop are compressive, they may actually add to the life of the component. However, if stresses are induced by processing, they may “fade” at a rate that increases with cycling stress (Fig. 22) (Ref 51).
Other Considerations Regarding Residual Stress and Fatigue. There are cases where the influence of residual stress on fatigue life is negligible. For hardened and ground specimens, a clear positive influence of compressive stresses and a negative influence of tensile stresses are often exhibited. Compressive residual stresses impart the maximum benefit when the volumes with the highest compressive residual stress coincide with the highest loaded component volumes. This explains why the influence of near-surface residual stresses is more pronounced for bending fatigue than in the case of tension-compression loading. For example, in the case of a normalized material with thin-reaching machining residual stresses, they may have very little impact on the resulting tension-compression loading S-N curves. Positive residual-stress effects are also more pronounced if detrimental starting conditions, such as decarburization or oxidation, exist in the case of hardened steels. Thus, the stability of residual stresses is of central importance; that is, their effect is more pronounced when less relaxation occurs during fatigue loading (Ref 2).

References cited in this section


45. S.G. Berkley, U.S. Patents 5,490,195 and 5,625,664
X-Ray Diffraction Residual Stress Measurement in Failure Analysis

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

The Effect of Manufacturing Processes on Residual Stress

Because of the enormous impact of manufacturing processes and loading history on resulting residual-stress states, they have to be taken into account in the case of failure analysis (Ref 2, 8). It is assumed, at least in principle, that in-service loads for a given component can be determined either by direct calculation, measurement, or finite-element modeling. X-ray diffraction can be used to measure the residual stress in as-manufactured components so as to enable the detection of stress states that may be potentially harmful to service life and to subsequently identify unfavorable process parameters. For example, consider the effects of what are identified in Fig. 23 (Ref 3) as gentle, conventional, and abusive grinding. It is important to note, in the case of abusive grinding (where grinder burn is present), that the surface residual stress may be neutral (as in this example) or even compressive; however, the life-limiting tensile residual stresses may be at very shallow depths below the surface. Thus, failure analysts must consider surface and subsurface residual stresses in their analysis.
Fig. 23 Subsurface residual-stress distribution after grinding hardened steel (stress measured in the direction of grinding). Source: Ref 3

References cited in this section


---

**X-Ray Diffraction Residual Stress Measurement in Failure Analysis**

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

---

**The Characterization of Stress Gradients Using XRD**

The XRD technique is generally applied to the measurement of residual stress in polycrystalline materials used in mechanical, electronic, and structural components. After machining, casting, shot peening, turning, heat treatment, and so on, the residual stress can vary significantly from the surface through the subsurface (Ref 2). These residual stresses must therefore be characterized at the surface and through the subsurface to evaluate the effects of the process and the stress gradients generated (Ref 52, 53). Changes in the process can then be evaluated via changes in the surface and subsurface stress state. To access subsurface measurement locations, material removal is required and can be performed using electro-polishing techniques. This technique is discussed in more detail in the section “Specimen Preparation” in this
article. Stress relaxation due to material removal and stress gradient corrections can subsequently be applied to stress versus depth results (Ref 3). Residual-stress characterization in both the surface and subsurface is a powerful and essential tool in failure analysis, process control, and optimization (Ref 6, 52).

The subsurface stress gradients shown in Fig. 24 reflect the effect of varying peening intensity on shot-peened components from the same population. The compressive residual stress is low at the surface for the unpeened sample and high at the surface for the peened ones. Normally, samples such as these are subsequently exposed to fatigue testing in order to select the optimal process. This information can then be used to establish a benchmark for quality control and quality assurance. In cases where the process is adequately controlled, the surface stresses can be used as an indication of the presence and/or effectiveness of the shot peening in line with the subsurface stress profiles confirmed via periodic (hourly/daily/weekly) audits. Indirect information about the effect of the loading history on the residual-stress state of components can also be obtained after they have been placed into service.

![Fig. 24 Residual-stress profiles on unpeened and peened samples](image)

**Fig. 24** Residual-stress profiles on unpeened and peened samples

Effects of In-Service Loads on Residual Stress. The residual stress versus depth plots in Fig. 25 indicate that the in-service loads have a profound effect on the near-surface stress state of the material in the gear. The service cycling introduced an increased compressive near-surface residual stress layer in the region of the pitch diameter. This compressive residual stress as a result of in-service cycling may, in fact, increase the resistance to fatigue of the gear at this location. This effect is, of course, localized and does not address the stress state of other locations, such as the root of the gear.

![Fig. 25 Comparison of the residual stress on the tooth pitch diameter found in two different types of hardened steel gears in new and used conditions using XRD. (a) Pinion gears. (b) Sun gears](image)

**Fig. 25** Comparison of the residual stress on the tooth pitch diameter found in two different types of hardened steel gears in new and used conditions using XRD. (a) Pinion gears. (b) Sun gears

The residual stress versus depth plot in Fig. 26 indicates that the in-service loads have the effect of reducing the compressive near-surface stress state of the material in the spring. In this case, the service cycling introduced probably decreases the resistance of the spring to fatigue failure.
Fig. 26 Comparison of the residual stress on the inner diameter of shot-peened coil springs in new and used conditions using XRD

The examples shown in Fig. 25 and 26 illustrate the potential effects of cold working on manufactured components prior to and subsequent to being placed into service.

References cited in this section


X-Ray Diffraction Residual Stress Measurement in Failure Analysis

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

Effects of Heat Treatment on Residual Stresses

Heat treatment processes are also commonly applied to engineering components to obtain the desired microstructure, such as in the case of quenching and/or tempering. Heat treatments may also be applied that have as their purpose the effect of relaxing micro- and/or macroresidual stresses. X-ray diffraction can be used effectively in the characterization of these processes as well.

Example 4: Use of XRD to Assess the Effects of Heat Treatment on Residual Stress in Steel Coil Springs. Shot-peened steel coil springs were heat treated for 45 min through a range of temperatures to observe stress relaxation effects with various tempering temperatures. The surface residual macrostress was measured prior to and subsequent to heat treatment. The purpose of measuring residual stresses on the same springs before and after heat treatment was to minimize the effect of potential variations in the as-manufactured residual-stress state. The data plotted in Fig. 27 were obtained.
Fig. 27 Plot of the change in the compressive residual stress due to heat treatment

One coil spring was kept as a control, and the reduction in compressive residual stress due to heat treatment was plotted for the remaining springs that were subject to various heat treatment temperatures. A trend of increasingly relaxed residual stress was observed for increased heat treatment temperature.

Similar studies can be conducted on a variety of components to assess the change in residual macrostress due to heat treatment. The effect may be a positive influence, if life-limiting (undesirable) residual stresses are relaxed. Conversely, the impact of life-extending (desirable) residual stresses may be reduced.

Example 5: Use of XRD to Evaluate the Effect of Varying Heat Treatment Temperature on Residual Stress for Iron Alloys. As this example demonstrates, XRD can also be applied to the evaluation of both micro- and macrostress simultaneously. Trends in the relaxation of macrostresses are evaluated using the associated shifts in the atomic lattice spacing and thus the diffraction angle, whereas the distribution of microstresses can be correlated to the XRD peak width.

The residual stress with varying heat treatment temperature for two iron alloys is plotted in Fig. 28. The magnitude (or absolute value) of the residual stress for these specimens tends to decrease for a given increase in heat treatment temperature. The effect is pronounced for the AISI 01 samples, because they had a significant residual stress prior to heat treatment. The residual stresses in the AISI 1070 samples were lower, on average, prior to heat treatment; thus, there were lower-magnitude macrostresses to relax. The scatter in these data (beyond the quoted error bars) can be attributed to the slightly different residual-stress state prior to heat treatment in the various samples used to represent each heat treatment temperature.

Fig. 28 X-ray diffraction residual stress versus heat treatment temperature for various iron alloys. Specimens were held at temperature for 1 h and furnace cooled.
Dislocations are typically introduced into materials by cold working. The formation of small crystallites (coherent domains) and the introduction of elastic microstrains both result in broadening of the XRD line profile (Ref 7). Line broadening is also a function of instrumental broadening, that is, the XRD instrument optics. Despite trends in macrostress (Fig. 28), a trend of decreased hardness with increased heat treatment temperature is often observed in iron alloys sample sets (Fig. 29a). An associated decrease in the XRD peak breadth is also often observed (Fig. 29b). In this example, the range of heat treatment temperatures remained below the annealing temperature for each alloy; thus the instrumental and particle size contributions to the broadening remained more or less constant. If the heat treatment temperature exceeds the annealing temperature, the observed XRD peak breadth may be reduced either by recovery, recrystallization, or grain growth (Ref 6). Heat treatment with slow cooling has the effect of relieving or reducing microstresses as well as decreasing the dislocation density, thus decreasing the hardness.

![Fig. 29 Effect of heat treatment temperature on (a) hardness (HRC) and (b) XRD peak integral breadth.](image)

The effects of these mechanisms can thus be observed in the XRD peak breadth. In general, this is used as a trending tool, and the data are most easily used in empirical form to assess the relative hardness, cold working, or dislocation density in a given material.

More detailed analyses of the source of XRD line broadening can be performed using Fourier space methods, such as the Warren-Averbach method (Ref 54), or real space methods, such as Voigt deconvolution (Ref 55). The XRD peak integral breadth and/or full width at half maximum is generally characterized as a function of depth with the use of electropolishing techniques in parallel with the measurement of residual macrostresses.

References cited in this section


X-Ray Diffraction Stress Measurements in Multiphase Materials and Composites

In many materials, the presence of more than one phase is not uncommon. Materials that are subject to heat treatment and/or have alloying elements may be composed of two, three, or more phases. For composite materials, the failure analyst should consider all constituent phases. The diffraction conditions and measurement parameters may differ for each phase, however, the additional effort is often worthwhile for a complete analysis.

For example, carbon steel containing over 0.5% C and other elements added to increase toughness, improve corrosion resistance, and so on can contain ferrite (and/or martensite), austenite, cementite, and/or other phases. The volume fraction of each phase can vary with alloying content, heat treatment, and even cold working; therefore, the stresses have to be evaluated for all phases of significant volume fraction. The residual stress in each phase and the weighted average stress (based on the volume fraction of each phase) should be calculated, because the fracture resistance of the material may vary locally, depending on the phase under consideration. For example, untempered martensite is more brittle than austenite; thus, when both phases are present in a component, each may behave quite differently during cyclic loading. The residual stress in all phases present in the material should thus be measured and analyzed.

X-Ray Diffraction Stress Measurements in Locations of Stress Concentration

Stress concentrations and stress raisers can be important contributors to the initiation and/or propagation of cracks and subsequent component or structural failure. Stress concentrations are ubiquitous, found in locations associated with geometric discontinuities such as sharp corners, weld “toes,” notches, undercuts, machining marks, scratches, microcracks, and so on, and naturally, crack tips themselves. The degree of care taken in the estimation, modeling, and ultimate management of the various stress concentration contributors in a given structure or component should be dependent on the criticality of the component in the structure.

Typically, stress concentration factors (SCFs) are either found in publications such as Peterson's Stress Concentration Factors (Ref 56) or determined via computer-based modeling using various techniques. The application of SCFs obtained using either of these two approaches may involve assumptions regarding component geometry, the SCF uniformity, and, in some cases, estimates of the residual stresses present. Experimental determination of SCFs using XRD is often possible for specific geometric features by performing stress measurements at concentration locations under an incremental loading regime (Ref 12, 31). Data can also be collected to determine the variation of the SCF due to inhomogeneity of the stress concentration geometry, that is, where the radius of curvature is nonuniform.

It is important to measure residual stresses local and adjacent to stress concentrations using XRD, because residual stresses are additive to applied stresses and thus influence the ultimate effect of the total stress present in the region of stress concentration (Ref 57). It has been shown that surface residual stresses can significantly affect crack opening behavior (thus, crack growth behavior) even when the crack has grown well beyond the zone of residual stress (Ref 58). This indicates that the residual compressive stress present at the surface can strongly influence the surface crack tip opening displacement throughout the entire loading cycle (Ref 58).

Example 6: Use of XRD to Evaluate Cracks in Steel Cargo Tiedown Sockets. Cracks initiating from the tip of the cloverleaf pattern in steel cargo tiedown sockets were observed by the builder following installation aboard several cargo vessels in various stages of construction. A root-cause failure analysis was launched, and three possible mechanisms of failure were considered: overload failure, fatigue fracture, and environmentally assisted cracking (stress-corrosion cracking, SCC). Failure mechanisms of fatigue fracture and environmentally assisted cracking were eliminated, because no evidence of SCC or fatigue fracture was observed; thus, the failure analysts focused on the possibility of overload
failure. Further tests showed that the overload failure mode and the transition from ductile to brittle fracture were facilitated by the combination of high brittleness of a carbon-rich transformed martensite layer introduced by flame cutting, increased hardness due to the cold-working coining process, and high residual stresses created by the welding process (Ref 59).

Finite-element models (Fig. 30) were used to predict residual stresses in the areas of high stress concentration introduced during the manufacturing (welding) process and additional applied stresses introduced by the “sagging” loads inherent in the ship hull construction and ballast distribution. The high SCF thus had the effect of locally amplifying the manufacturing-induced welding residual stresses and the in-service dead load stresses.

![Finite-element model showing maximum stress concentration in cloverleaf radius. The highest stress concentration is in the small black area surrounded by white.](image1)

To validate the “overload failure” postulate, measurements were performed in the field on cargo ships with the cracking problem. A typical residual-stress profile collected on the tiedown sockets installed on ships can be seen in Fig. 31.

![Typical residual-stress profile as a function of distance from the maximum stress concentration in the radius of a tiedown socket](image2)

X-ray diffraction was thus used to validate the postulated failure mechanism and effectively demonstrated the existence of residual weld-induced stresses in installed tiedown sockets (Ref 60). Subsequently, failure analysts were able to recommend corrective measures: the removal of the brittle, carbon-rich transformed martensite layer introduced by flame cutting, which facilitated the overload failure; and the application of a localized stress-relief heat treatment process. X-ray diffraction residual-stress measurements were then performed on heat treated tiedown sockets to verify the effectiveness of the localized heat treatment process applied.

**References cited in this section**
X-Ray Diffraction Residual Stress Measurement in Failure Analysis

J.A. Pineault, M. Belassel, and M.E. Brauss, Proto Manufacturing Ltd.

Acknowledgments

The authors would like to thank Debbie Aliya, Clayton O. Ruud, Stan Berkley, Robert Drake, and Fred Tonch for their helpful insight and careful review of the ideas presented in this article.

References


2. V. Hauk, Structural and Residual Stress Analysis by Nondestructive Methods, Elsevier, 1997


45. S.G. Berkley, U.S. Patents 5,490,195 and 5,625,664


47. C.O. Nonga et al., The Influence of Residual Stresses and Surface Finish on the Fatigue of Metal Matrix Composites, Proceedings of ICRS5 (Linkoping), 1997, p 95–100


50. G. Kuhn, Dr. Ing. thesis, University Karlsruhe (TH), 1984


54. B.E. Warren, X-Ray Diffraction, Addison-Wesley, 1969


59. X.J. Zhang and M. Gaudett, Steel Failure Analysis Results in Stronger Deck Sockets, Wavelengths—An Employee Digest of Events and Issues, Jan/Feb 1999

60. M. Zoccola, Division Trio Receive Navy Superior Service Award, Wavelengths—An Employee Digest of Events and Issues, Aug 1999

Metallographic Techniques in Failure Analysis

George F. Vander Voort, Buehler Ltd.

Introduction

METALLOGRAPHIC EXAMINATION is one of the most important procedures used by metallurgists in failure analysis. Development of powerful electron metallographic instruments, such as the scanning electron microscope, has not diminished the importance of light microscopy. Basically, the light microscope is used to assess the nature of the microstructure and its influence on the failure mechanism. The purpose in using the light microscope may be twofold. One purpose may be to determine the relationship between the microstructure and the crack path (in failures involving fracture) and/or the nature of corrosion or wear damage. The second purpose is to determine whether processing or service conditions have produced undesirable microstructural conditions that have contributed to the failure, such as abnormalities due to material quality, fabrication, heat treatment, and service conditions. Examples are given in this article to demonstrate such analytical work.

Conducting a materials failure analysis, a common activity for many metallurgists, requires a carefully planned series of steps (Ref 1, 2) designed to arrive at the cause of the problem. Proper implementation of light microscopy is of critical importance in failure analysis, and this article focuses on the use of metallographic techniques and examinations using the light microscope (LM) in failure analysis. Metallographic examination typically should follow nondestructive and macroscopic examination procedures and should precede use of techniques of electron microscopy.

Examination of fractured components should begin with the low-power stereomicroscope. Hand-held magnifying lenses are still widely used to study fractures but mainly in the field. While the light microscope has limited value for direct observation of fracture surfaces, a great deal can be learned by indirect examination, that is, by examination of the fracture profile and secondary cracking.

Detailed observation of the fracture surface is best accomplished by use of the scanning electron microscope (SEM) or by examination of replicas with the transmission electron microscope (TEM). However, lack of access to a SEM or TEM should not be viewed as a crippling obstacle to performing failure analysis, because such work was done successfully prior to the development of these instruments. In many studies, such equipment is not needed, while in other cases, they are very important tools. In most cases, a more thorough job can be accomplished using such tools.

The LM, on the other hand, is a virtually indispensable tool for the failure analyst. The SEM and TEM find application in microscopy when the required magnification/resolution exceed the capabilities of the LM (about 1000x), but the first tool of choice is the LM for microstructural examination and identification. Electron microscopy and LM are complementary tools. Microstructural examination can be performed with the SEM over the same magnification range as the LM, but examination with the latter is more efficient. Contrast mechanisms for viewing microstructures are different for LM and SEM. Many microstructures, for example, tempered martensite, exhibit poor contrast in the SEM and are best viewed by light microscopy. When atomic number contrast or topographic contrast is strong, the SEM provides good structural images, particularly above 500x (Ref 3). Again, because of the limitations and advantages of each instrument, they are complementary rather than competitive tools. All studies of microstructures and fractures should begin at the lowest magnification level, the unaided human eye, and progress upward, first using the stereomicroscope for fractures and the LM for fracture path and microstructural studies, before using electron metallographic equipment.

References cited in this section


Examination of Fractures

Microfractography is a relatively new field; its roots can be traced to the light optical fractographs published by Zapffe and coworkers (Ref 4) beginning in the early 1940s, although a few studies of historical value predated their efforts. Zapffe's work, however, was almost exclusively confined to observation of cleavage facets on rather brittle, coarse-grained specimens. The technique, basically an interesting academic exercise, did stimulate interest in fracture examination as part of failure analysis. However, the depth-of-field limitation of the light microscope has restricted its use for such work. Aside from the published light optical fractographs made by Zapffe (see Ref 5 for a review of many of these), very few optical fractographs of metallic materials have been published by others. Microfractography gained momentum with the development of TEM replication methods and became commonplace after the commercial introduction of the SEM in approximately 1965.

A flat, brittle fracture can be examined with the light microscope by orienting the fracture perpendicularly to the optical axis. It is best to start with a low-power objective; long-working distance types are preferred. Focusing reveals the limitations of the method, because only part of the fracture is in focus at any setting. Thus, photographs reveal only a portion of the fracture in focus, depending on the coarseness and orientation of the fracture facets. Figure 1 shows an example of a brittle fracture in a low-carbon steel examined in this manner. Figure 1 also shows a LM image of the fracture profile, a LM image of a replica of the fracture, SEM images of the fracture and a replica of the fracture, and a TEM replica of the fracture. Although Zapffe used bright-field illumination for this work, dark-field illumination often produces superior results. Figure 2 illustrates the use of bright-field and dark-field illumination for viewing a brittle fracture in an Fe-Cr-Al alloy, plus a SEM fractograph of the same area. Dark-field illumination is better at collecting the light scattered from the fracture features; glare is reduced, and image contrast is improved.

![Fig. 1 Illustration of a cleavage fracture in a quenched and tempered low-carbon steel examined using three direct methods and three replication methods. (a) LM cross section](image-url)
(nickel plated). Etched with Vilella's reagent. (b) LM fractograph (direct). (c) SEM fractograph (direct). (d) LM replica. (e) SEM replica. (f) TEM replica

Fig. 2 Light microscope fractographs taken with (a) bright-field and (b) dark-field illumination compared to (c) a SEM secondary-electron image fractograph of the same area. Sample is an Fe-Al-Cr alloy.

Examination of fracture replicas with the light microscope (Ref 5, 6, 7) can extend the use of the method only to a limited extent, because the replica collapses slightly, producing less depth of field. Also, with a replica, the risk of damaging the objective is eliminated.

Considerable information concerning the fracture mode and the relationship of the microstructure to the fracture path can be obtained by LM examination of the profile of partially fractured (Ref 8, 9, 10) or completely fractured (Ref 11, 12, 13, 14, 15) polished metallographic specimens. Such examinations have been conducted for many years, long before the development of electron metallographic techniques, and continue to be used because of the value of the method. If the fracture has progressed to complete rupture, so that only one side of the fracture is to be examined, it may be best to nickel plate the fracture to enhance edge retention. This is not required if the crack has not separated the component into two pieces, or if a secondary crack is to be examined.

Figure 3(b) shows a nickel-plated brittle impact fracture of a low-carbon steel with a ferrite-pearlite microstructure. The crack consists of numerous connected straight-line segments in the ferrite phase. Several subsurface cleavage cracks are also present. In comparison, a ductile impact fracture in a quenched and tempered carbon plate steel is shown in Fig. 3(a). Note that the fracture surface exhibits a much rougher appearance due to the linking up of the microvoids. Below the fracture surface, spherical ruptures are frequently seen, which are also indicative of a ductile fracture mechanism.

Fig. 3 Light micrographs of section profiles of (a) a nickel-plated ductile fracture and (b) a nickel-plated brittle fracture. Both are carbon steels etched with 2% nital.

Fatigue fractures can also be examined using fracture profiles, as illustrated in Fig. 4. This shows a low-cycle fatigue crack that has not propagated to final rupture. Note that the crack exhibits some minor branching and does not traverse the ferrite grains in a straight-line fashion, as with cleavage (Fig. 3b). Also, there are no spherical cavities near the crack, as observed with ductile fractures (Fig. 3a). The crack shows no preference for either the ferrite or pearlite during its growth, and the macroscopic appearance of the crack is rather flat, which is typical of fatigue cracks.
Fig. 4 Light micrograph of the path of a fatigue crack through a low-carbon steel specimen. Etched with 2% nital

Figure 5 shows partially broken and completely broken sensitized (649 °C, or 1200 °F, for 4 h) impact specimens of American Iron and Steel Institute (AISI) 304 stainless steel. Scanning electron microscope examination of the fracture face reveals extensive microvoid coalescence, that is, ductile rupture, although the impact strength (at -196 °C, or -320
°F) was only 40% of that of a nonsensitized sample. The partially broken sample reveals a fracture path that often follows the grain boundaries. Rupture cavities are observed behind and ahead of the crack; most are associated with coarse carbides. The fracture profile of the completely broken specimen reveals microvoid coalescence.

Fig. 5 Light micrographs of a partially broken and a completely broken specimen of sensitized (649 °C, or 1200 °F, for 4 h) AISI 304 austenitic stainless steel, and a SEM fractograph of the broken Charpy V-notch specimen

As a comparison to Fig. 5, Fig. 6 shows a similar fracture made in AISI 312 stainless steel weld metal that was aged at 816 °C (1500 °F) to transform the delta ferrite to sigma phase. In this case, the impact strength at room temperature was only 7% of that of an as-welded sample containing austenite and delta ferrite. Note that the fracture path is microscopically flatter and consists of numerous connected short straight-line segments. The SEM fractograph shows numerous small, flat fracture regions indicative of the embrittlement due to sigma. The high-magnification LM profile view of the completely fractured surface reveals the presence of extensive sigma along the crack path.
Fig. 6  Light micrographs of a partially broken and a completely broken specimen of AISI 312 stainless steel weld metal heat treated to transform the delta ferrite to sigma, and a SEM fractograph of the broken Charpy V-notch specimen

Examination of the crack path using cross sections is also very useful for study of fractures due to environmental problems. Figure 7 shows a stress-corrosion crack in a partially broken sample of solution-annealed AISI 304 stainless steel tested in boiling (151 °C, or 304 °F) magnesium chloride. The crack path is predominantly intergranular, but considerable transgranular fracture is also present. The SEM fractograph of the specimen clearly reveals the intergranular nature of the crack.
Fracture profile examination is also very useful in the study of failures due to liquid metal embrittlement (LME). Figure 8 shows the microstructure adjacent to a LME crack in a eutectoid steel where liquid copper has penetrated the grain boundaries at 1100 °C (2012 °F) while the sample was austenitic and under an applied tensile load. Light microscope examination reveals a discontinuous film of copper in the prior-austenite grain boundaries and an intergranular fracture path. Scanning electron microscope examination of the fracture also reveals the intergranular nature of the crack path.

Surface detail can also be studied by LM using taper sections (Ref 16, 17, 18). This method has been used to study wear phenomena, surface coatings, fatigue damage, and other fine surface detail. In this method, the surface is sectioned at a slight angle to the surface. Polishing on this plane produces a magnified view of the structure in the vertical direction. The degree of magnification is defined by the cosecant of the sectioning angle; an angle of 5° 43' produces a tenfold magnification.
Considerable progress has been made in applying the principles of quantitative metallography to the study of fractures (Ref 14, 15, 19, 20, 21, 22, 23). Much of this work has used measurements made on polished sections taken parallel to the crack-growth direction. This work provides new insight into fracture processes and should be useful in failure analysis, although its application to date has been limited mainly to research studies.

References cited in this section

Metallographic Techniques in Failure Analysis

George F. Vander Voort, Buehler Ltd.

Metallographic Specimen Preparation

Because the metallographer cannot predict in advance what the microstructural examination will reveal, specimen preparation must be perfect; otherwise, critical information can easily be lost. This basic truth has been proven over and over, yet is violated regularly. The preparation procedure and the prepared specimen must provide the following:

- Deformation induced by sectioning, grinding, and polishing must be removed or be shallow enough to be removed by the etchant.
- Coarse grinding scratches must be removed; even very fine polishing scratches may not be tolerable in examining failed parts.
- Pullout, pitting, cracking of hard particles, smear, and other preparation artifacts must be avoided.
- Relief (i.e., excessive surface-height variations between structural features of different hardness) must be minimized; otherwise, portions of the image are out of focus at high magnifications. Excessive relief invalidates image analysis measurements and is undesirable for wavelength-dispersive chemical analysis.
- The surface must be flat, particularly at edges (if they are of interest), or they cannot be examined. Edge preservation is of critical importance in failure studies, because many failures start at external surfaces.
- Coated or plated surfaces must be kept flat if they are to be examined, analyzed, measured, or photographed.
- An etchant should be used that reveals all of the structure at first. Later, it may be useful to use a selective etchant that reveals only the phase or constituent of interest, or at least produces strong contrast or color differences between two or more phases present, to improve the precision of microstructural measurements or to better reveal the relative presence of undesirable constituents or phases.

If these characteristics are met, then the true structure is revealed and can be interpreted, measured, analyzed, and recorded. The preparation method should be as simple as possible, should yield consistent, high-quality results in a minimum of time and cost, and must be reproducible.

Preparation of metallographic specimens (Ref 24) generally requires five major operations: sectioning, mounting (optional), grinding, polishing, and etching (optional).

Sectioning

It is certainly not uncommon in failure analysis to encounter large specimens. Indeed, the initial sectioning operation may be quite a challenge. Bulk samples for subsequent laboratory sectioning may be removed from larger pieces using methods such as core drilling, band- or hacksawing, flame cutting, or similar methods. Flame or torch cutting may be the only recourse in the field. If this is done, the torch-cut area must be well away from the area to be examined, because the heat from this operation severely alters the original microstructure for some distance from the cut. Subsequent cutting can be performed with laboratory devices that are much less damaging to the structure. Laboratory abrasive-wheel cutting is recommended to establish the desired plane of polish.

The most commonly used sectioning device in the metallographic laboratory is the abrasive cutoff machine. All abrasive-wheel sectioning should be performed wet. An ample flow of coolant, with an additive for corrosion protection and lubrication, should be directed uniformly into the cut. Wet cutting produces a smooth surface finish and, most importantly, guards against excessive surface damage caused by overheating. Figures 9(a) and (b) show the surface of quenched and tempered A2 tool steel (59 HRC) cut without using coolant. The cut surface was nickel plated for edge preservation. Figure 9(a) shows a light-etching surface zone extending to a depth of approximately 0.22 mm (0.009 in.), with a hardness of approximately 62.5 HRC. Beneath the light-etching surface zone is a region that was softer (53 to 56 HRC) and etches darker. The unaffected matrix is beneath this zone, off the edge of the micrograph. Figure 9(b) shows the extreme surface at high magnification. Incipient melting can be observed to a depth of approximately 10 μm. The light-etching zone contains untempered martensite, because the temperature in this region was high enough to reaustenitize the structure. The dark-etching zone beneath it saw temperatures below the lower critical but greater than the original tempering temperature. Abrasive wheels should be selected according to the manufacturer’s recommendations.
Wheels consist of abrasive particles, chiefly alumina or silicon carbide (SiC), and filler in a binder material that may be a resin, rubber, or a mixture of resin and rubber. Alumina is the preferred abrasive for ferrous alloys, and SiC is the preferred abrasive for nonferrous metals and minerals. Wheels have different bond strengths and are recommended based on the suitability of their bond strength and abrasive type for the material to be sectioned. In general, as the hardness of a material increases, abrasives become dull more quickly, and the binder must be adjusted to release the abrasives when they become dull, so that fresh abrasive particles are available to maintain cutting speed and efficiency. Consequently, these wheels are called “consumable” wheels, because they wear away with use. If they do not wear at the proper rate, dull abrasives rub against the region being cut, generating heat and altering the existing true microstructure. If this heat becomes excessive, it can lead to grain or particle coarsening, softening or phase transformations, and, in extreme cases, burning or melting. Different materials have different sensitivities to this problem, but the need to balance the wheel break-down rate with the hardness of the piece being sectioned produces the various recommendations listed for cutting different materials and metals at different hardnesses, such as steels.

