**Learning Objectives**
After careful study of this lecture you should be able to do the following:

1. Give a good review of historical development of material science.
2. Classify the material types.
3. Define and explain the meaning of nonmaterial.
4. Define the composite materials and its properties.
5. Explain and classify the advance materials.
6. Give differences between the materials types.

**HISTORICAL PERSPECTIVE**
Materials are probably more deep-seated in our culture than most of us realize. Transportation, housing, clothing, communication, recreation, and food production—virtually every segment of our everyday lives is influenced to one degree or another by materials. Historically, the development and advancement of societies have been intimately tied to the members’ ability to produce and manipulate materials to fill their needs. In fact, early civilizations have been designated by the level of their materials development (Stone Age, Bronze Age, Iron Age). The earliest humans had access to only a very limited number of materials, those that occur naturally: stone, wood, clay, skins, and so on. With time they discovered techniques for producing materials that had properties superior to those of the natural ones; these new materials included pottery and various metals. Furthermore, it was discovered that the properties of a material could be altered by heat treatments and by the addition of other substances. At this point, materials utilization was totally a selection process that involved deciding from a given, rather limited set of materials the one best suited for an application by virtue of its characteristics. It was not until relatively recent times that scientists came to understand the relationships between the structural elements of materials and their properties. This knowledge, acquired over approximately the past 100 years, has empowered them to fashion, to a large degree, the characteristics of materials. Thus, tens of thousands of different materials have evolved with rather specialized characteristics that meet the needs of our modern and complex society; these include metals, plastics, glasses, and fibers. The development of many technologies that make our existence so comfortable has been intimately associated with the accessibility of suitable materials. An advancement in the understanding of a material type is often the forerunner to the stepwise progression of a technology. For example, automobiles would not have been possible without the availability of inexpensive steel or some other comparable substitute. In our contemporary era, sophisticated electronic devices rely on components that are made from what are called semiconducting materials.

**CLASSIFICATION OF MATERIALS**
Solid materials have been conveniently grouped into three basic classifications: metals, ceramics, and polymers. This scheme is based primarily on chemical makeup and atomic structure, and most materials fall into one distinct grouping or another, although there are some intermediates. In addition, there are the composites, combinations of two or more of the above three basic material classes.

**Metals**
Materials in this group are composed of one or more metallic elements (such as iron, aluminum, copper, titanium, gold, and nickel), and often also nonmetallic elements (for example, carbon, nitrogen, and oxygen) in relatively small amounts. Atoms in metals and their alloys are arranged in a very orderly manner, and in comparison to the ceramics and polymers, are relatively dense. With regard to mechanical characteristics, these materials are relatively stiff and strong, yet are ductile (i.e., capable of large amounts of deformation without fracture), and are resistant to fracture, which accounts for their widespread use in structural applications. Metallic materials have large numbers of nonlocalized electrons; that is, these electrons are not bound to particular atoms. Many properties of metals are directly attributable to these electrons. For example, metals are extremely good conductors of electricity and heat, and are not transparent to visible light; a polished metal surface has a lustrous appearance. In addition, some of the metals (viz., Fe, Co, and Ni) have desirable magnetic properties.

**Ceramics**
Ceramics are compounds between metallic and nonmetallic elements; they are most frequently oxides, nitrides, and carbides. For example, some of the common ceramic materials include aluminum oxide (or alumina, Al₂O₃), silicon dioxide (or silica, SiO₂), silicon carbide (SiC), silicon nitride (Si₃N₄), and, in addition, what some refer to as the traditional ceramics—those composed of clay minerals (i.e., porcelain), as well as cement, and glass. With regard to mechanical behavior, ceramic materials are relatively stiff and strong—stiffnesses and strengths are comparable to those of the metal. In addition, ceramics are typically very hard. On the other hand, they are extremely brittle (lack ductility), and are highly susceptible to fracture. These materials are typically insulative to the passage of heat and electricity (i.e., have low electrical conductivities), and are more resistant to high temperatures and harsh environments than metals and polymers. With regard to optical characteristics, ceramics may be transparent, translucent, or opaque, and some of the oxide ceramics (e.g., Fe₂O₃) exhibit magnetic behavior.

**Polymers**
Polymers include the familiar plastic and rubber materials. Many of them are organic compounds that are chemically based on carbon, hydrogen, and other nonmetallic elements (viz., O, N, and Si). Furthermore, they have very large molecular structures, often chain-like in nature that have a backbone of carbon atoms. Some of the common and familiar polymers are polyethylene (PE), nylon, poly(vinyl chloride) (PVC), polycarbonate (PC), polystyrene (PS), and silicone rubber. These materials typically have low densities, whereas their mechanical characteristics are generally dissimilar to the metallic and ceramic materials—they are not as stiff nor as strong as
that are represented by different combinations of metals, ceramics, and polymers. Furthermore, some naturally-occurring materials are within a polymeric material (normally an epoxy or polyester). The glass fibers are relatively strong and stiff (but also brittle), whereas also considered to be composites—for example, wood and bone. However, most of those we consider in our discussions are synthetic (or man-made) composites. One of the most common and familiar composites is fiberglass, in which small glass fibers are embedded within a polymeric material (normally an epoxy or polyester). The glass fibers are relatively strong and stiff (but also brittle), whereas the polymer is ductile (but also weak and flexible). Thus, the resulting fiberglass is relatively stiff, strong, flexible, and ductile. In addition, it has a low density. Another of these technologically important materials is the “carbon fiberreinforced polymer” (or “CFRP”) composite—carbon fibers that are embedded within a polymer. These materials are stiffer and stronger than the glass fiber-reinforced materials, yet they are more expensive. The CFRP composites are used in some aircraft and aerospace applications, as well as high-tech sporting equipment (e.g., bicycles, golf clubs, tennis rackets, and skis/snowboards).

**ADVANCED MATERIALS**

Materials that are utilized in high-technology (or high-tech) applications are sometimes termed advanced materials. By high technology we mean a device or product that operates or functions using relatively intricate and sophisticated principles; examples include electronic equipment (camcorders, CD/DVD players, etc.), computers, fiber-optic systems, spacecraft, aircraft, and military rocketry. These advanced materials are typically traditional materials whose properties have been enhanced, and, also newly developed, high-performance materials. Furthermore, they may be of all material types (e.g., metals, ceramics, polymers), and are normally expensive. Advanced materials include semiconductors, biomaterials, and what we may term “materials of the future” (that is, smart materials and nanoengineered materials), which we discuss below.

**Semiconductors**

Semiconductors have electrical properties that are intermediate between the electrical conductors (viz. metals and metal alloys) and insulators (viz. ceramics and polymers). Furthermore, the electrical characteristics of these materials are extremely sensitive to the presence of minute concentrations of impurity atoms, for which the concentrations may be controlled over very small spatial regions. Semiconductors have made possible the advent of integrated circuitry that has totally revolutionized the electronics and computer industries over the past three decades.

**Biomaterials**

Biomaterials are employed in components implanted into the human body for replacement of diseased or damaged body parts. These materials must not produce toxic substances and must be compatible with body tissues (i.e., must not cause adverse biological reactions). All of the above materials—metals, ceramics, polymers, composites, and semiconductors—may be used as biomaterials.

**Materials of the Future**

**Smart Materials**

Smart (or intelligent) materials are a group of new and state-of-the-art materials now being developed that will have a significant influence on many of our technologies. The adjective “smart” implies that these materials are able to sense changes in their environments and then respond to these changes in predetermined manners—traits that are also found in living organisms. In addition, this “smart” concept is being extended to rather sophisticated systems that consist of both smart and traditional materials. Components of a smart material (or system) include some type of sensor (that detects an input signal), and an actuator (that performs a responsive and adaptive function). Actuators may be called upon to change shape, position, natural frequency, or mechanical characteristics in response to changes in temperature, electric fields, and/or magnetic fields. Four types of materials are commonly used for actuators: shape memory alloys, piezoelectric ceramics, magnetostrictive materials, and electro rheological/magnetorheological fluids. Shape memory alloys are metals that, after having been deformed, revert back to their original shapes when temperature is changed. Piezoelectric ceramics expand and contract in response to an applied electric field (or voltage); conversely, they also generate an electric field when their dimensions are altered. The behavior of magnetostrictive materials is analogous to that of the piezoelectrics, except that they are responsive to magnetic fields. Also, electro rheological and magnetorheological fluids are liquids that experience dramatic changes in viscosity upon the application of electric and magnetic fields, respectively. Materials/devices employed as sensors include optical fibers, piezoelectric materials (including some polymers), and microelectromechanical devices. For example, one type of smart system is used in helicopters to reduce aerodynamic cockpit noise that is created by the rotating rotor blades. Piezoelectric sensors inserted into the blades monitor blade stresses and deformations; feedback signals from these sensors are fed into a computer-controlled adaptive device, which generates noise-canceling antinoise.

**Composites**

A composite is composed of two (or more) individual materials, which come from the categories discussed above—viz., metals, ceramics, and polymers. The design goal of a composite is to achieve a combination of properties that is not displayed by any single material, and also to incorporate the best characteristics of each of the component materials. A large number of composite types exist that are represented by different combinations of metals, ceramics, and polymers. Furthermore, some naturally-occurring materials are also considered to be composites—for example, wood and bone. However, most of those we consider in our discussions are synthetic (or man-made) composites. One of the most common and familiar composites is fiberglass, in which small glass fibers are embedded within a polymeric material (normally an epoxy or polyester). The glass fibers are relatively strong and stiff (but also brittle), whereas the polymer is ductile (but also weak and flexible). Thus, the resulting fiberglass is relatively stiff, strong, flexible, and ductile. In addition, it has a low density. Another of these technologically important materials is the “carbon fiberreinforced polymer” (or “CFRP”) composite—carbon fibers that are embedded within a polymer. These materials are stiffer and stronger than the glass fiber-reinforced materials, yet they are more expensive. The CFRP composites are used in some aircraft and aerospace applications, as well as high-tech sporting equipment (e.g., bicycles, golf clubs, tennis rackets, and skis/snowboards).
**Nanoengineered Materials**

Until very recent times the general procedure utilized by scientists to understand the chemistry and physics of materials has been to begin by studying large and complex structures, and then to investigate the fundamental building blocks of these structures that are smaller and simpler. This approach is sometimes termed “topdown” science. However, with the advent of scanning probe microscopes, which permit observation of individual atoms and molecules, it has become possible to manipulate and move atoms and molecules to form new structures and, thus, design new materials that are built from simple atomic-level constituents (i.e., “materials by design”). This ability to carefully arrange atoms provides opportunities to develop mechanical, electrical, magnetic, and other properties that are not otherwise possible. We call this the “bottom-up” approach, and the study of the properties of these materials is termed “nanotechnology”; the “nano” prefix denotes that the dimensions of these structural entities are on the order of a nanometer ($10^{-9}$ m).
Learning Objectives
After careful study of this lecture you should be able to do the following:
1. Describe the difference in atomic/molecular structure between crystalline and noncrystalline materials.
2. Draw unit cells for face-centered cubic, body centered cubic, and hexagonal close-packed crystal structures.
3. Derive the relationships between unit cell edge length and atomic radius for face-centered cubic and body-centered cubic crystal structures.
4. Compute the densities for metals having face centered cubic and body-centered cubic crystal structures given their unit cell dimensions.
5. Given three direction index integers, sketch the direction corresponding to these indices within a unit cell.
6. Specify the Miller indices for a plane that has been drawn within a unit cell.

FUNDAMENTAL CONCEPTS
Solid materials may be classified according to the regularity with which atoms or ions are arranged with respect to one another. A crystalline material is one in which the atoms are situated in a repeating or periodic array over large atomic distances; that is, long-range order exists, such that upon solidification, the atoms will position themselves in a repetitive three-dimensional pattern, in which each atom is bonded to its nearest-neighbor atoms. All metals, many ceramic materials, and certain polymers form crystalline structures under normal solidification conditions. For those that do not crystallize, this long-range atomic order is absent; therefore solid materials found in nature in crystalline (single crystal and polycrystalline) and noncrystalline (amorphous).

Some of the properties of crystalline solids depend on the crystal structure of the material, the manner in which atoms, ions, or molecules are spatially arranged. There is an extremely large number of different crystal structures all having long-range atomic order; these vary from relatively simple structures for metals to exceedingly complex ones, as displayed by some of the ceramic and polymeric materials. The present discussion deals with several common metallic crystal structures. When describing crystalline structures, atoms (or ions) are thought of as being solid spheres having...
well-defined diameters. This is termed the *atomic hard sphere model* in which spheres representing nearest-neighbor atoms touch one another. In this particular case all the atoms are identical. Sometimes the term **lattice** is used in the context of crystal structures; in this sense “lattice” means a three-dimensional array of points coinciding with atom positions (or sphere centers). Lattice to be an array of points in space so arranged that each point has identical surroundings. The smallest unit, or unit cell, can be obtained by constructing planes through points, and the lines resulting from the intersection of the planes at lattice points define the unit cell. Figure above shows a unit cell in darker outline and defined by the cell parameters \( a, b, c \) and angles (not shown) \( \alpha, \beta, \gamma \) called lattice parameters. The angles are defined where \( \alpha \) is the angle between vectors \( a \) and \( b \), \( \beta \) the angle between \( a \) and \( c \), and \( \gamma \) is the angle between \( b \) and \( c \). It should be noticed that the unit cell so defined embodies the symmetry of the entire lattice. The entire lattice can be generated by simply translating the unit cell by \( |a| \) in the \( a \) direction, by \( |b| \) in the \( b \) direction, and by \( |c| \) in the \( c \) direction.

Thus translation becomes an important operation in understanding the long-range ordering represented by the lattice. The question as to how many different kinds of unit cells are necessary to fill all space by translation and how to accomplish this for all possible symmetries is a solved mathematical question for which we herein accept the solution without proof. The lattices that accomplish this task are called Bravais lattices, and there are 14 such Bravais lattices, as shown in Figure. These 14 Bravais lattices are organized into 7 crystal systems according to the basic symmetry that the lattice possesses: cubic, trigonal, hexagonal, orthorhombic, rhombohedral, monoclinic, and triclinic. Some of these systems can have different lattices: simple or primitive (P), body centered (B), and face centered (F). The number of lattice points for a unit cell, \( N \), is calculated by counting the points that bound and are interior to the cell and then considering the sharing of points by adjacent cells. For example, the eight lattice points at the cell corners in the unit cell in Figure above are each shared by eight adjacent cells (Nc), the points on the face of a cell by two cells (Nf), and of course, the interior points (Ni) belong solely to the cell in question. Hence the following relationship summarizes this:

\[
N = Ni + \frac{Nf}{2} + \frac{Nc}{8}
\]

Therefore \( N \) for SC, FCC and BCC are: 1, 4, and 2 respectively.

<table>
<thead>
<tr>
<th>Crystal System</th>
<th>Unit Cell Vectors</th>
<th>Unit Cell Angles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cubic</td>
<td>( a = b = c )</td>
<td>( \alpha = \beta = \gamma = 90^\circ )</td>
</tr>
<tr>
<td>Tetragonal</td>
<td>( a = b \neq c )</td>
<td>( \alpha = \beta = \gamma = 90^\circ )</td>
</tr>
<tr>
<td>Orthorhombic</td>
<td>( a \neq b \neq c )</td>
<td>( \alpha = \beta = \gamma = 90^\circ )</td>
</tr>
<tr>
<td>Hexagonal (trigonal)</td>
<td>( a = b \neq c )</td>
<td>( \alpha = \beta = 90^\circ, \gamma = 120^\circ )</td>
</tr>
<tr>
<td>Monoclinic</td>
<td>( a \neq b \neq c )</td>
<td>( \alpha = \beta = 90^\circ \neq \gamma )</td>
</tr>
<tr>
<td>Triclinic</td>
<td>( a \neq b \neq c )</td>
<td>( \alpha \neq \beta \neq \gamma \neq 90^\circ )</td>
</tr>
<tr>
<td>Rhombohedral</td>
<td>( a = b = c )</td>
<td>( \alpha = \beta = \gamma \neq 90^\circ )</td>
</tr>
</tbody>
</table>
Calculation the volume of the unit cell

The calculation of volume of unit cell depending in facts that the atoms considered as sphere geometry also we consider the side of cubic system $a$ there for the cubic side of FCC and BCC structure drive as the following:

\[(BD)^2 = (DC)^2 + (CB)^2\]
\[(4r)^2 = a^2 + a^2\]
\[4r = \sqrt{2} \ a\]
\[a = \frac{4r}{\sqrt{2}}\]

\[(DF)^2 = (DA)^2 + (AF)^2\]
\[(4r)^2 = a^2 + 2a^2\]
\[4r = \sqrt{3} \ a\]
\[a = \frac{4r}{\sqrt{3}}\]
Two other important characteristics of a crystal structure are the **coordination number** and the **atomic packing factor (APF)**. For metals, each atom has the same number of nearest-neighbor or touching atoms, which is the coordination number. For face-centered cubic, the coordination number is 12 and for body centered cubic 8. This may be confirmed by examination of Figure above for FCC; the front face atom has four corner nearest-neighbor atoms surrounding it, four face atoms that are in contact from behind, and four other equivalent face atoms residing in the next unit cell to the front, which is not shown. The APF is the sum of the sphere volumes of all atoms within a unit cell (assuming the atomic hard sphere model) divided by the unit cell volume—that is:

$$APF = \frac{\text{vol.of atoms per unit cell}}{\text{vol.of unit cell}}$$

As an example the APF for FCC structure calculated as follow:

Vol. of atoms per unit cell = $(4) \times \frac{4}{3} \pi r^3$

Vol. of unit cell = $(a)^3 = 16 r^3 \sqrt{2}$

Therefore the APF for FCC = 0.74 and for BCC is 0.68 (prove)

Depending on crystal structure we can simply calculated theoretical density as:

$$\sigma = \frac{n \times A}{V_c \times N_a}$$

Where $n$ is the no. of atoms per unit cell, $A$ atomic weight, $V_c$ vol. of unite cell and $N_a$ Avogadro's no.

