FUNDAMENTALS OF SEMICONDUCTOR MANUFACTURING AND PROCESS CONTROL

CONTENTS

PREFACE

In simple terms, manufacturing can be defined as the process by which raw materials are converted into finished products. The purpose of this book is to examine in detail the methodology by which electronic materials and supplies are converted into finished integrated circuits and electronic products in a high-volume manufacturing environment. This subject of this book will be issues relevant to the industrial-level manufacture of microelectronic device and circuits, including (but not limited to) fabrication sequences, process control, experimental design, process modeling, yield modeling, and CIM/CAM systems. The book will include theoretical and practical descriptions of basic manufacturing concepts, as well as some case studies, sample problems, and suggested exercises.

The book is intended for graduate students and can be used conveniently in a semester-length course on semiconductor manufacturing. Such a course may or may not be accompanied by a corequisite laboratory. The text can also serve as a reference for practicing engineers and scientists in the semiconductor industry.

Chapter 1 of the book places the manufacture of integrated circuits into its historical context, as well as provides an overview of modern semiconductor manufacturing. In the Chapter 2, we provide a broad overview of the manufacturing technology and processes flows used to produce a variety of semiconductor products. Various process monitoring methods, including those that focus on product wafers and those that focus on the equipment used to produce those wafers, are discussed in Chapter 3. As a backdrop for subsequent discussion of statistical process control (SPC), Chapter 4 provides a review of statistical fundamentals. Ultimately, the key metric to be used to evaluate any manufacturing process is cost, and cost is directly impacted by yield. Yield modeling is therefore presented in Chapter 5. Chapter 6 then focuses on the use of SPC to analyze quality issues and improve yield. Statistical experimental design, which is presented in Chapter 7, is a powerful approach for systematically varying controllable process conditions and determining their impact on output parameters which measure quality. Data derived from statistical experiments can then be used to construct process models that enable the analysis and prediction of manufacturing process behavior. Process modeling concepts are introduced in Chapter 8. Finally, several advanced process control topics, including run-by-run, supervisory control, and process and equipment diagnosis, are the subject of Chapters 9 and 10.

Each chapter begins with an introduction and a list of learning goals, and each concludes with a summary of important concepts. Solved examples are provided throughout, and suggested homework problems appear at the end of the chapter. A complete set of detailed solutions to all end-of-chapter problems has been prepared. This *Instructor's Manual* is available to all adopting faculty. The figures in the text are also available, in electronic format, from the publisher at the web site: *http://www.wiley.com/college/mayspanos*.

ACKNOWLEDGMENTS

G. S. May would like to acknowledge the support of the Steve W. Chaddick School Chair in Electrical and Computer Engineering at the Georgia Institute of Technology, which provided the environment that enabled the completion of this book. C. J. Spanos would like to acknowledge the contributions of the Berkeley students who, over the years, helped shape the material presented in this book.

1

INTRODUCTION TO SEMICONDUCTOR MANUFACTURING

OBJECTIVES

- Place the manufacturing of integrated circuits in a historical context.
- Provide an overview of modern semiconductor manufacturing.
- Discuss manufacturing goals and objectives.
- Describe manufacturing systems at a high level as a prelude to the remainder of the text.

INTRODUCTION

This book is concerned with the manufacturing of devices, circuits, and electronic products based on semiconductors. In simple terms, *manufacturing* can be defined as the process by which raw materials are converted into finished products. As illustrated in Figure 1.1, a manufacturing operation can be viewed graphically as a system with raw materials and supplies serving as its inputs and finished commercial products serving as outputs. In semiconductor manufacturing, input materials include semiconductor materials, dopants, metals, and insulators. The corresponding outputs include integrated circuits (ICs), IC packages, printed circuit boards, and ultimately, various commercial electronic systems and products (such as computers, cellular phones, and digital cameras). The types of processes that arise in semiconductor manufacturing include crystal

Fundamentals of Semiconductor Manufacturing and Process Control,
By Gary S. May and Costas J. Spanos

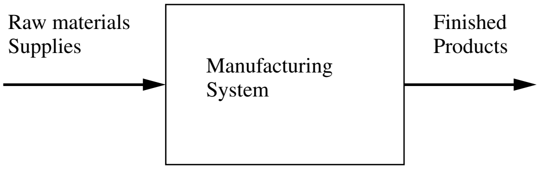


Figure 1.1. Block diagram representation of a manufacturing system.

growth, oxidation, photolithography, etching, diffusion, ion implantation, planarization, and deposition processes.

Viewed from a systems-level perspective, semiconductor manufacturing intersects with nearly all other IC process technologies, including design, fabrication, integration, assembly, and reliability. The end result is an electronic system that meets all specified performance, quality, cost, reliability, and environmental requirements. In this chapter, we provide an overview of semiconductor manufacturing, which touches on each of these intersections.

1.1. HISTORICAL EVOLUTION

Semiconductor devices constitute the foundation of the electronics industry, which is currently (as of 2005) the largest industry in the world, with global sales over one trillion dollars since 1998. Figure 1.2 shows the sales volume of the semiconductor device-based electronics industry since 1980 and projects sales to the year 2010. Also shown are the gross world product (GWP) and the sales volumes

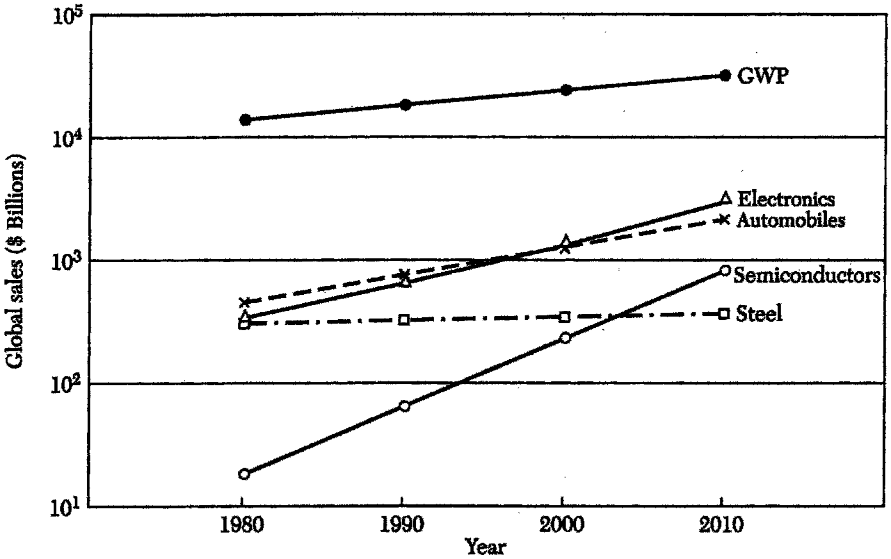


Figure 1.2. Gross world product (GWP) and sales volumes of various industries from 1980 to 2000 and projected to 2010 [1].

of the automobile, steel, and semiconductor industries [1]. If current trends continue, the sales volume of the electronic industry will reach three trillion dollars and will constitute about 10% of GWP by 2010. The semiconductor industry, a subset of the electronics industry, will grow at an even higher rate to surpass the steel industry in the early twenty-first century and to constitute 25% of the electronic industry in 2010.

The multi-trillion-dollar electronics industry is fundamentally dependent on the manufacture of semiconductor integrated circuits (ICs). The solid-state computing, telecommunications, aerospace, automotive, and consumer electronics industries all rely heavily on these devices. A brief historical review of manufacturing and quality control, semiconductor processing, and their convergence in IC manufacturing, is therefore warranted.

1.1.1. Manufacturing and Quality Control

The historical evolution of manufacturing, summarized in Table 1.1, closely parallels the industrialization of Western society, beginning in the nineteenth century. It could be argued that the key early development in manufacturing was the concept of *interchangeable parts*. Eli Whitney is credited with pioneering this concept, which he used for mass assembly of the cotton gin in the early 1800s [2]. In the late 1830s, a Connecticut manufacturer began producing cheap windup clocks by stamping out many of the parts out of sheets of brass. Similarly, in the early 1850s, American rifle manufacturers thoroughly impressed a British delegation by a display in which 10 muskets made in 10 different preceding years were disassembled, had their parts mixed up in a box, and subsequently reassembled quickly and easily. In England at that time, it would have taken a skilled craftsman the better part of a day to assemble a single unit.

The use of interchangeable parts eliminated the labor involved in matching individual parts in the assembly process, resulting in a tremendous time savings and increase in productivity. The adoption of this method required new forms of technology capable of much finer tolerances in production and measurement methods than those required by hand labor. Examples included the

Table 1.1. Major milestones in manufacturing history.

Year(s)	Event
1800–1850	Concept of interchangeable parts introduced
1850–1860	Advances in measurement and machining operations
1875	Taylor introduces scientific management principles
1900–1930	Assembly line techniques actualized by Ford
1924	Control chart introduced by Shewhart
Late 1920s	Dodge and Romig develop acceptance sampling
1950s	Computer numeric control and designed experiments introduced
1970s	Growth in the adoption of statistical experimental design
1980	Pervasive use of statistical methods in many industries

vernier caliper, which allowed workers to measure machine tolerances on small scales, and wire gauges, which were necessary in the production of clock springs. One basic machine operation perfected around this time was mechanical drilling using devices such as the turret lathe, which became available after 1850. Such devices allowed a number of tedious operations (hand finishing of metal, grinding, polishing, stamping, etc.) to be performed by a single piece of equipment using a bank of tool attachments. By 1860, a good number of the basic steps involved in shaping materials into finished products had been adapted to machine functions.

Frederick Taylor added rigor to the manufacturing research and practice by introducing the principles of *scientific management* into mass production industries around 1875 [3]. Taylor suggested dividing work into tasks so that products could be manufactured and assembled more readily, leading to substantial productivity improvements. He also developed the concept of standardized production and assembly methods, which resulted in improved quality of manufactured goods. Along with the standardization of methods came similar standardization in work operations, such as standard times to accomplish certain tasks, or a specified number of units that must be produced in a given work period.

Interchangeable parts also paved the way for the next major contribution to manufacturing: the *assembly line*. Industrial engineers had long noted how much labor is spent in transferring materials between various production steps, compared with the time spent in actually performing the steps. Henry Ford is credited for devising the assembly line in his quest to optimize the means for producing automobiles in the early twentieth century. However, the concept of the assembly line had actually been devised at least a century earlier in the flour mill industry by Oliver Evans in 1784 [2]. Nevertheless, it was not until the concept of interchangeable parts was combined with technology innovations in machining and measurement that assembly line methods were truly actualized in their ultimate form. After Ford, the assembly line gradually replaced more labor-intensive forms of production, such as custom projects or batch processing.

No matter what industry, no one working in manufacturing today can overemphasize the influence of the computer, which catalyzed the next major paradigm shift manufacturing technology. The use of the computer was the impetus for the concept of *computer numeric control* (CNC), introduced in the 1950s [4]. Numeric control was actually developed much earlier. The player piano is a good example of this technique. This instrument utilizes a roll of paper with holes punched in it to determine whether a particular note is played. The numeric control concept was enhanced considerably by the invention of the computer in 1943. The first CNC device was a spindle milling machine developed by John Parsons of MIT in 1952. CNC was further enhanced by the use of microprocessors for control operations, beginning around 1976. This made CNC devices sufficiently versatile that an existing tooling could be quickly reconfigured for different processes. This idea moved into semiconductor manufacturing more than a decade later when the machine communication standards made it possible to have factorywide production control.

The inherent accuracy and repeatability engendered by the use of the computer eventually enabled the concept of *statistical process control* to gain a foothold in manufacturing. However, the application of statistical methods actually had a long prior history. In 1924, Walter Shewhart of Bell Laboratories introduced the control chart. This is considered by many as the formal beginning of statistical quality control. In the late 1920s, Harold Dodge and Harry Romig, both also of Bell Labs, developed statistically based acceptance sampling as an alternative to 100% inspection. By the 1950s, rudimentary computers were available, and *designed experiments* for product and process improvement were first introduced in the United States. The initial applications for these techniques were in the chemical industry. The spread of these methods to other industries was relatively slow until the late 1970s, when their further adoption was spurred by economic competition between Western companies and the Japanese, who had been systematically applying designed experiments since the 1960s. Since 1980, there has been profound and widespread growth in the use of statistical methods worldwide, and particularly in the United States.

1.1.2. Semiconductor Processes

Many important semiconductor technologies were derived from processes invented centuries ago. Some of the key technologies are listed in Table 1.2 in chronological order. For the most part, these techniques were developed independently from the evolution of manufacturing science and technology. For example, the growth of metallic crystals in a furnace was pioneered by Africans living on the

Table 1.2. Major milestones in semiconductor processing history.

Year	Event
1798	Lithography process invented
1855	Fick proposes basic diffusion theory
1918	Czochralski crystal growth technique invented
1925	Bridgman crystal growth technique invented
1952	Diffusion used by Pfann to alter conductivity of silicon
1957	Photoresist introduced by Andrus; oxide masking developed by Frosch and Derrick; epitaxial growth developed by Sheftal et al.
1958	Ion implantation proposed by Shockley
1959	Kilby and Noyce invent the IC
1963	CMOS concept proposed by Wanlass and Sah
1967	DRAM invented by Dennard
1969	Self-aligned polysilicon gate process proposed by Kerwin et al.; MOCVD developed by Manasevit and Simpson
1971	Dry etching developed by Irving et al.; MBE developed by Cho; first microprocessor fabricated by Intel
1982	Trench isolation technology introduced by Rung et al.
1989	CMP developed by Davari et al.
1993	Copper interconnect introduced to replace aluminum by Paraszczak et al.

western shores of Lake Victoria more than 2000 years ago [5]. This process was used to produce carbon steel in preheated forced-draft furnaces. Another example is the lithography process, which was invented in 1798. In this first process, the pattern, or image, was transferred from a stone plate (*lithos*) [6]. The diffusion of impurity atoms in semiconductors is also important for device processing. Basic diffusion theory was described by Fick in 1855 [7].

In 1918, Czochralski developed a liquid–solid monocomponent growth technique used to grow most of the crystals from which silicon wafers are produced [8]. Another growth technique was developed by Bridgman in 1925 [9]. The Bridgman technique has been used extensively for the growth of gallium arsenide and related compound semiconductors. The idea of using diffusion techniques to alter the conductivity in silicon was disclosed in a patent by Pfann in 1952 [10]. In 1957, the ancient lithography process was applied to semiconductor device fabrication by Andrus [11], who first used photoresist for pattern transfer. Oxide masking of impurities was developed by Frosch and Derrick in 1957 [12]. In the same year, the epitaxial growth process based on chemical vapor deposition was developed by Sheftal et al. [13]. In 1958, Shockley proposed the method of using ion implantation to precisely control the doping of semiconductors [14].

In 1959, the first rudimentary integrated circuit was fabricated from germanium by Kilby [15]. Also in 1959, Noyce proposed the monolithic IC by fabricating all devices in a single semiconductor substrate and connecting the devices by aluminum metallization [16]. As the complexity of the IC increased, the semiconductor industry moved from NMOS (n-channel MOSFET) to CMOS (complementary MOSFET) technology, which uses both NMOS and PMOS (p-channel MOSFET) processes to form the circuit elements. The CMOS concept was proposed by Wanlass and Sah in 1963 [17]. In 1967, the dynamic random access memory (DRAM) was invented by Dennard [18].

To improve device reliability and reduce parasitic capacitance, the self-aligned polysilicon gate process was proposed by Kerwin et al. in 1969 [19]. Also in 1969, the metallorganic chemical vapor deposition (MOCVD) method, an important epitaxial growth technique for compound semiconductors, was developed by Manasevit and Simpson [20]. As device dimensions continued to shrink, dry etching was developed by Irving et al. in 1971 to replace wet chemical etching for high-fidelity pattern transfer [21]. Another important technique developed in the same year by Cho was molecular-beam epitaxy (MBE) [22]. MBE has the advantage of near-perfect vertical control of composition and doping down to atomic dimensions. Also in 1971, the first monolithic microprocessor was fabricated by Hoff et al. at Intel [23]. Currently, microprocessors constitute the largest segment of the industry.

Since 1980, many new technologies have been developed to meet the requirements of continuously shrinking minimum feature lengths. Trench technology was introduced by Rung et al. in 1982 to isolate CMOS devices [24]. In 1989, the chemical–mechanical polishing (CMP) method was developed by Davari et al. for global planarization of the interlayer dielectrics [25]. Although aluminum has been used since the early 1960s as the primary IC interconnect material, copper interconnect was introduced in 1993 by Paraszczak et al. to replace aluminum for minimum feature lengths approaching 100 nm [26].

1.1.3. Integrated Circuit Manufacturing

By the beginning of the 1980s, there was deep and widening concern about the economic well-being of the United States. Oil embargoes during the previous decade had initiated two energy crises and caused rampant inflation. The U.S. electronics industry was no exception to the economic downturn, as Japanese companies such as Sony and Panasonic nearly cornered the consumer electronics market. The U.S. computer industry experienced similar difficulties, with Japanese semiconductor companies beginning to dominate the memory market and establish microprocessors as the next target.

Then, as now, the fabrication of ICs was extremely expensive. A typical state-of-the-art, high-volume manufacturing facility at that time cost over a million dollars (and now costs several billion dollars) [27]. Furthermore, unlike the manufacture of discrete parts such as appliances, where relatively little rework is required and a yield greater than 95% on salable product is often realized, the manufacture of integrated circuits faced unique obstacles. Semiconductor fabrication processes consisted of hundreds of sequential steps, with potential yield loss occurring at every step. Therefore, IC manufacturing processes could have yields as low as 20–80%.

Because of rising costs, the challenge before semiconductor manufacturers was to offset large capital investment with a greater amount of automation and technological innovation in the fabrication process. The objective was to use the latest developments in computer hardware and software technology to enhance manufacturing methods. In effect, this effort in *computer-integrated manufacturing of integrated circuits* (IC-CIM) was aimed at optimizing the cost-effectiveness of IC manufacturing as *computer-aided design* (CAD) had dramatically affected the economics of circuit design.

IC-CIM is designed to achieve several important objectives, including increasing chip fabrication yield, reducing product cycle time, maintaining consistent levels of product quality and performance, and improving the reliability of processing equipment. Table 1.3 summarizes the results of a 1986 study by Toshiba that analyzed the use of IC-CIM techniques in producing 256-kbyte DRAM memory circuits [28]. This study showed that CIM techniques improved the manufacturing process on each of the four productivity metrics investigated.

Table 1.3. Results of 1986 Toshiba study.

Productivity Metric	Without CIM	With CIM
Turnaround time	1.0	0.58
Integrated unit output	1.0	1.50
Average equipment uptime	1.0	1.32
Direct labor hours	1.0	0.75

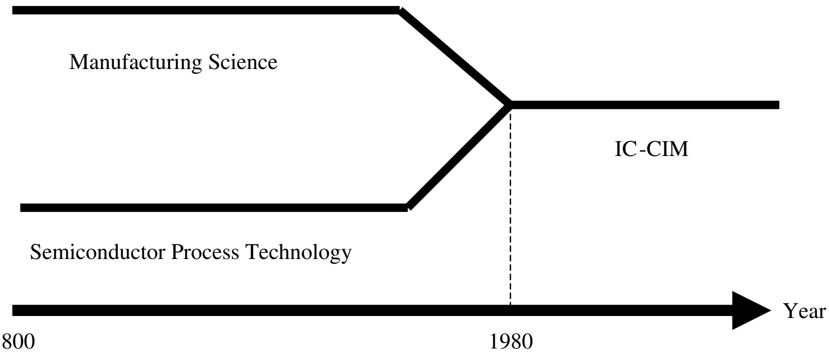


Figure 1.3. Timeline indicating convergence of manufacturing science and semiconductor processing into IC-CIM.

In addition to the demonstration of the effectiveness of IC-CIM techniques, economic concerns were so great in the early to mid-1980s that the Reagan Administration took the unprecedented step of partially funding a consortium of U.S. IC manufacturers—including IBM, Intel, Motorola, and Texas Instruments—to perform cooperative research and development on semiconductor manufacturing technologies. This consortium, SEMATECH, officially began operations in 1988 [29]. This sequence of events signaled the convergence of advances in manufacturing science and semiconductor process technology, and also heralded the origin of a more systematic and scientific approach to semiconductor manufacturing. This convergence is illustrated in Figure 1.3.

1.2. MODERN SEMICONDUCTOR MANUFACTURING

The modern semiconductor manufacturing process sequence is the most sophisticated and unforgiving volume production technology that has ever been practiced successfully. It consists of a complex series of hundreds of unit process steps that must be performed very nearly flawlessly.

This semiconductor manufacturing process can be defined at various levels of abstraction. For example, each process step has inputs, outputs, and specifications. Each step can also be modeled, either physically, empirically, or both. Much can be said about the technology of each step, and more depth in this area is provided in Chapter 2. At a higher level of abstraction, multiple process steps are linked together to form a process sequence. Between some of these links are inspection points, which merely produce information without changing the product. The flow and utilization of information occurs at another level of abstraction, which consists of various control loops. Finally, the organization of the process belongs to yet another level of abstraction, where the objective is to maximize the efficiency of product flow while reducing variability.

1.2.1. Unit Processes

It is difficult to discuss unit process steps outside the context of a process flow. Figures 1.4 and 1.5 show the major unit processes used in a simple process flow. These steps include oxidation, photolithography, etching, ion implantation, and metallization. We describe these steps briefly in this section via a simple sequence used to fabricate a *p*–*n* junction [1].

The development of a high-quality silicon dioxide (SiO_2) has helped to establish the dominance of silicon in the production of commercial ICs. Generally, SiO_2 functions as an insulator in a number of device structures or as a barrier to diffusion or implantation during device fabrication. In the fabrication of a *p*–*n* junction (Figure 1.4), the SiO_2 film is used to define the junction area. There are two SiO_2 growth methods, dry and wet oxidation, depending on whether

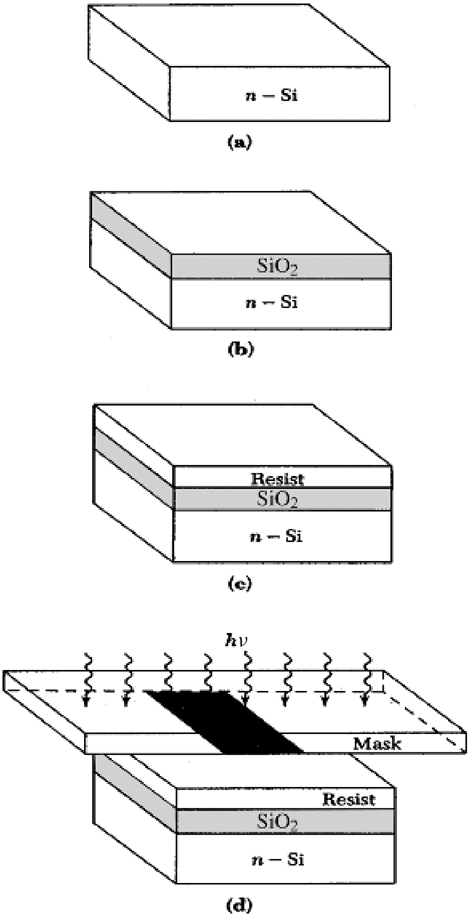


Figure 1.4. (a) A bare *n*-type silicon wafer; (b) an oxidized silicon wafer; (c) application of photoresist; (d) resist exposure through a mask [1].

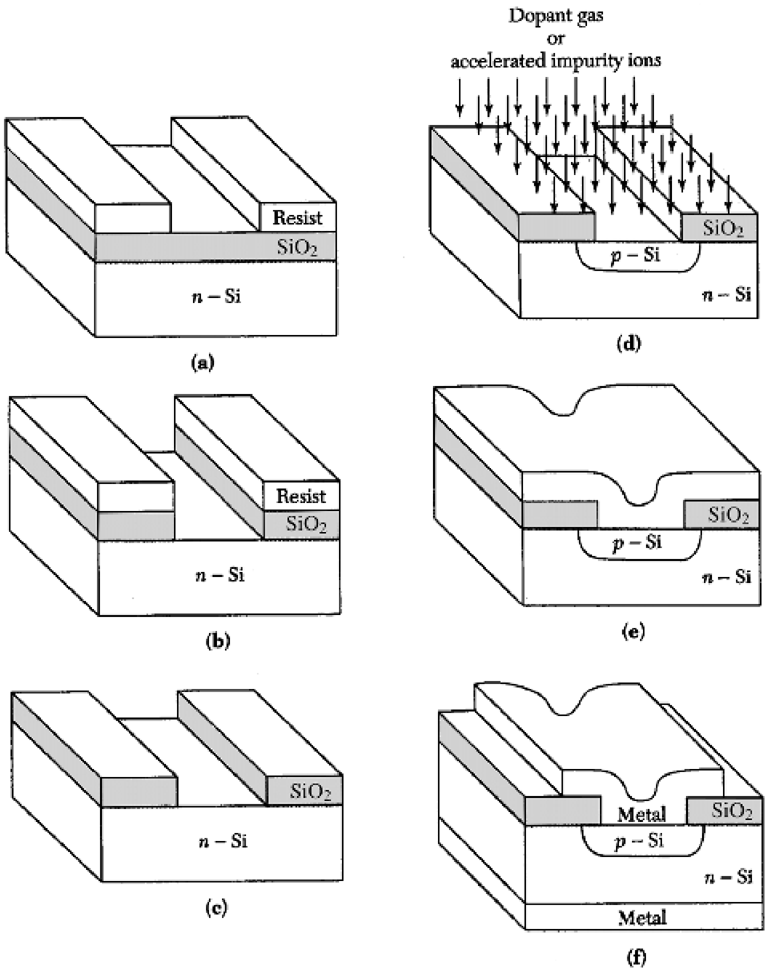


Figure 1.5. (a) Wafer after development; (b) wafer after SiO_2 removal; (c) result after photolithography; (d) formation of a *p–n* junction using diffusion or implantation; (e) wafer after metallization; (f) final product [1].

dry oxygen or water vapor is used. Dry oxidation is usually used to form thin oxides in a device structure because of its good $Si{-}SiO_2$ interface characteristics, whereas wet oxidation is used for thicker layers because of its higher growth rate. Figure 1.4a shows a section of a bare wafer ready for oxidation. After the oxidation process, a SiO_2 layer is formed all over the wafer surface. For simplicity, Figure 1.4b shows only the upper surface of an oxidized wafer.

Photolithography is used to define the geometry of the $p{-}n$ junction. After the formation of SiO_2, the wafer is coated with an ultraviolet (UV) light-sensitive material called *photoresist*, which is spun onto the wafer surface. Afterward (Figure 1.4c), the wafer is baked to drive the solvent out of the resist and to harden the resist for improved adhesion. Figure 1.4d shows the next step, which is to expose the wafer through a patterned mask using an UV light source. The exposed region of the photoresist-coated wafer undergoes a chemical reaction. The exposed area becomes polymerized and difficult to remove in an etchant. The polymerized region remains when the wafer is placed in a developer, whereas the unexposed region dissolves away. Figure 1.5a shows the wafer after the development. The wafer is again baked to enhance the adhesion and improve the resistance to the subsequent etching process. Then, an etch using hydrofluoric acid (HF) removes the unprotected SiO_2 surface (Figure 1.5b). Last, the resist is stripped away by a chemical solution or an oxygen plasma. Figure 1.5c shows the final result of a region without oxide (a window) after the lithography process. The wafer is now ready for forming the p–n junction by a diffusion or ion implantation process.

In diffusion, the wafer surface not protected by the oxide is exposed to a source with a high concentration of an opposite-type impurity. The impurity moves into the semiconductor crystal by solid-state diffusion. In ion implantation, the intended impurity is introduced into the wafer by accelerating the impurity ions to a high energy level and then implanting the ions in the semiconductor. The SiO_2 layer serves as barrier to impurity diffusion or ion implantation. After diffusion or implantation, the p–n junction is formed (Figure 1.5d).

After diffusion or ion implantation, a metallization process is used to form ohmic contacts and interconnections (Figure 1.5e). Metal films can be formed by physical vapor deposition or chemical vapor deposition. The photolithography process is again used to define the front contact, which is shown in Figure 1.5f. A similar metallization step is performed on the back contact without using a photolithography process.

1.2.2. Process Sequences

Semiconductor manufacturing consists of a series of sequential process steps like the one described in the previous section in which layers of materials are deposited on substrates, doped with impurities, and patterned using photolithography to produce ICs. Figure 1.6 illustrates the interrelationship between the major process steps used for IC fabrication. Polished wafers with a specific resistivity and orientation are used as the starting material. The film formation steps include thermally grown oxide films, as well as deposited polysilicon, dielectric, and metal films. Film formation is often followed by photolithography or impurity doping. Photolithography is generally followed by etching, which in turn is often followed by another impurity doping or film formation. The final IC is made by sequentially transferring the patterns from each mask, level by level, onto the surface of the semiconductor wafer.

After processing, each wafer contains hundreds of identical rectangular chips (or dies), typically between 1 and 20 mm on each side, as shown in Figure 1.7a. The chips are separated by sawing or laser cutting; Figure 1.7b shows a separated chip. Schematic top views of a single MOSFET and a single bipolar transistor

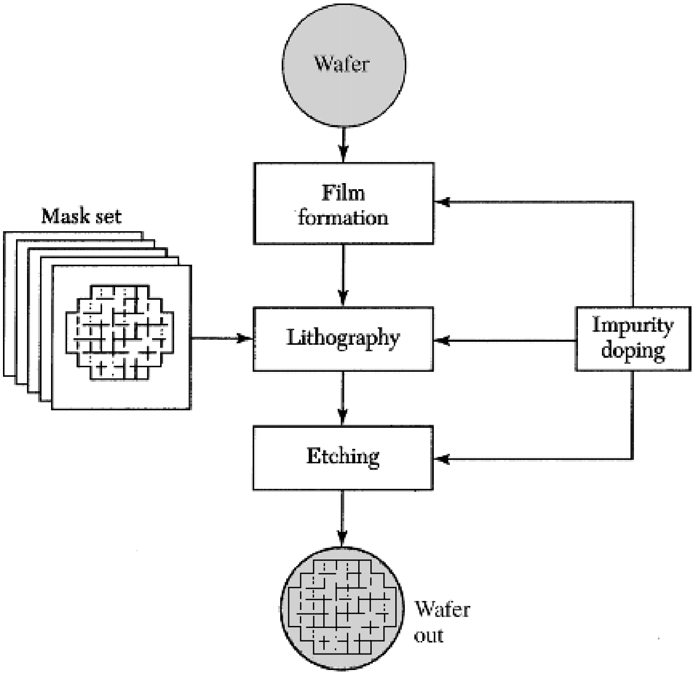


Figure 1.6. Flow diagram for generic IC process sequence [1].

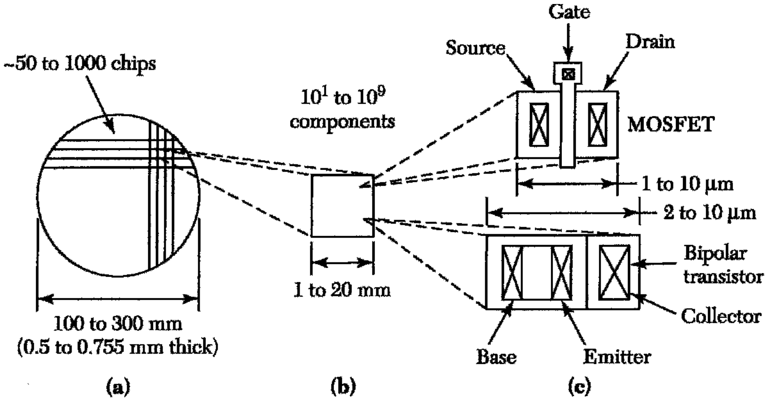


Figure 1.7. (a) Semiconductor wafer; (b) IC chip; (c) MOSFET and bipolar transistor [1].

are shown in Figure 1.7c. Inserted into this process sequence are various points at which key measurements are performed to ensure product quality.

1.2.3. Information Flow

The vast majority of quantitative evaluation of semiconductor manufacturing processes is accomplished via IC-CIM systems. The interdependent issues of ensuring high yield, high quality, and low cycle time are addressed by several critical capabilities in a state-of-the-art IC-CIM system: work-in-process (WIP) monitoring, equipment communication, data acquisition and storage, process/equipment modeling, and process control, to name only a few. The emphasis of each of these activities is to increase throughput and prevent potential misprocessing, but each presents significant engineering challenges in their effective implementation and deployment.

A block diagram of a typical modern IC-CIM system is shown in Figure 1.8. This diagram outlines many of the key features required for efficient information flow in manufacturing operations [28]. The lower level of this two-level architecture includes embedded controllers that provide real-time control and analysis of fabrication equipment. These controllers consist of personal computers and the associated control software dedicated to each individual piece of equipment. The second level of this IC-CIM architecture consists of a distributed local-area network of computer workstations and file servers linked by a common distributed database.

Equipment communication with host computers is facilitated by an electronics manufacturing standard called the *generic equipment model* (GEM). The GEM standard is used in both semiconductor manufacturing and printed circuit board assembly. This standard is based on the *semiconductor equipment communications standard* (SECS) protocol. SECS is a standard for communication between intelligent equipment and a host. The SECS standard has two components that define the communications protocol (SECS-I) and the messages exchanged (SECS-II), respectively. SECS-I specifies point-to-point communications over a high-speed messaging service interface. GEM is a standard set of SECS capabilities that can be selected by users as needed to coordinate equipment control in an automated factory. The GEM standard defines semiconductor

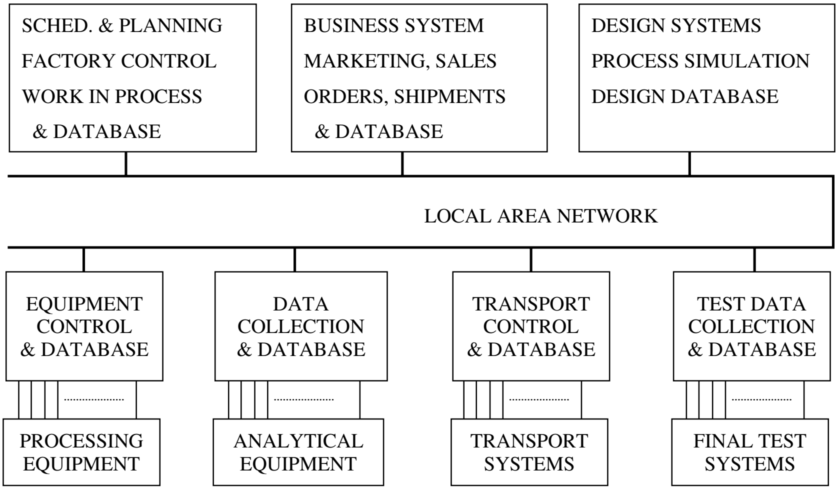


Figure 1.8. Two-level IC-CIM architecture [28].

equipment behavior as viewed through a communication link in terms of SECS-II messages communicated over that link. The GEM standard impacts equipment control and equipment–host communication and enables equipment to be integrated quickly and efficiently with a host computer [30].

The flow of information in this type of IC-CIM architecture enables equipment and process control at several levels. The highest level can be thought of as *supervisory control*, where the progression of a substrate is tracked from process to process. At this level, adjustments can be made to subsequent process steps to account for variation in previous procedures. The next lower level of control occurs on a *run-by-run* basis. For a single process, adjustments are made after each run to account for shifts and drifts that occur from wafer to wafer. Occasional shifts may occur when a new operator takes over or preventive maintenance is performed. A process may also experience drift due to equipment aging. *Real-time control* is at the lowest level of the hierarchy. In this case, adjustments are made to a process during a run to account for in situ disturbances. This hierarchy is diagramed in Figure 1.9.

1.2.4. Process Organization

As mentioned previously, the overall objective of process organization is to maximize the efficiency of product flow while minimizing variability and yield loss. Modern semiconductor factories [known as "fabs" (Fabrication Facilities)] are typically organized into *workcells*. In this approach, all the necessary equipment for completing a given process step is placed in the same room (see Figure 1.10). The workcell layout optimizes product flow, resulting in a minimal average distance traveled by semiconductor wafers as they migrate through the fabrication

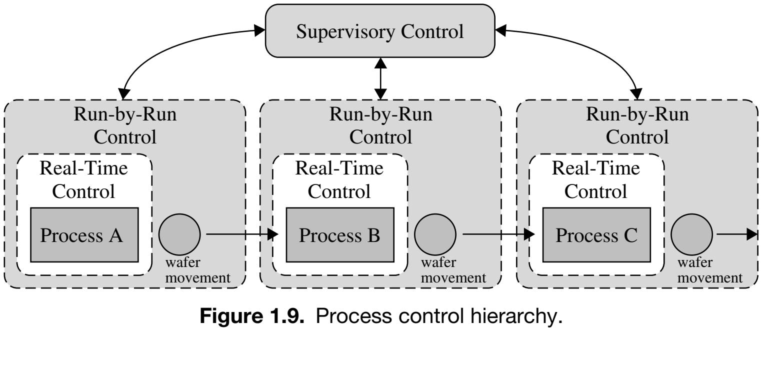


Figure 1.9. Process control hierarchy.

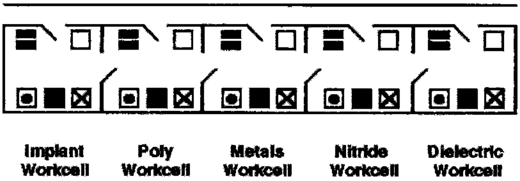


Figure 1.10. Workcell layout in a modern IC fabrication facility.

facility. This reduced distance translates into fewer chances for wafer mishandling and potential loss of product. Furthermore, the trend in equipment development since the mid-1990s has been toward single-wafer processing systems that enable enhanced reproducibility.

This modern IC factory also draws on powerful computing concepts, resulting in a highly flexible manufacturing system. In addition to the actual processing equipment, the factory consists of advanced in situ, postprocess, and end-of-the-line metrology and instrumentation necessary to quality control, equipment maintenance and diagnosis, rapid failure recovery, and inventory management. The physical factory is also augmented by simulation tools that allow various scenarios to be evaluated in a virtual manufacturing environment.

1.3. GOALS OF MANUFACTURING

From a systems-level perspective, semiconductor manufacturing intersects with design, fabrication, integration, assembly, and reliability. The fundamental goals of manufacturing are to tie all of these technologies together to achieve finished products with

- Low cost
- High quality
- High reliability

Cost is most directly impacted by yield and throughput. *Yield* is the proportion of products that meet the required performance specifications. Yield is inversely proportional to cost; that is, the higher the yield, the lower the cost. *Throughput* refers to the number of products processed per unit time. High throughput also leads to lower cost. The *quality* goal is virtually self-explanatory. It is obviously desirable to produce high-quality ICs that can be efficiently and repeatably mass-produced with a high degree of uniformity. Quality is derived from a stable and well-controlled manufacturing process. The *reliability* of electronic products is also impacted by the manufacturing process. High reliability results from the minimization of manufacturing faults. If each of the abovementioned goals is fulfilled, the end result is an IC that meets all specified performance, quality, cost, and reliability requirements.

1.3.1. Cost

Understanding the economics of IC manufacturing is important not only to the manufacturer but also to buyers and designers. A general rule of thumb is that IC fabrication, testing, and packaging each contribute about one-third of the total product cost. A variety of factors contribute to overall product costs, including the following [31]:

- Wafer processing cost

- Wafer processing yield
- Die size
- Wafer probe cost
- Probe yield
- Number of good dies
- Package cost
- Assembly yield
- Final test cost
- Final test yield

Wafer processing cost depends on wafer size, raw-wafer cost, direct labor cost, facility cost, and direct factory overhead (i.e., indirect labor costs, utilities, and maintenance). Direct costs (i.e., raw-wafer cost and direct labor) typically account for 10–15% of the wafer processing cost, with the remaining indirect costs accounted for by the equipment and facility depreciation, engineering support, facility operating costs, production control, and direct factory overhead. Increasingly, equipment costs contribute the lion's share, accounting for over 70% of the total indirect cost.

Currently, IC fabrication cost (excluding design costs) is about \$4/cm^2 at mature production levels [32]. The cost per IC to produce N chips (or equivalently, N circuit functions) is proportional to e^{kN}, where k is a constant proportional to the cost of assembly and testing [33]. The interplay between these factors and their impact on cost is illustrated in Figure 1.11. Cost per IC is minimized by maximizing both the number of chips per wafer and the proportion of good chips (also known as the *yield*).

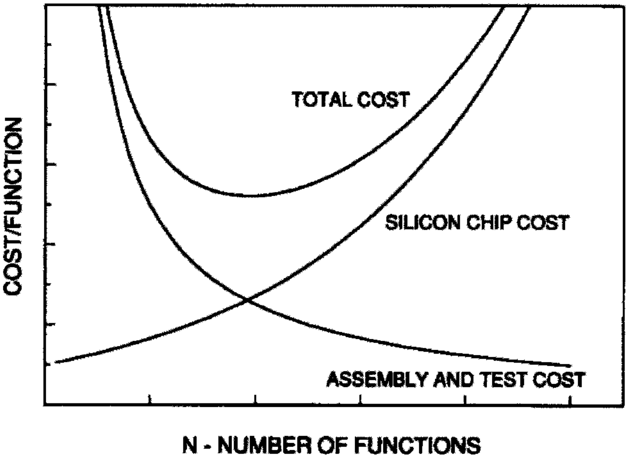


Figure 1.11. Cost per function versus number of functions [33].

1.3.2. Quality

Quality is among the most important factors in any manufacturing process. Understanding and improving quality are key ingredients to business success, growth, and enhanced competitiveness. Significant return on investment may be realized from adequate attention to continuous quality improvement as an integral part of an overall business strategy.

The term “quality” may be defined in many ways. The traditional definition is based on the notion that products must meet the requirements of those who use them. Thus, in this text, we adopt the simple definition of “fitness for use.” This definition encompasses two general aspects: quality of design and quality of conformance. *Quality of design* is affected by choices in fabrication materials, component specifications, product size, and other features. *Quality of conformance* addresses how well a product conforms to the specifications required by the design. Quality of conformance is impacted by the manufacturing process, equipment performance, competence and training of the workforce, and the implementation of quality control procedures.

