CHAPTER 1

INTRODUCTION

1.1 HISTORY OF RAILWAYS

The fundamental of rail transport consists of moving heavy goods and people speedily and safely through a prepared track that supports and guides the vehicle. The history of Rail transport dates back nearly 2500 years. The early Greeks, around 600 BC, observed that the parallel rut, or wagon-way, made by their two wheeled animal drawn carriages, not only guided the vehicle but also made the vehicle go faster. This prompted them to make smooth limestone ways for transportation of heavy material for building monuments. Wheeled vehicles pulled by men and animals, ran in these limestone grooves, which provided the track element, and this idea later graduated to running vehicles on stone/wooden rails. The early wagon-ways, in which horse pulled wagons ran on wooden rails, were primarily used for transportation of ore, coal and stones from the mines/queries to the nearest river front for onward transportation to the industry centers.

Various types of rails (like iron plated/L-shaped/edge rail) and wheels (flanged/unflanged) were experimented upon and finally, the flanged wheel proved its superiority due to its excellent performance on curves and composite iron/wood rails were eventually replaced by metal rails for their better durability and safety. The dramatic rise in fodder price in 1800s decreased the economic viability of goods transportation by horse drawn wagons. Motivated by James Watt’s stationary steam engine, Richard Trevithick developed the first full scale working railway steam locomotive in 1804, which was later refined by George Stephenson in year 1825, and accordingly, world’s first steam railway train was run on September 27, 1825, on the Stockton and Darlington Railway, connecting the collieries at Darlington with the port of Stockton on the river Tees, in northeast England.

Driven by Stephenson, the train hauled an 80-ton load of coal & flour and a passenger car carrying dignitaries, covered 15 Kilometers in two hours, reaching a maximum speed of
Kilometers per hour. The success of the Stockton & Darlington Railway encouraged more railways such as Liverpool & Manchester Railway (1830), Grand Junction Railway (1837) and many more. Initially, the main interest was in goods traffic but it was found that the passenger traffic was just as remunerative and railways quickly became vital for the swift movement of goods and labor that was needed for industrialization.

The earliest steam locomotives were small four wheeled locos and were less efficient and costly to operate. The high cost of crewing and fuelling, low mileage and declining availability of coal spurred on the development of diesel and electric locomotives. The world’s first diesel locomotive, modeled on Dr. Rudolf Diesel’s compression ignition engine (1892), was operated in year 1912 on the Winterthur-Romanshorn Railroad in Switzerland, while the first electric locomotive, working on three phase AC motors, was introduced by Charles Brown & Walter Bovary in year 1899 on the Bergdorf - Thune line, Switzerland.

1.2 RAIL TRANSPORT IN INDIA

The successful running of railways in England prompted the English to introduce railways in India. Accordingly, as per the directives of the then Governor General of India, Lord Harding, the first passenger train in India was introduced by Great Indian Peninsula Railway Company on April 16, 1853 at Bombay, between Bori-bundar and Thane. During years 1854 and 1910, a large number of railway lines were built such as Howrah – Hooghly (1854), Veyasarpaudy - Walajah Road (1856) Allahabad – Kanpur (1859) by private companies such as East Indian Railway Company, the Madras Railway Company, the Bombay Baroda and Central Indian Railway Company. In year 1908, the first electric locomotive made its appearance. In year 1923, based on the Acworth committee recommendations, the railways became a government enterprise. Consequent to attainment of independence and partition of the country on 15th August 1947, the railways were reorganized by partitioning the border railways and amalgamating all the existing railways such as Gaikwad’s Baroda State Railway, Bikaner State Railway, Cutch State Railway, Nizam’s State Railway and many others, spanning over 55,000 Kilometers. Thus the Indian Railways was born.
In the year 1951–52, for the sake of ease of administration, the entire railway system was divided in six zonal administrative units and between the years 1955 & 1966, these six zones were again reorganized to form nine zones on need basis. Gradually, railway electrification was done in 1987 and computerization of reservation was introduced. Another re-organization of railways between the years 2001 & 2003 increased the number of zones from nine to sixteen.

1.3 INDIAN RAILWAYS

The railways in India are controlled and operated by Indian Railways, which is under Railway Ministry’s control. The Minister for Railways is the head of the ministry and the Railway Board is the administrative authority. Indian Railways is spread over more than 64,460 route-kilometer and has total 7133 stations. IR ranks fourth in the world, in terms of route-kilometer length and some of its prominent attributes are:

(i) Indian Railways is divided into sixteen zones, which consist of number of divisions (total sixty-seven divisions). The General Manager is the senior most officer in zone, while Divisional Railway Managers are the in-charge of divisions.

(ii) The staff strength is approximately 1.32 million, including gazetted and non-gazetted cadres.

(iii) Indian Railways’ infrastructure includes the fixed assets such as track, signal & telecommunication installations, traction equipment, land & buildings and rolling stock such as wagons / coaches / locomotives.