As an example calculate the density of copper with an atomic radius 0.128 nm, atomic weight 63.5 g/mol and with FCC structure.

$$\sigma = \frac{n \times A}{V_c \times N_a} = \frac{(4\text{ atoms/unit cell} \times 63.5 \text{ g/mol})}{[16\sqrt{2} \times (1.28 \times 10^{-8} \text{ cm}^3) / \text{unit cell}] \times (6.023 \times 10^{23} \text{ atoms/mol})} = 8.89 \text{ g/cm}^3$$
POINT COORDINATES

The position of any point located within a unit cell may be specified in terms of its coordinates as fractional multiples of the unit cell edge lengths (i.e., in terms of \(a, b,\) and \(c\)). To illustrate, consider the unit cell and the point \(P\) situated therein as shown in Figure below. We specify the position of \(P\) in terms of the generalized coordinates \(q, r,\) and \(s\) where \(q\) is some fractional length of \(a\) along the \(x\) axis, \(r\) is some fractional length of \(b\) along the \(y\) axis, and similarly for \(s\). Thus, the position of \(P\) is designated using coordinates \(q\ r\ s\) with values that are less than or equal to unity. Furthermore, we have chosen not to separate these coordinates by commas or any other punctuation marks (which is the normal convention).

![Diagram of a unit cell with a point P located within it, showing generalized coordinates q, r, and s.](image)

**Location of Point Having Specified Coordinates**

For the unit cell shown in the accompanying sketch (a), locate the point having coordinates \(1/2\ 1/2\).

**Solution**

From sketch (a), edge lengths for this unit cell are as follows: \(a = 0.48\) nm, \(b = 0.46\) nm, and \(c = 0.40\) nm. Furthermore, in light of the above discussion, fractional lengths are \(q = 1/2, r = 1,\) and \(s = 1/2\). Therefore, first we move from the origin of the unit cell (point \(M\)) \(qa = 1/2(0.48\) nm\) = 0.12 nm units along the \(x\) axis (to point \(N\)), as shown in the (b) sketch. Similarly, we proceed \(rb = 1)(0.46\) nm\) = 0.46 nm parallel to the \(y\) axis, from point \(N\) to point \(O\). Finally, we move from this position, \(sc = 1/2(0.40\) nm\) = 0.20 nm units parallel to the \(z\) axis to point \(P\) as noted again in sketch (b). This point \(P\) then corresponds to the \(1/2\ 1/2\) point coordinates.
Naming Planes

The accepted system for naming planes is the Miller index notation. Naming planes is linked with finding the intersections of the planes with the basic lattice vectors that define the fundamental Bravais lattice for a structure. However, simply using intersections is sometimes cumbersome because interesting planes are often parallel to one or more of the unit cell lattice vectors. In this case the intersection is at infinity, and either the word or infinity symbol \( \infty \) needs to be carried along in the nomenclature. In order to obviate this situation, the reciprocals of the intercepts are taken so that \( \frac{1}{\infty} \) becomes 0. Fractions obtained after taking the reciprocal of the intercepts are cleared, and the resulting sets of usually three whole numbers (an exception to three is covered below) are placed in between rounded brackets ( ) indicative of specific planes. Figure below illustrates the Miller index system. Figure (a) shows the \( \mathbf{a}, \mathbf{b}, \mathbf{c} \) axes with the larger diagonal plane intercepts of \( a = 1, b = 1, \) and \( c = \infty \) (i.e., the plane is parallel to the \( \mathbf{c} \) axis). The reciprocals are \( 1/1, 1/1, \) and \( 1/\infty \), which yield the (110) plane. The small plane has intercepts of \( a = 1/3, b = 1/2, \) and \( c = 1 \). The corresponding reciprocals are \( 3, 2, 1, \) so the plane is the \( (321) \) plane. In Figure (b) the larger plane has intercepts of \( 1, 1, 1, \) so the plane is \( (111) \). The smaller plane has intercepts of \( 1/3, 1/2, 2/3 \). The reciprocals are \( 3, 2, 3/2 \) and, upon clearing fractions becomes the \( (643) \) plane. Figure (c) shows the shadowed plane with intercepts of \( \infty, 1/2, \infty \), which yields the \( (020) \) plane. We can imagine the planes perpendicular to and bisecting the shaded \( (020) \) plane. These planes would be the either the \( (200) \) plane or the \( (002) \) plane. These three planes comprise a family of planes denoted by \( \{200\} \). Similarly in Figure (c) the planes that bound the figure are \( \{100\} \), namely the family of \( (100) \) planes.

<table>
<thead>
<tr>
<th>Point Number</th>
<th>Fractional Lengths</th>
<th>Point Coordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( x ) axis</td>
<td>( y ) axis</td>
</tr>
<tr>
<td>1</td>
<td>1 0 0</td>
<td>0 0 0</td>
</tr>
<tr>
<td>2</td>
<td>1 1 0</td>
<td>0 0 0</td>
</tr>
<tr>
<td>3</td>
<td>1 1 0</td>
<td>0 0 0</td>
</tr>
<tr>
<td>4</td>
<td>0 1 0</td>
<td>0 0 0</td>
</tr>
<tr>
<td>5</td>
<td>1/3 1/3 1/3</td>
<td>0 0 0</td>
</tr>
<tr>
<td>6</td>
<td>0 1/2 0</td>
<td>0 0 0</td>
</tr>
<tr>
<td>7</td>
<td>1 0 1</td>
<td>0 0 0</td>
</tr>
<tr>
<td>8</td>
<td>1 1 1</td>
<td>0 0 0</td>
</tr>
<tr>
<td>9</td>
<td>0 1 1</td>
<td>0 0 0</td>
</tr>
</tbody>
</table>
These planes with the highest atom/molecule concentration also possess the highest bond density and thus electron density. Therefore all those properties that correlate with atom, bond, and/or electron density are determined by the low index planes of the material. It is easy to see why when describing the properties of a crystal it is important to also specify the direction in which the property was measured and the appropriate plane involved.

**Lattice Directions**

In order to name a direction, one must first construct a line parallel to the direction to be named, but that intersects the origin of the lattice vectors. Then at any point on the constructed line a perpendicular is dropped to each lattice vector. The intercepts to the lattice vectors cleared of fractions are the direction indexes. Figure below shows a cubic unit cell with low index planes and directions. Notice that the directions in square brackets are perpendicular to the planes in rounded brackets with the same indices.

**CLOSE-PACKED CRYSTAL STRUCTURES**

both face-centered cubic and hexagonal close-packed crystal structures have atomic packing factors of 0.74, which is the most efficient packing of equal-sized spheres or atoms. In addition to unit cell representations, these two crystal structures may be described in terms of close-packed planes of atoms (i.e., planes having a maximum atom or sphere-packing density); a portion of one such plane is illustrated in Figure below. Both crystal structures may be generated by the stacking of these close-packed planes on top of one another; the difference between the two structures lies in the stacking sequence. Let the centers of all the atoms in one close-packed plane be labeled A. Associated with this plane are two sets of equivalent triangular depressions formed by three adjacent atoms, into which the next close-packed plane of atoms may rest. Those having the triangle vertex pointing up are arbitrarily designated as B positions, while the remaining depressions are those with the down vertices, which are marked C.
A second close-packed plane may be positioned with the centers of its atoms over either \( B \) or \( C \) sites; at this point both are equivalent. Suppose that the \( B \) positions are arbitrarily chosen; the stacking sequence is termed \( AB \), which is illustrated in Figure (b). The real distinction between FCC and HCP lies in where the third close-packed layer is positioned. For HCP, the centers of this layer are aligned directly above the original \( A \) positions. This stacking sequence, \( ABABAB \ldots \), is repeated over and over.

**X-RAY DIFFRACTION: DETERMINATION OF CRYSTAL STRUCTURES**

Historically, much of our understanding regarding the atomic and molecular arrangements in solids has resulted from x-ray diffraction investigations; furthermore, x-rays are still very important in developing new materials. We will now give a brief overview of the diffraction phenomenon and how, using x-rays, atomic interplanar distances and crystal structures are deduced.

**The Diffraction Phenomenon**

Diffraction occurs when a wave encounters a series of regularly spaced obstacles that (1) are capable of scattering the wave, and (2) have spacings that are comparable in magnitude to the wavelength. Furthermore, diffraction is a consequence of specific phase relationships established between two or more waves that have been scattered by the obstacles. Consider waves 1 and 2 in Figure (a) which have the same wavelength and are in phase at point. Now let us suppose that both waves are scattered in such a way that they traverse different paths. The phase relationship between the scattered waves, which will depend upon the difference in path length, is important. One possibility results when this path length difference is an integral number of wavelengths. These scattered waves (now labeled and) are still in phase. They are said to mutually reinforce (or constructively interfere with) one another; and, when amplitudes are added, the wave shown on the right side of the figure results. This is a manifestation of **diffraction**, and we refer to a diffracted beam as one composed of a large number of scattered waves that mutually reinforce one another. Other phase relationships are possible between scattered waves that will not lead to this mutual reinforcement. The other extreme is that demonstrated in Figure (b), wherein the path length difference after scattering is some integral number of half wavelengths. The scattered waves are out of phase that is, corresponding amplitudes cancel or annul one another, or destructively interfere (i.e., the resultant wave has zero amplitude), as indicated on the extreme right side of the figure. Of course, phase relationships intermediate between these two extremes exist, resulting in only partial reinforcement.
X-Ray Diffraction and Bragg’s Law

X-rays are a form of electromagnetic radiation that has high energies and short wavelengths on the order of the atomic spacings for solids. When a beam of x-rays impinges on a solid material, a portion of this beam will be scattered in all directions by the electrons associated with each atom or ion that lies within the beam’s path. Let us now examine the necessary conditions for diffraction of x-rays by a periodic arrangement of atoms.

Consider the two parallel planes of atoms $A-A'$ and $B-B'$ in Figure below, which have the same $h$, $k$, and $l$ Miller indices and are separated by the interplanar spacing $d_{hkl}$. Now assume that a parallel, monochromatic, and coherent (in-phase) beam of x-rays of wavelength $\lambda$ is incident on these two planes at an angle $\theta$. Two rays in this beam, labeled 1 and 2, are scattered by atoms $P$ and $Q$. Constructive interference of the scattered rays 1' and 2' occurs also at an angle $\theta$ to the planes, if the path length difference between 1-$P$-1' and 2-$Q$-2' is equal to a whole number, $n$, of wavelengths. That is, the condition for diffraction is

$$ n\lambda = d_{hkl}\sin\theta + d_{hkl}\sin\theta $$

$$ n\lambda = 2d_{hkl}\sin\theta $$

1. X-Ray Diffraction and Bragg's Law

X-rays are a form of electromagnetic radiation that has high energies and short wavelengths on the order of the atomic spacings for solids. When a beam of x-rays impinges on a solid material, a portion of this beam will be scattered in all directions by the electrons associated with each atom or ion that lies within the beam's path. Let us now examine the necessary conditions for diffraction of x-rays by a periodic arrangement of atoms.

Consider the two parallel planes of atoms $A-A'$ and $B-B'$ in Figure below, which have the same $h$, $k$, and $l$ Miller indices and are separated by the interplanar spacing $d_{hkl}$. Now assume that a parallel, monochromatic, and coherent (in-phase) beam of x-rays of wavelength $\lambda$ is incident on these two planes at an angle $\theta$. Two rays in this beam, labeled 1 and 2, are scattered by atoms $P$ and $Q$. Constructive interference of the scattered rays 1' and 2' occurs also at an angle $\theta$ to the planes, if the path length difference between 1-$P$-1' and 2-$Q$-2' is equal to a whole number, $n$, of wavelengths. That is, the condition for diffraction is

$$ n\lambda = d_{hkl}\sin\theta + d_{hkl}\sin\theta $$

$$ n\lambda = 2d_{hkl}\sin\theta $$
Equation above is known as Bragg’s law; also, \( n \) is the order of reflection, which may be any integer (1, 2, 3, …) consistent with not exceeding unity. Thus, we have a simple expression relating the x-ray wavelength and interatomic spacing to the angle of the diffracted beam. If Bragg’s law is not satisfied, then the interference will be nonconstructive in nature so as to yield a very low-intensity diffracted beam. The magnitude of the distance between two adjacent and parallel planes of atoms (i.e., the interplanar spacing) is a function of the Miller indices \((h, k, l)\) as well as the lattice parameter(s). For example, for crystal structures that has cubic symmetry,

\[
d_{hkl} = \frac{a}{\sqrt{h^2 + k^2 + l^2}}
\]

In which \( a \) is the lattice parameter (unit cell edge length).

Bragg’s law is a necessary but not sufficient condition for diffraction by real crystals. It specifies when diffraction will occur for unit cells having atoms positioned only at cell corners. However, atoms situated at other sites (e.g., face and interior unit cell positions as with FCC and BCC) act as extra scattering centers, which can produce out-of-phase scattering at certain Bragg angles. The net result is the absence of some diffracted beams that, according to Bragg’s law, should be present. For example, for the BCC crystal structure, \( h \) and \( l \) must be even if diffraction is to occur, whereas for FCC, \( h, k, \) and \( l \) must all be either odd or even.

**Diffraction Techniques**

One common diffraction technique employs a powdered or polycrystalline specimen consisting of many fine and randomly oriented particles that are exposed to monochromatic x-radiation. Each powder particle (or grain) is a crystal, and having a large number of them with random orientations ensures that some particles are properly oriented such that every possible set of crystallographic planes will be available for diffraction. The diffractometer is an apparatus used to determine the angles at which diffraction occurs for powdered specimens; its features are represented schematically in Figure below. A specimen \( S \) in the form of a flat plate is supported so that rotations about the axis labeled \( O \) are possible; this axis is perpendicular to the plane of the page. The monochromatic x-ray beam is generated at point \( T \), and the intensities of diffracted beams are detected with a counter labeled \( C \) in the figure. The specimen, x-ray source, and counter are all coplanar. The counter is mounted on a movable carriage that may also be rotated about the \( O \) axis; its angular position in terms of is marked on a graduated scale. Carriage and specimen are mechanically coupled such that
a rotation of the specimen through is accompanied by a rotation of the counter; this assures that the incident and reflection angles are maintained equal to one another.

Collimators are incorporated within the beam path to produce a well-defined and focused beam. Utilization of a filter provides a near-monochromatic beam. As the counter moves at constant angular velocity, a recorder automatically plots the diffracted beam intensity (monitored by the counter) as a function of is termed the **diffraction angle**, which is measured experimentally. Figure below shows a diffraction pattern for a powdered specimen of lead. The high-intensity peaks result when the Bragg diffraction condition is satisfied by some set of crystallographic planes. These peaks are plane-indexed in the figure. Other powder techniques have been devised wherein diffracted beam intensity and position are recorded on a photographic film instead of being measured by a counter. One of the primary uses of x-ray diffractometry is for the determination of crystal structure. The unit cell size and geometry may be resolved from the angular positions of the diffraction peaks, whereas arrangement of atoms within the unit cell is associated with the relative intensities of these peaks.
Learning Objectives
After careful study of this lecture you should be able to do the following:
1. Describe both vacancy and self-interstitial crystalline defects.
2. Calculate the equilibrium number of vacancies in a material at some specified temperature, given the relevant constants.
3. For each of edge and screw, dislocations:
   a) Describe and make a drawing of the dislocation
   b) Note the location of the dislocation line, and
   c) Indicate the direction along which the dislocation line extends.

INTRODUCTION
Thus far it has been tacitly assumed that perfect order exists throughout crystalline materials on an atomic scale. However, such an idealized solid does not exist; all contain large numbers of various defects or imperfections. As a matter of fact, many of the properties of materials are profoundly sensitive to deviations from crystalline perfection; the influence is not always adverse, and often specific characteristics are deliberately fashioned by the introduction of controlled amounts or numbers of particular defects. By “crystalline defect” is meant a lattice irregularity having one or more of its dimensions on the order of an atomic diameter. Classification of crystalline imperfections is frequently made according to geometry or dimensionality of the defect. Several different imperfections are discussed, including point defects (those associated with one or two atomic positions), linear (or one-dimensional) defects, as well as interfacial defects, or boundaries, which are two-dimensional.

VACANCIES AND SELF-INTERSTITIALS
The simplest of the point defects is a vacancy, or vacant lattice site, one normally occupied from which an atom is missing (Figure below). All crystalline solids contain vacancies and, in fact, it is not possible to create such a material that is free of these defects. The necessity of the existence of vacancies is explained using principles of thermodynamics; in essence, the presence of vacancies increases the entropy (i.e., the randomness) of the crystal. The equilibrium number of vacancies for a given quantity of material depends on and increases with temperature according to

\[ N_v = N \exp\left(-\frac{Q_v}{kT}\right) \]

In this expression, \( N \) is the total number of atomic sites, \( Q_v \) is the energy required for the formation of a vacancy, \( T \) is the absolute temperature in kelvins, and \( k \) is the gas or Boltzmann’s constant. The value of \( k \) is \( (1.38 \times 10^{-23} \text{ J/K}) \), or \( (8.6174 \times 10^{-5} \text{ eV/K}) \), depending on the units of \( Q_v \). Thus, the number of vacancies increases exponentially with temperature; that is, as \( T \) in increases, so also does the expression \( \exp(-Q_v/kT) \). For most metals, the fraction of \( N_v/N \) vacancies just below the melting temperature is on the order of \( 10^{-4} \); that is, one lattice site out of 10,000 will be empty. A self-interstitial is an atom from the crystal that is crowded into an interstitial site, a small void space that under ordinary circumstances is not occupied. This kind of defect is also represented in Figure below. In metals, a self-interstitial introduces relatively large distortions in the surrounding lattice because the atom is substantially larger than the interstitial position in which it is situated. Consequently, the formation of this defect is not highly probable, and it exists in very small concentrations, which are significantly lower than for vacancies.