1.3.3. Variability

An alternative definition of quality of conformance is “the inverse of variability.” This definition implies that quality may be improved by reducing variation in the various figures of merit that define product performance. This reduced variability translates directly into lower manufacturing costs due to less misprocessing, rework, and waste. Thus, processes in which the degree of quality is repeatable with a high degree of uniformity are preferred.

Unfortunately, a certain amount of variability is inherent in every product. No two products are ever completely identical. For example, the dimensions of two thin metal films used for IC interconnect will vary according to the precise conditions and equipment used to deposit and pattern the films. Small variations might have negligible impact on the final product, but large variations can lead to final products that are unacceptable. *Quality improvement* is defined as the reduction of such variability in processes and products. Since variability is usually described in statistical terms, statistical methods are necessary for quality improvement efforts.

1.3.4. Yield

As previously mentioned, IC cost is minimized by maximizing both the number of chips produced (i.e., the throughput) and the proportion of functionally operational chips per wafer. The latter parameter is known as the *yield.* As a consequence of its direct impact on manufacturing cost, yield is perhaps the most important figure of merit in semiconductor manufacturing.

Yield improvement achieved over time is referred to as “yield learning.” Strategies for accelerated yield learning are critical for the economic viability of semiconductor manufacturing operations, as illustrated in Figure 1.12. Business goals

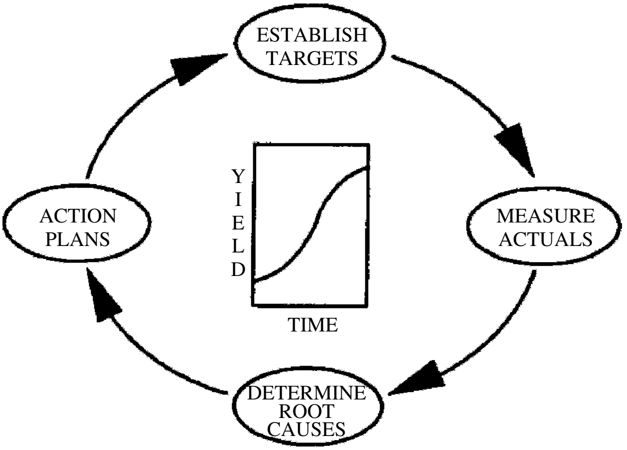


Figure 1.12. Yield learning cycle [33].

drive yield targets. The actual yield achieved is regularly monitored and tracked against those targets. The root causes of yield detractors are systematically identified and analyzed. Appropriate action plans to eliminate these causes are subsequently developed and implemented. Once target yields are achieved, they are modified (usually increased), and the learning cycle is repeated. It is of paramount importance to both reduce the duration of this cycle and optimize its efficiency.

1.3.5. Reliability

Another dimension of the quality of electronic products is their reliability. *Reliability* is a characteristic of a product that is associated with the probability that it will perform its intended function under specified conditions for a stated period of time. The enhancement of reliability is accomplished by failure-mode analysis, which is aimed at identifying the mechanisms for failure and translating this information into remedies that impact design and manufacturing processes. Reliability is usually quantified by statistical inference techniques applied to a suitable population of devices that have undergone extensive testing and failure-mode analysis. However, although the reliability of integrated circuits is often directly impacted by the manufacturing process, a detailed study of reliability and associated topics is beyond the scope of this text. Readers interested in a more thorough treatment are referred to Nash [34].

1.4. MANUFACTURING SYSTEMS

In general, manufacturing systems may be subdivided into two categories: (1) continuous-flow manufacturing and (2) discrete-parts manufacturing. *Continuous-flow* manufacturing involves chemical or physical processes that change the state of the part before the part is connected to other components to form a finished product. Most of the unit processes used in IC fabrication prior

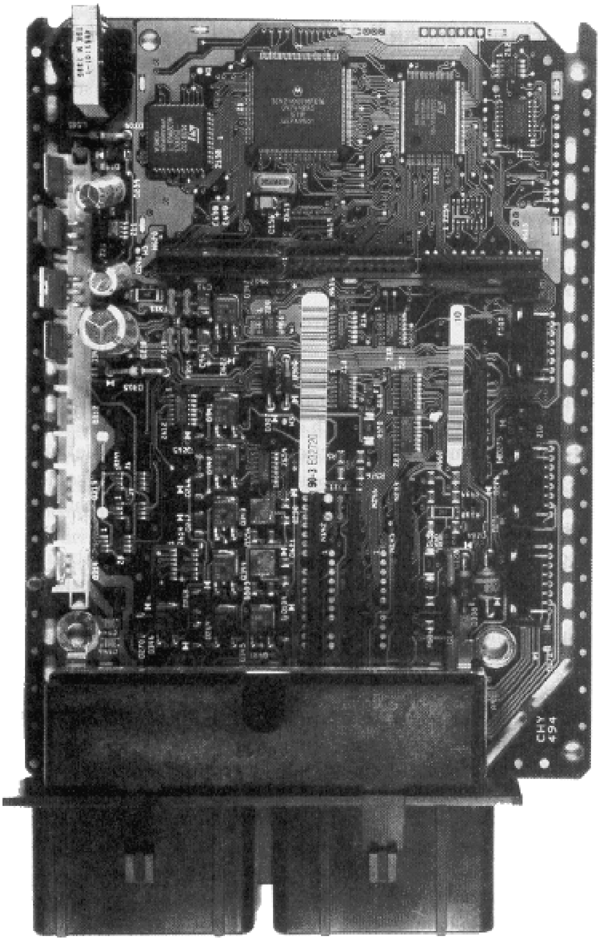


Figure 1.13. Printed circuit board for a single-board engine controller [35].

to wafer dicing and packaging are continuous flow manufacturing operations. *Discrete-parts manufacturing*, on the other hand, refers to the assembly of distinct pieces to yield a final product. In microelectronics manufacturing, an example of a product assembled using discrete parts manufacturing is a printed circuit board (PCB) populated by individual ICs (as shown in Figure 1.13).

1.4.1. Continuous Flow

Continuous-flow manufacturing refers to processing operations that do not involve assembly of discrete parts. For continuous-flow manufacturing operations, the process inputs (see Figure 1.1) are the semiconductor substrate and raw materials such as dopants, insulators, and metals. Continuous-flow processes consist of the steps such as those described in Section 1.2. These processes may be further subdivided into *batch* and *single-workpiece* (or *single-wafer*) operations.

1.4.1.1. Batch Processes

Batch processes are those that operate on multiple products simultaneously. In IC manufacturing, the items being processed are semiconductor wafers, and the batches are called "lots." State-of-the-art semiconductor manufacturing factories employ a plethora of batch fabrication equipment, such as furnaces for high-volume wafer processing. This facilitates factory throughputs on the order of tens of thousands of wafers processed per month.

An example of a batch process in semiconductor manufacturing is chemical vapor deposition (CVD; see Chapter 2). Figure 1.14 shows a schematic of a typical CVD furnace. In this type of hot-wall, reduced-pressure reactor, the quartz furnace tube is heated in three individual zones, and reactive gas is introduced at one end and pumped out the opposite end. The wafers are placed vertically side-by-side in a container known as a "boat." Gas reacting on the surface of the wafers causes the desired thin films to be deposited.

However, the high manufacturing throughput that is characteristic of batch processing is often achieved at the expense of uniformity and process control. In the case of CVD, for example, wafers farthest from the gas inlet may exhibit lower deposition rates as a result of the reduced availability of reactant gases, which are consumed by reactions closer to the inlet. This effect can be compensated for somewhat by increasing the deposition temperature in each subsequent reaction zone from the inlet.

1.4.1.2. Single Workpiece

Single-workpiece manufacturing operations involve individual items processed one at a time. In IC manufacturing, the workpiece is the semiconductor wafer. As wafer sizes have grown over the years, single-wafer processing approaches have proliferated. This has occurred for several reasons. First, scaling up batch tools and maintaining uniformity across the wafer surface becomes more difficult for wafers 200 mm in diameter and larger. At the same time, for submicrometer features, it is nearly impossible to maintain features size control across these large wafers. In addition, when only a single wafer is processed at a time, if any flaw in a process step is detected, it can be corrected before the next wafer is

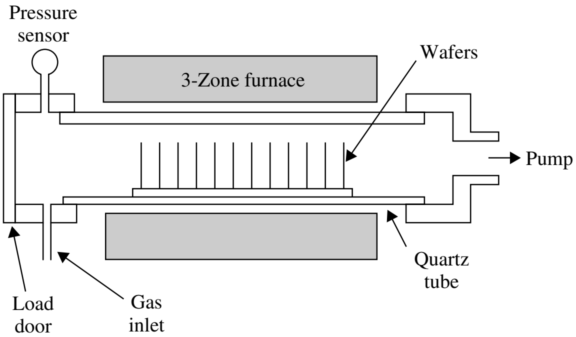


Figure 1.14. Example of a batch process: a CVD reactor [1].

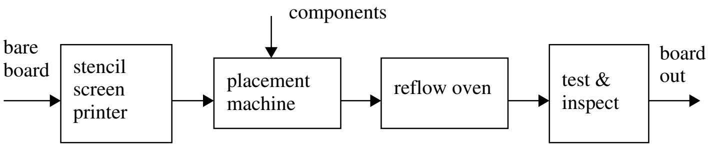


Figure 1.15. Surface mount printed circuit board assembly process.

processed. Finally, process development for large wafers has become increasingly expensive, and these costs are mitigated by the single wafer approach.

Therefore, semiconductor manufacturing has evolved from primarily a batch operation to an increasingly single-wafer operation. The advantages of single-wafer processing include (1) lower overall factory cost, (2) enhanced observability by in situ sensors for more robust process control, (3) rapid manufacturing cycle time, and (4) increased flexibility for manufacturing numerous products based on different technologies [36].

1.4.2. Discrete Parts

In discrete-parts microelectronics manufacturing, the inputs to the manufacturing system are the bare printed circuit board and the various circuit components. For example, in surface mount assembly, the manufacturing process consists of the following (Figure 1.15):

1. Screen-printing solder paste onto the bonding pads of the circuit board with a stencil printer
2. Placing the circuit components (ICs and passives) onto the pad locations using a placement machine
3. Melting the solder paste in a reflow oven to form the connection between components and the pads
4. Testing and inspecting the populated board for quality control

Following attachment of the ICs, the output of the process is the fully interconnected and populated circuit board.

The output of the process is a populated circuit board that is ready for integration into an electronic system.

1.5. OUTLINE FOR REMAINDER OF THE BOOK

In Chapter 2 we will provide a broad overview of the manufacturing technology and process flows used to produce a variety of semiconductor products. The individual unit processes used in fabricating ICs, as well as techniques for process integration and IC packaging, will be discussed. The unit processes include oxidation, photolithography, doping, etching, thin-film deposition, and planarization. The integrated process flows, which focus on silicon technology, include the complementary metal–oxide–semiconductor (CMOS), bipolar, and BiCMOS processes.

For all aspects of semiconductor manufacturing, testing and inspection are necessary to yield high-quality products. In Chapter 3, therefore, various process monitoring methods, including those that focus on product wafers and those that focus on the equipment used to produce those wafers, are discussed in detail. Maintaining quality involves the use of *statistical process control* (SPC). Since product variability is often described in statistical terms, statistical methods will necessarily play a central role in quality control and improvement efforts. Therefore, Chapter 4 will provide a review of statistical fundamentals.

Ultimately, the key metric to be used to evaluate any manufacturing process is cost, and cost is directly impacted by yield. *Yield* refers to the proportion of manufactured products that perform as required by a set of specifications. Yield is inversely proportional to the total manufacturing cost—the higher the yield, the lower the cost. *Yield modeling* is presented in Chapter 5. Chapter 6 will then focus on the use of SPC to analyze quality issues and improve yield.

A designed experiment is an extremely useful tool for discovering key variables that influence quality characteristics. *Statistical experimental design* is a powerful approach for systematically varying controllable process conditions and determining their impact on output parameters that measure quality. Data derived from such experiments can then be used to construct *process models* of various types that enable the analysis and prediction of manufacturing process behavior. Statistical experimental design is presented in Chapter 7, and process modeling concepts are introduced in Chapter 8.

Finally, several advanced process control topics are the subject of Chapters 9 and 10. These topics include run-by-run control of unit processes and supervisory control of process sequences, as well as the diagnosis of process and equipment malfunctions.

SUMMARY

In this chapter, we have provided background and motivation for the study of semiconductor manufacturing and process control. We have done so by surveying the history of integrated circuit processing, describing the attributes of manufacturing systems, and discussing the goals and objectives of modern electronics manufacturing operations. In so doing, this chapter has provided a foundation for the various issues relevant to semiconductor manufacturing that will be presented in the remainder of the book.

PROBLEMS

1.1. List the input and output parameters of a typical semiconductor manufacturing process.

1.2. What were the key milestones in the historical evolution of semiconductor manufacturing? How did the evolution of semiconductor process technology interact with and impact the development of manufacturing technology?

1.3. Describe the basic unit processes involved in IC fabrication and the sequences in which they are performed to yield products.

1.4. What is the significance of information flow within an IC factory?

1.5. Explain the differences between real-time, run-by-run, and supervisory control.

1.6. Why are IC factories organized into workcells?

1.7. List and prioritize the overall goals of IC manufacturing.

1.8. What is the difference between continuous-flow and discrete manufacturing processes? What role do each of these play in semiconductor manufacturing?

1.9. Why has semiconductor manufacturing evolved from batch operations toward single-wafer operations? What are the advantages and disadvantages of each approach?

REFERENCES

1. G. May and S. Sze, *Fundamentals of Semiconductor Fabrication*, Wiley, New York 2002.
2. G. Gunderson, *A New Economic History of America*, McGraw-Hill, New York, 1976.
3. D. Montgomery, *Introduction to Statistical Quality Control*, 3rd ed., Wiley, New York, 1997.
4. J. Stenerson and K. Curran, *Computer Numerical Control*, Prentice-Hall, Upper Saddle River, NJ, 1997.
5. D. Shore, “Steel-Making in Ancient Africa,” in *Blacks in Science: Ancient and Modern*, I. Van Sertima, ed., Transaction Books, New Brunswick, NJ, 1986, p. 157.
6. M. Hepher, “The Photoresist Story,” *J. Photo. Sci.* **12**, 181 (1964).
7. A. Fick, “Ueber Diffusion,” *Ann. Phys. Lpz.* **170**, 59 (1855).
8. J. Czochralski, “Ein neues Verfahren zur Messung der Kristallisationsgeschwindigkeit der Metalle,” *Z. Phys. Chem.* **92**, 219 (1918).
9. P. W. Bridgman, “Certain Physical Properties of Single Crystals of Tungsten, Antimony, Bismuth, Tellurium, Cadmium, Zinc, and Tin,” *Proc. Am. Acad. Arts Sci.* **60**, 303 (1925).
10. W. G. Pfann, *Semiconductor Signal Translating Device*, U.S. Patent 2,597,028 (1952).
11. J. Andrus, *Fabrication of Semiconductor Devices*, U.S. Patent 3,122,817 (filed 1957; granted 1964).
12. C. J. Frosch and L. Derrick, “Surface Protection and Selective Masking during Diffusion in Silicon,” *J. Electrochem. Soc.* **104**, 547 (1957).
13. N. N. Sheftal, N. P. Kokorish, and A. V. Krasilov, “Growth of Single-Crystal Layers of Silicon and Germanium from the Vapor Phase,” *Bull. Acad. Sci USSR, Phys. Ser.* **21**, 140, (1957).
14. W. Shockley, *Forming Semiconductor Device by Ionic Bombardment*, U.S. Patent 2,787,564 (1958).

15. J. S. Kilby, "Invention of the Integrated Circuit," *IEEE Trans. Electron Devices* **ED-23**, 648 (1976).
16. R. N. Noyce, *Semiconductor Device-and-Lead Structure*, U.S. Patent 2,981,877 (filed 1959, granted 1961).
17. F. M. Wanlass and C. T. Sah, "Nanowatt Logics Using Field-Effect Metal-Oxide Semiconductor Triodes," *Tech. Digest IEEE Int. Solid-State Circuit Conf.*, 1963, p. 32.
18. R. M. Dennard, *Field Effect Transistor Memory*, U.S. Patent 3,387,286 (filed 1967, granted 1968).
19. R. E. Kerwin, D. L. Klein, and J. C. Sarace, *Method for Making MIS Structure*, U.S. Patent 3,475,234 (1969).
20. H. M. Manasevit and W. I. Simpson, "The Use of Metal–Organic in the Preparation of Semiconductor Materials. I. Epitaxial Gallium-V Compounds," *J. Electrochem. Soc.* **116**, 1725 (1969).
21. S. M. Irving, K. E. Lemons, and G. E. Bobos, *Gas Plasma Vapor Etching Process*, U.S. Patent 3,615,956 (1971).
22. A. Y. Cho, "Film Deposition by Molecular Beam Technique," *J. Vac. Sci. Technol.* **8**, S31 (1971).
23. R. Slater, *Portraits in Silicon*, MIT Press, Cambridge, MA, 1987, p. 175.
24. R. Rung, H. Momose, and Y. Nagakubo, "Deep Trench Isolated CMOS Devices," *Tech. Digest. IEEE Int. Electron Devices Meet.*, 1982, p. 237.
25. B. Davari et al., "A New Planarization Technique, Using a Combination of RIE and Chemical Mechanical Polish (CMP)," *Tech. Digest IEEE Int. Electron Devices Meet.*, 1989, p. 61.
26. J. Paraszczak et al., "High Performance Dielectrics and Processes for ULSI Interconnection Technologies," *Tech. Digest IEEE Int. Electron Devices Meet.*, 1993, p. 261.
27. M. Dax, "Top Fabs of 1996," *Semiconductor Intl.* **19**(5) (1996).
28. D. Hodges, L. Rowe, and C. Spanos, "Computer-Integrated Manufacturing of VLSI," *Proc. IEEE/CHMT Intl. Electron. Manuf. Tech. Symp.*, 1989.
29. IEEE Electron Devices Society, *Fifty Years of Electron Devices*, IEEE, Piscataway, NJ, 2002.
30. J. Moyne, E. del Castillo, and A. Hurwitz, *Run-to-Run Control in Semiconductor Manufacturing*, CRC Press, Boca Raton, FL, 2001.
31. M. Penn, "Economics of Semiconductor Production," *Microelectron. J.* **23**, 255–265 (1992).
32. R. Tummala (ed.), *Fundamentals of Microsystems Packaging*, McGraw-Hill, New York, 2001.
33. A. Landzberg, *Microelectronics Manufacturing Diagnostics Handbook*, Van Nostrand Reinhold, New York, 1993.
34. F. Nash, *Estimating Device Reliability: Assessment of Credibility*, Kluwer Academic Publishers, Boston, MA, 1993.
35. W. Brown (ed.), *Advanced Electronic Packaging*, IEEE Press, New York, 1999.
36. M. Moslehi, R. Chapman, M. Wong, A. Paranjpe, H. Najm, J. Kuehne, R. Yeakley, and C. Davis, "Single-Wafer Integrated Semiconductor Device Processing," *IEEE Trans. Electron Devices* **39**(1) (Jan. 1992).

2

TECHNOLOGY OVERVIEW

OBJECTIVES

- Provide an overview of the critical unit processes in semiconductor manufacturing.
- Describe the integration of such processes into sequences for fabricating specific technology families.

INTRODUCTION

Planar fabrication technology is used extensively for integrated circuit manufacturing. In Section 1.2.1 we briefly described the major steps of a planar process. We provide a more thorough description of these steps, as well as their integration for particular technology families, in this chapter. However, this treatment is in no way intended to be comprehensive. More complete and detailed discussions can be found in several other texts, such as *Fundamentals of Semiconductor Fabrication* [1], for example.

2.1. UNIT PROCESSES

Chapter 1 provided an introduction to the key unit process steps in IC fabrication, including oxidation, photolithography, etching, ion implantation, and metallization. This was accomplished using the description of the process sequence used

Fundamentals of Semiconductor Manufacturing and Process Control,
By Gary S. May and Costas J. Spanos

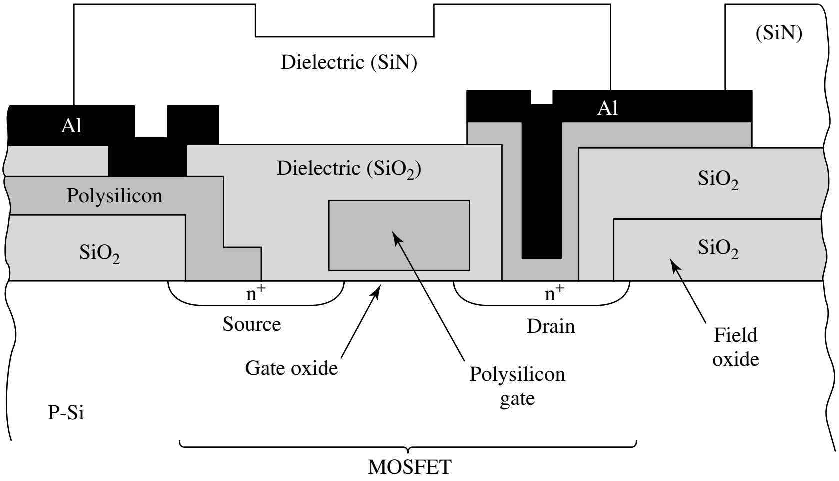


Figure 2.1. Cross section of a MOSFET [1].

to fabricate a *p–n* junction. Here, we describe each of these steps, as well as planarization, in more detail.

2.1.1. Oxidation

Many different kinds of thin films are used to fabricate discrete devices and integrated circuits, including thermal oxides, dielectric layers, polycrystalline silicon, and metal films. For example, a silicon *n*-channel MOSFET (Figure 2.1) uses all four groups of films. An important oxide layer is the gate oxide, under which a conducting channel can be formed between the source and the drain. A related layer is the field oxide, which provides isolation from other devices. Both gate and field oxides generally are grown by a thermal oxidation process because only thermal oxidation can provide the highest-quality oxides having the lowest interface trap densities.

Semiconductors can be oxidized by various methods, including thermal oxidation, electrochemical anodization, and plasma-enhanced chemical vapor deposition (PECVD; see Section 2.1.5). Among these, thermal oxidation is the most important for silicon devices. It is a key process in modern silicon IC technology. The basic thermal oxidation apparatus (shown in Figure 2.2) consists of a resistance-heated furnace, a cylindrical fused-quartz tube containing the silicon wafers held vertically in a slotted quartz boat, and a source of either pure dry oxygen or pure water vapor. Oxidation temperature is generally in the range of 900–1200°C, and the typical gas flowrate is about 1 L/min. The oxidation system uses microprocessors to regulate the gas flow sequence, to control the automatic insertion and removal of silicon wafers, to ramp the temperature up (i.e., to increase the furnace temperature linearly) from a low temperature to the oxidation temperature, to maintain the oxidation temperature to within ±1°C, and to ramp the temperature down when oxidation is completed.

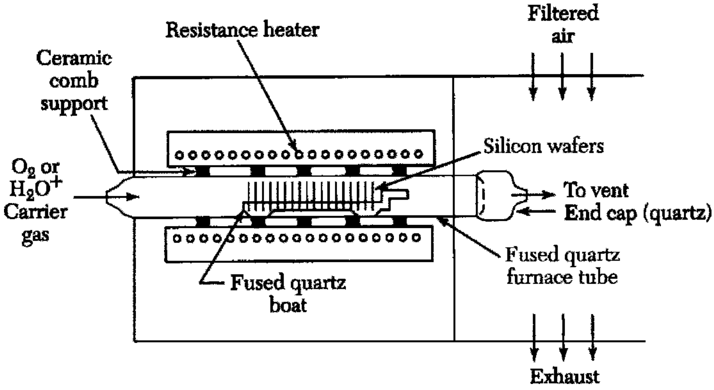


Figure 2.2. Schematic of an oxidation furnace [1].

2.1.1.1. Growth Kinetics

The following chemical reactions describe the thermal oxidation of silicon in oxygen ("dry" oxidation) and water vapor ("wet" oxidation), respectively:

$$\text{Si(solid)} + \text{O}_2\text{(gas)} \rightarrow \text{SiO}_2\text{(solid)} \tag{2.1}$$

$$\text{Si(solid)} + 2\text{H}_2\text{O(gas)} \rightarrow \text{SiO}_2\text{(solid)} + 2\text{H}_2\text{(gas)} \tag{2.2}$$

The silicon–silicon dioxide interface moves into the silicon during the oxidation process. This creates a new interface region, with surface contamination on the original silicon ending up on the oxide surface. As a result of the densities and molecular weights of silicon and silicon dioxide, growing an oxide of thickness x consumes a layer of silicon $0.44x$ thick (Figure 2.3).

The kinetics of silicon oxidation can be described on the basis of the simple model illustrated in Figure 2.4. A silicon slice contacts the oxidizing species (oxygen or water vapor), resulting in a surface concentration of C_0 molecules/cm^3 for these species. The magnitude of C_0 equals the equilibrium bulk concentration of the species at the oxidation temperature. The equilibrium concentration generally is proportional to the partial pressure of the oxidant adjacent to the oxide surface.

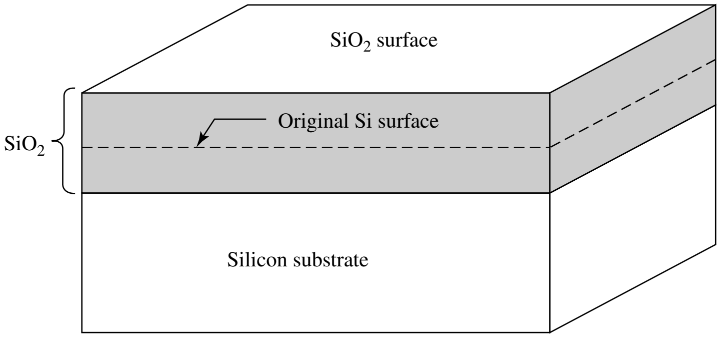


Figure 2.3. Movement of silicon–silicon dioxide interface during oxide growth [1].

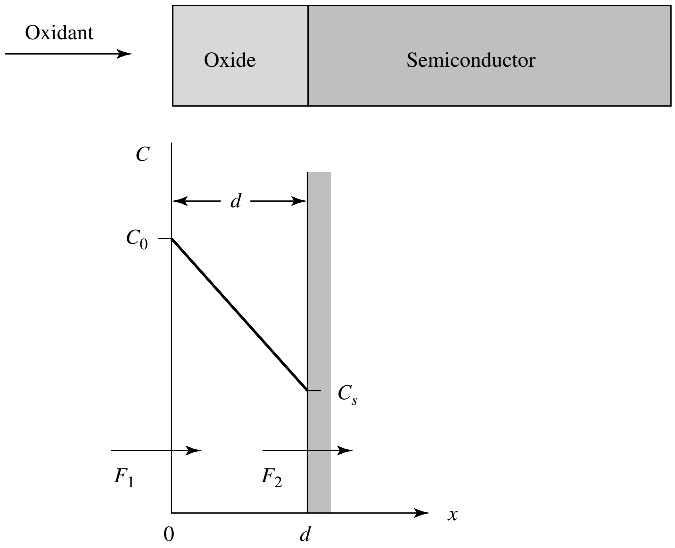


Figure 2.4. Basic model for the thermal oxidation of silicon [1].

At 1000°C and a pressure of 1 atm, the concentration C_0 is 5.2×10^{16} cm^{-3} for dry oxygen and 3×10^{19} cm^{-3} for water vapor.

The oxidizing species diffuses through the silicon dioxide layer, resulting in a concentration C_s at the surface of silicon. The flux F_1 can be written as

$$F_1 = D\frac{dC}{dx} \cong \frac{D(C_0 - C_s)}{x} \tag{2.3}$$

where D is the diffusion coefficient of the oxidizing species, and x is the thickness of the oxide layer already present.

At the silicon surface, the oxidizing species reacts chemically with silicon. Assuming the rate of reaction to be proportional to the concentration of the species at the silicon surface, the flux F_2 is given by

$$F_2 = \kappa C_s \tag{2.4}$$

where κ is the surface reaction rate constant for oxidation. At the steady state, $F_1 = F_2 = F$. Combining Eqs. (2.3) and (2.4) gives

$$F = \frac{DC_0}{x + (D/\kappa)} \tag{2.5}$$

The reaction of the oxidizing species with silicon forms silicon dioxide. Let C_1 be the number of molecules of the oxidizing species in a unit volume of the oxide. There are 2.2×10^{22} silicon dioxide molecules/cm^3 in the oxide, and one oxygen molecule (O_2) is added to each silicon dioxide molecule, whereas we add two water molecules (H_2O) to each SiO_2 molecule. Therefore, C_1 for oxidation in dry oxygen is 2.2×10^{22} cm^{-3}, and for oxidation in water vapor it is twice this number (4.4×10^{22} cm^{-3}). Thus, the growth rate of the oxide layer thickness is given by

$$\frac{dx}{dt} = \frac{F}{C_1} = \frac{DC_0/C_1}{x + (D/\kappa)} \tag{2.6}$$

This differential equation can be solved subject to the initial condition, $x(0) = d_0$, where d_0 is the initial oxide thickness; d_0 can also be regarded as the thickness of oxide layer grown in an earlier oxidation step. Solving Eq. (2.6) yields the general relationship for the oxidation of silicon:

$$x^2 + \frac{2D}{\kappa}x = \frac{2DC_0}{C_1}(t + \tau) \tag{2.7}$$

where $\tau \equiv (d_0^2 + 2Dd_0/\kappa)C_1/2DC_0$, which represents a time coordinate shift to account for the initial oxide layer d_0.

The oxide thickness after an oxidizing time t is given by

$$x = \frac{D}{\kappa}\left[\sqrt{1 + \frac{2C_0\kappa^2(t + \tau)}{DC_1}} - 1\right] \tag{2.8}$$

For small values of t, Eq. (2.8) reduces to

$$x \cong \frac{C_0\kappa}{C_1}(t + \tau) \tag{2.9}$$

and for larger values of t, it reduces to

$$x \cong \sqrt{\frac{2DC_0}{C_1}(t + \tau)} \tag{2.10}$$

During the early stages of oxide growth, when surface reaction is the rate limiting factor, the oxide thickness varies linearly with time. As the oxide layer becomes thicker, the oxidant must diffuse through the oxide layer to react at the silicon–silicon dioxide interface and the reaction becomes diffusion-limited. The oxide growth then becomes proportional to the square root of the oxidizing time, which results in a parabolic growth rate.

Equation (2.7) is often written in a more compact form

$$x^2 + Ax = B(t + \tau) \tag{2.11}$$

where $A = 2D/\kappa$, $B = 2DC_0/C_1$ and $B/A = \kappa C_0/C_1$. Using this form, Eqs. (2.9) and (2.10) can be written as

$$x = \frac{B}{A}(t + \tau) \tag{2.12}$$

for the linear region and as

$$x^2 = B(t + \tau) \tag{2.13}$$

for the parabolic region. For this reason, the term B/A is referred to as the *linear rate constant* and B is the *parabolic rate constant.* Experimentally measured results agree with the predictions of this model over a wide range of oxidation conditions. For wet oxidation, the initial oxide thickness d_0 is very small, or $\tau \cong 0$. However, for dry oxidation, the extrapolated value of d_0 at $t = 0$ is about 25 nm. Thus, the use of Eq. (2.11) for dry oxidation on bare silicon requires a value for τ that can be generated using this initial thickness. Table 2.1 lists the values of the rate constants for wet oxidation of silicon, and Table 2.2 lists the values for dry oxidation.

The temperature dependence of the linear rate constant B/A is shown in Figure 2.5 for both dry and wet oxidation and for (111)- and (100)-oriented silicon wafers [1]. The linear rate constant varies as exp $(-E_a/kT)$, where the activation energy E_a is about 2 eV for both dry and wet oxidation. This closely agrees with the energy required to break silicon–silicon bonds, 1.83 eV/molecule. Under a given oxidation condition, the linear rate constant depends on crystal orientation. This is because the rate constant is related to the rate of incorporation of oxygen atoms into the silicon. The rate depends on the surface bond structure of silicon atoms, making it orientation-dependent. Because the density of available bonds on the (111) plane is higher than that on the (100) plane, the linear rate constant for (111) silicon is larger.

Figure 2.6 shows the temperature dependence of the parabolic rate constant B, which can also be described by $\exp(-E_a/kT)$. The activation energy E_a is 1.24 eV for dry oxidation. The comparable activation energy for oxygen diffusion in fused silica is 1.18 eV. The corresponding value for wet oxidation, 0.71 eV, compares favorably with the value of 0.79 eV for the activation energy of diffusion of water in fused silica. The parabolic rate constant is independent of crystal orientation. This independence is expected because it is a measure of

Table 2.1. Rate constants for wet oxidation of silicon.

Temperature (°C)	A (μm)	B (μm^2/h)	τ (h)
1200	0.05	0.72	0
1100	0.11	0.51	0
1000	0.226	0.287	0
920	0.5	0.203	0

Table 2.2. Rate constants for dry oxidation of silicon.

Temperature (°C)	A (μm)	B (μm^2/h)	τ (h)
1200	0.04	0.045	0.027
1100	0.09	0.027	0.076
1000	0.165	0.0117	0.37
920	0.235	0.0049	1.4
800	0.37	0.0011	9.0
700	—	—	81.0

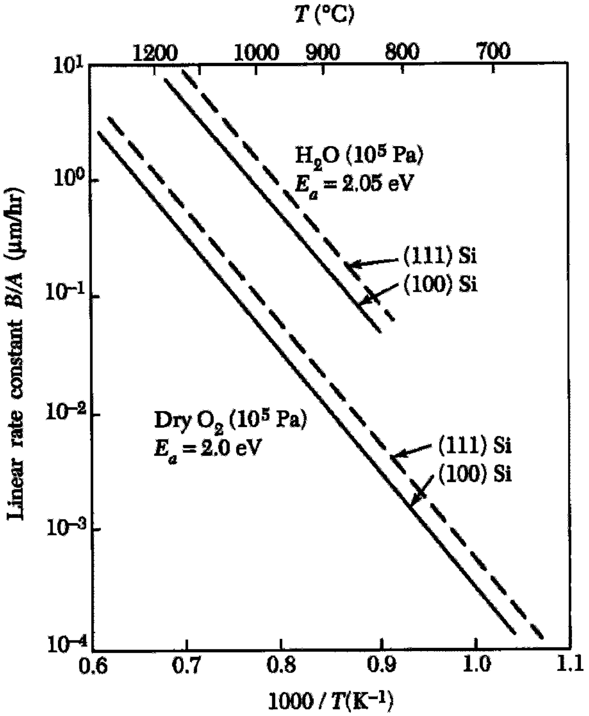


Figure 2.5. Linear rate constant versus temperature [1].

the diffusion process of the oxidizing species through a random network layer of amorphous silica.

Although oxides grown in dry oxygen have the best electrical properties, considerably more time is required to grow the same oxide thickness at a given temperature in dry oxygen than in water vapor. For relatively thin oxides such as the gate oxide in a MOSFET (typically ≤20 nm), dry oxidation is used. However, for thicker oxides such as field oxides (≥20 nm) in MOS integrated circuits, and for bipolar devices, oxidation in water vapor (or steam) is used to provide both adequate isolation and passivation.

2.1.1.2. Thin Oxide Growth

Relatively slow growth rates must be used to reproducibly grow thin oxide films of precise thickness. Approaches to achieve such slower growth rates include growth in dry O_2 at atmospheric pressure and lower temperatures (800–900°C); growth at pressures lower than atmospheric pressure; growth in a reduced partial pressures of O_2 by using a diluent inert gas, such as N_2, Ar, or He, together with the gas containing the oxidizing species; and the use of composite oxide films with the gate oxide films consisting of a layer of thermally grown SiO_2 and an overlayer of chemical vapor deposition (CVD) SiO_2. However, the mainstream approach for gate oxides 10–15 nm thick is to grow the oxide film at atmospheric pressure and lower temperatures (800–900°C). With this approach, processing

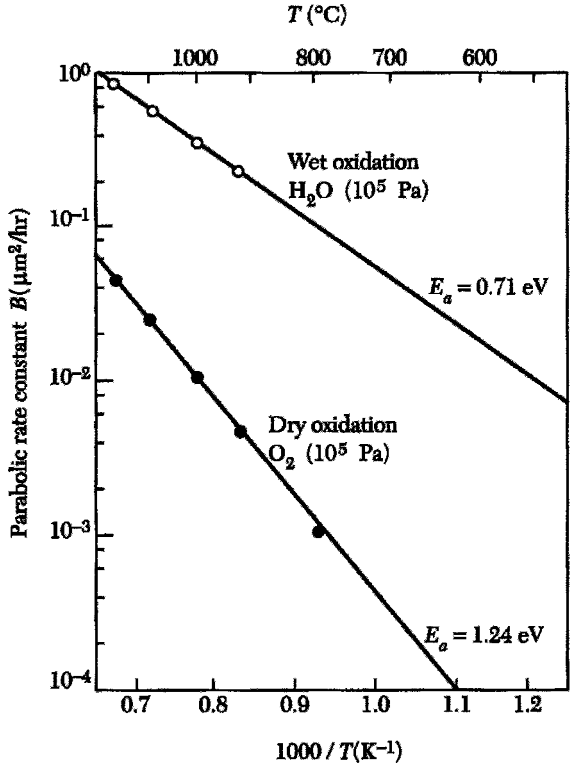


Figure 2.6. Parabolic rate constant versus temperature [1].

using modern *vertical* oxidation furnaces can grow reproducible, high-quality 10-nm oxides to within 0.1 nm across the wafer.

It was noted earlier that for dry oxidation, there is a rapid early growth that gives rise to an initial oxide thickness d_0 of about 20 nm. Therefore, the simple model given by Eq. (2.11) is not valid for dry oxidation with an oxide thickness $\leq$20 nm. For ultra-large-scale integration, the ability to grow thin (5–20 nm), uniform, high-quality reproducible gate oxides has become increasingly important.

In the early stage of growth in dry oxidation, there is a large compressive stress in the oxide layer that reduces the oxygen diffusion coefficient in the oxide. As the oxide becomes thicker, the stress will be reduced due to the viscous flow of silica and the diffusion coefficient will approach its stress-free value. Therefore, for thin oxides, the value of D/κ may be sufficiently small that we can neglect the term Ax in Eq. (2.11) and obtain

$$x^2 - d_0{}^2 = Bt \qquad (2.14)$$

where d_0 is equal to $\sqrt{2DC_0\tau/C_1}$, which is the initial oxide thickness when time is extrapolated to zero, and B is the parabolic rate constant defined previously. We therefore expect the initial growth in dry oxidation to follow a parabolic form.

2.1.1.3. Oxide Quality

Oxides used for masking are usually grown by wet oxidation. A typical growth cycle consists of a dry–wet–dry sequence. Most of the growth in such a sequence occurs in the wet phase, since the SiO_2 growth rate is much higher when water is used as the oxidant. Dry oxidation, however, results in a higher quality oxide that is denser and has a higher breakdown voltage (5–10 MV/cm). It is for these reasons that the thin gate oxides in MOS devices (see Section 2.2) are usually formed using dry oxidation.

MOS devices are also affected by charges in the oxide and traps at the SiO_2–Si interface. The basic classification of these traps and charges, shown in Figure 2.7, are interface-trapped charge, fixed-oxide charge, oxide-trapped charge, and mobile ionic charge. Interface-trapped charges (Q_{it}) are due to the SiO_2–Si interface properties and dependent on the chemical composition of this interface. The traps are located at the SiO_2–Si interface with energy states in the silicon-forbidden bandgap. The fixed charge (Q_f) is located within approximately 3 nm of the SiO_2–Si interface. Generally, Q_f is positive and depends on oxidation and annealing conditions, as well as on the orientation of the silicon substrate. Oxide-trapped charges (Q_{ot}) are associated with defects in the silicon dioxide. These charges can be created, for example, by X-ray radiation or high-energy electron bombardment. Mobile ionic charges (Q_m), which result from contamination from sodium or other alkali ions, are mobile within the oxide under raised-temperatures (e.g., >100°C) and high-electric-field operations. Trace contamination by alkali metal ions may cause stability problems in semiconductor devices operated under high-bias and high-temperature conditions. Under these

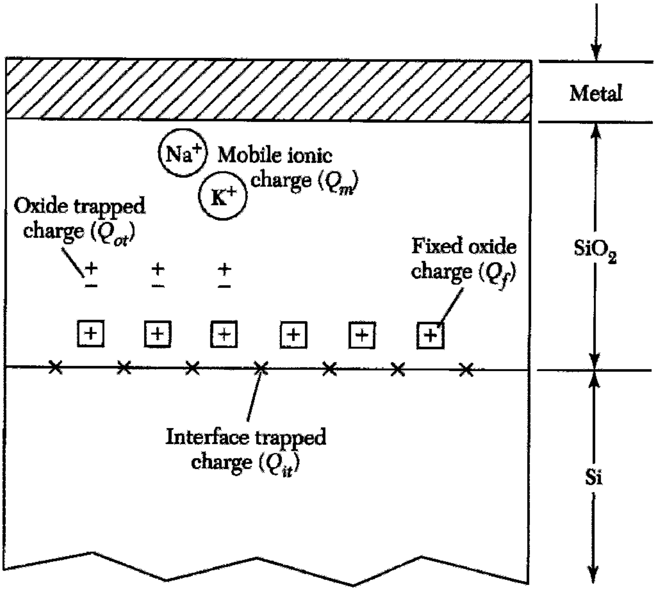

Metal
Na^+
Mobile ionic charge (Q_m)
K^+
Oxide trapped charge (Q_{ot})
Fixed oxide charge (Q_f)
SiO_2
Interface trapped charge (Q_{it})
Si

Figure 2.7. Description of charges associated with thermal oxides [1].

conditions mobile ionic charges can move back and forth through the oxide layer and cause threshold voltage shifts. Therefore, special attention must be paid to the elimination of mobile ions in device fabrication.