Precision saws are commonly used in metallographic preparation to section materials that are small, delicate, friable, extremely hard, or where the cut must be made as close as possible to a feature of interest, or where the cut width and material loss must be minimal. As the name implies, this type of saw is designed to make very precise cuts. They are smaller in size than the usual laboratory abrasive cutoff saw and use much smaller blades, typically from 7.6 to 20.3 cm (3 to 8 in.) in diameter. These blades can be of the nonconsumable type, made of copper-base alloys with diamond or cubic boron nitride abrasive bonded to the periphery of the blade, or they can be consumable blades using alumina or SiC abrasives with a rubber-based bond. Blades for the precision saws are much thinner than the abrasive wheels used in an abrasive cutter, and the load applied during cutting is much less. Consequently, less heat is generated during cutting, and damage depths are reduced. While pieces with a small section size that would normally be sectioned with an abrasive cutter can be cut with a precision saw, the cutting time is appreciably greater, but the depth of damage is much less. Precision saws are widely used for sectioning sintered carbides, ceramic materials, thermally sprayed coatings, printed circuit boards, electronic components, bone, teeth, and so on.

Mounting

The primary purpose of mounting metallographic specimens is for convenience in handling specimens of difficult shapes or sizes during the subsequent steps of metallographic preparation and examination. A secondary purpose is to protect and preserve extreme edges or surface defects during metallographic preparation. The method of mounting should in no way be injurious to the microstructure of the specimen. Pressure and heat are the most likely sources of injurious effects. The most common mounting method uses a device, called a mounting press, to provide the required pressure and heat to encapsulate the specimen with a thermosetting or thermoplastic mounting material. Common thermosetting resins include phenolic, diallyl phthalate, and epoxy, while methyl methacrylate is the most commonly used thermoplastic mounting resin. Both thermosetting and thermoplastic materials require heat and pressure during the molding cycle, but, after curing, mounts made of thermoplastic resins must be cooled under pressure to at least 70 °C (158 °F), while mounts made of thermosetting materials may be ejected from the mold at the maximum molding temperature. However, cooling
thermosetting resins under pressure to near-ambient temperature before ejection significantly reduces shrinkage gap formation. Never rapidly cool a thermosetting resin mount with water after hot ejection from the molding temperature. This causes the metal to pull away from the resin, producing shrinkage gaps that promote poor edge retention (Fig. 10a, b) because of the different rates of thermal contraction.

Fig. 10 Light micrographs of the surface of a carburized 8620 alloy steel specimen mounted in phenolic resin. Note the shrinkage gap (see arrows in a) that has reduced the edge flatness. In (b), taken at 1000×, decarburization at the surface has caused ferrite and pearlite to form, and this area is slightly out of focus. Specimen etched with nital

A thermosetting epoxy with special filler (Epomet, Buehler Ltd.) provides superior edge retention (Fig. 11a) compared to other resins (see Fig. 11(b) for a phenolic mount and Fig. 11(c) for methyl methacrylate) and is virtually unaffected by hot or boiling etchants, while phenolic resins are badly damaged. The thermoplastic resins, such as methyl methacrylate, produce a transparent mount, which is helpful when trying to grind to a specific feature, but provide poor edge retention (Fig. 11c). Electroless nickel plating is an effective method for improving edge retention, particularly for steels. However, if the area to be studied is white or with low contrast, it may be difficult to determine where the nickel plating ends and the surface begins, as shown in Fig. 11(d). Figure 12 shows an example of ion-nitrided hot work die steel with a brittle white-etching iron nitride surface layer that is quite visible when mounted in Epomet resin but would probably be very hard to detect if the surface had been plated with nickel.
Fig. 11  Light micrographs of specimens of 1215 carbon steel that were salt bath nitrided and mounted in different resins. (a) Epomet thermosetting epoxy resin. (b) Phenolic thermosetting resin. (c) Methyl methacrylate thermoplastic resin. (d) Electroless nickel plated and mounted Epomet in epoxy resin (resin not in the field of view). All four specimens were prepared in the same holder and were etched with nital. The arrows point to the nitrided surface layer.

Fig. 12  Light micrograph of an ion-nitrided H13 tool steel specimen mounted in epoxy thermosetting resin (Epomet). The arrows point to a white-etching iron nitride layer at
the surface that probably would not have been observed if the specimen was nickel plated for edge protection. Specimen etched with nital

Shrinkage gaps between specimen and mount are a prime cause of loss of edge retention, as discussed subsequently. Besides this, abrasives can become lodged in the gap and fall out, causing contamination problems in a subsequent step. Further, liquids can seep out of the gaps, despite best efforts to dry the specimen carefully, and obscure the microstructural details at the edge, or, worse yet, drip onto the objective (in an inverted microscope), causing loss of image clarity or even damage. Figure 13 shows a large shrinkage gap between a phenolic mount and a piece of 6061-T6 aluminum etched with dilute aqueous hydrofluoric acid. Nomarski differential interference contrast (DIC) reveals the curvature at the edge of the specimen and water stains (arrows) along the edge of the specimen. Figure 14 shows a high-speed steel specimen in a phenolic mount, where a large shrinkage gap is present, and the etchant, Vilella's reagent, has seeped out and now obscures the edge detail (arrows).

Fig. 13  Light micrograph showing a very large shrinkage gap between the phenolic resin mount (PM) and a specimen of 6061-T6 aluminum etched with aqueous 0.5% hydrofluoric acid. Note the metal flow at the specimen edge (revealed using Nomarski DIC illumination) and the water stains (arrows on the aluminum specimen).
Fig. 14 Light micrograph showing stain (arrows pointing up) from the etchant (Vilella's reagent) that seeped from the shrinkage gap (wide arrows pointing down) between the phenolic resin mount and the specimen of M2 high-speed steel

An advantage of compression mounting is production of a mount of a predictable, convenient size and shape. Further, considerable information can be engraved on the backside—this is always more difficult with unmounted specimens. Manual (hand) polishing is simplified, because the specimens are easy to hold. Also, placing a number of mounted specimens in a holder for semi- or fully-automated grinding and polishing is easier with standard mounts than for unmounted specimens. Mounted specimens are easier on the grinding/polishing surfaces than unmounted specimens.

Cold-mounting materials require neither pressure nor external heat and are recommended for mounting specimens that are sensitive to heat and/or pressure. Acrylic resins are a widely used castable resin, due to their low cost and short curing time, but they are generally unsatisfactory for failure studies, because shrinkage is a problem with acrylics. Epoxy resins, although more expensive than acrylics, are commonly used in failure studies, because epoxy physically adheres to specimens and can be drawn into cracks and pores, particularly if a vacuum impregnation chamber is employed and a low viscosity epoxy is used. Epoxies are very suitable for mounting fragile or friable specimens and corrosion or oxidation specimens. Dyes or fluorescent agents may be added to epoxies for the study of porous specimens such as thermal spray coated specimens. Epoxy resins are much more useful in failure analysis work than acrylic resins.

Most epoxies are cured at room temperature, and curing times can vary from 2 to 20 h. Some can be cured at slightly elevated temperatures in less time, as long as the higher temperature does not adversely affect the specimen. Acrylics do generate considerable heat during curing, and this can be strongly influenced by the molding technique used. Castable epoxy resins generate much less heat during curing, but this can vary substantially. The amount of heat generated increases as the epoxy volume increases and as the curing time decreases.

Edge Preservation

Edge preservation is the classic metallographic problem in failure analysis work, and many “tricks” have been promoted (most pertaining to mounting but some to grinding and polishing) to enhance edge flatness. These methods include the use of backup material in the mount, the application of coatings to the surfaces before mounting, or the addition of a filler material to the mounting resin. Plating of a compatible metal on the surface to be protected (electroless nickel has been widely used) is generally considered to be the most effective procedure. However, image contrast at an interface between a specimen and the electroless nickel may be inadequate for certain evaluations. Figures 11(a) and (d) show the surface of a specimen of 1215 free-machining steel that was salt bath nitrided. One specimen was plated with electroless nickel; both were mounted in epoxy resin. It is hard to tell where the nitrided layer stops for the plated specimen (Fig. 11d), which exhibits poor image contrast between the nickel and the nitrided surface. This is not a problem for the nonplated specimen (Fig. 11a).

Introduction of new technology has greatly reduced edge preservation problems. Mounting presses that cool the specimen to near-ambient temperature under pressure produce much tighter mounts. Gaps that form between specimen and resin are a major contributor to edge rounding, as shown in Fig. 10. Staining from bleed-out at shrinkage gaps obscures edge detail, (Fig. 14). Use of semiautomatic and automatic grinding/polishing equipment, rather than manual (hand) preparation, increases surface flatness and edge retention. To achieve the best results, particularly with a 200 mm (8 in.) diameter platen and a 125 mm (5 in.) diameter holder, the position of the specimen holder relative to the platen should be adjusted.
so that the outer edge of the specimen holder rotates out over the edge of the surface on the platen during grinding and polishing. This procedure can be used effectively with larger-diameter wheels, if the specimen holder diameter is large, relative to the platen. The use of “hard,” woven or nonwoven, napless surfaces for polishing with diamond abrasives (rather than softer cloths, such as canvas, billiard, and felt) maintains flatness. Rigid grinding discs (RGDs) yield surfaces with exceptional flatness. Final polishing with low-nap cloths for short times introduces very little rounding, compared to use of higher-nap, softer cloths.

These procedures produce better edge retention with all thermosetting and thermoplastic mounting materials. Nevertheless, there are still differences among the polymeric materials used for mounting. Thermosetting resins provide better edge retention than thermoplastic resins. Of the thermosetting resins, diallyl phthalate provides little improvement over the much-less-expensive phenolic compounds. The best results are obtained with an epoxy-based thermosetting resin that contains a special, hard filler material (Epomet). For comparison, Fig. 11 shows micrographs of a salt bath nitrided 1215 steel specimen mounted in a phenolic resin (Fig. 11b) and in methyl methacrylate (Fig. 11c) at 1000×. These specimens were prepared in the same specimen holder as those shown in Fig. 11(a) and (d), but neither displays acceptable edge retention at 1000×.

In the 1970s, very fine alumina spheres were mixed with liquid epoxy in an effort to improve edge retention. This is not a satisfactory procedure, because the particles are extremely hard (~2000 HV) and their grinding/polishing characteristics are incompatible with softer metals placed inside the mount. As a result, this product is no longer promoted for improving the edge retention of metallic specimens. Recently, a soft ceramic shot (~775 HV) was introduced that has grinding/polishing characteristics compatible with metallic specimens placed in the mount. Figure 15 shows an example of improving edge retention of annealed hot work die steel using soft ceramic shot in an epoxy mount.

Fig. 15  Good edge retention obtained in a cast epoxy mount containing soft ceramic shot filler. (Note the round particles in the epoxy at the top.) The specimen is annealed H13 hot work die steel, and it was etched with picral.

Following are general guidelines for obtaining the best possible edge retention. All of these factors contribute to the overall success, although some are more critical than others:

- Properly mounted specimens yield better edge retention than unmounted specimens, because rounding is difficult, if not impossible, to prevent at a free edge. Hot compression mounts yield better edge preservation than castable resins.
- Electrolytic or electroless plating of the surface of interest provides excellent edge retention. If the compression mount is cooled too quickly after polymerization, the plating may be pulled away from the specimen, leaving a gap. When this happens, the plating is ineffective for edge retention.
- Thermoplastic compression mounting materials are less effective than thermosetting resins. The best thermosetting resin for edge retention is an epoxy-based resin (Epomet) containing a hard filler material.
- Do not hot eject a thermosetting resin after polymerization and cool it quickly to ambient (e.g., by cooling it in water), because a gap forms between specimen and mount due to the differences in thermal contraction rates. Fully automated mounting presses cool the mounted specimen to near-ambient temperature under pressure, and this greatly minimizes gap formation due to shrinkage.
- Automated grinding/polishing equipment produces flatter specimens than manual (hand) preparation.
- Use the central force mode (defined later in this article) with an automated grinder/polisher, because this method provides better flatness than individual pressure mode (defined later in this article).
• Orient the position of the smaller-diameter specimen holder so that, as it rotates, its periphery slightly overlaps the periphery of the larger-diameter platen.
• Use pressure-sensitive-adhesive (PSA)-backed SiC grinding paper (when SiC is used) rather than water on the platen and a peripheral holddown ring, and PSA-backed polishing cloths rather than stretched cloths.
• Metal-bonded or resin-bonded diamond grinding discs produce excellent flat surfaces for a wide variety of materials.
• Use hard, napless surfaces for rough polishing (until the final polishing step(s)) and fine polishing. Use a napless or a low- to medium-nap cloth, depending on the material being prepared, for the final step(s), and keep it brief.
• Rigid grinding disks (RGD) produce excellent flatness and edge retention and should be used whenever possible.

Grinding

Grinding should commence with the finest grit size that establishes an initially flat surface and removes the effects of sectioning within a few minutes. An abrasive grit size of 180 or 240 in the American National Standards Institute/Coated Abrasive Manufacturers’ Institute (ANSI/CAMI) grading system (P180 or P280 in the FEPA, or Fédération Européenne des Fabricants de Produits Abrasifs system) is coarse enough to use on specimen surfaces sectioned by an abrasive cutoff wheel. Hacksawed, bandsawed, or other rough surfaces usually require abrasive grit sizes in the range of 120- to 180-grit (P120 to P180). Grinding must remove the damage created by sectioning (Fig. 16). If the initial grinding step does not remove this layer, the plane of polish may be within the zone of surface damage from cutting, and the true structure is not observed. The abrasive used for each succeeding grinding operation should be one or two grit sizes smaller than that used in the preceding step. A satisfactory grinding sequence might involve SiC papers with grit sizes of 240-, 320-, 400-, and 600-grit (P280, P400, P800, and P1200, respectively). This sequence is used in the “traditional” preparation approach.

Fig. 16 Light micrograph showing cutting damage (arrows at left) and a burr at the corner of a specimen of commercial-purity titanium (ASTM F67, grade 2) etched with modified Weck’s reagent and viewed with polarized light plus sensitive tint. The arrow along the top edge points to a surface layer containing mechanical twins.

As with abrasive-wheel sectioning, all grinding steps should be performed wet using water, provided that water has no adverse effects on any constituents of the microstructure. If water cannot be used during grinding, then some other non-aqueous coolant must be used, for example, kerosene or mineral spirits. Wet grinding minimizes specimen heating, prevents the abrasive from becoming loaded with metal removed from the specimen being prepared, and minimizes airborne metal-particle contamination and health problems.

Each grinding step, while producing damage itself, must remove the damage from the previous step. The depth of damage decreases with the abrasive size but so does the metal removal rate. For a given abrasive size, the depth of damage introduced is greater for soft materials than for hard materials. Grinding abrasive can become entrapped, or embedded, in the surface of specimens. This is especially true for soft, low-melting-point alloys ground using SiC paper. Embedding is more common with the finer-grit-size papers. Figure 17 illustrates embedding of SiC grinding paper abrasive in soft metals. The electron microprobe was used to study embedding of abrasives during grinding, and coating of the paper with candlewax or soap greatly reduced embedding (Ref 25).
For automated preparation using a multiple-specimen holder, the initial step is called planar grinding. This step must remove the damage from sectioning while establishing a common plane for all of the specimens in the holder, so that each specimen is affected equally in subsequent steps. Silicon carbide and alumina abrasive papers are commonly used for the planar grinding step and are very effective. Besides these papers, there are a number of other options available. One option is to planar grind the specimens with a conventional alumina grinding stone. This requires a special-purpose machine, because the stone must rotate at a high speed, ≥1500 rpm, to cut effectively. The stone must be dressed regularly with a diamond tool to maintain flatness, and embedding of alumina abrasive in specimens can be a problem. However, the approach provides high removal rates.

Polishing

Polishing is the final stage in producing a deformation-free surface that is flat, scratch-free, and mirrorlike in appearance. Such a surface is necessary for subsequent metallographic interpretation, both qualitative and quantitative. The polishing technique used should not introduce extraneous structures such as disturbed metal (Fig. 18), pitting (Fig. 19), dragging out of inclusions, “comet tailing” (Fig. 20), staining (Fig. 21), relief (height differences between different constituents, or between holes and constituents) (Fig. 22), or embedding (Fig. 23). Polishing usually is conducted in several stages. Rough polishing is conducted with 9, 6, or 3 μm diamond abrasives charged onto napless or low-nap cloths. For hard materials, such as through-hardened steels, ceramics, and cemented carbides, two rough-polishing steps may be required. The initial rough-polishing step may be followed by polishing with 1 μm diamond on a napless, low-nap, or medium-nap cloth. A compatible lubricant should be used sparingly to prevent overheating or deformation of the surface. Intermediate polishing should be performed thoroughly, so that final polishing may be of minimal duration.
Fig. 18 Light micrograph showing residual damage (arrows) from preparation that was not removed by the procedure when this specimen of commercial-purity titanium was prepared. The specimen was etched with Kroll’s reagent and photographed with Nomarski DIC illumination.

Fig. 19 Light micrograph showing pitting (arrows) after preparation of cold-drawn brass (Cu-20%Zn). Not etched

Fig. 20 Light micrograph illustrating “comet tails” emanating from hard nitrides on the surface of a prepared specimen of H13 tool steel. The specimen is unetched and viewed with Nomarski DIC.
Fig. 21 Light micrograph illustrating staining (arrow) on the surface of a prepared specimen of Ti-6\%Al-2\%Sn-4\%Zr-2\%Mo. The specimen was not etched.

Fig. 22 Light micrographs depicting (a) excessive and (b) low relief around voids in a braze between an austenitic stainless steel and Monel. The specimen was etched with glyceregia.
Manual polishing, or hand polishing, is usually conducted using a rotating wheel, where the operator rotates the specimen in a circular path counter to the wheel rotation direction. To obtain the best possible surfaces, it is necessary to use another step, typically with a 0.05 μm alumina or colloidal silica abrasive. This step can be performed on a wide variety of cloths. In the traditional approach, a medium-nap synthetic suede cloth is used. While this is still a useful cloth for many materials, increasing use is being made of synthetic polyurethane pads. This type of cloth is recommended when edge retention must be maximized.

The requirements of a good polishing cloth include the ability to hold the abrasive media, long life, absence of any foreign material that may cause scratches, and absence of any processing chemical (such as dye or sizing) that may react with the specimen. Many cloths of different fabrics, weaves, or naps are available for metallographic polishing. Napless or low-nap cloths are recommended for rough polishing with diamond abrasive compounds. Napless, low-, medium-, and, occasionally, high-nap cloths are used for final polishing. This step should be brief to minimize relief.

Mechanical polishing can be automated to a high degree using a wide variety of devices ranging from relatively simple systems to rather sophisticated, minicomputer, or microprocessor-controlled devices. Units also vary in capacity from a single specimen to a half-dozen or more at a time. These systems can be used for all grinding and polishing steps. These devices enable the operator to prepare a large number of specimens per day, with a higher degree of quality than by hand polishing and at reduced consumable costs. Automatic polishing devices produce the best surface flatness and edge retention. There are two automated approaches for holding specimens. Central force uses a specimen holder, with each specimen held in place rigidly. The holder is pressed downward against the preparation surface, with the force applied to the entire holder. Central force yields the best edge retention and specimen flatness. If the results after etching are inadequate, the specimens must be placed back in the holder, and the entire preparation sequence must be repeated. Instead of doing this, most metallographers repeat the final step manually and then re-etch the specimen. Individual force machines have a holder that does not hold the specimen rigidly. Specimens are placed inside a hole cut into a holder, and a piston comes down and presses the specimen against the working surface. Thus, the specimens can be removed and examined easily during the preparation cycle without losing planarity. This provides convenience if a step must be repeated, but the method is limited to mounted specimens, usually round mounts, and edge retention is not as good as with a central force holder.

Polishing usually involves the use of one or more of the following abrasives: diamond, alumina, and amorphous silicon dioxide in colloidal suspension. For certain materials, cerium oxide, chromium oxide, magnesium oxide, or iron oxide may be used, although these are used infrequently. With the exception of diamond, these abrasives are normally suspended in distilled water, but if the metal to be polished is not compatible with water, other suspensions, such as ethylene glycol, alcohol, kerosene, or glycerol, may be required. The diamond abrasive should be extended only with the product recommended by the manufacturer. Most diamond pastes and suspensions are water-based products, and these are suitable for most materials. However, oil-based diamond suspensions are needed to prepare materials sensitive to water.

Over the past forty years, a general procedure has been developed that is quite successful for preparing most metals and alloys. This method is based on grinding with SiC waterproof papers through a series of grits, then rough polishing with one or more diamond abrasive sizes, followed by fine polishing with one or more alumina suspensions of different particle size. This procedure is called the “traditional” method and is described in Table 1, listing equivalent ANSI/CAMI and FEPA grit sizes for the SiC paper.

---

**Table 1  The traditional method for preparing most metals and alloys**

<table>
<thead>
<tr>
<th>Surface</th>
<th>Abrasive/size</th>
<th>Load, N (lbf)</th>
<th>Speed, rpm/direction</th>
<th>Time, min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waterproof PSA paper</td>
<td>120/P120-grit SiC, water cooled</td>
<td>27 (6)</td>
<td>240–300 Comp</td>
<td>Until plane</td>
</tr>
<tr>
<td>Waterproof PSA paper</td>
<td>240/P280-grit SiC, water cooled</td>
<td>27 (6)</td>
<td>240–300 Comp</td>
<td>1–2</td>
</tr>
<tr>
<td>Waterproof PSA paper</td>
<td>320/P400-grit SiC, water cooled</td>
<td>27 (6)</td>
<td>240–300 Comp</td>
<td>1–2</td>
</tr>
<tr>
<td>Waterproof PSA paper</td>
<td>400/P800-grit SiC, water cooled</td>
<td>27 (6)</td>
<td>240–300 Comp</td>
<td>1–2</td>
</tr>
<tr>
<td>Waterproof PSA paper</td>
<td>600/P1200-grit SiC, water cooled</td>
<td>27 (6)</td>
<td>240–300 Comp</td>
<td>1–2</td>
</tr>
</tbody>
</table>

Fig. 23 Light micrograph showing 6 μm diamond (arrows) abrasive embedded in the surface of a partly prepared specimen of lead
Note: Comp, complementary (platen and specimen holder both rotate in the same direction).

This procedure is used for manual or automated preparation, although manual control of the force applied to a specimen would not be very consistent. Complementary motion means that the specimen holder is rotated in the same direction as the platen and does not apply to manual preparation, because this cannot be done. In manual preparation, the specimen is held still in grinding, aside from moving between the edge and the center. In manual polishing, the specimen is rotated clockwise, against the counterclockwise wheel rotation direction. Some machines can be set so that the specimen holder rotates in the direction opposite to that of the platen, called “contra.” This provides a more aggressive action but was not adopted when the traditional approach was automated. The traditional method is not rigid, because other polishing cloths may be substituted, and one or more of the polishing steps might be omitted. Times and pressures could be varied, as well, to suit the needs of the work or the material being prepared. This is the “art” of metallography.

New concepts and new preparation materials have been introduced that enable metallographers to shorten the process while producing better, more consistent results. Much of this effort has centered on reducing or eliminating the use of SiC paper in the grinding steps. In all cases, an initial grinding step must be used, but there is a wide range of materials that can be chosen instead of SiC paper. There is nothing wrong with the use of SiC for the first step, except that it has a short life. If an automated device is used that holds a number of specimens rigidly (central force), then the first step must remove the sectioning damage on each specimen and bring all of the specimens in the holder to a common plane. This first step is often called planar grinding. Silicon carbide paper can be used for this step, although more than one sheet may be needed. Alternatively, the metallographer could use alumina paper, an alumina stone on a dedicated high-speed grinder, metal- or resin-bonded diamond discs, stainless steel mesh cloth (diamond is applied during use), RGDs (diamond is applied during use), or lapping platens of several types (diamond is applied and becomes embedded in the surface during use).

In contemporary preparation methods, one or more steps using diamond abrasives on napless surfaces usually follow planar grinding. Pressure-sensitive-adhesive-backed silk, nylon, or polyester cloths are widely used. These give good cutting rates, maintain flatness, and minimize relief. Silk cloths provide the best flatness and excellent surface finishes relative to the diamond size used. Thicker harder, woven cloths are more aggressive, give nearly as good a surface finish, similar excellent flatness, and longer life than silk cloths. Synthetic chemotextile pads give excellent flatness and are more aggressive than silk. They are excellent for retaining second-phase particles and inclusions. Diamond suspensions are very popular with automated polishers, because they can be added easily during polishing, although it is still best to charge the cloth initially with diamond paste of the same size to get polishing started quickly. Final polishing could be performed with a very fine diamond size, such as 0.1 μm diamond, depending on the material, the metallographer's needs, and personal preferences. Otherwise, final polishing is performed with colloidal silica or with alumina slurries using napless or low- to medium-nap cloths. For some materials, such as titanium and zirconium alloys, an attack polishing solution is added to the abrasive slurry to enhance deformation and scratch removal and improve polarized light response. Contra rotation (head moves in the direction opposite to the platen) is preferred, because the slurry stays on the cloth better, although this does not work if the head rotates at a high number of revolutions per minute (>90). Examples of generic preparation practices for many metals and alloys are found in Tables 2, 3, 4.

### Table 2  Generic four-step contemporary practice for many metals and alloys

<table>
<thead>
<tr>
<th>Surface</th>
<th>Abrasive/size</th>
<th>Load, N (lbf)</th>
<th>Speed, rpm/direction</th>
<th>Time, min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waterproof PSA paper</td>
<td>120/P120-, 180/P180-, or 240/P280-grit SiC, water cooled</td>
<td>27 (6)</td>
<td>240–300</td>
<td>Until plane</td>
</tr>
<tr>
<td>Surface</td>
<td>Abrasive/size</td>
<td>Load, N (lbf)</td>
<td>Speed, rpm/direction</td>
<td>Time, min</td>
</tr>
<tr>
<td>--------------------------</td>
<td>----------------------------------------------------</td>
<td>---------------</td>
<td>----------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Silk cloth</td>
<td>9 μm polycrystalline diamond suspension</td>
<td>27 (6)</td>
<td>120–150 Comp</td>
<td>5</td>
</tr>
<tr>
<td>Synthetic woven cloth</td>
<td>3 μm polycrystalline diamond suspension</td>
<td>27 (6)</td>
<td>120–150 Comp</td>
<td>4</td>
</tr>
<tr>
<td>Synthetic short nap cloth</td>
<td>~0.05 μm colloidal silica or sol-gel alumina suspensions</td>
<td>27 (6)</td>
<td>120–150 Contra</td>
<td>2</td>
</tr>
</tbody>
</table>

Note: Comp, complementary (platen and specimen holder both rotate in the same direction). Contra, platen and specimen holder rotate in opposite directions.

### Table 3 Four-step contemporary practice for steels using a rigid grinding disc

<table>
<thead>
<tr>
<th>Surface</th>
<th>Abrasive/size</th>
<th>Load, N (lbf)</th>
<th>Speed, rpm/direction</th>
<th>Time, min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waterproof PSA paper</td>
<td>120/P120-, 180/P180-, or 240/P280-grit SiC, water cooled</td>
<td>27 (6)</td>
<td>240–300 Comp</td>
<td>Until plane</td>
</tr>
<tr>
<td>Rigid grinding disk(^a)</td>
<td>9 μm polycrystalline diamond suspension</td>
<td>27 (6)</td>
<td>120–150 Comp</td>
<td>5</td>
</tr>
<tr>
<td>Synthetic woven cloth</td>
<td>3 μm polycrystalline diamond suspension</td>
<td>27 (6)</td>
<td>120–150 Comp</td>
<td>4</td>
</tr>
<tr>
<td>Synthetic short nap cloth</td>
<td>~0.05 μm colloidal silica or sol-gel alumina suspensions</td>
<td>27 (6)</td>
<td>120–150 Contra</td>
<td>2</td>
</tr>
</tbody>
</table>

Note: Comp, complementary (platen and specimen holder both rotate in the same direction). Contra, platen and specimen holder rotate in opposite directions.
\(^a\) Disk for metals >175 HV

### Table 4 Four-step contemporary practice for nonferrous metals using a rigid grinding disc

<table>
<thead>
<tr>
<th>Surface</th>
<th>Abrasive/size</th>
<th>Load, N (lbf)</th>
<th>Speed, rpm/direction</th>
<th>Time, min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waterproof PSA paper</td>
<td>240/P280- or 320/P400-grit SiC, water cooled</td>
<td>22 (5)</td>
<td>240–300 Comp</td>
<td>Until plane</td>
</tr>
<tr>
<td>Rigid grinding disk(^a)</td>
<td>6 μm polycrystalline diamond suspension</td>
<td>22 (5)</td>
<td>120–150 Comp</td>
<td>5</td>
</tr>
<tr>
<td>Synthetic woven cloth</td>
<td>3 μm polycrystalline diamond suspension</td>
<td>22 (5)</td>
<td>120–150 Comp</td>
<td>4</td>
</tr>
<tr>
<td>Synthetic short nap cloth</td>
<td>~0.05 μm colloidal silica or sol-gel alumina suspensions</td>
<td>22 (5)</td>
<td>120–150 Contra</td>
<td>2</td>
</tr>
</tbody>
</table>

Note: Comp, complementary (platen and specimen holder both rotate in the same direction). Contra, platen and specimen holder rotate in opposite directions.
\(^a\) Disk for metals >50 HV

The starting SiC abrasive size is chosen based on the degree of surface roughness and depth of cutting damage and the hardness of the material. Never start with an abrasive size coarser than necessary to remove the cutting damage and...
achieve planar conditions in a reasonable time. A 1 μm diamond step can be added for more difficult-to-prepare materials using a napless cloth and a similar approach as the third step, but a 3 min polish. A similar scheme can be developed using RGDs. These discs are generally restricted to materials above a certain hardness level, such as 175 HV, although some softer materials can be prepared using them. The disc can also be used for the planar grinding step. An example of such a practice, applicable to nearly all steels (results are marginal for solution-annealed austenitic stainless steels), is given in Table 3.