Example: Calculate the equilibrium number of vacancies per cubic meter for copper at 1000°C. The energy for vacancy formation is 0.9 eV/atom; the atomic weight and density (at 1000°C) for copper are 63.5 g/mol and 8.4 g/cm³, respectively.

Solution:

\[ N = \frac{N_A \rho}{A} \]

\[ N = \frac{(6.023 \times 10^{23} \text{ atoms/mol})(8.4 \text{ g/cm}^3)(10^6 \text{ cm}^3/m^3)}{(63.5 \text{ g/mol})} \]

\[ = 8 \times 10^{28} \text{ atoms/m}^3 \]
Impurity point defects are found in solid solutions, of which there are two types: **substitutional** and **interstitial**. For the substitutional type, solute or impurity atoms replace or substitute for the host atoms (Figure below). There are several features of the solute and solvent atoms that determine the degree to which the former dissolves in the latter, as follows:

1. **Atomic size factor.** Appreciable quantities of a solute may be accommodated in this type of solid solution only when the difference in atomic radii between the two atom types is less than about ±15%. Otherwise the solute atoms will create substantial lattice distortions and a new phase will form.

2. **Crystal structure.** For appreciable solid solubility the crystal structures for metals of both atom types must be the same.

3. **Electronegativity.** The more electropositive one element and the more electronegative the other, the greater is the likelihood that they will form an intermetallic compound instead of a substitutional solid solution.

4. **Valences.** Other factors being equal, a metal will have more of a tendency to dissolve another metal of higher valency than one of a lower valency.

An example of a substitutional solid solution is found for copper and nickel. These two elements are completely soluble in one another at all proportions. With regard to the aforementioned rules that govern degree of solubility, the atomic radii for copper and nickel are 0.128 and 0.125 nm, respectively, both have the FCC crystal structure, and their electronegativities are 1.9 and 1.8 finally, the most common valences are for copper +1 (although it sometimes can be +2 ) and +2 for nickel. For interstitial solid solutions, impurity atoms fill the voids or interstices among the host atoms. For metallic materials that have relatively high atomic packing factors, these interstitial positions are relatively small. Consequently, the atomic diameter of an interstitial impurity must be substantially smaller than that of the host atoms. Normally, the maximum allowable concentration of interstitial impurity atoms is low (less than 10%). Even very small
Impurity atoms are ordinarily larger than the interstitial sites, and as a consequence they introduce some lattice strains on the adjacent host atoms. Carbon forms an interstitial solid solution when added to iron; the maximum concentration of carbon is about 2%. The atomic radius of the carbon atom is much less than that for iron: 0.071 nm versus 0.124 nm. There is also the possibility of forming charged point defects in ionic crystals, and these are known as Frenkel and Schottky defects. The Frenkel defect is an interstitial of a charged atom that, when formed, creates two regions of different polarity. It is often referred to as an interstitial pair defect. The Schottky defect is also a pair defect, but it is the absence of both ions. Overall charge neutrality must be maintained for the formation of charged defects.

**DISLOCATIONS – LINEAR DEFECTS**

A dislocation is a linear or one-dimensional defect around which some of the atoms are misaligned. One type of dislocation is shown in Figure below: an extra portion of a plane of atoms, or half-plane, the edge of which terminates within the crystal. This is termed an *edge dislocation*; it is a linear defect that centers around the line that is defined along the end of the extra half-plane of atoms. This is sometimes termed the *dislocation line*, which, for the edge dislocation is perpendicular to the plane of the page. Within the region around the dislocation line there is some localized lattice distortion. The atoms above the dislocation line are squeezed together, and those below are pulled apart; this is reflected in the slight curvature for the vertical planes of atoms as they bend around this extra half-plane. The magnitude of this distortion decreases with distance away from the dislocation line; at positions far removed, the crystal lattice is virtually perfect. Sometimes the edge dislocation is represented by the symbol \( \equiv \) which also indicates the position of the dislocation line. An edge dislocation may also be formed by an extra half-plane of atoms that is included in the bottom portion of the crystal; its designation is \( \equiv \).

Another type of dislocation, called a *screw dislocation*, exists, which may be thought of as being formed by a shear stress that is applied to produce the distortion shown in (Figure a) below: the upper front region of the crystal is shifted one atomic distance to the right relative to the bottom portion. The atomic distortion associated with a screw dislocation is also linear and along a dislocation line, line \( AB \) in (Figure b) below. The screw dislocation derives its name from the spiral or helical path or ramp that is traced around the dislocation line by the atomic planes of atoms. Sometimes the symbol \( \equiv \) is used to designate a screw dislocation. The magnitude and direction of the lattice distortion associated with a dislocation is expressed in terms of a *Burgers vector*, denoted by \( \vec{b} \). Burgers vectors are indicated in Figures above and below for edge and screw dislocations, respectively. Furthermore, the nature of a dislocation (i.e., edge or screw) is defined by the relative orientations of dislocation line and Burgers vector. For an edge, they are perpendicular, whereas for a screw, they are parallel. In addition, the Burgers vector is an element of the theory that has been developed to explain this type of deformation. Dislocations can be observed in crystalline materials using electron-microscopic...
techniques. In picture below, a high-magnification transmission electron micrograph, the dark lines are the dislocations. Virtually all crystalline materials contain some dislocations that were introduced during solidification, during plastic deformation, and as a consequence of thermal stresses that result from rapid cooling.

INTERFACIAL DEFECTS
Interfacial defects are boundaries that have two dimensions and normally separate regions of the materials that have different crystal structures and/or crystallographic orientations. These imperfections include external surfaces, grain boundaries, twin boundaries, stacking faults, and phase boundaries.

External Surfaces
One of the most obvious boundaries is the external surface, along which the crystal structure terminates. Surface atoms are not bonded to the maximum number of nearest neighbors, and are therefore in a higher energy state than the atoms at interior positions. The bonds of these surface atoms that are not satisfied give rise to a surface energy, expressed in units of energy per unit area (J/m² or erg/cm²). To reduce this energy, materials tend to minimize, if at all possible, the total surface area. For example, liquids assume a shape having a minimum area—the droplets become spherical. Of course, this is not possible with solids, which are mechanically rigid.

Grain Boundaries
Another interfacial defect, the grain boundary, as the boundary separating two small grains or crystals having different crystallographic orientations in polycrystalline materials. A grain boundary is represented schematically from an atomic perspective in Figure below. Within the boundary region, which is probably just several atom distances wide, there is some atomic mismatch in a transition from the crystalline orientation of one grain to that of an adjacent one. Various degrees of crystallographic misalignment between adjacent grains are possible. When this orientation mismatch is slight, on the order of a few degrees, then the term small- (or low-) angle grain boundary is used. The atoms are bonded less regularly along a grain boundary (e.g., bond angles are longer), and consequently, there is an interfacial or grain boundary energy similar to the surface energy described above. The magnitude of this energy is a function of the degree of misorientation, being larger for high-angle boundaries. Grain boundaries are more chemically reactive than the grains themselves as a consequence of this boundary energy. Furthermore, impurity atoms often preferentially segregate along these boundaries because of their higher energy state. The total interfacial energy is lower in large or coarse-grained materials than in fine-grained ones, since there is less total boundary area in the former. Grains grow at elevated
temperatures to reduce the total boundary energy. In spite of this disordered arrangement of atoms and lack of regular bonding along grain boundaries, a polycrystalline material is still very strong; cohesive forces within and across the boundary are present. Furthermore, the density of a polycrystalline specimen is virtually identical to that of a single crystal of the same material.

<table>
<thead>
<tr>
<th>Crystalline defects</th>
<th>0-dimension</th>
<th>1-dimension</th>
<th>2-dimension</th>
<th>3-dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Vacancies</td>
<td>Edge dislocation</td>
<td>External surface</td>
<td>Cracks and pores</td>
</tr>
<tr>
<td></td>
<td>Schottky and</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>frenkel</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Self-interstitial</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Impurities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Substitutional</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interstitial</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
INTRODUCTION

Corrosion involves the deterioration of a material as it reacts with its environment. Corrosion is the primary means by which metals deteriorate. Corrosion literally consumes the material reducing load carrying capability and causing stress concentrations. Corrosion is often a major part of maintenance cost and corrosion prevention is vital in many designs. Corrosion is not expressed in terms of a design property value like other properties but rather in more qualitative terms such as a material is immune, resistant, susceptible or very susceptible to corrosion. The corrosion process is usually electrochemical in nature, having the essential features of a battery. Corrosion is a natural process that commonly occurs because unstable materials, such as refined metals want to return to a more stable compound. For example, some metals, such as gold and silver, can be found in the earth in their natural, metallic state and they have little tendency to corrode. Iron is a moderately active metal and corrodes readily in the presence of water. The natural state of iron is iron oxide and the most common iron ore is Hematite with a chemical composition of Fe₂O₃. Rust, the most common corrosion product of iron, also has a chemical composition of Fe₂O₃. The difficulty in terms of energy required to extract metals from their ores is directly related to the ensuing tendency to corrode and release this energy. The electromotive force series (See table) is a ranking of metals with respect to their inherent reactivity. The most noble metal is at the top and has the highest positive electrochemical potential. The most active metal is at the bottom and has the most negative electrochemical potential.
MECHANISM OF CORROSION

Corrosion involves two chemical processes...oxidation and reduction. Oxidation is the process of stripping electrons from an atom and reduction occurs when an electron is added to an atom. The oxidation process takes place at an area known as the anode. At the anode, positively charged atoms leave the solid surface and enter into an electrolyte as ions. The ions leave their corresponding negative charge in the form of electrons in the metal which travel to the location of the cathode through a conductive path. At the cathode, the corresponding reduction reaction takes place and consumes the free electrons. The electrical balance of the circuit is restored at the cathode when the electrons react with neutralizing positive ions, such as hydrogen ions, in the electrolyte. From this description, it can be seen that there are four essential components that are needed for a corrosion reaction to proceed. These components are an anode, a cathode, an electrolyte with oxidizing species, and some direct electrical connection between the anode and cathode. Although atmospheric air is the most common environmental electrolyte, natural waters, such as seawater rain, as well as man-made solutions, are the environments most frequently associated with corrosion problems. A typical situation might involve a piece of metal that has anodic and cathodic regions on the same surface. If the surface becomes wet, corrosion may take place through ionic exchange in the surface water layer between the anode and cathode. Electron exchange will take place through the bulk metal. Corrosion will proceed at the anodic site according to a reaction such as

$$M \rightarrow M^{++} + 2e^-$$

Where M is a metal atom. The resulting metal cations (M++) are available at the metal surface to become corrosion products such as oxides, hydroxides, etc. The liberated electrons travel through the bulk metal (or another low resistance electrical connection) to the cathode, where they are consumed by cathodic reactions such as

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

The basic principles of corrosion that were just covered, generally apply to all corrosion situations except certain types of high temperature corrosion. However, the process of corrosion can be very straightforward but is often very complex due to variety of variable that can contribute to the process. A few of these variables are the composition of the material acting in the corrosion cell, the heat treatment and stress state of the materials, the composition of the electrolyte, the distance between the anode and the cathode, temperature, protective oxides and coating, etc.
FORMS OF CORROSION
It is convenient to classify corrosion according to the manner in which it is manifest. Metallic corrosion is sometimes classified into eight forms: uniform, galvanic, crevice, pitting, intergranular, selective leaching, erosion–corrosion, and stress corrosion. The causes and means of prevention of each of these forms are discussed briefly.

1-Uniform Attack
Uniform attack is a form of electrochemical corrosion that occurs with equivalent intensity over the entire exposed surface and often leaves behind a scale or deposit. In a microscopic sense, the oxidation and reduction reactions occur randomly over the surface. Some familiar examples include general rusting of steel and iron and the tarnishing of silverware. This is probably the most common form of corrosion. It is also the least objectionable because it can be predicted and designed for with relative ease.

2-Galvanic Corrosion
Galvanic corrosion occurs when two metals or alloys having different compositions are electrically coupled while exposed to an electrolyte. The less noble or more reactive metal in the particular environment will experience corrosion; the more inert metal, the cathode, will be protected from corrosion. For example, steel screws corrode when in contact with brass in a marine environment; or if copper and steel tubing are joined in a domestic water heater, the steel will corrode in the vicinity of the junction. Depending on the nature of the solution, one or more of the reduction reactions will occur at the surface of the cathode material. Again, the galvanic series (Table above) indicates the relative reactivities, in seawater, of a number of metals and alloys. When two alloys are coupled in seawater, the one lower in the series will experience corrosion. It is also worth noting from this series that some alloys are listed twice (e.g., nickel and the stainless steels), in both active and passive states. The rate of galvanic attack depends on the relative anode-to-cathode surface areas that are exposed to the electrolyte, and the rate is related directly to the cathode–anode area ratio; that is, for a given cathode area, a smaller anode will corrode more rapidly than a larger one. The reason for this is that corrosion rate depends on current density, the current per unit area of corroding surface, and not simply the current. Thus, a high current density results for the anode when its area is small relative to that of the cathode. A number of measures may be taken to significantly reduce the effects of galvanic corrosion. These include the following:

1. If coupling of dissimilar metals is necessary, choose two that are close together in the galvanic series.
2. Avoid an unfavorable anode-to-cathode surface area ratio; use an anode area as large as possible.
3. Electrically insulate dissimilar metals from each other.
4. Electrically connect a third, anodic metal to the other two; this is a form of cathodic protection.

3-Crevise Corrosion
Electrochemical corrosion may also occur as a consequence of concentration differences of ions or dissolved gases in the electrolyte solution, and between two regions of the same metal piece. For such a concentration cell, corrosion occurs in the locale that has the lower concentration. A good example of this type of corrosion occurs in crevices and recesses or under deposits of dirt or corrosion products where the solution becomes stagnant and there is localized depletion of dissolved oxygen. Corrosion preferentially occurring at these positions is called crevice corrosion (Figure
The crevice must be wide enough for the solution to penetrate, yet narrow enough for stagnancy; usually the width is several thousandths of an inch. The proposed mechanism for crevice corrosion is illustrated. After oxygen has been depleted within the crevice, oxidation of the metal occurs at this position. Electrons from this electrochemical reaction are conducted through the metal to adjacent external regions, where they are consumed by reduction—most probably reaction. In many aqueous environments, the solution within the crevice has been found to develop high concentrations of H\(^-\) and Cl\(^-\) ions, which are especially corrosive. Many alloys that passivate are susceptible to crevice corrosion because protective films are often destroyed by the H\(^-\) and Cl\(^-\) ions. Crevice corrosion may be prevented by using welded instead of riveted or bolted joints, using nonabsorbing gaskets when possible, removing accumulated deposits frequently, and designing containment vessels to avoid stagnant areas and ensure complete drainage.

4-Pitting

Pitting is another form of much localized corrosion attack in which small pits or holes form. They ordinarily penetrate from the top of a horizontal surface downward in a nearly vertical direction. It is an extremely insidious type of corrosion, often going undetected and with very little material loss until failure occurs. The mechanism for pitting is probably the same as for crevice corrosion in that oxidation occurs within the pit itself, with complementary reduction at the surface. It is supposed that gravity causes the pits to grow downward, the solution at the pit tip becoming more concentrated and dense as pit growth progresses. A pit may be initiated by a localized surface defect such as a scratch or a slight variation in composition. In fact, it has been observed that specimens having polished surfaces display a greater resistance to pitting corrosion. Stainless steels are somewhat susceptible to this form of corrosion; however, alloying with about 2% molybdenum enhances their resistance significantly.

5-Intergranular Corrosion

As the name suggests, intergranular corrosion occurs preferentially along grain boundaries for some alloys and in specific environments. The net result is that a macroscopic specimen disintegrates along its grain boundaries. This type of corrosion is especially prevalent in some stainless steels. When heated to temperatures between 500 and 800\(^\circ\)C (950 and 1450\(^\circ\)F) for sufficiently long time periods, these alloys become sensitized to intergranular attack. It is believed that this heat treatment permits the formation of small precipitate particles of chromium carbide by reaction between the chromium and carbon in the stainless steel. These particles form along the grain boundaries, as illustrated in Figure below. Both the chromium and the carbon must diffuse to the grain boundaries to form the precipitates, which leaves a chromium-depleted zone adjacent to the grain boundary. Consequently,
this grain boundary region is now highly susceptible to corrosion. Intergranular corrosion is an especially severe problem in the welding of stainless steels, when it is often termed weld decay. Stainless steels may be protected from intergranular corrosion by the following measures: (1) subjecting the sensitized material to a high-temperature heat treatment in which all the chromium carbide particles are redissolved, (2) lowering the carbon content below 0.03 wt% C so that carbide formation is minimal, and (3) alloying the stainless steel with another metal such as niobium or titanium, which has a greater tendency to form carbides than does chromium so that the Cr remains in solid solution.

6-Selective Leaching
Selective leaching is found in solid solution alloys and occurs when one element or constituent is preferentially removed as a consequence of corrosion processes. The most common example is the dezincification of brass, in which zinc is selectively leached from a copper–zinc brass alloy. The mechanical properties of the alloy are significantly impaired, since only a porous mass of copper remains in the region that has been dezincified. In addition, the material changes from yellow to a red or copper color. Selective leaching may also occur with other alloy systems in which aluminum, iron, cobalt, chromium, and other elements are vulnerable to preferential removal.