(iv) Three types of track gauges are prevalent: Broad Gauge (1,676mm.), Meter Gauge (1,000mm.) & Narrow Gauge (762mm.), which are about 85.61%, 10.56% & 3.81% (respectively) of the total route-kilometer (Indian Railways annual report & accounts, 2012).

(v) IR uses 25 KV AC traction through overhead catenary delivery, except on the Mumbai (Central Railway) & Kolkata suburban systems. As of March 2011, around 19,610 Route-Kilometer (31% of total length) is electrified.

(vi) All major routes have been provided with color light signals, but in some remote areas having less traffic density, semaphore signals are still in use. Automatic
block systems are in use on some of the high-traffic sections. Of the total 7133 railway stations, the signals at 4704 stations are interlocked. Advance electronic signaling equipment such as digital axle counters and electronic interlocking have been deployed to ensure safe and speedy train operations.

(vii) Railways have its own telecommunication network comprising of digital microwave and OFC communication systems. All the divisional / zonal headquarters, major centers and stations are linked with advanced telephony.

(viii) The locomotive fleet of IR is 9137 numbers strong, and comprises of approximately 4033 electric locomotives and 5137 diesel locomotives. The diesel locomotives are built at DLW, Varanasi, while the electric locos are built at CLW, Chittaranjan.

(ix) IR has about 2,00,000 freight wagons and more than 50,000 coaches. The wagons are open / covered / tank types and can carry loads of up to 22Tons. The passenger coaches have a speed potential of up to 160Kilometer per hour and are of ordinary / sleeper / air-conditioned / chair car / double-decker types. The coaches are built at ICF, Perambur and RCF, Kapurthala.

(x) IR has a mixed traffic system, that is, passenger (superfast/express/commuter) trains and freight trains run on the same tracks. Approximately 9,000 trains run every day carrying 200Lakh passengers while 2.5 Million Tongoods are transported daily. IR mainly carries heavy goods such as mineral ores, fertilizers, petrochemicals, iron & steel, coal and food grains.

(xi) Indian Railways has its own budget, which is separate from the general budget. The major source of earning is goods transportation, while the passenger operations are subsidized to a large extent, as the same are a loss-incurring venture. The operating ratio of IR was 95.28 during the year 2010–11, which had increased in comparison to its previous year.

1.4 HISTORY OF RAILWAY SIGNALING

In railway system, the vehicles are constrained to move on a fixed track and can’t be steered away as in the case of other transports. As such, trains are required to follow one another on the same track; otherwise, separate parallel paths would be required for every
train, which is practically impossible. Besides, trains are to be diverted, either for overtaking vehicles moving in the same direction or for crossing the vehicles from the opposite direction. Thus, a guiding system is required to streamline the movement of trains. Additionally, sometimes the track may get damaged or obstructed by some external agency; in such cases, it is vital that the train driver is warned in advance, so that he can control the train. Thus, the requirements of safety, speed and smooth operations in train running dictate that information about the section ahead should be communicated to the train driver in some form of audio–visual signal. Hence, the need of signaling for railways.

In the beginning, the train drives were required to be constantly on the lookout for obstructions, in order to be able to stop the train before colliding with it. But due to driver inexperience, bad brakes and rather tenuous rail-wheel contact, this system proved to be lacking. To address these issues, a flagman mounted on a horse preceded the trains and displayed flags / lanterns signals to direct the train drivers. But this restricted the speed of the train and was found impractical. After that, men (known as watchmen) standing beside the line used to convey the train drivers of the train ahead. However, they didn’t have any means to know the status of earlier train. Concern for safety of the trains led to the development of other systems of train movement control, such as the Time interval system and the Space interval system.

In time interval method, trains were run one after another, at fixed time intervals. However, the driver had no means to know whether the earlier train has reached its destination; besides, this ten-minute time interval severely restricted the line capacity. Thus, the time interval method of train movement control was not fully suitable for the high speed, high-density train operations. In space interval method, the entire track is distributed into sections and one train – one section method is followed. Thus, between two consecutive trains, there is a definite space interval. Entry into and exit from a block section is governed by the stationmaster through signals and communication system.

A signal is a medium to convey a particular predetermined meaning in nonverbal form. Mechanical or electrical equipment, fixed along the track, are used to inform the driver of
the status of line ahead. If the line is clear of obstructions, the signal conveys the driver to take the train forward, while if the conditions are not suitable for the train to go forward, the signal shall stop the train. These signals could be either a semaphore arm or a disc for displaying aspects during daytime, while an oil lamp with colored lens in front is used during nighttime. The red light indicates stop, while yellow and green lights signal the train to move forward.