2.1.2. Photolithography

Photolithography is the process of transferring patterns of geometric shapes on a mask to a thin layer of photosensitive material (called *photoresist*) covering the surface of a semiconductor wafer. These patterns define the various regions in an integrated circuit, such as the implantation regions, the contact windows, and the bonding pad areas. The resist patterns defined by the lithographic process are not permanent elements of the final device, but only replicas of circuit features. To produce circuit features, these resist patterns must be transferred once more into the underlying layers of the device. Pattern transfer is accomplished by an etching process that selectively removes unmasked portions of a layer (see Section 2.1.4).

Photolithography requires a clean processing room. The need for a cleanroom arises because dust particles in the air can settle on semiconductor wafers or lithographic masks and cause defects that result in circuit failure. For example, a dust particle on a semiconductor surface can disrupt the growth of an epitaxial film, causing the formation of dislocations. A dust particle incorporated into a gate oxide can result in enhanced conductivity and cause device failure due to low breakdown voltage. The situation is even more critical in photolithography. When dust particles adhere to the surface of a photomask, they behave as opaque patterns on the mask, and these patterns will be transferred to the underlying layer along with the circuit patterns on the mask. Figure 2.8 shows three dust particles on a photomask. Particle 1 may result in the formation of a pinhole in the underlying layer. Particle 2 is located near a pattern edge and may cause

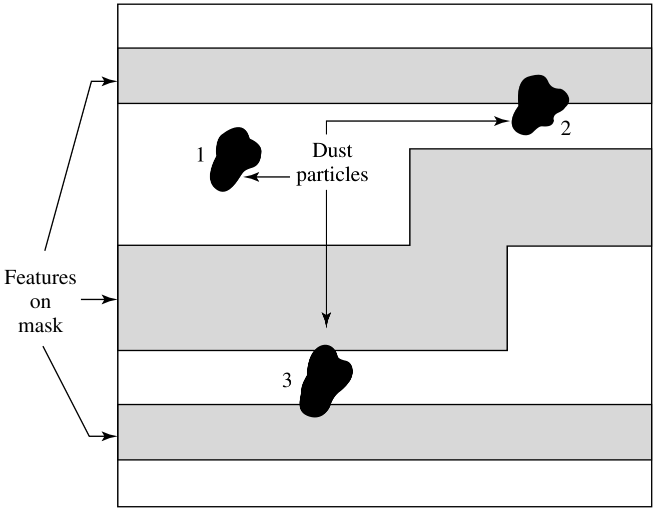


Figure 2.8. Various ways in which particles can interfere with photomask patterns [1].

a constriction of current flow in a metal runner. Particle 3 can lead to a short circuit between the two conducting regions and render the circuit useless.

In a cleanroom, the total number of dust particles per unit volume must be tightly controlled along with the temperature and humidity. There are two systems to define the classes of cleanroom. For the English system, the numerical designation of the class is taken from the maximum allowable number of particles 0.5 μm and larger per cubic foot of air. For the metric system, the class is taken from the logarithm (base 10) of the maximum allowable number of particles 0.5 μm and larger, per cubic meter. For example, a class 100 cleanroom (English system) has a dust count of 100 particles/ft^3 with particle diameters of 0.5 μm and larger, whereas a class M 3.5 cleanroom (metric system) has a dust count of $10^{3.5}$ or about 3500 particles/m^3 with particle diameters of 0.5 μm or larger.

Since the number of dust particles increases as particle size decreases, more stringent control of the cleanroom environment is required as the minimum feature lengths of ICs are reduced. For most IC fabrication areas, a class 100 cleanroom is required; that is, the dust count must be about four orders of magnitude lower than that of ordinary room air. However, for photolithography, a class 10 cleanroom or one with a lower dust count is required.

2.1.2.1. Exposure Tools

The pattern transfer process is accomplished by using a lithographic exposure tool. The performance of an exposure tool is determined by resolution, registration, and throughput. *Resolution* is the minimum feature dimension that can be transferred with high fidelity to a resist film on a semiconductor wafer. *Registration* is a measure of how accurately patterns on successive masks can be aligned (or overlaid) with respect to previously defined patterns on the wafer. *Throughput* is the number of wafers that can be exposed per unit time for a given mask level.

There are two primary optical exposure methods: shadow printing and projection printing. Shadow printing may have the mask and wafer in direct contact with one another (as in *contact printing*), or in close proximity (as in *proximity printing*). Figure 2.9a shows a basic setup for contact printing where a resist-coated wafer is brought into physical contact with a mask, and the resist is exposed by a nearly collimated beam of ultraviolet light through the back of the mask for a fixed time. The intimate contact between the resist and mask provides a resolution of ~1 μm. However, contact printing suffers from one major drawback—a dust particle on the wafer can be embedded into the mask when the mask makes contact with the wafer. The embedded particle causes permanent damage to the mask and results in defects in the wafer with each succeeding exposure.

To minimize mask damage, the proximity exposure method is used. Figure 2.9b shows the basic setup, which is similar to contact printing except that there is a small gap (10–50 μm) between the wafer and the mask during exposure. The small gap, however, results in optical diffraction at feature edges on the photomask; that is, when light passes by the edges of an opaque mask feature, fringes are formed and some light penetrates into the shadow region. As a result, resolution is degraded to the 2–5-μm range.

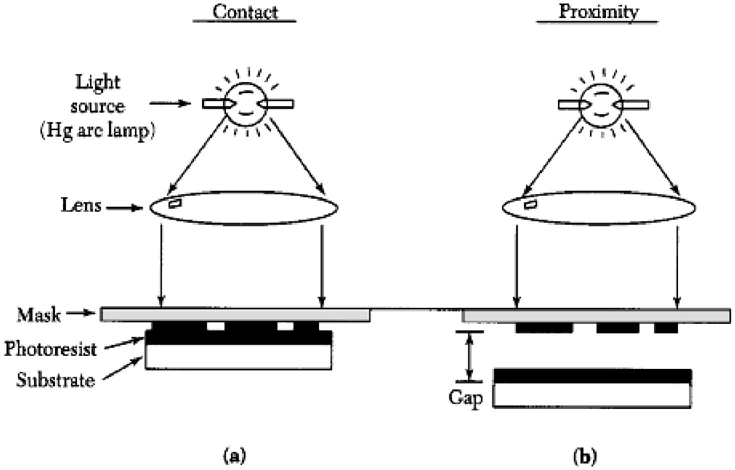


Figure 2.9. Optical shadow printing techniques: (a) contact printing; (b) proximity printing [1].

In shadow printing, the minimum linewidth, or critical dimension (CD), that can be printed is approximately

$$\mathrm{CD} \cong \sqrt{\lambda g} \tag{2.15}$$

where λ is the wavelength of the exposure radiation and g is the gap between the mask and the wafer and includes the thickness of the resist. For $\lambda = 0.4$ μm and $g = 50$ μm, the CD is 4.5 μm. If we reduce λ to 0.25 μm (a wavelength range of 0.2–0.3 μm is in the deep-UV spectral region) and g to 15 μm, the CD becomes 2 μm. Thus, there is an advantage in reducing both λ and g. However, for a given distance g, any dust particle with a diameter larger than g potentially can cause mask damage.

To avoid the mask damage problem associated with shadow printing, projection printing tools have been developed to project an image of the mask patterns onto a resist-coated wafer many centimeters away from the mask. To increase resolution, only a small portion of the mask is exposed at a time. The small image area is scanned or stepped over the wafer to cover the entire wafer surface. Figure 2.10a shows a 1 : 1 wafer scan projection system. A narrow, arc-shaped image field ~1 mm in width serially transfers the slit image of the mask onto the wafer. The image size on the wafer is the same as that on the mask.

The small image field can also be stepped over the surface of the wafer by two-dimensional translations of the wafer only, whereas the mark remains stationary. After the exposure of one chip site, the wafer is moved to the next chip site and the process is repeated. Figures 2.10b and 2.10c show the partitioning of the wafer image by *step-and-repeat projection* with a ratio of 1 : 1 or at a demagnification ratio M : 1 (e.g., 10 : 1 for a 10 times reduction on the wafer), respectively. The 1 : 1 optical systems are easier to design and fabricate than a 10 : 1 or a 5 : 1 reduction system, but it is much more difficult to produce defect-free masks at 1 : 1 than it is at a 10 : 1 or a 5 : 1 demagnification ratio.

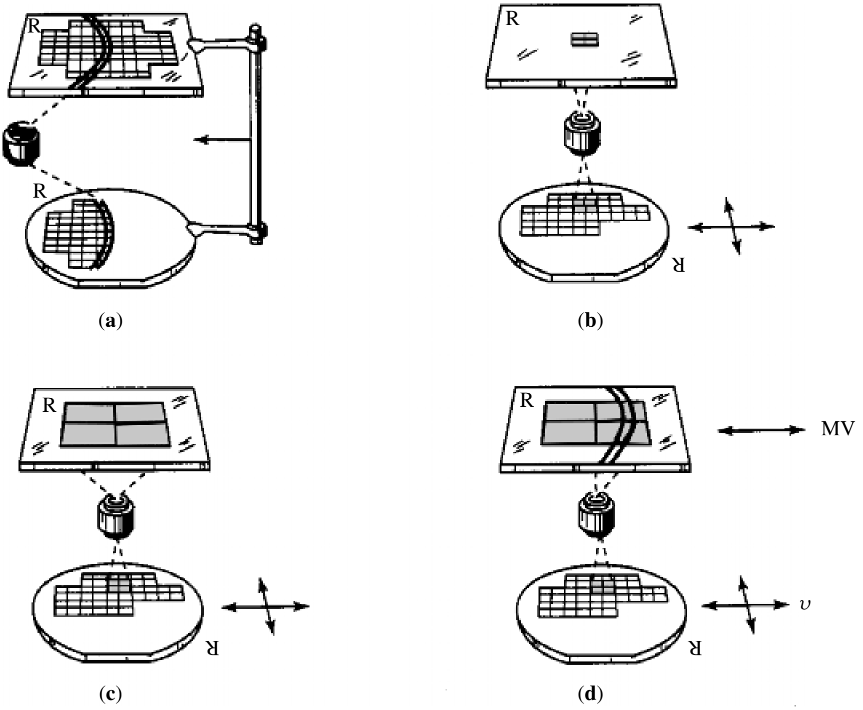


Figure 2.10. Image partitioning techniques for projection printing: (a) annual field wafer scan; (b) 1 : 1 step-and-repeat; (c) *M* : 1 step-and-repeat; and (d) *M* : 1 step-and-scan [1].

Reduction projection lithography can also print larger wafers without redesigning the stepper lens, as long as the field size (i.e., the exposure area onto the wafer) of the lens is large enough to contain one or more ICs. When the chip size exceeds the field size of the lens, further partitioning of the image on the reticle is necessary. In Figure 2.10d, the image field on the reticle can be a narrow, arc shape for $M : 1$ step-and-scan projection lithography. For the step-and-scan system, we have two-dimensional translations of the wafer with speed v, and one-dimensional translation of the mask with M times that of the wafer speed.

The resolution of a projection system is given by

$$l_m = k_1 \frac{\lambda}{\text{NA}} \tag{2.16}$$

where k_1 is a process-dependent factor and NA is the numerical aperture, which is given by

$$\text{NA} = \overline{n} \sin\theta \tag{2.17}$$

where $\overline{n}$ is the index of refraction in the image medium (usually air, where $\overline{n} = 1$) and θ is the half-angle of the cone of light converging to a point image at the

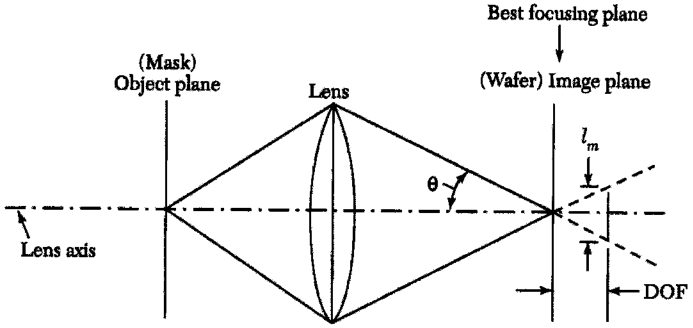


Figure 2.11. Illustration of DoF [1].

wafer, as shown in Figure 2.11. Also shown in the figure is the depth of focus (DoF), which can be expressed as

$$\text{DoF} = \frac{\pm l_m/2}{\tan\theta} \approx \frac{\pm l_m/2}{\sin\theta} = k_2\frac{\lambda}{(\text{NA})^2} \tag{2.18}$$

where k_2 is another process-dependent factor.

Equation (2.16) indicates that resolution can be improved (i.e., smaller l_m) by either reducing the wavelength, increasing NA, or both. However, Eq. (2.18) indicates that the DoF degrades much more rapidly by increasing NA than by decreasing λ. This explains the trend toward shorter-wavelength sources in optical lithography.

2.1.2.2. Masks

Masks used for semiconductor manufacturing are usually reduction reticles. The first step in maskmaking is to use a computer-aided design (CAD) system in which designers can completely describe the circuit patterns electrically. The digital data produced by the CAD system then drive a pattern generator, which is an electron-beam lithographic system (see Section 2.1.2.5) that transfers the patterns directly to electron-sensitized mask. The mask consists of a fused-silica substrate covered with a chrominum layer. The circuit pattern is first transferred to the electron-sensitized layer (electron resist), which is transferred once more into the underlying chrominum layer for the finished mask. The patterns on a mask represent one level of an IC design. The composite layout is broken into mask levels that correspond to the manufacturing process sequence, such as the isolation region on one level, the gate region on another, and so on. Typically, 15–20 different mask levels are required for a complete IC process cycle.

The standard-size mask substrate is a fused-silica plate 15 × 15 cm square, 0.6 cm thick. This size is needed to accommodate the lens field sizes for 4 : 1 or 5 : 1 optical exposure tools, whereas the thickness is required to minimize pattern placement errors due to substrate distortion. The fused-silica plate is needed for its low coefficient of thermal expansion, its high transmission at shorter

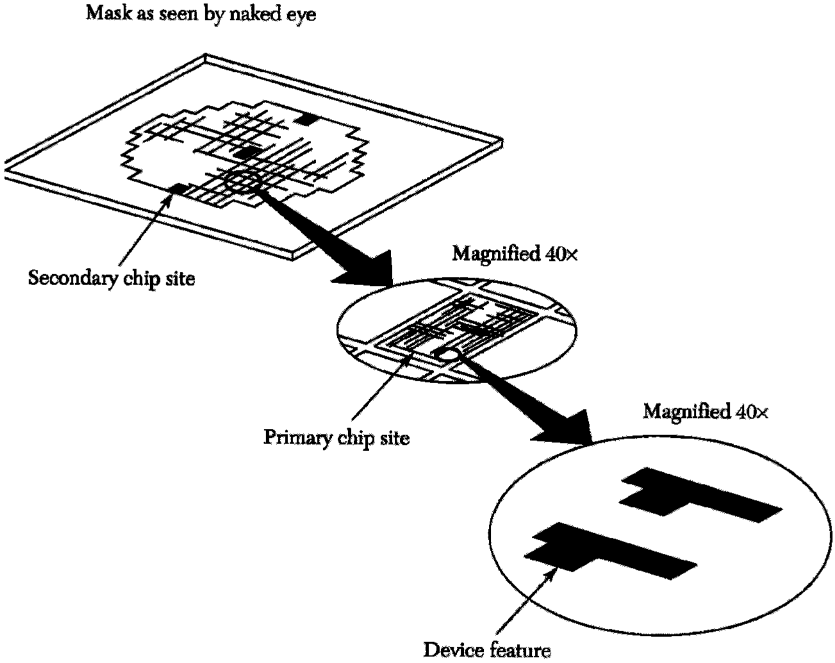


Figure 2.12. A typical IC photomask [1].

wavelengths, and its mechanical strength. Figure 2.12 shows a mask on which patterns of geometric shapes have been formed. A few secondary-chip sites, used for process evaluation, are also included in the mask.

One of the major concerns about masks is the defect density. Mask defects can be introduced during the manufacture of the mask or during subsequent lithographic processes. Even a small mask defect density has a profound effect on the final IC yield. *Yield* is defined as the ratio of good chips per wafer to the total number of chips per wafer (see Chapter 5). Inspection and cleaning of masks are important to achieve high yields on large chips. An ultraclean processing area is mandatory for photolithographic processing.

2.1.2.3. Photoresist

Photoresist is a radiation-sensitive compound that can be classified as positive or negative, depending on how they respond to radiation. For *positive resists*, the exposed regions become more soluble and thus more easily removed in the development process. The result is that the patterns formed in the positive resist are the same as those on the mask. Positive photoresists consist of three components: a photosensitive compound, a base resin, and an organic solvent. Prior to exposure, the photosensitive compound is insoluble in the developer solution. After exposure, the photosensitive compound absorbs radiation in the exposed pattern areas, changes its chemical structure, and becomes soluble in the developer solution. After development, the exposed areas are removed.

With *negative resists*, exposed regions become less soluble, and the patterns formed in the negative resist are the reverse of the mask patterns. Negative photoresists are polymers combined with a photosensitive compound. After exposure, the photosensitive compound absorbs the optical energy and converts it into chemical energy to initiate a polymer crosslinking reaction. This reaction causes crosslinking of the polymer molecules. The crosslinked polymer has a higher molecular weight and becomes insoluble in the developer solution. After development, the unexposed areas are removed. One major drawback of a negative photoresist is that in the development process, the whole resist mass swells by absorbing developer solvent. This swelling action limits the resolution of negative photoresists.

Figure 2.13a shows a typical exposure response curve and image cross section for a positive resist. The response curve describes the percentage of resist remaining after exposure and development versus the exposure energy. As the exposure energy increases, the solubility gradually increases until at a threshold energy E_T, the resist becomes completely soluble. The sensitivity of a positive resist is defined as the energy required to produce complete solubility in the exposed region. Thus, E_T corresponds to the sensitivity. In addition to E_T, a parameter γ, the contrast ratio, is defined to characterize the resist

$$\gamma \equiv \left[\ln \left(\frac{E_T}{E_1} \right) \right]^{-1} \tag{2.19}$$

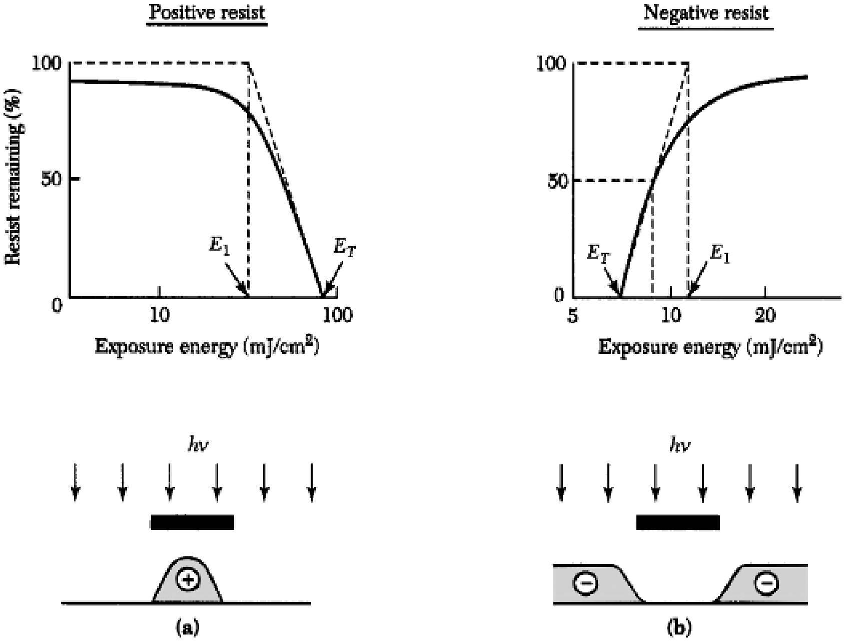


Figure 2.13. Exposure response curve and cross section of resist image after development for (a) positive photoresist and (b) negative photoresist [1].

where E_1 is the energy obtained by drawing the tangent at E_T to reach 100% resist thickness, as shown in Figure 2.13a. A larger γ implies a higher solubility of the resist with an incremental increase of exposure energy and results in sharper images.

The image cross section in Figure 2.13a illustrates the relationship between the edges of a photomask image and the corresponding edges of the resist images after development. The edges of the resist image are generally not at the vertically projected positions of the mask edges because of *diffraction*. The edge of the resist image corresponds to the position where the total absorbed optical energy equals the threshold energy E_T.

Figure 2.13b shows the exposure response curve and image cross section for a negative resist. The negative resist remains completely soluble in the developer solution for exposure energies lower than E_T. Above E_T, more of the resist film remains after development. At exposure energies twice the threshold energy, the resist film becomes essentially insoluble in the developer. The sensitivity of a negative resist is defined as the energy required to retain 50% of the original resist film thickness in the exposed region. The parameter γ is defined similarly to γ in Eq. (2.19), except that E_1 and E_T are interchanged. The image cross section for the negative resist (Figure 2.13b) is also influenced by the diffraction effect.

2.1.2.4. Pattern Transfer

Figure 2.14 illustrates the steps to transfer IC patterns from a mask to a silicon wafer that has an insulating SiO_2 layer formed on its surface. The wafer is placed in a cleanroom, which typically is illuminated with yellow light (since photoresists are not sensitive to wavelengths greater than 0.5 μm). To ensure satisfactory adhesion of the resist, adhesion promoter is then applied. The most common adhesion promoter for silicon ICs is hexamethylene–disiloxane (HMDS). After the application of this adhesion layer, the wafer is held on a vacuum spindle, and liquidous resist is applied to the center of wafer. The wafer is then rapidly accelerated up to a constant rotational speed, which is maintained for about 30 s. Spin speed is generally in the range of 1000–10,000 rpm [revolutions per minute (r/min)] to coat a uniform film about 0.5–1 μm thick, as shown in Figure 2.14a. The thickness of photoresist is correlated with its viscosity.

After spinning, the wafer is "soft-baked" (typically at 90–120°C for 60–120 s) to remove solvent from the photoresist and to increase resist adhesion to the wafer. The wafer is aligned with respect to the mask in an optical lithographic system, and the resist is exposed to ultraviolet light, as shown in Figure 2.14b. If a positive photoresist is used, the exposed resist is dissolved in the developer, as shown on the left side of Figure 2.14c. Photoresist development is usually done by flooding the wafer with the developer solution. The wafer is then rinsed and dried. After development, "postbaking" at approximately 100–180°C may be required to increase the adhesion of the resist to the substrate. The wafer is then put in an ambient that etches the exposed insulation layer but does not attack the resist, as shown in Figure 2.14d. Finally, the resist is stripped (using solvents or plasma oxidation), leaving behind an insulator image that is the same as the opaque image on the mask (left side of Figure 2.14e). For negative photoresist,

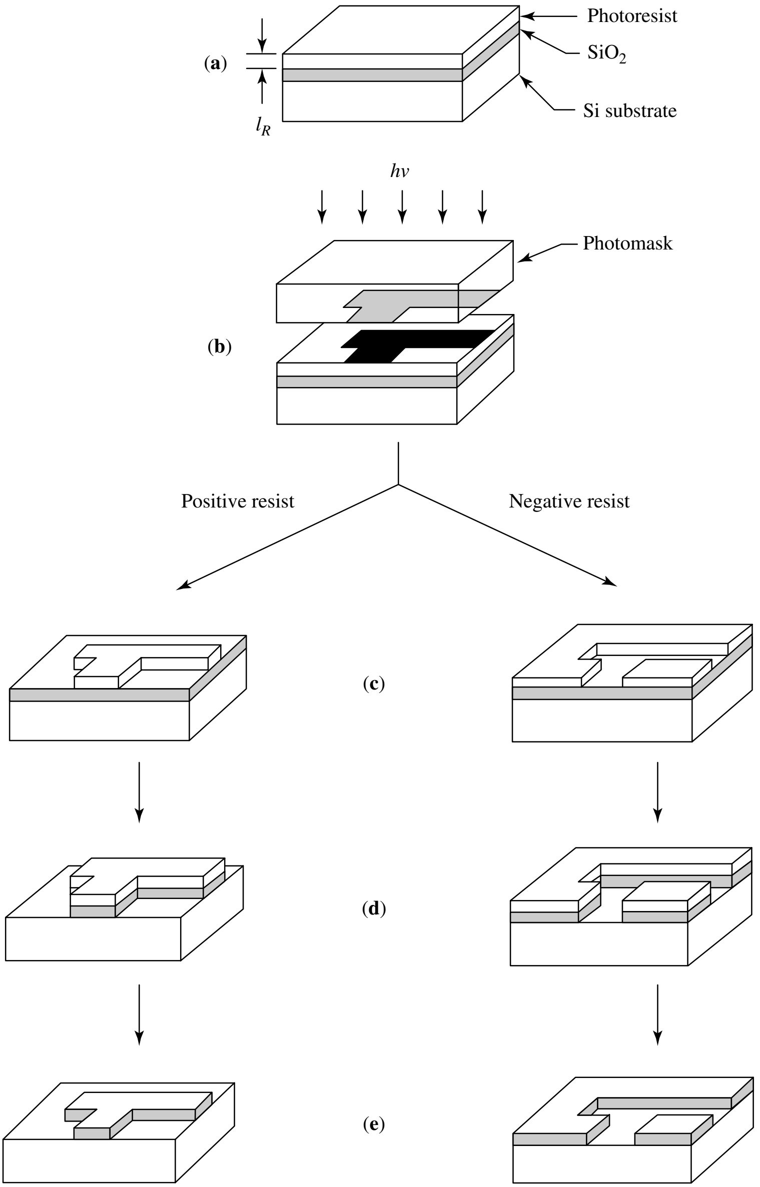


Figure 2.14. Photolithographic pattern transfer process: (a) photoresist application; (b) exposure; (c) development; (d) etching; (e) resist stripping [1].

the procedures described are also applicable, except that the unexposed areas are removed. The final insulator image (right side of Figure 2.14e) is the reverse of the opaque image on the mask.

The insulator image can be used as a mask for subsequent processing. For example, *ion implantation* (Section 2.1.3) can be done to dope the exposed semiconductor region, but not the area covered by the insulator. The dopant pattern is a duplicate of the design pattern on the photomask for a negative photoresist or is its complementary pattern for a positive photoresist. The complete circuit is fabricated by aligning the next mask in the sequence to the previous pattern and repeating the lithographic transfer process.

2.1.2.5. E-Beam Lithography

Optical lithography is so widely used because it has high throughput, good resolution, low cost, and ease of operation. However, due to deep-submicrometer IC process requirements, optical lithography has some limitations that have not yet been solved. Although we can use PSM or OPC to extend its useful lifespan, the complexity of mask production and mask inspection cannot be easily resolved. In addition, the cost of the masks is very high. Therefore, we need to find alternatives to optical lithography to process deep-submicrometer or nanometer ICs.

Electron-beam (or *e-beam*) lithography is used primarily to produce photomasks. Relatively few tools are dedicated to direct exposure of the resist by a focused electron beam without a mask. Figure 2.15 shows a schematic of an e-beam lithography system. The electron gun is a device that can generate a beam of electrons with a suitable current density. A tungsten thermionic emission cathode or single-crystal lanthanum hexaboride (LaB_6) is used for the electron gun. Condenser lenses are used to focus the electron beam to a spot size 10–25 nm in diameter. Beam blanking plates that turn the electron beam on and off, and beam deflection coils are computer-controlled and operated at MHz or higher

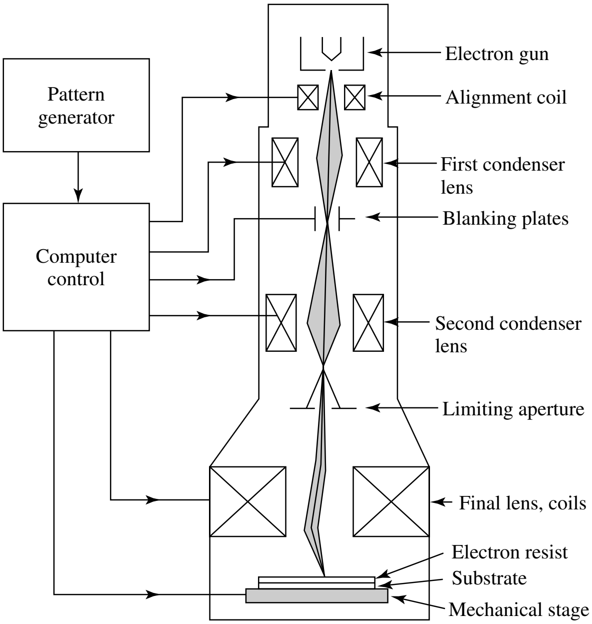


Figure 2.15. E-beam lithography system [1].

rates to direct the focused electron beam to any location in the scan field on the substrate. Because the scan field (typically 1 cm) is much smaller than the substrate diameter, a precision mechanical stage is used to position the substrate to be patterned.

The advantages of electron-beam lithography include the generation of submicrometer resist geometries, highly automated and precisely controlled operation, depth of focus greater than that available from optical lithography, and direct patterning on a semiconductor wafer without using a mask. The disadvantage is that electron-beam lithographic machines have low throughput—approximately 10 wafers per hour at less than 0.25 μm resolution. This throughput is adequate for the production of photomasks, for situations that require small numbers of custom circuits, and for design verification. However, for maskless direct writing, the machine must have the highest possible throughput, and therefore, the largest beam diameter possible consistent with the minimum device dimensions.

There are two ways to scan the focused electron beam: raster scan and vector scan. In a *raster scan* system, resist patterns are written by a beam that moves through a regular mode, vertically oriented, as shown in Figure 2.16a. The beam scans sequentially over every possible location on the mask and is blanked (turned off) where no exposure is required. All patterns on the area to be written must be

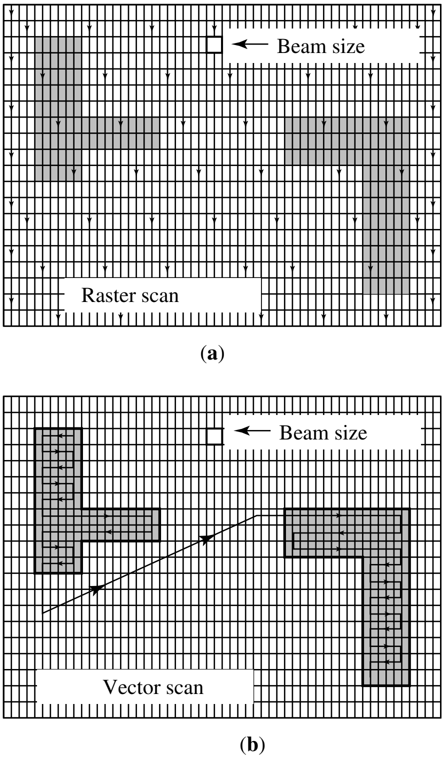


Figure 2.16. (a) Raster scan and (b) vector scan systems [1].

subdivided into individual addresses, and a given pattern must have a minimum incremental interval that is evenly divisible by the beam address size. In the *vector scan* system (Figure 2.16b), the beam is directed only to the requested pattern features and jumps from feature to feature, rather than scanning the whole chip, as in raster scan. For many chips, the average exposed region is only 20% of the chip area, which saves time.

Electron resists are polymers. The behavior of an e-beam resist is similar to that of a photoresist; that is, a chemical or physical change is induced in the resist by irradiation. This change allows the resist to be patterned. For a positive electron resist, the polymer–electron interaction causes chemical bonds to be broken (chain scission) to form shorter molecular fragments. As a result, the molecular weight is reduced in the irradiated area, which can be dissolved subsequently in a developer solution that attacks the low-molecular-weight material. Common positive electron resists include poly(methyl methacrylate) (PMMA) and poly(butene-1 sulfone) (PBS). Positive electron resists can achieve resolution of 0.1 μm or better. For a negative electron resist, the irradiation causes radiation-induced polymer linking. The crosslinking creates a complex three-dimensional structure with a molecular weight higher than that of the nonirradiated polymer. The nonirradiated resist can be dissolved in a developer solution that does not attack the high-molecular-weight material. Polyglycidylmethacrylate–coethylacrylate (COP) is a common negative electron resist. COP, like most negative photoresists, also swells during development, so the resolution is limited to about 1 μm.

While resolution is limited by diffraction of light in optical lithography, in e-beam lithography, the resolution is not impacted by diffraction (because the wavelengths associated with electrons of a few keV and higher energies are less than 0.1 nm), but by electron scattering. When electrons penetrate the resist and underlying substrate, they undergo collisions. These collisions lead to energy losses and path changes. The incident electrons spread out as they travel until either all of their energy is lost or they leave the material because of backscattering. Because of backscattering, electrons can irradiate regions several micrometers away from the center of the exposure beam. Since the dose of a resist is the sum of the irradiations from all surrounding areas, the electron-beam irradiation at one location will affect the irradiation in neighboring locations. This phenomenon is called the *proximity effect*. The proximity effect places a limit on the minimum spacings between pattern features. To correct for the proximity effect, patterns are divided into smaller segments. The incident electron dose in each segment is adjusted so that the integrated dose from all its neighboring segments is the correct exposure dose. This approach further decreases the throughput of the electron-beam system, because of the additional computer time required to expose the subdivided resist patterns.

2.1.2.6. X-Ray Lithography

X-ray lithography (XRL) is a potential candidate to succeed optical lithography for the fabrication of integrated circuits with feature sizes less than 100 nm. XRL

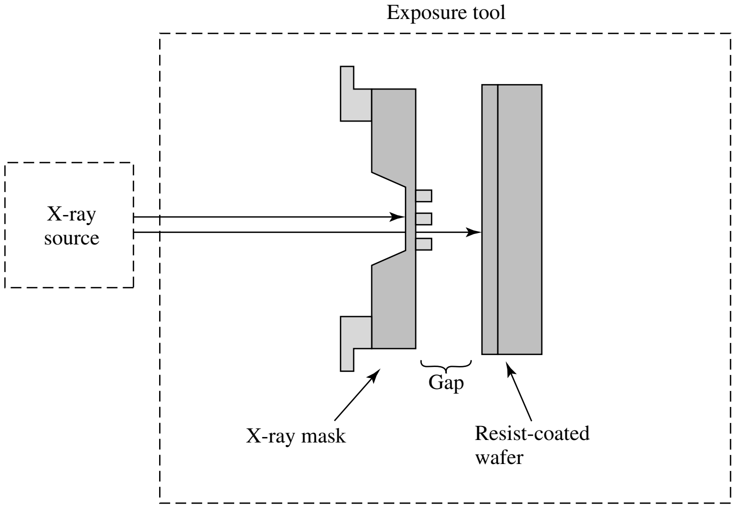


Figure 2.17. Schematic of x-ray lithography system.

uses a shadow printing method similar to optical proximity printing. Figure 2.17 shows a schematic of an XRL system. The X-ray wavelength is about 1 nm, and the printing is through a 1× mask in close proximity (10–40 μm) to the wafer. Since X-ray absorption depends on the atomic number of the material and most materials have low transparency at $\lambda \cong 1$ nm, the mask substrate must be a thin membrane (1–2 μm thick) made of low-atomic-number material, such as silicon carbide or silicon. The pattern itself is defined in a thin (~0.5 μm), relatively high-atomic-number material, such as tantalum, tungsten, gold, or one of their alloys, which is supported by the thin membrane.

Masks are the most difficult and critical element of an XRL system, and the construction of an X-ray mask is much more complicated than that of an optical photomask. To avoid absorption of the X rays between the source and mask, the exposure generally takes place in a helium environment. The X rays are produced in a vacuum, which is separated from the helium by a thin vacuum window (usually of beryllium). The mask substrate will absorb 25–35% of the incident flux and must therefore be cooled. An X-ray resist 1 μm thick will absorb about 10% of the incident flux, and there are no reflections from the substrate to create standing waves, so antireflection coatings are unnecessary.

Electron-beam resists can be used as X-ray resists because when an X ray is absorbed by an atom, the atom goes to an excited state with the emission of an electron. The excited atom returns to its ground state by emitting an X ray having a wavelength different from that of the incident X ray. This X ray is absorbed by another atom, and the process repeats. Since all the processes result in the emission of electrons, a resist film under X-ray irradiation is equivalent to one being irradiated by a large number of secondary electrons from any of the other processes.

2.1.3. Etching

As discussed in Section 2.1.2, photolithography is the process of transferring patterns to photoresist covering the surface of a semiconductor wafer. To produce circuit features, these resist patterns must be transferred into the underlying layers of the device. Pattern transfer is accomplished by an etching process that selectively removes unmasked portions of a layer.

2.1.3.1. Wet Chemical Etching

Wet chemical etching is used extensively in semiconductor processing. Prior to thermal oxidation (Section 2.1.1) or epitaxial growth (Section 2.1.5), semiconductor wafers are chemically cleaned to remove contamination that results from handling and storing. Wet chemical etching is especially suitable for blanket etches (i.e., over the whole wafer surface) of polysilicon, oxide, nitride, metals, and III–V compounds.

The mechanisms for wet chemical etching involve three essential steps, as illustrated in Figure 2.18; the reactants are transported by diffusion to the reacting surface, chemical reactions occur at the surface, and the products from the surface are removed by diffusion. Both agitation and the temperature of the etchant solution will influence the etch rate, which is the amount of film removed by etching per unit time. In IC processing, most wet chemical etches proceed by immersing the wafers in a chemical solution or by spraying the wafers with the etchant solution. For immersion etching, the wafer is immersed in the etch solution, and mechanical agitation is usually required to ensure etch uniformity and a consistent etch rate. Spray etching has gradually replaced immersion etching

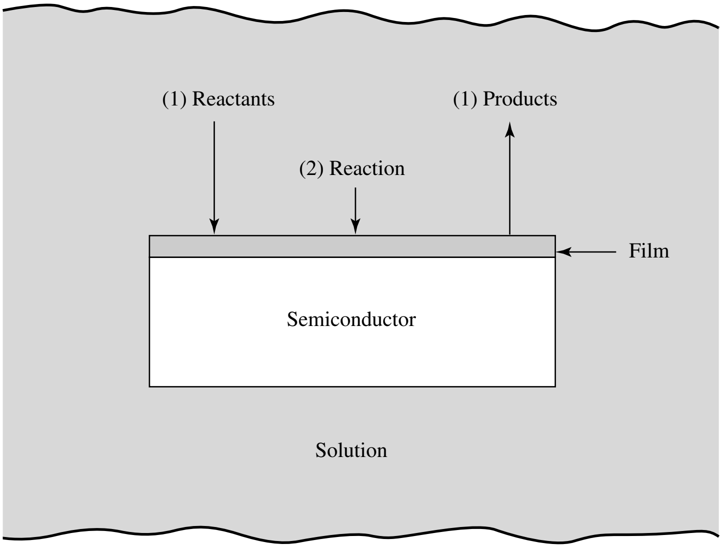


Figure 2.18. Basic mechanisms in wet chemical etching [1].

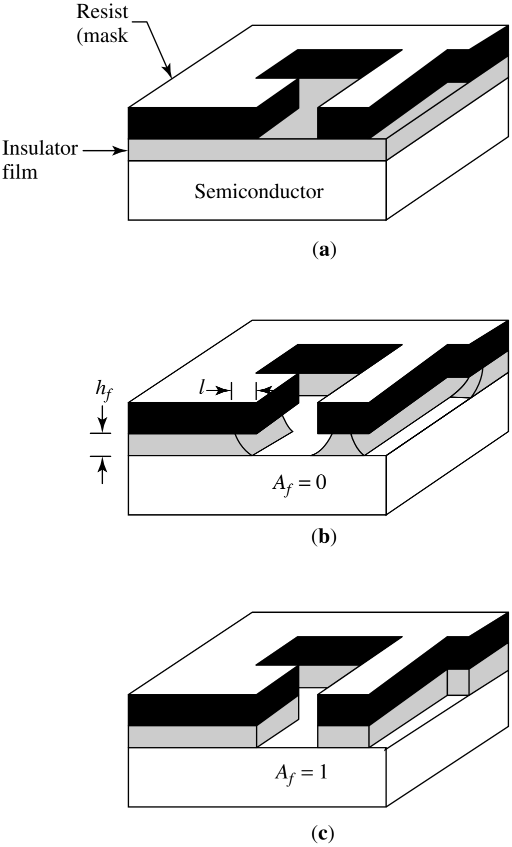
because it greatly increases the etch rate and uniformity by constantly supplying fresh etchant to the wafer surface.

In semiconductor production lines, highly uniform etch rates are important. Etch rates must be uniform across a wafer, from wafer to wafer, from run to run, and for any variations in feature sizes and pattern densities. Etch rate uniformity is given by

$$\text{Etch rate uniformity (\%)} = \frac{(\text{max. etch rate} - \text{min. etch rate})}{\text{max. etch rate} + \text{min. etch rate}} \times 100\% \qquad (2.20)$$

2.1.3.2. Dry Etching

In pattern transfer operations, a resist pattern is defined by a photolithographic process to serve as a mask for etching of its underlying layer (Figure 2.19a). Most of the layer materials (e.g., SiO_2, Si_3N_4, and deposited metals) are amorphous or polycrystalline thin films. If they are etched in a wet etchant, the etch rate



(a)

(b)

(c)

Figure 2.19. Comparison between wet and dry etching: (a) resist pattern; (b) isotropic etching; (c) anisotropic etching [1].

is generally isotropic (i.e., the lateral and vertical etch rates are the same), as illustrated in Figure 2.19b. If h_f is the thickness of the layer material and l the lateral distance etched underneath the resist mask, we can define the degree of anisotropy (A_f) by

$$A_f \equiv 1 - \frac{l}{h_f} = 1 - \frac{R_l t}{R_v t} = 1 - \frac{R_l}{R_v} \tag{2.21}$$

where t is time and R_l and R_v are the lateral and vertical etch rates, respectively. For isotropic etching, $R_l = R_v$ and $A_f = 0$.