The planar grinding step could also be performed using a 45 μm metal-bonded or a 30 μm resin-bonded diamond disc or with a RGD and 15 or 30 μm diamond, depending on the material. Rigid grinding discs contain no abrasive; they must be charged during use, and suspensions are the easiest way to do this. Polycrystalline diamond suspensions are favored over monocrystalline synthetic diamond suspensions for most metals and alloys due to their higher cutting rate. Again, a 1 μm diamond step can be added for difficult materials or to ensure generation of the required degree of perfection in the surface finish.

Rigid grinding discs designed for soft metals and alloys are used in a similar manner. These discs are quite versatile and can be used to prepare harder materials as well, although their wear rate is greater when used to prepare very hard materials. A generic four-step practice is given in Table 4 for soft metals and alloys.

The planar grinding step can be performed with the 30 μm resin-bonded diamond disc or with a second RGD and 15 or 30 μm diamond, depending on the metal or alloy. For some very difficult metals and alloys, a 1 μm diamond step on a synthetic woven cloth (similar to step 3 but for 3 min) could be added, and/or a brief vibratory polish (use the same cloths and abrasives as for step 4) may be needed to produce perfect publication-quality images.

**Electrolytic Polishing**

Electrolytic polishing can be used to prepare specimens with deformation-free surfaces. The technique offers reproducibility and speed. In most cases, the published instructions for electrolytes tell the user to grind the surface to a 600-grit (P1200) finish and then electropolish for approximately 1 to 2 min. However, the depth of damage after a 600-grit (P1200) finish may be several micrometers, but most electropolishing solutions remove only about 1 μm/min. In this case, the deformation is not completely removed. In general, electropolished surfaces tend to be wavy rather than flat, and focusing may be difficult at high magnifications. Further, electropolishing tends to round edges associated with external surfaces, cracks, or pores. In two-phase alloys, one phase polishes at a different rate than another, leading to excessive relief. In some cases, one phase may be attacked preferentially, and inclusions are usually attacked. Consequently, electrolytic polishing is not recommended for failure analysis or image analysis work, except possibly as a very brief step at the end of a mechanical polishing cycle to remove whatever minor damage may persist.

**References cited in this section**


**Metallographic Techniques in Failure Analysis**

George F. Vander Voort, Buehler Ltd.

**Examination of Microstructures**

The second main use of light microscopy is to determine the microstructure of the material in question to evaluate its influence on the failure. Such examination is first performed at the origin of the failure to detect any anomalies that may have arisen from inadequate material quality, fabrication or heat treatment deficiencies, or alterations due to service conditions. In failures that do not involve fracture, for example, certain types of wear or corrosion failures, the relationship of the microstructure to the observed damage is assessed at the damage sites (Ref 26, 27). Examples of such work are given in the following to illustrate the value of light microscopy.
Figure 24(a) shows a micrograph of an unusual tensile test fracture that was obtained during the evaluation of a carbon steel (0.25% C, 0.63% Mn, 0.27% Si) casting. Instead of the usual cup-and-cone tensile fracture, the surface was at an angle, approximately 45°, to the tensile axis and was rather rough. Scanning electron microscope examination (Fig. 24b) of the fracture revealed numerous voids typical of shrinkage cavities. Light microscope examination (Fig. 24c) also reveals these cavities. Neither SEM or LM examination by itself was satisfactory for determining why the fracture was unusual, but together they provide a more complete picture. The shrinkage cavities caused failing yield strength, elongation, and reduction of area results.

![Figure 24](image)

**Fig. 24** (a) Macrograph of fracture, (b) SEM fractograph, and (c) light micrograph showing shrinkage cavities in an unusual tensile fracture from a carbon steel casting. The microstructure was revealed using nital.

Another example of a material quality problem is shown in Fig. 25(a) and (b). These micrographs show a seam that was the cause of rejection for a forged pitman arm. In this case, the seam is not perpendicular to the surface because of the metal flow during forging.

![Figure 25](image)

**Fig. 25** Light micrographs of two cross-sectional views of a seam found on a closed-die forged pitman arm showing decarburization and internal oxidation. Etched with 2% nital

Another common material problem is the presence of decarburization that may be present on as-rolled stock or may form during heat treatment or, in some cases, during service. Light microscopy is the generally accepted method for detecting decarburization and measuring its extent (Ref 28). Figures 26(a), (b), and (c) show examples of decarburized AISI 5160 spring steel in the as-rolled condition (Fig. 26a) and after heat treatment (Fig. 26b, c). Decarburization of hardened coil springs is not desired, because it reduces the fatigue life of the spring.
Fig. 26 Light micrographs of decarburization observed on cross sections of as-rolled and heat treated AISI 5160H alloy steel spring. (a) Nickel plating on top of scale on an as-rolled specimen. (b) Partial decarburization at the surface of a hardened specimen. (c) Free ferrite and partial decarburization at the surface of a hardened specimen. Etched with 2% nital

A less common example of decarburization is shown in Fig. 27, which shows the surface of a decarburized solution-annealed austenitic manganese steel. At locations where the carbon content is below approximately 0.50%, epsilon martensite is observed (Ref 29, 30). This structure does not possess the remarkable work-hardening capacity typical of such alloys.

Fig. 27 Light micrograph showing epsilon martensite at the surface of a decarburized (less than 0.5% C) austenitic manganese steel specimen. Etched with 2% nital/20% sodium metabisulfite

Figure 28 provides another example of inadequate material quality. This shows a defect observed on the polished inside diameter of an AISI 420 stainless steel mold. Sectioning of the mold at the defect and light microscopy examination revealed a large silicate inclusion that caused the defect.
Numerous failures have been traced to problems occurring during heat treatment. The following examples show how light microscopy was employed to analyze such failures.

Figure 29 shows the operating surface of an AISI L6 tool steel punch that exhibited poor service life. Examination of the microstructure revealed that it was under-austenitized in heat treatment. The hardness was several points HRC below the expected value.

Figure 30 shows the microstructure of a roll made from AISI O1 tool steel that cracked during quenching. The quench cracks were located adjacent to deep stamp marks, and the microstructure contained considerable retained austenite. No residual carbide, normally present in this grade when properly heat treated, was present. Consequently, an excessively high austenitizing temperature had been employed. The hardness was 56/57 HRC, which was increased to 62/64 HRC after the specimen was refrigerated in liquid nitrogen.
Fig. 30 Light micrograph of overaustenitized AISI O1 tool steel containing coarse plate martensite and substantial unstable retained austenite. Specimen etched with nital

Figure 31 shows two views of the microstructure of a jewelry-striking die made from AISI S7 tool steel that cracked soon after being placed in service. The views show a coarse zone at the surface, consisting of coarse plate martensite and retained austenite. The carbon content at the surface was 0.79%, while the interior was 0.53%. The die had been lightly carburized to a depth of approximately 0.5 mm (0.020 in.) due to improper furnace atmosphere control.

Fig. 31 Light micrographs of an AISI S7 tool steel jewelry-striking die that failed due to the presence of a carbon-enriched surface layer that contained coarse plate martensite and unstable retained austenite. Specimen etched with nital

Figure 32 shows the microstructure of a cracked carburized tread of a track wheel made from AISI 1035 carbon steel. The 60 cm (24 in.) diameter track wheel was carburized at a higher temperature than usual, and the diffusion cycle after carburizing had been omitted. Microstructural examination revealed that cracking followed the grain-boundary carbide network present due to the improper carburizing practice.
Fig. 32 Light micrographs of a carburized AISI 1035 track wheel that cracked due to the presence of an extensive grain-boundary carbide film. Specimen etched with nital

Figure 33 shows the microstructure of an AISI D2 tool steel powder metallurgy die that was deformed at one end after heat treatment. Light microscopy revealed that the steel was locally melted in this region, apparently due to flame impingement during austenitization.

Fig. 33 Light micrograph of a melted region found on an AISI D2 powder metallurgy die after heat treatment. Specimen etched with Marble's reagent

As a final example of a heat-treatment-related failure, Fig. 34 shows the microstructure of an AISI D2 draw die insert that exhibited galling and chipping after limited service. Examination of the microstructure revealed that the austenitizing temperature was well above the recommended 1010 °C (1850 °F) temperature, high enough to cause liquidation at the grain boundaries. Note the nearly complete grain-boundary network of skeletal carbides, similar in appearance to an as-cast condition.
Failures may also arise during fabrication processes. The next several examples illustrate such problems and the use of light microscopy. A wide variety of problems can occur; these examples illustrate only a few of many such problems. Figure 35 shows planar and through-thickness views of cracks observed in several reaustenitized zones in a forged steel specimen. An electric pencil had been used to identify the component. The cracked regions were present on the face placed against the grounding plate. Apparently, arcing occurred along the edges touching the grounding plate, producing enough heat to locally reaustenitize the steel. On cooling, the hardenability was sufficient to form as-quenched martensite in these regions. The transformation-related expansion caused cracking in these reaustenitized spots.

Figure 36 shows an example of central bursting (“chevron cracking”) in extruded, as-rolled AISI 4615 alloy steel. Although optimal extrusion parameters can generally minimize such failures, the ductility of the material is also an
important variable. Because this alloy has substantial hardenability, it is difficult to prevent formation of bainite and martensite in small, as-rolled section sizes. In this case, the microstructure consisted of ferrite, bainite, and martensite (the arrows point to microcracks present in martensite patches).

![Image](image.png)

**Fig. 36** (a) Central bursting during extruding of AISI 4615 alloy steel specimen was promoted by (b) the presence of bainite and martensite (arrows point to microcracks in the martensite) in the as-rolled stock. Specimen etched with 4% picral, followed by 2% nital.

Final grinding of hardened tool steel components is an important processing step that must be carefully controlled. Improper grinding practices can produce a fine network of surface cracking and a characteristic scorch pattern easily revealed by macroetching (Ref 31, 32). Examination of the microstructure at the cracks reveals their shallow nature and a back-tempered condition at the surface, as shown in Fig. 37. In some cases, a shallow, reaustenitized light-etching layer of as-quenched martensite is found at the surface (Ref 31) above the back-tempered zone. In many cases of abusive grinding, microstructural examination reveals that the die was not tempered. As-quenched tool steels are very difficult to grind without causing such cracking. In the example shown, however, the AISI O1 tool steel die was properly tempered, and one must conclude that the grinding operation was at fault.
Electrical discharge machining (EDM) is widely used to produce cavities in tool steels. Because it is a spark-erosion process that generates considerable temperature and localized melting, it must be rigorously controlled. After EDM, the cavity surface generally is stoned to remove melted surface layers, and the part is tempered. However, numerous failures have been observed in EDM-processed components (Ref 31, 32). Figure 38 shows a classic example of the microstructure of a failed part that was machined by this technique but not properly posttreated. This was a plastic mold made from AISI S7 tool steel, where pitting was observed on the cavity after polishing. The microstructure at the pit shows a large remelted surface layer (note dendrites). Beneath this layer is a white-etching zone of as-quenched martensite, typical of such failures. Beneath this is a back-tempered zone, where the temperature was below the upper critical temperature for the steel. In many such EDM-related failures, the as-cast layer is not present, but the as-quenched region is always observed.
Fig. 38  (a) Pitting on this mold cavity (arrow) was observed after polishing of the AISI S7 plastic mold and was caused by the use of improper post-EDM procedures. (b) The classic appearance of such failures. There is a large, remelted surface layer above a reaustenitized, untempered zone. Specimen etched with nital

Failures also occur due to service conditions, and the next several examples illustrate some of the many problems that can occur. These examples are provided to illustrate the value of light microscopy in such studies. Figure 39 shows the microstructure of an as-cast 25%Cr-12%Ni heat-resisting alloy used as a hook for holding a basket of parts during austenitizing and water quenching. The alloy contained delta ferrite that transformed to sigma, producing a discontinuous network that caused the hook to crack. Electrolytic etching was used to reveal only the sigma phase.

Fig. 39  Light micrograph showing sigma phase revealed by selective etching with 10N KOH (electrolytic). The brittle sigma phase caused extensive cracking in a 25% Cr-12% Ni cast heat treatment basket hook.
Figure 40 shows surface damage due to wear on a 4485 alloy steel medart roll. The specimen was plated with electroless nickel for edge retention. Some of the nickel can be observed in the crack. Note that the surface layer has been reaustenitized due to service-generated heat. Oxidation at the surface and flow in the layer can also be observed.

Fig. 40  Light micrograph showing wear damage at the surface of a 4485 alloy steel medart roll. The surface was nickel plated for edge retention and etched with nital.

It is not unusual to observe as-quenched martensitic layers produced on the surface of steels subjected to heavy wear conditions. Figure 41 shows such a layer on the surface of a scrap chopper knife made from a proprietary wear-resistant tool steel. Cracking and spalling may be produced at such layers because of their brittle nature. Such a condition is often observed on steel railroad rails where cracking and spalling has occurred. Figure 42 shows an example of surface cracking near a spall on a 136 lb/yd rail. Microstructural examination revealed three regions at the surface. The outer zone is featureless, typical of heavily deformed as-quenched martensite. The hardness was approximately 60 HRC. Beneath this is a zone containing the same white-etching martensite plus a network of ferrite. Beneath this zone is the normal pearlitic structure of an as-rolled rail steel. Tempering of the sample produced a tempered martensite structure with spheroidized carbides in the outer zones containing the white-etching constituent.

Fig. 41  Light micrograph of the surface of a badly worn steel chopper knife made from an air-hardenable tool steel showing a reaustenitized surface zone and a back-tempered region below it. The large angular particles are nitrides. Specimen etched with nital.
Fig. 42 Light micrograph of a white-etching surface layer formed on a rail head due to frictional heat. This specimen was taken adjacent to a spalled area. Specimen etched with picral

References cited in this section


Scanning Electron Microscopy

L.S. Chumbley, Iowa State University; L.D. Hanke, Materials Evaluation and Engineering Inc.

Introduction

THE SCANNING ELECTRON MICROSCOPE (SEM) is one of the most versatile instruments for investigating the microstructure of metallic materials. Compared to the optical microscope, it expands the resolution range by more than one order of magnitude to approximately 10 nm in routine instruments, with ultimate values below 3 nm. Useful magnification thus extends beyond 10,000× up to 100,000×, closing the gap between the optical and the transmission electron microscope. Compared to optical microscopy, the depth of focus, ranging from 1 μm at 10,000× to 2 mm (0.08 in.) at 10×, is larger by more than two orders of magnitude.
Development of SEM Technology

Development of what is now recognized as a scanning electron microscope was begun in 1938 by Ardenne (Ref 1) who hypothesized that an image of a sample surface could be generated by subjecting the sample to a beam of electrons and then monitoring the current that resulted from the sample. The success of these initial experiments led to further work that improved and modified his original design, until the first commercially successful SEM was marketed by Cambridge in the early 1960s. Since that time, the use of SEMs has expanded dramatically until SEMs are commonly found in use today in industry, education, and research applications.

The success of the SEM is directly related to the versatility of the instrument. Typical SEMs have the ability to deliver clear, high-magnification images while also providing the user with the ability to obtain chemical compositions from samples. Recent developments (Ref 2) also have greatly increased the capability of the SEM to obtain crystallographic information from a sample. Thus, the analytical capabilities of an SEM cover a wide range of information, and accessories can be chosen to suit the particular interests of the user.

Scanning electron microscopes have been found particularly useful in failure analysis investigations. This is true for a number of reasons, the most prominent being the ability of the SEM to image large, non-flat samples from low to high magnifications (approximately 10× to greater than 100,000×). The ability to obtain in-focus images of rough samples over a large change in vertical height is termed depth of field and it is this trait that gives SEM images their characteristic three-dimensional appearance. The large depth of field is made possible by the relationship between the small size of the electron probe used as related to the size of the imaging pixel determined by the operating magnification. This is explained in more detail in the following section.

While all SEMs exhibit similar characteristics, a number of variations exist, depending on the specific application and/or environment in which they are used. By changing the type of electron source used, the signals to be processed, and the type of vacuum systems, an SEM can be designed to enhance particular capabilities. For example, proper choice of the electron source can be used to enhance imaging, while changing the types of x-ray detectors can provide better chemical analysis. Modifying the vacuum system can allow examination of a wide range of sample types, including nonconductive or wet samples. Thus, a number of types of SEMs now exist with slightly different capabilities.

A conventional SEM typically uses a heated filament to produce electrons. These thermionic sources use tungsten or lanthanum hexaboride (LaB$_6$) as the filament material, tungsten being less expensive and more robust while LaB$_6$ is brighter and lasts longer. Scanning electron microscopes that use a high electric field to remove electrons from a tungsten filament are termed cold-field-emission SEMs or high-resolution SEMs and have the highest resolution capabilities of all SEMs. A compromise is reached in the thermal-field-emission (or Schottky gun) SEM, where a combination of heat and electric field are used to produce electrons. These machines have better resolution than conventional SEMs, while being easier to use than the high-resolution machines.

A conventional SEM also requires that the sample be electrically conductive to prevent a charge buildup in the sample that affects the incoming primary and emitted secondary electrons, resulting in a poor, distorted image that is constantly changing in contrast and location. The sample should also be dry when placed in the vacuum chamber and subjected to the beam, since outgassing will similarly prevent a clear image from being obtained. These restrictions on the sample can be overcome by changing the design of the vacuum system and the choice of detection system used to image the sample. Microscopes adapted in this way are termed low- or variable-pressure microscopes. In this type of microscope, the pressure in the sample chamber is raised to a value on the order of 0.1 to 1 torr. Interaction of the electron beam with gas molecules in the region where the beam strikes the sample effectively creates a positively charged “cloud” of ions above the surface of the sample. This positive “cloud” offsets the negative charge buildup that occurs in insulating or poorly conductive samples and allows images of these types of samples to be obtained using back-scattered electrons. Modern microscopes can be purchased that allow the user to operate in either the high-vacuum, conventional imaging mode or in the low- or variable-pressure mode, making them extremely flexible for all types of investigations.

If even greater capability to image nonconductive or wet samples is desired, an environmental SEM can be purchased. This type of SEM operates at vacuum levels of 20 torr and allows both backscattered and secondary images to be obtained from the sample surface. This is made possible by the use of an electron detector that operates on the principle of induced current. In this instance, the interaction of the electron signals from the sample surface with the gas molecules above the sample induces a current in the detector, which consists of two parallel conductive plates, one placed above the sample and one that acts as the sample holder. By proper selection of where the current is measured, a secondary or backscattered image can be obtained.

References cited in this section

Scanning Electron Microscopy

L.S. Chumbley, Iowa State University; L.D. Hanke, Materials Evaluation and Engineering Inc.

Operation

All SEMs may be divided into three basic systems of components, namely, the electron optical column, signal detection and display equipment, and the vacuum system.

Electron-Optical Column

A schematic of a typical SEM column is shown in Fig. 1. Electrons are generated by the electron gun and focused and controlled by a series of electromagnetic lenses and coils. The resultant beam is collimated and defined by various apertures placed in the beam path.

Fig. 1  Schematic showing the general layout of a typical scanning electron microscope

The electron gun consists of a filament, usually of tungsten or LaB₆, placed within a metal cap called the Wehnelt. A current is applied to the filament, causing resistance heating and resulting in emission of light, heat, and electrons. The electrons are attracted down the microscope column by the anode, which is positively biased with respect to the filament. The Wehnelt cylinder placed around the filament carries a slight negative charge with respect to the filament. This charge produces a slight repulsive force on the emitted electrons, causing them to be focused at a point near the exit hole of the filament/Wehnelt gun assembly. The size of this focused crossover point is one factor in determining the ultimate spatial resolution of the instrument. Field emission SEMs have a smaller crossover point, which is the reason for their higher resolving capabilities.

Once the focused electron beam passes the anode, it enters the first set of electromagnetic lenses, collectively called the condenser lens. The condenser lens control is given a variety of labels by manufacturers, such as resolution, probe current, and so forth. Regardless of the name, the condenser lens acts to demagnify the electron beam, producing an even
smaller probe for subsequent imaging purposes. As the strength of the condenser lens increases, the demagnification increases and the probe size becomes progressively smaller. This results in increased resolution. However, the overall signal strength will drop as a finer probe size is only accomplished with a concomitant loss in signal strength for a fixed objective aperture size. Modern digital microscopes can compensate for this signal loss to a certain extent by frame averaging, which allows an image to be acquired over the course of any number of scans. An improper understanding of the trade-off between resolution and signal strength, as determined by the condenser lens, is the most common error associated with SEM operation by a novice. A strong signal, which gives a clear image on the viewing screen, may have poor resolution, whilst a poor, snowy image on the view screen may produce excellent images when a photograph is taken, due to the higher resolution.

After passing through the condenser lens assembly, the electron beam enters the objective lens. The objective lens is responsible for focusing the beam on the surface of the sample, and the control dial for this lens is universally termed “focus” by SEM manufacturers. Unlike the condenser lens, the operation of the objective lens is such that it is always set to produce a focused beam image at the surface of the sample. As the distance from the sample to the objective lens varies, the lens is continually readjusted by the operator to once again bring the probe into focus on the sample surface. Thus, the objective lens current is determined by the distance of the sample to the bottom of the lens. This distance, termed the working distance, is usually given as a readout somewhere on the microscope console. Demagnification of the beam using the objective lens to produce a smaller probe (and therefore higher resolution) is therefore only possible by decreasing the working distance and refocusing the image. However, a decrease in the working distance results in a corresponding loss in depth of field.

Within the objective lens are a series of coils that allow scanning of the beam across the surface of the sample in various ways (scan coils, deflection coils) and correction of any astigmatism associated with the beam (stigmation coils). Also contained within the region of the objective lens is the objective aperture. The objective aperture serves to define the extent of the beam and determines the angular spread. A small aperture will have a small angular spread, which is good for increasing depth of field. However, the smaller the aperture the less the signal that reaches the sample from the condenser lens.

Once the beam strikes the sample, the resolution and depth of field obtained is determined by the relationship between the probe size, the angle of convergence of the beam to produce this probe, and the pixel size on the sample. The pixel size is determined by the magnification and may be thought of as a discreet area on the sample where data is taken and transferred to a corresponding area on a computer screen. This complex relationship is best understood by considering a few simple examples. First consider the schematic shown in Fig. 2. For the sake of convenience, assume an SEM where the beam scans the surface in a square array, transferring information from the sample to the viewing screen 1,000 times per line before dropping to the next line. For a 1,000 line scan, there are, therefore, 1,000,000 pixels of information transferred from the sample to the view screen per scan. For a total scan length on the surface of the sample of 1 cm × 1 cm, each pixel is therefore 0.01 mm square. Magnification is determined simply by dividing the length of the scan of the display screen by the length of scan on the sample. For a 10 cm long scan on the display screen, magnification is just:

\[
10 \text{ cm}/1 \text{ cm} = 10 \times \text{magnification}
\]  

(Eq 1)

\[\text{Magnification} = \frac{\text{display length}}{\text{sample length}} \div L = 10x\]

Fig. 2 Illustration of how the information obtained from a square raster on the sample is transferred to the viewing screen to produce magnification
Magnification is increased by decreasing the length of the scan on the sample. The number of pixels scanned remains constant, being 1000 points × 1000 lines. Thus, as magnification increases, the sample pixel size will become progressively smaller. When the electron beam is on the order of the sample pixel size, a loss of resolution occurs and the image is unresolved and blurry, even though the focus (i.e. the setting of the objective lens) is correct. However, as long as the probe size stays less than the pixel size on the specimen, a clear, resolved image will be produced.

If the probe size is significantly smaller than the pixel size, then an in-focus image can be obtained over a considerable vertical distance in a rough sample. This is illustrated in Fig. 3, which schematically shows a view of the electron beam as it strikes a sample. The beam converges to a point and then diverges again, once the image plane is passed. The size of the pixel element on the sample (as determined by the magnification), is shown. It is clear from the illustration that even though the size of the electron beam increases above and below the focus point, that size is still smaller than the pixel element size for a considerable vertical distance. This means that the entire sample surface between the indicated vertical distance will appear in focus. This depth of focus (or depth of field) is what gives SEM images their characteristic three-dimensional appearance. Note that if a larger objective aperture was inserted, the angle of convergence/divergence would be greater, as shown by the dotted lines in Fig. 3. For this greater angle, the probe size becomes greater than the pixel size over a smaller vertical distance, and the depth of field is reduced.

Fig. 3 Illustration showing how depth of focus is dependent on specimen pixel size and aperture size

Successful operation of the SEM, then, depends upon determining which parameter is important for any particular investigation and then optimizing the microscope to obtain the desired information. In a failure analysis investigation, where an extremely rough surface exists, depth of field may be the most important parameter. In this case, depth of field may be enhanced by having a long working distance, low magnification, and a small objective aperture. If one desires to image fatigue striations and high resolution is required, this can be obtained by a strong condenser lens (i.e. high resolution or low probe current) and a short working distance.

**Signal Detection and Display**

A wide range of signals is available for collection in any SEM. The most common signals used are the secondary and primary backscattered electron signals (SE, BSE) for imaging and the characteristic x-ray signal emitted from the sample for composition identification.
Secondary electrons are low-energy electrons emitted from the sample as a result of coulombic interactions between the incident electrons from the beam and electrons present in the sample. Emission of secondary electrons is such that as the angle between the beam and the surface increases, secondary emission increases (Fig. 4a). Thus, secondary electrons are ideal for imaging rough samples (such as fracture surfaces), where the angle between the beam and the surface is constantly varying. For a smoothly polished sample, where the angle remains constant across the surface of the sample, little or no contrast is seen using secondary electrons.

Due to their low energy, secondary electrons are typically collected to a photomultiplier/scintillator detector by placing a positive bias on the detector. For each pixel where the beam interacts with the sample, a signal is collected, processed, and displayed as an intensity level on a CRT (technically, a cathode ray tube; as a practical matter, a high resolution black and white or monochrome TV) screen. By coupling the scan of the beam to the scan of the CRT, a point by point correspondence exists between the surface roughness of the sample and the intensity observed on the screen. As the beam from the electron gun scans across a rough surface, the number of emitted secondary electrons continually changes, and the collected signal varies correspondingly. Thus, the intensity displayed on the CRT level changes from white (high-electron-emission) to black (low-electron-emission), and black and white contrast results in an image being produced in a manner exactly analogous to a black and white TV image.

The strength of the secondary electron signal often may be increased by tilting the sample slightly toward the detector. This results in more efficient collection of the secondary signal and can improve the signal-to-noise ratio without decreasing resolution. This may be helpful when trying to get the utmost resolution possible.

Backscattered Electrons. Imaging is somewhat different for primary backscattered electrons (BSE), which are incident electrons that have undergone elastic Rutherford scattering from the nuclei of the atoms that comprise the sample. For this signal, as the size of the nuclei gets larger (i.e., higher atomic number), the BSE emission increases in a nearly linear fashion (Fig. 4b). Thus, for a flat, polished sample containing numerous phases, where little or no contrast may be seen using secondary electrons, an image of high contrast may still be produced using the BSE signal. This is illustrated in Fig. 5, where a layered structure of various metals is imaged using secondary and backscattered electrons. While the layers can be clearly seen in the etched sample, it is only in the BSE image that the large difference in atomic number of the components is discerned. If a polished, homogeneous, single-phase sample were placed in an SEM (e.g., an unetched silicon wafer), little or no contrast would be observed using either technique.

**Fig. 4** Emission of (a) secondary electrons as a function of angle of incidence and (b) backscattered electrons as a function of atomic number

Due to their low energy, secondary electrons are typically collected to a photomultiplier/scintillator detector by placing a positive bias on the detector. For each pixel where the beam interacts with the sample, a signal is collected, processed, and displayed as an intensity level on a CRT (technically, a cathode ray tube; as a practical matter, a high resolution black and white or monochrome TV) screen. By coupling the scan of the beam to the scan of the CRT, a point by point correspondence exists between the surface roughness of the sample and the intensity observed on the screen. As the beam from the electron gun scans across a rough surface, the number of emitted secondary electrons continually changes, and the collected signal varies correspondingly. Thus, the intensity displayed on the CRT level changes from white (high-electron-emission) to black (low-electron-emission), and black and white contrast results in an image being produced in a manner exactly analogous to a black and white TV image.

The strength of the secondary electron signal often may be increased by tilting the sample slightly toward the detector. This results in more efficient collection of the secondary signal and can improve the signal-to-noise ratio without decreasing resolution. This may be helpful when trying to get the utmost resolution possible.

Backscattered Electrons. Imaging is somewhat different for primary backscattered electrons (BSE), which are incident electrons that have undergone elastic Rutherford scattering from the nuclei of the atoms that comprise the sample. For this signal, as the size of the nuclei gets larger (i.e., higher atomic number), the BSE emission increases in a nearly linear fashion (Fig. 4b). Thus, for a flat, polished sample containing numerous phases, where little or no contrast may be seen using secondary electrons, an image of high contrast may still be produced using the BSE signal. This is illustrated in Fig. 5, where a layered structure of various metals is imaged using secondary and backscattered electrons. While the layers can be clearly seen in the etched sample, it is only in the BSE image that the large difference in atomic number of the components is discerned. If a polished, homogeneous, single-phase sample were placed in an SEM (e.g., an unetched silicon wafer), little or no contrast would be observed using either technique.
Fig. 5 Example of a layered structure of copper, niobium, and zirconium. (a) Imaged using secondary electrons. (b) Imaged using backscattered electrons. Copper appears dark gray, zirconium is light gray, and niobium is white.