7-Erosion–Corrosion
Erosion–corrosion arises from the combined action of chemical attack and mechanical abrasion or wear as a consequence of fluid motion. Virtually all metal alloys, to one degree or another, are susceptible to erosion–corrosion. It is especially harmful to alloys that passivate by forming a protective surface film; the abrasive action may erode away the film, leaving exposed a bare metal surface. If the coating is not capable of continuously and rapidly reforming as a protective barrier, corrosion may be severe. Relatively soft metals such as copper and lead are also sensitive to this form of attack. Usually it can be identified by surface grooves and waves having contours that are characteristic of the flow of the fluid. The nature of the fluid can have a dramatic influence on the corrosion behavior. Increasing fluid velocity normally enhances the rate of corrosion. Also, a solution is more erosive when bubbles and suspended particulate solids are present. Erosion–corrosion is commonly found in piping, especially at bends, elbows, and abrupt changes in pipe diameter—positions where the fluid changes direction or flow suddenly becomes turbulent. Propellers, turbine blades, valves, and pumps are also susceptible to this form of corrosion. One of the best ways to reduce erosion–corrosion is to change the design to eliminate fluid turbulence and impingement effects. Other materials may also be utilized that inherently resist erosion. Furthermore, removal of particulates and bubbles from the solution will lessen its ability to erode.
8-Stress Corrosion

Stress corrosion, sometimes termed stress corrosion cracking, results from the combined action of an applied tensile stress and a corrosive environment; both influences are necessary. In fact, some materials that are virtually inert in a particular corrosive medium become susceptible to this form of corrosion when a stress is applied. Small cracks form and then propagate in a direction perpendicular to the stress with the result that failure may eventually occur. Failure behavior is characteristic of that for a brittle material, even though the metal alloy is intrinsically ductile. Furthermore, cracks may form at relatively low stress levels, significantly below the tensile strength. Most alloys are susceptible to stress corrosion in specific environments, especially at moderate stress levels. For example, most stainless steels stress corrode in solutions containing chloride ions, whereas brasses are especially vulnerable when exposed to ammonia. The stress that produces stress corrosion cracking need not be externally applied; it may be a residual one that results from rapid temperature changes and uneven contraction, or for two-phase alloys in which each phase has a different coefficient of expansion. Also, gaseous and solid corrosion products that are entrapped internally can give rise to internal stresses. Probably the best measure to take in reducing or totally eliminating stress corrosion is to lower the magnitude of the stress. This may be accomplished by reducing the external load or increasing the cross-sectional area perpendicular to the applied stress. Furthermore, an appropriate heat treatment may be used to anneal out any residual thermal stresses.

CORROSION PREVENTION

Some corrosion prevention methods were treated relative to the eight forms of corrosion; however, only the measures specific to each of the various corrosion types were discussed. Now, some more general techniques are presented; these include material selection, environmental alteration, design, coatings, and cathodic protection. Perhaps the most common and easiest way of preventing corrosion is through the judicious selection of materials once the corrosion environment has been characterized. Standard corrosion references are helpful in this respect. Here, cost may be a significant factor. It is not always economically feasible to employ the material that provides the optimum corrosion resistance; sometimes, either another alloy and/or some other measure must be used. Changing the character of the environment, if possible, may also significantly influence corrosion. Lowering the fluid temperature and/or velocity usually produces a reduction in the rate at which corrosion occurs. Many times increasing or decreasing the concentration of some species in the solution will have a positive effect; for example, the metal may experience passivation. Inhibitors are substances that, when added in relatively low concentrations to the environment, decrease its corrosiveness. Of course, the specific inhibitor depends both on the alloy and on the corrosive environment. There are several mechanisms that may account for the effectiveness of inhibitors. Some react with and virtually eliminate a chemically active species in the solution (such as dissolved oxygen). Other inhibitor molecules attach themselves to the corroding surface and interfere with either the oxidation or the reduction reaction, or form a very thin protective coating. Inhibitors are normally used in closed systems such as automobile radiators and steam boilers. Several aspects of design consideration have already been discussed, especially with regard to galvanic and crevice corrosion and erosion–corrosion. In addition, the design should allow for complete drainage in the case of a shutdown, and easy washing. Since dissolved oxygen may enhance the corrosivity of many solutions, the design should, if possible, include provision for the exclusion of air. Physical barriers to corrosion are applied on surfaces in the form of films and coatings. A large diversity of metallic and nonmetallic coating materials are available. It is essential that the coating maintain a high degree of surface adhesion, which undoubtedly requires some preapplication surface treatment. In most cases, the coating must be virtually nonreactive in the corrosive environment and resistant to mechanical damage that exposes the bare metal to the corrosive environment. All three material types—metals, ceramics, and polymers—are used as coatings for metals.
Learning Objectives
After careful study of this lecture you should be able to do the following:
1. Define engineering stress and engineering strain.
2. State Hooke’s law, and note the conditions under which it is valid.
3. Define Poisson’s ratio.
4. Given an engineering stress–strain diagram, determine (a) the modulus of elasticity, (b) the yield strength (0.002 strain offset), and (c) the tensile strength, and (d) estimate the percent elongation.
5. For the tensile deformation of a ductile cylindrical specimen, describe changes in specimen profile to the point of fracture.
6. Compute ductility in terms of both percent elongation and percent reduction of area for a material that is loaded in tension to fracture.
7. Be able to compute true stress and true strain values.

Stress, Strain and Elastic Moduli
Consider a solid rod with length $L_0$, diameter $D_0$ and cross sectional area $A_0$ as depicted in Figure below (a). When tensile forces, those stretching the solid, are applied to this specimen, the rod becomes extended, or stretched as shown in Figure below (b). At the same time it becomes narrower. Its new length, diameter and cross sectional area are $L$, $D$ and $A$. The effects of the applied load can only be meaningfully compared between samples if we compare the force experienced per unit cross sectional area of the sample. In other words, we need to compare forces experienced by the same area on different samples. We define engineering tensile stress as the applied force per unit original cross sectional area of the sample where the force is perpendicular to the area,

$$\sigma = \frac{F}{A_0}$$

Tensile stresses pull away the faces of application. The forces are at right angles to the faces. If the applied forces are pushing the faces inwards, compressing the sample, then the stresses are said to be compressive. The actual stress, or instantaneous stress, is the force per instantaneous unit area. If during the application of the force $F$ the cross sectional area is $A$, then true stress $\rho$ is defined as,

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

When comparing the amounts of extension different samples exhibit under a given tensile stress, it is useful to compare the extensions per unit original length. If $\Delta L = L - L_0$ is the extension under an applied tensile load then engineering strain is defined as,

$$\varepsilon = \frac{\Delta L}{L_0} \cdot \frac{L - L_0}{L_0}$$

Engineering strain is based on extensions with respect to the original length $L_0$. Consider what happens when we increase the tensile load $F$ by a small amount $\delta F$ as shown in Figure below (c). The length $L$ changes by $\delta L$. The instantaneous incremental increase in length per unit length $\delta L/L$ is defined as incremental true strain, True strain is then the summation of all incremental true strains from original length, $L_0$, to final length $L$,

$$\varepsilon_i = \ln(1 + \varepsilon)$$

An elastic behavior in which the strain exhibited by a body increases with the applied stress. If the stress is removed, the strain returns to zero. This type of elastic behavior is called Hooke's law. The proportionality is usually written as

$$\sigma = E\varepsilon$$
where $E$ is an elastic modulus that is called Young's modulus. Young's modulus depends on the nature of bonding between the atoms or molecules in the solid inasmuch as the strain is a result of the stretching of bonds between atoms. A solid that has been strained by an applied stress is said to be *elastically deformed* if it returns to its original shape after the removal of the stress. By convention, a positive stress pulls, or creates tension, and a negative stress pushes, or creates compression. Thus, compression results in a negative strain or compression as we expect. When a solid is extended along the direction of an applied tensile stress, along $z$, it becomes narrower in the perpendicular directions. The longitudinal stress, $\sigma_z$, therefore induces not only a longitudinal strain, $\varepsilon_z$, but also a lateral strain, $\varepsilon_x$, along $x$ in a direction perpendicular to $z$. The lateral strain along $x$, along any direction perpendicular to $z$, is defined as $\varepsilon_x = \Delta D / D_o = (D - D_o) / D_o$ which is a negative quantity for tensile stresses and positive for compressional stresses. The ratio of lateral to longitudinal strain is defined as the *Poisson's ratio*, $\nu$, for that material,

$$\nu = -\frac{\varepsilon_x}{\varepsilon_z} = -\frac{\text{Lateral strain}}{\text{Longitudinal strength}}$$

There are other forms of stresses and strains in engineering besides those tensile stresses. When opposite forces act tangentially on opposite faces of a brick-like solid as in Figure below, the resulting stresses are called shear stresses as they tend to shear the solid. *Shear stress*, denoted by $\tau$, is defined as the tangential force per unit surface area on which it acts,

$$\tau = \frac{F_t}{A}$$

where $F_t$ is the tangential force acting on surface area $A$. Shear stresses cause a solid to become sheared or skewed. In Figure below, shear stresses acting on a brick-type solid induce the opposite surfaces, separated by distance $L$, to be relatively displaced by $\Delta x$ along the forces (along $x$). Greater is $\Delta x$ with respect to $L$, bigger is the extent of skewing. Shear strain, $\gamma$, is defined by
\[ \gamma = \frac{\Delta x}{L} = \tan(\theta) \approx \theta \]

where \( \theta \) is the shearing angle. As typically \( \theta \) is small, \( \tan \theta \approx \theta \) and shear strain is simply this angle of shear. All experiments indicate that for a given material and for small shear strains, the shear deformation is elastic and obeys Hooke's law. \( G \) called the shear modulus (also known as modulus of rigidity) relates the shear stress and strain in elastic shear deformation via

\[ \tau = G\gamma \]

Another form of stress arises by the uniform application of forces over the whole surface of the body as, for example, when the body is immersed deep into ocean. Force per unit area in this case is called pressure, \( P \). Even though pressure is always compressing a body, it is nonetheless, denoted as a positive quantity. Suppose that an applied pressure changes the volume by an amount \( \Delta V \) as shown in Figure below. Then volume strain, \( \Delta \), is defined by

\[ \Delta = \frac{\Delta V}{V} \]

For small volume strains, the volume deformation is elastic and obeys Hooke's law with an elastic modulus that is called the bulk modulus, \( K \):

\[ P = -K\Delta \]

The negative sign in Equation therefore ensures that the bulk modulus, which is a material property, is tabulated as a positive quantity. Elastic deformation is said to occur instantaneously. When a stress is applied to a body, the bonds between atoms become stretched or compressed almost immediately. Furthermore, when the stress is removed, the solid instantaneously and the induced strain occurs returns to its original shape and size.
Stress-Strain Characteristics, Brittle and Ductile Materials

Tensile stress–strain characteristics of a material are typically examined using a tensile test machine. The specimen, in the shape of a rod as depicted in Figure below, is pulled by a progressively increasing tensile load until it is fractured.

A typical stress vs. strain result for a ductile polycrystalline metal (such as an aluminum alloy) from a tensile test is shown in Figure below. Initially the stress–strain characteristic is linear and represents the elastic deformation of the specimen. The extension of the sample is due to the stretching of the interatomic bonds. As the load is increased, the bonds are stretched even more to balance the applied tensile forces. The slope of this linear region from the origin represents Young's modulus or elastic modulus, \( E \). If we unload the sample anywhere in the linear region, say at point \( A \), the sample size and shape return to their original values along \( OA \).

Thus, after unloading, strain always returns to \( O \) in this linear or elastic regime. As the load is increased further eventually a point \( Y \) is reached where the deformation is no longer elastic. Beyond point \( Y \) the sample yields permanently by deforming plastically under load. Point \( Y \) marks the end of the elastic regime and the onset of plastic deformation in which deformation involves material flow and is therefore permanent. If we were to unload the specimen in this plastic region, say at point \( P \), then the sample would recover along a line parallel to the elastic region. This elastic recovery, along \( PO' \), represents the stretched bonds recoiling back (springing back) to their equilibrium lengths. But the specimen is left with a permanent plastic deformation \( \varepsilon_{pl} \), corresponding to \( OO' \). Reapplying the load would elongate the specimen along the line \( O'P \) just as if it had started at \( O' \). In the plastic deformation region, from point \( Y \) onwards, the deformation of the specimen is permanent and is due to material flow, or relative motion of atomic layers, aided by motions of dislocations in
the crystal grains as described later. Permanent deformation is due to the breaking of bonds and relative motions of atomic layers in contrast to bond stretching in the elastic region. We should note that in the plastic region the bonds are still stretched and when the specimen is unloaded (say at point $P$) these stretched bonds uncoil. It is this uncoiling of stretched bonds that returns the stress–strain behavior along a line $(PO)$ almost parallel to the elastic deformation line. As the load is increased, the specimen is plastically deformed further and eventually a point $M$ is reached when there appears to be less stress needed for further plastic deformation. At this point $M$, the specimen develops a small constriction or a neck at one location. The actual stress in the neck region is greater than anywhere else and all further deformation henceforth takes place in the neck which gets thinner and longer. Eventually the neck fractures at point $F$. From $Y$ to $M$ the plastic deformation is uniform throughout the central region of the specimen rod but from $M$ to $F$ it occurs locally in the neck region where the stresses are greatest due to the small cross sectional area of the neck. As soon as the specimen fractures at $F$ the stretched bonds recoil which corresponds to the recovery of elastic deformation along the line $FB$. If we were to bring and fit together the two fracture pieces we would find that the gauge marks have moved apart from their original separation $L_o$ to $L_f$ which is the separation at fracture. The plastic fracture strain, $\varepsilon_f$, or the strain after fracture and after elastic recovery, is the permanent deformation $OB$. The stress at fracture, $\sigma_f$, is that stress at the instant of fracture $F$. If the load at instant of fracture is $F_f$ then $\sigma_f = F_f / A_o$. After fracture, the two broken pieces can be refitted together to obtain the cross sectional area at fracture $A_f$ (at the neck) by measuring the diameter $D_f$ at the neck. The true stress $\sigma_f$ at fracture is then $F_f / A_f$ and is more than $\sigma_f$, the apparent engineering stress. It seems unusual that there appears to be less stress needed to deform the material from point $M$ to $F$. This is only a reflection of the less load needed from $M$ to $F$ not the actual stress in the specimen. The true stress, force per unit instantaneous area, actually increases because the cross sectional area of the neck decreases with deformation. If we were to plot the true stress vs. true strain behavior we would find the dashed curve shown in Figure below. The true stress–strain behavior increases monotonically and does not exhibit a maximum point $M$. We always need to increase the true stress to further deform the sample. The points $M'$ and $F$ on the true stress–strain curve correspond to $M$ and $F$. We should note that the relationship $\varepsilon = \ln(1+\varepsilon)$ assumes uniform straining of the specimen and therefore does not apply beyond point $M$ because from $M$ to $F$ extension is localized to the neck region. The true strain from $M$ to $F$ is larger than the engineering strain as sketched in Figure below. It is apparent from Figure that once point $M$ is reached necking occurs and the specimen proceeds to fracture in this neck region with less and less load. Point $M$ therefore represents the maximum tensile stress that can be applied to the specimen without fracturing it. It is called the ultimate tensile strength or just tensile strength ($\sigma_{TS}$ or $TS$) of the specimen and it is an important engineering design quantity. The yield point $Y$ that defines the elastic limit or the onset of yielding and plastic deformation defines the yield strength of the material, $\sigma_y$ or $YS$, which is another important design parameter. Yield strength represents the resistance of the material against plastic deformation. Higher is the yield strength, greater is the stress that can be applied without permanently deforming the specimen. Although in some materials the yield point $Y$ is reasonably well defined in some other materials the transition from elastic to plastic deformation does not occur at such a well defined point. Moreover, there is some uncertainty involved in locating the true yield point $Y$ from elastic to plastic deformation regimes. It is therefore convenient to adopt a standard definition for yield strength that can be applied across the board for many materials. This is the 0.2% offset (or 0.2% proof) yield strength, $\sigma_{y(0.2\%)}$. We take a strain 0.002 (0.2%) along the strain axis and draw a line parallel to the elastic deformation line. This line cuts the stress–strain curve at point $Y'$ which defines the offset yield strength. According to Figure below, the plastic strain at fracture, $\varepsilon_f$, represents the extent of plastic deformation a material exhibits before fracturing. The greater is $\varepsilon_f$, the more ductile is the material. The ductility is defined as the amount of plastic deformation that is exhibited by a material at fracture. There are two quantitative definitions based on percent specimen elongation and percent cross sectional area reduction. Percent elongation is plastic strain at fracture, $\varepsilon_f$, expressed as percentage,

$$EL\% = \frac{L_f - L_o}{L_o} \times 100\% = \varepsilon_f \times 100\%$$
If $A_f$ is the cross sectional area at fracture then $A_o - A_f$ is the reduction in the cross sectional area. Percent area reduction $AR\%$, is defined by

$$AR\% = \frac{A_o - A_f}{A_o} \times 100\%$$

For a given material $EL\%$ and $AR\%$ values are different. Further, since most of the plastic elongation occurs in the small necked region rather than throughout the whole gauge length, $L_o$, $EL\%$ is higher for shorter $L_o$ which must therefore be stated in comparing $EL\%$ values between materials (for standard specimens, $L_o = 2$ in.). $AR\%$, on the other hand, does no depend on $L_o$ or $A_o$ and is therefore more frequently used to express ductility. Ductility is obviously an important design parameter because we have to know the maximum amount of deformation a component can experience before fracture. In shaping metals, as in forging, ductility defines the limit of maximum allowable plastic deformation. We can differentiate between ductile and brittle materials by comparing their stress–strain characteristics and their percent elongations or plastic strains at fracture ($\varepsilon_f$) as shown in Figure below. A ductile material exhibits an ultimate tensile strength point $M$ and considerable plastic deformation before fracture at point $F$. It has typical values of $EL\%$ in excess of 10%. Some materials, for example some cast irons and some metal alloys, exhibit only moderate ductility. Although it is possible to define yield strength, the fracture occurs either before or around the ultimate tensile strength point $M$. Brittle materials do not exhibit any marked plastic deformation and their fracture strains are less than a few percent. Many ceramics, glasses and some metals, such as gray cast irons, are brittle materials. The fracture occurs suddenly which may be on the elastic region. Brittle fracture invariably involves induced crack propagation across the sample at sufficiently large stresses and the position of the fracture point, $F_{brittle}$, usually depends on the surface conditions of the specimen e.g. the presence of surface flaws, surface cracks etc. For example, in a tensile test on a brittle material, it would not be unusual to have fracture by induced crack propagation across the specimen originating from one of the indentations of the extensometer grip pins. For brittle materials the tensile strength is usually taken as the fracture stress $\sigma_f$ at $F_{brittle}$ but its value is highly variable from sample to sample depending on various imperfections and flaws not only on the surface but also in the bulk of the brittle material. We will see later that failure in brittle materials invariably involves crack propagation across the specimen. The fracture starts from a tiny, invisible, crack on the surface.
Dislocation Glide and Crystal Slip

The atomic origin of plastic deformation in metals involves the breaking of bonds and relative motions of layers of atoms. However, the layers of atoms do not all move together because this would require the breaking of all the bonds between this layer and the neighboring layers all at the same time which requires far too much energy. The actual plastic flow is the result of breaking and making bonds one line of atoms at a time through the motion of dislocations. A crystal with an edge dislocation is shown in Figure below (a) where atom A is on the dislocation line running into the paper. The sign $\perp$ represent the edge dislocation and the position of the dislocation line where the vertical line is the extra half plane of atoms in the crystal. The atoms $A$, $B$, $B'$, $C$, $C'$ around the dislocation line are displaced from their normal equilibrium positions which means that bonds between them are already strained. For example, the bond $AB$ is shorter and bond $BC$ is longer than the equilibrium bond length. When a shear stress is applied to the crystal as shown in Figure below (b), the shear strain causes atoms $A$ and $B$ to be slightly displaced towards the right and atom $C$ to be slight displaced towards the left. Atom $A$ therefore gets closer to $C$ whereas atom $B$ moves away from $C$. The bond $BC$ becomes even more stretched. If the shear strain is sufficient to get $A$ close to $C$ and $B$ further away from $C$ then the stretched bond $BC$ breaks and a new bond $AC$ is formed. Consequently the dislocation becomes shifted to atom $B$ by one interatomic distance as depicted in Figure below (c).