The major disadvantage of wet etching in pattern transfer is the undercutting of the layer underneath the mask, resulting in a loss of resolution in the etched pattern. In practice, for isotropic etching, the film thickness should be about one-third or less of the resolution required. If patterns are required with resolutions much smaller than the film thickness, anisotropic etching (i.e., $1 \geq A_f > 0$) must be used. In practice, the value of A_f is chosen to be close to unity. Figure 2.19c shows the limiting case where $A_f = 1$, corresponding to $l = 0$ (or $R_l = 0$).

To achieve a high-fidelity transfer of the resist patterns required for ultra-large-scale integration (ULSI) processing, dry etching methods have been developed. Dry etching is synonymous with plasma-assisted etching, which denotes several techniques that use plasma in the form of low-pressure discharges. Dry-etch methods include plasma etching, reactive-ion etching (RIE), sputter etching, magnetically enhanced RIE (MERIE), reactive-ion-beam etching, and high-density plasma (HDP) etching.

A *plasma* is a fully or partially ionized gas composed of equal numbers of positive and negative charges and a different number of un-ionized molecules. Plasma is produced when an electric field of sufficient magnitude is applied to a gas, causing the gas to break down and become ionized. The plasma is initiated by free electrons that are released by some means, such as field emission from a negatively biased electrode. The free electrons gain kinetic energy from the electric field. In the course of their travel through the gas, the electrons collide with gas molecules and lose their energy. The energy transferred in the collision causes the gas molecules to be ionized (i.e., to free electrons). The free electrons gain kinetic energy from the field, and the process continues. Therefore, when the applied voltage is larger than the breakdown potential, a sustained plasma is formed throughout the reaction chamber.

Plasma etching is a process in which a solid film is removed by a chemical reaction with ground-state or excited-state neutral species. The process is often enhanced or induced by energetic ions generated in a gaseous discharge. Plasma etching proceeds in five steps, as illustrated in Figure 2.20: (1) the etchant species is generated in the plasma, (2) the reactant is then transported by diffusion through a stagnant gas layer to the surface, (3) the reactant is adsorbed on the surface, (4) a chemical reaction (along with physical effects such as ion bombardment) follows to form volatile compounds, and (5) the compounds are desorbed from the surface, diffused into the bulk gas, and pumped out by the vacuum system.

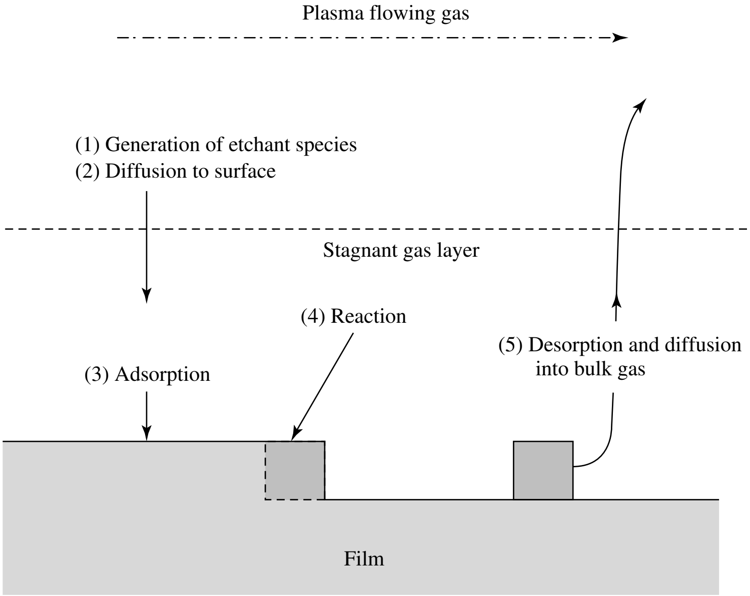


Figure 2.20. Basic steps in dry etching [1].

Plasma etching is based on the generation of plasma in a gas at low pressure. Two basic methods are used: physical methods and chemical methods. The former includes sputter etching, and the latter includes pure chemical etching. In physical etching, positive ions bombard the surface at high speed; small amounts of negative ions formed in the plasma cannot reach the wafer surface and therefore play no direct role in plasma etching. In chemical etching, neutral reactive species generated by the plasma interact with the material surface to form volatile products. Chemical and physical etch mechanisms have different characteristics. Chemical etching exhibits a high etch rate, and good selectivity (i.e., the ratio of etch rates for different materials) produces low ion bombardment–induced damage but yields isotropic profiles. Physical etching can yield anisotropic profiles, but it is associated with low etch selectivity and severe bombardment-induced damage. Combinations of chemical and physical etching give anisotropic etch profiles, reasonably good selectivity, and moderate bombardment-induced damage. An example is reactive-ion etching (RIE), which uses a physical method to assist chemical etching or creates reactive ions to participate in chemical etching.

Plasma reactor technology in the IC industry has changed dramatically since the first application of plasma processing to photoresist stripping. A reactor for plasma etching contains a vacuum chamber, pump system, power supply generators, pressure sensors, gas flow control units, and *endpoint detector* (see Chapter 3). Each etch tool is designed empirically and uses a particular combination of pressure, electrode configuration and type, and source frequency to control the two primary etch mechanisms—chemical and physical. Higher etch rates and tool automation are required for most etchers used in manufacturing.

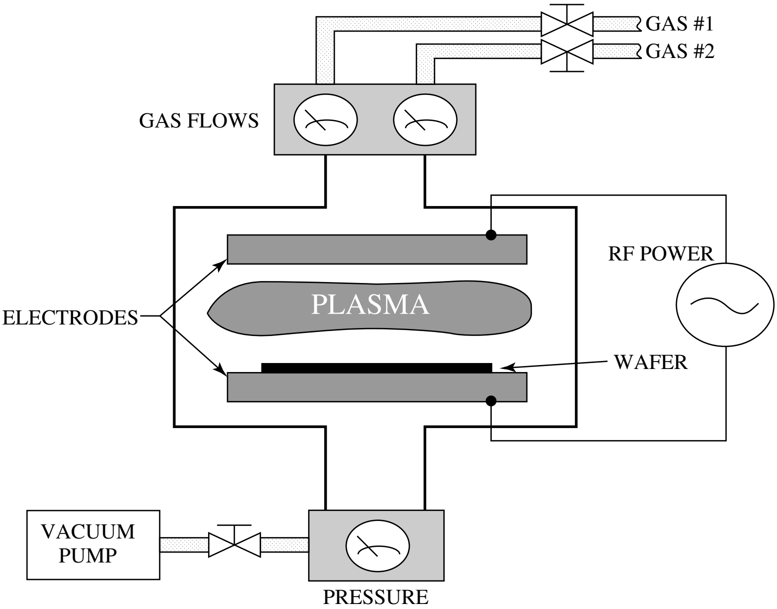


Figure 2.21. Typical reactive-ion etching system.

RIE has been extensively used in the microelectronic industry. In a parallel-plate diode system (Figure 2.21), a radio frequency (RF), capacitively coupled bottom electrode holds the wafer. This allows the grounded electrode to have a significantly larger area because it is, in fact, the chamber itself. The larger grounded area combined with the lower operating pressure (<500 mTorr) causes the wafers to be subjected to a heavy bombardment of energetic ions from the plasma as a result of the large, negative self-bias at the wafer surface. The etch selectivity of this system is relatively low compared with traditional barrel etch systems because of strong physical sputtering. However, selectivity can be improved by choosing the proper etch chemistry.

2.1.4. Doping

Impurity doping is the introduction of controlled amounts of impurities into semiconductors to change their electrical properties. Diffusion and ion implantation are the two key methods of impurity doping. Both diffusion and ion implantation are used for fabricating discrete devices and integrated circuits because these processes generally complement each other. For example, diffusion is used to form a deep junction (e.g., a twin well in CMOS), whereas ion implantation is used to form a shallow junction (e.g., a source–drain junction of a MOSFET).

Until the early 1970s, impurity doping was performed by diffusion at elevated temperatures, as shown in Figure 2.22a. In this method the dopant atoms are placed on or near the surface of the wafer by deposition from the gas phase of the dopant or by using doped-oxide sources. The doping concentration decreases

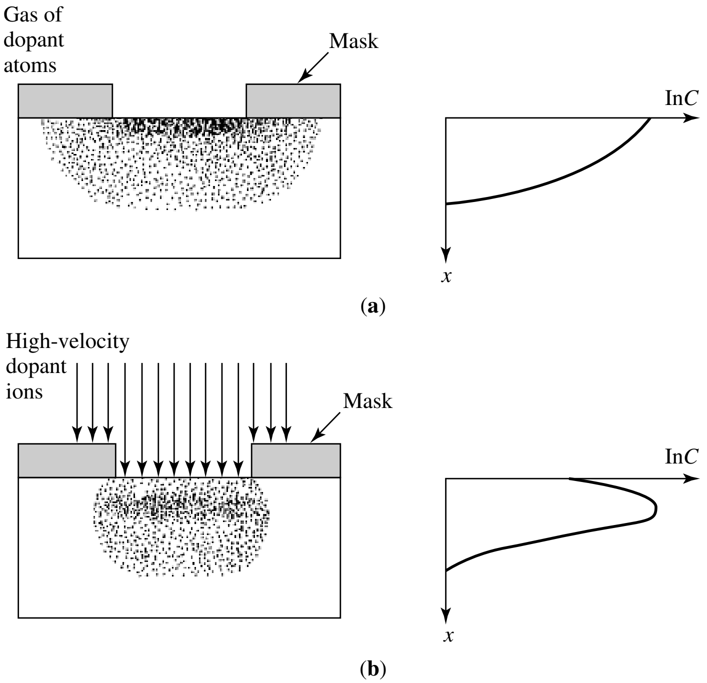


Figure 2.22. (a) Diffusion and (b) ion implantation techniques for impurity doping [1].

monotonically from the surface, and the profile of the dopant distribution is determined mainly by the temperature and the diffusion time. Since the 1970s, doping operations have been performed chiefly by ion implantation, as shown in Figure 2.22b. In this process the dopant ions are implanted into the semiconductor by means of an ion beam. The doping concentration has a peak distribution inside the semiconductor, and the profile of the dopant distribution is determined mainly by the ion mass and the implanted ion energy.

2.1.4.1. Diffusion

Diffusion of impurities is accomplished by placing semiconductor wafers in a carefully controlled, high-temperature quartz-tube furnace and passing a gas mixture that contains the desired dopant through it. The number of dopant atoms that diffuse into the semiconductor is related to the partial pressure of the dopant impurity in the gas mixture. For diffusion in silicon, boron is the most popular dopant for introducing a *p*-type impurity, whereas arsenic and phosphorus are used extensively as *n*-type dopants. These dopants can be introduced in several ways, including solid sources (e.g., BN for boron, As_2O_3 for arsenic, and P_2O_5 for phosphorus), liquid sources (BBr_3, $AsCl_3$, and $POCl_3$), and gaseous sources (B_2H_6, AsH_3, and PH_3). However, liquid sources are most commonly used. A schematic diagram of the furnace and gas flow arrangement for a liquid source is shown in Figure 2.23. This arrangement is similar to that used for thermal

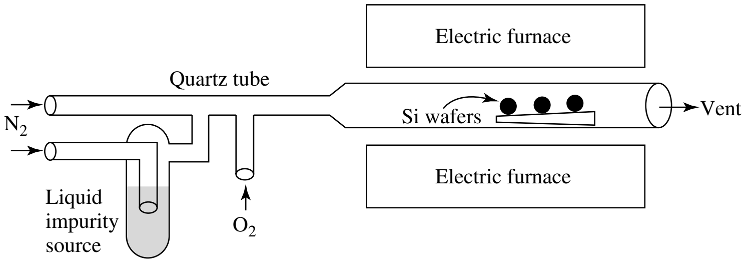


Figure 2.23. Schematic of a diffusion system [1].

oxidation. An example of the chemical reaction for phosphorus diffusion using a liquid source is

$$4POCl_3 + 3O_2 \rightarrow 2P_2O_5 + 6Cl_2 \uparrow \tag{2.22}$$

The P_2O_5 is then reduced to phosphorus by silicon according to

$$2P_2O_5 + 5Si \rightarrow 4P + 5SiO_2 \tag{2.23}$$

and the phosphorus is released and diffuses into the silicon, and Cl_2 is vented.

Diffusion in a semiconductor is the atomic movement of the diffusant (dopant atoms) in the crystal lattice by vacancies or interstitials. The diffusion process is similar to that of charge carriers (electrons and holes). Let a flux F be defined as the number of dopant atoms passing through a unit area in a unit time and C as the dopant concentration per unit volume. Then

$$F = -D\frac{\partial C}{\partial x} \tag{2.24}$$

where the proportionality constant D is the *diffusion coefficient* or *diffusivity*. Note that the basic driving force of the diffusion process is the concentration gradient dC/dx. The flux is proportional to the concentration gradient, and the dopant atoms will move (diffuse) away from a high-concentration region toward a lower-concentration region.

If Eq. (2.24) is substituted into the one-dimensional continuity equation under the condition that no materials are formed or consumed in the host semiconductor, the result is

$$\frac{\partial C}{\partial t} = -\frac{\partial F}{\partial x} = \frac{\partial}{\partial x}\left(D\frac{\partial C}{\partial x}\right) \tag{2.25}$$

When the concentration of the dopant atoms is low, the diffusion coefficient can be considered to be independent of doping concentration, and Eq. (2.25) becomes

$$\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2} \tag{2.26}$$

Equation (2.26) is often referred to as *Fick's diffusion equation* or *Fick's law*. Note that the diffusion coefficient varies with temperature. Over the temperature ranges commonly used in semiconductor manufacturing, the diffusion coefficient can be expressed as

$$D = D_0 \exp\left(\frac{-E_a}{kT}\right) \tag{2.27}$$

where D_0 is the diffusion coefficient in cm^2/s extrapolated to infinite temperature and E_a is the activation energy in eV. The values of E_a are typically found to be between 0.5 and 2 eV for interstitial diffusion. For vacancy diffusion, E_a is much larger, usually between 3 and 5 eV.

The diffusion profile of the dopant atoms is dependent on the initial and boundary conditions. There are two important cases to consider: (1) constant-surface-concentration diffusion and (2) constant-total-dopant diffusion. In the first case, impurity atoms are transported from a vapor source onto the semiconductor surface and diffused into the semiconductor wafers. The vapor source maintains a constant level of surface concentration during the entire diffusion period. In the second case, a fixed amount of dopant is deposited onto the semiconductor surface and is subsequently diffused into the wafers.

For case 1, the initial condition at $t = 0$ is

$$C(x, 0) \tag{2.28}$$

which indicates that the dopant concentration in the host semiconductor is initially zero. The boundary conditions are

$$C(0, t) = C_s \tag{2.29a}$$

$$C(\infty, t) = 0 \tag{2.29b}$$

where C_s is the surface concentration (at $x = 0$), which is independent of time. The second boundary condition states that at long distances from the surface there are no impurity atoms.

The solution of Fick's equation that satisfies the initial and boundary conditions is

$$C(x, t) = C_s \operatorname{erfc}\left(\frac{x}{2\sqrt{Dt}}\right) \tag{2.30}$$

where *erfc* is the complementary error function and $\sqrt{Dt}$ is the diffusion length. The diffusion profile for the constant surface concentration condition is shown in Figure 2.24a, where, on both linear (upper) and logarithmic (lower) scales, the normalized concentration as a function of depth for three values of the diffusion length corresponding to three consecutive diffusion times and a fixed D for a given diffusion temperature are plotted. Note that as the time progresses, the dopant penetrates deeper into the semiconductor.

The total number of dopant atoms per unit area of the semiconductor is given by

$$Q(t) = \int_0^\infty C(x, t)dx \tag{2.31}$$

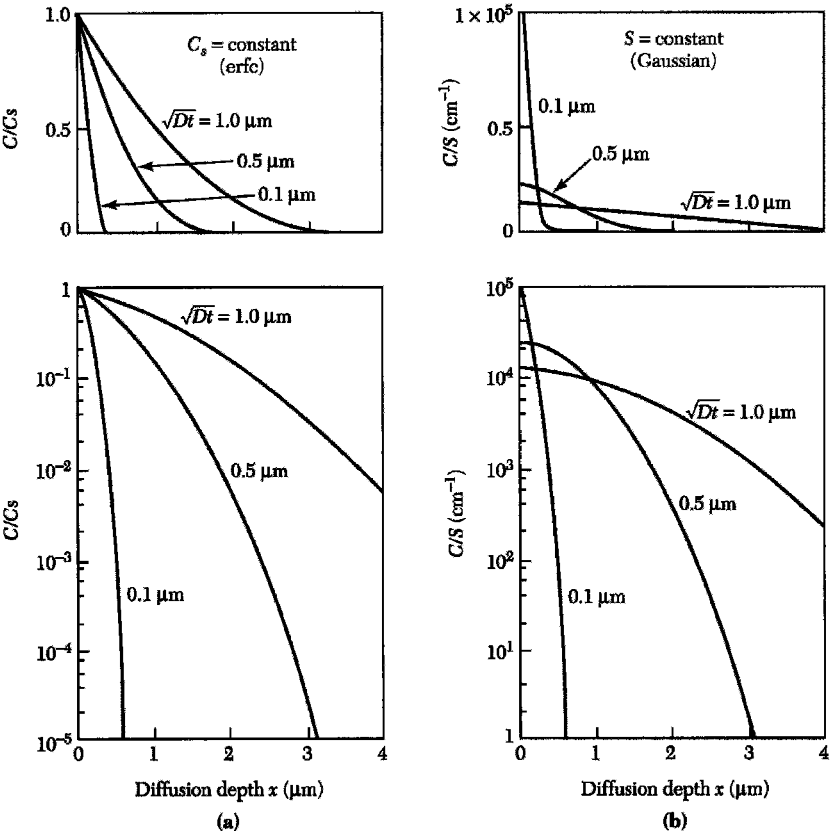


Figure 2.24. Diffusion profiles: (a) normalized erfc versus distance for successive diffusion times; (b) normalized Gaussian function versus distance [1].

Substituting Eq. (2.30) into Eq. (2.31) yields

$$Q(t) = \frac{2}{\sqrt{\pi}} C_s \sqrt{Dt} \cong 1.13 C_s \sqrt{Dt} \tag{2.32}$$

The quantity $Q(t)$ represents the area under one of the diffusion profiles of the linear plot in Figure 2.24a. These profiles can be approximated by triangles with height C_s and base $2\sqrt{Dt}$. This leads to $Q(t) \cong C_s\sqrt{Dt}$, which is close to the exact result obtained from Eq. (2.32).

Now consider constant total dopant diffusion. For this case, a fixed (or constant) amount of dopant is deposited onto the semiconductor surface in a thin layer, and the dopant subsequently diffuses into the semiconductor. The initial condition is the same as in Eq. (2.28). The boundary conditions are

$$\int_0^\infty C(x,t) = S \tag{2.33a}$$

$$C(\infty, t) = 0 \tag{2.33b}$$

where S is the total amount of dopant per unit area. The solution of the diffusion equation that satisfies these conditions is

$$C(x,t) = \frac{S}{\sqrt{\pi D t}} \exp\left(-\frac{x^2}{4Dt}\right) \tag{2.34}$$

This expression is the Gaussian distribution. Since the dopant will move into the semiconductor as time increases, in order to keep the total dopant S constant, the surface concentration must decrease. This is the case, since the surface concentration is given by Eq. (2.34) with $x = 0$:

$$C(x,t) = \frac{S}{\sqrt{\pi D t}} \tag{2.35}$$

Figure 2.24b shows the dopant profile for a Gaussian distribution where the normalized concentration (C/S) as a function of the distance for three increasing diffusion lengths is plotted. Note the reduction of the surface concentration as the diffusion time increases.

In IC processing, a two-step diffusion process is commonly used, in which a *predeposition* diffused layer is first formed under the constant-surface-concentration condition (case 1, above). This step is followed by a *drive-in* diffusion (also called *redistribution* diffusion) under constant total dopant conditions. For most practical cases, the diffusion length $\sqrt{Dt}$ for the predeposition diffusion is much smaller than the diffusion length for the drive-in diffusion. Therefore, the predeposition profile can be considered a delta function at the surface, and the extent of the penetration of the predeposition profile can be regarded as negligibly small compared with that of the final profile that results from the drive-in step.

2.1.4.2. Ion Implantation

As discussed above, diffusion and ion implantation are the two key methods of impurity doping. Since the early 1970s, many doping operations have been performed by ion implantation, which is shown in Figure 2.22b. In this process the energetic dopant ions are implanted into the semiconductor by means of an ion beam. The doping concentration has a peak distribution inside the semiconductor and the profile of the dopant distribution is determined mainly by the ion mass and energy.

Implantation energies are typically between 1 keV and 1 MeV, resulting in ion distributions with average depths ranging from 10 nm to 10 μm. Ion doses vary from 10^{12} ions/cm^2 for threshold voltage adjustment in MOSFETs to 10^{18} ions/cm^2 for the formation of buried insulating layer. Note that the dose is expressed as the number of ions implanted into 1 cm^2 of the semiconductor surface area. The main advantages of ion implantation are its more precise control and reproducibility of impurity dopings and its lower processing temperature compared with those of the diffusion process.

Figure 2.25 shows schematically a medium-energy ion implantor. The ion source has a heated filament to break up source gas, such as BF_3 or AsH_3, into

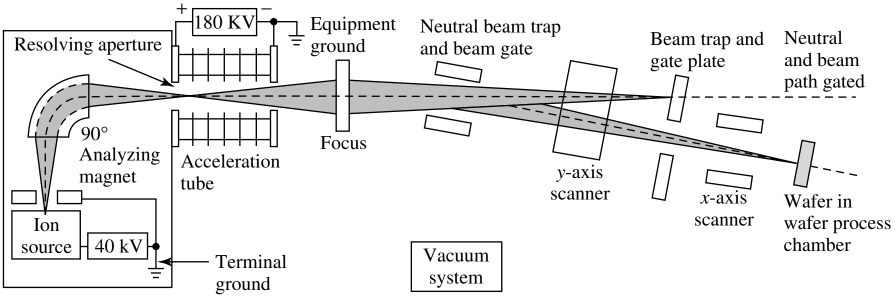


Figure 2.25. Schematic of ion implantor [1].

charged ions (B^+ or As^+). An extraction voltage (around 40 kV) causes the charged ions to move out of the ion source chamber into a mass analyzer. The magnetic field of the analyzer is chosen such that only ions with the desired mass : charge ratio can travel through it without being filtered. The selected ions then enter the acceleration tube, where they are accelerated to the implantation energy as they move from high voltage to ground. Apertures ensure that the ion beam is well collimated. The pressure in the implantor is kept below 10^{-4} Pa to minimize ion scattering by gas molecules. The ion beam is then scanned over the wafer surface using electrostatic deflection plates and is implanted into the semiconductor substrate.

The energetic ions lose their energies through collision with electrons and nuclei in the substrate and finally come to rest at some depth within the lattice. The average depth can be controlled by adjusting the acceleration energy. The dopant dose can be controlled by monitoring the ion current during implantation. The principal side effect is the disruption or damage of the semiconductor lattice due to ion collisions. Therefore, a subsequent annealing treatment is needed to remove these damages.

The total distance that an ion travels in coming to rest is called its *range* (R) and is illustrated in Figure 2.26a. The projection of this distance along the axis of incidence is called the *projected range* (R_p). Because the number of collisions per unit distance and the energy lost per collision are random variables, there will be a spatial distribution of ions having the same mass and the same initial energy. The statistical fluctuations in the projected range are called the *projected straggle* (σ_p). There is also a statistical fluctuation along an axis perpendicular to the axis of incidence, which is called the *lateral straggle* ($\sigma_\perp$).

Figure 2.26b shows the ion distribution. Along the axis of incidence, the implanted impurity profile can be approximated by a Gaussian distribution function

$$n(x) = \frac{S}{\sqrt{2\pi}\sigma_p} \exp\left[-\frac{(x - R_p)^2}{2\sigma_p^2} \right] \tag{2.36}$$

where S is the ion dose per unit area. This equation is similar to Eq. (2.34) for constant total dopant diffusion, except that the quantity $4Dt$ is replaced by $2\sigma_p^2$

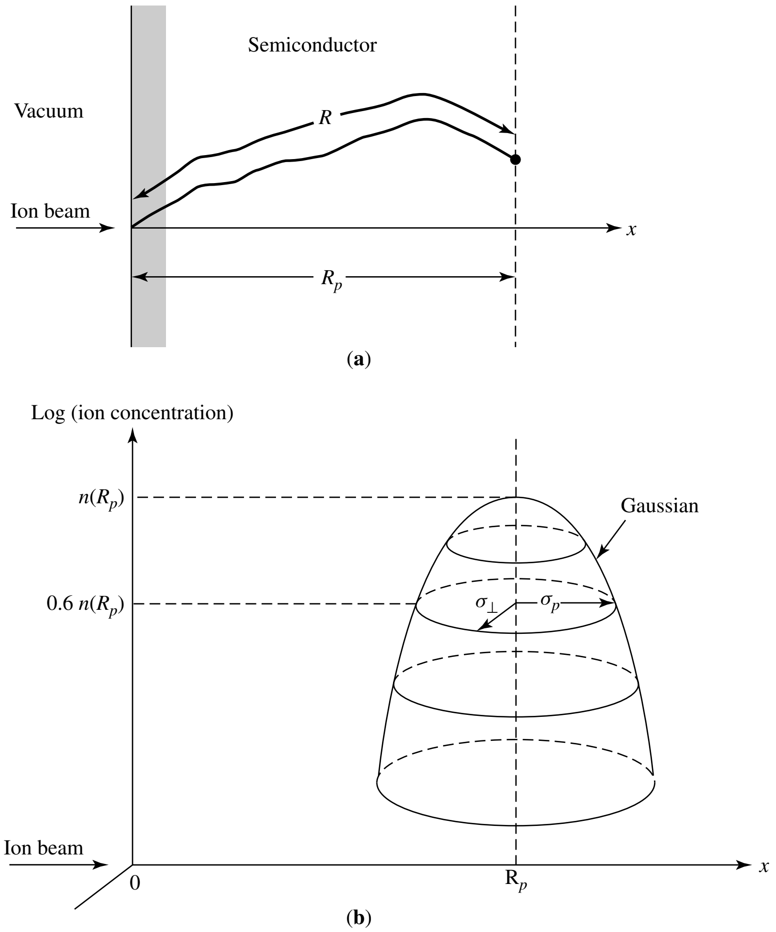


Figure 2.26. (a) Ion range and projected range; (b) two-dimensional distribution of implanted ions [1].

and the distribution is shifted along the x axis by R_p. Thus, for diffusion, the maximum concentration is at $x = 0$, whereas for ion implantation the maximum concentration is at the projected range. The ion concentration is reduced by 40% from its peak value at $(x - R_p) = \pm\sigma_p$, by one decade at $\pm 2\sigma_p$, by two decades at $\pm 3\sigma_p$, and by five decades at $\pm 4.8\sigma_p$.

2.1.5. Deposition

Many different types of thin films are used to manufacture integrated circuits, including thermal oxides, dielectric layers, epitaxial layers, polycrystalline silicon, and metal films. This section addresses two of the various techniques for depositing such films: physical vapor deposition and chemical vapor deposition.

2.1.5.1. Physical Vapor Deposition

The most common methods of physical vapor deposition (PVD) of metals are evaporation, electron-beam evaporation, plasma spray deposition, and sputtering. Metals and metal compounds can be deposited by PVD. Evaporation occurs when a source material is heated above its melting point in an evacuated chamber. The evaporated atoms then travel at high velocity in straight-line trajectories. The source can be melted by resistance heating, by radio frequency (RF) heating, or with a focused electron beam (or e-beam). Evaporation and e-beam evaporation were used extensively in earlier generations of integrated circuits, but they have been replaced by sputtering for modern ICs.

In ion-beam sputtering, a source of ions is accelerated toward the target and impinges on its surface. Figure 2.27a shows a standard sputtering system. The sputtered material deposits on a wafer that is placed facing the target. The ion current and energy can be independently adjusted. Since the target and wafer are placed in a chamber that has lower pressure, more target material and less contamination are transferred to the wafer.

One method to increase the deposition rate in sputtering is to use a third electrode that provides more electrons for ionization. Another method is to use a magnetic field, such as in *electron cyclotron resonance* (ECR) systems, to capture and spiral electrons, increasing their ionizing efficiency in the vicinity of the sputtering target. This technique, referred to as *magnetron sputtering*, has found widespread applications for the deposition of aluminum and its alloys at a rate that can approach 1 μm/min.

Long-throw sputtering is another technique used to control the angular distribution. Figure 2.27b shows a long-throw sputtering system. In standard sputtering configurations, there are two primary reasons for a wide angular distribution of incident flux at the surface: (1) the use of a small target to substrate separation d_{ts} and (2) scattering of the flux by the working gas as the flux travels from the target to the substrate. These two factors are linked because a small d_{ts} is needed to achieve good throughput, uniformity, and film properties when there is substantial gas scattering. A solution to this problem is to sputter at very low

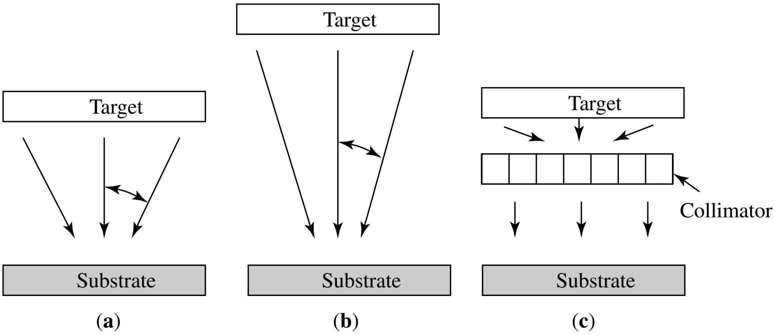


Figure 2.27. (a) Standard sputtering; (b) long-throw sputtering; (c) sputtering through a collimator [1].

pressures, a capability that has been developed using a variety of systems, which can sustain the magnetron plasma under more rarefied conditions. These systems allow for sputtering at working pressures of less than 0.1 Pa. At these pressures, gas scattering is less important, and the target–substrate distance can be greatly increased. From a simple geometric argument, this allows the angular distribution to be greatly narrowed, which permits more deposition at the bottom of high-aspect features such as contact holes.

Contact holes with large aspect ratio are difficult to fill with material, mainly because scattering events cause the top opening of the hole to seal before appreciable material has deposited on its floor. This problem can be overcome by collimating the sputtered atoms by placing an array of collimating tubes just above the wafer to restrict the depositing flux to normal ±5°. Sputtering with a collimator is shown in Figure 2.27c. Atoms whose trajectory is more than 5° from normal are deposited on the inner surface of the collimators.

2.1.5.2. Chemical Vapor Deposition

Chemical vapor deposition (CVD), also known as *vapor-phase epitaxy* (VPE), is a process whereby an epitaxial layer is formed by a chemical reaction between gaseous compounds. CVD can be performed at atmospheric pressure (APCVD) or at low pressure (LPCVD). Figure 2.28 shows three common susceptors for epitaxial growth. Note that the geometric shape of the susceptor provides the name for the reactor: horizontal, pancake, and barrel susceptors—all made from graphite blocks. Susceptors in epitaxial reactors are analogous to crucibles in the

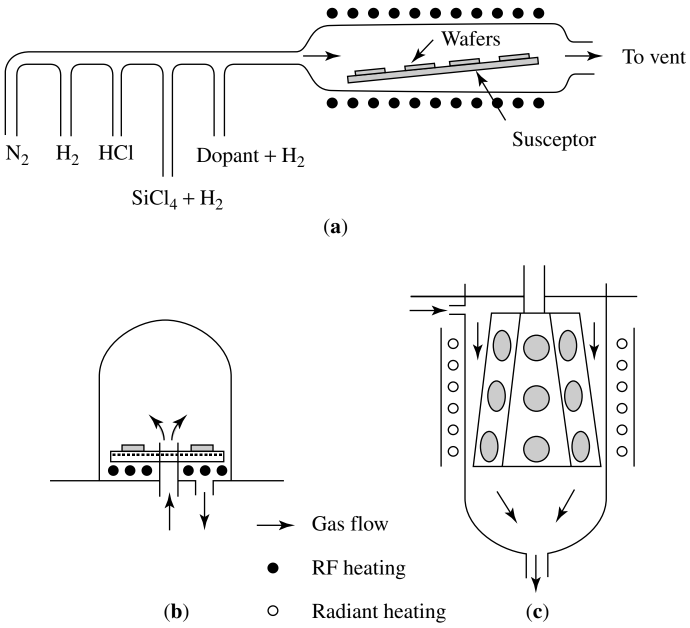


Figure 2.28. Common susceptors for CVD: (a) horizontal; (b) pancake; (c) barrel [1].

crystal growing furnaces. Not only do they mechanically support the wafer, but in induction-heated reactors, they also serve as the source of thermal energy for the reaction. The mechanism of CVD involves a number of steps: (1) the reactants (gases and dopants) are transported to the substrate region; (2) they are transferred to the substrate surface, where they are adsorbed; (3) a chemical reaction occurs, catalyzed at the surface, followed by growth of the epitaxial layer; (4) the gaseous products are desorbed into the main gas stream; and (5) the reaction products are transported out of the reaction chamber.

CVD is attractive for metallization because it offers coatings that are conformal, has good step coverage, and can coat a large number of wafers at a time. The basic CVD setup is the same as that used for deposition of dielectrics and polysilicon (see Figure 1.14). Low-pressure CVD (LPCVD) is capable of producing conformal step coverage over a wide range of topographical profiles, often with lower electrical resistivity than that from PVD. One of the major new applications of CVD metal deposition for integrated circuit production is in the area of refractory metal deposition. For example, tungsten's low electrical resistivity (5.3 $\mu\Omega \cdot$ cm) and refractory nature make it a desirable metal for use in IC fabrication.

2.1.6. Planarization

The development of chemical–mechanical polishing (CMP) has become important for multilevel interconnection technology because it is the only method that allows global planarization (i.e., a flat surface across the whole wafer). It also offers other advantages, including reduced defect density and the avoidance of plasma damage (which would occur in an RIE-based planarization system).

The CMP process consists of moving the sample surface against a pad that carries slurry between the sample surface and the pad. Abrasive particles in the slurry cause mechanical damage on the sample surface, loosening the material for enhanced chemical attack or fracturing off the pieces of surface into a slurry where they dissolve or are swept away. The process is tailored to provide an enhanced material removal rate from high points on surfaces. Mechanical grinding alone may theoretically achieve the desired planarization, but is undesirable because of extensive associated damage to the material surface. There are three main parts of the process: (1) the surface to be polished; (2) the pad, which is the key medium enabling the transfer of mechanical action to the surface being polished; and (3) the slurry, which provides both chemical and mechanical effects. Figure 2.29 shows a typical CMP setup.

2.2. PROCESS INTEGRATION

An integrated circuit is an ensemble of active (e.g., transistors) and passive devices (e.g., resistors, capacitors, and inductors) formed on and within a single-crystal semiconductor substrate and interconnected by a metallization pattern.

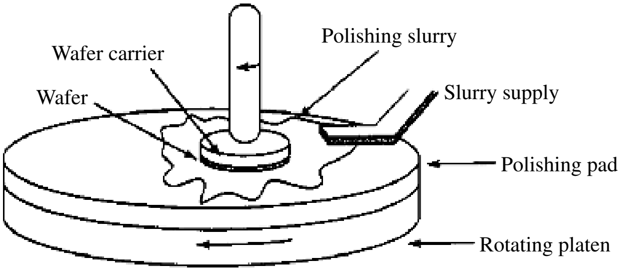


Figure 2.29. CMP schematic [1].

ICs have enormous advantages over discrete devices, including (1) reduction of the interconnection parasitics, (2) full utilization of a semiconductor wafer's area, and (3) drastic reduction in processing cost. In this section, we discuss the manner in which the basic processes described in previous portions of this chapter are combined to fabricate ICs. We consider three major IC technologies associated with two transistor families (viz., bipolar junction transistors and metal–oxide–semiconductor field-effect transistors, or MOSFETs): bipolar, CMOS, and BiCMOS. In addition, we will discuss the packaging of ICs by various techniques.

Figure 2.30 illustrates the interrelationship between the major process steps used for IC fabrication. Polished wafers with a specific resistivity and orientation are used as the starting material. The film formation steps include thermally grown

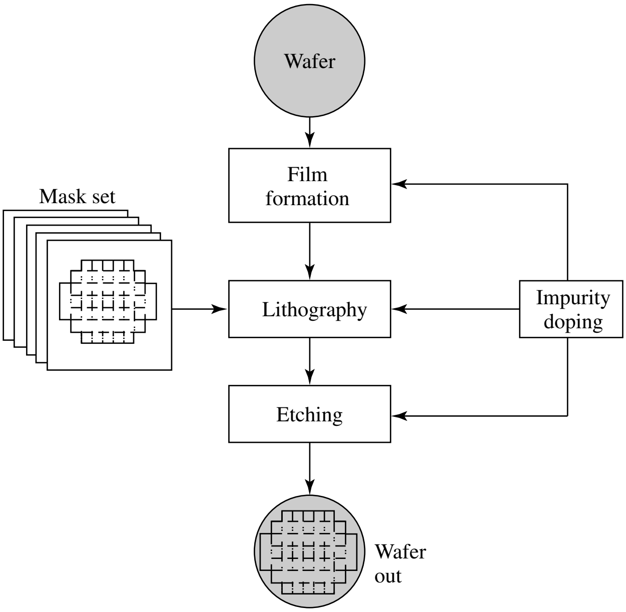


Figure 2.30. Schematic diagram of IC fabrication [1].

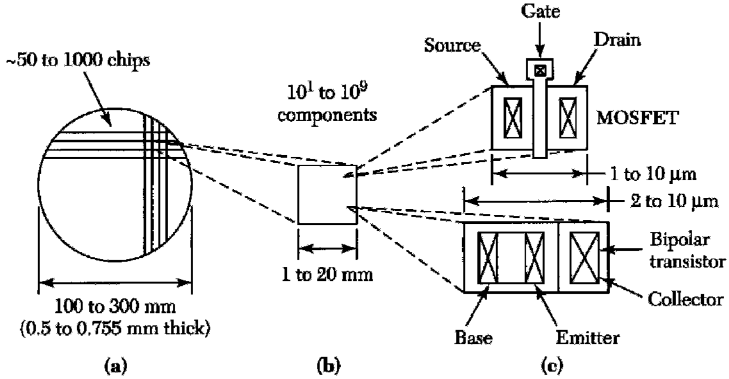


Figure 2.31. (a) Semiconductor wafer; (b) IC chip; (c) MOSFET and bipolar transistor [1].

oxide films (Section 2.1.1), deposited polysilicon, dielectric, and metal films (Section 2.1.5). Film formation is often followed by lithography (Section 2.1.2) or impurity doping (Section 2.1.4). Lithography is generally followed by etching (Section 2.1.3), which in turn is often followed by another impurity doping or film formation. The final IC is made by sequentially transferring the patterns from each mask, level by level, onto the surface of the semiconductor wafer.

After processing, each wafer contains hundreds of identical rectangular chips (or dies), typically between 1 and 20 mm on each side, as shown in Figure 2.31a. The chips are separated by sawing or laser cutting. Figure 2.31b shows a separated chip. Schematic top views of a single MOSFET and a single bipolar transistor are shown in Figure 2.31c to give some perspective of the relative size of a component in an IC chip. Prior to chip separation, each chip is electrically tested. Good chips are selected and packaged to provide an appropriate thermal, electrical, and interconnection environment for electronic applications.

2.2.1. Bipolar Technology

The majority of bipolar transistors used in ICs are of the $n-p-n$ type because the higher mobility of minority carriers (electrons) in the base region results in higher-speed performance than can be obtained with $p-n-p$ types. Figure 2.32 shows a perspective view of an $n-p-n$ bipolar transistor in which lateral isolation is provided by oxide walls and vertical isolation is provided by the n^+-p junction. The lateral oxide isolation approach reduces not only the device size but also the parasitic capacitance because of the smaller dielectric constant of silicon dioxide (3.9, compared with 11.9 for silicon).

For an $n-p-n$ bipolar transistor, the starting material is a p-type, lightly doped ($\sim 10^{15}$ cm^{-3}), $\langle 111 \rangle$- or $\langle 100 \rangle$-oriented, polished silicon wafer. Because the junctions are formed inside the semiconductor, the choice of crystal orientation is not as critical as for MOS devices (see Section 2.2.2). The first step is to form a buried layer. The main purpose of this layer is to minimize the series resistance of the collector. A thick oxide (0.5–1 μm) is thermally grown

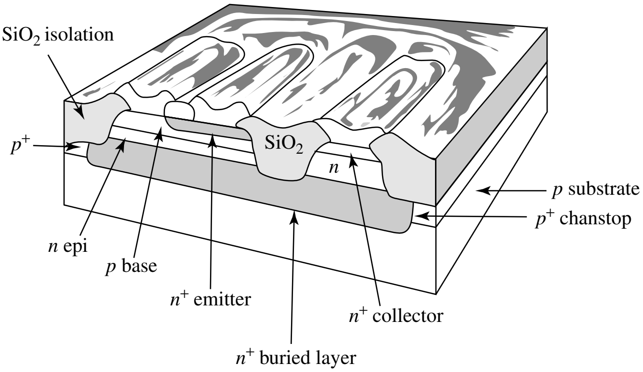


Figure 2.32. Oxide-isolated bipolar transistor [1].

on the wafer, and a window is then opened in the oxide. A precisely controlled amount of low-energy arsenic ions (~30 keV, ~10^{15} cm^{-2}) is implanted into the window region to serve as a predeposit (Figure 2.33a). Next, a high-temperature (~1100°C) drive-in step forms the n^+-buried layer, which has a typical sheet resistance of 20 Ω/□.

The second step is to deposit an n-type epitaxial layer. The oxide is removed and the wafer is placed in an epitaxial reactor for epitaxial growth. The thickness and the doping concentration of the epitaxial layer are determined by the ultimate use of the device. Analog circuits (with their higher voltages for amplification) require thicker layers (~10 μm) and lower dopings (~5×10^{15} cm^{-3}), whereas digital circuits (with their lower voltages for switching) require thinner layers (~3 μm) and higher dopings (~2×10^{16} cm^{-3}). Figure 2.33b shows a cross-sectional view of the device after the epitaxial process.