1.4.1 MECHANICAL SIGNALS

Locally worked lower quadrant semaphore signals were first installed in year 1842 in London. The signals were named according to their function/location. Thus, the main signal allowing the train to home into the station was called Home signal, while the signals at the outer periphery of the station were called Outer/Distant signal. The signal used to start the train was called Starter while a signal in advance of the starter signal, which allowed entry into the block section ahead, was called the Advanced Starter.

Mechanical signals are worked by levers through steel wires. The semaphore signals are of two types: Two positions (Lower Quadrant) and Three Positions (Upper Quadrant). These signals have different day & night aspects – arm in daytime and light in nighttime. Thus, the driver has to learn two sets of indications, which is tedious. The visibility of these signals is restricted, more so during fog/nighttime. The mechanical parts require more maintenance effort and the technology is almost obsolete. This limitation has been removed by the innovations in electric circuits technology.

1.4.2 ELECTRICAL COLOR LIGHT SIGNALS

Color light signals were first used in United States in year 1904 on Boston Elevated Railway. Color light signals comprise of an electric lamp having a colored Fresnel lens in front to focus the beam. Each lamp is associated with a relay, and is lit according to the state of the section ahead. As shown in Fig. 1.1, the CLS are of three types – two aspects (Red/Green; Red/Yellow), three aspects (Red/Yellow/Green; Yellow/Yellow/green) and four aspects (Red/Yellow/Yellow/Green). For high speed signaling, four aspect signals are predominantly used.
The advantage with color light signals is that they show the same aspect during day/night; besides their maintenance requirements are quite meager.

1.5 RAILWAY SIGNALING IN INDIA

The evolution of signaling in India closely followed the progress in signaling in United Kingdom & America. The initial signaling was of mechanical type (wire operated signals & locally operated points, followed by rod operated points), which later gave way to electromechanical signaling (signal / point machines / color light signals operated by levers) and eventually, electrically operated color light signaling was provided. With the advent of push button panels, the electrical operation of signals / points was popularized. The various milestones in the advancement of railway signaling in India are enumerated in Table 1.1.

Table 1.1 Milestones in Indian Railway Signaling History

<table>
<thead>
<tr>
<th>Year</th>
<th>Milestone</th>
</tr>
</thead>
<tbody>
<tr>
<td>1872</td>
<td>3-position Block instrument for single line provided on Madras Railway</td>
</tr>
<tr>
<td>1885</td>
<td>Home signal provided on the Great Indian Peninsula Railway</td>
</tr>
<tr>
<td>1889</td>
<td>Signal Engineer posted on East Indian Railway</td>
</tr>
<tr>
<td>1892</td>
<td>Fish tailed arms called Warner provided below the outer signal arm</td>
</tr>
<tr>
<td>1910</td>
<td>Power Signaling introduced at Howrah</td>
</tr>
<tr>
<td>1920</td>
<td>Isolation of lines made mandatory</td>
</tr>
<tr>
<td>1927</td>
<td>Train Describer System installed between Churchgate &amp;...</td>
</tr>
</tbody>
</table>
1.6 THE TAO OF SIGNALING

The main purpose of signaling is twofold – ensuring safety in train operations and increasing the line capacity of the section, with minimum of investment.

1.6.1 TENETS OF SIGNAL ENGINEERING

A well-designed signaling system should invariably encompass the following essential features:

(i) Fail Safe Feature: Failure in any of the components / sub-systems shall result in safe side failure of the system.

(ii) One Aspect One indication: Aspects shall be distinctive, unambiguous and communicate the same intelligence to the driver, under all circumstances.

(iii) Sighting distance of a signal: Signal should be visible to the driver from a distance minimum required to react to the most restrictive aspect of the signal.

(iv) Normal danger system: One operation, one clearance. Second clearance should require normalizing of preceding operation.

(v) Visibility of signal aspect from place of operation: The stationmaster should know the result of his action of operating signals.
(vi) Continuous visibility of signal to driver: Enables the driver to react, if there is any change in the aspect of signal, after first viewing.

(vii) Block signaling to be independent of yard signaling: Block working should not be affected by the operations in the yard. Ensures flexibility in operations.

(viii) Uniformity of signaling on a section: Driver should not have to remember different types of signals on a single section, to avoid confusion.

1.6.2 CONSTITUENTS OF SIGNALING SYSTEM

The components of a signaling system are as shown in Fig. 1.2:

![Yard layout with various signal functions](image)

Signal: Signal is the indicator provided for the driver of a train and conveys the state of section ahead.

Point: Point and crossings are used for allowing the trains to switch from one track to another.

Track circuit: The track circuit has a voltage source at one end and a relay at the other end of the rails. Normally the relay is in picked up condition, due to voltage flowing through the rails. When the two rails are shorted by the train’s wheels & axle, the relay supply is cut off and it drops. The relay position (picked up / dropped) is used to indicate the vacancy / occupancy of tracks.

Level Crossing: A level crossing or railroad crossing is the intersection of railway line and road on same level. To protect the train from road vehicles, mechanically / electrically operated gates or barriers are installed. The Level Crossing gates are interconnected with
the signals protecting it, and can be opened only when the concerned rail signals are at ON position.