If the shear stress is continued to be applied then the dislocation motion repeats itself until this reaches the end of the crystal where it causes a step equal to an interatomic spacing as illustrated in the sequence of events (a) to (e) in Figure below. The crystal is then said to have slipped. The motion of a dislocation under an applied shear stress is termed dislocation glide. An applied shear stress to a crystal, if sufficiently large, can generate and glide a dislocation from one end of the crystal to the other as illustrated in Figure below (a) to (e). When a shear stress is applied to a crystal, as in (a), it can generate a dislocation at $A$ as in (b). As long as the shear stress is applied, this dislocation can glide from $A$ to $B$ and so on, as in (c) and (d). When eventually the dislocation line reaches the end of the crystal, it generates a step equal to the interatomic separation. Plastic
deformation by dislocation gliding is analogous the locomotion of a caterpillar which is depicted in the same series (a) to(e) in Figure below. The caterpillar moves by lifting up one leg at a time. There is yet a further analogy. It is quite difficult to move a carpet on the floor by simply pulling it. It is possible, however, to shift it by first making a ruck at one end and then pushing this ruck to the end of the carpet somewhat similar to the way in which caterpillar's "ruck" (hump) moves forward.

It is apparent that edge dislocations play an important role in plastic deformation in metals. Figure below illustrates macroscopically, without the atomic details, how an applied shear stress to a crystal in (a) generates a dislocation in (b) which glides on a plane called the slip plane and in a direction called the slip direction. In (c) the dislocation reaches the end of the crystal and causes a slip, a permanent displacement. (d) is simply a different perspective of (c). The top and bottom portions of the crystal have been permanently displaced with respect to each other. The crystal has slipped. There is now a permanent shear strain as indicated by the dashed lines in (c). A slip plane is an atomic plane on which a dislocation can glide. The step at the surface of a slipped crystal is said to be along a slip line because the dislocation is a line defect and when it reaches the surface it causes a step along a line as shown in Figure below(d). The magnitude and direction of the slip, due to the motion of one dislocation entirely through the crystal as in Figure below(d), is defined in terms of the displacement of the lower portion of the crystal with respect to the upper portion and is the Burgers vector, \( \mathbf{b} \). It is apparent that for an edge dislocation induced slip \( \mathbf{b} \) is an atomic spacing in the slip plane along the slip direction. When shear forces are applied to the perfect crystal, a dislocation is generated which is then moved from \( A \) to \( B \) and so on all the way to the end of the crystal. The crystal then becomes displaced by one atomic spacing. Work is done by the shear forces in moving the dislocation across the crystal and slipping the crystal. Majority of this work, some 95% of it, generates lattice vibrations or heat while the dislocation is being "pushed" across the crystal. When we plastically deform a metal piece it gets warmer for this reason. When we repeatedly bend a piece of metal such as a metal clothes' hanger back and forward we can feel that the piece gets warmer.
Screw dislocations also contribute to plastic deformation, the only difference is in the direction of dislocation glide and crystal slip as illustrated in Figure below (a) and (b). For a screw dislocation the dislocation line and the Burgers vector, or slip direction, are parallel whereas the two are at right angles for an edge dislocation. This means that the slip plane is well defined for an edge dislocation, but there are a number of choices for the screw dislocation. Any plane that contains the dislocation line and the Burgers vector in Figure below is possible. For example the screw dislocation in Figure below (c) starts gliding on plane $A$ and then changes its glide direction and glides on plane $B$. When it reaches the end of the crystal it causes a step with a slip line that is bent. Planes $A$ and $B$ are both slip planes because the slip direction, Burgers vector $\mathbf{b}$, lies in both planes. A screw dislocation can therefore cross-slip which means that it can change its glide from one plane to another that is parallel to Burgers vector, as in Figure below (c). This is not the case for an edge dislocation. Screw dislocations have more choices for planes on which they can glide than edge dislocations and hence their contribution to plastic flow is more substantial.

Example 1-1: Young's modulus and Poisson's ratio

Consider a 10 meter steel wire of diameter 3 mm that is carrying a tensile load of 1.2 kN (equivalent to a mass of about 120 kg). If the elastic modulus and Poisson's ratio of steel are 210 GPa and 0.27 what is the new length and diameter of the steel wire assuming that the elongation is elastic? Compare the true and engineering strains and also true and engineering stresses.

Solution

The applied tensile stress (engineering stress) along the wire taken as the $z$-direction is

$$\sigma = \frac{F}{A} = \frac{F}{\pi \left( \frac{D}{2} \right)^2} = \frac{120 \times 10^3 \text{N}}{\pi (1.5 \times 10^{-3} \text{m})^2} = 169.77 \text{ MPa}$$

The resulting engineering strain along $z$ can be found from $\sigma = E\varepsilon_z$,

$$\varepsilon_z = \frac{\sigma}{E} = \frac{169.77 \times 10^6 \text{Pa}}{210 \times 10^9 \text{Pa}} = 8.084 \times 10^{-4}$$

If $L$ is new length of the wire, then

$$\varepsilon_z = \frac{L - L_0}{L_0}$$

which gives

$$L = L_0(1 + \varepsilon_z) = (10 \text{ m})(1 + 8.084 \times 10^{-4}) = 10.0081 \text{ m}$$

The extension $L - L_0$ is therefore 8.1 mm nearly 1 cm.
Problems

1. What was the original length of a wire that has been strained by 30% and whose final length is 1 m?

2. The initial diameter of a wire is 2 cm and needs to be reduced to 1 cm. Calculate the amount of cold work (reduction in area in percent) which is necessary.

3. Calculate the initial diameter of a wire that has been longitudinally strained by 30% and whose final diameter is 0.1 cm. Assume no volume change.

4. What force is needed to plastically deform a wire of 2 cm diameter whose yield strength is 40 MPa?

5. Calculate the ductility of a wire (that is, its percent area reduction at fracture during tensile stressing) whose initial diameter was 1 cm and whose diameter at fracture is 0.8 cm.
6. Calculate the true stress at fracture for a metal rod whose engineering fracture strength is 450 MPa and whose
diameter at fracture was reduced by plastic deformation from 1 to 0.8 cm.

7. Compare engineering strain with true strain and engineering stress with true stress for a material whose initial
diameter was 2 cm and whose final diameter at fracture is 1.9 cm. The initial length before plastic deformation
was 10 cm. The applied force was $3 \times 10^4$ N. Assume no volume change during plastic deformation.

8. A metal plate needs to be reduced to a thickness of 4 cm by involving a rolling mill. After rolling, the elastic
properties of the material cause the plate to regain some thickness. Calculate the needed separation between the
two rollers when the yield strength of the material after plastic deformation is 60 MPa and the modulus of
elasticity is 124 GPa.

9. In the following relation  
$$
\varepsilon_t = \int_0^1 \frac{d_l}{l} = \ln \left( \frac{l}{l_0} \right) = \ln \left( \frac{A_o}{A_i} \right).
$$
Show in mathematical terms for what condition
(pertaining to a possible change in volume) this relation is true.

10. Show that the true and engineering stress and strain are related by
$$
\sigma_t = \sigma (1 + \varepsilon) \\
\varepsilon_t = \ln (1 + \varepsilon)
$$
For the case when no volume changes occurs during deformation, that is, before the onset of necking.
Learning Objectives
After careful study of this lecture you should be able to do the following:
1. Describe the relation between relative permittivity and refractive index.
2. Describe the atoms polarization that takes place due to EM.
3. Drive a dispersion relation.
4. Describe the relation between dispersion and light extinction.
5. Explaining the real and imaginary part of refractive index.

Relative Permittivity and Refractive Index
The refractive index of materials in general depends on the frequency, or the wavelength, of light. This wavelength dependence follows directly from the frequency dependence of the relative permittivity $\varepsilon_r$. Figure (1) shows what happens to an atom in the presence of an electric field $E$ which may be due to a light wave passing through this location, or simply an applied field. We consider the induced electric dipole moment at this atom. In the absence of an electric field and in equilibrium, the center of mass $C$ of the orbital motions of the electrons coincides with the positively charged nucleus at $O$, and the net electric dipole moment is zero as indicated in Figure (1-a). Suppose that the atom has $Z$ number of electrons orbiting the nucleus and all the electrons are contained within a given shell. In the presence of the electric field $E$, however, the light electrons become displaced in the opposite direction to the field so that their center of mass $C$ is shifted by some distance $x$ with respect to the nucleus $O$ which we take to be the origin as shown in Figure (1-b). As the electrons are "pushed" away by the applied field, the Coulombic attraction between the electrons and nuclear charge "pulls in" the electrons. The force on the electrons, due to $E$, trying to separate them away from the nuclear charge is $ZeE$. The restoring force $F_r$ which is the Coulombic attractive force between the electrons and the nucleus, can be taken to be proportional to the displacement $x$ provided that the latter is small. The reason is that $F_r = F_r(x)$ can be expanded in powers of $x$ and for small $x$ only the linear term matters. The restoring force $F_r$ is obviously zero when $C$ coincides with $O$ ($x = 0$). We can write $F_r = -\beta x$ where $\beta$ is a constant and the negative sign indicates that $F_r$ is always directed towards the nucleus $O$. First consider applying a dc field. In equilibrium, the net force on the negative charge is zero or $ZeE = \beta x$ from which $x$ is known. Therefore the magnitude of the induced electronic dipole moment is given by

$$p_{\text{induced}} = (Ze)x = (Ze^2/\beta)E$$  \hspace{1cm} (1)

As expected, $p_{\text{induced}}$ is proportional to the applied field. The electronic dipole moment in equation (1) valid under static conditions, i.e. when the electric field is a dc field. Suppose that we suddenly remove the applied electric field polarizing the atom. There is then only the restoring force, $-\beta x$, which always acts to pull the electrons towards the nucleus, $O$. The equation of motion of the negative charge center is then (force = mass $\cdot$ acceleration)

$$-\beta x = Zm_eE x/dt^2$$

Thus, by solving this differential equation, the displacement at any time is a simple harmonic motion, that is

$$x(t) = x_0 \cos(\omega_t)$$

where the angular frequency of oscillation $\omega_t$ is

$$\omega_t = \left[\frac{\beta}{Zm_e}\right]^{1/2}$$  \hspace{1cm} (2)
In essence, this is the oscillation frequency of the center of mass of the electron cloud about the nucleus and $x_0$ is the displacement before the removal of the field. After the removal of the field, the electronic charge cloud executes simple harmonic motion about the nucleus with a natural frequency $\omega_o$ determined by equation (2); $\omega_o$ is also called the resonance frequency. Of course, the oscillations die out with time because there is an inevitable loss of energy from an oscillating charge cloud. An oscillating electron is like an oscillating current and loses energy by radiating electromagnetic waves; all accelerating charges emit radiation. Consider now the presence of an oscillating electric field due to an electromagnetic wave passing through the location of this atom as in figure (1-b). The applied field oscillates harmonically in the $+x$ and $-x$ directions, that is $E = E_o \exp (j\omega t)$. This field will drive and oscillate the electrons about the nucleus. There is again a restoring force $F_r$ acting on the displaced electrons trying to bring back the electron shell to its equilibrium placement around the nucleus. For simplicity we will again neglect energy losses. Newton’s second law for $Ze$ electrons with mass $Zm_e$ driven by $E$ is given by,

$$Zm_e \frac{d^2x}{dt^2} = -ZeE_o \exp(j\omega t) - \beta x$$

The solution of this equation gives the instantaneous displacement $x(t)$ of the center of mass of electrons from the nucleus $(C$ from $O)$,

$$x = x(t) = \frac{eE_o \exp(j\omega t)}{m_e(\omega_o^2 - \omega^2)}$$

The induced electronic dipole moment is then simply given by $p_{\text{induced}} = -(Ze)x$. The negative sign is needed because normally $x$ is measured from negative to positive charge whereas in figure (1) it is measured from the central nucleus. By definition, the electronic polarizability $\alpha_e$ is the induced dipole moment per unit electric field,

$$\alpha_e = \frac{p_{\text{induced}}}{E} = \frac{Ze^2}{m_e(\omega_o^2 - \omega^2)}$$

Thus, the displacement $x$ and hence electronic polarizability $\alpha_e$ increase as $\omega$ increases. Both become very large when $\omega$ approaches the natural frequency $\omega_o$. In practice, charge separation $x$ and hence polarizability $\alpha_e$ do not become infinite at $\omega = \omega_o$ because two factors impose a limit. First is that at large $x$, the system is no longer linear and the above analysis is not valid. Secondly, there is always some energy loss. Given that the
polarizability is frequency dependent as in Equation (3), the effect on the refractive index $n$ is easy to predict. The simplest (and a very “rough”) relationship between the relative permittivity $\varepsilon_r$ and polarizability $\alpha_e$ is

$$\varepsilon_r = 1 + \frac{N}{\varepsilon_o} \alpha_e$$

Where $N$ is the number of atoms per unit volume. Given that the refractive index $n$ is related to $\varepsilon_r$ by $n^2 = \varepsilon_r$, it is clear that $n$ must be frequency dependent, i.e.

$$n^2 = 1 + \left( \frac{NZ\varepsilon^2}{\varepsilon_o m_e} \right) \frac{1}{\omega_o^2 - \omega^2} \tag{4}$$

We can also express this in terms of the wavelength $\lambda$. If $\lambda_o = \frac{2\pi c}{\omega_o}$ is the resonance wavelength, then Equation (4) is equivalent to

$$n^2 = 1 + \left( \frac{NZ\varepsilon^2}{\varepsilon_o m_e} \right) \left( \frac{\lambda_o}{2\pi c} \right)^2 \frac{\lambda^2}{\lambda^2 - \lambda_o^2} \tag{5}$$

This type of relationship between $n$ and the frequency $\omega$, or wavelength $\lambda$, is called a dispersion relation. The refractive index $n$ decreases as the wavelength $\lambda$ increases above and away from the resonance wavelength $\lambda_o$, as illustrated schematically in Figure (2). Although the above treatment is grossly simplified, it does nonetheless emphasize that $n$ will always be wavelength dependent and will exhibit a substantial increase as the frequency increases towards a natural frequency of the polarization mechanism. In the above example, we considered the electronic polarization of an isolated atom with a well-defined natural frequency $\omega_o$. In the crystal, however, the atoms interact, and further, we also have to consider the valence electrons in the bonds. The overall result is that $n$ is a complicated function of the frequency or the wavelength. In many applications, it is useful to describe the dispersion relationship by an empirical formula. Over a limited wavelength range, the empirical equation may be relatively simple.