The third step is to form the lateral oxide isolation region. A thin oxide pad (~50 nm) is thermally grown on the epitaxial layer, followed by a silicon nitride deposition (~100 nm). If nitride is deposited directly onto the silicon without the thin oxide pad, the nitride may cause damage to the silicon surface during subsequent high-temperature steps. Next, the nitride–oxide layers and about half of the epitaxial layer are etched using a photoresist as mask (Figures 2.33c and 2.33d). Boron ions are then implanted into the exposed silicon areas (Figure 2.33d).

The photoresist is removed, and the wafer is placed in an oxidation furnace. Since the nitride layer has a very low oxidation rate, thick oxides will be grown only in the areas not protected by the nitride layer. The isolation oxide is usually grown to a thickness such that the top of the oxide becomes coplanar with the original silicon surface to minimize the surface topography. This oxide isolation process is called *local oxidation of silicon* (LOCOS). Figure 2.34a shows the cross section of the isolation oxide after removal of the nitride layer. Because of segregation effects, most of the implanted boron ions are pushed underneath the isolation oxide to form a p^+ layer. This is called the p^+*-channel stop* (or *chanstop*), because the high concentration of p-type semiconductor will prevent surface inversion and eliminate possible high-conductivity paths (or channels) among neighboring buried layers.

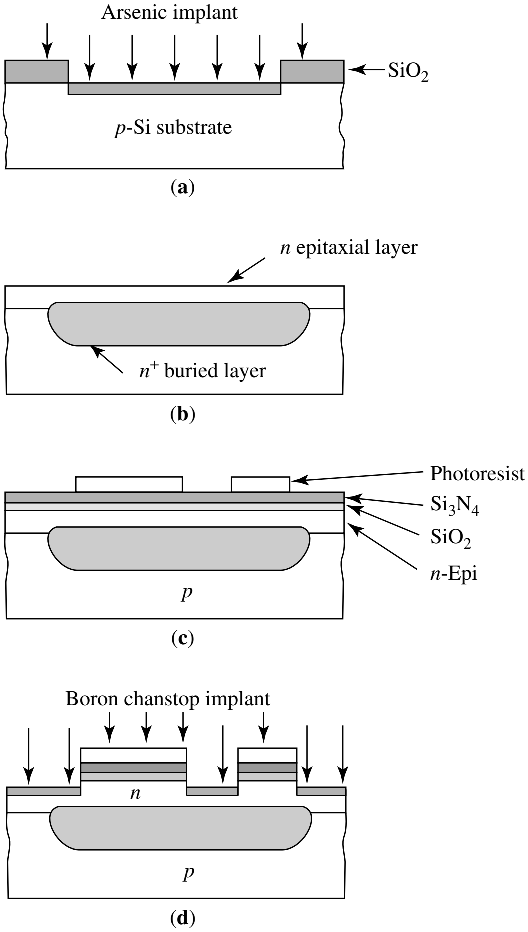


Figure 2.33. Cross-sectional views of bipolar transistor fabrication: (a) buried-layer implantation; (b) epitaxial layer; (c) photoresist mask; (d) channel-stop layer [1].

The fourth step is to form the base region. A photoresist is used as a mask to protect the right half of the device. Then, boron ions ($\sim 10^{12}$ cm^{-2}) are implanted to form the base regions, as shown in Figure 2.34b. Another lithographic process removes all the thin pad oxide except for a small area near the center of the base region (Figure 2.34c). The fifth step is to form the emitter region. As shown in Figure 2.34d, the base contact area is protected by a photoresist mask. Then, a low-energy, high-arsenic-dose ($\sim 10^{16}$ cm^{-2}) implantation forms the n^+-emitter and n^+-collector contact regions. The photoresist is removed, and a final metallization step forms the contacts to the base, emitter, and collector, as shown in Figure 2.32.

In this basic bipolar process, there are six film formation operations, six lithographic operations, four ion implantations, and four etching operations. Each operation must be precisely controlled and monitored. Failure of any one of the

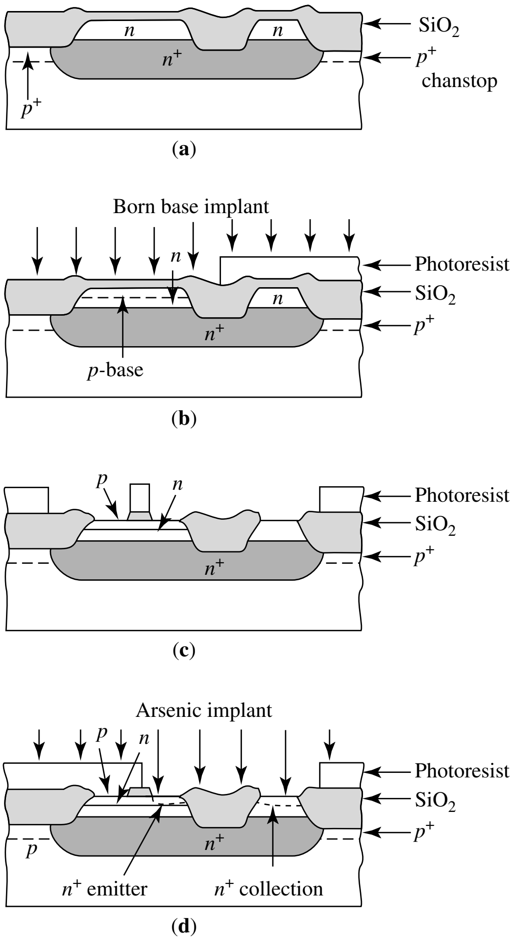


Figure 2.34. Cross-sectional views of bipolar transistor fabrication: (a) oxide isolation; (b) base implantation; (c) removal of thin oxide; (d) emitter−collector implant [1].

operations generally will render the wafer useless. The doping profiles of the completed transistor along a coordinate perpendicular to the surface and passing through the emitter, base, and collector are shown in Figure 2.35. The emitter profile is abrupt because of the concentration-dependent diffusivity of arsenic. The base doping profile beneath the emitter can be approximated by a Gaussian distribution for limited-source diffusion. The collector doping is given by the epitaxial doping level ($\sim 2 \times 10^{16}$ cm^{-3}) for a representative switching transistor.

2.2.2. CMOS Technology

The MOSFET is the dominant device used in modern integrated circuits because it can be scaled to smaller dimensions than other types of devices. The dominant technology for MOSFET is *complementary MOSFET* (CMOS) technology, in

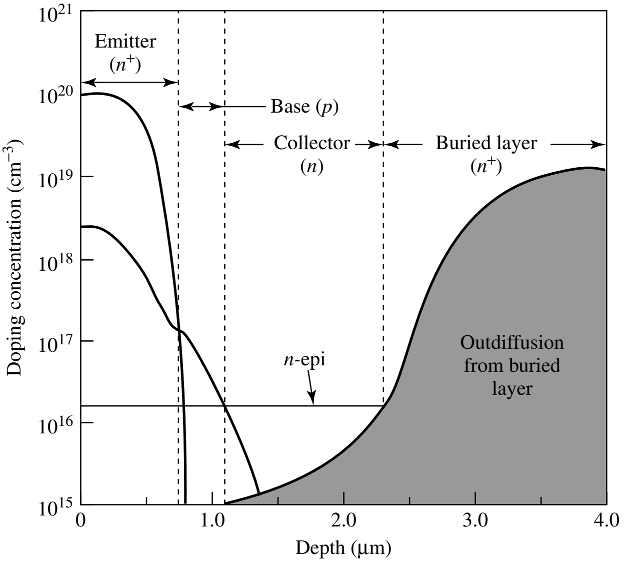


Figure 2.35. *n–p–n* bipolar transistor doping profile [1].

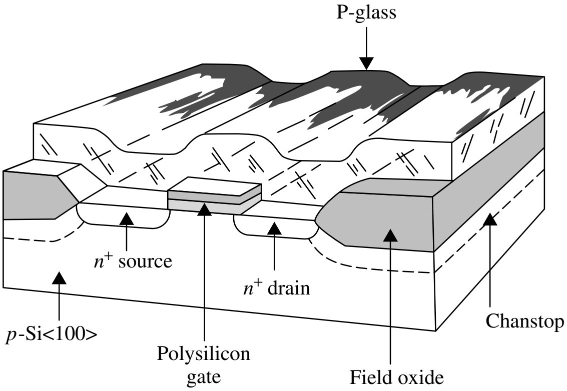


Figure 2.36. *n*-channel MOSFET [1].

which both n-channel and p-channel devices (NMOS and PMOS, respectively) are provided on the same chip. CMOS technology is particular attractive because it has the lowest power consumption of all IC technology. Figure 2.36 shows a perspective view of an n-channel MOSFET prior to final metallization. The top layer is a phosphorus-doped silicon dioxide (P-glass) that is used as an insulator between the polysilicon gate and the gate metallization and also as a gettering layer for mobile ions.

2.2.2.1. Basic NMOS Fabrication Sequence

In an NMOS process, the starting material is a p-type, lightly doped ($\sim 10^{15}$ cm^{-3}), $\langle 100 \rangle$-oriented, polished silicon wafer. The first step is to form the oxide isolation region using LOCOS technology. The process sequence for this step is similar to that for the bipolar transistor. A thin-pad oxide (~35 nm) is thermally grown, followed by a silicon nitride (~150 nm) deposition (Figure 2.37a). The active-device area is defined by a photoresist mask and a boron chanstop layer and is then implanted through the composite nitride–oxide layer (Figure 2.37b). The nitride layer not covered by the photoresist mask is subsequently removed by etching. After stripping the photoresist, the wafer is placed in an oxidation furnace to grow an oxide (called the *field oxide*), where the nitride layer is removed, and to drive in the boron implant. The thickness of the field oxide is typically 0.5–1 μm.

The second step is to grow the gate oxide and to adjust the threshold voltage. The composite nitride–oxide layer over the active-device area is removed, and

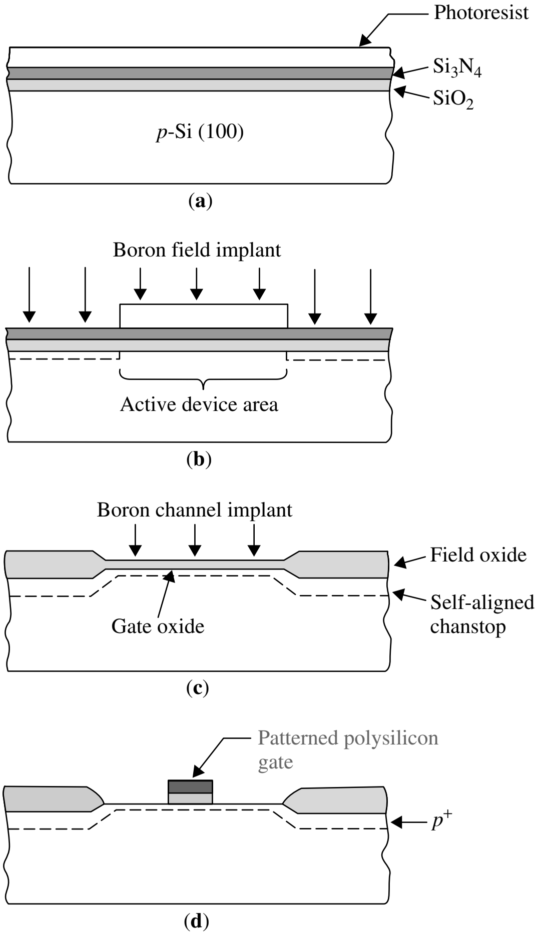


Figure 2.37. NMOS fabrication sequence: (a) formation of SiO_2, Si_3N_4, and photoresist layers; (b) boron implant; (c) field oxide; (d) gate [1].

a thin-gate oxide layer (less than 10 nm) is grown. For an enhancement-mode n-channel device, boron ions are implanted in the channel region, as shown in Figure 2.37c, to increase the threshold voltage to a predetermined value (e.g., +0.5 V). For a depletion-mode n-channel device, arsenic ions are implanted in the channel region to decrease the threshold voltage (e.g., −0.5 V).

The third step is to form the gate. A polysilicon is deposited and is heavily doped by diffusion or implantation of phosphorus to a typical sheet resistance of 20–30 Ω/□. This resistance is adequate for MOSFETs with gate lengths larger than 3 μm. For smaller devices, polycide, a composite layer of metal silicide and polysilicon such as W-polycide, can be used as the gate materials to reduce the sheet resistance to about 1 Ω/□.

The fourth step is to form the source and drain. After the gate is patterned (Figure 2.37d), it serves as a mask for the arsenic implantation (~30 keV, ~5 × 10^{15} cm^{-2}) to form the source and drain (Figure 2.38a), which are self-aligned

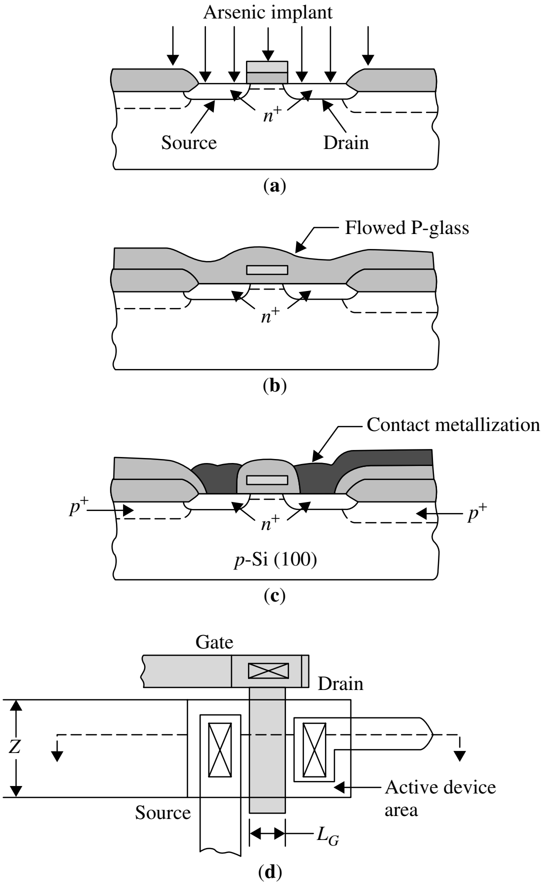


Figure 2.38. NMOS fabrication sequence: (a) source and drain; (b) P-glass deposition; (c) MOSFET cross section; (d) MOSFET top view [1].

with respect to the gate. At this stage, the only overlapping of the gate is due to lateral straggling of the implanted ions (for 30 keV As, $\sigma_\perp$ is only 5 nm). If low-temperature processes are used for subsequent steps to minimize lateral diffusion, the parasitic gate–drain and gate–source coupling capacitances can be much smaller than the gate–channel capacitance.

The last step is metallization. P-glass is deposited over the entire wafer and is flowed by heating the wafer to give a smooth surface topography (Figure 2.38b). Contact windows are defined and etched in the P-glass. A metal layer, such as aluminum, is then deposited and patterned. A cross-sectional view of the completed MOSFET is shown in Figure 2.38c, and the corresponding top view is shown in Figure 2.38d. The gate contact is usually made outside the active-device area to avoid possible damage to the thin-gate oxide.

2.2.2.2. CMOS Fabrication Sequence

The MOS process forms the foundation for CMOS technology. Figure 2.39a shows a CMOS inverter. The gate of the upper PMOS device is connected to the gate of the lower NMOS device. For the CMOS inverter, in either logic state, one device in the series path from V_{DD} to ground is nonconductive. The current that flows in either steady state is a small leakage current, and only when both devices are on during switching does a significant current flow through the inverter. Thus, the average power dissipation is on the order of nanowatts. Low power consumption is the most attractive feature of the CMOS circuit.

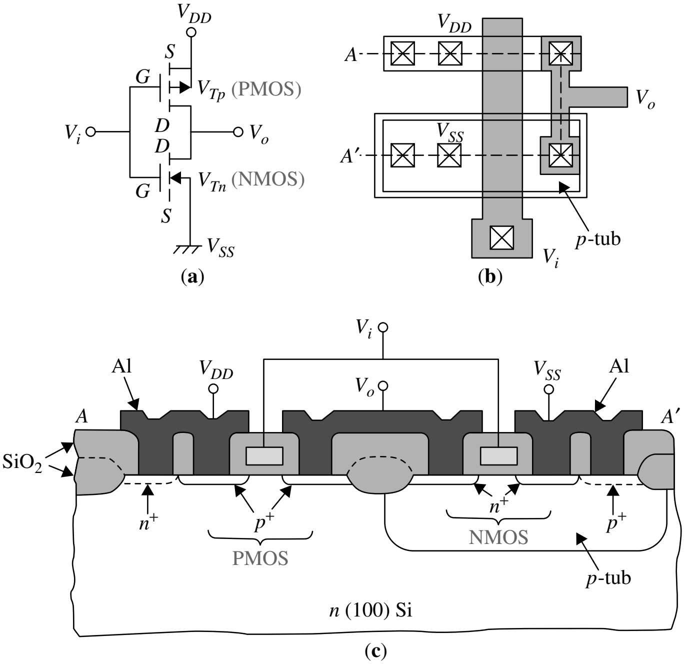


Figure 2.39. CMOS inverter: (a) circuit diagram; (b) layout; and (c) cross section [1].

Figure 2.39b shows a layout of the CMOS inverter, and Figure 2.39c shows the device cross section along the A–A' line. In the processing, a p tub (also called a *p well*) is first implanted and subsequently driven into the n substrate. The p-type dopant concentration must be high enough to overcompensate the background doping of the n substrate. The subsequent processes for the n-channel MOSFET in the p tub are identical to those described previously. For the p-channel MOSFET, $^{11}B^+$ or $^{49}(BF_2)^+$ ions are implanted into the n substrate to form the source and drain regions. A channel implant of $^{75}As^+$ ions may be used to adjust the threshold voltage and a n^+ chanstop is formed underneath the field oxide around the p-channel device. Because of the p tub and the additional steps needed to make the p-channel MOSFET, the number of steps to make a CMOS circuit is essentially double that to make an NMOS circuit. Thus, there is a trade-off between the complexity of processing and a reduction in power consumption.

Instead of the p tub described above, an alternate approach is to use an n tub formed in p-type substrate, as shown in Figure 2.40a. In this case, the n-type dopant concentration must be high enough to overcompensate for the background doping of the p substrate (i.e., $N_D > N_A$). In both the p-tub and the n-tub approaches, the channel mobility will be degraded because mobility is determined by the total dopant concentration $(N_A + N_D)$. A more recent approach using two separated tubs implanted into a lightly doped substrate is shown in Figure 2.40b. This structure is called a "twin tub." Because no overcompensation is needed in either of the twin tubs, higher channel mobility can be obtained.

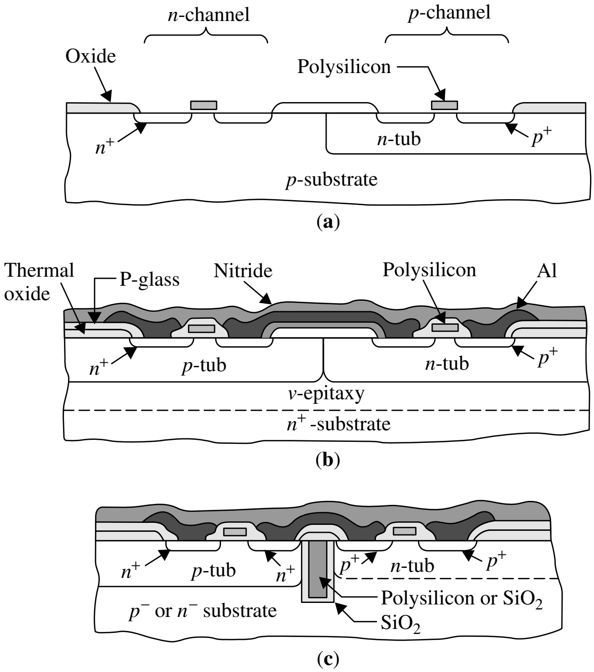


Figure 2.40. Various CMOS structures: (a) n-tub; (b) twin-tub; (c) refilled trench [1].

All CMOS circuits have the potential for a problem called *latchup* that is associated with parasitic bipolar transistors. These parasitic devices consist of the *npn* transistor formed by the NMOS source–drain regions, *p* tub, and *n*-type substrate, as well as the *pnp* transistor formed by the PMOS source-drain regions, *n*-type substrate, and *p* tub. Under appropriate conditions, the collector of the *pnp* device supplies base current to the *npn* and vice versa in a positive feedback arrangement. This latchup current can have serious negative repercussions in a CMOS circuit.

An effective processing technique to eliminate latchup is to use deep-trench isolation, as shown in Figure 2.40c. In this technique, a trench with a depth deeper than the well is formed in the silicon by anisotropic reactive sputter etching. An oxide layer is thermally grown on the bottom and walls of the trench, which is then refilled by deposited polysilicon or silicon dioxide. This technique can eliminate latchup because the *n*-channel and *p*-channel devices are physically isolated by the refilled trench. The detailed steps for trench isolation and some related CMOS processes are now considered.

Well Formation. The well of a CMOS circuit can be a single well, a twin well, or a retrograde well. The twin-well process exhibits some disadvantages. For example, it needs high-temperature processing (above 1050°C) and a long diffusion time (longer than 8 h) to achieve the required well depth of 2–3 μm. In this process, the doping concentration is highest at the surface and decreases monotonically with depth. To reduce the process temperature and time, high-energy implantation is used (i.e., implanting the ion to the desired depth instead of diffusion from the surface). The profile of the well in this case can have a peak at a certain depth in the silicon substrate. This is called a *retrograde well.*

The advantage of high-energy implantation is that it can form the well under low-temperature and short-time conditions. Hence, it can reduce the lateral diffusion and increase the device density. The retrograde well offers some additional advantages over the conventional well: (1) because of high doping near the bottom, the well resistivity is lower than that of the conventional well, and latchup can be minimized; (2) the chanstop can be formed at the same time as the retrograde well implantation, reducing processing steps and time; and (3) higher well doping in the bottom can reduce the chance of punchthrough from the drain to the source.

Isolation. The conventional MOS isolation process has some disadvantages that make it unsuitable for deep-submicrometer ($\leq$0.25-μm) fabrication. The high-temperature oxidation of silicon and long oxidation time result in the encroachment of the chanstop implantation (usually boron for *n*-MOSFET) to the active region and cause a threshold voltage shift. The area of the active region is reduced because of the lateral oxidation. In addition, the field oxide thickness in submicrometer-isolation spacings is significantly less than the thickness of field oxide grown in wider spacings. Trench isolation technology can avoid these problems.

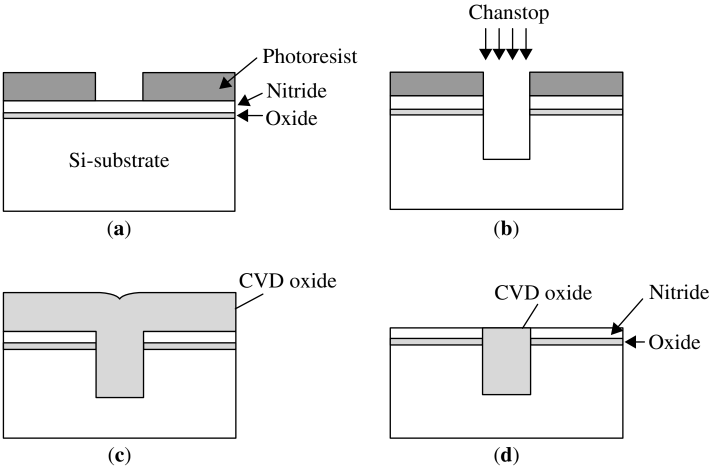


Figure 2.41. Shallow-trench isolation: (a) patterning on nitride–oxide films; (b) dry etching and chanstop implantation; (c) CVD oxide to refill; (d) surface after CMP [1].

An example is *shallow trench* (depth less than 1 μm) isolation, shown in Figure 2.41. After patterning (Figure 2.41a), the trench area is etched (Figure 2.41b) and then refilled with oxide (Figure 2.41c). Before refilling, a channel stop implantation can be performed. Since the oxide has overfilled the trench, the oxide on the nitride should be removed. Chemical–mechanical polishing is used to remove the oxide on the nitride and to get a flat surface (Figure 2.41d). Because of its high resistance to polishing, the nitride acts as a stop layer for the CMP process. After the polishing, the nitride layer and the oxide layer can be removed by H_3PO_4 and HF, respectively. This initial planarization step at the beginning is helpful for the subsequent polysilicon patterning and planarizations of the multilevel interconnection processes.

Gate Engineering. If n^+-polysilicon is used for both PMOS and NMOS gates, the threshold voltage for PMOS has to be adjusted by boron implantation. This makes the channel of the PMOS a buried type, as shown in Figure 2.42a. The buried-type PMOS suffers serious short-channel effects as the device size shrinks below 0.25 μm. The most noticeable phenomena for short-channel effects are threshold voltage rolloff, drain-induced barrier lowering, and the large leakage current at the OFF state. To alleviate these problems, the n^+-polysilicon can be changed to p^+-polysilicon for the PMOS devices. Due to the workfunction difference (1.0 eV from n^+- to p^+-polysilicon), a surface p-type channel device can be achieved without the boron V_T adjustment implantation. Hence, as the technology shrinks to 0.25 μm and less, dual-gate structures are required: p^+-polysilicon gate for PMOS and n^+-polysilicon for NMOS (Figure 2.42b).

To form the p^+-polysilicon gate, ion implantation of BF_2 is commonly used. However, boron penetrates easily from the polysilicon through the oxide into the silicon substrate at high temperatures, resulting in a V_T shift. This penetration

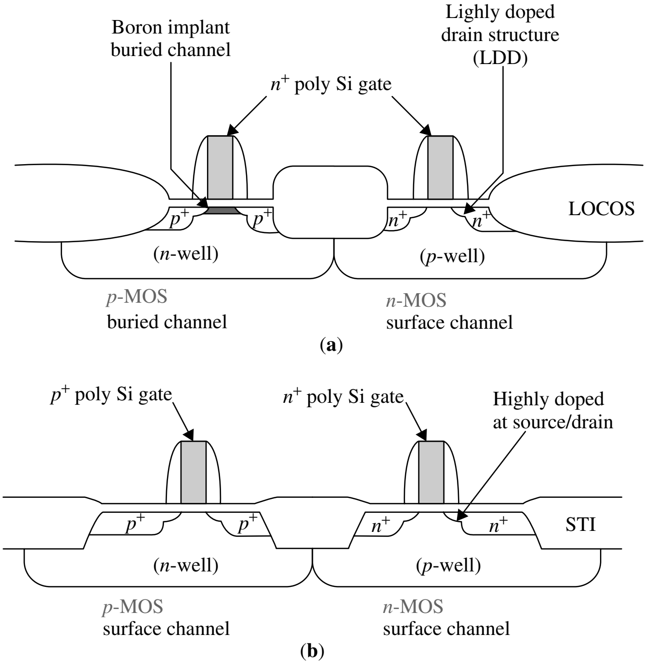


Figure 2.42. (a) Conventional CMOS structure with a single polysilicon gate; (b) advanced CMOS structure with dual polysilicon gates [1].

is enhanced in the presence of a F atom. There are methods to reduce this effect: use of rapid thermal annealing to reduce the time at high temperatures and, consequently, the diffusion of boron; use of nitrided oxide to suppress the boron penetration, since boron can easily combine with nitrogen and becomes less mobile; and the creation of a multilayer of polysilicon to trap the boron atoms at the interface of the two layers.

2.2.3. BiCMOS Technology

BiCMOS is a technology that combines both CMOS and bipolar device structures in a single IC. The reason to combine these two different technologies is to create an IC chip that has the advantages of both CMOS and bipolar devices. We know that CMOS exhibits advantages in power dissipation, noise margin, and packing density, whereas bipolar technology shows advantages in switching speed, current drive capability, and analog capability. As a result, for a given design rule, BiCMOS can have a higher speed than CMOS, better performance in analog circuits than CMOS, a lower power dissipation than bipolar, and a higher component density than bipolar.

BiCMOS has been widely used in many applications. Early on, it was used in static random access memory (SRAM) circuits. Currently, BiCMOS technology has been successfully developed for transceiver, amplifier, and oscillator

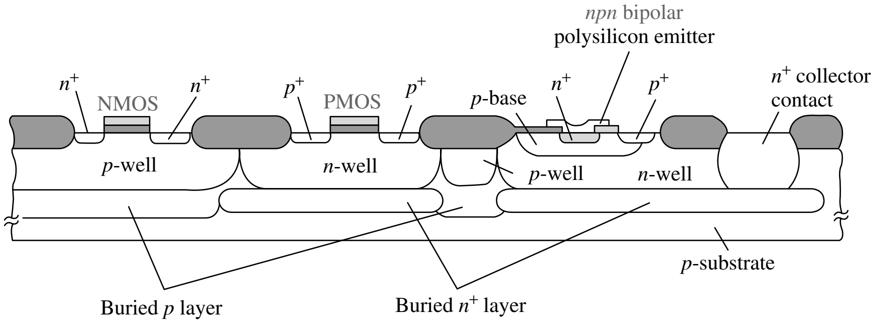


Figure 2.43. BiCMOS device structure [1].

applications in wireless communication equipment. Most BiCMOS processes are based on standard CMOS process with some modifications, such as adding masks for bipolar transistor fabrication. The example shown in Figure 2.43 is for a high-performance BiCMOS process based on the twin-well CMOS approach.

The initial material is a p-type silicon substrate. An n^+-buried layer is formed to reduce collector resistance. The buried p layer is formed by ion implantation to increase the doping level and prevent punchthrough. A lightly doped n-epi layer is grown on the wafer, and a twin-well process for the CMOS is performed. To achieve high performance for the bipolar transistor, four additional masks are needed: the buried n^+ mask, the collector deep n^+ mask, the base p mask, and the polyemitter mask. The p^+ region for base contact can be formed with the p^+ implant in the source–drain implantation of the PMOS, and the n^+ emitter can be formed with the source/drain implantation of the NMOS. The additional masks and longer processing time compared with a standard CMOS process are the main drawbacks of BiCMOS.

2.2.4. Packaging

Before finished ICs can be put to their intended use in various commercial electronic systems and products (such as computers, cellular phones, and digital cameras), several other key processes must take place. These include both *electrical testing* and *packaging*. Testing, which is discussed in detail in Chapter 3, is clearly necessary to ensure high-quality products. The term *packaging* refers to the set of technologies and processes that connect ICs with electronic systems. A useful analogy is to consider an electronic product as the human body. Like the body, these products have "brains," which are analogous to ICs. Electronic packaging provides the "nervous system," as well as the "skeletal system." The package is responsible for interconnecting, powering, cooling, and protecting the IC.

Overall, electronic systems consist of several levels of packaging, each with distinctive types of interconnection devices. Figure 2.44 depicts this packaging hierarchy. Level 0 consists of on-chip interconnections. Chip-to-printed circuit board or chip-to-module connections constitute level 2, and board-to-board

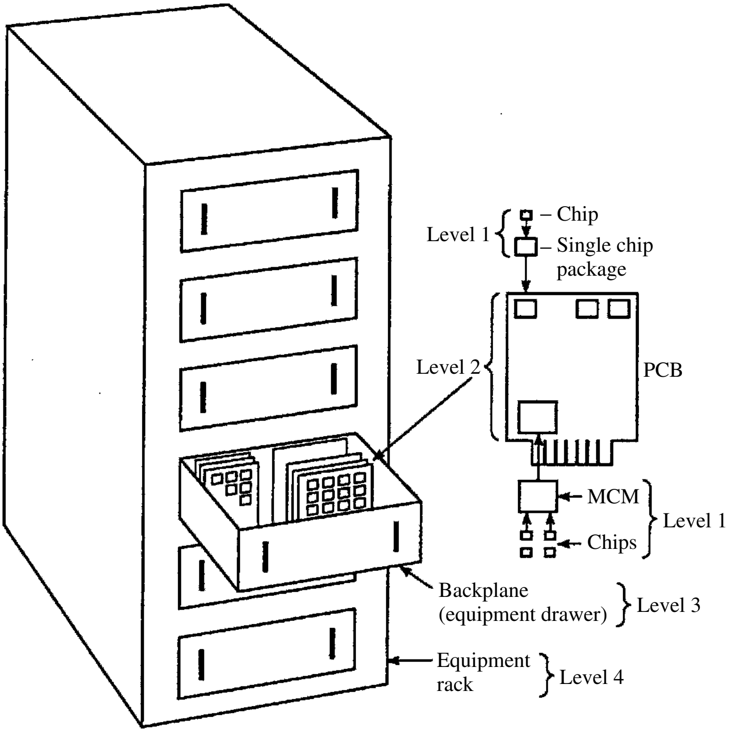


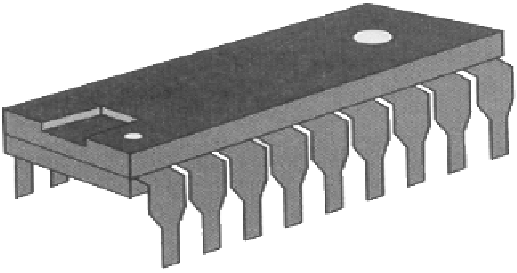
Figure 2.44. Electronic packaging hierarchy [2].

interconnections make up level 3. Levels 4 and 5 consist of connections between subassemblies and between systems (such as computer to printer), respectively.

2.2.4.1. Die Separation

After functional testing, individual ICs (or dies) must be separated from the substrate. This is the first step in the packaging process. In a common method that has been used for many years, the substrate wafer is mounted on a holder and scribed in both the x and y directions using a diamond scribe. This is done along scribe borders of 75–250 μm in width that are formed around the periphery of the dies during fabrication. These borders are aligned with the crystal planes of this substrate if possible. After scribing, the wafer is removed form the holder and placed upside-down on a soft support. A roller is then used to apply pressure, fracturing the wafer along the scribe lines. This must be accomplished with minimal damage to the individual die.

More modern die separation processes use a diamond saw, rather than a diamond scribe. In this procedure, the wafer is attached to an adhesive sheet of mylar film. The saw is then used to either scribe the wafer or to cut completely through it. After separation, the dies are removed from the mylar. The separated dice are then ready to be placed into packages.



Plastic dual-in-line-package

Figure 2.45. Dual-inline package [2].

2.2.4.2. Package Types

There are a number of approaches to the packaging of single ICs. The *dual-inline package* (DIP) (Figure 2.45), is the package most people envision when they think of integrated circuits. The DIP was developed in the 1960s, quickly became the primary package for ICs, and has long dominated the electronics packaging market. The DIP can be made of plastic or ceramic; the latter is called the *CerDIP*. The CerDIP consists of a DIP constructed of two pieces of sandwiched ceramic with leads protruding from between the ceramic plates.

In the 1970s and 1980s, *surface-mount packages* were developed in response to a need for higher-density interconnect than the DIP approach could provide. In contrast to DIPs, the leads of a surface-mounted package do not penetrate the printed circuit board (PCB) on which it is mounted. This means that the package can be mounted on both sides of the board, thereby allowing higher density. One example of such a package is the quad flatpack (QFP) (Figure 2.46), which has leads on all four sides to further increase the number of input/output (I/O) connections.

More recently, the need for rapidly increasing numbers of I/O connections has led to the development of pin-grid array (PGA) and ball-grid array (BGA) packages (Figures 2.47 and 2.48, respectively). PGAs have an I/O density of about 600, and BGAs can have densities greater than 1000, as compared to

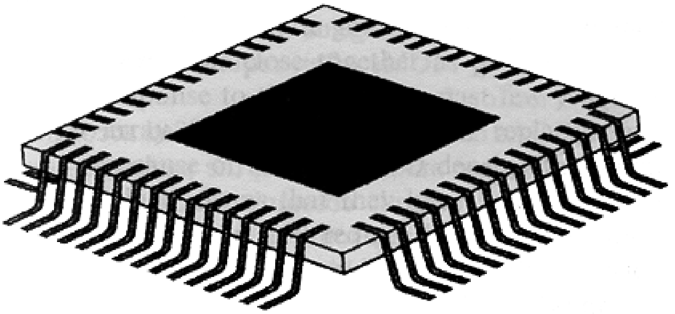


Figure 2.46. Quad flatpack [2].

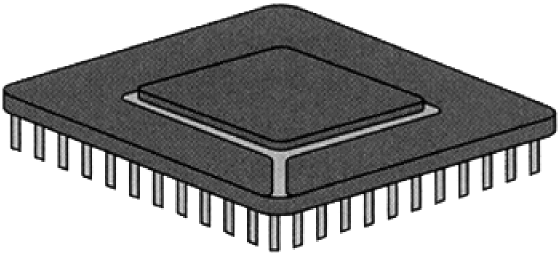


Figure 2.47. Pin-grid array [2].

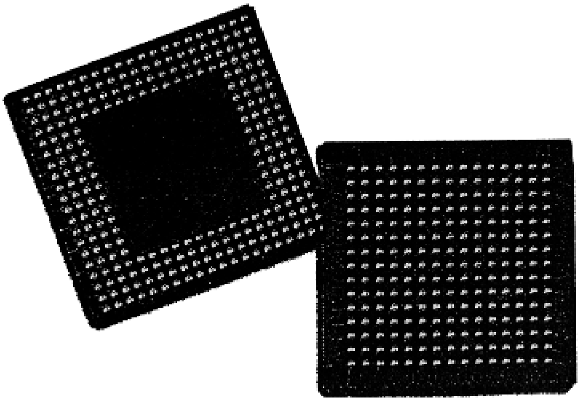


Figure 2.48. Ball-grid array [2].

~200 for QFPs. BGAs can be identified by the solder bumps on the bottom of the package. With QFPs, as the spacing between leads becomes tighter, the manufacturing yield decreases rapidly. The BGA allows higher density and takes up less space than the QFP, but its manufacturing process is inherently more expensive.

The most recent development in packaging is the chip-scale package (CSP), which is shown in Figure 2.49. CSPs, defined as packages no larger than 20% greater than the size of the IC die itself, often take the form of miniaturized ball-grid arrays. They are designed to be *flipchip-mounted* (see Section 2.2.4.3) using conventional equipment and solder reflow. CSPs are typically manufactured in a process that creates external power and signal I/O contacts and encapsulates the finished silicon die prior to dicing the wafer. Essentially, CSPs provide an interconnection framework for ICs so that before dicing, each die has all the functions (i.e., external electrical contacts, encapsulation of the finished silicon) of a conventional, fully packaged IC. Two essential features of this approach are that the leads and interposer layer (an added layer on the IC used to provide electrical functionality and mechanical stability) are flexible enough so that the packaged device is compliant with the test fixture for full testing and burning, and the package can accommodate the vertical nonplanarity and thermal expansion and contraction of the underlying printed circuit board during assembly and operation.

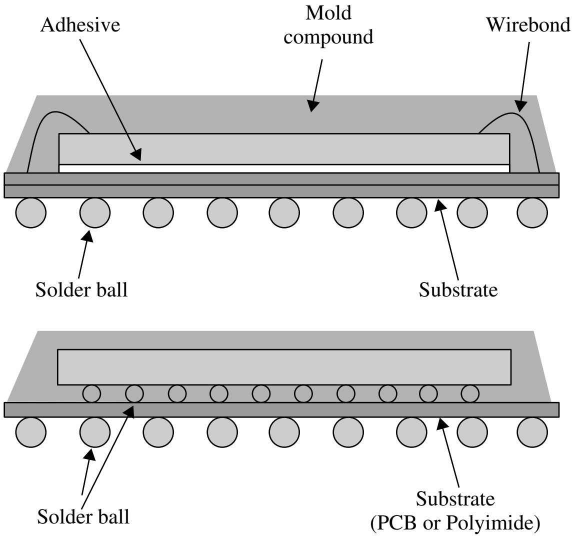


Figure 2.49. Two typical chip-scale packages [3].

2.2.4.3. Attachment Methods

An IC must be mounted and bonded to a package, and that package must be attached to a printed circuit board before the IC can be used in an electronic system. Methods of attaching ICs to PCBs are referred to as *level 1 packaging*. The technique used to bond a bare die to a package has a significant effect on the ultimate electrical, mechanical and thermal properties of electronic system being manufactured. Chip-to-package interconnection is generally accomplished by either *wire bonding, tape-automated bonding* (TAB), or *flipchip bonding* (see Figure 2.50).

Wire bonding is the oldest attachment method and is still the dominant technique for chips with fewer that 200 I/O connections. Wire bonding requires connecting gold or aluminum wires between chip bonding pads and contact points on the package. ICs are first attached to the substrate using a thermally conductive adhesive with their bonding pads facing upward. The Au or Al wires are then attached between the pads and substrate using *ultrasonic, thermosonic*, or *thermocompression* bonding [1]. Although automated, this process is still time-consuming since each wire must be attached individually.

Tape-automated bonding (TAB) was developed in the early 1970s and is often used to bond packages to PCBs. In TAB, chips are first mounted on a flexible polymer tape (usually polyimide) containing repeated copper interconnection patterns. The copper leads are defined by lithography and etching, and the lead pattern can contain hundreds of connections. After the IC pads have been aligned to metal interconnection stripes on the tape, attachment takes place by thermocompression. Gold bumps are formed on either side of the die or tape and are used to bond the die to the leads on the tape.