1.6.3 INTERLOCKING

An interconnection of signaling functions such as signals / points / track circuits, operated from a panel or lever frame, either mechanically or electrically or both, such that their operation must take place in proper sequence to ensure safety is termed as interlocking. The relationship between signaling functions is effected through the rods attached to levers (in case of mechanical interlocking) or relays (in case of electrical interlocking).

1.6.3.1 Essentials of interlocking

The Levers / buttons provided for the operation and controls of Signals, Points, Track circuits, Level crossing gates and Block instrument shall be so interconnected that:

(i) All the points and interlocked gates should be set / closed and locked, and the line should be clear up to the next signal, before the signal is taken off.

(ii) No points can be altered nor any closed gates be opened, once the signal has been switched to off condition.

(iii) No two conflicting signals can be taken off at the same time.

(iv) Interlocking of points should not allow conflicting train movement.

Based on the compliment of signaling required and the speed of running through Trains, Standards of Interlocking have been classified in four categories - Standard I / II / III / IV, each having an associated speed potential and signaling paraphernalia.

1.6.3.2 Types of interlocking

Based on the type of signaling equipment, interlocking is of three types:

Mechanical Interlocking: In mechanical interlocking, the signaling functions are interconnected through mechanical means, such as locks and dogs between the bars connected to the function levers.

Electrical Interlocking: In Electrical Interlocking, the functions are controlled through electromagnetic devices called relays. It consists of complex signal circuits constituted by
relays. The relay positions are used to interlock signaling functions. Based on the movement flexibility required, electrical interlocking can either be Panel Interlocking (PI) – all the points are individually set, or Route Relay Interlocking (RRI) – the points in entire route are set automatically.

Electronic Interlocking: In Electronic Interlocking, the relay logic is replicated with software logic, and consists of CPUs, which give commands for operating the signaling functions as per the stored predefined software logic. For providing redundancy, the EI can have two / three processors, in wired hot standby mode, and the action is taken through polling (two out of two / two out of three).

1.6.4 BLOCK WORKING

Signaling block systems enable safe & efficient operation of trains. Block working systems are defined by the type of signaling equipment being used and the set of train operation rules being employed for controlling the train movement and are mainly of three types – Absolute Block System (the section between two stations is considered as one block and only one train is allowed in it at one time), Automatic Block System (the section between two stations is divided into many blocks, each governed by a signal, and one train is allowed in each block) and Moving Block System (where the limits of blocks are not fixed and vary in real time, according to the train movement).

1.7 OPERATION OF SIGNALING SYSTEMS

In mechanical signaling, the functions are operated through levers fixed on a lever frame. In electromechanical signaling, circuit-controllers, fitted on the levers are used to electrically control the functions. In electrical signaling, the signals / points / track circuits / LC gates are controlled electrically and operated through an operation cum indication panel, which is installed in the stationmaster’s room. The entire yard layout is replicated on the panel; push buttons for individual signals / points / routes and their corresponding indications are also provided. For clearing a signal, the relevant route entrance and exit buttons are simultaneously pressed, which automatically sets and locks all the points and if all the required predefined conditions are fulfilled, the relevant signal is cleared. The signaling functions are operated through relays, which are wired as per
The signaling circuit. Integrated Power Supply system (with battery backup), provides various voltages for different signaling functions such as 110 Volt AC for signals, 110 Volt DC for points, 110 Volt DC (stepped down to 2 Volt at site) for track circuits and 60 Volt DC / 24 Volt DC for relays.

1.8 CONCLUSION

Signaling systems are designed to work in fail-safe mode that is if any signal equipment fails, then no unsafe situation should be created. Since signaling ensures safe and speedy train operations, reliability of signaling equipment, in particularly relays, is of utmost importance, and has been made more critical in the modern times by the rapid increase in traffic density and the unrelenting demand of faster trains. Thus, it is of utmost importance to study the functioning and performance specifics of relays in detail.

1.9 LITERATURE REVIEW

Railway signaling is a niche subject and hence not much literature was available for review. Besides, no research work has earlier been done on metal - carbon relays, and hence practically no published research papers / conference proceedings were available for appraisal. However, literature review of the related available resources such as books, journals, articles, Indian Railway metal-to-carbon relays specifications, Indian Railway annual reports, Indian Railway Signal Engineering Manual, operating manuals, relay manufacturers datasheets / Quality Assurance Programs and web resources, was undertaken.