![Figure 2](image)

**Dispersion and Light Extinction**

Consider an atom in a material as in Figure (1) that is experiencing an alternating field $E$ that oscillates harmonically in the $+x$ and $-x$ directions, that is $E = E_o \exp(i\omega t)$. The electrons are driven by this field. There is a restoring force $F_r$ acting on the displaced electrons trying to bring back the electron shell to its equilibrium placement around the nucleus. This force $F_r$ is proportional to the displacement $x$ and is always directed towards the center $O$; it can be written as $-\beta x$. Oscillating electrons are equivalent to an oscillating current which radiates energy like an antenna. This is an effective loss of energy, just like a frictional force. Further, some of the electron oscillations can be coupled to crystal vibrations and increase their energy, that is, energy will be transferred from electron oscillations to heat. All energy losses are proportional to the velocity $dx/dt$ and the equivalent frictional force per electron and per unit electron mass is $\gamma dx/dt$. Thus, Newton’s second law for Ze electrons with mass $Zm_e$ is given by,
where $Zm_o \omega_o^2$ is the force constant $\beta$. Solving this equation we obtain the instantaneous displacement $x = x(t)$ of the center of mass $C$ of the electron shell from the nucleus $O$ in Figure (1). Once we know $x(t)$ we can easily find the electronic polarizability $\alpha_e$,

$$\alpha_e = \frac{Ze^2}{m_e[\omega_o^2 - \omega^2 + j\gamma\omega]}$$

Where $\omega_o = (\beta/Zm_o)^{1/2}$. It is a resonance frequency where $\alpha_e$ peaks. We can easily separate this complex $\alpha_e$ into real and imaginary parts as

$$\alpha_e = \alpha_e' - j\alpha_e''$$

where

$$\alpha_e' = \frac{Ze^2 (\omega_o^2 - \omega^2)}{m_e[(\omega_o^2 - \omega^2)^2 - (\gamma\omega)^2]}$$

and

$$\alpha_e'' = \frac{Ze^2 (\gamma\omega)}{m_e[(\omega_o^2 - \omega^2)^2 - (\gamma\omega)^2]}$$

The frequency dependencies of the real and imaginary parts of $\alpha_e$ are shown in Figure (3). It is important to recognize that $\alpha_e''$ is directly proportional to the loss coefficient $\gamma$ which means that the imaginary part $\alpha_e''$ represents the loss in the medium. The real part is not significantly affected by $\gamma$ if $\omega$ is sufficiently smaller than $\omega_o$. At resonance (at $\omega = \omega_o$) however, $\alpha_e'$ does not peak to infinity; its maximum is controlled by the loss mechanism. Since $\alpha_e$ is a complex quantity, so is $\varepsilon$ and hence the refractive index. Consider the simplest (and very “rough”) relationship between the relative permittivity $\varepsilon$ and polarizability $\alpha_e$,

$$\varepsilon = 1 + \frac{N}{\varepsilon_o} \alpha_e = 1 + \frac{N}{\varepsilon_o} (\alpha_e' - j\alpha_e'')$$
where \( N \) is the number of atoms per unit volume. Thus, the relative permittivity is a complex quantity, that is
\[
\varepsilon_r = \varepsilon_r' - j \varepsilon_r'' \tag{12}
\]
\[
\varepsilon_r' = 1 + \frac{N}{\varepsilon_0} \alpha_e' \quad \text{and} \quad \varepsilon_r'' = \frac{N}{\varepsilon_0} \alpha_e'' \tag{13}
\]

The real part is the usual relationship between the relative permittivity and polarizability when loss is neglected. Clearly the imaginary part represents the loss, the extinction of light as it passes through the material. There are always some losses in all polarization processes. For example, when the ions of an ionic crystal are displaced from their equilibrium positions by an alternating electric field and made to oscillate, some of the energy from the electric field is coupled and converted to lattice vibrations (intuitively, “sound” and heat). These losses are generally accounted for by describing the whole medium in terms of a complex relative permittivity (or dielectric constant) \( \varepsilon_r \) as in Equation (12) where the real part \( \varepsilon_r' \) determines the polarization of the medium with losses ignored and the imaginary part \( \varepsilon_r'' \) describes the losses in the medium. For a lossless medium, obviously \( \varepsilon_r = \varepsilon_r' \). The loss \( \varepsilon_r'' \) depends on the frequency of the wave and usually peaks at certain natural (resonant) frequencies. If the medium has a finite conductivity (e.g., due to a small number of conduction electrons), then there will be a Joule loss due to the electric field in the wave driving these conduction electrons. This type of light attenuation is called free carrier absorption. In such cases, \( \varepsilon_r'' \) and \( \sigma \) are related by
\[
\varepsilon_r'' = \sigma/(\varepsilon_0 \omega) \tag{14}
\]
where \( \varepsilon_0 \) is the absolute permittivity and \( \sigma \) is the conductivity at the frequency of the EM wave. Since \( \varepsilon_r \) is a complex quantity, we should also expect to have a complex refractive index. An EM wave that is traveling in a medium and experiencing attenuation due to absorption can be generally described by a complex propagation constant \( k \), that is
\[
k = k' - j k'' \tag{15}
\]
where \( k' \) and \( k'' \) are the real and imaginary parts. If we put Equation (15) into the expression for an ideal traveling wave
\[
E = E_0 \exp(j(\omega t - kz))
\]
we will find the following
\[
E = E_0 \exp(-k''z) \exp(j(\omega t - k'z)) \tag{16}
\]
The amplitude decays exponentially while the wave propagates along \( z \) as in Figure (4).
The real $k'$ part of the complex propagation constant (wavevector) describes the propagation characteristics, e.g. phase velocity $v = \omega / k'$. The imaginary $k''$ part describes the rate of attenuation along $z$. The intensity $I$ at any point along $z$ is

$$I \propto |\mathbf{E}|^2 \propto \exp(-2k''z)$$

so that the rate of change in the intensity is

$$dI/dz = -2k'' I$$

Where the negative sign represents attenuation. Suppose that $k_o$ is the propagation constant in vacuum. This is a real quantity as a plane wave suffers no loss in free space. The complex refractive index $N$ with real part $n$ and imaginary part $K$ is defined as the ratio of the complex propagation constant in a medium to propagation constant in free space,

$$\frac{k}{k_o} = N = n - jK = k/k_o = (1/k_o)[k' - jk'']$$

$$n = k'/k_o \quad \text{and} \quad K = k''/k_o \quad \text{-------------------------- (17)}$$

The real part $n$ is simply and generally called the refractive index and $K$ is called the extinction coefficient. In the absence of attenuation,

$$k'' = 0, \quad k = k' \quad \text{and} \quad N = n = k/k_o = k'/k_o$$

We know that in the absence of loss, the relationship between the refractive index $n$ and the relative permittivity $\varepsilon$ is $n = \sqrt{\varepsilon}$. This relationship is also valid in the presence of loss except that we must use complex refractive index and complex relative permittivity, that is,

$$N = n - jK = \sqrt{\varepsilon} = \sqrt{(\varepsilon' - je'')} \quad \text{-------------------------- (18)}$$

By squaring both sides we can relate $n$ and $K$ directly to $\varepsilon'$ and $\varepsilon''$. The final result is

$$n^2 + K^2 = \varepsilon' \quad \text{and} \quad 2nK = \varepsilon'' \quad \text{-------------------------- (19)}$$

Optical properties of materials are typically reported either by showing the frequency dependencies of $n$ and $K$ or $\varepsilon'$ and $\varepsilon''$. Clearly we can use Equation (19) to obtain one set of properties from the other. Figure (5) shows the complex relative permittivity and the complex refractive index of crystalline silicon in terms of the photon energy $\hbar \omega$ (or frequency). For photon energies below the bandgap energy, both $\varepsilon''$ and $K$ are negligible and $n$ is close to about 3.7. Both $\varepsilon'$ and $K$ increase and change strongly as the photon energy becomes greater than 3 eV; far beyond the bandgap energy (1.1 eV). Notice that both $\varepsilon'$ and $n$ peak around $\hbar \omega \approx 3.5$ eV. If we know the frequency dependence of the real part $\varepsilon'$ of the relative permittivity of a material, we can also determine the frequency dependence of the imaginary part $\varepsilon''$; and vice versa. This may seem remarkable but it is true provided that we know the frequency dependence of either the real or imaginary part over as wide a range of frequencies as possible (ideally from dc to infinity) and the material is linear, i.e. it has a relative permittivity that is independent of the applied field; the polarization response must be linearly proportional to the applied field. The relationships that relate the real and imaginary parts of the relative permittivity are called Kramers-Kronig relations. If $\omega'(\omega)$ and $\varepsilon''(\omega)$ represent the frequency dependencies of the real and imaginary parts, then one can be determined from the other as depicted schematically in Figure (6). The optical properties $n$ and $K$ can be determined by measuring the reflectance from the surface of a material as a function of polarization and the angle of incidence (based on Fresnel’s equations). It is instructive to mention that the reflection and
transmission coefficients that are normally given by Fresnel’s equations are based in using a real refractive index, that is neglecting losses. We can still use the Fresnel reflection and transmission coefficients if we simply use the complex refractive index N instead of n. For example, consider a light wave traveling in free space incident on a material at normal incidence $\theta_i = 90^\circ$). The reflection coefficient is now,

\[
R = \frac{N-1}{N+1} = \frac{n - jK - 1}{n - jK + 1}
\]

The reflectance is then

\[
R = \frac{(n - jK - 1)^2}{n - jK + 1} = \frac{(n - 1)^2 + K^2}{(n + 1)^2 + K^2}
\]

which reduce to the usual forms when the extinction coefficient $K = 0$.

Kramers-Kronig relations allow frequency dependencies of the real and imaginary parts of the relative permittivity to be related to each other. The material must be a linear system.
Examples

Ex. 1: Complex refractive index of Si
Spectroscopic ellipsometry measurements on a silicon crystal at a wavelength of 496 nm show that the real and imaginary parts of the complex relative permittivity are 18.661 and 0.630 respectively. Find the complex refractive index. Find the reflectance and the absorption coefficient $\alpha$ at this wavelength. What is the phase velocity?

Solution
We know that $\varepsilon_r' = 18.661$ and that $\varepsilon_r'' = 0.630$. Thus, from Equation (19), we have

$$n^2 + K^2 = 18.661$$

and

$$2nK = 0.630$$

We can take $K$ from the second equation and substitute for it in the first equation,

$$n^2 + \left(\frac{0.630}{2n}\right)^2 = 18.661$$

This is a quadratic equation in $n^2$ that can be easily solved on a calculator to find, $n = 4.319$. Once we know $n$, we can find $K = 1/2n = 0.0729$. If we simply square root the real part of $\varepsilon_r$, we would still find $n = 4.320$, because the extinction coefficient $K$ is small. The reflectance of the Si crystal is

$$R = \frac{(n - 1)^2 + K^2}{(n + 1)^2 + K^2} = \frac{(4.319 - 1)^2 + 0.0729^2}{(4.319 + 1)^2 + 0.0729^2} = 0.3895$$

which is the same as simply using $(n - 1)^2/(n + 1)^2 = 0.3894$, because $K$ is small.

The absorption coefficient $\alpha$ describes the loss in the light intensity $I$ via $I = I_0 \exp(-\alpha z)$ so that

$$\alpha = 2k'' = 2k_sK = 2(2\pi/496 \times 10^{-9})(0.0729) = 1.85 \times 10^6 \text{ m}^{-1}$$

Almost all of this absorption is due to band-to-band absorption (photogeneration of electron-hole pairs).

The phase velocity is given by

$$v = c/n = (3 \times 10^8 \text{ m s}^{-1})/4.32 = 6.94 \times 10^7 \text{ m s}^{-1}$$

EX. 2: Complex refractive index of InP
An InP crystal has a refractive index (real part) $n$ of 3.549 at a wavelength of 620 nm (photon energy of 2 eV). The reflectance of the air-InP crystal surface at this wavelength is 0.317. Calculate the extinction coefficient $K$ and the absorption coefficient $\alpha$ of InP at this wavelength.

Solution
The reflectance $R$ is given by

$$R = \frac{(n - 1)^2 + K^2}{(n + 1)^2 + K^2}$$

which on rearranging gives,

$$K = \left[\frac{(n - 1)^2 - R(n + 1)^2 - 2RN}{R - 1}\right]^{1/2}$$

i.e.

$$K = \left[\frac{(3.549 - 1)^2 - (0.317)(3.549 + 1)^2 - 2(0.317)(3.549)}{0.317 - 1}\right]^{1/2}$$
i.e. \( K = 0.302 \)

The absorption coefficient is

\[
\alpha = 2k_0K = 2(2\pi/620 \times 10^9)(0.302) = 6.1 \times 10^6 \text{ m}^{-1}
\]

**EX.3: Free carrier absorption coefficient and conductivity**

Consider a semiconductor sample with a conductivity \( \sigma \), refractive index \( n \). Show that the absorption coefficient due to free carrier absorption (due to conductivity) is given by

\[
\alpha = \left( \frac{1}{ce_\varepsilon} \right) \frac{\sigma}{n}
\]

**Solution**

The relationship between the conductivity and the absorption coefficient is given by

\[
\varepsilon_r'' = \sigma/(e_\varepsilon \omega)
\]

The relationship between the imaginary part \( \varepsilon_r'' \) of the relative permittivity and the extinction coefficient \( K \) is

\[
2nK = \varepsilon_r''
\]

where \( n \) is the refractive index (the real part of \( N \)). Since the absorption coefficient from the above example is

\[
\alpha = 2k'' = 2k_0K = 2 \left( \frac{2\pi}{\lambda} \right) \left( \frac{\varepsilon_r''}{2n} \right)
\]

thus,

\[
\alpha = \left( \frac{\omega}{c} \right) \frac{\varepsilon_r''}{n}
\]

where \( \omega \) is the angular frequency of the EM radiation, \( \omega = 2\pi c/\lambda \). Substituting for \( \sigma \) in terms of \( \varepsilon_r'' \) gives,

\[
\alpha = \left( \frac{1}{ce_\varepsilon} \right) \frac{\sigma}{n}
\]

**EX.4: Absorption, extinction coefficient and conductivity**

Intrinsic germanium has a resistivity of about 0.5 \( \Omega \) m. Calculate the imaginary part \( \varepsilon_r'' \) of the relative permittivity at a wavelength of 10 \( \mu \)m where the refractive index is 4. Find the attenuation coefficient \( \alpha \) due to free carrier absorption.

**Solution**

The frequency \( \omega \) is

\[
\omega = 2\pi c/\lambda = (2\pi \text{ rad})(3 \times 10^8 \text{ m s}^{-1})/(10 \times 10^{-6} \text{ m}) = 1.88 \times 10^{14} \text{ rad s}^{-1}
\]

The relationship between the conductivity and \( \varepsilon_r'' \) is given by

\[
\varepsilon_r'' = \sigma/(e_\varepsilon \omega) = (0.5 \Omega \text{ m})^{-1}/[(8.85 \times 10^{-12} \text{ F m}^{-1})(1.88 \times 10^{14} \text{ rad s}^{-1})]
\]

i.e.

\[
\varepsilon_r'' = 0.0012
\]

The absorption coefficient \( \alpha \) due to free carriers is given by

\[
\alpha = \left( \frac{1}{ce_\varepsilon} \right) \frac{\sigma}{n} = \left[ \frac{1}{(3 \times 10^8 \text{ m s}^{-1})(8.85 \times 10^{-12} \text{ F m}^{-1})} \right] \left( 0.5 \Omega \text{ m}^{-1} \right)^{-1} = 188 \text{ m}^{-1}
\]
Learning Objectives
After careful study of this lecture you should be able to do the following:
1. Describe the four possible electron band structures for solid materials.
2. Explain the energy band structure in solids.
3. Explain the relation between the electron mobility and conduction.
4. Effects of impurities and temperature on conductivity.