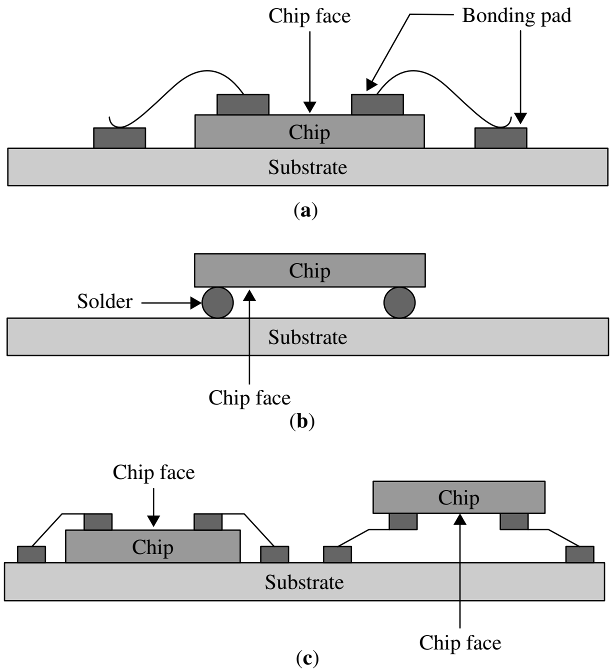


Figure 2.50. (a) Wire; (b) flipchip; (c) tape-automated bonding [2].

Flipchip bonding is a direct interconnection approach in which the IC is mounted upside-down onto a module or printed circuit board. Electrical connections are made via solder bumps (or solderless materials such as epoxies or conductive adhesives) located over the surface of the chip. Since bumps can be located anywhere on the chip, flipchip bonding ensures that the interconnect distance between the chip and package is minimized. The I/O density is limited only by the minimum distance between adjacent bond pads.

SUMMARY

In this chapter, we have provided an overview of the critical unit processes in IC fabrication and described the integration of these unit processes into sequences for fabricating and packaging ICs. In the next chapter, we will discuss how these processes are monitored to facilitate quality control.

PROBLEMS

2.1. Assuming that a silicon oxide layer of thickness x is grown by thermal oxidation, show that the thickness of silicon being consumed is $0.44x$. The molecular weight of Si is 28.9 g/mol, and the density of Si is 2.33 g/cm^3. The corresponding values for SiO_2 are 60.08 g/mol and 2.21 g/cm^3.

2.2. A silicon sample is oxidized in dry O_2 at 1200°C for one hour.

(a) What is the thickness of the oxide grown?

(b) How much additional time is required to grow 0.1 μm more oxide in wet O_2 at 1200°C?

2.3. Find the parameter γ for the photoresists shown in Figure 2.13.

2.4. Calculate the Al average etch rate and etch rate uniformity on a 200-mm-diameter silicon wafer, assuming that the etch rates at the center, left, right, top, and bottom of the wafer are 750, 812, 765, 743, and 798 nm/min, respectively.

2.5. The electron densities in RIE and HDP systems range within 10^9–10^{10} and 10^{11}–10^{12} cm^{-3}, respectively. Assuming that the RIE chamber pressure is 200 mTorr and HDP chamber pressure is 5 mTorr, calculate the ionization efficiency in RIE reactors and HDP reactors at room temperature. The ionization efficiency is the ratio of the electron density to the density of molecules.

2.6. For a boron diffusion in silicon at 1000°C, the surface concentration is maintained at 10^{19} cm^{-3} and the diffusion time is 1 h. If the diffusion coefficient of boron at 1000°C is 2×10^{14} cm^2/s, find $Q(t)$ and the gradient at $x = 0$ and at a location where the dopant concentration reaches 10^{15} cm^{-3}.

2.7. Arsenic was predeposited by arsine gas, and the resulting total amount of dopant per unit area was 1×10^{14} atoms/cm^2. How long would it take to drive the arsenic in to a junction depth of 1 μm? Assume a background doping of $C_B = 1 \times 10^{15}$ atoms/cm^3, and a drive-in temperature of 1200°C. For As diffusion, $D_0 = 24$ cm^2/s, and $E_a = 4.08$ eV.

2.8. Assume 100-keV boron implants on a 200-mm silicon wafer at a dose of 5×10^{14} ions/cm^2. The projected range and project straggle are 0.31 and 0.07 μm, respectively. Calculate the peak concentration and the required ion-beam current for 1 min of implantation.

2.9. In a CMP process, the oxide removal rate and the removal rate of a layer underneath the oxide (called a *stop layer*) are $1r$ and $0.1r$, respectively. To remove 1 μm of oxide and a 0.01-μm stop layer, the total removal time is 5.5 min. Find the oxide removal rate.

REFERENCES

1. G. May and S. Sze, *Fundamentals of Semiconductor Fabrication*, Wiley, New York, 2003.
2. W. Brown, ed., *Advanced Electronic Packaging*, IEEE Press, New York, 1999.
3. R. Tummala, ed., *Fundamentals of Microsystems Packaging*, McGraw-Hill, New York, 2001.

3

PROCESS MONITORING

OBJECTIVES

- Survey various sensor metrology and methods of monitoring IC fabrication processes.
- Place this metrology in the context of process needs.
- Identify key measurement points in the process flow.
- Differentiate between wafer state and equipment state measurements.

INTRODUCTION

In Chapter 2, the basic unit processes used in fabricating an integrated circuit, as well as the process flows for several major IC technologies, were discussed in detail. In order for these processes to repeatably produce reliable, high-quality devices and circuits, each unit process must be strictly controlled. Many diagnostic tools are used to maintain systematic control. Such control requires that the key output variables for each process step (i.e., those that are correlated with product functionality and performance) be carefully monitored.

Process monitoring enables operators and engineers to detect problems early on to minimize their impact. The economic benefit of effective monitoring systems increases with the complexity of the manufacturing process. Manufacturing line monitors consist of extremely sophisticated metrology equipment that can be divided into tools characterizing the state of features on the semiconductor

Fundamentals of Semiconductor Manufacturing and Process Control,
By Gary S. May and Costas J. Spanos

wafers themselves and those that describe the status of the fabrication equipment operating on those wafers. The issues involved in understanding and implementing both wafer state and equipment state measurements will be discussed in detail in this chapter.

3.1. PROCESS FLOW AND KEY MEASUREMENT POINTS

When we monitor a physical system, we observe that system's behavior. On the basis of these observations, we take appropriate actions to influence that behavior in order to guide the system to some desirable state. Semiconductor manufacturing systems consist of a series of sequential process steps in which layers of materials are deposited on substrates, doped with impurities, and patterned using photolithography to produce sophisticated integrated circuits and devices.

As an example of such a system, Figure 3.1 depicts a typical CMOS process flow (refer to Section 2.2.1 for more details). Inserted into this flow diagram in various places are symbols denoting key measurement points. Clearly, CMOS technology involves many unit processes with high complexity and tight tolerances. This necessitates frequent and thorough inline process monitoring to assure high-quality final products.

The measurements required may characterize physical parameters, such as film thickness, uniformity, and feature dimensions; or electrical parameters, such as resistance and capacitance. These measurements may be performed directly on product wafers, either directly or using test structures, or alternatively, on nonfunctional monitor wafers (or "dummy" wafers). In addition to these, some measurements are actually performed "in situ," or *during* a fabrication step. When a process sequence is complete, the product wafer is diced, packaged, and subjected to final electrical and reliability testing.

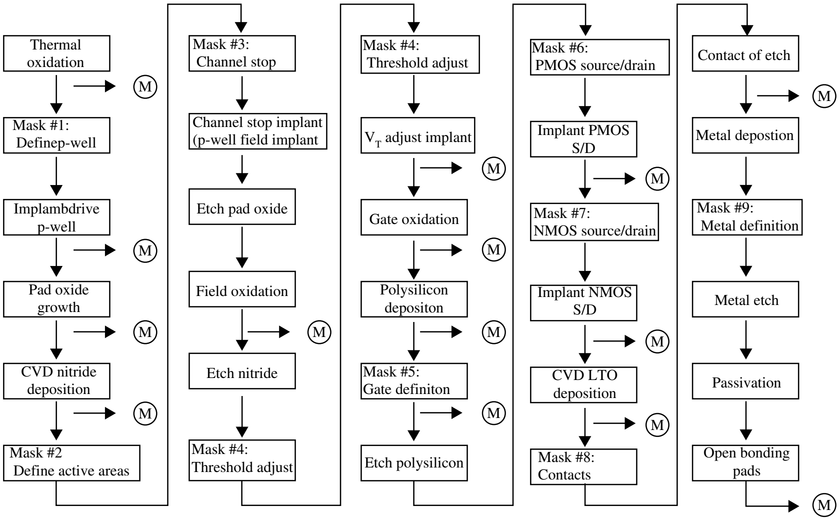


Figure 3.1. CMOS process flow showing key measurement points (denoted by "M").

3.2. WAFER STATE MEASUREMENTS

There is no substitute for regular inspection of products during manufacturing to ensure high quality. Inspections can reveal contamination, structural flaws, or other problems. Such investigations must not be limited to visual inspections, however, since not all processes have a visible effect on electronic products. Thin-film deposition and ion implantation are two important examples of this. In addition, with ever-increasing levels of integration, features on wafers become smaller and more difficult to inspect. As a result, visual inspection must be supplemented by sophisticated physical and electrical measurements of various characteristics that describe the state of a wafer.

Wafer state characterization includes the measurement of the physical parameters related to each manufacturing process step. Examples include

Lithography

- Linewidth
- Overlay
- Print bias
- Resist profiles

Etch

- Etch rate
- Selectivity
- Uniformity
- Anisotropy
- Etch bias

Deposition or Epitaxial Growth

- Sheet resistance
- Film thickness
- Surface concentration
- Dielectric constant
- Refractive index

Diffusion or Implantation

- Sheet resistance
- Junction depth
- Surface concentration

The total collection of such measurements relate to the physical characteristics of product wafers, and these physical characteristics can be correlated with the electrical performance of devices and circuits. The following sections describe wafer state measurements and the corresponding measurement apparatus in greater detail.

3.2.1. Blanket Thin Film

We begin the discussion of wafer state measurements with those measurements that are performed on blanket thin films. The term "blanket" is used to differentiate wafers that have been uniformly coated by a thin film from those in which the film has been patterned using photolithography and etching.

3.2.1.1. Interferometry

Optical metrology provides fast and precise measurements of film thickness and optical constants. In semiconductor manufacturing, *interferometry* (sometimes called *reflectometry*) is a widely used optical method for measuring such parameters. Single- or multiple wavelength interferometers are commonly used for both in situ and postprocess measurements of film thickness. In this method, a light source, usually a laser, is focused on a semiconductor wafer while a detector measures the reflected light intensity. The wafer consists of a parallel stack of partially transparent thin films. The reflected light intensity varies as a function of time depending on the thickness of the top layer due to constructive and destructive interference caused by multiple reflections.

To illustrate the basic concept, consider Figure 3.2, which shows a film of uniform thickness d and index of refraction n, with the eye of the observer focused on point a. The film is illuminated by broad source of monochromatic light S. There is a point P on the source such that two rays (represented by the single and double arrows) can leave P and enter the eye after traveling through point a. These two rays follow different paths, one reflected from the upper surface of the film and the other from the lower surface. Whether point a appears bright or dark depends on the nature of the interference (i.e., constructive or destructive) between the two waves that diverge from a.

The two factors that impact the nature of the interference are differences in optical path length and phase changes on reflection. For the two rays to combine to give maximum intensity, we must have

$$2dn\cos\theta = (m + 0.5)\lambda \tag{3.1}$$

where $m = 0, 1, 2, \ldots$ and θ is the angle of the refracted beam relative to the surface normal. The term 0.5λ accounts for the phase change that occurs on reflection since a phase change of 180° is equivalent to half a wavelength. The condition for minimum intensity is

$$2dn\cos\theta = m\lambda \tag{3.2}$$

Equations (3.1) and (3.2) hold when the index of refraction of the film is either greater or less than the indices of the media on *each* side of the film. Therefore,

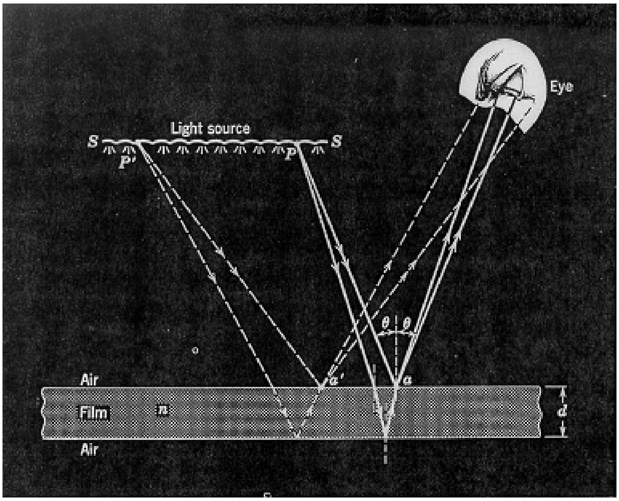
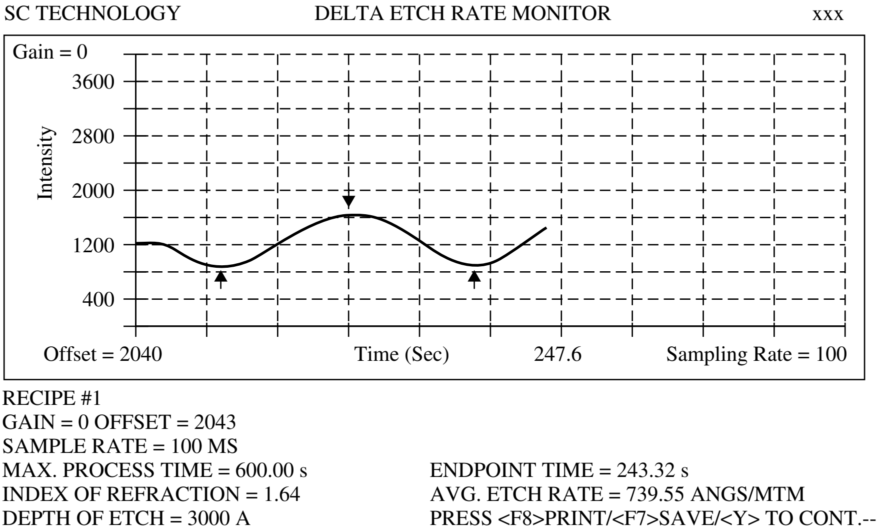


Figure 3.2. Interference by reflection from a thin film [1].


SC TECHNOLOGY DELTA ETCH RATE MONITOR xxx

RECIPE #1
GAIN = 0 OFFSET = 2043
SAMPLE RATE = 100 MS
MAX. PROCESS TIME = 600.00 s
INDEX OF REFRACTION = 1.64
DEPTH OF ETCH = 3000 A
ENDPOINT TIME = 243.32 s
AVG. ETCH RATE = 739.55 ANGS/MTM
PRESS <F8>PRINT/<F7>SAVE/<Y> TO CONT.--

Figure 3.3. Sample interferogram used for plasma etch monitoring [2].

if the index of refraction is known, the thickness of the film may be computed by simply counting peaks or valleys in the reflected waveform. An example of such a waveform (or *interferogram*) appears in Figure 3.3.

Interferometry becomes more complex when applied to stacks of several thin films. The overall goal, however, is still to obtain film thickness information from the time-varying reflected intensity signal. The reflected light intensity is given by [7]

$$I_r(d, \lambda) = I_0(\lambda) r(d, \lambda, \phi_1, \phi_2, \ldots, \phi_N) \tag{3.3}$$

where I_0 is the incident light intensity, r is the reflection coefficient, d is the thickness of the top layer, and ϕ_i are physical constants (i.e., thicknesses and refractive indices) associated with the lower films in the film stack.

The reflected intensity is monitored using a detector consisting of a light-sensitive transducer, such as a photodiode, in conjunction with an optical filter or diffraction grating to select the wavelength(s) of interest. The output of the detector corresponding to a particular wavelength is of the form

$$y_\lambda(kT) = \alpha(\lambda, kT)A(\lambda, kT)I_0(\lambda, kT)r(d(kT), \lambda) + e_\lambda(kT) \tag{3.4}$$

where T denotes the sampling period, k is an integer, α represents losses in the optical system, A is the gain of the detector, and e_λ is measurement noise. The physical parameters ϕ_i are considered to be fixed and known in this formulation and are not shown. For multiple-wavelength (or *spectroscopic*) measurements, this expression is repeated for each wavelength used. For p wavelengths, in matrix form, this is written as

$$\mathbf{y}(kT) = \text{diag}(\mathbf{h}(kT)\mathbf{r}(d(kT)) + e(kT) \tag{3.5}$$

where diag($\mathbf{x}$) represents a matrix with the elements of the vector $\mathbf{x}$ along the diagonal and

$$\mathbf{y}(kT) = [y_{\lambda_1}(kT) \cdots y_{\lambda_p}(kT)] \tag{3.6}$$

$$\begin{aligned}\mathbf{h}(kT) = [&\alpha(\lambda_1, kT)A(\lambda_1, kT)I_0(\lambda_1, kT) \cdots \alpha(\lambda_p, kT)\\ &A(\lambda_p, kT)I_0(\lambda_p, kT)]\end{aligned} \tag{3.7}$$

$$\mathbf{r}(d(kT)) = [r(d(kT), \lambda_1) \cdots r(d(kT), \lambda_p)]^T \tag{3.8}$$

$$\mathbf{e}(kT) = [e_{\lambda_1}(kT) \cdots e_{\lambda_p}(kT)]^T \tag{3.9}$$

where the superscript T represents the transpose operation.

To obtain film thickness or the rate of change of thickness (i.e., etch or deposition rate), the detector output is processed in one of two ways: (1) extrema counting or (2) least-squares fitting. Extrema counting takes advantage of the fact that the reflected light intensity varies approximately periodically with both the wavelength of the incident light and the thickness of the top film. The distance between peaks and valleys is a known function of the top film thickness. Thus, if many wavelengths are available, thickness can be determined by counting the peaks in a plot of reflectance versus wavelength. If only a single wavelength is available, the movement of peaks and valleys over time during in situ measurements indicates that a specific amount of material has been etched or deposited. This provides the average etch or deposition rate between successive minima and maxima.

To use the least-squares approach, at each timepoint, the following nonlinear optimization problem is posed:

$$\min_d[(\mathbf{y}(kT) - \text{diag}(\mathbf{h})\mathbf{r}(d))^T(\mathbf{y}(kT) - \text{diag}(\mathbf{h})\mathbf{r}(d))] \tag{3.10}$$

The film thickness is then that for which the minimum is achieved. Etch rate or deposition rate is then calculated from the resulting thickness versus time curve.

The final variation of interferometry we will discuss briefly is one that is particularly applicable to thickness monitoring during plasma etching. During etching, the emission from the plasma itself may be used as the light source. As this light is reflected from the etched film and underlying film surfaces while the thickness of the etched film decreases, the optical path difference between light rays varies and the changing constructive and destructive interference results in periodic signals in the same manner as previously described. If a charge-coupled device (CCD) camera is placed in such a way that it can view these signals (see Figure 3.4), each pixel of the CCD camera then acts as an individual interferometer monitoring a different part of the wafer. This arrangement is called *full-wafer interferometry* [5].

3.2.1.2. Ellipsometry

Ellipsometry is a widely used measurement technique based on the polarization changes that occur when light is reflected from or transmitted through a medium. Changes in polarization are a function of the optical properties of the material (i.e., its complex refractive indices), its thickness, and the wavelength and angle of incidence of the light beam relative to the surface normal. When multiple light beams of varying wavelength are used, the technique is referred to as *spectroscopic ellipsometry* (SE). SE, which can be used to make in situ or postprocess measurements, is a fundamentally more accurate technique than interferometry for obtaining film thickness and optical dielectric function information. In general, SE measurements are performed at an off-normal angle with respect to the sample. In this configuration, the measurement is sensitive to the polarization state of both the incident and reflected waves.

Figure 3.5 shows an unpolarized beam of light falling on a dielectric surface. In this case, the dielectric is glass. The electric field vector for each wavetrain

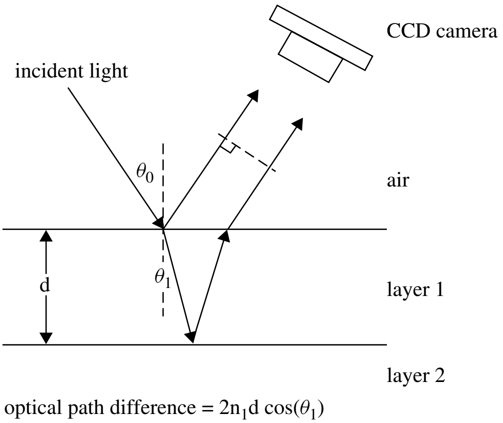


Figure 3.4. Schematic of full-wafer interferometry [5].

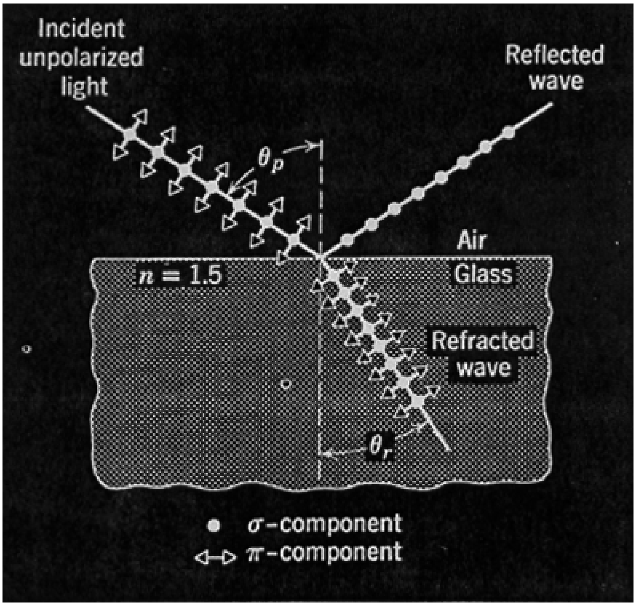


Figure 3.5. Illustration of components of polarization [1].

in the beam can be resolved into two components—one perpendicular to the plane of incidence (i.e., the plane of the figure) and another parallel to this plane. The perpendicular component, represented by the dots, is the σ component (or "*s* component"). The parallel component, represented by the arrows, is the π component (or "*p* component"). On average, for completely unpolarized incident light, these two components are of equal amplitude. However, if the incident beam is polarized (as is the case in ellipsometry), this is no longer true.

In the most common configuration, linearly polarized light is incident on the surface, and the elliptical polarization status of the reflected light is analyzed. Measured ellipsometry data are usually written in the form of the ratio (ρ) of the *total reflection coefficients* for *s* and *p* polarization (R^s and R^p, respectively). In other words

$$\rho = R^p/R^s = \tan(\psi)e^{i\Delta} \tag{3.11}$$

where tan(ψ) is the ratio of the magnitude of the *p*-polarized light to the *s*-polarized reflected light and Δ is the difference in phase shifts on reflection for the *p* and *s* polarizations, respectively.

Another set of expressions called the *Fresnel equations* relate [Eq. (3.11)] to the bulk complex dielectric function (ε). The dielectric function represents the degree to which the material may be polarized by an applied external electric field, and as a complex number, it is expressed as

$$\varepsilon = \varepsilon_1 + j\varepsilon_2 \tag{3.12}$$

where ε_1 and ε_2 are the real and imaginary parts, respectively. For heterogeneous samples consisting of multiple layers, the dielectric function determined by ellipsometry is an average over the region penetrated by the incident light called the *effective dielectric function*, $\langle\varepsilon\rangle$. If the sample structure is not too complicated, $\langle\varepsilon\rangle$ can be simulated by appropriate models (such as the "ambient–film–substrate" model). In this case, film and substrate properties can be separated, and film properties (i.e., thickness or dielectric function) can be determined as follows.

Because there are a maximum of two independent optical parameters (Ψ and Δ) measured at each wavelength, the maximum number of unknowns that can be determined from a single spectral measurement is $2w$, where w is the number of wavelengths scanned. Thus far, we have discussed the index of refraction as if it were a single parameter. However, in general, the *complex index of refraction* (N) consists of a real part (n) and an imaginary part (k), or

$$N = n - jk \tag{3.13}$$

where k is the *extinction coefficient*, which is a measure of how rapidly the intensity decreases as light passes through a material. The dielectric function is related to the complex index of refraction by the relationship

$$\varepsilon = N^2 \tag{3.14}$$

Therefore, we can obtain values for n and k in terms of ε_1 and ε_2 using

$$n = \sqrt{\tfrac{1}{2}\left[(\varepsilon_1^2 + \varepsilon_2^2)^{1/2} + \varepsilon_1\right]} \tag{3.15}$$

$$k = \sqrt{\tfrac{1}{2}\left[(\varepsilon_1^2 + \varepsilon_2^2)^{1/2} - \varepsilon_1\right]} \tag{3.16}$$

As mentioned above, the complex index of refraction is related to the total reflection coefficients by the Fresnel equations, which are given by [6]

$$R^p = \frac{r_{12}^p + r_{23}^p \exp(-j2\beta)}{1 + r_{12}^p r_{23}^p \exp(-j2\beta)} \tag{3.17}$$

$$R^s = \frac{r_{12}^s + r_{23}^s \exp(-j2\beta)}{1 + r_{12}^s r_{23}^s \exp(-j2\beta)} \tag{3.18}$$

where the Fresnel reflection coefficients at the individual interfaces are of the form

$$r_{12}^p = \frac{N_2 \cos\phi_1 - N_1 \cos\phi_2}{N_2 \cos\phi_1 + N_1 \cos\phi_2} \tag{3.19}$$

$$r_{12}^s = \frac{N_1 \cos\phi_1 - N_2 \cos\phi_2}{N_1 \cos\phi_1 + N_2 \cos\phi_2} \tag{3.20}$$

and

$$\beta = 2\pi\left(\frac{d}{\lambda}\right) N_2 \cos\phi_2 \tag{3.21}$$

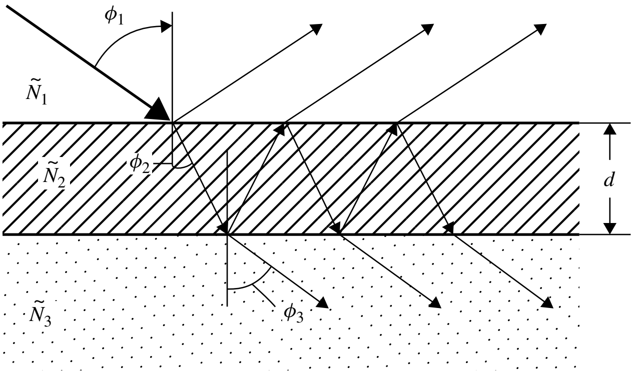


Figure 3.6. Reflections and transmissions in ambient (1), film (2), and substrate (3) [6].

All subscripts and angles mentioned in Eqs. (3.17)–(3.21) are described in Figure 3.6.

Thus, materials with finite light absorption have two unknowns (ε_1 and ε_2, or equivalently, n and k), at each wavelength and one additional unknown in the film thickness. Thus, the total number of unknowns is $2w + 1$. Because this number of unknowns is one too many to be determined from spectroscopic ellipsometry data, it is necessary to employ a dispersion model. Such a model describes the functional dependence of n and k on λ based on P fitting parameters. Therefore, the total number of unknowns becomes $P + 1$. As long as $2w > P + 1$, film thickness and the optical constants may be determined simultaneously by numerically iterating the $P + 1$ fitting parameters to fit spectra [7].

For example, for a thin film on a substrate, the usual objective is to determine thickness d for a known substrate and film dielectric function. To do so, the value of d is found that minimizes the function

$$\sum_{\lambda} |\langle\varepsilon\rangle - \langle\varepsilon\rangle_{\text{calc}}|^2 \tag{3.22}$$

(or similar functions using ρ, or ψ and Δ) [8]. Here, the first term represents measured values, and the second term represents theoretically calculated values. This expression can be minimized using well-known procedures such as Newton's method or the Levenberg–Marquardt algorithm [9].

3.2.1.3. Quartz Crystal Monitor

As described in Chapter 2, the deposition of metals such as aluminum is often accomplished using the evaporation technique. The deposition rate during evaporation operations is commonly measured using a device known as a *quartz crystal monitor.* This device is a vibrating crystal sensor that is allowed to oscillate at its resonant frequency as the frequency is monitored. This resonant frequency then shifts as a result of mass loading as additional mass from the evaporated metal is deposited on top of the crystal. When enough material has been added, the resonant frequency shifts by several percent. By feeding the frequency measurements to the mechanical shutters of the evaporation system, the thickness of the deposited layer, as well as its time rate of change, can be readily monitored. The sensing elements needed to detect such shifts are quite inexpensive and easy to replace. This method is effective for a wide range of deposition rates.

3.2.1.4. Four-Point Probe

The *four-point probe* is an instrument used to measure the resistivity and sheet resistance of diffused layers. As depicted schematically in Figure 3.7, this technique requires a fixed current to be injected into the wafer surface through two outer probes. The resulting voltage is measured between two inner probes. If the probes have a uniform spacing (s, in cm), and the sample is infinite, then the resistivity in Ω·cm is given by [11]

$$\rho = 2\pi s V/I \tag{3.23}$$

for $t \gg s$ and

$$\rho = (\pi t/\ln 2)V/I \tag{3.24}$$

for $s \gg t$. For shallow layers such as this, Eq. (3.24) means that the sheet resistance (R_s) is then given by

$$R_s = \rho/t = (\pi/\ln 2)V/I = 4.53V/I \tag{3.25}$$

Although the approximations used in Eqs. (3.24) and (3.25) are valid for shallow diffused layers in silicon, different correction factors must be used for sheet resistance measurements on bulk wafers.

It should be noted that monitor wafers used for sheet resistance measurements can also be used to determine junction depth (x_j). After the wafers are diffused or implanted with dopants, the thickness of the diffused region is defined as the junction depth. This parameter may be determined from sheet resistance measurements by replacing t with x_j in Eq. (3.25).

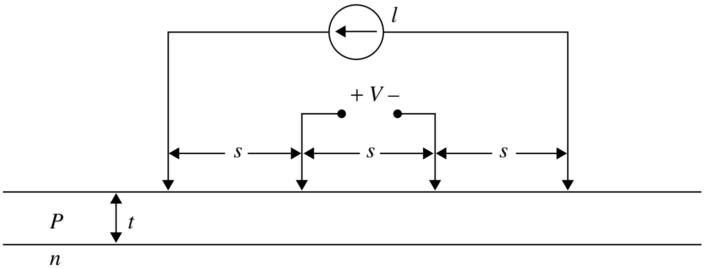


Figure 3.7. Schematic of four-point probe measurement [11]. In this example, the sheet resistance of a *p*-type epitaxial layer of thickness *t* on an *n*-type substrate is measured.

3.2.2. Patterned Thin Film

We continue our discussion of wafer state measurements with those measurements that are performed primarily on wafers that have previously been patterned using photolithography and etching to form specific structures or devices.

3.2.2.1. Profilometry

Profilometry is a very common method of postprocess film thickness measurement. In this technique, a step feature in the grown or deposited film is first created, either by masking during deposition or by etching afterward. The profilometer then drags a fine stylus across the film surface (see Figure 3.8). When the stylus encounters a step, a signal variation (based on a differential capacitance or inductance technique) indicates the step height. This information is then displayed on a chart recorder or cathode-ray tube (CRT) screen. Films of thicknesses greater than 100 nm can be measured with this instrument. The measurement of thinner films is difficult because of vibration, surface roughness, and the precision required in leveling the instrument. Some more recently developed surface profilometers use atomic force microscopy (see Section 3.2.2.2 below).

3.2.2.2. Atomic Force Microscopy

Atomic force microscopy (AFM) is a method for measuring surface properties and/or profiles with atomic-scale topographical definition. In this technique, a sharp tip built at the end of a soft cantilever arm is vibrated perpendicular to

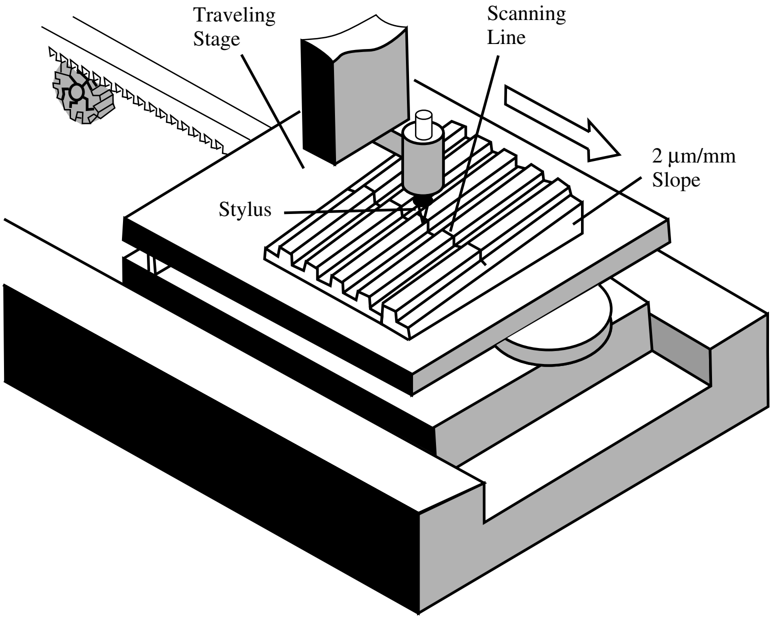


Figure 3.8. Schematic of surface profilometer [12].

the surface at close to the resonant frequency of the cantilever–tip mass as the probe tip traverses laterally across the feature to be characterized. The tip is in atomically close proximity to the surface, so a van der Waals electrostatic force is created between them. This force, which has a strong dependence on the gap between the tip and surface topography, modifies the resonant frequency of the system. The changes in resonance are monitored by an interferometric detection technique that provides a corresponding displacement signal, resulting in a direct measure of the atomic-scale surface topography. A schematic of an AFM system is shown in Figure 3.9. Figure 3.10 shows a typical AFM scan of a

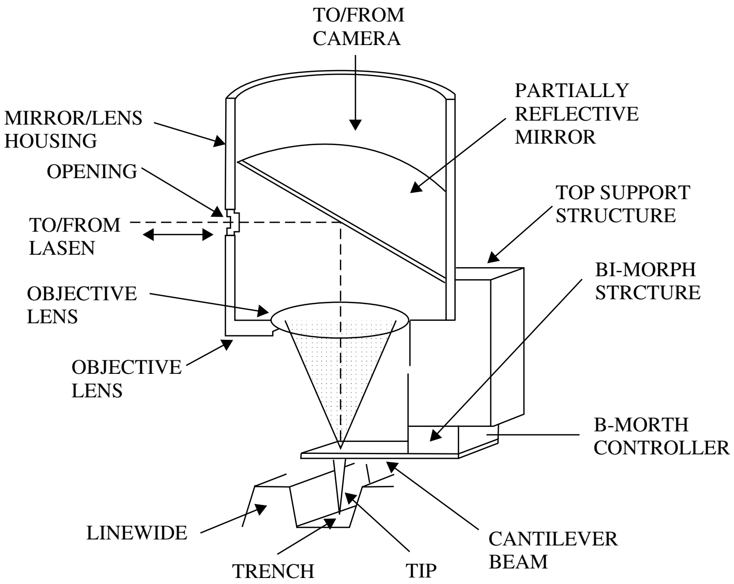


Figure 3.9. Schematic of atomic force microscopy system [13].

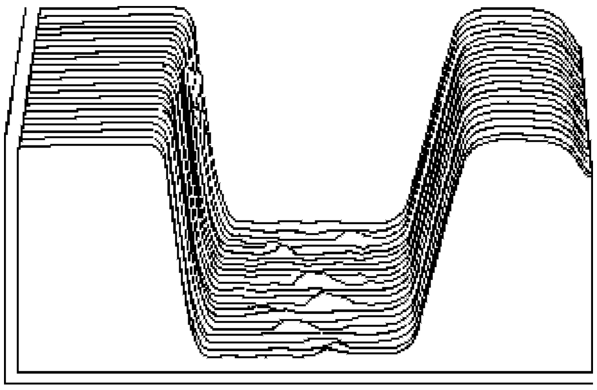


Figure 3.10. Typical AFM image of a surface feature (in this case, a trench) [13].

surface structure. One disadvantage of this technique compared to conventional methods is its low throughput.

3.2.2.3. Scanning Electron Microscopy

Scanning electron microscopy (SEM) is a key technique for assessing minimum feature size in semiconductor manufacturing. The minimum feature size is often expressed in terms of the critical dimension (CD) or minimum *linewidth* that can be resolved by the photolithography system. The decrease in linewidths toward the scale of fractions of a micrometer has rendered conventional optical microscopes nearly obsolete. However, linewidth measurements based on SEM can overcome the limitations of optical techniques for submicrometer geometry features.

The fine imaging capability of the SEM is due to the fact that the wavelength of electrons is four orders of magnitude less than that of optical systems. At such small wavelengths, diffraction effects are usually negligible and spatial resolution is excellent. Features as small as 100 nm can be readily resolved [13]. The electron beam may be based on thermionic or field emission sources. A schematic of a typical field emission SEM is shown in Figure 3.11.

As shown in this figure, the electron gun consists of a tip, first anode, and second anode. A voltage is established between the tip and first anode to facilitate field emission from the tip. An accelerating voltage is then applied between the tip and the second anode to accelerate the electrons. The electron beam emitted from the tip passes through the aperture provided at the center of the first anode, is accelerated, and passes through the center aperture of the second

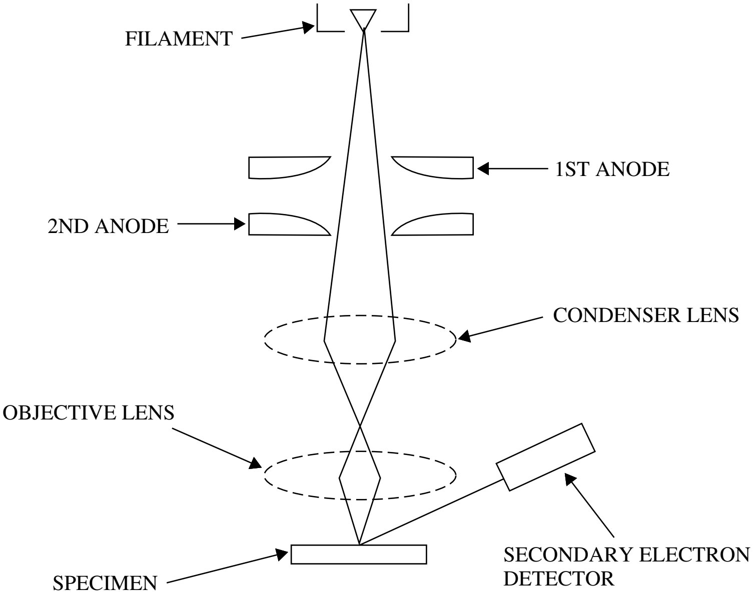


Figure 3.11. Schematic of field emission SEM optics [13].

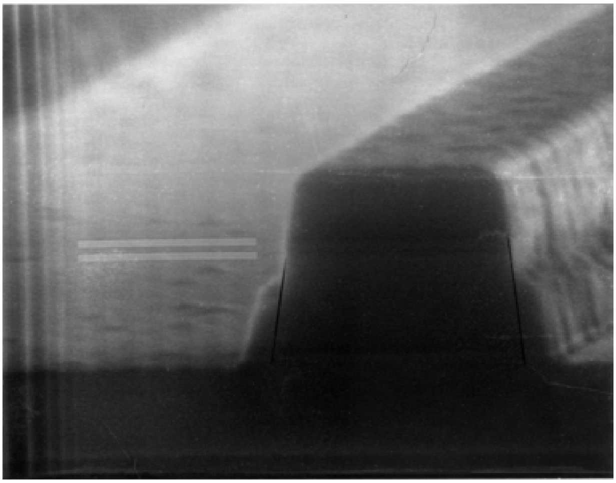


Figure 3.12. Sample SEM output (the parallel lines are calibration marks).

anode to the condenser lens. Electron beams are collected by the condenser lens and aggregated into a small spot on the objective lens. Figure 3.12 shows a typical digital photo output of an SEM. The CD of the feature is usually determined by an arbitrary edge criterion. While lateral resolution offers a tremendous benefit, it must be pointed out that SEM still suffers from several disadvantages, including high cost, low throughput (only ~30 wafers per hour), and the destructive nature of the measurement (i.e., wafers must be cleaved to expose the feature to be imaged).

3.2.2.4. Scatterometry

Scatterometry is another optical measurement technique. It is used for patterned features based on an analysis of the light diffracted (or *scattered*) from a periodic structure such as a grating of photoresist lines. Figure 3.13 shows a schematic of an angle-resolved scatterometer, which measures the intensity of the light diffracted as a function of incident angle and polarization. Scatterometry is used to characterize surface roughness, defects, particle density on the surface, film thickness, or the CD of the periodic structure.

The most common type of angle-resolved scatterometer is called a "2θ" scatterometer due to the two angles (incident and measurement) associated with the method. An incident laser is focused on a sample and scanned through some range of incident angles (θ_i). The light is scattered by the periodic patterns into distinct diffraction orders at angular locations specified by the grating equation

$$\sin\theta_i + \sin\theta_n = n\lambda/d \tag{3.26}$$

where θ_i is taken to be negative, θ_n is the angular location of the nth diffraction order, λ is the wavelength of the incident light, and d is the spatial period (or pitch) of the periodic structure. Because of the complex interaction between the

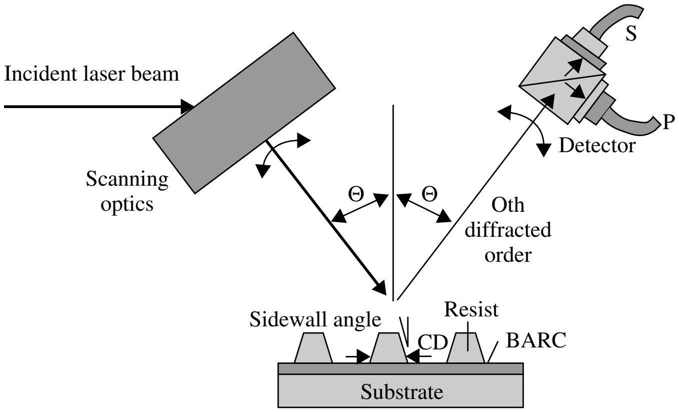


Figure 3.13. Schematic of a 2θ angle-resolved scatterometer [14].

incident light and the periodic features, the fraction of power diffracted into each order is a function of the dimensions of the structure and thus may be used to characterize them.