Unlike the secondary electrons that can be attracted to a detector, the high energy BSE signal can only be collected by placing a detector in a position where it is likely to be struck by the emitted electrons. The most advantageous position is directly over the top of the sample. If the detector is designed such that different signals can be collected from different sides of the detector (usually divided into halves or quadrants), enhanced imaging called topographical imaging is possible by skillful use of the signals. By manipulating the signals received by different halves or quadrants of the BSE detector, small surface features that are indistinct in a secondary image can be imaged very well with BSE (Fig. 6). Topographical imaging works best on single phase samples that are nearly flat and possess only slight surface roughness, such as pits or ledges.
Topographical imaging is possible due to the highly directional nature of the elastically scattered BSE signal. This characteristic, while useful in many investigations as was just mentioned, also can cause confusion when using the BSE signal to observe rough samples. In this instance, it should be remembered that the image obtained is not directly related to the degree of surface roughness (as is the case for the secondary image) but is a function of how that roughness is angled with respect to the detectors rather than the incident beam. When imaging a rough sample using BSE, the image can be altered radically by changing how the detected BSE signals are manipulated. The signals from all detectors should be added together when viewing BSE images from a rough sample. This should be remembered when using low-pressure or variable-pressure SEMs, where only a BSE signal is available for imaging in the low-pressure mode. This is especially true if the sample is not single phase, since the difference in BSE emission due to varying atomic number adds an additional complication. Having noted these potential problems, BSE imaging of rough surfaces often can still provide a rapid identification of areas that may deserve closer scrutiny. An example of this is shown in Fig. 7, where BSE imaging quickly reveals material of high atomic number embedded in the surface. This particular example is of a failed superalloy shaft. Lead from a nearby failed journal bearing, invisible in the secondary image, is readily apparent in the BSE image.
Characteristic x-rays are emitted due to collisions between incoming electrons and bound inner-shell electrons from the atoms that comprise the sample. If the incident electron has enough energy, the collision causes the ejection of the bound electrons. One way the atom decays to the ground state is by replacing the inner shell electron with a higher shell electron. Energy balance is restored by the emission of an x-ray whose energy is equal to the difference in energy levels of the electrons involved. Since the energy levels of all the electrons of all atoms are different, the energy of the emitted x-ray, termed the critical excitation energy, $E_c$, will be characteristic for that atom. By measuring the energies or the wavelengths of the x-rays emitted from a sample, the composition can be determined. The precision and accuracy of the determination depends on the type of x-ray detector used and how the analysis is carried out. Detectors that determine composition based on the wavelength of the emitted x-ray are termed wavelength dispersive spectrometers or WDS. Using WDS and pure elemental standards for the analysis, typical accuracies of 5% with a precision of 0.001 wt% can be obtained. However, these systems are the most expensive, and obtaining this level requires a fairly skilled operator. The more prevalent detectors determine composition based on x-ray energy. These energy dispersive spectrometers (EDS) are fairly inexpensive, easy to operate, and robust. A careful analysis using elemental standards can produce results of 5% accuracy with a precision of 0.1 wt%.

When conducting an x-ray analysis, it is important to realize that the area sampled is much larger than the size of the electron probe. This is due to spreading of the beam through collisions with the sample atoms, (Fig. 8). The total volume of the sample affected by the beam is termed the interaction volume. The shape of the interaction volume depends primarily on the accelerating voltage and the atomic number, Z, of the target material. Depending on these parameters, the shape may vary from being roughly semispherical for low voltages and/or high Z, to bottle-shaped, as shown in Fig. 8. It is clear that for the conditions of Fig. 8, x-rays may be emitted from the sample over an area considerably larger than the size of the electron probe. A reasonable estimate for the smallest area that can be sampled using x-rays is a 1 μm cube.
The condition of the sample surface and the orientation of the surface with respect to the detector greatly affect both the size of the interaction volume and the numbers obtained. Quantitative EDS analyses should only be conducted on flat, polished samples to reduce the possibility of absorption of low energy x-rays by the sample. Analyses from rough surfaces can vary greatly, depending on whether the site sampled is facing toward or away from the EDS detector. Thus, EDS data on rough samples should only be used for qualitative investigations.

X-ray data can be collected for a single point or for a designated area. Point analysis is useful to investigate the compositions of discrete phases, corrosion products, deposits present on a surface, and similar situations. An area analysis averages out point-to-point variations. In either case, it must be remembered that only the composition of a thin surface layer is analyzed, and the results obtained may not be representative of the bulk material.

The working distance at which the signal is collected is another critical factor in quantitative analysis. The EDS detector is usually sighted for best signal collection by the manufacturer at a certain working distance from the objective lens. Acquiring a signal from a sample not at this optimum working distance may result in low and/or inaccurate counts. The manufacturer of the microscope and the EDS system should be consulted to determine the optimum conditions for x-ray collection.

Finally, it should be remembered that a certain amount of overvoltage, approximately $1.5 \times E_c$ (the critical excitation voltage of the x-ray), is required to obtain a desired x-ray. For example, in order to obtain copper $k\alpha$ x-rays of $E_c = 8.04$ keV, the SEM accelerating voltage must be set greater than or equal to $(1.5 \times 8.04)$, or approximately 12.04 keV minimum. If the voltage is too low, these x-rays will not be generated or detected, even though copper may be present.

This leads to a critical point for the failure analyst. During a failure investigation, analyses are often conducted over a number of areas on a number of samples in different laboratories at different times. Before making comparisons regarding similarity or differences among x-ray spectra, one must consider whether the spectra were obtained under the same conditions, especially with respect to accelerating voltage.

Additional signals that can be monitored in an SEM are the specimen current and the electron backscattered pattern (EBSP). Similar in many respects to the specimen current, charge collection techniques are used extensively in the semiconductor industry (Ref 3) whilst the EBSP signal is used to determine crystal structure and orientation (Ref 4). More information concerning the use of these signals is available (Ref 5).

**Vacuum System**

All SEMS require a vacuum system to allow an electron beam to be produced and directed onto the target material. The constraints placed on the vacuum system depend on the type of SEM. With the advent of low-pressure and environmental SEMs, the design of vacuum systems has changed from a simple vacuum chamber to a multistage chamber with various levels of vacuum produced in different parts of the electron-optical column.

A high vacuum is desired in the gun chamber of all SEMs to minimize filament oxidation and prolong life. This is especially true in high resolution cold field emission SEMs, where oxidation of the tip can seriously reduce the emission of electrons. These SEMs have the most stringent vacuum requirements, and usually a circuit is supplied that allows the operator to heat the filament tip for a short period of time to degas any oxygen that adsorbs on the tip during operation. SEMs with LaB$_6$ filaments have the next highest requirements while tungsten filaments have the lowest requirements.

Isolation valves that separate the gun chamber from the sample chamber allow a high vacuum to be maintained in the gun area while samples are being inserted or removed. Separate pumps (such as titanium ion-getter pumps) attached to the gun continue to operate while the sample chamber is vented to atmosphere. These features are standard for field emission and LaB$_6$ guns and add extra complexity (and cost) to these systems. Separate pumps and an isolated gun chamber are often omitted from tungsten thermionic SEMs, which is one reason why they are the least expensive of the SEMs types.
sample insertion port can often be purchased as an accessory that allows the entire column to remain under vacuum with the electron gun in operation while a sample is inserted through an air-lock into the specimen chamber. Such ports can significantly improve productivity by eliminating the need to turn the microscope off and vent the chamber each time the operator wishes to examine a new sample.

The introduction of low-pressure and environmental SEMs in recent years has resulted in even more complex vacuum systems. Since these microscopes require a given pressure to be maintained in the sample chamber, an even greater sophistication is involved in the vacuum design. The ability to maintain a high vacuum in the gun region and a poor vacuum in the sample chamber is achieved using a series of apertures and vacuum pumps.

References cited in this section


Scanning Electron Microscopy

L.S. Chumbley, Iowa State University; L.D. Hanke, Materials Evaluation and Engineering Inc.

Specimen Preparation

Preparation of samples for observation using an SEM has undergone a revolution in recent years with the introduction of low-pressure and environmental SEMs. Prior to their introduction, SEM samples were required to be electrically conductive to enable the electrons from the beam to be carried away, thus preventing charging of the sample and poor imaging. Samples were also required to be dry in order to maintain the vacuum level in the sample chamber and prevent outgassing that would affect the electron beam and again result in poor imaging. These stringent conditions often required samples to be prepared by a variety of methods in order to obtain a useable image. However, nonconductive samples, both dry and moist, can now be imaged successfully depending on the type of microscope used. Thus, sample preparation has become microscope dependent. When purchasing an instrument, consider as an additional cost factor the types of samples to be examined and the necessity for (or elimination of) sample preparation equipment. For example, in failure analysis investigations, parts are often submitted that contain residual oil or grease or show nonconductive corrosion products. For cases such as these, the additional cost of a low-pressure SEM might be offset by the time and money saved in eliminating sample preparation. Furthermore, in some investigations, SEM examination of greasy, dirty, or corroded parts in the as-received condition may be desired to, for example, identify constituents in the corrosion products in situ.

For a conventional SEM, the only constraint placed on examining a dry, conductive sample is that the sample fit in the SEM chamber. The size of the chamber needed should be carefully considered before a purchase is made since a larger, more costly chamber may allow some samples to be examined directly without sectioning. For small samples that need to be secured in some type of resin for polishing and etching, conductive mounting materials are available for purchase. If a nonconductive mount such as normal bakelite is used for polishing, a conductive path from the sample to the SEM specimen holder or stage must be made to prevent sample charging. This is achieved using metal or carbon tape, or by painting a path on the side of the mount using a conductive solution such as Dag (Acheson Colloids Co.). (Dag is available from various manufacturers including Buehler Inc. of Lake Bluff, IL; Struers, Inc. of Westlake, OH, and Structure Probe of West Chester, PA).

If a nonconductive sample is to be examined using a conventional SEM, several methods exist to produce a conductive layer on the surface. The two most common are evaporation and sputter coating. Evaporation coating consists of heating a metal filament to a high temperature in a vacuum chamber to cause evaporation to occur. Atoms that leave the filament travel a line-of-sight trajectory to the surface of the sample where they are adsorbed. A thin coating consisting of a few monolayers of atoms can quickly result, producing a conductive layer across the entire surface of the sample. The heating is usually achieved in a wire filament by passing an electric current through it or, in a small metal sample, by using a levitation induction coil. Carbon evaporation is also possible using a slightly modified process. In this instance, a carbon filament is placed between two electrodes in the vacuum chamber. A current is rapidly introduced through the filament,
causing the carbon fibers to heat and burn out. A large amount of evaporated carbon is produced during the moment of filament burnout.

Evaporation methods are very easy and reliable provided a clean vacuum chamber is maintained, since degradation of the vacuum can cause the quality of the coatings to suffer. Typical metals used for coating are gold and platinum, with finer grained coatings being produced as the evaporation temperature increases. One drawback of metallic films is that the coating produced will be detected in any subsequent x-ray investigations. If chemical analysis of the sample is required, it is best to coat the sample with carbon, which is more difficult to detect using EDS and is less likely to obscure any additional elemental peaks. This is assuming, of course, that one is not wishing to determine the composition of carbon in the sample.

Sputter coating is similar in that the result is a few monolayers of a metal being produced across the surface of the sample. However, in this instance the metal atoms are not removed by evaporation but rather by bombardment from heavy argon ions in a manner analogous to sand blasting. In sputter coating, the sample is placed in a vacuum chamber beneath a metallic target, usually of gold or gold-platinum. After a vacuum is achieved, a partial pressure of argon is introduced into the chamber and a voltage placed across the target material and the sample. The target material is negatively biased with respect to the sample. The presence of the electric field causes an argon plasma to form, producing argon ions. These positive ions bombard the negative target, knocking off atoms of gold (or gold-platinum) that travel by line-of-site to the sample beneath and adhere to the surface. The thickness of the film depends on the time of deposition and the deposition rate that results in forming the plasma.

Many times in failure investigations, it is impractical or impossible to insert a sample into the SEM chamber. In cases such as these, the surface can still be examined at high magnifications using an SEM by making a surface replica of the surface. Production of a replica involves placing a thin sheet of a suitable polymeric material, such as cellulose acetate, on the surface to be examined (Fig. 9). The cellulose acetate is then soaked with acetone to soften the material, and capillary forces between the film and the surface pull the sheet tightly to the surface. Once the cellulose acetate has dried, the material is pulled away from the surface, producing a mirror replication of the surface in the form of the now somewhat stiff acetate. This film is made conductive either by sputter coating or evaporation, and the resultant sample is examined in the SEM. Many times this same technique can be used to remove small particles or inclusions from the surface of a sample. In this instance the particles adhere more tightly to the acetate film than they do to the substrate, and, when the film is removed, the particles can be found clinging to the surface of the acetate. Termed an extraction replica, this technique is useful for removing particles from a rough surface so that more accurate quantitative information can be obtained from the particles using x-ray analysis.
Fig. 9 Schematic showing method for producing extraction replicas and surface replicas for SEM examination

New liquid epoxy resins are also available for making surface replicas for SEM investigations. A small portion of the resin is mixed and applied to the surface. Some manufacturers market kits where the resin is mixed automatically when using an application gun. Different mixtures are available, providing different curing times at various temperatures. Upon drying, the polymeric material is simply peeled from the surface to be examined. Resolution of the cured resin is on the order of 0.1 μm and some mixtures are available that can be examined directly in the SEM.

As discussed in other sections of this Volume, more than one analyst may be interested in examining the specimen(s) involved in a failure investigation. Other interested parties should be given the opportunity to examine the specimen before it is prepared for examination in the SEM (or altered in any other way).

Scanning Electron Microscopy

L.S. Chumbley, Iowa State University; L.D. Hanke, Materials Evaluation and Engineering Inc.

Application of SEM in Fractography
Fractography is one of the most common applications of SEM. The large depth of focus, the possibility of changing magnification over a wide range, very simple nondestructive specimen preparation with direct inspection, and the three-dimensional appearance of SEM fractographs make the SEM an indispensable tool in failure studies and fracture research. Although the interpretation of microfractographs requires practice and understanding of fracture mechanisms, there are only a small number of basic features that are clearly recognizable and indicative of a particular mode of failure:

- Dimpled fracture, typical of overstress failures of ductile metals and alloys
- Cleavage facets, typical of transgranular brittle fracture of body-centered cubic (bcc) and hexagonal close-packed (hcp) metals and alloys
- Brittle intergranular fracture typical of temper-embrittled steel, in which fracture is due to segregation of an embrittling species to grain boundaries, to intergranular stress-corrosion cracking, or to hydrogen embrittlement
- Stage II striations, typical of fatigue failure

These typical forms of fracture surfaces are briefly described in the following section, “Typical Fracture Surfaces.” More extensive coverage on the interpretation of fracture surfaces is contained in the Section “Fracture” in this Volume. Image dependence on microscope type and operating parameters is discussed in the final section of this article.

**Typical Fracture Surfaces**

The appearances of fracture surfaces are directly related to the mechanism by which the failure occurs. If the surface remains unaffected and undamaged by events subsequent to the actual failure, it is often a simple matter to determine the failure mode by use of an SEM. In cases where the surface is altered after the initial failure, the case may not be so straightforward. A few typical examples that illustrate these points are shown subsequently. Other sections of this Volume provide extensive further discussion of these fracture mechanisms.

Ductile failures result when a material is plastically deformed, usually in tension, to the point of failure. In the early stages of deformation, small holes called microvoids form, usually around weak points in the material, such as inclusions or second phase particles. As deformation proceeds, the microvoids grow and begin to link (or coalesce), creating larger voids. This process continues until failure occurs. The result of this sequence of events can be observed on the fracture surface of a ductile failure, as shown in Fig. 10. Note the presence of round inclusions at the bottom of several coalesced voids, which appear as small depressions.

![Fig. 10 Typical fracture surface exhibiting ductile failure in a mild carbon steel](image)

**Fig. 10 Typical fracture surface exhibiting ductile failure in a mild carbon steel**

Brittle failures typically occur with very little plastic deformation near the formation of the crack prior to rupture. In a brittle fracture, the material may fail by cleavage along specific atom planes or directions, or it may fail due to separation of microstructure at specific sites caused by embrittling impurities or corrosive attack. Regardless of the reason, brittle failures can usually be identified by the sharp, angular appearance of the fracture surface. Figure 11 shows examples of three types of brittle fracture. Figure 11(a) shows the surface of a steel that was fractured below the ductile/brittle transition temperature. The cleavage of the material is apparent, with river markings indicating propagation of the fracture along nearly parallel sets of planes.
Fig. 11 Brittle fractures. (a) Fracture of mild carbon steel below the ductile/brittle transition temperature. Note the appearance of river lines on the faces of the cleavage surfaces. (b) Fracture of a soda-lime glass. Note similarity of river lines to those of (a). (c) Intergranular stress-corrosion cracking in brass. The grain structure is clearly revealed as the material fails along the grain boundaries.

This image should be compared to that of Fig. 11(b), which shows the fracture surface of a common soda-lime glass. Note that both brittle materials exhibit similar river lines marking propagation of the crack along converging lines and pointing back to the origin of the crack. However, the amorphous structure of the glass does not produce the flat cleavage surfaces apparent in the steel. Instead, extremely large, smoothly scalloped regions exist. These are often found in brittle fractures of glasses and polymers.

The brittle failure of the steel and glass is remarkably different from that of Fig. 11(c). Although Fig. 11(c) still exhibits a sharply angular structure common to brittle failures in metals, in this instance the appearance of the surface is due to failure along grain boundaries. The faceted appearance of the grains is quite distinctive and is indicative in this case of intergranular failure due to stress-corrosion cracking.

Mixed Mode Failure. In many instances, the fracture consists of both ductile and brittle regions. An example of such a mixed mode fracture, consisting of ductile tear and brittle cleavage, is shown in Fig. 12. This example is a good illustration of the difference in scale of the surface features associated with ductile and brittle failures.

Fig. 12 Mixed mode fracture in a mild carbon steel cooled just to the ductile/brittle transition

Fatigue Failures. Not all failures will exhibit the characteristic ductile or brittle features displayed in the previous images. This is especially true for fatigue failure where the growth of a crack occurs over an extended period of time as the piece is cycled from one stress state to another while in use. In fatigue, crack propagation occurs with each cycle, and the characteristic features associated with a ductile or brittle failure do not appear. Instead the surface of a classic fatigue failure will be marked by a series of striations, observed at high magnification, which mark the advancement of the crack front with each stress cycle. This is illustrated in Fig. 13 for an aluminum alloy, Fig. 14 for a Type 304 stainless steel, and Fig. 15 for a Grade 250 maraging steel.
Fig. 13 Ductile fatigue striations in aluminum alloy 2024-T3. (a) The uniformity of the crack-propagation process is well illustrated. There was little or no interaction of the fracture process with the inclusion within the rectangle. The long ridges are believed to be high-angle steps at subgrain boundaries. 650×. (b) Higher-magnification view of the region outlined by the rectangle in (a), showing the continuity of the fracture path through and around the inclusion. Fracture “wake lines” are visible and indicate that the direction of crack propagation was toward the lower right corner of the fractograph. These wake lines originate at the inclusion and run normal to the crack front in the direction of crack propagation. 2600×

Fig. 14 Fatigue fracture in type 304 stainless steel tested at room temperature. The vertical secondary cracks in (a) are grain-boundary separations. 39×. (b) The well-defined striations resulted from the planar slip characteristic of stainless steels. 810×
Fig. 15 Fatigue fracture in 18% Ni, grade 250, maraging steel tested at room temperature. (a) Extensive secondary cracking can be seen in a lower-magnification fractograph. 126×. (b) The cracking is clearly evident in a higher 1440× magnification fractograph. Secondary cracks formed at the roots of many fatigue striations. The range of stress intensity ($\Delta K$) at the crack tip was 76.9 MPa · $\sqrt{m}$ (70 ksi · $\sqrt{in.}$).

The striations that are formed at very low crack-growth rates (less than $5 \times 10^{-6}$ in. per cycle) can be difficult to resolve and often cannot be distinguished from the network of slip lines and slip bands associated with plastic deformation at and near the fracture surface. Under these circumstances, measurement of striation spacing is difficult, if not impossible to measure. In steels and other materials, fatigue striations that may form at ordinary crack-growth rates also are not always as well-defined as in aluminum alloys. This is the more common situation, because the fatigue striations are often obscured by subsequent damage to the crack surface as crack propagation proceeds. In this instance the surface of the fracture appears smeared and the fatigue striations may be difficult or impossible to see. The smeared nature of the fracture surface may still permit an identification of the failure as being fatigue related, especially if the suspected fatigue region can be compared to the point of final overload within the sample. An example of this is shown in Fig. 16(a), where the regions of fatigue and final ductile overload are easily identified.

Fig. 16 Fatigue failures. (a) Fatigue (upper right) changing to ductile failure (lower left). (b) Fatigue fracture of class 30 gray iron. (c) Single overload fracture of class 30 gray iron

Comparisons such as that shown in Fig. 16(a) often are the last resort for determining the nature of the failure mode. This is especially true in the fatigue of a brittle material such as cast iron, where the fatigue surface appears almost indistinguishable from the final failure (Fig. 16b and 16c). In cases such as this, identification of the true failure mode may be extremely difficult.

Dependence of Image on Microscope Type and Operating Parameters
A previous section noted that the appearance of any surface imaged with an SEM is influenced to a great extent by the type of SEM employed and the operating parameters used for the investigation. A few specific examples are shown below.

Low-Pressure Microscope. The biggest advantage of low-pressure microscopes is that they allow the imaging of nonconductive surfaces, eliminating the coating that would be required for clear imaging in a conventional SEM. This is illustrated by Fig. 17, which shows a section of PVC tube that failed due to fatigue. The image obtained shows excellent contrast without necessitating that the sample be coated in any way. The chamber pressure for this particular example was 0.3 torr. The importance of selecting a suitable chamber pressure is illustrated in the micrographs of Fig. 18, which present a series of images of nonconductive particulate taken under various conditions using a modern low-pressure SEM.

Fig. 17  Fatigue failure of a nonconductive polyvinyl chloride pipe imaged in the uncoated state using a low-pressure microscope
Fig. 18 Nonconductive material imaged using (a) conventional imaging, (b) low-pressure imaging with 20 Pa gas pressure, and (c) using 270 Pa gas pressure

Higher Resolution Field Emission Microscope. Field emission microscopes offer the ability to image samples at higher resolution than a conventional thermionic (LaB$_6$ or tungsten) microscope. Figure 19 shows a comparison of pearlite lamellae imaged using an inexpensive tungsten-filament microscope and a field emission instrument under identical operating conditions (15 kV; 16 mm working distance; moderate condenser lens strength). The field emission machine clearly produces a better image. However, this comes at an increase in instrument cost. The question often arises when
purchasing a new microscope: at what point is the increased cost justified? This question cannot be definitively answered for all samples and situations. Typical sample condition, operator expertise, and the information to be obtained all must be considered in purchasing to determine what the normal operating mode will be. For the example given in Fig. 19, the inexpensive tungsten-filament instrument still provides much of the same information as the (much more costly) field emission machine. If most investigations will be conducted at magnifications of 10,000× or below, the wisest choice may be to purchase the less expensive and easier to maintain thermionic machine.

![Fig. 19 Comparison of (a) conventional tungsten filament SEM to (b) field emission instrument. Samples imaged at 10,000×](image)

High Accelerating Voltage versus Low Voltage. Accelerating voltage also plays a large role in the appearance of the observed secondary electron image. As accelerating voltage increases, the interaction volume shape changes from semispherical to bottle shaped. Most secondary electrons are emitted from within 100 Å of the surface, so as the neck gets smaller and smaller with increasing voltage the relative number of secondaries coming from near the surface becomes less. Conversely, as accelerating voltage decreases, the secondary signal from the near-surface region increases. This results in an image with greater surface information than is possible at the higher voltage. This effect is illustrated in Fig. 20. The images were obtained using a convention tungsten-filament SEM operated under identical conditions except for accelerating voltage.

![Fig. 20 Affect of accelerating voltage on imaging. Sample is gold-coated ZnS powders imaged using (a) 25 kV and (b) 5 kV. Note enhanced surface information at 5 kV. Both 20,000×](image)

Surface films are also best evaluated at lower accelerating voltage. Figure 21 shows a defect area in an anodized coating on a titanium component for a medical device. Using light microscopy, a small white spot was visible on the blue
anodized coating. With the SEM at 20 keV, the spot was barely detectable. The defect area was easily characterized as a void in the anodized coating by inspection using a 2 keV accelerating voltage.

![Image](image1.png)

Fig. 21 Examination of a surface film at (a) 20 kV and (b) 2 kV

Figures 20 and 21 clearly show the advantages in imaging that can be obtained by going to lower voltages. However, it must be remembered that if one wants to simultaneously obtain chemical compositions using x-ray analysis, the accelerating voltage must be sufficient to excite these x-rays. Thus, it may not be possible to obtain the best imaging conditions while also obtaining the best x-ray signal. It should also be noted that in some instances, using higher accelerating voltages may be useful in obtaining images of features that are invisible using other techniques. Figure 22 is a good illustration of this. Optical observation of this painted surface revealed only a small surface bump, as did low voltage imaging using the SEM. Imaging using 30 keV, however, revealed microscopic particles under the surface at the bump that could be identified as iron. Thus, the higher voltage allowed identification of these near-surface contaminant particles when other methods failed.

![Image](image2.png)

Fig. 22 Near-surface contaminant particles visible using 30 kV accelerating voltage

Comparison and Recommendations

The preceding discussion is meant to convey that a large number of factors need to be considered when deciding which instrument is best suited for any particular application. In many failure applications, the need for SEM work is limited to observations below 10,000x magnification. Essentially, any SEM should be able to obtain good images at these magnifications. Purchase decisions, therefore, are usually based on other factors than mere magnification considerations. Types of samples examined, the size of the samples, information required, and so on, all play a role in choosing an
instrument, with cost usually being the overriding consideration. If 95% of the specimens examined are large, metallic
samples where only low magnifications are desired, an inexpensive W-filament machine equipped with an EDS detector
provides extremely good value for the amount of funds expended. As sample variety increases to include insulating
polymer and ceramic materials, a low-pressure SEM may be more advantageous. If higher resolutions are required for a
considerable amount of work, the field emission machines may need to be purchased.
The wide range of capabilities now available in SEMs makes it possible to purchase machines especially designed for
specific applications. Only the operator can determine which SEM features are necessary and which are optional. A
skilled operator with a thorough knowledge of the operation of the instrument may be able to obtain images beyond the
capability of a novice. Thus, ease of operation may be the final factor in determining the best instrument for any given
application.

Scanning Electron Microscopy

L.S. Chumbley, Iowa State University; L.D. Hanke, Materials Evaluation and Engineering Inc.

Acknowledgments

Paul Stevenson of Perkins Engines, Ltd. is gratefully acknowledged for his assistance in acquiring several images.

References

2.  D. Dingley and Baba-Kishi, Use of Electron Back-Scatter Diffraction Patterns for the Determination of Crystal
    Symmetry Elements, Scanning Electron Microsc., 2, 1986, p 386
    Press, New York, 1986
4.  B.L. Adams, S.I. Wright and K. Kunze, Orientation Imaging: The Emergence of a New Microscopy, Met Trans
    A, April 1993, p 819–831

Chemical Characterization of Surfaces

John G. Newman, Physical Electronics, Inc., Evans PHI Analytical Laboratory

Introduction

MANY ANALYTICAL TECHNIQUES are available for the study and characterization of surfaces. These techniques
provide data about the physical topography, physical properties, chemical composition, and chemical structure of the
surfaces under study. Most of these techniques are based on bombarding the surface with photons, x-rays, ions, neutrons,
or electrons and analyzing the radiation emitted and/or reflected from the surface. Other techniques use other interactions,
such as physical probing of the surface.
A few key techniques have become standardized and are available to most failure analysts, either through an in-house
laboratory or from an outside service laboratory. The most common analytical methods for chemical characterization of
surfaces are shown in Table 1. The techniques to be applied to a particular failure depend on the type and size of the
sample, the depth of analysis, the type of information sought, the ease of performing the analysis, the allowable
destruction of the sample in either preparation or analysis, and the cost/time required. The information required about the
surfaces in a failure analysis varies from failure to failure. No one technique can fully characterize a surface, but a full
characterization is seldom required to solve a particular problem.

### Table 1 Evaluation techniques for chemical characterization of surfaces

<table>
<thead>
<tr>
<th>Technique</th>
<th>Information</th>
<th>Analysis depth</th>
<th>Analysis area</th>
<th>Detection limit</th>
<th>Ease of use</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDS</td>
<td>Elemental</td>
<td>&lt;5 μm</td>
<td>&lt;1 μm</td>
<td>&lt;0.1 at.%</td>
<td>Easy</td>
</tr>
<tr>
<td>WDS</td>
<td>Elemental</td>
<td>&lt;5 μm</td>
<td>&gt;1 μm</td>
<td></td>
<td>Easy</td>
</tr>
<tr>
<td>AES</td>
<td>Elemental, chemical structure</td>
<td>&lt;5 nm</td>
<td>&gt;100 nm</td>
<td>&lt;0.5 at.%</td>
<td>Moderate</td>
</tr>
<tr>
<td>XPS</td>
<td>Elemental, molecular structure</td>
<td>&lt;2 nm</td>
<td>&lt;5 μm</td>
<td>&lt;1 ppm</td>
<td>Moderate</td>
</tr>
<tr>
<td>TOF-SIMS</td>
<td>Elemental, molecular structure</td>
<td>&lt;1 μm</td>
<td>&gt;1 μm</td>
<td>&lt;1 ppm</td>
<td>Moderate</td>
</tr>
<tr>
<td>FTIR</td>
<td>Chemical structure</td>
<td>&lt;5 μm</td>
<td>&gt;10 μm</td>
<td>&lt;100 ppm</td>
<td>Moderate</td>
</tr>
<tr>
<td>Raman</td>
<td>Chemical structure</td>
<td>&gt;1 μm</td>
<td>&gt;1 μm</td>
<td>&lt;0.1 at.%</td>
<td>Difficult</td>
</tr>
</tbody>
</table>

Note: EDS, energy-dispersive spectroscopy; WDS, wavelength-dispersive spectroscopy; AES, Auger electron spectroscopy; XPS, x-ray photoelectron spectroscopy; TOF-SIMS, time-of-flight secondary ion mass spectrometry; FTIR, Fourier transform infrared (spectroscopy)

Understanding the various analytical techniques allows an analyst to select the most appropriate method(s) to obtain the
data needed for each failure. In many cases, a combination of analytical techniques may be required to evaluate the
physical and chemical nature of the surface under study. This article covers the chemical characterization of surfaces by
Auger electron spectroscopy (AES), x-ray photoelectron spectroscopy (XPS), and time-of-flight secondary ion mass
spectrometry (TOF-SIMS). Chemical characterization of surfaces by energy-dispersive spectroscopy (EDS) instrumentation, which is commonly a module integrated with modern scanning electron microscopes, is discussed in
many other articles in this Volume; an overview of EDS and wavelength-dispersive spectroscopy is provided in the article
“Chemical Analysis of Metals in Failure Analysis” in this Volume. Fourier transform infrared spectroscopy is used more
extensively in the analysis of polymers (see the article “Characterization of Plastics in Failure Analysis” in this Volume).
Table 2 is a summary chart of techniques discussed in this article. In this brief review, it is not possible to develop all the
capabilities of each technique, but some of the more important attributes are highlighted for a preliminary insight into the
strength and usefulness of these techniques for chemical characterization of surfaces.