OHM’S LAW
One of the most important electrical characteristics of a solid material is the ease with which it transmits an electric current. Ohm’s law relates the current $I$—or time rate of charge passage—to the applied voltage $V$ as follows:

$$V = IR$$

where $R$ is the resistance of the material through which the current is passing. The units for $V$, $I$, and $R$ are, respectively, volts (J/C), amperes (C/s), and ohms (V/A). The value of $R$ is influenced by specimen configuration, and for many materials is independent of current. The resistivity is independent of specimen geometry but related to $R$ through the expression

$$\rho = \frac{RA}{l}$$

where $l$ is the distance between the two points at which the voltage is measured, and $A$ is the cross-sectional area perpendicular to the direction of the current. The units for $\rho$ are ohm-meters (Ω -m). From the expression for Ohm’s law resistivity can be expressed as

$$\rho = \frac{VA}{ll}$$

ELECTRICAL CONDUCTIVITY
Sometimes, electrical conductivity $\sigma$ is used to specify the electrical character of a material. It is simply the reciprocal of the resistivity, or

$$\sigma = \frac{1}{\rho}$$

and is indicative of the ease with which a material is capable of conducting an electric current. The units for $\sigma$ are reciprocal ohm-meters [(Ω-m)$^{-1}$ or mho/m]. The following discussions on electrical properties use both resistivity and conductivity. In addition to above Equation, Ohm’s law may be expressed as

$$J = \rho \xi$$

in which $J$ is the current density, the current per unit of specimen area $I/A$, and $\xi$ is the electric field intensity, or the voltage difference between two points divided by the distance separating them; that is,

$$\xi = \frac{V}{l}$$

Solid materials exhibit an amazing range of electrical conductivities, extending over 27 orders of magnitude; probably no other physical property experiences this breadth of variation. In fact, one way of classifying solid materials is according to the ease with which they conduct an electric current; within this classification scheme there are three groupings: conductors, semiconductors, and insulators. Metals are good conductors, typically having conductivities on the order of $10^7$ (Ω-m)$^{-1}$. At the other extreme are materials with very low conductivities, ranging between $10^{-10}$ and $10^{-20}$ (Ω-m)$^{-1}$; these are electrical insulators. Materials with intermediate conductivities, generally from $10^{-6}$ to $10^4$ (Ω-m)$^{-1}$ are termed semiconductors.
ENERGY BAND STRUCTURES IN SOLIDS

In all conductors, semiconductors, and many insulating materials, only electronic conduction exists, and the magnitude of the electrical conductivity is strongly dependent on the number of electrons available to participate in the conduction process. However, not all electrons in every atom will accelerate in the presence of an electric field. The number of electrons available for electrical conduction in a particular material is related to the arrangement of electron states or levels with respect to energy, and then the manner in which these states are occupied by electrons. A thorough exploration of these topics is complicated and involves principles of quantum mechanics that are beyond the scope of this lecture; the ensuing development omits some concepts and simplifies others. Concepts relating to electron energy states, their occupancy, and the resulting electron configuration for isolated atoms are known. By way of review, for each individual atom there exist discrete energy levels that may be occupied by electrons, arranged into shells and subshells. Shells are designated by integers (1, 2, 3, etc.), and subshells by letters (s, p, d, and f). For each of s, p, d, and f subshells, there exist, respectively, one, three, five, and seven states. The electrons in most atoms fill only the states having the lowest energies, two electrons of opposite spin per state, in accordance with the Pauli Exclusion Principle. The electron configuration of an isolated atom represents the arrangement of the electrons within the allowed states. Let us now make an extrapolation of some of these concepts to solid materials. A solid may be thought of as consisting of a large number, say, \( N \), of atoms initially separated from one another, which are subsequently brought together and bonded to form the ordered atomic arrangement found in the crystalline material. At relatively large separation distances, each atom is independent of all the others and will have the atomic energy levels and electron configuration as if isolated. However, as the atoms come within close proximity of one another, electrons are acted upon, or perturbed, by the electrons and nuclei of adjacent atoms. This influence is such that each distinct atomic state may split into a series of closely spaced electron states in the solid, to form what is termed an electron energy band. The extent of splitting depends on interatomic separation (Figure 1) and begins with the outermost electron shells, since they are the first to be perturbed as the atoms coalesce. Within each band, the energy states are discrete, yet the difference between adjacent states is exceedingly small. At the equilibrium spacing, band formation may not occur for the electron subshells nearest the nucleus, as illustrated in Figure 2b. Furthermore, gaps may exist between adjacent bands, as also indicated in the figure; normally, energies lying within these band gaps are not available for electron occupancy. The conventional way of representing electron band structures in solids is shown in Figure 2a. The number of states within each band will equal the total of all states contributed by the \( N \) atoms. For example, an s band will consist of \( N \) states, and a p band of \( 3N \) states. With regard to occupancy, each energy state may accommodate two electrons, which must have oppositely directed spins. Furthermore, bands will contain the electrons that resided in the corresponding levels of the isolated atoms; for example, a 4s energy band in the solid will contain those isolated atom’s 4s electrons. Of course, there will be empty bands and, possibly, bands that are only partially filled. The electrical properties of a solid material are a consequence of its electron band structure—that is, the arrangement of the outermost electron bands and the way in
Figure 2

which they are filled with electrons. Four different types of band structures are possible at 0 K. In the first (Figure 3a), one outermost band is only partially filled with electrons. The energy corresponding to the highest filled state at 0 K is called the Fermi energy as indicated. This energy band structure is typified by some metals, in particular those that have a single $s$ valence electron (e.g., copper). Each copper atom has one $4s$ electron; however, for a solid comprised of $N$ atoms, the $4s$ band is capable of accommodating $2N$ electrons. Thus only half the available electron positions within this $4s$ band are filled. For the second band structure, also found in metals (Figure 3b), there is an overlap of an empty band and a filled band. Magnesium has this band structure. Each isolated Mg atom has two $3s$ electrons. However, when a solid is formed, the $3s$ and $3p$ bands overlap. In this instance and at 0 K, the Fermi energy is taken as that energy, below which, for $N$ atoms, $N$ states are filled, two electrons per state. The final two band structures are similar; one band (the valence band) that is completely filled with electrons is separated from an empty conduction band, and

Figure 3

an energy band gap lies between them. For very pure materials, electrons may not have energies within this gap. The difference between the two band structures lies in the magnitude of the energy gap; for materials that are insulators, the band gap is relatively wide (Figure 3c), whereas for semiconductors it is narrow (Figure 4d). The Fermi energy for these two band structures lies within the band gap—near its center.

CONDUCTION IN TERMS OF BAND AND ATOMIC BONDING MODELS

At this point in the discussion, it is vital that another concept be understood—namely, that only electrons with energies greater than the Fermi energy may be acted on and accelerated in the presence of an electric field. These are the electrons that participate in the conduction process, which are termed free electrons. Another charged electronic entity called a hole is found in semiconductors and insulators. Holes have energies less than $E_f$ and also participate in electronic conduction. As the ensuing discussion reveals, the electrical conductivity is a direct function of the numbers of free electrons and holes. In addition, the distinction between conductors and nonconductors (insulators and semiconductors) lies in the numbers of these free electron and hole charge carriers.
Metals
For an electron to become free, it must be excited or promoted into one of the empty and available energy states above $E_f$. For metals having either of the band structures shown in Figures 3a and 3b, there are vacant energy states adjacent to the highest filled state at $E_f$. Thus, very little energy is required to promote electrons into the low-lying empty states, as shown in Figure 4. Generally, the energy provided by an electric field is sufficient to excite large numbers of electrons into these conducting states. For the metallic bonding model, it was assumed that all the valence electrons have freedom of motion and form an “electron gas,” which is uniformly distributed throughout the lattice of ion cores. Although these electrons are not locally bound to any particular atom, nevertheless, they must experience some excitation to become conducting electrons that are truly free. Thus, although only a fraction is excited, this still gives rise to a relatively large number of free electrons and, consequently, a high conductivity.

Insulators and Semiconductors
For insulators and semiconductors, empty states adjacent to the top of the filled valence band are not available. To become free, therefore, electrons must be promoted across the energy band gap and into empty states at the bottom of the conduction band. This is possible only by supplying to an electron the difference in energy between these two states, which is approximately equal to the band gap energy $E_g$. This excitation process is demonstrated in Figure 5. For many materials this band gap is several electron volts wide. Most often the excitation energy is from a nonelectrical source such as heat or light, usually the former. The number of electrons excited thermally (by heat energy) into the conduction band depends on the energy band gap width as well as temperature. At a given temperature, the larger the $E_g$ the lower is the probability that a valence electron will be promoted into an energy state within the conduction band; this results in fewer conduction electrons. In other words, the larger the band gap, the lower is the electrical conductivity at a given temperature. Thus, the distinction between semiconductors and insulators lies in the width of the band gap; for semiconductors it is narrow, whereas for insulating materials it is relatively wide. Increasing the temperature of either a semiconductor or an insulator results in an increase in the thermal energy that is available for electron excitation. Thus, more electrons are promoted into the conduction band, which gives rise to an enhanced conductivity. The conductivity of insulators and semiconductors may also be viewed from the perspective of atomic bonding models. For electrically
insulating materials, interatomic bonding is ionic or strongly covalent. Thus, the valence electrons are tightly bound to or shared with the individual atoms. In other words, these electrons are highly localized and are not in any sense free to wander throughout the crystal. The bonding in semiconductors is covalent (or predominantly covalent) and relatively weak, which means that the valence electrons are not as strongly bound to the atoms. Consequently, these electrons are more easily removed by thermal excitation than they are for insulators.

ELECTRON MOBILITY
When an electric field is applied, a force is brought to bear on the free electrons; as a consequence, they all experience acceleration in a direction opposite to that of the field, by virtue of their negative charge. According to quantum mechanics, there is no interaction between an accelerating electron and atoms in a perfect crystal lattice. Under such circumstances all the free electrons should accelerate as long as the electric field is applied, which would give rise to an electric current that is continuously increasing with time. However, we know that a current reaches a constant value the instant that a field is applied; indicating that there exist what might be termed “frictional forces,” which counter this acceleration from the external field. These frictional forces result from the scattering of electrons by imperfections in the crystal lattice, including impurity atoms, vacancies, interstitial atoms, dislocations, and even the thermal vibrations of the atoms themselves. Each scattering event causes an electron to lose kinetic energy and to change its direction of motion, as represented schematically in Figure 6. There is, however, some net electron motion in the direction opposite to the field, and this flow of charge is the electric current. The scattering phenomenon is manifested as a resistance to the passage of an electric current. Several parameters are used to describe the extent of this scattering; these include the drift velocity and the mobility of an electron. The drift velocity \( v_d \) represents the average electron velocity in the direction of the force imposed by the applied field. It is directly proportional to the electric field as follows:

\[
v_d = \mu_e \xi
\]

The constant of proportionality \( \mu_e \) is called the electron mobility, which is an indication of the frequency of scattering events; its units are square meters per volt-second (m²/V-s). The conductivity \( \sigma \) of most materials may be expressed as

\[
\sigma = n |e| \mu_e
\]

where \( n \) is the number of free or conducting electrons per unit volume (e.g., per cubic meter), and \( |e| \) is the absolute magnitude of the electrical charge on an electron (1.6×10⁻¹⁹ C). Thus, the electrical conductivity is proportional to both the number of free electrons and the electron mobility. The drift mobility is determined by the mean scattering time \( \tau \) of the conduction electrons and is given by

\[
\mu = e \tau / m_e
\]

where \( m_e \) is the mass of the electron (effective mass of a conduction electron in the crystal).
ELECTRICAL RESISTIVITY OF METALS

As mentioned previously, most metals are extremely good conductors of electricity; room-temperature conductivities for several of the more common metals are contained in Table below. Again, metals have high conductivities because of the large numbers of free electrons that have been excited into empty states above the Fermi energy. Thus $n$ has a large value in the conductivity expression. At this point it is convenient to discuss conduction in metals in terms of the resistivity, the reciprocal of conductivity; the reason for this switch in topic should become apparent in the ensuing discussion. Since crystalline defects serve as scattering centers for conduction electrons in metals, increasing their number raises the resistivity (or lowers the conductivity). The concentration of these imperfections depends on temperature, composition, and the degree of cold work of a metal specimen. In fact, it has been observed experimentally that the total resistivity of a metal is the sum of the contributions from thermal vibrations, impurities, and plastic deformation; that is, the scattering mechanisms act independently of one another. This may be represented in mathematical form as follows:

$$\rho_{\text{total}} = \rho_t + \rho_i + \rho_d$$

in which $\rho_t$, $\rho_i$, and $\rho_d$ represent the individual thermal, impurity, and deformation resistivity contributions, respectively. Equation above is sometimes known as Matthiessen's rule. The influence of each variable on the total resistivity is demonstrated in Figure 7, a plot of resistivity versus temperature for copper and several copper-nickel alloys in annealed and deformed states. The additive nature of the individual resistivity contributions is demonstrated at -100°C.

<table>
<thead>
<tr>
<th>Metal</th>
<th>Electrical Conductivity [Ω·m] $^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silver</td>
<td>$6.8 \times 10^7$</td>
</tr>
<tr>
<td>Copper</td>
<td>$6.0 \times 10^7$</td>
</tr>
<tr>
<td>Gold</td>
<td>$4.3 \times 10^7$</td>
</tr>
<tr>
<td>Aluminum</td>
<td>$3.8 \times 10^7$</td>
</tr>
<tr>
<td>Brass (70Cu-30Zn)</td>
<td>$1.6 \times 10^7$</td>
</tr>
<tr>
<td>Iron</td>
<td>$1.0 \times 10^7$</td>
</tr>
<tr>
<td>Platinum</td>
<td>$0.94 \times 10^7$</td>
</tr>
<tr>
<td>Plain carbon steel</td>
<td>$0.6 \times 10^7$</td>
</tr>
<tr>
<td>Stainless steel</td>
<td>$0.2 \times 10^7$</td>
</tr>
</tbody>
</table>

Table 1: Electrical conductivity for some metals at 20°C.
Temperature Coefficient of Resistivity:
As noted in table 1, electrical conductivity values (and resistivity values) are typically reported at 20 °C. This is done because the conductivity and resistivity of material is temperature dependant. The conductivity of most materials decreases as temperature increases. Alternately, the resistivity of most material increases with increasing temperature. The amount of change is material dependant but has been established for many elements and engineering materials. The reason that resistivity increases with increasing temperature is that the number of imperfection in the atomic lattice structure increases with temperature and this hampers electron movement. These imperfections include dislocations, vacancies, interstitial defects and impurity atoms. Additionally, above absolute zero, even the lattice atoms participate in the interference of directional electron movement as they are not always found at their ideal lattice sites. Thermal energy causes the atoms to vibrate about their equilibrium positions. At any moment in time many individual lattice atoms will be away from their perfect lattice sites and this interferes with electron movement. When the temperature coefficient is known, an adjusted resistivity value can be computed using the following formula:

$$\rho = \rho_o [1 + \alpha_o (T - T_o)]$$

Where: \(\alpha_o\) is the temperature coefficient of resistivity; \(\rho_o\) is the resistivity at the reference temperature \(T_o\).

Influence of Impurities
For additions of a single impurity that forms a solid solution, the impurity resistivity \(\rho_i\) is related to the impurity concentration \(C_i\) in terms of the atom fraction (at% \(\times 100\)) as follows:

$$\rho_i = A_C_i (1 - C_i)$$

where \(A\) is a composition-independent constant that is a function of both the impurity and host metals. The influence of nickel impurity additions on the room temperature resistivity of copper is demonstrated in Figure 8, up to 50 wt% Ni; over this composition range nickel is completely soluble in copper. Again, nickel atoms in copper act as scattering centers, and increasing the concentration of nickel in copper results in an enhancement of resistivity.

Matthiessen’s Rule
The drift mobility \(\mu\) of conduction electrons in a conductor is controlled by the mean free time \(\tau\) between the scattering events that randomize the electron’s velocity. A conduction electron, on average, travels for a time \(\tau\), gains velocity from the applied electric field and then collides with a scattering center and looses the velocity gained from the field. These velocity randomizing collisions prevent the electron from accelerating for ever and hence result in a drift velocity that is parallel to the applied electric field. Drift mobility is simply the drift
velocity per unit applied field. The drift mobility is related to the mean free time \( \tau \) between the scattering events; this mean collision time is also called the relaxation time. \( 1/\tau \) is the mean frequency of electron scattering events, or collisions. It is also the mean probability per unit time that the electron will be scattered. The relationship between the drift mobility \( \mu \) and the mean collision time \( \tau \) is

\[
\mu = \frac{e \tau}{m_e} \quad \text{(A)}
\]

where \( e \) is the electronic charge and \( m_e \) is the mass of the electron in the crystal (its effective mass). Suppose that \( l \) is the electron mean free path associated with a particular scattering process in a metal (such as scattering from atomic vibrations of the crystal) as shown in Figure 9. If \( u \) is the mean speed of the conduction electrons, then the mean free path \( l = u \tau \). The mean speed \( u \) in a metal is relatively temperature insensitive and is determined by the random motions of the conduction electrons around the Fermi energy \( E_F \). Thus, \( 1/2meu^2 \) is on the order of \( E_F \) (several electron volts) and \( u \) is therefore on the order of \( \sim 10^6 \) m s\(^{-1}\). The drift mobility as controlled by this scattering process is given by Equation (B).

\[
\mu = \frac{el}{m \mu} \quad \text{(B)}
\]

The corresponding resistivity \( \rho \) is,

\[
\rho = \frac{1}{en\mu} = \frac{m \mu}{e\tau l} \quad \text{(C)}
\]

where \( n \) is the conduction electron concentration. Clearly if the mean free path \( l \) for a particular scattering mechanism is very long, the corresponding resistivity resulting from that process is very low (and insignificant). Right after the scattering process, it is assumed that the velocity of the electron is in a random direction, that is the scattering process randomizes the electron’s velocity. The mean free path \( l \) in Equation (B) assumes that once the electron travels a distance \( l \) it becomes scattered and its velocity right after the scattering is random. Stated differently, the direction of the electron after the scattering process is totally independent of its direction before scattering. Thus, it takes a mean distance \( l \) to randomize the electron’s velocity. If the scattering process does not totally randomize the velocity, then we have to consider how many collisions, say \( \nu \), it takes for the velocity to be randomized. The effective mean free path is then longer by a factor \( \nu \) so that the drift mobility is,

\[
\mu = \frac{evl}{m \mu} \quad \text{(D)}
\]
where \( \nu_l \) is the effective mean free path, that is the mean free path for totally randomizing the electron’s velocity. What happens if there are more than one type of scattering mechanisms operating on the conduction electrons? Figure 10 shows how an electron can be scattered by thermal vibrations of the atoms and also by impurity atoms in the crystal. We now effectively have two types of mean free times between collisions. One, \( \tau_T \), for scattering from thermal vibrations only and the other, \( \tau_I \), for scattering from impurities only as indicated in Figure 10. In other words, \( \tau_T \) is the mean time between scattering events arising from thermal vibrations alone and \( \tau_I \) is the mean time between scattering events arising from collisions with impurities alone. In general, an electron may become scattered by both processes so that the effective mean free time \( \tau \) between any two scattering events will be less than the individual scattering times \( \tau_T \) and \( \tau_I \). The electron will therefore be scattered when it collides with either an atomic vibration or with an impurity atom. Since \( 1/\tau_T \) is the frequency of collisions with lattice vibrations alone and \( 1/\tau_I \) is the frequency of collisions with the impurities alone, then the total (overall) frequency of collisions is the sum of these two:

\[
\frac{1}{\tau} = \frac{1}{\tau_T} + \frac{1}{\tau_I} \quad \text{(E)}
\]

In writing Equation (E) for the various frequencies (probabilities) we make the reasonable assumption that, to a greater extent, the two scattering mechanisms are essentially independent. Here the effective mean scattering time \( \tau \) is clearly smaller than \( \tau_T \) and \( \tau_I \). In unit time, the overall number of collisions (1/\( \tau \)) is the sum of number of collisions with thermal vibrations alone (1/\( \tau_T \)) and the number of collisions with impurities alone (1/\( \tau_I \)). By multiplying Equation (E) by 1/\( \mu \), we can find the overall mean free path \( l \) in terms of the mean free path for collisions with thermal vibrations alone, \( l_T \), and mean free path for collisions with impurities alone, \( l_I \),

\[
\frac{1}{l} = \frac{1}{l_T} + \frac{1}{l_I} \quad \text{(F)}
\]

The drift mobility \( \mu \) depends on the effective scattering time \( \tau \) via \( \mu = e\tau / m_e \) so that the above can be just as well be written in terms of the drift mobilities determined by the various scattering mechanisms. In other words,

\[
\frac{1}{\mu} = \frac{1}{\mu_T} + \frac{1}{\mu_I} \quad \text{(G)}
\]
where \( \mu \) is the lattice scattering limited drift mobility and \( \mu_I \) is the impurity scattering limited drift mobility where by definition \( \mu = e\tau /m_e \) and \( \mu_I = e\tau_I /m_e \). The effective (or overall) resistivity, \( \rho \), of the material is simply \( 1/(\rho_{\mu}) \) or

\[
\rho = \frac{1}{en\mu} + \frac{1}{en\mu_I}
\]

which can be written as

\[
\rho = \rho_T + \rho_I
\]

where we have defined \( 1/(en\mu) \) as resistivity due to scattering from thermal vibrations and \( 1/(en\mu_I) \) as the resistivity due to scattering from impurities, i.e. \( \rho_T = 1/(en\mu) \) and \( \rho_I = 1/(en\mu_I) \).