Capturing diffracted light "signatures" (such as those depicted in Figure 3.14) is just the first phase of scatterometry. In the subsequent analytical phase, a diffraction model is used to interpret the experimental signatures in terms of key parameters such as CD or film thickness. Doing so requires a library of theoretical signatures for comparison to the measured data. The generation of such a library is accomplished by first specifying nominal film stack dimensions and the expected variation of each parameter to be measured. A computerized diffraction model is then used to produce the library of scatter signatures that encompasses all combinations of these parameters for subsequent analysis.

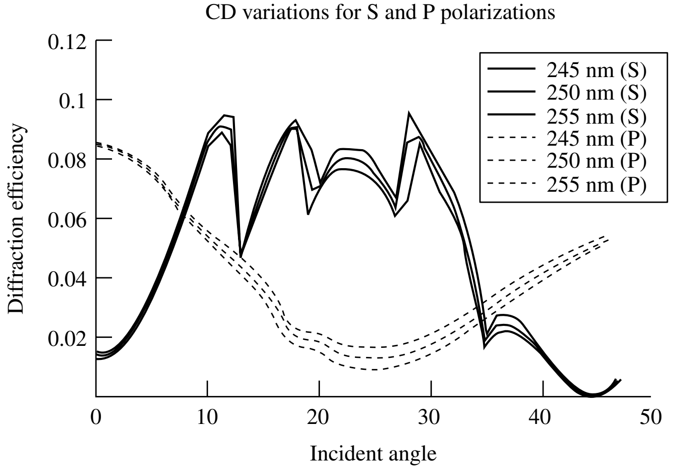


Figure 3.14. Sample scatterometry signatures for 5-nm photoresist CD variations [14].

3.2.2.5. Electrical Linewidth Measurement

Another CD characterization technique depends on direct-current electrical measurement. The most common test structure for this measurement is shown in Figure 3.15. In this configuration, two structures are combined to perform resistance measurements. The upper portion, a four-terminal Van der Pauw structure, is used to measure sheet resistance. This structure is designed to account for doping or film thickness variations. The lower structure is a four-terminal crossbridge linear resistor pattern used to determine the average linewidth (W).

The length of the line segment (L) between pads 4 and 6 is known. When a known current is applied through pads 3 and 5, the resulting voltage is measured at pads 4 and 6. The average linewidth may then be calculated as the product of the measured sheet resistance and length, divided by the measured resistance (V_{46}/I_{35}). The key advantages of electrical linewidth measurement are resolutions on the order of 1 nm and short cycle time. The main disadvantages are the requirement that the film be conductive and the need for physical contact with the wafer.

3.2.3. Particle/Defect Inspection

Contamination is a major concern in semiconductor manufacturing, and billions of dollars are spent annually by manufacturers in order to reduce it. Contamination often takes the form of particles that can appear on the surface of wafers and cause defects in devices or circuits. The fraction of the product that is sensitive to particles depends in part on the particle size. A general rule of thumb is that particles as small as one-tenth the size of a structure can cause the structure to fail. With the industry currently immersed in manufacturing devices with submicrometer features, even nanometer-scale particles are of great concern. Inspection and characterization of particles are therefore critical.

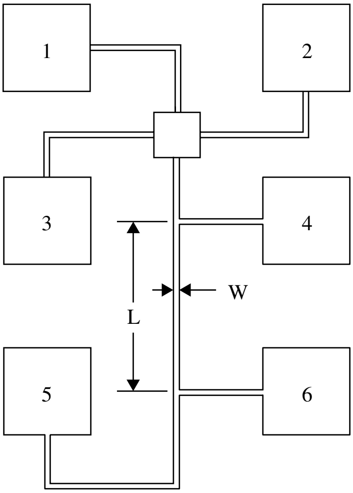


Figure 3.15. Electrical linewidth measurement test structure [13].

3.2.3.1. Cleanroom Air Monitoring

One method of controlling particulate contamination is performing manufacturing operations in a *cleanroom* environment, such as the one schematically depicted in Figure 3.16. Air enters the cleanroom through high-efficiency particulate air (HEPA) or ultra-high-efficiency particulate air (ULPA) filters. The air is forced to flow laminarly (as opposed to turbulently) so that lateral dispersion of contaminants generated in the room is minimized. Cleanrooms are categorized by their "class," which quantifies the number of particles of a given size per cubic foot of air. Various aspects of cleanroom performance affect product quality, and as feature sizes continue to decrease, cleanroom specifications are likewise becoming progressively tighter.

Despite the use of cleanrooms, semiconductor fabrication processes, as well as manufacturing personnel themselves, still generate materials that can contaminate products. Such contamination may originate from process gases and vapors, process liquids, processes that break up bulk material (such as sputtering), deposition processes, metallic impurities, wafer handling, or tool wear, to name just a few. The usual methods for quantitatively determining cleanroom air quality involve sampling via optical particle counters and sampling onto "witness plates" that are later read by surface particle counters.

In the latter approach, a preinspected clean silicon wafer is placed in a location to be monitored. After a fixed time period, the plate is removed and reinspected. The particles per unit area added to the plate are counted. Surface particle counters can inspect an entire plate within minutes with nearly complete detection of particles of sizes a low as fractions of micrometers.

For gases, liquids, and many types of surfaces, optical particle counters are used. Using these devices, particles are illuminated as they pass through a focused

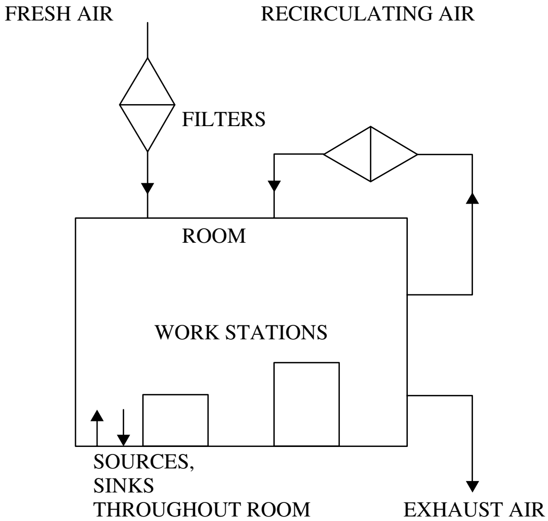


Figure 3.16. Cleanroom schematic [13].

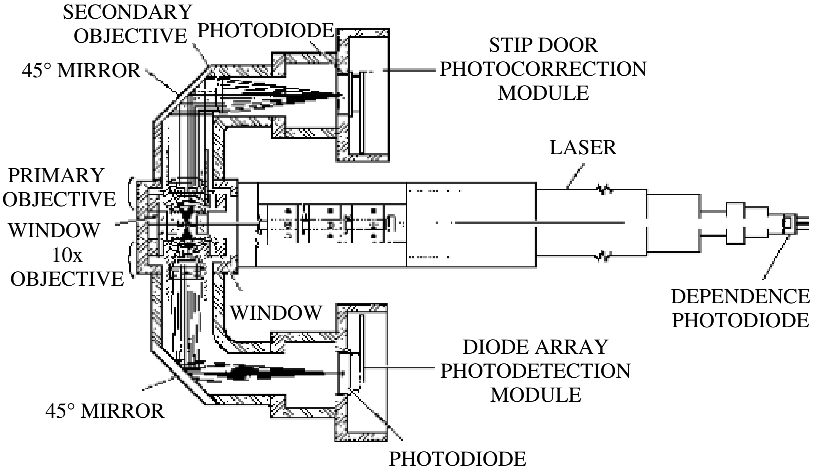


Figure 3.17. Optical particle counter [13].

laser beam (see Figure 3.17). The light scattered from the particles is then measured and correlated with the number of particles present. The amount of light scattered into the sensing element will depend on the light (intensity, wavelength, polarization), the characteristics of the particles (size, shape, orientation, refractive index), and the measurement geometry (position and solid angle subtended by the optics with respect to the beam and the particle). In addition to cleanroom monitoring, this technique is also used for *in situ particle monitoring* (ISPM) inside of processing equipment that produces particles, such as ion implantation or sputtering equipment.

3.2.3.2. Product Monitoring

In addition to monitoring contamination in the ambient environment, it is perhaps more crucial to monitor particles that actually wind up on the wafer surface, since these are the particles that can cause circuit defects. Experience has shown that most processing-related defects tend to occur in a few layers of the complete process [15]. For CMOS processes, for example, defects in the gate oxide and interconnect layers represent the vast majority of all defects.

To control the formation of such defects, special *inline monitoring* techniques are required. These techniques involve inspection of product wafers at various stages in the process. Two common approaches for local defects are "surfscan" and image evaluation. The surfscan technique uses scattered laser light and analyzes reflections to count the particles on the wafer surface (see Figure 3.18). Surfscan is usually applied to unpatterned wafers. Image evaluation techniques, on the other hand, make use of automated inspection equipment to check the occurrence of local defects on patterned wafers at several critical points in the manufacturing process.

Generic particle counts are useful, but limited. In order to assess the impact of the presence of defects caused by particles, specially designed test structures

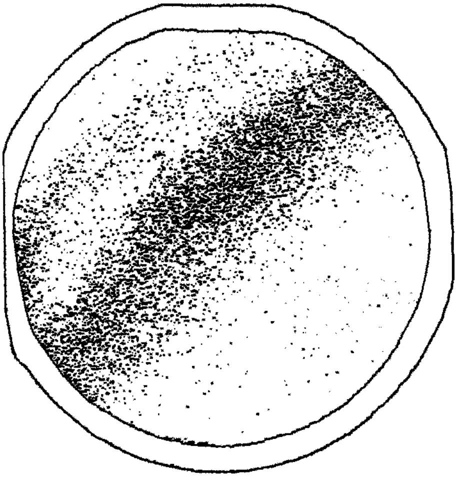


Figure 3.18. Sample surfscan [15].

are used. These structures, also known as *process control monitors* (PCMs), include single transistors, single lines of conducting material, MOS capacitors, via chains, and interconnect monitors. Product wafers typically contain several PCMs distributed across the surface, either in die sites or in the scribe lines between die (see Figure 3.19).

Process quality can be checked at various stages of manufacturing through inline measurements on PCM structures. Three typical interconnect test structures are shown in Figure 3.20. Using such test structures, measurements are performed to assess the presence of defects, which can be inferred by the presence of short

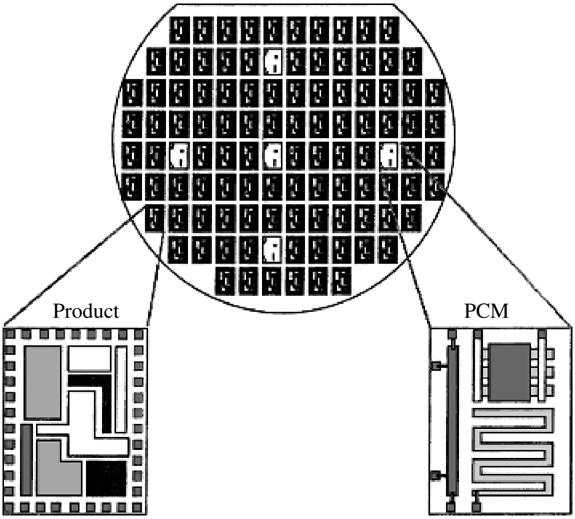


Figure 3.19. Configuration of products and PCMs on a typical wafer [15].

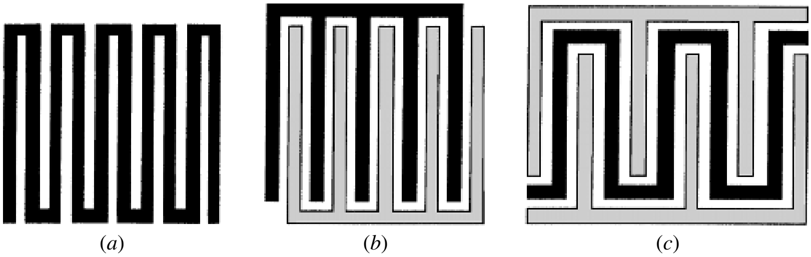


Figure 3.20. Basic test structures for interconnect layers: (a) meander structure; (b) double-comb structure; (c) comb–meander–comb structure [15].

circuits or open circuits using simple resistance measurements. For example, the meander structure facilitates the detection of open circuits through increased end-to-end resistance of the meander. The double-comb structure can likewise be used to detect shorts (short circuits), since any extra conducting material bridging the two combs will reduce the resistance between combs significantly. The comb–meander–comb structure combines the capabilities of the other two structures and permits the detection of both shorts and opens. Various combinations of widths of lines and spaces in these test structures allow the collection of statistics on defects of various sizes.

3.2.4. Electrical Testing

In the preceding section, the concept of test structures for process monitoring was introduced. Although this introduction was presented in the context of particle and defect monitoring, it should not be construed that this is the only use of test structures. In fact, electrical measurements performed on test structures are a major mechanism for assessing *yield* (see Chapter 5) and other indicators of product performance as well. Such measurements are performed on an inline basis and also at the conclusion of the fabrication process. In addition, electrical testing of the final product is crucial to ensure quality. These concepts are discussed in more detail below.

3.2.4.1. Test Structures

Figures 3.20–3.26 are examples of electrical test structures used for process monitoring. However, these by no means represent a comprehensive set, as dozens of possible structures exist for monitoring hundreds of process variables.

Figure 3.21 shows a high-density bipolar transistor chain used to monitor the leakage current between transistor terminals (emitter–base leakage, emitter–collector leakage, etc.). The emitters and bases are wired into parallel chains. Collectors are contacted via the substrate, which eliminates metal short interference in the emitter–collector leakage test. The collectors are also wired to the second level of metal to test collector isolation leakage. Transistor chains can also be used to monitor base–base shorts, as shown in Figure 3.22. In this example, shorts due to polysilicon bridging can be detected by forming the polysilicon

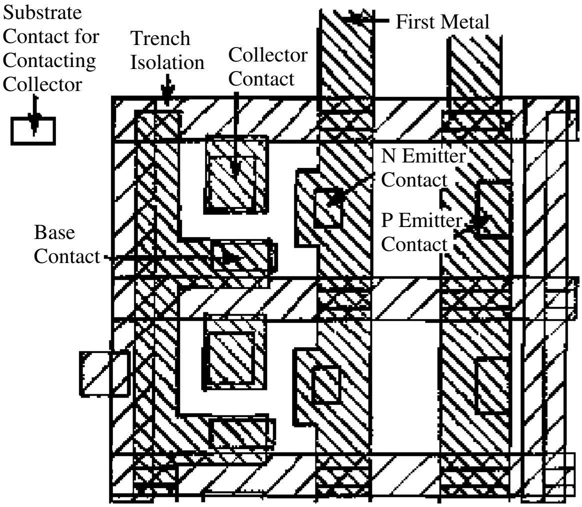


Figure 3.21. Bipolar transistor chain [13].

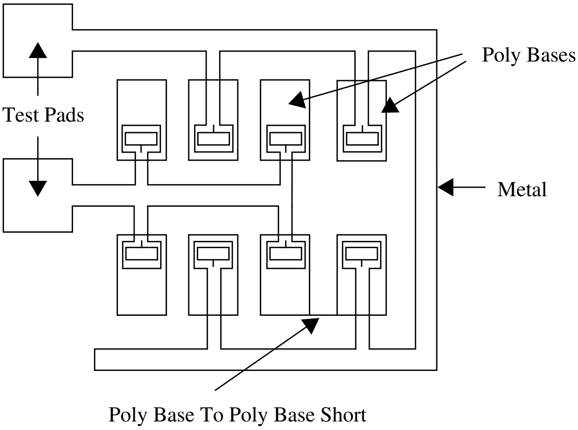


Figure 3.22. Polysilicon base to polysilicon base short chain [13].

bases on field oxide to eliminate the possibility of shorts through the substrate. It is also important to monitor transistor contacts for open circuits. Series-type chains similar to the meander structure in Figure 3.20 can be used for this purpose by connecting the contacts for the various transistor terminals. Figure 3.23 shows an example of a collector contact chain. Note that although the structures depicted in Figures 3.21–3.23 were designed for bipolar circuits, analogous structures can be fabricated to evaluate MOS circuits by wiring up chains connecting their source, drain, and gate terminals in similar chains.

Figure 3.24 is a typical example of a via chain structure used to test connectivity between metal layers. This chain also includes a first-level metal stripe

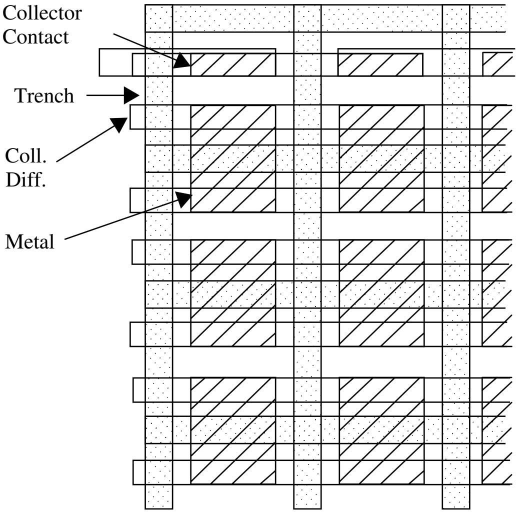


Figure 3.23. Collector contact chain [13].

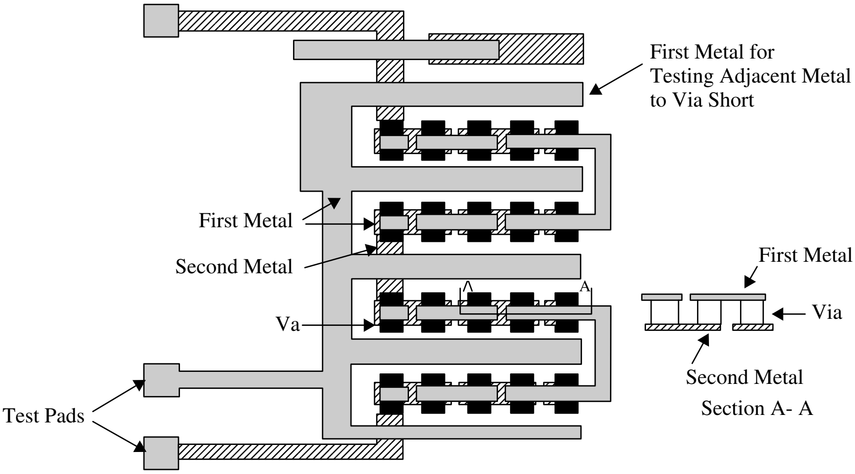


Figure 3.24. Via chain [13].

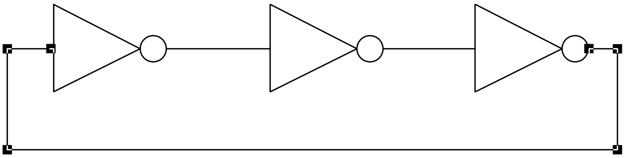


Figure 3.25. Ring oscillator.

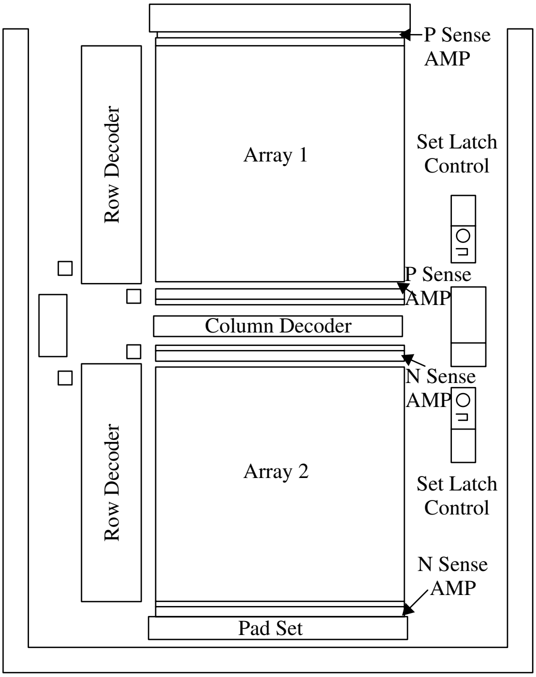


Figure 3.26. Array diagnostic monitor [13].

running parallel to the chain as a mask misalignment monitor. An adjacent metal stripe runs on every level, but never along the full length of the chain. They instead appear at certain sections, alternating with each other.

In addition to defect monitoring, test structures are also used to assess functional characteristics of the semiconductor devices and their dependence on processing conditions. These can be individual devices or simple subcircuits. A common example of such a structure is a *ring oscillator*, which is used to measure speed and capacitive loading effects. A ring oscillator is essentially a chain of inverters (see Figure 3.25). It is formed by connecting an odd number of inverters in a loop. In general, a ring with N inverters will oscillate with a period of $2N\tau_p$ and a frequency of $1/2N\tau_p$, where τ_p is the propagation delay through a single inverter. Inverter chains can also be used to monitor transistor current gain or voltage drops across transistors [13].

An example of a more elaborate functional test structure is the array diagnostic monitor (ADM) shown in block diagram form in Figure 3.26. The ADM, which is used to assess CMOS DRAM circuits, has DC and AC diagnostic capabilities. It is essentially a simplified, yet fully functional duplicate version of a memory array. ADM testing allows for rapid process feedback and ultimately translates into accelerated process improvement.

3.2.4.2. Final Test

Functional testing at the completion of manufacturing is the final arbiter of process quality and yield. The purpose of final testing is to ensure that all products perform to the specifications for which they were designed. For integrated circuits, the test process depends a great deal on whether the chip tested is a logic or memory device. In either case, automated test equipment (ATE) is used to apply a measurement stimulus to the chip and record the results. The major functions of the ATE are input pattern generation, pattern application, and output response detection. A block diagram for a basic ATE is shown in Figure 3.27.

For logic devices, during each functional test cycle, input vectors are sent through the chip by the ATE in a timed sequence. Output responses are read and compared to expected results. This sequence is repeated for each input pattern. It is often necessary to perform such tests at various supply voltages and operating temperatures to ensure device operation at all potential regimes. The number and sequence of failures in the output signature are indicative of manufacturing process faults.

The test process for memory products is very similar to that used for logic. However, one important variation is the availability of the redundancy technique. For dynamic RAM circuits, a widely used approach is to add a few extra word and/or bit lines that can replace faulty lines in the main array. Replacement of these faulty lines is accomplished by fusing them to redirect a bad word or bit address to a redundant line. Testing the redundant lines requires two passes. During the first pass, the addresses of errors are recorded and stored. As long as the number of faults is less than the number of extra lines, the chip is repairable. Although redundancy adds considerable cost and complexity to testing, the yield benefit achieved more than compensates for this.

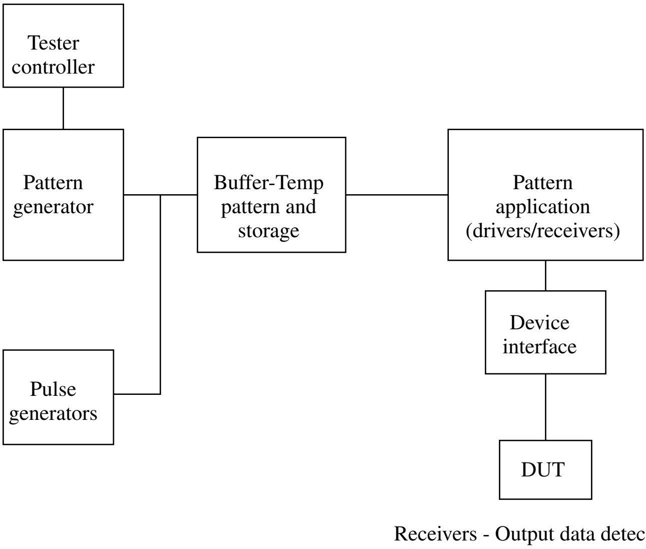

Tester controller
Pattern generator
Buffer-Temp pattern and storage
Pattern application (drivers/receivers)
Device interface
Pulse generators
DUT
Receivers - Output data detection

Figure 3.27. Block diagram of basic test system (DUT = “device under test”) [13].

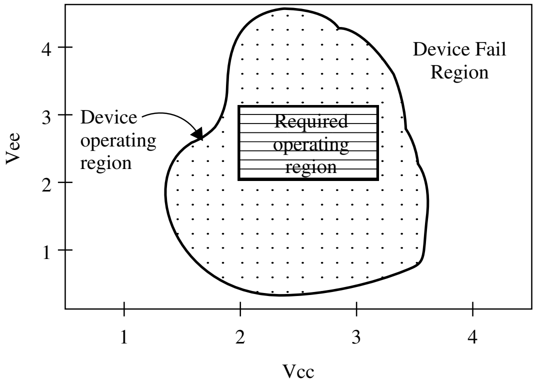


Figure 3.28. Example of two-dimensional voltage *shmoo* plot for hypothetical bipolar chip [13].

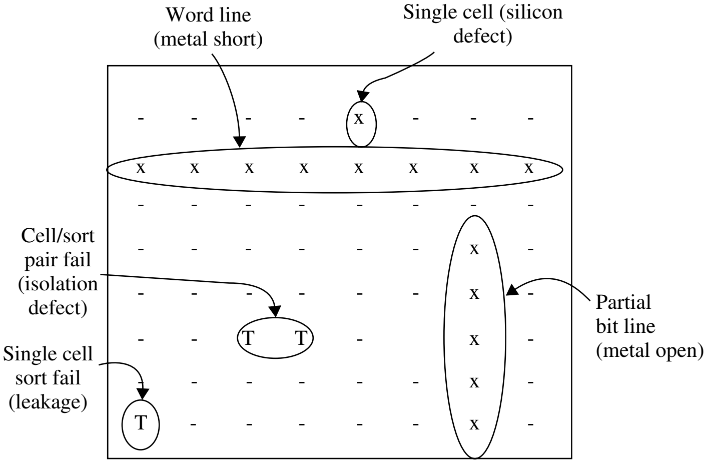

Word line (metal short)
Single cell (silicon defect)
Cell/sort pair fail (isolation defect)
Single cell sort fail (leakage)
Partial bit line (metal open)
x
T
T T

Figure 3.29. Cell map showing examples of failure patterns and defect types [13].

Test results may be expressed in a variety of ways. A couple of examples are shown in Figures 3.28 and 3.29. Figure 3.28 shows a plot of a two-dimensional plot called a "shmoo" plot for a hypothetical bipolar product. In a shmoo plot, the outlined shaded region is where the device is intended to operate, while the blank area outside represents the failure region. Another typical test output is the cell map shown in Figure 3.29. Cell maps are very useful in identifying and isolating device failures, particularly in memory arrays. In addition, the patterns generated in the cell map may be compiled, catalogued, and later compared to a library of existing defect types, thereby aiding in the diagnosis of faults.

3.3. EQUIPMENT STATE MEASUREMENTS

Rather than characterizing the state of the product wafers themselves, equipment monitors measure the status of tools while they are processing these wafers. Such monitors are the most immediate measure of process quality and therefore provide the shortest feedback loop for maintaining control. In other words, the impact of out-of-control conditions can be minimized if such conditions are promptly identified by tool monitors and immediate corrective action is taken.

Certain physical parameters are routinely measured as a part of equipment monitoring. The following are a few commonly monitored process variables at various stages of the manufacturing process:

Lithography

- Exposure energy
- Exposure dose and intensity
- Time
- Magnification
- Aperture

Wet Stations

- Fluid level
- Temperature gradients
- Flowrates
- Development/etch rates
- Time to endpoint

Deposition

- Gas flowrate
- Pressure
- Temperature

Implantation

- Accelerating voltage
- Beam current

Diffusion

- Source composition
- Pressure
- Flowrate
- Temperature

The combined effects of these tool variables eventually lead to measurable impact on the characteristics of product wafers. The process engineer must therefore have available reliable methods for monitoring these variables in order to facilitate process control. The following sections describe several equipment state measurements used for monitoring such characteristics.

3.3.1. Thermal Operations

Thermal operations refer to any process step that occurs at an elevated temperature. Examples include epitaxial growth, chemical vapor deposition, evaporation, and annealing. This subsection describes the measurement of key process variables during these operations.

3.3.1.1. Temperature

In situ measurements of conditions such as temperature can be used to infer the quality of the wafers being produced in thermal processes. In many types of thermal processing equipment, temperature is measured using a thermocouple embedded in the wafer holder (or *susceptor*). A *thermocouple* is a circuit consisting of a pair of wires made of different metals joined at one end (the "sensing junction") and terminated at the other end (the "reference junction") in such a way that the terminals are both at a known reference temperature. Leads from the reference junction to a load resistance (i.e., an indicating meter) complete the thermocouple circuit. Due to the *thermoelectric effect* (or Seebeck effect), a current is induced in the circuit whenever the sensing and reference junctions are different temperatures. This current varies linearly with the temperature difference between the junctions.

In some cases (such as in rapid thermal processes), the use of a thermocouple is not possible because there is no susceptor. Alternative temperature sensors used in such situations include thermopiles and optical pyrometers. A *thermopile*, which also operates via the Seebeck effect, consists of several sensing junctions made of the same material pairs located in close proximity and connected in series in order to multiply their output.

The second alternative method to the thermocouple is pyrometry. Pyrometers operate by measuring the radiant energy received in a certain band of energies, assuming that the source is a graybody of known emissivity. The input energy can then be converted to a source temperature using the Stefan–Boltzmann relationship [16]. Most commercial systems monitor the mid-infrared band (3–6 μm). One major issue in using pyrometry is that the effective emissivity of the source must be accurately known. The effective emissivity includes both intrinsic and extrinsic contributions. *Intrinsic emissivity* is a function of the material, surface finish, temperature, and wavelength. *Extrinsic emissivity* is affected by the amount of radiant energy from other sources reflected back to the spot being measured (which can increase the apparent temperature). In addition, the presence of multiple layers of different thin-film materials can also alter the apparent emissivity due to interference effects.

3.3.1.2. Pressure

Pressure in vacuum systems used in thermal operations can be measured using a variety of transducers, including capacitance manometers, thermal conductivity gauges, and ionization gauges. Capacitance manometers are mechanical gauges that sense the deflection caused by the pressure difference between the chamber to be measured and a reference volume. These devices detect the movement of a thin metal diaphragm to do so. Although they can be used to detect pressures as low as 1 mTorr, they are also often used to measure pressures as high as 1 Torr.

Thermal conductivity gauges derive the thermal conductivity of the ambient gas by passing a current through a wire and measuring its temperature. Pressure may then be inferred from the conductivity measurement.

However, neither mechanical deflection nor thermal conductivity gauges are able to measure pressures below 1 mTorr. This type of application requires an *ionization gauge*, which operates using an electron stream to ionize the gas in the gauge and an electric field to collect the ions. The ion current is a function of the pressure in the chamber. The pressure that can be measured in this way is limited only by the ability to sense small ion currents, so ionization gauges can detect pressures as low as 10^{-12} Torr.

3.3.1.3. Gas Flow

Thermal systems of various types, as well as plasma etchers, require controlled rates of introduction of process gases into the reaction chambers. This is most commonly achieved using an instrument called a *mass flow controller*, which consists of a flowmeter, a controller, and a valve, and it is located between the gas source and the chamber itself. Gas flow is measured in units of volume/time. The most common unit is the *standard cubic centimeter per minute* (sccm), defined as the flux of one cm^3 of gas per minute at 273 K.

There are two primary types of mass flowmeters: (1) the differential pressure type and (2) the thermal type. The differential pressure flowmeter relates a pressure drop at a physical flow restriction to rate of mass flow. The thermal type, which is more widely used in semiconductor manufacturing, relies on the ability of a flowing gas to transfer heat. As shown in Figure 3.30, the thermal flowmeter consists of a larger gas flow tube in parallel with a small sensor. A heating coil is wrapped around the sensor midway along its length, and temperature sensors are placed both upstream and downstream of the heated point. Flowing gas causes the temperature distribution in the sensor tube to change as a result of thermal transfer between the heated wall and the gas stream. The temperature downstream from the heated region becomes higher than the upstream temperature as the flowing gas conducts heat away. It can be shown that the rate of mass flow (m_f) is then given by

$$m_f = (\kappa W_h\ \Delta T)^{1.25} \tag{3.27}$$

where W_h is the heater power, ΔT is the temperature difference between the two sensors, and κ is a constant that depends on the heat transfer coefficients and the specific heat, density, and thermal conductivity of the gas. Assuming that the remaining parameters remain constant over the flow range of interest, the mass flowrate can be obtained by measuring the temperature difference. The two temperature sensors, which are usually *resistance thermometers* (see Section 3.3.2.1),

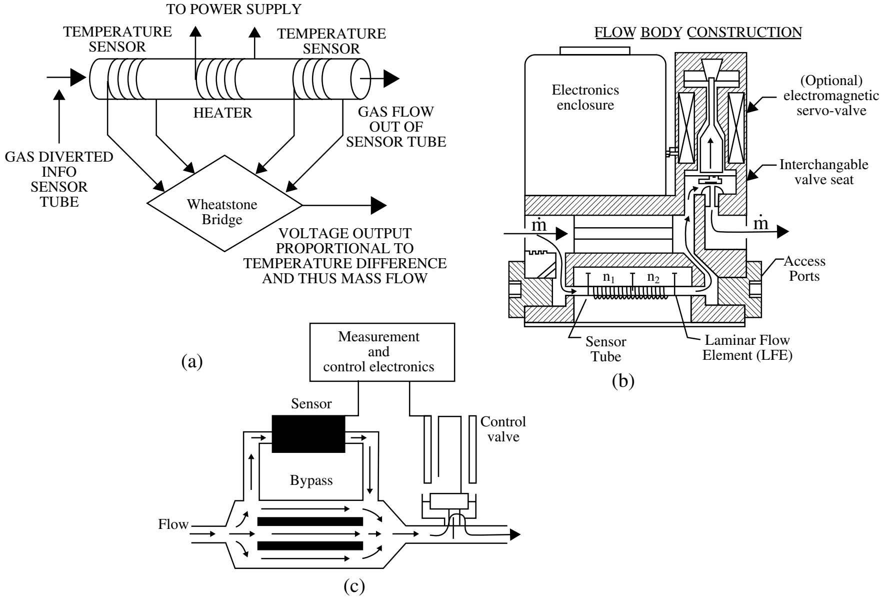

TO POWER SUPPLY
TEMPERATURE SENSOR
TEMPERATURE SENSOR
HEATER
GAS FLOW OUT OF SENSOR TUBE
GAS DIVERTED INFO SENSOR TUBE
Wheatstone Bridge
VOLTAGE OUTPUT PROPORTIONAL TO TEMPERATURE DIFFERENCE AND THUS MASS FLOW
(a)
FLOW BODY CONSTRUCTION
Electronics enclosure
(Optional) electromagnetic servo-valve
Interchangable valve seat
ṁ
ṁ
n1
n2
Access Ports
Sensor Tube
Laminar Flow Element (LFE)
(b)
Measurement and control electronics
Sensor
Control valve
Bypass
Flow
(c)

Figure 3.30. Mass flow controller: (a) operational principles; (b) cross-sectional drawing; (c) schematic diagram [12].

are connected to one port of an unbalanced Wheatstone bridge, and the temperature difference is converted into a voltage signal. As the flowrate is determined, its value is compared to a setpoint value and adjusted as necessary to maintain that value by the controller.

3.3.2. Plasma Operations

As discussed in Chapter 2, plasma etching has emerged as a critical process in the production of integrated circuits. This emergence has stemmed from a continuous need to fabricate devices with extremely small dimensions. However, without sufficient online monitoring and control, etch equipment can produce unacceptably large volumes of defective products, leading to millions of dollars lost as a result of misprocessing. As a result, in addition to the measurements described above for thermal operations, plasma etching systems often employ some unique supplemental equipment monitoring devices.

3.3.2.1. Temperature

In many plasma etching systems, the process temperature is controlled by means of a system that removes heat from the lower electrode by circulating deionized water. This closed-loop recirculation system is sometimes referred to as a "chiller." The chiller maintains a preset temperature, often room temperature. This temperature is monitored using a standard resistance thermometer device (RTD).

RTDs have either conductive or semiconductive elements for which the resistivity (ρ) versus temperature characteristic is given by the well-known relationship

$$\rho(T) = \rho_0(1 + \alpha\ \Delta T) \tag{3.28}$$

where T is the temperature in degrees Kelvin, ρ_0 is the resistivity at some reference temperature T_0, α is the temperature coefficient of resistivity, and $\Delta T = T - T_0$ is the change in temperature relative to the reference. This relationship provides an accurate temperature measurement with a precision of 0.01 K.

3.3.2.2. Pressure

Pressure in plasma etching chambers is measured using capacitance manometers, as described in Section 3.3.1.2.

3.3.2.3. Gas Flow

Gas flowrates in plasma etching systems are monitored using mass flow controllers, as described in Section 3.3.1.3.

3.3.2.4. Residual Gas Analysis

Mass spectroscopy is a well-established scheme for monitoring plasma etching systems by analyzing the residual gas composition in the etch process chamber. The fundamental principle by which a mass spectrometer operates is based on the separation of gas molecules by atomic mass. An etch system continuously depletes its chamber gases during processing. At the beginning of an etch, the gas in the chamber consists of a mixture of process gas and that resulting from the etch. Toward the end, the gas in the chamber will resemble its mixture prior to etching. This information may be used to detect the etch endpoint using *residual gas analysis* (RGA).

There are two main methods for mass spectroscopic monitoring of plasmas: flux analysis and partial-pressure analysis. *Flux analysis* involves sampling the plasma directly by coupling the emission of plasma particles through a small aperture into the ion optics of a mass spectrometer. This method is primarily a research tool and is best suited for plasma species and energy analysis. On the other hand, *partial-pressure analysis* is accomplished by simple vacuum connections between the spectrometer and the plasma chamber. Because of its simplicity, partial-pressure analysis is the method of choice in most production systems used in semiconductor manufacturing.

Figure 3.31 is a schematic diagram of a *quadrupole mass spectrometer* (QMS), the main apparatus used for partial-pressure analysis. Depending on the operating pressure of the plasma system, there is either a high or low conductance connection between the etch and QMS chamber, which is usually differentially pumped. This results in the dynamic response of a pressure change in the QMS chamber ($\Delta P_Q(t)$) differing from the pressure change in the etch chamber ($\Delta P_D(t)$). For dynamic measurements, it can be shown that [17]

$$P_D(t) = \left(1 + \frac{S_Q}{C_T}\right) P_Q(t) + \frac{V_Q}{C_T}\frac{d[P_Q(t)]}{dt} - \left(1 + \frac{S_Q}{C_T}\right) P_B \tag{3.29}$$

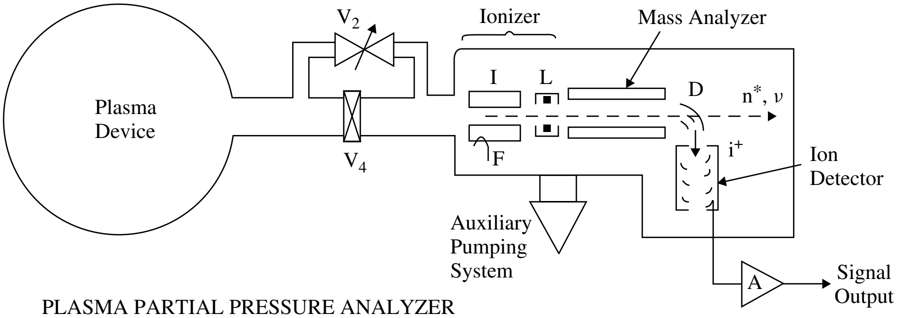


Figure 3.31. Schematic diagram of QMS system used for partial-pressure analysis [17].

where a QMS chamber with volume V_Q is pumped with a pump speed of S_Q and is connected to the etch chamber through a tube with conductance C_T. The last term represents the background pressure (P_B) correction in the QMS. The quantities S_Q and C_T are a function of the gas temperature, pressure, mass, and viscosity of the chamber gas mixture. Equation (3.29) usually must be solved numerically.

Figure 3.32 is an example of the results of RGA using a QMS system for the etching of a GaAs/AlGaAs metal–semiconductor–metal structure in a BCl_3/Cl_2 plasma [18]. The time evolution of the RGA signals from the various reaction product species are clearly evident, indicating the usefulness of this technique for etch process monitoring.

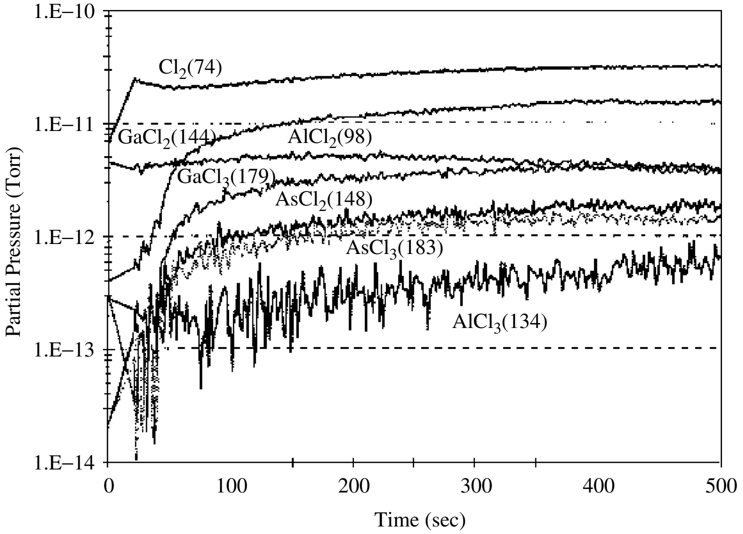


Figure 3.32. RGA signals from a BCl_3/Cl_2 etch of a GaAs/AlGaAs structure (numbers in parentheses represent the atomic mass of the species).