### Table 2 Summary chart of techniques discussed in this article

<table>
<thead>
<tr>
<th>Feature</th>
<th>AES</th>
<th>XPS</th>
<th>TOF-SIMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probe beam</td>
<td>Electrons</td>
<td>X-ray photons</td>
<td>Ions</td>
</tr>
<tr>
<td>Analyzed beam</td>
<td>Electrons</td>
<td>Electrons</td>
<td>Ions</td>
</tr>
<tr>
<td>Average sampling depth</td>
<td>5 nm</td>
<td>5 nm</td>
<td>2 nm</td>
</tr>
<tr>
<td>Detection limits</td>
<td>$10^{-3}$</td>
<td>$10^{-4}$</td>
<td>$10^{-6}$</td>
</tr>
<tr>
<td>Spatial resolution</td>
<td>10 nm</td>
<td>5–10 μm</td>
<td>150 nm</td>
</tr>
<tr>
<td>Information</td>
<td>Mostly elemental, SEM photos</td>
<td>Elemental, chemical</td>
<td>Elemental, molecular</td>
</tr>
<tr>
<td>Strengths</td>
<td>Ultimate small area analysis, imaging</td>
<td>Ease of use, quantification</td>
<td>Chemical and molecular analysis, imaging</td>
</tr>
<tr>
<td>Limitations</td>
<td>Semiconductive</td>
<td>Very few</td>
<td>Quantification difficult</td>
</tr>
<tr>
<td>Major applications</td>
<td>Semiconductors, electronics</td>
<td>All industries</td>
<td>Polymers, contamination, trace metal analysis</td>
</tr>
</tbody>
</table>

---

Chemical Characterization of Surfaces

John G. Newman, Physical Electronics, Inc., Evans PHI Analytical Laboratory
Overview of Surface Analysis

What Is Surface Analysis? The “surface” of a sample can mean different things to different people. To those performing analytical tests at the air/specimen interface of samples, the word surface can be anywhere from the top monolayer to as deep as several micrometers into the sample, depending on the depth of analysis of the technique that is being applied. Although there are scores of different techniques currently being used to study the surfaces of materials, only a fraction of these can be classified as true surface analysis tools that derive the majority of their analytical signal from the top few atomic layers. Three such techniques include AES, XPS—also known as electron spectroscopy for chemical analysis (ESCA)—and TOF-SIMS. The electron spectroscopy techniques of AES and XPS have depths of analysis, on average, of approximately 5 nm, while TOF-SIMS is even more surface sensitive, deriving the majority of its signal from the top 2 nm. Because these techniques are so surface sensitive, they are often employed in failure analysis or general surface characterization. Typical applications include identification of thin layers of contaminants on surfaces or at interfaces, evaluation of cleaning processes, and the identification of stains and discolorations. Successful failure analysis can often be a matter of piecing together the different bits of information that different techniques can provide. Each of these three surface analysis techniques provides a different view of the sample surface and thus, a different piece of the puzzle. This article describes the basic theory behind each of the different techniques, the types of data produced from each, and some typical applications. Table 2 summarizes the different features of these techniques to allow for at-a-glance comparisons. Also discussed are the different types of samples that can be analyzed and the special sample-handling procedures that must be implemented when preparing to do failure analysis using these surface-sensitive techniques.

Analytical Considerations. Proper design of experiment and the choice of samples for analysis are important factors in the likelihood of success when determining the cause of a failure. Failures can occur in the manufacturing of a device or some time after the device has been in use. In the ideal situation, good versus bad samples can be compared. This allows for differences in elemental concentrations and chemistries to be observed and related to their effects on device failure. Attempts are then made to link these differences to a known step in the manufacturing process or suspected contaminant in the end use of the product. The cause of failure may be obvious from the contaminants observed, or it may be extremely subtle. If a contaminant is found, but its root source is not known, then the next step, if possible, is to analyze samples and materials extracted at various steps in the manufacturing cycle or end use of the product. In other instances, a direct comparison of a failed material to a good sample is difficult or even impossible, such as in interface delamination problems, where the bonding failure does not occur on good samples. In these cases, data from the failed area should be compared to reference materials to infer what should or should not be present. Reference materials can include starting substrate materials (before and after cleaning processes); pure compounds and coatings used in the manufacturing of the device; dried solvents, oils, greases, and cleaning solution residues used in processing; and any material that may have come in contact with the device. In all cases, the surface analyst must rely on the knowledge of the process engineer to provide as much detail as possible regarding the nature of the failure and the history of the sample of interest. Doing analysis on samples where no processing and handling history are provided can result in the incorrect interpretation of the data and results that are very misleading.

Sample-Handling Issues. In no other class of analytical techniques is sample handling as critical an issue. Because AES, XPS, and TOF-SIMS provide information from the top few atomic layers, extreme care must be taken to keep the surface of interest from being covered up with extraneous contamination. Physically touching a sample with un gloved and even gloved hands is a common source of handling contamination. With ungloved hands, finger oils, salts, hand lotions, and other contaminants are easily transferred to touched surfaces. Many hand lotions contain siloxane, commonly referred to as silicone, as one of their main ingredients. Siloxanes are known to easily spread across surfaces. Thus, touching any part of a sample with siloxane-contaminated hands can lead to large areas of the sample becoming contaminated. Many assume that if sample handling is performed with gloved hands, their samples are safe from contamination. Unfortunately, the surfaces of many gloves are also laden with oils, salts and/or lubricants that can be transferred to the sample during handling (Ref 1). Even so-called clean-room gloves are not necessarily clean of such contaminants, because they are manufactured more to be particle-free than contaminant-free. Common surface contaminants on powder-free gloves include Si, Cl, Na, and Zn. Touching the area to be investigated with unclean tools can also introduce impurities onto the area of interest. In general, nothing should ever come in physical contact with the area of interest. Even the air surrounding the sample should be kept free of smoke and particles. Caution should also be taken to make sure that samples do not come in contact with common plastic bags during storage or shipping, because additives in the plastic can transfer to the sample and be detected with surface-sensitive tools. Clean, plastic petri dishes (polystyrene or polypropylene are common), glass containers, and even plain white typing paper are all acceptable ways of storing or shipping samples that will later be analyzed by one or more surface analysis methods. If tape must be used to secure samples during shipping, avoid using tapes that are laden with silicones, such as many conducting carbon tapes. Many clear “cellophane” tapes are fairly clean of silicones, but make sure to avoid placing the tape in the vicinity of the area of interest on the sample.
Sample Types. Most solid materials, if they are vacuum compatible and are of the proper geometric size to be accepted into the surface analysis instrument, can be analyzed by one or more of these surface analysis tools. This includes, but is not limited to, wafers, sheets, films, coatings, foils, chunks, powders, wires, tubes, printed circuit boards, computer chips, and other whole or partial pieces of devices. While surface analysis tools are used to analyze solid surfaces, dried residues from liquids and even low-vapor-pressure oils and greases can also be looked at. A minute drop or thin layer of the liquid or oil of interest can be placed on a clean substrate (silicon wafer, aluminum foil, etc.) and allowed to dry in air prior to its analysis.

One consideration, however, is the electrical property of the sample—is it conducting or insulating? In general, AES is better adapted to running conducting and semiconducting materials, while both conducting and insulating materials are readily analyzed by XPS and TOF-SIMS.

Reference cited in this section


Chemical Characterization of Surfaces

John G. Newman, Physical Electronics, Inc., Evans PHI Analytical Laboratory

Auger Electron Spectroscopy

Auger electron spectroscopy is a surface analysis technique used to determine the elemental composition of the top few atomic layers of a surface or exposed interface in a solid material. Auger electron spectroscopy can detect all elements except hydrogen and helium and can provide semi-quantitative information, with detection limits of 0.1 to 1 at.% for most elements. The technique is named after Pierre Auger, who first described the process in 1925. The most attractive attribute of AES is its ability to analyze extremely small features. Modern AES instruments with field-emission electron sources can produce secondary electron micrographs with spatial resolutions as small as 10 nm and can characterize sample features as small as 25 nm.

Auger electron spectroscopy is performed under ultrahigh vacuum conditions, using an electron beam typically in the 3 to 25 keV range. The AES process begins with electron bombardment of the sample material causing an atom to eject an inner-shell electron, thus forming a vacancy. A second electron from a higher shell fills this inner-shell vacancy in an energy gain process. This energy can then cause the ejection of a third electron, referred to as an Auger electron. The energy of the escaping Auger electron is analyzed by an electron spectrometer. Because each element has a unique set of electron energies surrounding the nucleus, it also has a unique set of Auger peaks, and the resulting spectrum provides a fingerprint of the probed surface. The kinetic energy of the Auger electrons is typically between 40 and 2500 eV. In this energy regime, electrons can travel only a short distance before interacting with other atoms and losing energy. This short distance is referred to as the escape depth of the electron. Escape depths range from 0.5 to 10 nm, depending on the kinetic energy of the emitted electron and the material being analyzed. It is this small range of escape depths that gives AES its surface sensitivity. Those electrons that are close enough to the surface to escape without loss of energy are detected as Auger electron peaks. Those electrons that lose energy before leaving the sample surface add to the background of the spectrum.

Scanning Auger microscopy is accomplished by scanning an electron beam across the surface of a sample while measuring resultant electron signals. This scanning process generates secondary electron microscopy (SEM) images, backscattered electron images, and Auger maps. Secondary electron microscopy images, which provide a topographic view of the sample by detecting low-energy electrons emitted from the surface, are used to locate specific areas for more detailed study. Backscattered electron images, involving higher-energy electrons that have undergone scattering processes before escaping from the sample, reveal atomic number contrast and crystallographic information. Auger maps, obtained by measuring the emitted Auger electron intensity while scanning the electron beam, reveal the lateral distribution of elements across the sample surface.

Because of the uncompensated loss of secondary electrons during electron bombardment, AES typically has a difficult time with insulating materials and, therefore, is primarily used for analyzing conducting and semiconducting solids. However, AES analysis of some inorganic insulating materials is also possible.

A competing process to Auger electron generation is that of x-ray emission, the analytical signal for the technique known as energy dispersive x-ray spectroscopy (EDS). In general, x-ray emission is more pronounced for heavier elements,
while the probability of Auger electron emission is greater from lighter elements. While AES can detect elements above He in the periodic table, EDS is usually limited to elements heavier than beryllium. Auger electron spectroscopy is also more appropriate for submicrometer particles, because the analyzed volume of material with AES is much less than that for EDS. In EDS, the x-ray emission is generated from regions a micrometer or more in width and depth, even when much-finer electron probes are used.

The nomenclature used to identify the various Auger peaks consists of listing the three electron shells involved in the formation of the Auger electron. For example, for low-atomic-number elements, the most probable transitions occur when a K-level electron is ejected by the primary electron beam, an L-level electron drops into the vacancy, and another L-level electron is ejected. This is referred to as a KLL Auger electron. Because of its excellent analytical spot size, AES is highly used in semiconductor, electronics, computer disk drive, and metallurgical applications where extremely small area analyses (spectra, depth profiling, or mapping) are required. The basic types of AES data, which are briefly described subsequently, include:

- Scanning electron micrographs
- Survey spectra
- Depth profiles
- Maps and line scans

For further information, see the Selected References at the end of this article for recommended reading on AES.

Scanning Electron Micrographs. Because the electron beam used in modern-day AES systems is similar to those used in scanning electron microscopes, AES systems can provide very high-spatial-resolution secondary electron micrographs. These images are used for the location and documentation of the areas of interest (Fig. 1).

**Fig. 1** Scanning electron microscopy photo of the surface of a 300-series stainless steel sample obtained from AES instrument. Field of view, 1 μm

Survey Spectra. Typically, the first piece of analytical data obtained in an AES experiment is a survey spectrum. A survey spectrum is a plot of AES signal intensity versus the kinetic energy of the detected electrons. Survey spectra show which elements are present and, after applying sensitivity factors (Ref 2), can be used to quantify the amount of each element detected. Survey spectra are typically displayed on a differentiated scale of signal intensity to better observe the small Auger peaks present on a large, sloping background (Fig. 2). In some instances, when excessive atmospheric or handling contamination obscures much of the underlying surface of interest, inert gas sputtering can be employed to remove at least some of the obstructing contamination layer. Care must be taken to prevent sputter removing too much of the sample, or the surface of interest will also be removed. Typically, only 0.5 to 3 nm is presputtered (relative to the sputter rate of SiO₂) for this type of “subsurface” analysis.
Fig. 2 Auger electron spectroscopy survey spectra obtained from features observed in secondary electron microscopy photo of stainless steel (Fig. 1). Point 1, large particle feature near center of Fig. 3. Point 2, small particle feature. Point 3, control point off of defects.

Depth Profiles. Compositional depth profiling is possible with AES with the aid of an inert gas sputter ion beam. The ion beam is used to slowly remove material in a controlled manner, thus exposing a new surface to be analyzed. By sequentially sputtering and taking AES data, a compositional depth profile may be generated that shows the AES signal for selected elements as a function of sputter time (depth). Exact sputter depths are often difficult to obtain, because exact sputter rates are seldom known, and, on rough surfaces, the sputter crater depths are difficult to measure. Therefore, sputter rates are often specified relative to a known sputter rate on a reference material, such as thermally grown SiO₂ on single-crystal silicon. Practical maximum profile depths are on the order of a couple micrometers or less. Compositional depth profiling is particularly effective for thin-film analysis and for solving materials problems that require identifying the thickness of a thin surface layer, the composition of a thin film deposit, the presence of interdiffusion between thin films, or the presence of contamination at an interface between two layers.

Maps and Line Scans. Because of its excellent spatial resolution, elemental mapping is one of the most common modes of operation with AES. A map is a two-dimensional display of Auger electron signal intensity for a select element from a given area on the sample surface. High intensities (bright areas) on the map indicate that more of that particular element is present at that point than at lower-intensity (darker) points (Fig. 3). Mapping can be an excellent way of detecting localized (unknown) contamination. After first obtaining survey spectra from the area of interest, Auger maps are obtained for the elements of interest. The maps are then studied to identify localized areas that have low signal for those elements. Additional spectra can then be obtained in these low-intensity areas to determine if other localized elements are also present. Line scans are similar to maps, except that the data is obtained in one direction only.
Fig. 3 Auger electron spectroscopy map of calcium contamination on stainless steel surface. Field of view, 1 μm

Reference cited in this section


**Chemical Characterization of Surfaces**

John G. Newman, Physical Electronics, Inc., Evans PHI Analytical Laboratory

---

**X-Ray Photoelectron Spectroscopy**

X-ray photoelectron spectroscopy, also known as ESCA, is a surface analysis technique that can be used to determine both the elemental and chemical composition of the outermost atomic layers of a solid material. With the exception of H and He, all elements can be detected. Detection limits are, on average, approximately 0.1 at.%; however many of the heavier elements can be detected down to about 0.01 at.% (100 ppm). By monitoring subtle shifts in the atomic binding energies of the emitted photoelectrons, chemical speciation can be obtained from both organic and inorganic materials. Similar to AES, the average depth of analysis is approximately 5 nm.

X-ray photoelectron spectroscopy is accomplished by flooding the sample with x-rays of a known energy (typically Mg Kα at 1253.6 eV or monochromated Al Kα at 1486.7 eV). Absorption of these x-rays by the sample atoms causes photoelectrons to be emitted. The kinetic energy of the emitted photoelectrons is measured with an electron spectrometer. To a first approximation, the kinetic energy is determined from the following equation:

\[
KE = h\nu - BE
\]

where \(KE\) is the measured kinetic energy of the emitted photoelectron, \(h\nu\) is the energy of the x-rays being used, and \(BE\) is the atomic binding energy associated with the emitted photoelectron. To maintain consistency in the spectral energy scale when using different x-ray energies, the data are displayed on a binding-energy, rather than kinetic-energy, scale. The binding energy associated with a peak is then used to establish its elemental identity and chemical state. The incoming x-rays penetrate microns into the surface of the sample. However, the escape depths for these low-kinetic-energy (less than 1500 eV) photoelectrons are less than 10 nm, thus making XPS an extremely surface-sensitive technique.

Quantification is possible with the use of elemental sensitivity factors (Ref 3) that have been determined empirically and found to be in agreement with the current theoretical models for quantification of XPS data.

If the sample is insulating, x-ray bombardment can create a positive-charge buildup on the surface of the sample, causing all of the spectral peaks to shift to apparently higher binding energies. To minimize this charging phenomenon, low-energy electrons or a combination of electron and ion floods (Ref 4) may be added to the sample surface. All peaks are then referenced to a peak of known energy, such as C because hydrocarbons are present on most surfaces. This charge correction method allows one to obtain useful chemical-state information from insulating samples.
Common materials analyzed by XPS include metals, lubricants, semiconductors, metal oxides, glasses, ceramics, catalysts, plastics/polymers, coatings/thin films, and paper. Because the technique is so flexible in its sample-handling capabilities, the industries served by XPS are quite varied. Application areas include bond pads, paint and other thin-film adhesion problems, cleanliness concerns, corrosion analysis, identification of debris and discolorations, characterization of polymer surface functionalization, lubricant thickness measurements, optical and other thin-film profiling, and determination of oxidation state and oxide thickness of alloys. Types of XPS data, as briefly described subsequently, include:

- Survey spectra
- High-resolution spectra
- Depth profiles
- Angle-dependent analysis
- Maps and line scans

See the Selected References at the end of this article for recommended reading on XPS.

Survey Spectra. Similar to AES, the first piece of analytical data obtained in an XPS experiment is typically a survey spectrum (Fig. 4). Survey spectra are acquired to determine which elements are present and, after applying sensitivity factors (Ref 3), can be used to quantify the amount of each element detected. Survey spectra are obtained using high-sensitivity analyzer conditions to maximize the possibility of detecting low-level species.

![X-ray photoelectron spectroscopy survey spectrum of stainless steel surface](image)

**Fig. 4** X-ray photoelectron spectroscopy survey spectrum of stainless steel surface

High-Resolution Spectra. Once it is determined what species are present on a surface, high-resolution scans are typically acquired on specific elements of interest to monitor binding-energy shifts that can provide chemical information. These scans are obtained using much higher energy-resolution conditions on the spectrometer compared to the survey spectra. Because the peak shapes obtained in this mode are more accurate, quantification is also better. It is on these high-resolution scans where curve-fits and other mathematical routines are performed to extract chemical-state information and quantification. An example of the type of localized chemical bonding information that can be obtained with XPS is shown in the high-resolution C spectrum of polyethylene terephthalate (PET) (Fig. 5). Three relatively large peaks and one smaller peak are observed. The largest peak in the spectrum, located at a binding energy of 284.8 eV, is assigned to hydrocarbon-type carbon (C-C or C-H) where C is bound only to other C or H. In PET, this peak is associated with the nonoxidized carbons in the aromatic ring. At about 286.3 eV, a C-O peak is observed that is characteristic of the ethylene carbons in PET that are singly bound to oxygen. Located at 288.7 eV is a peak associated with the carboxylate carbons.
(ester O=C-O) that are bound to two oxygens. A small satellite peak is also observed at about 291.5 eV. This peak is indicative of the aromaticity within the PET molecule.

**Fig. 5** X-ray photoelectron spectroscopy high-resolution spectrum of polyethylene terephthalate (PET)

Depth Profiles. Elemental and, in some cases, chemical depth profiling is possible with the use of an inert gas sputter ion gun. The ion gun is used to slowly remove material, thus exposing a new surface to be analyzed. By sequentially sputtering and taking XPS data, a compositional depth profile can be generated (Fig. 6). If the chemical matrix of the sample is not severely damaged by the sputter ion beam, chemical information may also be obtained as a function of depth. However, many oxides and especially organics are very susceptible to chemical modification during ion bombardment. Similar to AES, exact sputter rates are often difficult to determine. Therefore, sputter rates are often specified relative to a known sputter rate on a reference material, such as thermally grown SiO₂ on single-crystal silicon.

**Fig. 6** X-ray photoelectron spectroscopy compositional depth profile of stainless steel

Angle-Dependent Analysis. Angle-dependent XPS is a method for nondestructively analyzing the in-depth chemical gradients in the top surface layers (<10 nm) of a smooth sample. By changing the sample tilt with respect to the energy analyzer (takeoff angle), the effective depth of analysis can easily be changed. This technique can be useful in cases where ion bombardment is known to modify the chemical makeup of the material being probed. It has been greatly used...
for studying such things as the depth of modification of plasma- and corona-treated polymers, the oxidation of metals and semiconductors, polymer-metal adhesion, additive surface migration, and the lubrication of various materials.

Maps and Line Scans. Many newer-generation XPS instruments are capable of obtaining spatially resolved information from a sample surface by acquiring photoelectron maps and line scans. A photoelectron map is a two-dimensional display of photoelectron intensity for a specific element from a given area on the sample surface. High intensities (bright areas) on the map indicate that more of that particular element or chemistry is present at that point than at lower-intensity (darker) points. Line scans are similar to maps, except that the data is obtained in one direction only. Spatial resolutions for state-of-the-art instruments are on the order of 5 to 10 μm.

References cited in this section


---

**Chemical Characterization of Surfaces**

John G. Newman, Physical Electronics, Inc., Evans PHI Analytical Laboratory

**Time-of-Flight Secondary Ion Mass Spectrometry**

Time-of-flight secondary ion mass spectrometry is an analytical technique that uses a high-energy, primary ion beam to probe the surface of a solid material. The instrument is typically operated in the “static” mode for obtaining elemental and molecular-chemical information from both organics and inorganics. In this mode of operation, the sample integrity and chemistry are preserved by applying extremely low primary ion doses (less than $1 \times 10^{12}$ ions/cm²) during the entire experiment. This ensures that roughly less than 0.1% of the surface atoms or molecules are ever struck and damaged by the primary ion beam. Time-of-flight secondary ion mass spectrometry is the most surface sensitive of the surface analytical techniques, with a depth of analysis of only about 2 nm. The technique has extremely good detection sensitivities, with detection limits for most elements in the parts-per-thousand to parts-per-million range. By using a finely focused ion beam (typically Ga or In) it is also possible to record the lateral distribution of chemical species across a surface with micron to submicron spatial resolution. Because secondary ion intensities vary dramatically from element to element and are highly dependent on the matrix from which they are sputtered, quantification with TOF-SIMS can be extremely difficult. Thus, in most cases, the technique is used more for qualitative purposes than for quantitative analyses. However, in select cases where appropriate standards are available (e.g., metals on silicon), accurate quantification can be performed (Ref 5).

When elemental depth profiles are required, most TOF-SIMS systems can also be operated as a “dynamic” SIMS instrument. In the dynamic mode, very high primary ion doses are used in order to obtain in-depth information. However, in this mode of operation, very little chemical specificity can be gleaned, due to the destructive nature of the primary ion beam. Therefore, the following discussion only focuses on the static mode of operation.

In a TOF-SIMS experiment, the desorption/ejection of “secondary” ions from the surface of a material is initiated by a short pulse (~1 ns) of primary ions that impinge on the surface at high angles of incidence. The momentum transfer from the primary beam to the solid initiates a “collision cascade” within the solid, much like a microscopic billiard game. A portion of this momentum is redirected back toward the surface, resulting in the ejection of atomic and molecular ions. Greater than 90% of these secondary ions originate from the outermost one to two layers of the solid, thus defining TOF-SIMS as an extremely surface-sensitive technique. By applying a potential between the sample surface and the mass analyzer, the desorbed secondary ions are extracted into a time-of-flight mass spectrometer where their masses are separated in flight time, with very high accuracy, based on their mass-to-charge ratio (m/z). The resulting mass spectrum, typically in the 1 to 2000 dalton range, allows for the unambiguous identification of chemical moieties that can often be correlated to the original surface structure. An example of the type of chemical/molecular information that can be obtained from secondary ion mass spectra is shown in the positive ion spectrum of PET (Fig. 7). The spectrum shows a variety of large molecular cluster ions that are identified as fragments of the PET polymer chain. This provides information on the long-range molecular makeup of the sample as well as providing a unique fingerprint for this material.
On insulating materials, the loss of secondary electrons during ion bombardment can lead to a positive charge buildup and loss of signal from the sample surface. Therefore, auxiliary electron sources are often used to supply electrons to the sample surface to help neutralize the excess charge. Because TOF-SIMS can easily handle insulating as well as conducting materials, the types of samples analyzed by TOF are often similar to those analyzed with XPS. However, because of its much smaller analytical probe size, TOF-SIMS systems can analyze features as small as a few micrometers in width and are commonly used for mapping distributions of elements and molecules across surfaces. Application areas for TOF-SIMS include organic and inorganic contaminant identification, cleaning studies, surface segregation and modifications, corrosion, discolorations, drug distribution, diffusion studies, and chemical characterization. Two types of secondary ion data are simultaneously obtained in a TOF-SIMS experiment:

- A total-area mass spectrum
A total secondary ion image

As the focused primary ion beam is digitally rastered across the sample surface, complete mass spectra of ion intensities versus mass-to-charge ratio are obtained at every pixel (256 × 256) within the raster. The summation of these spectra produces a total area mass spectrum (Fig. 8). From the same acquisition, a total ion image is also obtained that shows the lateral distribution of secondary ion signals from across the area of analysis (Fig. 9). High intensities (bright areas) on the image indicate that more secondary ion signal is present at those points than at lower-intensity (darker) points. The variation in secondary ion intensity can be due to topographical effects or from differences in chemical composition. The total ion image is used to selectively analyze the mass spectra (chemistries) from areas that show differing amounts of brightness. Conversely, the total ion mass spectrum can be used to select specific elements or molecules for display in secondary ion maps that show the relative localized abundance of these species. Because both positive and negative secondary ions are created during ion bombardment, two separate experiments must be performed for complete characterization of a given sample.

Fig. 8  Time-of-flight secondary ion mass spectrometry positive ion spectrum of stainless steel surface
Example: Stainless Steel Analysis

To exemplify the strengths of AES, XPS, and TOF-SIMS, data obtained from the surface of a slightly corroded stainless steel sheet are presented for each technique. Each method provides a different view of the surface and a different piece of the sample characterization puzzle. Note that in these examples not every type of analysis possible is shown for each technique (e.g., survey spectra and Auger maps but no depth profile are shown for AES).

Example 1: AES Analysis of Stainless Steel. Figure 1 shows a high-magnification secondary electron micrograph (100,000×) obtained from an AES instrument on the as-received surface of a piece of 300-series stainless steel sheet. The micrograph shows a series of what appear to be extremely small particles or defects, most of which are only about 25 nm in diameter. This SEM image was used to define points on which AES was to be performed. In this case, one large and one small particle (points 1 and 2, respectively) were analyzed along with a control region (point 3) off of the defects. Prior to analysis, a light sputter etch (approximately 1.5 nm removed) was performed to remove some of the atmospheric contamination to better observe the underlying stainless components. All three of the survey spectra (Fig. 2) show the presence of the expected C, O, Cr, Fe and Ni from the air-exposed stainless steel surface. In addition, the spectrum from point 1 (larger defect) shows the presence of a large amount of Ca along with a smaller Na peak. The AES spectrum from the small defect (point 2) shows the presence of S and Na contaminants but also shows a much different Fe to Cr peak ratio than the spectrum from the control area (point 3). Quantification of the elements detected from these three points is shown in Table 3. Note that for each point, the concentrations are normalized to 100% in an assumption that all elements were detected. Auger maps were then acquired for several of the elements of interest, including Ca, Cr, Fe, and O. The map for Ca is shown in Fig. 3, which reveals a lateral distribution of Ca within a 1 μm field of view. Although scores of particle defects were observed in the SEM image, only a few of them appear to have significant Ca, as shown in the Ca...
Auger map. Most of the smaller particles correspond to larger concentrations of Fe, compared to the background area (point 3, Fig. 2). This indicates that localized corrosion has occurred within these 25 nm-sized defects, with an obvious potential cause being the presence of Na.

### Table 3 Auger electron spectroscopy atomic concentrations (at.%). 1.5 nm sputter removed

<table>
<thead>
<tr>
<th>Sample</th>
<th>C</th>
<th>O</th>
<th>Na</th>
<th>S</th>
<th>Ca</th>
<th>Cr</th>
<th>Fe</th>
<th>Ni</th>
</tr>
</thead>
<tbody>
<tr>
<td>Point 1, large defect</td>
<td>13.1</td>
<td>30.2</td>
<td>3.5</td>
<td>nd</td>
<td>38.0</td>
<td>3.4</td>
<td>10.0</td>
<td>1.9</td>
</tr>
<tr>
<td>Point 2, small defect</td>
<td>8.6</td>
<td>35.2</td>
<td>5.6</td>
<td>0.5</td>
<td>nd</td>
<td>9.7</td>
<td>37.9</td>
<td>2.6</td>
</tr>
<tr>
<td>Point 3, control area</td>
<td>21.4</td>
<td>33.4</td>
<td>nd</td>
<td>nd</td>
<td>nd</td>
<td>16.7</td>
<td>24.1</td>
<td>4.5</td>
</tr>
</tbody>
</table>

nd, not detected

Example 2: XPS Analysis of Stainless Steel. Figure 4 shows a survey spectrum obtained from a millimeter-sized area of the as-received 300-series stainless steel. Several elements are observed, including C, N, O, Si, Cr, Fe, and Ni. Integrating the area under the peaks and applying empirically derived sensitivity factors yields the atomic concentration values shown in Table 4.