C. R. Kothari, in his book - Research Methodology: Methods & Techniques - deals with the various steps involved in the research process that are constantly overlapping and are not mutually exclusive. The first step is defining the research problem, which involves wording the problem in precise terminologies. For this, extensive survey of literature connected with the problem, like journals, published research papers, conference proceedings, and books, is undertaken. Then the hypothesis, which is a provisional supposition regarding the problem is formulated. Then the research design for experimenting and collection of evidence is developed, and this includes determining the
sample size and method of data collection. Then the experiments are conducted and the test data is analyzed using a proper analysis tool to establish relationships / patterns. Based on the findings, hypothesis testing is done by various statistical tests such as Chi-Square tests, t-test, F-test and many others and is either accepted or rejected. The result is generalized and thesis is prepared.

R. C. Sharma, in his book - Art& Science of Railway Signaling- has explained that low potential electromagnetic relays are extensively used in operating and interlocking signaling functions such as points, signals etc. These relays can be categorized as shelf type / plug-in type, track relay / line relay, proved / non-proved type, DC Relays/AC relays. Pick up value is the current just sufficient to close all the front contacts of relays while Drop Away value is the current at which all the front contacts of a relay just open. Front contact is the contact, which is made with the arm contact on energization of relay. The contact, which touches the arm contact when the relay is de-energized, is known as Back Contact. The relays used in AC traction territory are AC Immunized so that they do not alter their position & get influenced by induced EMF. The various types of Q series relays used on Indian Railways are – QN1 Relay (line relay used for all controls and detections), QNA1 Relay (AC immunized relay used in railway electrified area), QTA2 Relay (AC immunized track relay), QSPA1 Relay (slow to pick up and slow to release, AC immunized relay used with track relays), and QECX Relay (current sensing relays to check status of lamps).

Kanakasabapathi K. S. & Ramasamy V., in their technical report - Techniques of manufacturing of Q series relays to BRS specifications - provided the details of producing Q series relays. In the manufacturing of armature first the piece is blanked to the outer dimensions, the recesses for the residual pin and return pin seating are formed and then two circular holes inside the recesses and the rectangular slot are pierced. After this the armature is bent roughly to the angle required and then broaching of the armature is done on the Vertical Broaching Machine. For making the heelpiece, the piece is blanked, two holes are pierced, bent, sheared and angle is set. Then the heelpieces are milled to perfect finish. For Contact springs, first the pieces are blanked and pierced, getting the contour of the springs along with the holes and slots. Then dimples are formed.
The relay coil is wound on the Coil Winding Machine. The bobbin mould is fixed with the core in the spindle, a starting lead wire using gas torch is prepared and lead is soldered. At the end of the winding the ending is prepared and the wire is secured with adhesive tape. The specification details on the coil are printed and the coil is stored for 48 hours at 80°C in the oven. After the coil cools down, it is inspected and tested for resistance.

M/s. Westinghouse rail systems have detailed the General Information on style Q-Relays. The metal-to-carbon relays have a phenolic thermoset base moulding on which various parts of relays are fitted and is covered by a transparent polycarbonate cover held in place by an internal support strap. The contacts are made of silver impregnated graphite and silver, which are fixed on contact springs made of phosphor bronze, and having specified shaped ends, which extend through the relay base such that they make electrical contact with the plug board connections when the relay is plugged in. The operating arms operate the moving contacts are through armature movement. The magnet assembly consists of a cylindrical core, heelpiece and armature. Armature return torque is provided by an adjustable helical spring, front contact spring pressure and gravity. A small phosphor bronze residual pin, riveted to the armature, prevents sticking on release. A bobbin is used for winding the coil and this coil is fixed around the iron core. The relays have contact ratings of 7Amperes continuous and 2Amperes switching, and are designed to function from -10°C to +70°C.

European Rail Research Institute has discussed the failure analysis of metal-to-carbon relays in their report - Performance of Signaling Relays: Metal-to-carbon Relays. The mechanism & philosophy of current conduction through silver-SIG contacts is such that when the contacts are closed, the conducting area may consist of three distinct areas – full conducting area, quasi-electric spots (areas covered with thin films through which current passes easily) and resistive spots (areas having thick non-conducting films). Due to small number of spots making electric contact, the current density increases. This in turn leads to rise in temperature at the contact site. This increase in temperature adversely affects the current density. To prevent this, the load will have to be increased. However, there would be more breakage in contacts and a more powerful coil shall be required. Due to
the superficial corrosion films, the actual contact area goes down and thus, the contact resistance increases. Thus, the reliability of contacts from HCR point of view is affected by two typical causes:

- A cause that does not exist at the beginning of useful life of the contact but develops with time, becoming crucial after a reliability time. For example, surface deformation due to material transfer & formation of high resistivity films.
- A cause of failure existing from the beginning, such as dust particles and level of impregnation.