The final result in Equation (I) simply states that the effective resistivity \( \rho \) is the sum of two contributions. First, \( \rho_T = 1/(en\mu) \), is the resistivity due to scattering by thermal vibrations of the host atoms. For those near-perfect pure metal crystals, this is the dominating contribution. As soon as we add impurities, however, there is an additional resistivity, \( \rho_I = 1/(en\mu_I) \), which arises from the scattering of the electrons from the impurities. The summation rule of resistivities from different scattering mechanisms as endowed in Equation (I) is called Matthiessen's Rule. It can be generalized as follows: Suppose that \( l_i \) is the effective mean free path (for randomizing collisions) associated with a particular electron scattering mechanism of type \( i \), and if \( \rho_i \) is the corresponding resistivity, then the overall mean free path \( l \) and resistivity \( \rho \) are

\[
\frac{1}{l} = \sum \frac{1}{l_i} \quad \text{and} \quad \rho = \sum \rho_i
\]

Each mean free path in Equation (J) is an effective mean free path that is the mean intercollision distance after which the velocity of the electron is random. If it takes \( v_i \) number of collisions to randomize the electron’s velocity, then the effective mean free path is \( v_i l_i \).
Learning Objectives
After careful study of this lecture you should be able to do the following:

1. Determine the magnetization of some material given its magnetic susceptibility and the applied magnetic field strength.
2. Briefly explain the nature and source of (a) diamagnetism, (b) paramagnetism, and (c) ferromagnetism.
3. (a) Describe magnetic hysteresis; (b) explain why ferromagnetic and ferrimagnetic materials experience magnetic hysteresis; and (c) explain why these materials may become permanent magnets.

INTRODUCTION

Modern technology would be unthinkable without magnetic materials and magnetic phenomena. Magnetic tapes or disks (for computers, video recorders, etc.) motors, generators, telephones, transformers, permanent magnets, electromagnets, loudspeakers, and magnetic strips on credit cards are only a few examples of their applications. At least five different kinds of magnetic materials exist. They have been termed para-, dia-, ferro-, ferrri-, and antiferromagnetics. A qualitative as well as a quantitative distinction between these types can be achieved in a relatively simple way by utilizing a method proposed by Faraday. The magnetic material to be investigated is suspended from one of the arms of a sensitive balance and is allowed to reach into an inhomogeneous magnetic field (Figure 1). Diamagnetic materials are expelled from this field, whereas para-, ferro-, antiferro-, and ferrimagnetics are attracted in different degrees. It has been found empirically that the apparent loss or gain in mass, that is, the force, $F$, on the sample exerted by the magnetic field, is:

$$F = V \chi \mu_0 H \frac{dH}{dx}$$

Where $V$ is the volume of the sample, $\mu_0$ is a universal constant called the permeability of free space ($1.257 \times 10^{-6}$ H/m or Vs/Am), and $\chi$ is the susceptibility, which expresses how responsive a material is to an applied magnetic field. Characteristic values for $\chi$ are given in Table 1. The term $dH/dx$ is the change of the magnetic field strength $H$ in the $x$-direction. The field strength $H$ of an electromagnet (consisting of helical windings of a long, insulated wire as seen in the lower portion of Figure 1) is proportional to the current, $I$, which flows through this coil, and on the number, $n$, of the windings (called turns) that have been used to make the coil. Further, the magnetic field strength is inversely proportional to the length, $L$, of the solenoid. Thus, the magnetic field strength is expressed by:
The field strength is measured (in SI units) in “Amp-turns per meter” or shortly, in A/m. The magnetic field can be enhanced by inserting, say, iron, into a solenoid, as shown in Figure 1. The parameter which expresses the amount of enhancement of the magnetic field is called the permeability $\mu$. The magnetic field strength within a material is known by the names magnetic induction (or magnetic flux density) and is denoted by $B$. Magnetic field strength and magnetic induction are related by the equation:

$$B = \mu \mu_0 H.$$  

The SI unit for $B$ is the tesla (T) and that of $\mu_0$ is henries per meter (H/m or Vs/Am). The relationship between the susceptibility and the permeability is

$$\mu = 1 + \chi.$$  

For empty space and, for all practical purposes, also for air, one defines $\chi = 0$ and thus $\mu = 1$. The susceptibility is small and negative for diamagnetic materials. As a consequence, $\mu$ is slightly less than 1. For para- and antiferromagnetic materials, $\chi$ is again small, but positive. Thus, $\mu$ is slightly larger than 1. Finally, $\chi$ and $\mu$ are large and positive for ferro- and ferrimagnetic materials. The magnetic constants are temperature-dependent, except for diamagnetic materials, Further, the susceptibility for ferromagnetic materials depends on the field strength, $H$. In free (empty) space, $B$ and $\mu_0 H$ are identical. Inside a magnetic material the induction $B$ consists of the free-space component ($\chi_0 H$) plus a contribution to the magnetic field ($\mu_0 M$) which is due to the presence of matter [Figure 2(a)], that is,

$$B = \mu_0 H + \mu_0 M,$$

where $M$ is called the magnetization of the material. Which relate with $\chi$ and $H$ by:

$$M = \chi H.$$  

$H$, $B$, and $M$ are actually vectors. Specifically, outside a material, $H$ (and $B$) point from the north to the South Pole. Inside of a ferro- or paramagnetic material, $B$ and $M$ point from the south to the north;

<table>
<thead>
<tr>
<th>Material</th>
<th>$\chi$ (SI) unitless</th>
<th>$\chi$ (cgs) unitless</th>
<th>$\mu$ unitless</th>
<th>Type of magnetism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bi</td>
<td>$-165 \times 10^{-6}$</td>
<td>$-13.13 \times 10^{-6}$</td>
<td>0.99983</td>
<td>Diamagnetic</td>
</tr>
<tr>
<td>Ge</td>
<td>$-71.1 \times 10^{-6}$</td>
<td>$-5.66 \times 10^{-6}$</td>
<td>0.99993</td>
<td></td>
</tr>
<tr>
<td>Au</td>
<td>$-34.4 \times 10^{-6}$</td>
<td>$-2.74 \times 10^{-6}$</td>
<td>0.99996</td>
<td></td>
</tr>
<tr>
<td>Ag</td>
<td>$-25.3 \times 10^{-6}$</td>
<td>$-2.016 \times 10^{-6}$</td>
<td>0.99997</td>
<td></td>
</tr>
<tr>
<td>Be</td>
<td>$-23.2 \times 10^{-6}$</td>
<td>$-1.85 \times 10^{-6}$</td>
<td>0.99998</td>
<td></td>
</tr>
<tr>
<td>Cu</td>
<td>$-9.7 \times 10^{-6}$</td>
<td>$-0.77 \times 10^{-6}$</td>
<td>0.99999</td>
<td></td>
</tr>
<tr>
<td>Superconductors$^d$</td>
<td>$-1.0$</td>
<td>$-8.0 \times 10^{-2}$</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$\beta$-Sn</td>
<td>$+2.4 \times 10^{-6}$</td>
<td>$+0.19 \times 10^{-6}$</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Al</td>
<td>$+20.7 \times 10^{-6}$</td>
<td>$+1.65 \times 10^{-6}$</td>
<td>1.00002</td>
<td>Paramagnetic</td>
</tr>
<tr>
<td>W</td>
<td>$+77.7 \times 10^{-6}$</td>
<td>$+6.18 \times 10^{-6}$</td>
<td>1.00008</td>
<td></td>
</tr>
<tr>
<td>Pt</td>
<td>$+264.4 \times 10^{-6}$</td>
<td>$+21.04 \times 10^{-6}$</td>
<td>1.00026</td>
<td></td>
</tr>
<tr>
<td>Low carbon steel</td>
<td></td>
<td></td>
<td>$5 \times 10^7$</td>
<td>Ferromagnetic</td>
</tr>
<tr>
<td>Fe-3%Si (grain-oriented)</td>
<td>Approximately the same as $\mu$</td>
<td>because of $\chi = \mu - 1$</td>
<td>$4 \times 10^4$</td>
<td></td>
</tr>
<tr>
<td>Ni-Fe-Mo (supermalloy)</td>
<td></td>
<td></td>
<td>$10^6$</td>
<td></td>
</tr>
</tbody>
</table>

Table 1
Figure 2

$B$ was called above to be the magnetic flux density in a material, that is, the magnetic flux per unit area. The magnetic flux $\Phi$ is then defined as the product of $B$ and area $A$, that is, by

$$\Phi = B \cdot A.$$ 

Finally, we need to define the magnetic moment $\mu_m$ (also a vector) through the following equation:

$$M = \frac{\mu_m}{V},$$

Which means that the magnetization is the magnetic moment per unit volume.

**Diamagnetism**

Ampère postulated more than one hundred years ago that so-called molecular currents are responsible for the magnetism in solids. He compared these molecular currents to an electric current in a loop-shaped piece of wire which is known to cause a magnetic moment. Today, we replace Ampère’s molecular currents by orbiting valence electrons. To understand diamagnetism, a second aspect needs to be considered. A current is induced in a wire loop whenever a bar magnet is moved toward (or from) this loop. The current thus induced causes, in turn, a magnetic moment that is opposite to the one of the bar magnet (Figure 3). Diamagnetism may then be explained by postulating that the external magnetic field induces a change in the magnitude of the atomic currents, i.e., the external field accelerates or decelerates the orbiting electrons, so that their magnetic moment is in the opposite direction to the external magnetic field. In other words, the responses of the orbiting electrons counteract the external field [Figure 2(c)].
Paramagnetism

Paramagnetism in solids is attributed to a large extent to a magnetic moment that results from electrons which spin around their own axis; see Figure 4(a). The spin magnetic moments are generally randomly oriented so that no net magnetic moment results. An external magnetic field tries to turn the unfavorably oriented spin moments in the direction of the external field, but thermal agitation counteracts the alignment. Thus, **spin paramagnetism** is slightly temperature-dependent. It is generally weak and is observed in some metals and in salts of the transition elements. Free atoms (dilute gases) as well as rare earth elements and their salts and oxides possess an additional source of paramagnetism. It stems from the magnetic moment of the orbiting electrons; see Figure 4(b). Without an external magnetic field, these magnetic moments are, again, randomly oriented and thus mutually cancel one another. As a result, the net magnetization is zero. However, when an external field is applied, the individual magnetic vectors tend to turn into the field direction which may be counteracted by thermal agitation. Thus, **electron-orbit paramagnetism** is also temperature-dependent. From the above-said it becomes clear that in paramagnetic materials the magnetic moments of the electrons eventually point in the direction of the external field, that is, the magnetic moments enhance the external field [see Figure 2(a)]. On the other hand, diamagnetism counteracts an external field [see Figure 2(c)].

![Figure 4](image)

Thus, para- and diamagnetism oppose each other. Solids that have both orbital as well as spin paramagnetism are consequently paramagnetic (since the sum of both paramagnetic compounds is commonly larger than the diamagnetism). Rare earth metals are an example of this. In many other solids, however, the electron orbits are essentially coupled to the lattice. This prevents the orbital magnetic moments from turning into the field direction. Thus, electron-orbit paramagnetism does not play a role, and only spin paramagnetism remains. The possible presence of a net spin-paramagnetic moment depends, however, on whether or not the magnetic moments of the individual spins cancel each other. Specifically, if a solid has completely filled electron bands, then a quantum mechanical rule, called the **Pauli principle**, requires the same number of electrons with spins up and with spins down [Figure 4(a)]. The Pauli principle stipulates that each electron state can be filled only with two electrons having opposite spins. The case of completely filled bands thus results in a cancellation of the spin moments and no net paramagnetism is expected. Materials in which this occurs are therefore diamagnetic (no orbital and no spin paramagnetic moments). Examples of filled bands are intrinsic semiconductors, insulators, and ionic crystals such as NaCl. In materials that have partially filled bands, the electron spins are arranged according to **Hund’s rule** in such a manner that the total spin moment is maximized. For example, for an atom with eight valence \( d \)-electrons, five of the spins may point up and three spins point down, which results in a net number of two spins up; Figure 5. The atom then has two units of (para-) magnetism or, as it is said, two Bohr magnetons per atom. The **Bohr magneton** is the smallest unit (or quantum) of the magnetic moment and has the value:

\[
\mu_B = \frac{e\hbar}{4\pi m} = 9.274 \times 10^{-24} \left(\frac{J}{T}\right) = (A \cdot m^2).
\]
Ferromagnetism

Figure 6 depicts a ring-shaped solenoid consisting of a newly cast piece of iron and two separate coils which are wound around the iron ring. If the magnetic field strength in the solenoid is temporarily increased (by increasing the current in the primary winding), then the magnetization (measured in the secondary winding with a flux meter) rises slowly at first and then more rapidly, as shown in Figure 7 (dashed line). Finally, $M$ levels off and reaches a constant value, called the saturation magnetization, $M_s$. When $H$ is reduced to zero, the magnetization retains a positive value, called the remanent magnetization, or remanence, $M_r$. It is this retained magnetization which is utilized in permanent magnets. The remanent magnetization can be removed by reversing the magnetic field strength to a value $H_c$, called the coercive field. Solids having a large combination of $M_r$ and $H_c$ are called hard magnetic materials (in contrast to soft magnetic materials, for which the area inside the loop of Figure 7 is very small). A complete cycle through positive and negative $H$-values as shown in Figure 7 is called a hysteresis loop. The saturation magnetization is temperature-dependent. Above the Curie temperature, $T_c$, ferromagnetics become paramagnetic. The spins of unfilled $d$-bands spontaneously align parallel to each other below $T_c$, that is, they align within small domains (1-100µm in size) without the presence of an external magnetic field; Figure 8(a). The individual domains are magnetized to saturation. The spin direction in each domain is, however, different, so that the individual magnetic moments for virgin ferromagnetic materials as a whole cancel each other and the net magnetization is zero. An external magnetic field causes those domains whose spins are parallel or nearly parallel to the external field to grow at the expense of the unfavorably aligned domains; Figure 8(b). When the entire crystal finally contains only one single domain, having spins aligned parallel to the external field direction then the material is said to have reached technical saturation magnetization, $M_s$ [Figure 8(c)]. An increase in temperature progressively destroys the spontaneous alignment, thus reducing the saturation magnetization, Figure 8(d).
Antiferromagnetism
Antiferromagnetic materials exhibit, just as ferromagnetics, a spontaneous alignment of spin moments below a critical temperature (called the Néel temperature). However, the responsible neighboring atoms in antiferromagnetics are aligned in an antiparallel fashion (Figure 9). Actually, one may consider an antiferromagnetic crystal to be divided into two interpenetrating sublattices, A and B, each of which has a spontaneous alignment of spins. Figure 9 depicts the spin alignments for two manganese compounds. (Only the spins of the manganese ions contribute to the antiferromagnetic
behavior.) Figure 9(a) implies that the ions in a given \{110\} plane possess parallel spin alignment, whereas ions in the adjacent plane have antiparallel spins with respect to the first plane. Thus, the magnetic moments of the solid cancel each other and the material as a whole has no net magnetic moment. Most antiferromagnetics are found among ionic compounds such as MnO, MnF₂, FeO, NiO, and CoO. They are generally insulators or semiconductors. Additionally, manganese and chromium are antiferromagnetic.

**Ferrimagnetism**

Ferrimagnetic materials such as NiO. Fe₂O₃ or FeO . Fe₂O₃ are of great technical importance. They exhibit a spontaneous magnetic moment (Figure 8) and hysteresis (Figure 7) below a Curie temperature, just as iron, cobalt, and nickel do. In other words, ferrimagnetic materials possess, similar to ferromagnetics, small domains in which the electron spins are spontaneously aligned in parallel. The main difference to ferromagnetics is, however, that ferrimagnetics are ceramic materials (oxides); they are therefore poor electrical conductors. To explain the spontaneous magnetization in ferrimagnetics, Néel proposed that two sublattices, say A and B, should exist in these materials (just as in antiferromagnetics), each of which contains ions whose spins are aligned parallel to each other. Again, the spins of the ions on the A sites are antiparallel to the spins of the ions on the B sites. The crucial point is that each of the sublattices contains a different amount of magnetic ions. This causes some of the magnetic moments to remain uncancelled. As a consequence, a net magnetic moment results. Ferrimagnetic materials can thus be described as *imperfect antiferromagnetics*. nickel ferrite has two uncancelled spins and therefore two Bohr magnetons per formula unit. This is essentially observed by experiment.
.