### Table 4 X-ray photoelectron spectroscopy atomic concentrations (at.%)

<table>
<thead>
<tr>
<th>Sample</th>
<th>C</th>
<th>N</th>
<th>O</th>
<th>Si</th>
<th>Cr</th>
<th>Fe</th>
<th>Ni</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stainless steel</td>
<td>44.6</td>
<td>1.6</td>
<td>40.2</td>
<td>0.1</td>
<td>0.9</td>
<td>12.5</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Although a certain amount of organic contamination is expected due to atmospheric exposure, the amount of C found on the as-received surface is relatively high. Some of this organic may have been introduced from surface or vapor-phase transfer of plastic additives or contaminants from the plastic bag in which the steel was stored. The C high-resolution spectrum (Fig. 10) shows a C peak that can be mathematically fit into at least three types of carbon, with energies (but not peak ratios) similar to that of the spectrum of PET (Fig. 5). The largest peak, occurring at 284.8 eV, is from hydrocarbon-type carbon. A second peak, located at 286.3 eV, is attributed to either C-O (ether or hydroxyl) or possibly C-N (organic-based N), because a small amount of N is observed in the survey spectrum. At 288.4 eV, a peak typically associated with carboxylate chemistry (O=C-O) is also observed. Thus, the XPS reveals which organic functional groups are present on the stainless steel and suggests possible contaminants. However, further characterization with a technique such as TOF-SIMS would be needed before definite conclusions should be drawn regarding the identity of the organic.
The ratio of Cr to Fe is often used as an indication of the “goodness” of passivation (Ref 6), with ratios of 1 to 2 being optimal (Ref 7) for well-passivated stainless steel. Clearly, this stainless steel sample has a very poor Cr to Fe ratio. Well-passivated stainless steel contains a thin Cr oxide layer as the outermost surface, protecting the underlying metallic Cr, Fe, and Ni from becoming oxidized. If the Cr oxide surface layer is performing properly, the XPS high-resolution spectral analysis should be able to see through the passivation layer and detect the presence of subsurface metallic Fe that has been protected from oxidation. On this particular stainless steel sample, little, if any, metallic Fe is observed (Fig. 11). Figure 12 shows a high-resolution Fe spectrum obtained from a well-passivated stainless steel. Note that the Fe metal component on the well-passivated steel is much more prominent compared to this “corroded” steel surface.
Fig. 12 X-ray photoelectron spectroscopy high-resolution iron spectrum obtained from well-passivated stainless steel surface

Figure 6 shows the compositional depth profile of the 300-series stainless steel sample. The sputter rate for the profile, relative to the sputter rate of SiO$_2$, is 2.1 nm/min, with data points collected every 0.8 nm. The Fe and Cr energy windows for the first several cycles of the profile are shown in Fig. 13 and 14, respectively. Both graphs show primarily oxide (higher binding-energy peaks) for the first few sputter cycles, after which the metallic form (lower binding-energy peaks) becomes dominant. These binding-energy shifts allow for easy mathematical separation of the oxidized and metallic forms within the profile (Ref 8, 9). The rapid decrease in surface C in the first half-minute of sputtering indicates that this contaminant is present as a thin layer on the surface of the stainless steel rather than in thick, uneven patches. Just under the C layer, Fe (oxide) and O become the primary components, with relatively low amounts of Cr being observed. Below the Fe oxide, a thin Cr oxide-enriched layer is detected, under which the bulk constituents of the stainless steel are found. On a well-passivated steel surface, a Cr oxide layer, rather than Fe oxide, would dominate the outermost surface region. The XPS results on this sample are consistent with those of the AES that found localized regions of corrosion occurring.

Fig. 13 X-ray photoelectron spectroscopy montage display of iron in the first eight sputter cycles of the depth profile (Fig. 6)
Fig. 14 X-ray photoelectron spectroscopy montage display of Cr in the first nine sputter cycles of the depth profile (Fig. 6)

Example 3: TOF-SIMS Analysis of Stainless Steel. The positive and negative mass spectra obtained from a 50 μm area on the surface of the as-received stainless steel sample are shown in Fig. 8 and 15. The positive ion spectrum shows the presence of the expected stainless components (Fe, Cr, Ni, and Mn) along with a variety of other species. Inorganic impurities include Na, Mg, K, and Ca while the organic contaminants observed include generic hydrocarbon fragments and polydimethyl siloxane (PDMS), often referred to as silicone. (Note that XPS and AES found little, if any, Si on the stainless steel surface.) A characteristic PDMS fragmentation pattern is seen with peaks at positive ion masses 73, 133, 147, 207, 221, 281, 295, 325, and 355. Between 1000 and 2000 daltons, an oligomer distribution for the PDMS is observed that provides an indication of the molecular weight of this contaminant. When looking at TOF-SIMS spectra, two things should be kept in mind. First, this technique is only sampling the outer few atomic layers of the sample and thus highlights any impurities present on the surface. Second, secondary ion yields play a huge role in how intense a peak for a particular specie appears. For example, on this steel sample, relatively large peaks for alkali and alkali earth impurities are observed. However, these elements have some of the highest positive secondary ion yields, because they all have relatively low electron affinities. Thus, they give strong peaks, even though their concentrations may be very low.
Fig. 15 Time-of-flight secondary ion mass spectrometry negative ion spectrum of stainless steel surface. Positive ion spectrum is in Fig. 8.

The negative ion spectrum provides complimentary information, showing the presence of several species not observed in the positive spectrum. Inorganic impurities seen on the stainless steel surface include O, OH, F, Cl, and NO$_x$ (nitrate). Note that the negative ion spectrum shows little evidence for the stainless steel matrix material, because the transition metals have relatively poor negative secondary ion yields. Organic contaminants include general low-mass hydrocarbon fragments; PDMS at masses 59, 60, 75, 119, 149, and 223; and fatty carboxylates (masses 255, 281, and 283). Possible sources of the fatty carboxylate contaminant include finger oils, lubricants, and polymer additives. The finding of a carboxylate on the steel is consistent with the XPS carbon high-resolution spectrum. However, it is clear from the TOF-SIMS data that there are multiple sources of organic contamination on the surface, not just one.

Another important feature of TOF-SIMS instrumentation is its ability to perform high-mass resolution analyses. In many cases, peaks with the same nominal mass must be separated and their masses accurately measured for proper peak identification. For example, Fig. 16 shows an expanded scale of the positive ion mass spectrum centered around mass 63. Two peaks are observed that are separated by less than 0.1 mass unit. From their exact masses, the peaks can be positively identified as the 63 isotope of Cu and a C$_5$H$_3$ fragment. This capability is especially important for separating and identifying inorganic masses from organic contaminants that are typically present on a surface.
Figure 9 is the total positive ion image from the steel surface and shows the variation in secondary ion signal from the 50 μm square area where the spectra were collected. The point-to-point variation in signal can be caused by differences in chemistry or can be due simply to topographical effects, because sputter and secondary ion yields can be altered by surface morphology and changes in the effective angle of incidence. The fact that topography effects can be observed in the total ion image can be very useful in locating small surface features of interest. Secondary ion images for select elements, such as Cr, Fe, and Na are shown in Fig. 17(a–c). The Fe and especially the Cr maps are very similar to the total ion image. However, note that the vertical dark line or “scratch” seen in the left side of the Cr image is nicely filled in with a strong Fe signal. This indicates that the protective Cr passivation layer has been lost in this region and thus is an active site for corrosion to occur. The Na image (Fig. 17c) reveals both small and larger regions of Na contaminating the surface. The presence of ionic species on steel can be especially detrimental, because localized areas of electrochemical activity can be created that are known to accelerate corrosion (Ref 10).
Fig. 17  Time-of-flight secondary ion mass spectrometry images of 50 by 50 μm stainless steel surface area. (a) Map of Cr (b) Map of Fe, (c) Map of Na

Summary of Example of Stainless Steel Analysis with AES, XPS, and TOF-SIMS. The previously mentioned results show three separate analytical views of a stainless steel surface. Auger electron spectroscopy provided the best spatially resolved data and showed the presence of localized corrosion. X-ray photoelectron spectroscopy provided a view of the average surface chemistry and substantiated the AES results by showing that the Cr to Fe ratio was relatively low and that the expected Cr oxide passivation layer was lacking on this surface. The XPS also showed a lack of metallic Fe indicating that the surface was more fully oxidized than expected. Finally, the TOF-SIMS showed what organic contaminants, as well as other low-level impurities, were present in the top few atomic layers.

References cited in this section


Chemical Characterization of Surfaces

John G. Newman, Physical Electronics, Inc., Evans PHI Analytical Laboratory

Acknowledgments

The author would like to thank the rest of the Evans PHI and Physical Electronics Laboratory staff for their help in the preparation of this work.

Chemical Characterization of Surfaces

John G. Newman, Physical Electronics, Inc., Evans PHI Analytical Laboratory

References


Selected References

General


AES, XPS, and TOF-SIMS


Quantitative Fractography

Arun M. Gokhale, Georgia Institute of Technology

Introduction

A FRACTURE SURFACE is generated by the deformation and fracture processes operative at numerous length scales ranging from nanometers to millimeters. The morphology and topography of the fracture surface are generally dictated by the material microstructure and active failure mechanisms. Therefore, quantitative characterization of the fracture surface
geometry, that is, quantitative fractography, can provide useful information regarding the microstructural features and failure mechanisms that govern the material fracture. The objective of quantitative fractography is to describe the geometric attributes of the topographic and microstructural features present in the fracture surface in quantitative terms, such as number, size, length, area, spacing, and orientation.

In the 1940s, fracture surfaces were studied using light microscopy (Ref 1). Such observations were useful for qualitative description of fracture surface topography. However, low magnifications and depth of field limited the utility of the technique for quantitative characterization. Development of the scanning electron microscope (SEM), and more recently, of powerful digital image analysis equipment, has led to significant advances in quantitative fractography. Numerous quantitative correlations between the material properties such as strength, ductility, toughness, and fatigue life, and the quantitative fractographic parameters have been reported in the literature during the past decade or so (Ref 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16). Modern quantitative fractographic techniques can be divided into the following categories:

- Profilometry-based quantitative fractographic techniques
- SEM-based quantitative fractographic methods
- Three-dimensional fracture surface reconstruction methods
- Internal fractographic techniques

Fracture profilometry concerns characterization of fracture profiles generated by intersections of the fracture surface with metallographic planes (Ref 17, 18, 19, 20, 21, 22). A fracture profile can be observed by using a standard metallurgical light microscope or an SEM. An important advantage of fracture profilometry is that the underlying microstructure can be observed simultaneously and, therefore, it can be correlated to the fracture path. Nonetheless, a fracture profile is essentially a lower-dimensional section through a nonplanar fracture surface present in the three-dimensional (3D) space. Consequently, the geometric characteristics of the features observed in a fracture profile do not necessarily represent the true values in the corresponding fracture surface. Therefore, it is essential to use unbiased statistical stereological relationships (Ref 23, 24) for estimation of the geometric attributes of the features in the fracture surface from the measurements performed on the fracture profiles. Significant progress has been made in the development of practical, unbiased, and efficient stereological techniques during the last decade or so (Ref 25, 26, 27, 28, 29). Several geometric parameters that quantify various aspects of fracture surface geometry can be estimated from the profilometric measurements by using general and unbiased stereological relationships. Characteristics such as fracture surface roughness (Ref 21, 22), fracture surface anisotropy (Ref 4, 30), extent of overlaps (Ref 4), average asperity height (Ref 6, 31), fraction of fracture path through different constituents or phases (Ref 32), and fracture surface fractal dimension (Ref 33, 34) can be estimated from profilometric data.

A fracture surface can be conveniently observed using an SEM. In most cases, no special specimen preparation is required for SEM fractography. Fractography by SEM is particularly useful for quantitative characterization of specific features (for example, dimples, striations, pores, facets, inclusions, etc.) present in the fracture surface (Ref 35, 36). The features of interest are characterized via estimation of their number density, average size, shape, spacing, orientation, and the distributions of these parameters, as required. Fractography by SEM is also useful for estimation of the fractions of fracture surface area generated due to different failure mechanisms, such as cleavage fracture, ductile fracture, and intergranular fracture (Ref 2, 37, 38). It is, however, important to recognize that a SEM fractograph is essentially a projected image of the nonplanar fracture surface present in the 3D space. Clearly, the geometric characteristics measured from such a projected image do not necessarily represent the corresponding true values in the fracture surface. Thus, it is essential to use general stereological relationships to estimate the true feature-specific fractographic attributes from the measurements performed on the SEM fractographs.

It is possible to characterize the true 3D geometry of a fracture surface by using stereo-pair based imaging (Ref 39, 40, 41, 42, 43, 44, 45, 46, 47). Modern image processing techniques permit reconstruction of the 3D fracture surface and its quantitative characterization using stereo-pair SEM data in an efficient manner (Ref 41, 42, 43). The basic image analysis techniques are described in detail in the article “Image Analysis” in Materials Characterization, Volume 10 of ASM Handbook. The electron backscatter diffraction (EBSD) technique of crystallographic orientation imaging and its combination with stereo-pair based SEM fractography is useful for characterization of crystallographic orientations of the features present in the fracture surface (Ref 46). Direct simultaneous observations of a nonplanar fracture surface in 3D space and the 3D microstructure just beneath it can be extremely useful in understanding and quantifying the failure modes and the relationships between the microstructure and the fracture path. Such observations are possible using a combination of SEM fractography, serial sectioning, and image analysis (Ref 48, 49).

Internal quantitative fractography concerns quantitative characterization of the microstructural damage features such as microcracks, particle cracks and debonds, microvoids, delaminations, and so forth and their evolution in the 3D microstructure as a function of stress, strain, stress state, number of stress cycles, and so on (Ref 50, 51, 52, 53, 54, 55, 56, 57). Quantitative characterization of the microstructural damage evolution is useful for understanding the local fracture processes at lower length scales and their contributions to the global fracture. Such data are required for
development of internal state variables-based predictive fracture models that are applicable under multiaxial loading conditions (Ref 58, 59). However, the internal quantitative fractographic parameters can be estimated by using the same stereological methods that are available for characterization of 3D microstructures from metallographic sections and projections. Therefore, the internal quantitative fractographic techniques are covered in the article on “Quantitative Metallography” in Metallography and Microstructures, Volume 9 of ASM Handbook.

Different categories of quantitative fractographic techniques are described in the following sections in this article. In each case, sufficient methodological details, governing equations, and practical examples are given. The next section is devoted to the fractographic techniques that are based on fracture profilometry. It is followed by a section describing methods based on SEM fractography. The final section addresses procedures for 3D fracture surface reconstruction.

References cited in this section


---

**Quantitative Fractography**

Arun M. Gokhale, Georgia Institute of Technology

---

**Profilometry-Based Quantitative Fractography**

A fracture profile is a tortuous irregular line generated by the intersection of a fracture surface with a sectioning plane. Several experimental methods are available for generation of fracture profiles. These experimental methods yield the \((x, y)\) coordinates of the set of closely spaced points in the fracture profile. From these basic data, important geometric attributes of the fracture profile can be computed. Stereological equations can be then used for unbiased statistical estimation of the geometric attributes of the corresponding fracture surface from the computed fracture profile parameters. The experimental methods for profile generation, computation of fracture profile parameters, and the stereological procedures for estimation of fracture surface attributes from the profilometric data are described in the following subsections.

**Experimental Methods for Profile Generation**

Physical metallographic sectioning of the specimen, sectioning of fracture surface replicas, and nondestructive profile generation by using stereo pairs and other methods are the important experimental techniques used for generation of fracture profiles. These methods are described as follows.

Metallographic Technique for Fracture Profile Generation. Depending on the angular orientation of the sectioning plane with respect to the fracture surface, it is possible to generate horizontal (Ref 8, 58, 59), slanted (Ref 60), conical (Ref 61), and vertical section (Ref 17, 18, 21, 22) fracture profiles. The vertical section fracture profiles are most widely used as they are easy to generate using well-known metallographic procedures, and general stereological equations are available for estimation of fracture surface characteristics from the measurements performed on vertical section fracture profiles (Ref 4, 21, 22). Figure 1 shows a schematic fracture surface, a sectioning plane, and a schematic representation of the fracture profile generated by intersection of the fracture surface with the sectioning plane. Figure 2(a) shows a fracture profile generated from the fracture surface of a tensile test specimen of a metal-matrix composite (MMC) containing unidirectionally aligned alumina fibers in the matrix of an aluminum alloy. Figures 2(b) and 2(c) show a fracture profile from the tensile fracture surface of a low-alloy steel specimen. These fracture surfaces were electroplated with copper. As illustrated by Fig. 2, an important advantage of profilometry is that the fracture profile and the underlying microstructure can be observed simultaneously, and therefore, the fracture path can be correlated to the microstructure.
Fig. 1 A schematic fracture surface, a sectioning plane, and the fracture profile generated by intersection of the sectioning plane with the fracture surface. Source: Ref 4
Fig. 2 Fracture profiles. (a) Fracture profile generated from the fracture surface of a metal-matrix composite containing alumina fibers distributed in the matrix of an aluminum alloy. (b) Fracture profile generated from the fracture surface of a tensile test specimen of low-alloy steel. (c) A high magnification field of view of the fracture profile in (b)

Nevertheless, metallographic fracture profile generation is destructive of the fracture surface and sample. Consequently, additional detailed scrutiny of special areas in the fracture surface is not possible once it is coated and cut. However, the fracture surface can be carefully inspected by SEM in advance, and the images of areas of interest (or the image montage of the whole fracture surface) can be grabbed and stored in the memory of an image analysis computer before coating and cutting take place for fracture profilometry (Ref 48). The experimental procedures for fracture profile generation are straightforward. When the specimen is fractured, two nonplanar fracture surfaces are produced. After the completion of SEM observations and image storage, one or both surfaces can be electrolytically coated (see Fig. 2) to preserve the edge on subsequent sectioning (Ref 62). The coated specimen is then mounted metallographically and prepared according to conventional metallographic procedures (Ref 62).

The complete information regarding the fracture profile topography is contained in the set of \((x, y)\) coordinates of the points in the profile. These data can be obtained by digitizing the fracture profile via either interactive or automatic digital image analysis (Ref 22). Some popular image manipulation software packages also have these capabilities. Typically, the process involves tracing the fracture profile using an electronic mouse for interactive measurements, or appropriate image segmentation for automatic measurements. The image analyzer records the coordinates of closely spaced points in the fracture profile at preselected fixed intervals. The profile is thus approximated by a series of straight-line segments as illustrated in Fig. 3. As explained in the subsequent section, numerous geometric characteristics of the fracture profile can be computed from these data.

Fig. 3 Digitization of the fracture profile yields \((x, y)\) coordinates of a set of closely spaced points on the profile. The profile is then approximated by a series of straight-line segments joining the adjacent points.

Replica Technique For Fracture Profile Generation. Planar fracture profiles can be also generated by sectioning replicas of the fracture surface (Ref 63). In this manner, the fracture surface itself can be preserved for subsequent studies. The suitable foil materials and procedures for coating and striping replicas are well established. The replicas can be sectioned by using a microtome, which is also a useful instrument for serial sectioning of the replicas for 3D fracture surface reconstruction. However, care must be taken to minimize the distortion of the replica during the sectioning process. Dimples in fractured 2024 aluminum alloy have been studied using microtomed slices from replicas of the fracture surface (Ref 64). Even though the replica technique is quite attractive, its applications are limited to relatively smooth surfaces in order to avoid tearing the replica.

Nondestructive Profile Generation Methods. The techniques for nondestructive fracture profile generation include laser profilometry (Ref 65), atomic force microscopy (Ref 66), confocal scanning microscopy (Ref 67), light profile microscopy (Ref 68), interferometry (Ref 69), stylus profilometry (Ref 70), and SEM stereo-pair based methods (Ref 71). For laser profilometry (Ref 65), the fracture surface is illuminated using a thin slit of laser at an angle of 45 degrees to the surface, and the surface is observed at an angle of 90 degrees. The projected slit of laser light appears as a straight line if the surface is flat and as a progressively more undulating rough line as the roughness of the surface increases. The output image of the laser slit is grabbed using an image analyzer and digitized to obtain the \((x, y)\) coordinates of closely spaced points along the profile, and the profile characteristics are computed from these data.
Estimation of Geometric Attributes of Fracture Profiles

All experimental profilometric methods yield \((x, y)\) coordinates of a set of closely spaced points in the fracture profile under investigation. From these basic data, numerous geometric attributes of the fracture profile can be calculated in a fairly straightforward manner as to be described. Fracture Profile Roughness Parameters. An important topographic attribute of a fracture profile is its roughness. The motivation for the estimation of roughness is its likely correlation with fracture toughness, the extent of crack deflection, and fatigue-crack growth behavior. Several parameters have been proposed to quantify fracture profile roughness. The most important ones are those that can be defined uniquely and can be used to estimate the corresponding fracture surface roughness attributes in an unbiased manner.

The lineal profile roughness parameter, \(R_L\), is perhaps the most important profile roughness parameter for quantitative fractographic analysis. It is defined as the ratio of the true total fracture profile length, \(L\), and its apparent projected length, \(L_{\text{app}}\) (Ref 17, 18, 19, 20, 21):

\[
R_L = \frac{\lambda}{L}
\]

\((\text{Eq 2})\)

Figure 4(a) shows a schematic fracture surface area \(S\) enclosed in a cylindrical reference volume. The axis of the cylinder \((z\)-axis) is along the direction perpendicular to the average topographic plane of the fracture surface. In the present context, this direction is called vertical axis (Ref 21), and any sectioning plane containing (or parallel to) this vertical axis is called a vertical sectioning plane (see Fig. 4b). A fracture profile generated by a vertical sectioning plane is called a vertical section fracture profile (see Fig. 4c). Note that the vertical axis can be identified in the vertical sections. To calculate the roughness parameter \(R_L\), it is necessary to measure the total length of the fracture profile \(\lambda\). Once the fracture profile is digitized, the total profile length \(\lambda\) is simply given by the sum of the lengths of the straight line segments joining the adjacent digitized points in the profile (see Fig. 3), and therefore, the parameter \(R_L\) can be calculated.

In general, \(R_L\) can have any value from one to infinity. For a randomly oriented fracture profile that does not have any reentrant regions or folds (Ref 20), \(R_L\) is equal to \(\pi/2\). Experimental values of \(R_L\) reported in the literature are in the range of 1.05 to 3.75. For fatigue crack growth studies, the value of the parameter \(R_L\) in the crack growth direction can be used to calculate the geodesic crack length from the experimentally measured conventional projected crack length (Ref 72).
For ductile fracture of ultrahigh-strength steels where the fracture micromechanisms and microstructure lead to smooth dimple interfaces, the lineal profile roughness parameter exhibits quantitative correlation with the critical value of the crack opening displacement, $\delta_{ic}$ (Ref 73).

**Fig. 4** Vertical section planes. $L$, cylinder diameter. (a) A schematic fracture surface of area $S$ enclosed in the cylindrical reference volume. (b) Fracture surface in (a) sectioned by vertical sectioning planes. The vertical planes are the sectioning planes that contain (or are parallel to) the vertical axis. (c) A vertical section fracture profile generated by a vertical plane. The vertical axis can be identified in the vertical sections. $L$, apparent projected length of fracture profile.

It is important to recognize that the geometry of a fracture profile depends on the geometry and topography of the fracture surface as well as on the orientation of the sectioning plane with respect to the fracture surface. Therefore, the value of $R_L$ (and of other profile parameters) can systematically vary with the orientation of the vertical sectioning plane (Ref 20, 21). As illustrated in Fig. 5, if the fracture surface anisotropy is not symmetric with respect to the vertical axis, then different vertical plane orientations yield significantly different values of $R_L$, which has been observed experimentally (Ref 22). It follows that the value of $R_L$ is not unique for a given fracture surface, and contrary to El-Soudani’s earlier theorem (Ref 19), a higher value of $R_L$ (or of any other profile roughness parameter) on a single sectioning plane does not necessarily imply a rougher fracture surface.

**Fig. 5** Schematic illustrating why vertical planes of different orientations may yield fracture profiles having different $R_L$ values.

Profile configuration parameter, $R_P$, is another descriptor of fracture profile roughness. $R_P$ is equal to the average height of the peaks on the fracture profile divided by the average spacing of the peaks (Ref 74), and it also can be calculated from the digitized profile data. Vertical roughness parameter, $R_V$, is defined as the total projected length of the fracture profile.
on the y-axis (i.e., vertical axis for the vertical section fracture profiles) divided by the apparent projected length of the fracture profile on the x-axis (i.e., mean line through the fracture profile). According to Underwood and Banerji (Ref 75), for the fracture profiles that do not have overlaps (reentrant segments), \( R_V \) is equal to \( 2R_P \).

It is important to recognize that the roughness parameters \( R_L \), \( R_P \), and \( R_V \) are dimensionless, and therefore they do not contain information about the length scale of the topographic details on the fracture surface. These dimensionless parameters have the same values for geometrically similar fracture profiles that differ only in scale. A profile roughness parameter that is length-scale dependent is the average value of peak-to-trough height, \( H_a \), in the fracture profile. The parameter \( H_a \) can also be calculated from the digitized profile data. Another important length scale dependent profile roughness parameter is the average asperity height \( \cdot h^* \) in the fracture profile. The asperity heights have to be measured with respect to some reference base line, which is not always naturally defined in the complex tortuous fracture profiles.

A least-square straight line through the fracture profile segment may be used as a base line, but the position of such a base line is dependent on the length of the profile segment. Therefore the calculated average asperity height may vary with the increase in the length of the profile segment under analysis. However, analysis of a number of fracture profiles generated by fatigue crack propagation has revealed that although initially the calculated average asperity height increases with the length of the base line, it quickly approaches a saturation value (Ref 6, 31) for baseline lengths larger than about 500 \( \mu m \) (see Fig. 6). For the nickel-base alloy 718, the saturation value of the average asperity height \( \cdot h^* \) for fatigue crack propagation fracture profiles correlates with the roughness-induced crack closure stress intensity \( K_{cl} \) (see Fig. 7). Recently, a quantitative correlation between the fatigue crack propagation threshold and the standard deviation of the fracture surface asperity height distribution has been reported for some nickel-base superalloys (Ref 5). Through the extrapolation, the linear correlation enables calculation of an “intrinsic” threshold corresponding to zero roughness, which is in agreement with the one predicted by theoretical models.

**Fig. 6** Variation of the calculated average asperity height with the baseline length for the fatigue crack propagation fracture profile of a IN-718 nickel-base alloy having a coarse grain structure (grain size, 23 \( \mu m \)). Observe that average asperity height reaches a saturation value for baseline lengths larger than 500 \( \mu m \). Source: Ref 6, 31
Fig. 7 Plot of crack-closure stress intensity versus average asperity height for fatigue crack propagation in IN-718 nickel-base alloy. The data include specimens of two different grain sizes tested at three different ΔK levels. Source: Ref 6

Angular Orientation Distribution of the Fracture Profile. The angular orientation of a line element on a vertical-section fracture profile is specified by angle $\alpha$ between the line element and the vertical axis. The angle $\alpha$ can be calculated for each straight-line segment connecting the consecutive points of the digitized profile (see Fig. 3). The orientation distribution function of the fracture profile, $f(\alpha)$, is defined such that $f(\alpha)d\alpha$ is the fraction of profile length in the orientation range $\alpha$ to $(\alpha + d\alpha)$, where $0 \leq \alpha \leq \pi$. The orientation distribution function, $f(\alpha)$ quantifies the anisotropy of the fracture profile. For a randomly oriented fracture profile, $f(\alpha)$ is equal to $[1/\pi]$, as all angular orientations are equally likely in such a profile. The orientation distribution function can be computed in histogram form from the experimentally measured values of the orientation angles of the line elements in the fracture profile. The profile orientation distribution function is required for estimation of the fracture surface roughness as well as for estimation of the orientation distribution of surface elements in the fracture surface from the profilometric data. In general, the profile angular orientation distribution may vary with the orientation of the vertical section. Figure 8 shows experimentally measured profile orientation distribution functions, lineal profile roughness parameters, and profile structure factors for fracture profiles observed in three vertical sections of tensile fracture surface of a metal-matrix composite (MMC) generated by uniaxial load perpendicular to the fibers. The three vertical sections are mutually at an angle of 120° to one another, and as they are vertical sections, all of them contain the vertical axis. The MMC contains aligned alumina fibers in an aluminum-alloy matrix (Ref 22).
Fig. 8 Profile orientation distribution functions, $R_L$, and profile structure factors $Ψ$ for fracture profiles observed in three vertical sections of the tensile fracture surface of a metal-matrix composite (MMC) generated by uniaxial loading perpendicular to the fibers. The three vertical sections are mutually at an angle of 120° to one another, and all three sections contain the vertical axis. The MMC contains aligned alumina fibers in an aluminum alloy matrix. Source: Ref 22

Extent of Overlaps in the Fracture Profile. Fracture surfaces, and therefore, fracture profiles may contain reentrant segments or overlaps. Figure 9 depicts a schematic fracture profile containing overlapped (or reentrant) segments. Figure 10(a) shows the fracture profile from the fracture surface of copper that failed in creep, whereas Fig. 10(b) shows a typical room-temperature tensile fracture profile from the same material (Ref 4, 76). Observe that the creep fracture profile contains significant overlapped/reentrant segments, whereas the tensile fracture profile contains no overlaps. The overlaps are obviously not observed in SEM images; profilometry is an important experimental technique to observe and quantify the fracture surface overlaps. The extent of overlaps in the fracture profile is quantified by the profile overlap parameter $O_P$, which is simply equal to the fraction of the profile length containing overlapped or reentrant segments. The profile overlap parameter can be computed via interactive image analysis, where the operator identifies the overlapped segments using an electronic mouse. The lengths of the overlapped segments are then measured by using the digitized coordinate data. The angular orientation distribution of the overlapped segments and their roughness can be also computed from the same data.

Fig. 9 A schematic fracture profile containing overlapped (or reentrant) segments
Fracture surface overlaps. (a) Overlapped/reentrant segments in the fracture profile from the fracture surface of a copper specimen that failed in creep. Source: Ref 4, 76 (b) A typical room temperature tensile fracture profile of copper does not show overlaps.