Douglas A. Kimber, Eric Bauer and Xuemei Zhang, in their book - Practical System Reliability - have provided insight into the reliability of systems. Reliability of mechanisms is expressed as the likelihood that they shall work without failure in a particular situation for a unit period. This indicates that their failure chances changes with time as per a likelihood operand. The various dependability indices are MTTF, MTBF, availability and many others. Mean time To Failure (MTTF) is used for non-repairable systems and is the time of the first failure of the system to occur. It is the minimum time that any device shall work effectively, without failures. For reliability probability \( R(t) \) the MTTF of a system is,

\[
MTTF = \int_0^\infty R(t)\,dt
\]  

(1.1)

V. L. Popov, in his book - Contact Mechanics and Friction - has discussed the contact theory between engineering surfaces. The Hertzian contact theory states that when a rigid sphere of radius \( R \) comes in contact with elastic half-space, the displacement of the points on the surface in the contact area is given by,

\[
u_z = d - \frac{r^2}{2R}
\]  

(1.2)

where, \( u_z \) is the vertical displacement and \( d \) is the indentation depth.

For a Hertzian pressure distribution, the quadratic distribution of the vertical displacement \( u_z \) is given by,
\[ u_c = \frac{1}{E^*} \frac{\pi \rho_0}{4a} \left( 2a^2 - r^2 \right) \] (1.3)

where, \( E^* \) is the elastic moduli of the surfaces, contact radius is \( a \) and \( \rho_0 \) is stress on contact area. Therefore, it follows that,

\[ \frac{1}{E^*} \frac{\pi \rho_0}{4a} \left( 2a^2 - r^2 \right) = d - \frac{r^2}{2R} \] (1.4)

Hence, solving for \( a \) and \( d \),

\[ a = \frac{\pi \rho_0 R}{2E^*} \quad \text{and} \quad d = \frac{\pi a \rho_0}{2E^*} \] (1.5)

Thus, the contact radius is given by,

\[ a = \sqrt{Rd} \] (1.6)

and the displacement is related to Force as,

\[ F = \frac{4}{3} E^* R^{1/2} d^{3/2} \] (1.7)

From the above, the pressure at the middle of matingzone as well as contact radius can be calculated as a function of the normal force as,

\[ \rho_0 = \left( \frac{6FE^*^{3/2}}{\pi R^2} \right)^{1/3} \quad \text{and} \quad a = \left( \frac{3FR}{4E^*} \right)^{1/3} \] (1.8)

M/s. Advanced Probing Systems Incorporation, in their technical paper - Fundamentals of contact resistance (part-I) - have discussed the current conduction through closed contacts. The contact theory states that no metal surface, howsoever polished, is completely flat at microscopic level. There are always some hills or asperities on the surfaces. The magnitude of asperities depends on the level of polishing and the intrinsic properties of metals. Besides, due to exposure to atmosphere, there are chemical or impurities films on the surface of metals. The films may be either thin films, caused by impurities, or thick films such as oxides / sulphides. Thus, when two metal contacts meet under normal loading, the current has to pass through the asperities and films. If viewed on a microscopic level, the actual conducting area consists of three distinct regions. First is the full conducting region, where both the metals are in contact, and where the
asperities have ruptured the films, and hence, the current passes through without any restriction. Then there are the semi-conducting regions, where the current has to pass through thin films, and thus experiences the film resistance. And finally, there are some non-conducting regions, covered totally with insulating films, where the current cannot pass. Thus, the real conducting area of contacts is very small in comparison to the apparent contact area. The regions where the current is able to pass are termed as α-spots. The conduction of current through metal contacts is only through these α-spots and hence, to increase the conductivity, the number of α-spots should be increased.

L. Kogut & K. Komvopoulos, in their research paper - Electrical contact resistance theory for conductive rough surfaces separated by a thin insulating film - have discussed the behavior of resistance under the effect of superficial films. When two metallic surfaces, covered by thin insulating film, are in contact, the current flow is through tunnel effect. Only those electrons, which possess power enough for overcoming the film barricade, shall pass over from one surface to another, thus, introducing a tunnel or film resistance. In such scenario, the surface roughness of the metallic surfaces becomes quite significant and the film resistance dominates the contact resistance. Many asperity micro-contacts are made when two surface mate. The presence of insulating film makes the contact resistance changes due to load more prominent, as the film resistance is mostly non-ohmic. Similarly, increment in applied current makes the variation less pronounced. The film resistance is exactly relational to surface unevenness, the dielectric constant of materials, height of energy barrier and width of the film; and inversely proportional to the current flow through the surfaces, load acting on them and the magnitude of superficial peaks / valleys.

William Everett Wilson, in his thesis - Surface separation and contact resistance considering sinusoidal elastic-plastic multi-scale rough surface contact - has proposed various models for explaining the contact mechanics, based on the Hertzian contact theory. The statistical contact model, proposed by Greenwood & Williamson, considers the interaction between a perfectly flat elastic plane and another one having spherically shaped asperities of similar radii, presuming that each asperity behaves independently of its neighbors. However, this model is not practical as it assumes the surface to have a single
radius of curvature and neglects the effects of different scales of features on a surface. To address these issues, Majumdar & Bhushan developed a fractal model, based on the fact that a surface is multi-scalar in nature and each surface scale is related by the fractal equations (represented by Weierstrass-Mandelbrot function) and the radius of curvature of asperities is dependent upon the size of contact. However, this model has limitations that all the scales can never be perfectly described by a single fractal equation. To remove this anomaly, various multi-scale models, based on the Archard’s protuberance upon protuberance scheme (a sphere of a certain radius coated with hemispheres of another radius which are all then coated with smaller hemispheres of a third radius) have been proposed. The Jackson & Streator model, which uses a series of stacked three-dimensional sinusoidal waves for the multiple scales is based on the assumptions that the smaller asperities are stacked on top of larger asperities, load is distributed equally over all asperities, and a smaller asperity level is not capable of extending the contact area beyond that capable of the larger scale.

Roland E. Ott, in his thesis - Thermal and electrical resistance of metal contacts - states that Electrical Contact Resistance is the combined effect of many factors, such as electrical, mechanical & chemical. All metallic surfaces have hills & valleys and impurities films on them, and hence, a current passing through a pair of mating contacts, shall pass through these constrictions and films and experience constriction and film resistances. However, the current can take another possible path – through the gaps between the contact surfaces. This current would consist of those electrons that can overcome the distance between the contact surfaces. J. Frenkel has shown that for a gap of 10Å and a mean free path of electrons equal to 100Å, the resistance of the contact per unit surface area is approximately 10⁻²Ohms. Any further increase in the gap prohibits all possibility of conduction. Since it is seen that practically, the gaps normally occurring between metal contacts are of the order of 100Å or more, there is absolutely no possibility of any thermionic current to pass, at room temperatures. This establishes that for all practical purposes, at ambient temperatures, all the measurable current between contacts flows through the hills and the films on these hills on the contact surfaces.
In his research paper - *Effect of connection design on the contact resistance of high power overlapping bolted joints* - Milenko Braunovic opines that on a microscopic level, all surfaces have many hills known as asperities and some dirt/chemical films on them. Hence, the actual conduction of current is through the localized metallic contacts, which are formed by the surface asperities penetrating the superficial films. Increasing the load shall enhance the contact points’ quantity and extent. The obvious area $A_a$ is much larger than $A_r$, the actual mating area. However, the electrical conduction is not only from one contact point; there are many such locations inside the mating area, as demonstrated by Williamson. The resistance generated by obstruction to current is termed as constriction resistance. It is based on hardness and electrical resistivity of contact metals, as given in the formula:

$$ R_c = \frac{\rho_1 + \rho_2}{4a} $$

where $\rho_1$ & $\rho_2$ are resistivity of the materials touching each other while $a$ is the contact zone radius.

Quanfeng Song, Wenqi Zhang & Niels Bay, in their research paper - *An experimental study determines the electrical contact resistance in resistance welding* - have discussed the relationship of contact resistance and temperature & pressure. The contact resistance comprises of constriction & layer resistance. When two materials meet, under normal pressure, the CR is inversely proportional to pressure. Increasing the load enlarges the real contact area and helps in breaking the superficial films. This not only decreases the constriction resistance, but also lowers the film resistance, thus reducing the overall contact resistance. When the pressure is extremely high, the surfaces mate completely (as the layers are fully broken) and the constriction & layer resistance become negligible. As temperature rises, the contact area becomes sizeable and hence, the constriction resistance decreases. On the other hand, at elevated temperatures, metallic bulk resistance increases and hence, the constriction resistance enhances. Besides, films are readily broken when metals are really hot, which decreases the layer resistance; contrary, the film growth is exponential at increased temperatures, thus enhancing the resistance. The final picture is a combined effect of all the above phenomena.
Swati Desai, in her book - Quantitative Techniques - has discussed the fundamentals of hypothesis testing. The hypothesis testing involves determining it’s accuracy and can be done either through the Sampling Theory approach (analysis of available sampling data) or through Bayesian approach (subjective probability estimates are also considered). Null Hypothesis states that there is no difference between the sample and population, and is denoted by $H_0$, while Alternative Hypothesis indicates that difference exists between the parameter & statistic and is denoted by $H_a$. The $H_0$ & $H_a$ are complementary to each other. Two kinds of errors can be committed while testing the hypothesis – denunciation of correct assumption (Type I Error) and approval of incorrect supposition (Type II error).

Ian M. Watt, in his book - The principles and practice of electron Microscopy - has discussed the fundamentals of Scanning Electron Microscopy. The basic principle of Scanning Electron Microscopy involves scanning a precise narrow electron beam in a raster pattern over the sample surface. Due to this, secondary electrons having very low energy are ejected while the high-energy electrons are reflected. This resultant electron stream is used for creating the surface topography. The equipment comprises of – one electric-optical column, which is a conventional triode gun with a tungsten filament and delivers a total of electron current up to 250 microampere at energies 1 to 30KeV. A double condenser lens system forms a more than hundred times de-magnified electron source close after the second lens. The final condenser lens/objective lens projects the diminished image of the electron crossover as a spot focused on the specimen surface. The electron collection system consists of a Perspex cylindrical light guide with a hemispherical end coated with a phosphor and a thin aluminum film, maintained at high positive potential. The specimen chamber & stage accommodates the specimen holder while the vacuum pumps evacuate the electron-optical column and the specimen chamber to a high vacuum of the order of 1 x $10^{-4}$ mbar.