Correlation between Fracture Profile Path and Microstructure. An important advantage of fracture profilometry using metallographic sections is that the fracture profile and underlying microstructure can be observed simultaneously, which is useful for characterizing correlations between the microstructure and the fracture path. A simple parameter that quantifies such a correlation is the fraction of fracture profile length through a phase or constituent of interest, \( P_f \). In one study (Ref 32), the fraction of fracture profile length through Si particles was observed to be 0.16 in the vertical section fracture profile of a tensile fracture surface of chill cast A356 aluminum alloy having a volume fraction of Si particle equal to 0.074. In this case, as \( P_f \) is significantly higher than the silicon particle volume fraction, the fracture profile appears to preferentially traverse through the silicon particles present in the interdendritic eutectic regions of the cast alloy. Another profile parameter, \( P(k) \), represents the probability of a microstructural constituent \( i \) being associated with a fracture mode \( k \). The probability parameter is defined as (Ref 77):

\[
P_i(k) = \frac{\sum (\lambda_i)_k}{\sum (\lambda_i)_\text{profile}}
\]  

(Eq 3)

where the denominator denotes the total profile length through constituent \( i \), and the numerator represents only that portion of the profile length through constituent \( i \) that is associated with fracture mode \( k \). Three-dimensional analogs to the preceding linear parameters are also proposed. The index \( k \) for fracture modes is assigned according to \( k = 0 \), cleavage fracture; \( k = 1 \), quasi-cleavage fracture; \( k = 2 \), quasi-dimple fracture; and \( k = 3 \), dimple fracture. A linear fracture path preference index, \( Q_i \), has also been proposed for each microstructural constituent \( i \), according to (Ref 77):

\[
Q_i = \frac{\sum (\lambda_i)_\text{profile}}{\sum (L_i)_{||}}
\]  

(Eq 4)

where \( \sum (\lambda_i)_{\text{profile}} \) is the total length of the fracture profile that runs through the \( i \)th constituent, and \( \sum (L_i)_{||} \) is the total intercept length through the microstructure across the \( i \)th constituent and along a straight test line parallel to the effective fracture direction. This parameter can be recognized as a form of the lineal profile roughness parameter \( R_L \) for the corresponding phase or constituent.

Fractal Characteristics of Fracture Profiles. It is well known that the lengths of many naturally occurring irregular lines such as coastlines, mountain ranges, fracture profiles, and so forth depend on the size of the measuring unit or the yardstick used for the length measurements. This concept is illustrated in Fig. 11, which shows that the measured length \( \lambda(\eta) \) of the irregular curve increases as the size of the measuring unit \( \eta \) decreases. Mandelbrot (Ref 79) has given the following relationship between the measured length \( \lambda(\eta) \) and the yardstick length \( \eta \):

\[
\lambda(\eta) = \lambda_0[\eta]^{(1-D)}
\]  

(Eq 5)
In the Eq 5, the dimensionless constant $D$ is called fractal dimension of the line and can be interpreted as a measure of the rate at which new topographic details are resolved as the length $\eta$ of the measuring unit decreases. It is a characteristic of the geometry of the irregular curve. For a straight line, $D$ is obviously equal to one. $D$ increases with the increase in the roughness and tortuosity of a line. Therefore, $D$ can be used as a parameter to quantify roughness and tortuosity of irregular lines such as fracture profiles. In Eq 5, the constant $\lambda_0$ has no physical significance, and it does not have dimensions of length (Ref 78). Further, Eq 5 is based on an assumption that the irregular line is self-similar at all length scales. An equation analogous to Eq 5 can also be written for the surface area of an irregular nonplanar complex surface such as a fracture surface, where the surface area $S(\eta^2)$ is postulated to increase with the area of the measuring unit ($\eta^2$), which then leads to similar definition for the fractal dimension of the surfaces, $D_S$: 

$$S(\eta^2) = S_0[\eta^2]^{(2-D_S)/2} \tag{Eq 6}$$

$D_S$ can be used to characterize tortuosity and roughness of nonplanar complex fracture surfaces.

For characterization of vertical section fracture profiles, Eq 5 can be cast in terms of the lineal profile roughness parameter, $R_L$, by dividing both sides of this equation by the apparent projected profile length, $L$, (see Eq 2) to obtain the following result (Ref 10):

$$R_L(\eta) = \lambda^* \eta^{(1-D)} \tag{Eq 7}$$

$R_L(\eta)$ is the lineal vertical section profile roughness parameter measured with the measuring unit of length $\eta$, $\lambda^*$ is a constant, and $D$ is the fractal dimension of the fracture profile. For a digitized fracture profile, the length of the measuring unit $\eta$ is simply equal to the average distance between the consecutive digitized points in the profile. From the same digitized data of the $(x, y)$ coordinates of the points on the fracture profile, $R_L(\eta)$ can be computed for different values of $\eta$ simply by removing the alternate points (which doubles the value of $\eta$) and recalculating $R_L(\eta)$; this procedure can be repeated until only two points are left on the fracture profile (Ref 10). Eqs 5 and 7 predict that the plot of $\ln \{R_L(\eta)\}$ versus $\ln \{\eta\}$ should be linear with the slope of $(1 - D)$. Thus, the fractal dimension $D$ can be calculated from the slope of such a linear log-log plot using this “compass” method (Ref 79). The fractal dimension of self-similar fracture profiles can be also measured by using the “Minkowski sausage” method (Ref 80), the box counting method (Ref 81), the variation method (Ref 82), and numerous other techniques, which are reviewed elsewhere (Ref 33, 34, 83). The experimental vertical section profilometric data show that, although there is a regime of length scales ($\eta$ values) where the
plot of ln\{R_L(\eta)\} versus ln\{\eta\} is linear, deviations from linearity do exist at low and high values of \(\eta\) (Ref 4, 10, 33, 72, 75). Figure 12a illustrates the experimental trend in a schematic manner, and Fig. 12(b) shows the behavior of one data set pertaining to a fracture profile from tensile fracture surface of AISI 4340 steel (Ref 4). The experimental data show that the profile roughness parameter reaches a saturation value \(R_L(0)\) at small values of \(\eta\). The slope of the linear portion of the ln\{R_L(\eta)\} versus ln\{\eta\} plot is used to compute the fractal dimension of the fracture profile. Vertical section fracture profiles have been used in numerous studies for measurement of profile fractal dimension (Ref 4, 6, 8, 9, 10, 11, 12, 16, 17, 22, 30, 31, 33, 34, 73, 75, 76).

**Fig. 12** Dependence of profile roughness parameter \(R_L\) on ruler length \(\eta\). (a) Schematic illustration of the experimental trend of dependence of profile roughness parameter on the ruler length \(\eta\). (b) Plot of \(R_L(\eta)\) versus \(\eta\) for fracture profile from the tensile fracture surface of AISI 4340 steel. Source: Ref 4

Another popular method for measurement of fractal dimension is the slit-island method proposed by Mandelbrot (Ref 8). In this method, the fracture surface is coated or plated with a protective adherent coating material to avoid damage to the fractured material during sectioning and polishing. The technique is based on the analysis of the features observed on metallographic planes passing horizontally through the fracture surface as illustrated in Fig. 13(a). Two major types of features observed in these horizontal sections are islands of fractured material of interest surrounded by plating and/or mounting material (see Fig. 13b) and lakes of mounting/plating material within the islands of the fractured material of interest (see Fig. 13c). These features are called slit-islands and slit-lakes, respectively. The boundaries of these islands and lakes are the horizontal fracture profiles. In this technique, the areas and perimeters of the individual slit islands are measured. For representative sampling it is necessary to measure the perimeters and areas of more than 100 islands. For self-similar fracture surfaces, the log-log plot of perimeter versus area is expected to be linear (see Fig. 13d), whose slope \(\gamma\) is measured. The fractal dimension of the fracture surface is equal to \((1 + 2\gamma)\). This method has been used to estimate the fractal dimension of fracture surfaces of numerous materials fractured under different conditions (Ref 8, 10, 33, 34, 47, 58, 59, 84, 85).
Estimation of Fracture Surface Parameters from Profilometric Data

In general, a material fracture surface is a tortuous nonplanar surface of complex geometry present in 3D space. A fracture surface may contain reentrant segments (folds), as well. The \((x, y, z)\) coordinates of the set of points in the fracture surface contain complete information about its global topography and morphology. However, in practice, it is of interest to quantify the global fracture surface topography in terms of only few parameters, which can be defined uniquely and can be estimated in an unbiased manner using practical measurement procedures. It is also important that these parameters are physically meaningful, so that they can be correlated to the dominant failure mechanisms, microstructure, and material properties. In this context, the most meaningful global attributes of a global fracture surface are its roughness, tortuosity, anisotropy, and extent of reentrant regions (folds). The work of fracture of brittle solids is related to the roughness of the fracture surface, because a rougher fracture surface has a higher surface area. The fracture surface tortuosity is a measure of crack deflection, and it may be useful for correlations with crack growth resistance, toughness, and impact energy. The extent of reentrant regions has been shown to correlate well with the extent of creep damage. Numerous quantitative parameters have been devised to quantify important global attributes of fracture surface topography. These parameters are defined, and procedures for their estimation from profilometric data are described in subsequent sections.
Fracture Surface Roughness. Figure 4(a) shows a schematic fracture surface of area $S$ enclosed in a cylindrical reference volume. The axis of the cylinder (z-axis) is along the direction perpendicular to the average topographic plane of the fracture surface. Recall that this direction is the vertical axis. The area of the base of the cylinder $A$ is equal to the apparent projected area of the fracture surface $S$ on its average topographic plane. The surface roughness parameter, $R_S$, is defined as the ratio of the true area of the fracture surface, $S$, and its apparent projected area, $A$ (Ref 19, 20, 21):

$$ R_S = \frac{S}{A} \quad \text{(Eq 8)} $$

Note that overlapped segments of projected area are not added for calculating the apparent projected area, $A$. The roughness parameter, $R_S$, is essentially determined by the true area of the fracture surface, $S$. The rougher the fracture surface, the higher is the value of $R_S$. For a flat fracture surface, $R_S$ is equal to 1. On the other hand, for a randomly oriented fracture surface that does not have any folds, $R_S$ is equal to 2. In general, $R_S$ can have any value from 1 to $\infty$. A strong correlation has been observed between $R_S$ and the stress intensity amplitude, $\Delta K$, during fatigue crack growth (Ref 12). A quantitative correlation between impact energy and $R_S$ is proposed when cleavage fracture is the dominant failure mode (Ref 10). Interestingly, $R_S$ has also been used to quantify the textural index of the geomembranes (Ref 9).

Two general methods are available for experimental measurement of $R_S$ that are free of any statistical bias. The first is the technique involving measurements on the vertical section fracture profiles, and the second is based on the direct estimation of fracture surface area using the $(x, y)$ coordinate points in the fracture profile, and from these data, the apparent projected area of the fracture surface enclosed in a cylindrical reference plane is derived (Ref 21):

$$ R_S = \cdot R_L \cdot \Psi^* \quad \text{(Eq 9)} $$

In Eq (9), $R_L \cdot \Psi^*$ is the average value of the product of the two quantities $R_L$ and $\Psi$ measured from the vertical section fracture profiles, which is obtained by averaging the product $(R_L \cdot \Psi)$ over the angular orientations and locations of the vertical sectioning planes. $R_L$ is the lineal vertical section profile roughness parameter defined in Eq 2, and $\Psi$ is the profile structure factor defined by the following equation:

$$ \Psi = \int_0^\pi \left[ \sin \alpha + \left( \frac{\pi}{2} - \alpha \right) \cos \alpha \right] f(\alpha) d\alpha \quad \text{(Eq 10)} $$

In Eq 10, $\alpha$ is the angle between a line element in the vertical section fracture profile and the vertical axis ($0 \leq \alpha \leq \pi$), and $f(\alpha)$ is the angular orientation distribution function of the line elements in the profile, such that $f(\alpha) d\alpha$ is equal to the fraction of profile length in the orientation range $\alpha$ to $(\alpha + d\alpha)$. The orientation angle $\alpha$ can be calculated for each straight line segment joining consecutive digitized $(x, y)$ coordinate points in the fracture profile, and from these data, the distribution function $f(\alpha)$ can be computed in a histogram form. In practice, it is sufficient to divide the angular range of 0 to $180^\circ$ (i.e., $\pi$) into 18 classes having a class interval of $10^\circ$. In such a case, the following discretized form of Eq 10 can be used to compute the profile structure factor $\Psi$ (Ref 22):

$$ \Psi = \sum_{i=1}^{18} a_i h_i \quad \text{(Eq 11)} $$

In Eq 11, $h_i$ is the fraction of profile length having the orientation angle $\alpha$ in the range $10(i - 1)$ to $10i^\circ$. The integer $i$ takes the values from 1 to 18, representing the 18 classes. The coefficients $a_i$ are pure numbers whose values are given in Table 1. Thus, $\Psi$ can be calculated from the $(x, y)$ coordinate data of the digitized profile. The lineal roughness parameter $R_L$ can be also calculated from the same data set as explained previously.

<table>
<thead>
<tr>
<th>$i$</th>
<th>$a_i$</th>
<th>$i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.565</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>1.5232</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>1.4508</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>1.3599</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>1.2655</td>
<td>14</td>
</tr>
</tbody>
</table>
Estimation of (a) $R_5$ using Eq 10 involves averaging the product of the measured parameters $R_L$ and $\Psi$ over the angular orientation of the vertical sectioning planes. Therefore, it is of interest to know how many different orientations should be used to obtain a reliable and precise estimate of $R_5$ from measurements on the vertical section fracture profiles. To answer this question, theoretical calculations and extensive computer simulations that led to the following useful findings were performed (Ref 22):

- For the fracture surface of a material having an isotropic microstructure with the fracture generated by a predominantly uniaxial applied load (for example, uniaxial tensile test), the anisotropy of the fracture surface is expected to be symmetric with respect to the loading direction in a statistical sense. In such a case, the measurements of $R_L$ and $\Psi$ on a single vertical section fracture profile of any one orientation can provide a reliable estimate of $R_5$ when the vertical axis and the loading direction have approximately the same orientation.
- For the fracture surface of a material having an anisotropic microstructure that has a symmetry axis with the fracture generated by the load applied along the symmetry axis, the anisotropy of the fracture surfaces is expected to be symmetric with respect to the loading direction. In such a case also, the measurements of $R_L$ and $\Psi$ on a single vertical section fracture profile of any one orientation can provide a reliable estimate of $R_5$ when the vertical axis and the loading direction have approximately the same orientation.
- In the case of the fracture surfaces of materials that have highly anisotropic microstructures and no symmetry axis, $R_L$ and $\Psi$ can vary significantly and systematically with the orientation of the vertical sections. In such cases, three vertical sections mutually at an angle of 120° are sufficient for estimation of a reliable value of $R_5$. For such fracture surfaces, the average value of the product ($R_L \cdot \Psi$) on the three vertical planes mutually at 120° is a reliable estimate of the fracture surface $R_5$.
- Even for highly anisotropic fracture surfaces, if the axis of fracture surface anisotropy is known, then even two perpendicular vertical sections (one parallel to axis of anisotropy and the other perpendicular to it) are sufficient for a reliable estimation of $R_5$ using Eq 10. In such a case, the average value of the product ($R_L \cdot \Psi$) on the two planes is a reliable estimate of the fracture surface roughness $R_5$.

Figure 8 shows three experimentally measured profile orientation distribution functions in the histogram format $h_i$, lineal profile roughness parameters, and profile structure factors for the fracture profiles observed in three vertical sections of the tensile fracture surface of an MMC generated by uniaxial load perpendicular to the fibers (Ref 22). These three vertical sections are mutually at an angle of 120° to one another, and all three sections contain the vertical axis, which is the direction perpendicular to the mean plane of the fracture surface. Observe that the values of $R_L$ and $\Psi$ are different for the fracture profiles on the three vertical sectioning planes. In this case, the average value of the product ($R_L \cdot \Psi$) on these three planes is 2.17, which is equal to the fracture surface toughness parameter $R_5$ of the fracture surface of the composite. The surface roughness parameter is a useful descriptor of fracture surface topography. It can be used to compute the fractal dimension of the fracture surface (Ref 11, 16, and 33), and it is required for estimation of the true number density and other feature specific attributes of the features (pores, inclusions, particles, etc.) in the fracture surface (Ref 20).

Equation 10 has been used for estimation of the surface roughness of numerous materials including metals (Ref 4, 6, 11, 12, 32, 52, 76 and 87), MMCs (Ref 22), polymers (Ref 86), geosynthetic membranes (Ref 9), and concrete (Ref 16). It has been successfully used to quantify the roughness of the fracture surfaces generated under quasi-static tensile (Ref 22) and compressive (Ref 16) loading conditions, impact loading conditions (Ref 11), cyclic loading conditions (Ref 6, 12, and 87), and high-temperature creep (Ref 4 and 76). Baggerly (Ref 12 and 87) studied fatigue crack propagation and the corresponding fracture surfaces of SAE D4018 and D5506 cast irons and demonstrated a linear correlation between the fracture surface roughness parameter $R_5$ and the stress intensity amplitude $\Delta K$ (see Fig. 14).
Two other parameters sensitive to surface roughness are the surface volume and the topographic index (Ref 75). Surface volume is defined as the volume per unit surface area enclosed by the surface and a horizontal reference plane located at the peak of the highest summit. Experimentally, this parameter is evaluated as the distance between the reference plane and the mean plane through the surface. The topographic index, $\xi$, is defined in terms of the peaks nipped off by a horizontal sectioning plane. Experimentally, $\xi = \Delta/\rho$, where $\Delta$ equals the average separation of asperity contacts and $\rho$ equals the average asperity contact spot radius. An average $\xi$, obtained from different levels of the sectioning plane, would probably be more useful. Other parameters for characterizing some aspects of rough surfaces have been proposed, but the motivation has been mostly for applications other than fractography.

Fracture Surface Anisotropy. The surface elements on a fracture surface may be isotropically oriented, or they may exhibit preferred orientations (i.e., anisotropy). However, in most of the material fracture surfaces, there is always some degree of anisotropy. The extent of anisotropy can provide information concerning the fracture processes and the relationship between the anisotropy of the fracture path and the underlying microstructure. In most of the practical cases, it is of interest to know the distribution of the elevation angles pertaining to different surface elements in the fracture surface. Scriven and Williams (Ref 89) developed a stereological procedure for the estimation of orientation distribution of a set of equiaxed planar facets in 3D space from measurements of the angular distribution of the traces of these facets observed in a metallographic plane under the conditions that the anisotropy of the facets has a symmetry axis and the angular distribution of the traces is measured in a metallographic plane containing the symmetry axis. Underwood and Banerji (Ref 72 and 75) adopted this procedure for estimation of the angular orientation distribution of the surface elements in fracture surfaces. Subsequently, this method has been generalized for the estimation of the angular distribution of any arbitrary fracture surface that may or may not have a symmetry axis, may or may not be composed of facets, and may not necessarily be random (Ref 90). Fortunately, the same formal equation as that of Scriven and Williams (Ref 89) gives the relationship between the discretized histogram form of the distribution of the elevation angles and the discretized histogram form of the average angular orientation distribution function of the line elements in the vertical section fracture profiles. Therefore, the formal numerical unfolding procedure developed by Scriven and Williams is applicable with a somewhat different interpretation of the orientation distribution functions. Scriven and Williams have given a recursive formula and tables of coefficients from which the 3D angular distribution $Q_r$ can be obtained from the experimental profile angular distribution, $Q_r$. The essential working equations are:

$$Q_r = Q(rh) - Q((r - 1)h) \quad (Eq \, 12)$$

where $r = 1, 2, 3 \ldots 18$ (for 0 to 90°) and $h$ is a constant interval value of 5° and $Q(0) = 0$. Moreover:

$$Q(rh) = \frac{\sum_{s=1}^{r} b_{rs} G_s}{\sum_{s=1}^{18} b_{18s} G_s} \quad (Eq \, 13)$$

where the coefficients $b_{rs}$ are tabulated for each successive value of $r$, and the values of $G_s$ are experimentally obtained from $s = r$. The denominator is a constant, with the coefficients $b_{18s}$ tabulated in Table III in Ref 89 under $r = 18$. The values of $Q_r$ give the distribution of the elevation angle, $\theta$, per unit angular interval in $\theta$, in the form of a histogram.
total surface area without overlaps is obtained simply by multiplying the \( O \) values by the cosine of the angle corresponding to the class, and summing over all classes (Ref 72 and 75).

Extent of Fracture Surface Overlaps. The fracture surfaces may contain reentrant or overlapped segments. The reentrant regions can be observed in vertical section fracture profiles (see Fig. 9 and 10a), but they cannot be detected by SEM fractography. The overlap parameter, \( O \), quantifies the extent of overlapped or reentrant regions present in a fracture surface, and it is defined as follows:

\[
O = \frac{S_\text{RW}}{S} \tag{Eq 14}
\]

In this equation, \( S_\text{RW} \) is the total surface area of the reentrant or overlapped segments of the fracture surface, and \( S \) is the total area of the fracture surface. Equation 14 can be cast into the following alternate form by expressing the surface areas in terms of the corresponding roughness parameters:

\[
O = 1 - \frac{(R_S)_\text{WO}}{R_S} \tag{Eq 15}
\]

In Eq (15), \( R_S \) is the fracture surface roughness parameter as defined by Eq 8, and it can be estimated using Eq 9, as explained earlier. The parameter \((R_S)_\text{WO}\) is essentially the roughness of non-overlapped regions of the fracture surface, and it is defined as the ratio of the total surface area of the non-overlapped regions of the fracture surface and the apparent projected area, \( A \). The parameter \((R_S)_\text{WO}\) can be estimated from the vertical section fracture profiles using the following equation:

\[
(R_S)_\text{WO} = \frac{1}{\Psi_\text{WO}} \cdot (R_L)_\text{WO} \tag{Eq 16}
\]

In Eq 16, \((R_L)_\text{WO}\) is the lineal profile roughness parameter calculated by using only the length of the non-overlapped portion of the vertical section fracture profile, and \( \Psi_\text{WO} \) is the profile structure factor (see Eq 10) calculated by using the orientation distribution function of only the non-overlapped element of the fracture profile. The fraction overlap has been useful in the study of creep fracture surfaces, where the extent of overlaps increases with the decrease in the applied stress (Ref 76).

Fractal Behavior of Fracture Surfaces. The fractal dimension of self-similar fracture surfaces can be measured by using the slit-island method discussed earlier or by vertical section profilometry. If the vertical axis of the fracture surface is not an axis of symmetry for the anisotropy, then vertical planes of different angular orientations can lead to fracture profiles having different roughness and fractal dimension values. It can be shown that the fractal dimension, \( D_S \), of a self-similar fracture surface is related to the average value \( D_L \) of the vertical section fracture profile fractal dimension, \( D_L \), by the following simple equation (Ref 4, 33, and 91).

\[
D_S = 1 + \cdot D_L \tag{Eq 17}
\]

If the fracture surface is isotropic or the vertical axis is a symmetry axis for fracture surface anisotropy, then all vertical section fracture profiles are statistically similar. In such a case, the measurement of \( D_L \) on a single vertical section fracture profile can yield a reliable estimate of \( \cdot D_L \), and therefore, of \( D_S \). Numerous attempts have been made to measure the fractal dimension of fracture profiles and surfaces and to correlate these measures to the material properties (Ref 4, 6, 8, 9, 10, 11, 12, 16, 17, 22, 30, 31, 33, 34, 47, 58, 59, 73, 75, 76, 84, 85, 87, and 88). Unfortunately, the fractal dimension data, correlations with mechanical properties (toughness in particular), and the conclusions reached by different researchers are often not in agreement with one another, which has led to some controversies that are yet to be completely resolved.

The values of the fractal dimension of similar specimens measured by the slit-island method and on the fracture profile of a vertical plane often differ significantly (Ref 58, 84, 85, and 92). This may be due to experimental errors or due to a variety of other reasons. For anisotropic fracture surfaces, the vertical section profile fractal dimension can vary systematically with the orientation of the vertical plane. In such a case, the fracture surface fractal dimension can be estimated only from the statistical average value of the vertical fracture profile fractal dimension (Eq 17), and therefore, it may be necessary to perform the profile fractal dimension measurements on vertical planes of several different orientations. Consequently, the fractal dimension measured from a single vertical section fracture profile may not agree with that obtained by the slit-island method. It must be pointed out that the fractal dimension of a fracture surface may vary from one location to another as well. For the impact fracture surfaces of low-alloy steel, it has been shown that the fractal dimension near the Charpy-notch (i.e., fracture initiation region) is higher than that away from the notch (Ref 4). In this case, the average fractal dimension for the whole fracture surface does not correlate with the impact energy, but the fractal dimension in the initiation region at the Charpy notch correlates nicely with the impact energy.

The slit-island method is based on the fundamental assumption that all the islands observed in the horizontal sections are self-similar. This assumption may not be valid for some fracture surfaces (Ref 92), and in such cases, the measured values of the fractal dimension may be erroneous. The assumption of self-similarity of the islands is certainly violated if the surface fractures is self-affine and not self-similar. Yet another difficulty associated with the slit-island method is that for the same fracture surface, different values of the fractal dimension are obtained depending on whether the area-perimeter measurements are performed on the slit-islands or slit-lakes (Ref 58). Further, the island perimeter values are very sensitive to the magnification at which the measurements are performed.
As mentioned earlier, the plot of $\ln \{R_L(\eta)\}$ versus $\ln(\eta)$ is linear only in certain length-scale regimes (that may depend on the nature of the fracture surface/profile), where the yardstick length $\eta$ is not too high or too low. Therefore, the profile fractal dimension calculated by including the data beyond the linear regime may be erroneous. Interestingly, for the slit-island technique as well, at high and low values of $\eta$, the measured fractal dimension is observed to depend on the length of the yardstick length $\eta$ used to measure the perimeters of the islands (Ref 93). It can be concluded that the observed discrepancies between the fractal dimensions measured by different methods can be due to several different reasons, and more research is needed on the measurement techniques and the interpretation of the fracture surface fractal dimension.

References cited in this section


72. K. Banerji, Ph.D. dissertation, Georgia Institute of Technology, 1987


Quantitative Fractography
Arun M. Gokhale, Georgia Institute of Technology

SEM Quantitative Fractography

The SEM is routinely used for characterization of features observed in material fracture surfaces. A fracture surface may contain microstructural features such as voids, inclusions, precipitates, and/or topographic features such as dimples, cleavage plane facets, grain boundary facets, and striations that are generated by specific fracture micromechanisms such as ductile fracture, cleavage fracture, intergranular fracture, and so forth. Once the features of interest are uniquely identified, they can be quantitatively characterized via estimation of their number density, average size, spacing, and area fraction, and the distributions of these attributes, as required. It must be emphasized that the material fracture surfaces are of stochastic nature. Therefore, no two SEM fields of view from the same fracture surface are exactly alike. Consequently, the number density of features of interest (as well as other attributes) observed in an SEM field of view (or fractograph) usually varies from one field to another in a statistical manner. Thus, the feature specific attributes estimated from the measurements performed on a single field of view (or a single SEM fractograph) are usually not representative of the whole fracture surface region of interest. Consequently, quantitative characterization of the fracture surface based on measurements performed on a single SEM fractograph (or a field of view) is not useful. The most meaningful measures are the statistically representative averages obtained by performing the measurements on numerous randomly positioned fields of view (or fractographs). Further, as the SEM image is a projected image of the fracture surface, it is necessary to relate the attributes measured from the SEM fractographs to their corresponding true values in the fracture surfaces. In the following subsections, the feature-specific fractographic parameters are defined and the procedures for their estimation are described.

Number Density of Features in Fracture Surface. The number density of features of interest (dimples, inclusions, voids, etc.), $N_f$, is the statistical average number of features of interest per unit true area of the fracture surface. The three steps involved in the estimation of $N_f$ are:

1. Unbiased sampling of the fields of view from the population of fields contained in the fracture surface (or a region of the fracture surface of interest) for performing the measurements
2. Correct counting of the number of features of interest in each sampled field of view and calculation of the average number of features per field of view
3. Estimation of the true number density by dividing the average number of features per field of view by the true average fracture surface area corresponding to one SEM field of view (or one SEM fractograph).

The most efficient and unbiased sampling of the fracture surface can be achieved by using a systematic random sampling scheme (Ref 23 and 24). For such a sampling, the location of the first measurement field is chosen at random in the fracture surface region of interest, and the remaining fields are sampled at fixed distance intervals, so as to uniformly cover the complete area of the fracture surface region of interest (see Fig. 15). For example, one may sample every fifth or tenth field of view for performing the measurements.
Once the features of interest are correctly identified in a field of view, counting the number of features in the field is relatively straightforward, but some difficulties are created due to the “edge effect.” Some features of interest may intersect the boundaries of the field of view, which raises the question whether to include them, exclude them, or count them as in the number count. Gundersen (Ref 94) has shown that none of these are statistically unbiased and correct procedures. The correct statistical procedure to account for the edge effect requires the use of an unbiased counting frame (Ref 94). An unbiased counting frame of area, $A_C$, smaller than the area of the field of view is superimposed as shown in Fig. 16. The counting frame has two edges shown in solid lines. The other two edges are shown in broken lines. Every feature that is completely contained in the counting frame and does not intersect either the solid or broken lines is counted as one. Every feature that intersects the solid line(s) in counted as zero, and every feature that intersects the broken line(s) is counted as one provided it does not also intersect the solid line(s). When correctly counted using this counting rule, there are six features in the counting frame shown in Fig. 16, including four features completely contained in the frame and two features that intersect the broken lines but not the solid line(s). Using this procedure, the number of features in each sampled field of view can be correctly counted, and from these data, the statistically representative average number of features $N$ in the SEM image area $A_C$ can be calculated, which yields an estimate of the statistical average number of features per unit area of SEM image, $N_s$.

As the measurements are performed only on a statistical sample from a large population of fractographic fields, although the sample average value is an unbiased estimator of the population average, there is always a statistical sampling error. This sampling error in the estimated value of $N_s$, is inversely proportional to the square root of the number fields of view on which the measurements are performed. It can be estimated by using standard statistical techniques described elsewhere (Ref 23 and 24). Performing the measurements on more fields of view can reduce the sampling error in the estimated $N_s$. 

**Fig. 15** Systematic random sampling of fields of view

**Fig. 16** Counting frame for unbiased counting number of features. Source: Ref 94
The statistical average number of features per unit area of SEM image, $N_S$, is not equal to the average number of features per unit true area of fracture surface, $N_f$. This is because the SEM image is a projected image, and therefore the area $A_C$ of the counting frame superimposed on the SEM image is equal to the projected area of the corresponding fracture surface segment on whose image the counting frame is overlaid. In other words, in the calculation of $N_S$, the normalization is with respect to the projected area of the fracture surface, whereas in $N_f$, it is with respect to the true area of the fracture surface. Recall that the fracture surface roughness parameter, $R_s$, is equal to the true area of the fracture surface divided by its projected area. Therefore, the average number of features per unit true area of fracture surface $N_f$ is related to the measurable parameters $N_S$ and $R_s$ through the following equation:

$$N_f = N_s/R_s$$  \hfill (Eq 18)

The fracture surface roughness parameter, $R_s$, thus converts the measurements performed on the SEM images to the desired true number density of the features on the fracture surface, $N_f$. If the fracture surface contains significant overlaps or reentrant regions, then a further overlap correction is necessary, and Eq 18 takes the following form (Ref 4):

$$N_f = N_s[R_s(1 - O)]$$  \hfill (Eq 19)

In Eq (19), $O$ is the fraction of fracture surface area having overlaps as defined in Eq 14, which can be also estimated from fracture profilometric measurements.

The number density of features in the fracture surface, $N_f$, can be used to quantify the affinity of the features of interest to the fracture path. For such an analysis, it is necessary to estimate the number of the same category of features per unit area of a random plane, $N_A$, through the bulk microstructure by using stereological techniques (Ref 23 and 24). If $N_f$ is significantly higher than $N_A$, then it can be concluded that the fracture path preferentially goes through the features under investigation. On the other hand, if $N_f$ and $N_A$ have comparable values, then the fracture path has no preference for the features under the study. As an example, consider the role of processing defects in the tensile failure of a unidirectional continuous Al$_2$O$_3$ fiber-reinforced metal-matrix composite (MMC) having an aluminum-lithium alloy matrix (Ref 14, 30, and 95). The virgin material contains processing defects like voids and inclusions. On this MMC, the tensile tests were performed with the tensile loading direction perpendicular to the fibers. The number of voids and inclusions per unit area of these tensile fracture surfaces were measured, and the number density of these defects were also measured in a random metallographic plane through the bulk microstructure. These data are given in Table 2. The number density of voids in the fracture surface is almost an order of magnitude larger than that in a random plane through the bulk microstructure. Further, although both the voids and the inclusions have a comparable number density in a random metallographic plane through the bulk microstructure, the number density of voids is higher than that of the inclusions in the fracture surface by almost a factor of two. Therefore, it can be concluded that the fracture path preferentially goes through the processing voids, and the voids are more deleterious to the tensile fracture than the inclusions.

Table 2 Number density of processing defects in the tensile fracture surface of metal-matrix composite specimen with loading direction perpendicular to the fibers

<table>
<thead>
<tr>
<th>Type of processing defect</th>
<th>Number density (per mm$^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>In the tensile fracture surface</td>
</tr>
<tr>
<td>Voids</td>
<td>15.5</td>
</tr>
<tr>
<td>Inclusions</td>
<td>7.5</td>
</tr>
</tbody>
</table>

Source: Ref 14, 30

Area Fraction of Features in Fracture Surface. In addition to the number density, it is often of interest to estimate the fraction (or percentage) of fracture surface area occupied by a phase or constituent of interest. Further, when the global fracture is governed by multiple fracture micromechanisms or failure modes such as cleavage fracture, intergranular fracture, ductile fracture, and so forth, it is of interest to measure the fraction (or percentage) of fracture surface area generated by each micromechanism or failure mode. For example, Fig. 17 depicts an SEM fractograph of a wrought aluminum alloy containing two distinctly different types of regions. In this fracture surface, the conventional ductile fracture process generates the regions with dimples, whereas the faceted regions are generated by an intergranular fracture mechanism. The area fraction, $A_A$, generated due to a specific mechanism or of a specific constituent in the SEM image can be measured manually by superimposing a test point grid on the SEM fractographs/images and using the following classical stereological relationship (Ref 23 and 24), or it can be measured automatically using digital image analysis:

$$A_A = \bullet P_{p\bullet}$$  \hfill (Eq 20)

$\bullet P_{p\bullet}$ is the statistical average value of the fraction of test points on the grid contained in the constituent of interest or in the region generated by the specific failure mode. Note that the special counting frame used for number density measurements (Fig. 16) is not required for the area fraction measurements. The average value $\bullet P_{p\bullet}$ must be obtained by performing the measurements on numerous fields of view.
Fig. 17 Scanning electron microscope fractograph of fracture surface of a wrought aluminum alloy. Observe that there are distinct regions generated by a ductile fracture process (regions with dimples) and intergranular fracture process (facets).

The measurement fields can be sampled in an unbiased and efficient manner by using the systematic random sampling scheme (Ref 23 and 24) discussed earlier, and the sampling error in the estimated value of \( P_\ell \) can be calculated by using the standard statistical methods (Ref 23 and 24). The quantity, \( A_\lambda \), is the fraction of the SEM image area occupied by the phase of interest or generated by a specific failure mechanism, and in general it is not equal to the fraction of fracture surface area, \( F \), occupied by the phase of interest or generated by a specific failure mechanism, because in the SEM, only projected areas are observed.

The true area fraction, \( F \), representing the fraction of the fracture surface area is related to the corresponding SEM parameter, \( A_\lambda \), through the following equation involving the roughness parameters:

\[
F = \frac{R_s}{(R_s)_\beta} A_\lambda
\]  
(Eq 21)

\( R_s \) is the roughness parameter for the overall fracture surface, and \((R_s)_\beta\) is the roughness parameter only for the specific constituent \( \beta \) in the fracture surface or for the regions generated by a specific failure mechanism, defined in an analogous manner. Both \( R_s \) and \((R_s)_\beta\) can be estimated from vertical section fracture profiles using Eq 9. However, if the roughness of the constituent of interest \((R_s)_\beta\) is approximately the same as that of the overall fracture surface, then the following approximate equation holds. In such a case, it is not necessary to measure the roughness parameters:

\[
F \approx A_\lambda
\]  
(Eq 22)

The area fractions of the fracture surface generated by different failure mechanisms affect the fracture related properties such as plane strain fracture toughness and ductility. These area fractions are in turn governed by the microstructural geometry and test conditions. For example, in a partially recrystallized 7050 wrought aluminum alloy, the microstructure consists of large recrystallized grains and unrecrystallized regions containing subgrains. The fracture path goes through the recrystallized grains, the unrecrystallized regions, the high-angle grain boundaries, and numerous other microstructural features (Ref 48). It is observed that a higher area fraction of fracture surface through the unrecrystallized regions leads to a higher value of \( K_{ic} \). Figure 18 depicts the plot of \([K_{ic}]^2\) versus the area fraction of fracture surface plastic zone through the unrecrystallized regions for a series of specimens having different degrees of recrystallization. There is a linear quantitative correlation between \([K_{ic}]^2\) and the area fraction of the fracture through the unrecrystallized regions for the two sets of specimens studied (Ref 48). Another example is shown in Fig. 19, where the ultimate tensile strength (UTS) of high-pressure die-cast AM60 magnesium alloy is correlated to the area percentage of the gas (air) porosity observed on the tensile fracture surfaces (Ref 3).
Fig. 18 A plot of \([K_{ic}]^2\) versus the area fraction of the fracture surface in the plastic zone through the unrecrystallized regions containing subgrains for two sets of specimens of partially recrystallized 7050 Al alloy (Ref 48)

Fig. 19 Quantitative correlation between the ultimate tensile strength and the area percentage of voids on the corresponding fracture surfaces of high-pressure die-cast AM60 magnesium alloy specimens having the same dendrite arm spacing. Source: Ref 3

Average Size and Shape of Features in the Fracture Surface. Estimation of the average size and shape of features such as dimples, facets, and asperities in the fracture surface is of practical interest. Considerable effort has been devoted to measurements of dimple size and shape (Ref 35, 64, 75, and 77), and average asperity height and width (Ref 5, 6, 31, and 90). The measurements have been performed on SEM fractographs (Ref 72 and 96), on stereo-pair fractographs (Ref 97), and on fracture profiles (Ref 5, 6, 31, 72, 77, and 90). The true average size of the dimples (or any other features of interest) can be specified in terms of their average area \(\Omega\) in the fracture surface (not in the SEM projected image). If the true number density of the dimples (or any other features of interest), \(N_t\), and their area fraction in the fracture surface \(F\)
are estimated, then the average area of the dimples (or of any other features such as facets, inclusions, etc.) \( \Omega \) can be efficiently estimated by using the following relationship:

\[
\Omega = \frac{F}{N_f}
\]

(Eq 23)

If the whole fracture surface consists of ductile dimpled fracture only, then \( F \) is equal to one. In such a case, assuming that the dimples completely cover the dimpled fracture surface, Eq 23 can be simplified as follows:

\[
\Omega = \frac{1}{N_f}
\]

(Eq 24)

Alternately, if the number density of dimples is expressed in terms of true unit area of the dimpled fracture regions, then assuming that the dimples completely cover the dimpled fracture regions, Eq 24 can be used, even if the complete fracture surface does not consist of dimpled fracture. In such a case, the counting frame (Fig. 16) should be placed such that it is completely contained in the dimpled fracture region(s), and in Eq 18, \( R_s \) should be replaced by the surface roughness parameter corresponding to the dimpled fracture regions of the fracture surface. Once \( \Omega \) is estimated, its square root, \( \Omega^{1/2} \), can be used as a measure of the lineal average dimple size. Note that with the appropriate interpretation of the different terms, Eq 23 and 24 are equally applicable to other features such as grain facets, inclusions, and so forth.

As an example of application of Eq 23 and 24, consider the dimples in the tensile fracture surface of a permanent mold cast A356 aluminum alloy widely used for automotive applications (see Fig. 20). In this fracture surface, the number density of the dimples, \( N_S \), measured from the SEM fractographs has a value of \( 3.6 \times 10^{-2} \) per \( \mu m^2 \), and the surface roughness parameter \( R_s \) is equal to 1.9. Substituting these values in Eq 18 and 24 yields average true dimple surface area, \( \Omega \), equal to 53 \( \mu m^2 \), which is 89% higher than the measured average projected area of 28 \( \mu m^2 \). Therefore significant improvements in the estimated dimple size values are realizable when roughness parameters are available to correct the SEM measurements (Ref 75). Interestingly, the present procedure for estimation of true average dimple surface area does not require the measurements of individual dimple sizes; only the number density of dimples is required. Therefore, the procedure is very efficient.

**Fig. 20 Dimples in the ductile fracture surface of a permanent mold cast A356 Al-alloy**

Another measure of the average feature size is based on their average perimeter in the fracture surface \( L_{true} \), which can be estimated in a straightforward manner using the following equation (Ref 75):

\[
L_{true} = \frac{\pi R_L}{2N_f} \frac{P_L}{N_S}
\]

(Eq 25)

In Eq 25, \( R_L \) is the average lineal vertical section profile roughness parameter, \( N_S \) is the number of dimples (or any other features of interest) per unit area of SEM image, and \( P_L \) is the average number of intersections between random test lines placed on the SEM images and the boundaries of the dimples or other features of interest. For the dimples in the fracture surface of the A356 aluminum alloy shown in Fig. 20, \( P_L \) is equal to 0.5 per \( \mu m \), \( R_L \) is equal to 1.5, and \( N_S \) is equal to \( 3.6 \times 10^2 \) per \( \mu m^2 \). Substituting these values in Eq 25 yields the value of the average perimeter of the dimples \( L_{true} \) equal to \( 32.7 \mu m \). If the dimple edges can be assumed to be convex (not necessarily circular), then the average width of the dimples \( W \), is related to their average perimeter, \( L_{true} \), by the following equation (Ref 98):

\[
W = \frac{L_{true}}{\pi}
\]

(Eq 26)

Thus, for the dimples in the fracture surface of the A356 aluminum alloy, the average width of the dimples, \( W \), is equal to 10.4 \( \mu m \). On the other hand, the average projected dimple width given by \( \frac{P_L}{2N_S} \) is equal to 6.9 \( \mu m \). Again, there is a significant difference between the true and the projected measures of the dimple characteristic.
Other dimple/feature characteristics of interest are the depth (or height, $H$) of the dimple/feature and the height-to-width ratio, $\xi$. The height-to-width ratio of the dimples (or other features of interest) can be directly measured by using stereophotogrammetry (Ref 97). Measurements on fracture profiles have also been used for the estimation of the height-to-width ratios of dimples (Ref 64, 72, 77, and 75), and asperities (Ref 6, 31, and 90) on fracture surfaces. The dimple/asperity heights and widths can be extracted directly from the set of $(x, y)$ coordinates of the points on the digitized fracture profiles, from which the height-to-width ratios can be calculated (Ref 6, 31, 72, 75, and 90).

Estimation of Striation Spacing in the Fatigue Fracture Surfaces. Fatigue striations are observed in the fatigue fracture surfaces of numerous materials. The close correlation between the striation spacing, macroscopic crack growth rate ($da/dN$), and effective stress intensity range, $\Delta K$, is well known. Consequently, measurement of striation spacing is an important aspect of the quantitative characterization of fatigue fracture surfaces. Figure 21 shows fatigue striations in the fracture surface of a fatigue test of a permanent-mold cast modified A356 aluminum alloy tested at 0.4% strain amplitude. The fatigue striations are observed in the fracture regions through the aluminum-rich dendrites but not in the interdendritic regions. Further, the local striation normals for all colonies are not along the same direction.

**Fig. 21  Fatigue striations in a cast A356 aluminum alloy. (a) 500×. (b) 1500×**

Figure 22 illustrates the relationship of the true striation spacing on the fatigue fracture surface to the projections of the striations observed in the SEM fractographs. The problem of estimation of true striation spacing from the measurements performed on the projected SEM images has been analyzed (Ref 99). To estimate the average projected striation spacing normal to the striations, $\lambda_{pr}$, superimpose a straight test line on a striation colony and count the number of striations intersecting the test line. The test line should be oriented normal to the striations. Repeat these measurements on numerous randomly sampled striation colonies in the region of interest, each time placing the test line normal to the striations. From these measurements, calculate the average number of striations intersected per unit test line length, $N_{lt}$. The average projected striation spacing normal to the striations $\lambda_{pr}$ is equal to $1/N_{lt}$:
The true average striation spacing normal to the striations \( \lambda \) is related to its projected measure \( \lambda_{pr} \) through the lineal roughness parameter, \( R_L \), as follows:

\[
\lambda = R_L \lambda_{pr} = \frac{R_L}{N_L}
\]

(Eq 28)

Fig. 22 Geometric considerations for striation spacing measurements. \( L_T' \) test line length; \( N \), number of striations; \( l_i \); true striation spacing; \( l_{meas} \) measured striation spacing in crack propagation direction. Primes indicate quantities in projection planes. Source: Ref 75

For the fracture surface of the fatigue specimen of the A356 aluminum alloy shown in Fig. 21, the profile roughness parameter, \( R_L \), measured from vertical section fracture profile was observed to be 1.4. The average number of intersections of the striations with the test lines normal to them, \( \frac{N}{N_L} \), was estimated to be equal to 0.83 per \( \mu \)m. For these data, Eq 28 yields a value of 1.7 \( \mu \)m for the average striation spacing normal to the striations, \( \lambda \). In the cast aluminum alloys, where the fatigue cracks initiate at large pore(s) and/or processing defects such as internal oxides, the number of cycles for fatigue failure strongly correlates with the average striation spacing normal to the striations, \( \lambda \). Figure 23 depicts such a quantitative correlation for a set of specimens of a permanent mold cast A356 aluminum alloy tested at 0.5% strain amplitude having different defect populations.

Fig. 23 Correlation between striation spacing and fatigue life of permanent mold cast modified A356 aluminum alloy specimens tested at 0.5% strain amplitude

For fatigue fracture surfaces, another parameter of interest is the average true striation spacing along the global crack growth direction, \( \lambda_{CG} \), which can be estimated by superimposing a set of test lines on the striation colonies such that the test lines are parallel to the global crack growth direction and counting the number of striations intersected by the test
The average value of the number of striations intersected per unit test line length, \( \cdot (N_{L})_{CG} \cdot \), can be calculated from these data. The average projected striation spacing along the crack growth direction, \( \cdot (\lambda_{CG})_{P} \cdot \), and the corresponding true measure \( \cdot \lambda_{CG} \cdot \) can be calculated by using the following equations:

\[
\begin{align*}
\cdot (\lambda_{CG})_{P} &= 1/\cdot (N_{L})_{CG} \cdot \\
\cdot \lambda_{CG} &= R_{L} \cdot (\lambda_{CG})_{P} = R_{L}/\cdot (N_{L})_{CG} \cdot
\end{align*}
\]

(Eq 29)

(Eq 30)

If the striation colonies are oriented randomly with respect to the global crack growth direction, then the average true striation spacing normal to the striations, \( \cdot \lambda \cdot \), is related to the average true striation spacing along the global crack growth direction, \( \cdot \lambda_{CG} \cdot \), as follows:

\[
\cdot \lambda_{CG} = (\pi/2) \cdot \lambda
\]

(Eq 31)

References cited in this section


35. A.S. Argon and J. Im, Separation of Second Phase Particles in Spheroidized 1045 Steel, Cu-0.6% Cr Alloy, and Maraging Steel in Plastic Straining, Metall. Trans., Vol 6, 1975, p 839–851


72. K. Banerji, Ph.D. dissertation, Georgia Institute of Technology, 1987


Quantitative Fractography

Three-Dimensional Fracture Surface Reconstruction

Fracture surfaces are usually complex tortuous nonplanar surfaces in 3D space. Direct quantitative characterization of the fracture surface topography and its graphic representation is possible via 3D reconstruction of the fracture surface. The techniques for fracture surface reconstruction include SEM stereo-pair based imaging (Ref 39, 40, 41, 42, 43, 44, 45, 47, 63, and 100), serial sectioning (Ref 48 and 49), fracture surface topography analysis (FRASTA) using confocal scanning microscopy (Ref 67), and atomic force microscopy (Ref 66). In addition, the electron backscatter diffraction (EBSD) and image processing can be combined with the stereo-photogrammetry and serial sectioning to obtain both crystallographic and morphological information (Ref 46 and 101). Further, reconstruction of the 3D fracture surface and the 3D microstructure beneath it from a stack of vertical serial sections can provide detailed information on correlation of microstructure to the fracture path (Ref 48 and 49).

SEM Stereo-Pair Imaging-Based Methods. A SEM stereo pair consists of two photographs (or images) of the same field taken at small tilt angles with respect to the normal. In Fig. 24, the points A and B appear at A' and B', and A'' and B'' in stereo pictures taken at tilt angles ±α. The lengths A'B' and A''B'' can be measured either from the two photos separately or from the stereo image directly. The difference A'B' - A''B'' is called the parallax, Δx. The difference in the z-coordinates (elevation) of the points A and B in the fracture surface, Δz can be calculated from the measured parallax, Δx, tilt angle α, and average magnification of the field, M, by using the following equation (Ref 39):

\[
\Delta z = \left[ \frac{1}{2M \sin \alpha} \right] \Delta x
\]

(Eq 32)
Equation 33 assumes that the image is created by a parallel incident beam, which is a good approximation only at high magnifications. Consequently, a low magnification can lead to a spherical distortion error in some regions of the reconstructed topographic map of the fracture surface. To minimize such errors, for metallic fracture surfaces, a magnification lower than 500× should be avoided for stereo-pair imaging (Ref 102). Further, to avoid an excessive excitation depth, the beam voltage must be low (Ref 43). The tilt angle is usually selected in the range of 5 to 10°. The \((x, y)\) coordinates of the points on the fracture surface can be measured with a grid superimposed on the images or from the pixel coordinates if an image analyzer is used. Combining the parallax data with the \((x, y)\) coordinates then generates the \((x, y, z)\) coordinates of the points in the fracture surface from which the topography of the fracture surface can be reconstructed using appropriate algorithms.

A fracture surface can be mapped in a semiautomatic interactive manner by using a mirror type stereometer that can be linked to an image analyzer (Ref 47, 63, and 97). In this technique, a point light source of the stereometer is adjusted to lie at a chosen \((x, y)\) position. If the floating marker point is not at the height level of the fracture surface at \((x, y)\) point, then the marker appears to float out of contact with the fracture surface. The operator adjusts the point marker until the marker appears to be positioned on the fracture surface, which then determines the \(z\)-coordinate of the chosen \((x, y)\) point. A foot switch can be used to send the \((x, y, z)\) coordinates of the point to a computer. Figure 25 shows a contour map of a fracture surface and some fracture profiles obtained in this manner (Ref 63).
As the two images of a stereo pair basically differ only in the local relative shifts due to the parallax, modern pattern recognition and image analysis techniques can be utilized to automatically detect the location of the same point on the fracture surface in the two images, from which the parallax and therefore the \( z \)-coordinate of the point can be computed. Recently, Hein and coworkers (Ref 41, 42, 43 and 45) have utilized an area-matching algorithm based on image cross-correlation function (Ref 103) for detection of the same point in the two stereo-pair images. First, a small window is extracted from each image of the stereo pair at the same absolute location. A cross correlation operation is performed on these images, which utilizes the gray-scale information at each pixel of the windows. As the stereo-pair images basically differ only in the local relative shifts, a maximum appears in the cross-correlation function identifying the pair of points in the stereo pair that originated from the same point in the fracture surface. The position of the maximum relative to the center of the cross-correlation function is a measure of the parallax. However, the image quality and contrast can significantly affect the ability of the software to identify the matching pixels in the two images of the stereo pair, leading to noisy elevation data. A median filter may be useful to reduce such a noise.

Using this information and the \((x, y)\) pixel coordinates, the \((x, y, z)\) coordinates of closely spaced points on the fracture surface can be computed, which can be then used for the reconstruction of the fracture surface. Hein (Ref 43) has given image macroroutines for stereo-pair image editing, reconstruction of the fracture surface, and quantitative analysis that can be used with the image analysis software developed by National Institute of Health (NIH) (Ref 104). Note that the NIH image analysis software is distributed as freeware.

Once the \( x \)-, \( y \)-, and \( z \)-coordinates of closely spaced points in the fracture surface are available, numerous descriptors of the fracture surface geometry can be calculated from these data. Further, the reconstructed 3D fracture surface can be visualized by using the surface-rendering image processing algorithms (Ref 105). Stereo-pair imaging and fracture surface reconstruction have been used for numerous applications in fractography including identification of the boundaries of the stretch zone in the fracture surfaces of plane-strain fracture toughness test specimens of 7050 aluminum alloy (Ref 106), estimation of the roughness and fractal dimension of the fracture surfaces (Ref 43, 47), estimation of the distribution of asperity heights on the fracture surfaces (Ref 5), and characterization of fatigue-fracture morphologies (Ref 40).

The stereo-pair-based methods are nondestructive with regard to the fracture surface. They also allow detailed scrutiny of some areas and wider spacings in regions of less interest. It is also possible to obtain a roughness profile from a fracture surface along a curved or meandering path, a task that is very difficult to accomplish by sectioning methods. However, no information is provided on any subsurface cracking or possible interactions between microstructure and fracture path.
Further, overlapped (reentrant) regions of the fracture surface are difficult to characterize using the stereo-pair based methods.

Electron Backscatter Diffraction (EBSD) for Fracture Surface Crystallography. Conventional fractography provides morphological and topographic information, but it does not enable determination of the crystallographic orientations of the features present in the fracture surface, which is crucial for understanding some fracture processes and micro-mechanisms. The combination of EBSD technique and SEM stereo-pair imaging presents an opportunity to interrogate the morphology and crystallography of the features like planar facets in the fracture surfaces, which can be very useful for the study of the failure modes such as cleavage fracture (Ref 46, 101, and 107). However, the EBSD determines only three of the five degrees of freedom required to unambiguously describe a facet in 3D space; the information concerning the other two degrees of freedom must be inferred indirectly or measured separately from serial sections perpendicular to the fracture surface (Ref 101).

Recently, a combination of EBSD, SEM stereo-pair imaging, and serial sectioning has been successfully used to measure the orientation distribution of the cleavage facets in the fracture surfaces of C-Mn and Mn-Mo low-alloy ferritic steels generated by Charpy impact tests performed at different temperatures on the specimens from weld metals (Ref 46, and 101). In these fracture paths, the transgranular cleavage fracture is the dominant failure mode, which is accompanied by the microvoid-induced ductile fracture. The EBSD and SEM stereo-pair measurements revealed that all cleavage facets were of [001]-plane orientations, and there was no evidence of cleavage along any other family of planes. The study also demonstrated that even when the information concerning all the five degrees of freedom of the cleavage facets is not available, the cleavage facet plane could still be determined under certain conditions. Nonetheless, additional research is necessary to further develop and refine this promising technique.

Three-Dimensional Laser Confocal Scanning Microscopy. Laser confocal scanning microscopy is another nondestructive technique for reconstruction of the fracture surface topography. The confocal microscope captures a series of images at different focusing levels and extracts the image of the band in focus at the given level (Ref 65 and 81). For the fracture surface, it essentially generates “horizontal” contour plots at different elevation levels. To obtain an image of the whole fracture surface, either the specimen is moved on a computer-controlled scanning stage in a raster scan, or the laser beam is moved with scanning mirrors to move the focused spot across the specimen. In both the cases, the image on each plane is constructed pixel by pixel in the computer memory as the scan proceeds. The images of the in-focus bands at successive planes are then combined to reconstruct the well-focused 3D topographic map of the fracture surface by using image analysis techniques. Laser-confocal scanning microscopy has been combined with appropriate computer codes for precision matching of the conjugate fracture surfaces (Ref 67). The resulting data are used to recreate the microfracture events as the crack propagates to generate the fracture surface. Under certain conditions, such fracture surface topographic analysis (FRASTA) can provide useful information concerning fracture micromechanisms, crack nucleation and growth, and fracture toughness from the analysis of the fracture surfaces alone. Therefore, FRASTA is a useful tool for analyzing fracture surfaces of specimens tested under controlled laboratory conditions, as well as for analyzing field failures.

Vertical Serial Sectioning of Fracture Surface and Associated Local Microstructure. Direct correlation of the local 3D microstructure above and beneath the fracture surface to the fracture path, fractures surface morphologies, and failure modes is of considerable practical interest in the study of relationships between the fracture-related properties and 3D microstructure. Slant metallographic sections through the fracture surface (Ref 60 and 68) have been used to understand the correlation of the local microstructure to the fracture path in a qualitative manner. However, a precision matching of the conjugate fracture surfaces of the specimen and associated local 3D microstructure using a combination of stereo-pair-based SEM imaging, exhaustive vertical section serial sectioning, and digital image processing is essential to quantitatively and unambiguously correlate the local topography and morphology of the fracture surface with the associated local 3D microstructure. To the best of the author's knowledge, such a detailed quantitative fractographic investigation has never been carried out.

In a recent study (Ref 48 and 49), a combination of SEM fractography (projected images), exhaustive vertical-section serial sectioning, and precision matching of the montage of the SEM fractographs with the local 3D microstructure above and beneath the fracture surface has been carried out for the plastic zone in the fracture surface of plane-strain fracture toughness test specimen machined from a partially recrystallized 7050 Al-alloy. First, a “montage” of contiguous digital SEM fractographs of the plastic zone was created by precision matching of the adjacent fractographic fields by using digital image analysis. Next, a series of closely spaced vertical serial sections through the fracture surface were generated using standard metallographic procedures. The vertical sections were aligned and precisely matched, and then, using image analysis, the fracture profiles and the local microstructures in the serial sections were precisely associated with their corresponding locations in the fractographic montage. In this way, the exact local microstructure above and beneath any local region of the fracture surface could be matched with the corresponding local fracture surface features and morphologies. For example, Fig. 26 shows the two SEM fractographs of the same local region of the plastic zone of the conjugate fracture surfaces and the two vertical serial sections through this region, which illustrates the local fracture involving void nucleation and growth around the constituent particles (inclusions) present in the recrystallized grains.
Fig. 26 The two matching SEM fractographs of the same field in the two conjugate fracture surfaces and the two vertical serial sections through the same location above and beneath the fracture surface. The combination of the fractographs and serial sections clearly reveal the fracture segment generated due to the nucleation and growth of void(s) around the constituent particles present in the recrystallized grains.

This technique was used to associate different fracture surface morphologies generated by different features in the local 3D microstructure and the corresponding fracture micromechanisms. The resulting information was used to estimate the percentage of the fracture path in the plastic zone through different microstructural features such as recrystallized grains, unrecrystallized regions, grain boundaries between recrystallized grains, grain boundaries between recrystallized and unrecrystallized regions, and so forth of the partially recrystallized microstructure of the 7050 aluminum alloy. In this way, the role of different microstructural features in the generation of the fracture surface was quantitatively characterized.

References cited in this section


Quantitative Fractography
Arun M. Gokhale, Georgia Institute of Technology

Acknowledgments
The Division of Materials Research, U.S. National Science Foundation (NSF) supported this work through the research grant DMR-9816618. K.L. Murty is the program manager for this NSF grant. Additional financial support was provided by the Air Force Office of Scientific Research (AFOSR) through the research grant F49620-01-1-0045, whose program manager is C.S. Hartley, and by United States Consortium for Automotive Research (USCAR) project (DOE) through a research grant administered by American Foundry Society for which the program managers are Richard Osborne, D.E. Penrod, and Joseph Santner. The financial support is gratefully acknowledged. The research results and data of numerous students and scientists from the author’s research group have been used in this article. The author would like to acknowledge and recognize the contributions of his students W.J. Drury, Surendra Mishra, W.T. Whited, N.U. Deshpande, M.D. Dighe, and G.R. Patel.

Quantitative Fractography
Arun M. Gokhale, Georgia Institute of Technology

References


72. K. Banerji, Ph.D. dissertation, Georgia Institute of Technology, 1987


90. A.M. Gokhale, private research, Georgia Institute of Technology, Atlanta, 2001


95. W.J. Drury, M.S. dissertation, Georgia Institute of Technology, 1988