Joseph Goldstein et al, in their book - Scanning Electron Microscopy and X-ray Microanalysis - have deliberated over the fundamentals of Energy Dispersive X-ray spectroscopy. The determination of identification and the percentage volume of individual chemical substances present on the surface of a sample is done through the Energy Dispersive X-ray spectroscopy. In this, a concentrated electron ray is blasted on the
sample surface due to which certain X-rays are emitted. These X-rays are distinctive of the chemical element from which they are released. Measurement of the energy of these rays provides a comprehensive picture of the elements and their quantum, present on the surface. The primary components of an EDX setup consist of an X-ray generation mechanism and a detector system for detecting these rays. The signals thus generated are forwarded to a processor and analyzer for mapping the atomic character of the specimen plane.

M/s. Physical Electronics, in their technical literature on TOF-SIMS, have enumerated its essentials. The technique for ascertaining the contour of thin films and scattering of chemicals in it is the Time Of Flight Secondary Ion Mass Spectrometry. In this, the elements are recognized by the time their electrons take to cover the distance from the specimen surface to the detector. Since the travel time is directly proportional to the mass/charge of an ion, hence, by estimating the time, the chemical properties of the unknown element can be determined. The TOF-SIMS apparatus comprises of a high-energy electron beam, the flight path (to direct the particle beam) and the mass detection and analyzing system.

Phillip J. Ross, in his book, Taguchi Techniques for Quality Engineering, has explained the essentials of Design Of Experiments. The systematic exercise of determining the factors (and their parametric values) responsible for affecting the process output is termed as Design Of Experiments. Data collected from a designed experiment is analyzed statistically to determine the effect of the individual or combination of variables and is used in evaluating the parameters/properties of physical objects, chemical formulations, structures, components and materials. The first step in DOE of any product or process is to develop the statement of the problem giving the details of the problem. The next step is to define a clear and precise objective for experiment. After this, the quality parameters to be measured and how and by whom the measurement shall be done is determined. Next, all the factors which effect the quality parameters are listed through Cause & Effect diagram, brainstorming and other tools. After this, the values and number of levels for the selected control factors are fixed. Then the appropriate orthogonal array for the experiment is selected from the Orthogonal Array design and selection tables. The
interaction effects are also enumerated. After this, the factors are assigned to various columns through the Orthogonal Array factor assignment tables, which also indicate the columns, which shall contain the interaction effects. Then the experiment is conducted as per the trial design and results are analyzed through ANOVA tables. The main and interaction effects plots are also studied, and the factors having the maximum effect on the product/process are determined. Finally, validation test is conducted with the modified values of the factors, which confirms the trial results.

M. A. Stephens, in his research paper - EDF Statistics for Goodness of Fit & Some Comparisons - has debated on the various tests of goodness of fit. The attributes of data are checked with this test, with the help of a statistical model, which compares the variation in the experimental and anticipated data. These tests are engaged for testing for normality (Anderson - Darling test) and to investigate that the sources of two diverse samples are similar. The Anderson–Darling test is a test for checking whether the data follows a normal/log-normal/Weibull/exponential or logistic distribution. It is a statistical test that looks for the absence of normality and indicates this with a small P-value. The Anderson - Darling is a one-sided test and makes use of the specific distribution in calculating critical values. The test requires adequate amount of data to prove the goodness of fit, is very sensitive and has great accuracy.

Forrest W. Breyfogle III, in his book - Six sigma, Smarter solutions using statistical methods - has described the various distributions for reliability. Relay is a non-repairable device and hence special statistical methods are used for testing their reliability. The various methods depend on the type of distribution the failure mechanism & data follow, such as normal, log-normal, and Weibull. The normal distribution is used when the failure time follows normal distribution, and in this, the mean and standard deviation of data is considered. The lognormal distribution is applicable when the logarithms of the failure time data are normally distributed. The two-parameter Weibull distribution explains all the three cases of failure time distribution – decreasing, increasing and constant failure rates of devices, and is expressed as,

\[ F(t) = 1 - \exp\left[-\left(\frac{t}{k}\right)^b\right] \]  

(1.10)
withformfactor $b$ & gauge parameter $k$. The value $b$ is dependent on the shape of the probability density function curve for various failure rates. Numerically, it is less than 1 for early life decreasing failure rate, equal to 1 for constant failure rate and is greater than 1 for increasing failure rate. The factor $k$ is a token of characteristic life of the device.