3.3.2.5. Optical Emission Spectroscopy

Optical emission spectroscopy (OES) is one of the oldest and most popular methods of plasma etch monitoring. Fundamentally, OES is a bulk measure of the optical radiation of the plasma species. Since emissions can emanate from etch reactants as well as products, OES measurements are most often used to obtain the average optical intensity at a particular wavelength above the wafer. By setting an optical spectrometer to monitor the intensity at a wavelength associated with a particular reactant or byproduct species, OES serves as a noninvasive, real-time etch endpoint detector. Quantitative measurement of the species concentrations is not required for this purpose. Instead, the intensity of the emission from the key species, perhaps along with its time derivative, can be used empirically to determine the proper point to discontinue the etch process.

A series of such measurements for a particular etch process is referred to as an "endpoint trace," a curve representing the intensity of the optical emission of the key species over time. An example of such a trace is illustrated in Figure 3.33, which depicts fluorine and CN emission intensities during silicon nitride etching. At the beginning of the etch, the gas in the chamber consists of a mixture of process gas and that resulting from the etch. At the end of the etch, the gas mixture again resembles its mixture prior to the start of the process. Therefore, the etch endpoint is characterized by a sharp change in the intensity of the endpoint trace.

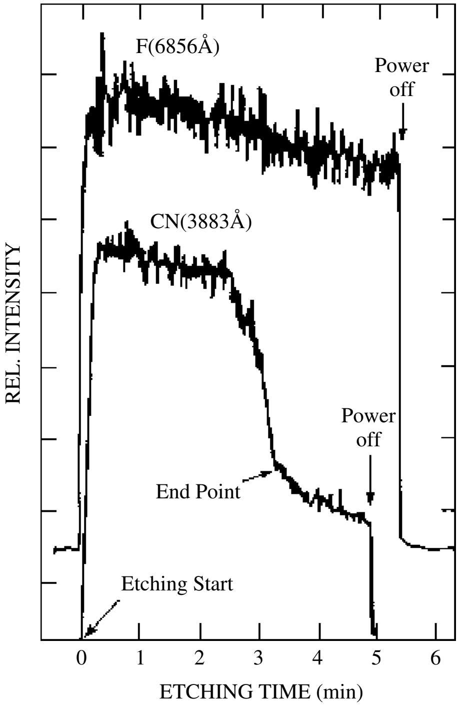


Figure 3.33. OES endpoint trace showing the intensity of the emission of key species in a silicon nitride etch process [17].

OES measurements not only reflect the chemistry of the plasma but also inherently have embedded in them information concerning the operational status of the plasma equipment, pattern density on the substrate, and nonideal fluctuations in the processing conditions (gas flow, pressure, etc.). It is therefore also possible to use OES signals to monitor and diagnose etch equipment problems.

3.3.2.6. Fourier Transform Infrared Spectroscopy

Infrared (IR) spectroscopy is a widely used method for identifying organic compounds, such as those that may result from the etching of polymer films. This method is based on the absorption of infrared radiation by molecules at characteristic wavelengths. Radiation causes various components of such molecules to vibrate and rotate. Since the frequency of vibration/oscillation is dependent on the nature of the chemical bonds present, the presence or absence of absorption in certain well-defined regions of the IR spectrum can be used to determine the presence or absence of chemical groups. The intensity of the absorption peaks is proportional to the amount of material present. Computer databases and search routines are usually used to identify compounds.

In Fourier transform infrared (FTIR) spectroscopy, an infrared source is sent through a beamsplitter to the surface of the wafer being etched and to a movable mirror. The reflected radiation from both surfaces is added and sent to a detector. The distance of the mirror path is swept, and the intensity of the reflected beam as a function of the position of the mirror is monitored. The intensity of the IR peaks can then be used to determine the composition of the film on the wafer surface. An example of typical FTIR output is provided in Figure 3.34.

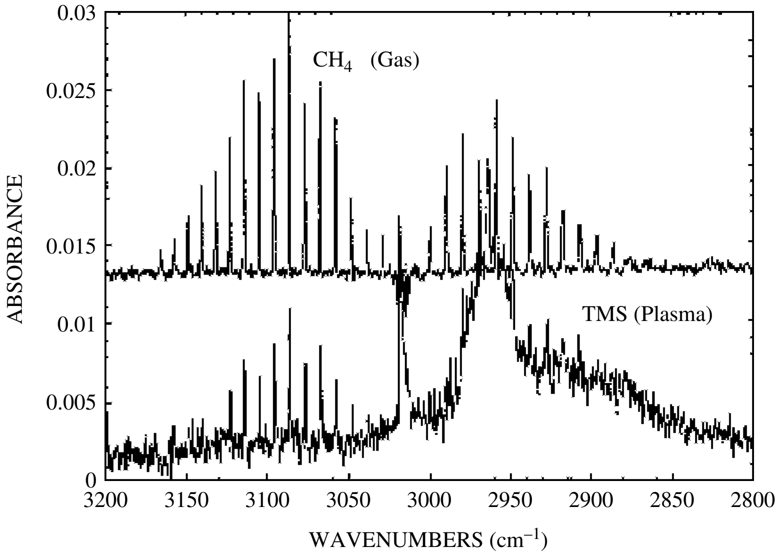


Figure 3.34. IR spectra of CH_4 and tetramethylsilane (TMS) in an electron cyclotron resonance plasma system [19].

3.3.2.7. RF Monitors

Historically, the only aspect of radiofrequency (RF) power monitored in plasma systems is the power delivered from the RF supply to the matching network. This is typically expressed in terms of forward and reflected RF power. The addition of an RF plasma impedance sensor between the matching network and plasma electrode, however, allows new electrical variables to be monitored and controlled. This allows problems such as poor RF connections, electrode condition, and changes in process gas mixture to be detected more easily [20].

Monitoring these parameters facilitates inferences regarding the state of the etch system, such as the degree of ionization of the presence of chamber wall coatings. Etch endpoint can also be detected using changes in RF impedance during the etch cycle. Figure 3.35 shows an example of RF data that can be gathered by a plasma impedance sensor.

3.3.3. Lithography Operations

The success of pattern transfer in photolithographic operations is determined by interactions between four constituents. Those constituents (and examples of the relevant process variables in each) are (1) the wafer (reflectivity, pattern density,

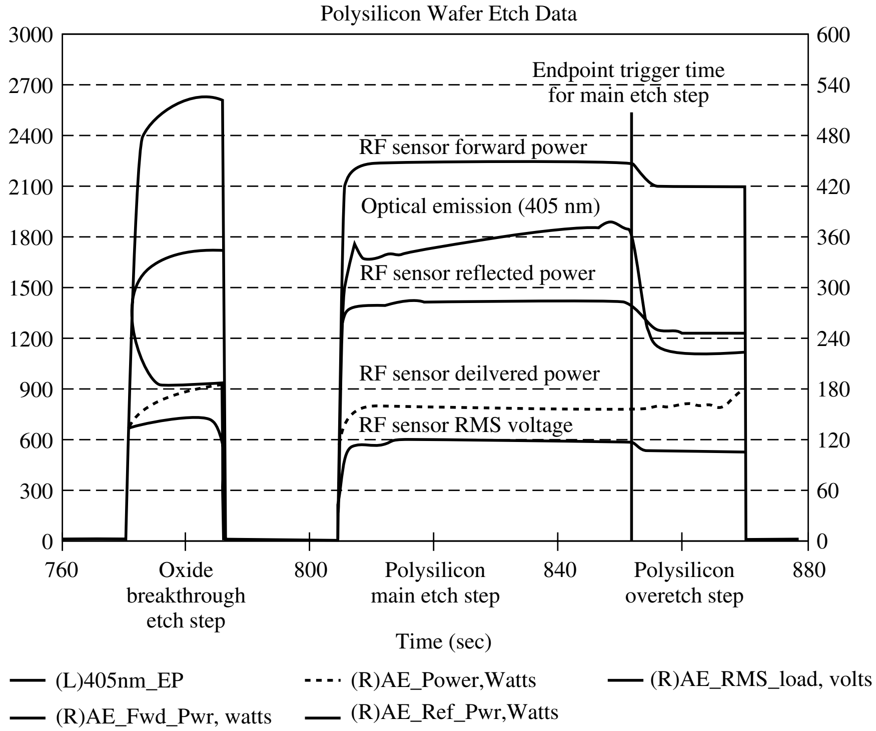


Figure 3.35. Set of RF waveforms gathered during a polysilicon etch [20].

topography), (2) the photoresist (thickness, uniformity, age), (3) the exposure tool (mask variance, wavelength, exposure dose, lens characteristics, barometric pressure), and (4) the developer (concentration, temperature).

The key output measurement in photolithography is linewidth or critical dimension, which is a wafer state variable (see Section 3.2.2). Nevertheless, the CD is significantly impacted by several equipment state variables that must also be monitored to ensure quality. For example, resist thickness and uniformity are controlled in part by the spin speed and ramp of spin coaters (as well as the coating solvent and viscosity of the resist). The primary equipment state measurements, however, are related to the exposure tool. In most modern exposure tools, the monitoring of equipment state variables occurs internally. For example, barometric pressure and lens characteristic changes are now monitored as part of the tool package.

3.3.4. Implantation

In modern ion implantation systems, it is important to monitor and carefully control the dose of the implant. This is accomplished in the end station by placing the wafer undergoing implantation in a Faraday cup, which is simply a cage that captures all the charge that enters it. The ion current into the wafer is measured by connecting an ammeter between the Faraday cup and ground. The dose is the time integral of this current divided by the wafer area.

Accurate measurement of the dose requires that precautions be taken against errors due to secondary-electron ejection. This process involves the creation of large numbers of electrons, many of which have sufficient energy to escape the wafer when a high-energy ion strikes the wafer surface. To prevent secondary-electron dose measurement errors, the wafer is biased with a small positive voltage. This bias (usually tens of volts) is sufficient to attract all the secondary electrons back to the surface of wafer, where they are absorbed.

Another problem often seen in implanting through a photoresist mask is outgassing. Ions striking the surface break apart organic molecules in the resist, leading to the formation of gaseous hydrogen that evolves from the surface and leaves behind carbon. Not only can this hardened carbonize layer be difficult to remove, but outgassing can raise the pressure in the end station enough to cause neutralization of the ion beam through impact with the H_2 molecules, which can result in significant dose rate measurement errors. Modern cryopumps are very effective in pumping away H_2 and other photoresist outgassing products, but these cryopumps must be regenerated at regular intervals to maintain adequate vacuum levels, and this impacts throughput. The beam neutralization effect of outgassing is controlled in some implant systems by the use of a feedback loop that corrects the observed signal at the Faraday cup in response to changes in the beamline pressure. Other systems avoid these problems by monitoring the ion-beam current during those portions of the implant operation when the ion beam is not impinging on the wafer surface and the photoresist outgassing rate is low.

3.3.5. Planarization

As discussed in Chapter 2, planarization operations employ chemical–mechanical polishing (CMP) systems. In CMP systems, some of the key equipment state measurements that must be performed include characterization of polishing pads, determining the condition of the slurry, and endpoint detection.

The condition of CMP polishing pads is a key indicator of removal rate since the porosity of the pad determines the slurry arrival rate at the surface of the wafer. Glazing of the pad tends to occur after several runs, which slows the removal rate. The solution of this problem is frequent conditioning of the pad to obtain consistent roughness. Care must be taken in pad conditioning, however, since processed wafers show greatly increased particle counts immediately after pad conditioning [16].

Slurries for CMP applications consist of particles suspended in various liquids (depending on the specific material being polished). By measuring and controlling the pH of the slurry, particle agglomeration is minimized. In addition, for oxide CMP, the polishing rate increases with increasing pH, particle concentration, and particle size. Therefore monitoring each of these qualities of the slurry is important.

Since CMP is a process for reducing thickness at selected locations on the wafer, it is necessary to identify when a suitable degree of overall planarization has been achieved and the process has reached its endpoint. One method of endpoint detection involves monitoring the current supplied to the motor of the wafer carrier. This motor current monitoring technique is a production-proven method that works well when polishing down to a stop layer (such as polishing a CVD–SiO_2 film on a silicon nitride stop layer in shallow-trench isolation processes) [12]. Circuitry such as a current shunt or Hall effect probe is used to monitor the current supplied to the motor that rotates the wafer carrier. Since the carrier is driven at a constant rotational speed to maintain a constant polishing rate, the drive current is varied to compensate for any load changes on the motor. This makes the current sensitive to frictional changes at the wafer surface. The largest changes occur when one material has been polished away, leaving a layer that has different polishing characteristics. Therefore, substantial changes in drive current are indicative of process endpoint.

SUMMARY

This chapter has provided a survey of sensor metrology and methods of monitoring semiconductor manufacturing processes. After identifying key measurement points in the process flow and differentiating between wafer state and equipment state measurements, a description of such measurement techniques ensued. Measuring key process and equipment state variables enables operators and engineers to ascertain product quality. However, conclusions regarding quality can be drawn only after this measurement data have been collected and analyzed. Methods for data analysis involve the application of various statistical tools. The fundamental concepts that support these tools are the subject of the next chapter.

PROBLEMS

3.1. A thin film of silicon dioxide covers a silicon wafer. Plane lightwaves of variable wavelengths are incident normal to the film. When one views the reflected wave, it is noted that complete destructive interference occurs at 600 nm and constructive interference occurs at 700 nm. Calculate the thickness of the SiO_2 film.

3.2. The correction factor for sheet resistance when thick materials are being measured with a Four-point probe is shown in Figure P3.2. Equation (3.24) must be multiplied by this factor to obtain accurate R_S values from $I-V$ measurements. Given that a uniformly doped silicon layer with a thickness equal to the probe spacings is measured and $V/I = 45$, compute R_S.

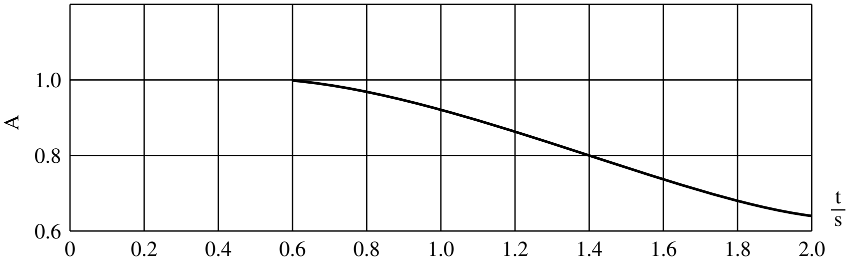


Figure P3.2

3.3. Describe four techniques for measuring the linewidth of patterned features on a substrate. Why is accurate linewidth measurement more difficult on wafer surfaces than on masks?

3.4. A 200-mm-diameter silicon wafer contains chips that are 0.25 cm^2. The wafer is initially clean and is then exposed to room air containing 1000 particles/ft^3 of diameter 0.5 μm and larger. On average, how long will it take to deposit one particle per chip, assuming a laminar air flow of 30 m/min?

3.5. Explain why a thermal conductivity gauge will not work in an ultrahigh vacuum.

3.6. The major source of uncertainty in pyrometry is uncertainty in emissivity. Planck's radiation law gives the spectral radiant exitance as a function of wavelength and temperature ($M_\lambda(\lambda, T)$) as

$$M_\lambda(T) = \varepsilon(\lambda)\frac{c_1}{\lambda^5(e^{c_2/\lambda T} - 1)}$$

where $\varepsilon(\lambda)$ is the wavelength-dependent emissivity of the emitting body and c_1 and c_2 are the first and second radiation constants (given by 3.7142 ×

10^{-16} W-m^2 and 1.4388 × 10^{-2} m·K, respectively). If the wafer temperature is 1000°C, what wavelength is most desirable to minimize the effect of this uncertainty?

3.7. An ion implanter has a beam current of 30 mA. The wafer holder can accommodate thirty 100-mm-diameter wafers. For a 5-min implant at a 130 keV implant energy, compute the dose received by the wafers.

REFERENCES

1. D. Halliday and R. Resnick, *Physics*, NY: Wiley, New York, 1978.
2. J. Pope, R. Woodburn, J. Watkins, R. Lachenbruch, and G. Viloria, "Manufacturing Integration of Real-Time Laser Interferometry to Isotropically Etch Silicon Oxide Films for Contacts and Vias," *Proc. SPIE Conf. Microelectronic Processing*, Vol. 2091, 1993, pp. 185–196.
3. S. Maung, S. Banerjee, D. Draheim, S. Henck, and S. Butler, "Integration of In-Situ Spectral Ellipsometry with MMST Machine Control," *IEEE Trans. Semiconduct. Manuf.* **7**(2), (May 1994).
4. T. Vincent, P. Khargonekar, and F. Terry, "An Extended Kalman Filtering-Based Method of Processing Reflectometry Data for Fast *In-Situ* Etch Rate Measurements," *IEEE Trans. Semiconduct. Manuf.* **10**(1), (Feb. 1997).
5. K. Wong, D. Boning, H. Sawin, S. Butler, and E. Sachs, "Endpoint Prediction for Polysilicon Plasma Etch via Optical Emission Interferometry," *J. Vac. Sci. Technol. A.* **15**(3), (May/June 1997).
6. H. Tompkins and W. McGahan, *Spectroscopic Ellipsometry and Reflectometry*, Wiley, New York, 1999.
7. F. Yang, W. McGahan, C. Mohler, and L. Booms, "Using Optical Metrology to Monitor Low-K Dielectric Thin Films," *Micro* **31–38** (May 2000).
8. J. McGilp, D. Weaire, and C. Patterson (eds), *Epioptics*, Springer-Verlag, New York, 1995.
9. W. Press, B. Flannery, S. Teukolsky, and W. Vetterling, *Numerical Recipes in C*, Cambridge Univ. Press, Cambridge, MA, 1988.
10. B. Stutzman, H. Huang, and F. Terry, "Two-Channel Spectroscopic Reflectometry for *In-Situ* Monitoring of Blanket and Patterned Structures During Reactive Ion Etching," *J. Vac. Sci. Technol. B.* **18**(6), (Nov./Dec. 2000).
11. R. Jaeger, *Introduction to Microelectronic Fabrication*, Addison-Wesley, Reading, MA, 1993.
12. S. Wolf and R. Tauber, *Silicon Processing for the VLSI Era*, Lattice Press, Sunset Beach, CA, 2000.
13. A. Landzberg, *Microelectronics Manufacturing Diagnostics Handbook*, Van Nostrand Reinhold, New York, 1993.
14. C. Raymond, "Angle-Resolved Scatterometry for Semiconductor Manufacturing," *Microlithogry. World* (winter 2000).
15. J. Pineda de Gyvez and D. Pradhan, *Integrated Circuit Manufacturability*, IEEE Press, Piscataway, NJ, 1999.

16. S. Campbell, *The Science and Engineering of Microelectronic Fabrication*, Oxford Univ. Press, New York, 2001.
17. D. Manos and D. Flamm, *Plasma Etching: An Introduction*, Academic Press, San Diego, CA, 1989.
18. D. Stokes and G. May, “Real-Time Control of Reactive Ion Etching Using Neural Networks,” *IEEE Trans. Semiconduct. Manuf.* **13**(4), 469–480 (Nov. 2000).
19. P. Raynaud, T. Amilis, and Y. Segui, “Infrared Absorption Analysis of Organosilicon/Oxygen Plasmas in a Microwave Multipolar Plasma Excited by Distributed Electron Cyclotron Resonance,” *Appl. Surf. Sci.* **138–139**, 285–291 (1999).
20. C. Almgren, “The Role of RF Measurements in Plasma Etching,” *Semicond. Intl.* 99–104 (Aug. 1997).

4

STATISTICAL FUNDAMENTALS

OBJECTIVES

- Explain, in general terms, the issues surrounding process variability.
- Introduce the statistical fundamentals necessary for analyzing semiconductor manufacturing processes.
- Describe and differentiate between discrete and continuous probability distributions.
- Discuss the concepts of sampling, estimation, statistical significance, confidence intervals, and hypothesis testing.

INTRODUCTION

In Chapter 3, various monitoring tools used to generate data necessary for process control were presented. For high-volume semiconductor manufacturing, such testing and inspection methods are essential for producing high-quality ICs. The term “quality” here refers to the fitness of a product for its designated use. In this sense, quality requires conformance of all products to a set of specifications and the reduction of any variability in the manufacturing process. A key metric for process quality is product *yield*, which is discussed in Chapter 5. Maintaining quality involves the use of *statistical process control* (SPC), which is the subject of Chapter 6. Since product variability is often described in statistical

Fundamentals of Semiconductor Manufacturing and Process Control,
By Gary S. May and Costas J. Spanos

terms, statistical methods necessarily play a central role in quality control and yield improvement efforts. Therefore, this chapter provides a concise review of some basic statistical fundamentals, along with appropriate examples from the semiconductor manufacturing domain.

In terms of semiconductor manufacturing processes, the most relevant aspect of quality is *quality of conformance*, or how well manufactured products conform to the specifications and tolerances required by their design and intended use. Every semiconductor device or circuit possesses a number of elements that collectively describe its fitness for use. These elements are referred to as *quality characteristics*.

Perhaps the major barrier to perfecting quality in a manufacturing environment is *variability*. Variability is inherent in every product—no two products are ever identical. For example, the dimensions of two thin films used for interconnect will vary according to the precise conditions and equipment used to deposit and pattern the films. Small variations might have negligible impact on the final product, but large variations can lead to final products that are unacceptable. *Quality improvement* may be defined as the reduction of such variability in processes and products.

Statistics allow engineers to make decisions about a process or population based on the analysis of a sample from that population. For example, two well-known statistics are the *sample average* and *sample variance*. Suppose that $x_1, x_2, \ldots, x_n$ are observations in a sample of size n. The statistic used to estimate the mean value (μ) of this population based on the sample is the sample average (x), which is given by

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.1}$$

The variance (σ^2), or spread, in a dataset is a statistic that can be estimated by the sample variance (s^2):

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{4.2}$$

The square root of the sample variance is known as the *sample standard deviation*. Generally, the larger the variance, the greater the spread in the sample data.

Statistical methods provide the principal means by which products are sampled, tested, and evaluated in a manufacturing environment. In the remainder of the chapter, various statistical methodologies are introduced as tools for use in quality control and improvement.

4.1. PROBABILITY DISTRIBUTIONS

A *probability distribution* is a mathematical model that relates the value of a random variable to its probability of occurrence. There are two types of probability distributions: discrete and continuous. *Discrete* distributions are used to describe random variables that can take on only certain specific values, such as the number of defects on a semiconductor wafer. On the other hand, when the random variable can have any value on a continuous scale (such as linewidth in a sample population of interconnect), the probability distribution is *continuous*. Examples of discrete and continuous probability distributions are shown in Figure 4.1.

4.1.1. Discrete Distributions

The discrete distribution is characterized by a series of vertical lines whose height represents the probability (Figure 4.1a). The probability that a random variable x is equal to a specific value x_i is given by

$$P\{x = x_i\} = p(x_i) \tag{4.3}$$

Two examples of discrete probability distributions that arise frequently in manufacturing applications are the binomial distribution and the Poisson distribution.

4.1.1.1. Hypergeometric

Let N represent the size of a finite population of items. Suppose that D of these items (where $D \leq N$) fall into a specific class of interest, such as the number of defective items in the population. If a random sample of n items is selected from the population without replacement, then the number of items in the sample that belong in the class of interest (x) is a random variable that follows the *hypergeometric* distribution. The probability of selecting x items belonging to the class is given by

$$P(x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}} \tag{4.4}$$

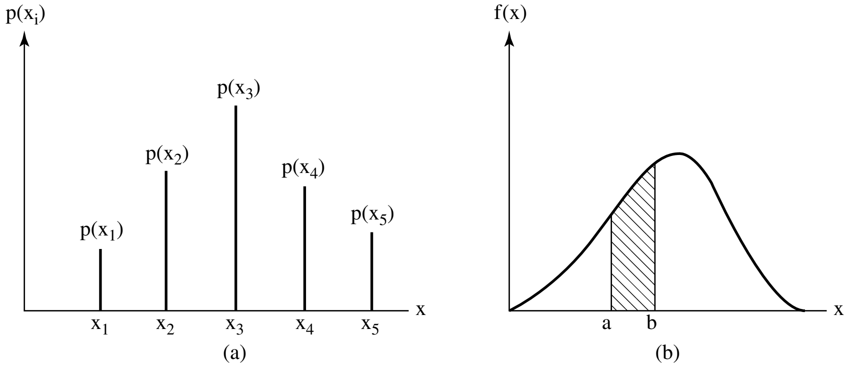


Figure 4.1. (a) Discrete and (b) continuous probability distributions [1].

where

$$\binom{a}{b} = \frac{a!}{b!(a-b)!}.$$

The mean (μ) and variance (σ^2) of the binomial distribution are

$$\mu = \frac{nD}{N} \tag{4.5}$$

$$\sigma^2 = \frac{nD}{N}\left(1 - \frac{D}{N}\right)\left(\frac{N-n}{N-1}\right) \tag{4.6}$$

The hypergeometric distribution is an appropriate model for encountering defective samples when selecting a random sample of n items without replacement from a population N of these items, of which D are nonconforming or defective. In semiconductor manufacturing, this is analogous to selecting a sample of n dies from a lot of wafers containing N total dies, D of which are known to be defective.

Example 4.1. Suppose that a lot of wafers contains 100 dies, 5 of which are known to be defective. If 10 of these dies are selected at random for inspection, what is the probability of finding less than two defective dies in the sample?

Solution: Here, $N = 100$, $n = 10$, and $D = 5$. To find the probability of less than two defective dies, we apply Eq. (4.4) as follows:

$$P(x < 2) = P(x \le 1) = P(0) + P(1)$$
$$= \frac{\binom{5}{0}\binom{95}{10}}{\binom{100}{10}} + \frac{\binom{5}{1}\binom{95}{9}}{\binom{100}{10}} = 0.923$$

Therefore, the probability of finding less than two defective dies is 92.3%.

4.1.1.2. Binomial

Suppose that a process consists of n independent trials. Each trial has two possible outcomes: “success” or “failure.” Trials with these characteristics are called *Bernoulli* trials. Let p be the probability of success for any given trial (thus, $0 < p < 1$). If p is constant, then the probability of achieving x successes in n trials is

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0, 1, \ldots, n \tag{4.7}$$

The mean (μ) and variance (σ^2) of the binomial distribution are

$$\mu = np \tag{4.8}$$

$$\sigma^2 = np(1-p) \tag{4.9}$$

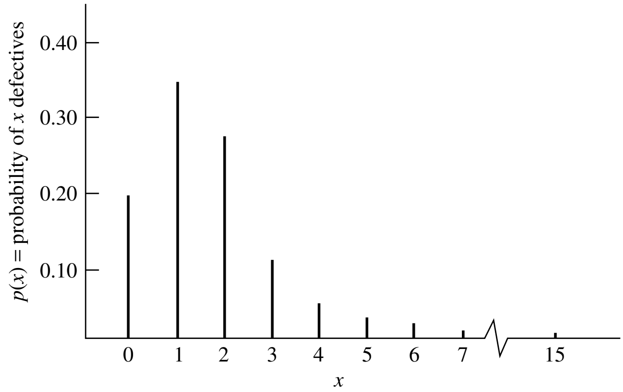


Figure 4.2. Binomial distribution with $p = 0.10$ and $n = 15$ [1].

The binomial model is used for sampling from an infinite population, and p represents the fraction of defective or nonconforming parts in that population. In this situation, x is the number of nonconforming parts identified in a random sample of n items. A typically shaped binomial distribution corresponding to $p = 0.10$ and $n = 15$ is shown in Figure 4.2.

Example 4.2. Suppose that a wire bonding process has an average of 1% defective bonds. If an inspector selects a random sample of 100 bonds, what is the probability of more than two of the bonds being defective?

Solution: In this case, $n = 100$ and $p = 0.01$. To find the probability of greater than two defective bonds, we apply Eq. (4.7) as follows:

$$P(x) = \binom{100}{x}(0.01)^x(0.99)^{100-x} \qquad x = 0, 1, \ldots, 100$$

Note that

$$\begin{aligned} P(x > 2) &= 1 - P(x \leq 2) = P(0) + P(1) + P(2) \\ &= \sum_{x=0}^{2}\binom{100}{x}(0.01)^x(0.99)^{100-x} \\ &= (0.99)^{100} + 100(0.01)^1(0.99)^{99} + 4950(0.01)^2(0.99)^{98} \cong 0.92 \end{aligned}$$

Therefore, the probability of finding more than two defective bonds is $1 - 0.92 = 0.08$ (8%).

An important random variable used in statistical process control is the *sample fraction nonconforming* $(\hat{p})$, which is

$$\hat{p} = \frac{x}{n} \tag{4.10}$$

This variable is the ratio of defective items to sample size. The probability distribution for $\hat{p}$ is derived from the binomial, since

$$P(\hat{p} \leq a) = P\left(\frac{x}{n} \leq a\right) = P(x \leq na) = \sum_{x=0}^{na} \binom{n}{x} p^x (1-p)^{n-x} \qquad (4.11)$$

where $[na]$ = greatest integer less than or equal to na. It can be shown that the mean and variance of $\hat{p}$ are $\mu(\hat{p}) = p$ and $\sigma^2(\hat{p}) = [p(1-p)]/n$, respectively.

4.1.1.3. Poisson

Another important discrete distribution is the Poisson distribution, which is characterized by the expression

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad (4.12)$$

where x is an integer and λ is a constant > 0. The mean and variance of the Poisson distribution are

$$\mu = \lambda \qquad (4.13)$$

$$\sigma^2 = \lambda \qquad (4.14)$$

respectively. The Poisson distribution is used to model the number of defects that occur in a single product. To illustrate, consider the following example.

Example 4.3. Suppose that the number of wire bonding defects that occur has a Poisson distribution with $\lambda = 4$. What is the probability that a randomly selected package will have two or fewer defects?

Solution: Applying (4.12) gives $P\{x \leq 2\} = \sum_{x=0}^{2}(e^{-4}4^x)/x! = 0.238$.

The Poisson distribution corresponding to $\lambda = 4$ is shown in Figure 4.3. The Poisson distribution is known for its skewed shape (i.e., the long "tail" to

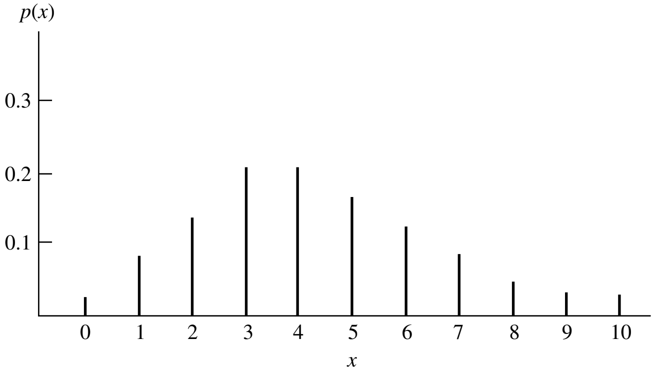


Figure 4.3. Poisson distribution with λ = 4 [1].

the right). As λ becomes larger, the shape of the distribution becomes more symmetric.

The Poisson distribution can be derived as a limiting form of the binomial distribution. In a binomial distribution with parameters n and p, as n approaches infinity and p approaches zero in such a way that $\lambda = np$ is a constant, then the Poisson distribution results.

4.1.1.4. Pascal

Like the binomial distribution, the Pascal distribution is based on a series of Bernoulli trials. For a sequence of independent trials, each with a probability of success (or failure) given by p, let x denote the trial in which the rth success (or failure) occurs. Under these circumstances, x is a Pascal random variable with the following probability distribution

$$P(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \tag{4.15}$$

where $r \geq 1$ is an integer and $x \geq r$. The mean and variance of the Pascal distribution are

$$\mu = \frac{r}{p} \tag{4.16}$$

$$\sigma^2 = \frac{r(1-p)}{p^2} \tag{4.17}$$

respectively.

There are two special cases of the Pascal distribution that are of interest in semiconductor manufacturing applications. The first is when $r > 0$ and not necessarily an integer. The resulting distribution in this case is called the *negative binomial* distribution, which is particularly useful in modeling IC yield (see Chapter 5). The second special case occurs when $r = 1$, which results in the *geometric* distribution. This is the distribution of the number of Bernoulli trials until the first success.

4.1.2. Continuous Distributions

A continuous distribution provides the probability that x lies in a specific interval (Figure 4.1b). This can be computed by integrating the continuous distribution between the endpoints of the interval. The probability that x is between a and b is given by

$$P\{a \leq x \leq b\} = \int_a^b f(x)\,dx \tag{4.18}$$

Two examples of continuous distributions that are important in statistical process control are the normal distribution and the exponential distribution. Each is described in more detail below.

4.1.2.1. Normal

The normal distribution is undoubtedly the most important and best known probability distribution in applied statistics. The *probability density function* for a normally distributed random variable x is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{4.19}$$

The notation $x \sim N(\mu, \sigma^2)$ is often used to imply that x is normally distributed with mean μ and variance σ^2. The normal distribution has a symmetric bell shape, as shown in Figure 4.4.

A useful graphic to interpret the value of the standard deviation of the normal distribution is shown in Figure 4.5. This figure shows that 68.26% of the area under a normal curve lies in the interval $\mu \pm 1\sigma$, 95.46% of the area lies in the interval $\mu \pm 2\sigma$, and 99.73% of the area lies in the interval $\mu \pm 3\sigma$.

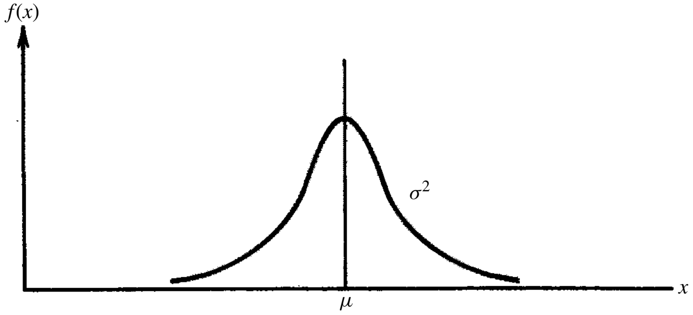


Figure 4.4. The normal distribution [1].

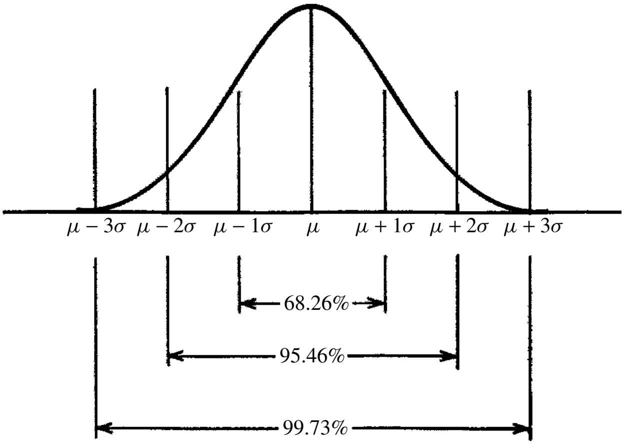


Figure 4.5. Areas under the normal distribution [1].

The *cumulative normal distribution* is defined as the probability that x is less than or equal to some value a, or

$$P(x \leq a) = F(a) = \int_{-\infty}^{a} f(x)\, dx \tag{4.20}$$

Unfortunately, this integral cannot be evaluated in closed form. Instead, the following change of variables is used:

$$z = \frac{x - \mu}{\sigma} \tag{4.21}$$

This allows the integral in Eq. (4.20) to be evaluated independently of μ and σ^2. In other words

$$P(x \leq a) = P\left\{z \leq \frac{a - \mu}{\sigma}\right\} = \Phi\left(\frac{a - \mu}{\sigma}\right) \tag{4.22}$$

where Φ is the cumulative distribution function of the *standard normal distribution* (i.e., the normal distribution with $\mu = 0$ and $\sigma = 1$). A table of values for the cumulative standard normal distribution function can be found in Appendix B.

Example 4.4. The linewidth of the interconnect for a given process has a mean value of $\mu = 40$ μm and a standard deviation of $\sigma = 2$ μm. What is the probability that a particular line will have a width of at least 35 μm?

Solution: We want to compute $P\{x \geq 35\}$. Note that $P\{x \geq 35\} = 1 - P\{x \leq 35\}$. To evaluate this probability, we standardize x and use the table in Appendix B.

$$P\{x \leq 35\} = P\left\{z \leq \frac{35 - 40}{z}\right\} = P\{z \leq -2.5\} = \Phi(-2.5) = 0.0062$$

The required probability is therefore

$$P\{x \geq 35\} = 1 - P\{x \leq 35\} = 1 - 0.0062 = 0.9938$$

One useful property of the normal distribution pertains to linear combinations of normally distributed random variables. If $x_1, x_2, \ldots, x_n$ are normally and independently distributed with means $\mu_1, \mu_2, \ldots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2 \ldots, \sigma_n^2$, respectively, then the distribution of

$$y = a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

is normal with mean

$$\mu_y = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n \tag{4.23}$$

and variance

$$\sigma_y^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2 \tag{4.24}$$

where $a_1, a_2, \ldots, a_n$ are constants.

4.1.2.2. Exponential

The exponential distribution is widely used in reliability engineering as a model for the time to failure of a component or system. The probability density function for a random variable x that has this distribution is

$$f(x) = \lambda e^{-\lambda x} \tag{4.25}$$

where $\lambda > 0$ is a constant. A graph of the density function appears in Figure 4.6. The mean and variance of the exponential distribution are

$$\mu = \frac{1}{\lambda} \tag{4.26}$$

$$\sigma^2 = \frac{1}{\lambda^2} \tag{4.27}$$

respectively. The cumulative exponential distribution function is

$$F(a) = P(x \le a) = \int_0^a \lambda e^{-\lambda t}\, dt = 1 - e^{-\lambda a} \quad a \ge 0 \tag{4.28}$$

The parameter λ is used to model the *failure rate* of a system, and the mean of the distribution ($1/\lambda$) is called the *mean time to failure.*

Example 4.5. An electronic component has a useful lifetime that is described by an exponential distribution with a failure rate of 10^{-4} per hour (i.e., $\lambda = 10^{-4}$). What is the probability that this component will fail before its expected life?

Solution: We want to compute $P\{x \le 1/\lambda\}$. We evaluate this probability as follows:

$$P\left\{x \le \frac{1}{\lambda}\right\} = \int_0^{1/\lambda} \lambda e^{-\lambda t} dt = 1 - e^{-1} = 0.6321$$

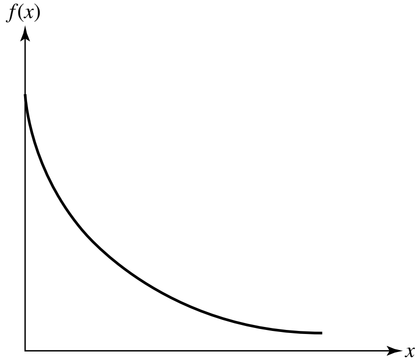


Figure 4.6. The exponential distribution.

There is an important relationship between the exponential and Poisson distributions. If the Poisson distribution is assumed to model the number of occurrences of a failure in the interval $(0, t]$, then applying Eq. (4.12) gives

$$P(x) = \frac{e^{-\lambda t}(\lambda t)^x}{x!} \tag{4.29}$$

If $x = 0$, there are no failures in the interval $(0, t]$, and, $P(0) = e^{-\lambda t}$. This may also be regarded as the probability that the first failure occurs *after* time t, or

$$P\{y > t\} = P(0) = e^{-\lambda t} \tag{4.30}$$

where y is a random variable representing the time interval until the first failure. Since

$$F(t) = P\{y \leq t\} = 1 - e^{-\lambda t} \tag{4.31}$$

and

$$f(y) = dF(y)/dy \tag{4.32}$$

we can conclude that

$$f(y) = \lambda e^{-\lambda y} \tag{4.33}$$

is the distribution of the interval to the first failure. Note that Eq. (4.33) is just the exponential distribution with parameter λ. Therefore, if the number of failures has a Poisson distribution with parameter λ, then the *interval between* failures is exponential with parameter λ.

4.1.3. Useful Approximations

For certain process control applications, approximating one probability distribution with another can significantly simplify the analysis. This approach is particularly useful in situations when the original distribution is complex or not well tabulated. Two such approximations are presented in the following.

4.1.3.1. Poisson Approximation to the Binomial

The Poisson distribution can be derived as a limiting form of the binomial distribution when p approaches zero and n approaches infinity with $\lambda = np$ constant. This implies that for small values of p and large values of n, the Poisson distribution with $\lambda = np$ can be used to approximate the binomial distribution. This approximation is usually reasonable for $p < 0.1$, but the larger the n and the smaller the p, the better the approximation.

4.1.3.2. Normal Approximation to the Binomial

The binomial distribution was previously defined as a sum of n Bernoulli trials, each with an associated probability of success p. If n is large, then the central limit theorem may be used to justify a normal approximation to the binomial distribution with mean np and variance $np(1 - p)$. In other words

$$P(x = a) = \binom{n}{a} p^a(1 - p)^{n-a} = \frac{1}{\sqrt{2\pi np(1 - p)}} e^{-(1/2)[(a-np)^2/np(1-p)]} \tag{4.34}$$

Since the binomial distribution is discrete and the normal distribution is continuous, the following continuity correction is commonly applied to this approximation

$$P(x = a) \cong \Phi\left(\frac{a + \frac{1}{2} - np}{\sqrt{np(1 - p)}}\right) - \Phi\left(\frac{a - \frac{1}{2} - np}{\sqrt{np(1 - p)}}\right) \tag{4.35}$$

where Φ is the standard normal cumulative distribution function. Probability intervals are evaluated similarly. In other words

$$P(a \le x \le b) \cong \Phi\left(\frac{b + \frac{1}{2} - np}{\sqrt{np(1 - p)}}\right) - \Phi\left(\frac{a - \frac{1}{2} - np}{\sqrt{np(1 - p)}}\right) \tag{4.36}$$

This approximation is satisfactory for $p \approx \frac{1}{2}$ and $n > 10$. For larger values of p, larger values of n are required. In general, the approximation is inadequate for $p < 1/(n + 1)$ or $p > n/(n + 1)$, or for values of the random variable outside the interval $np \pm 3\sqrt{np(1 - p)}$.

Since the binomial distribution can be approximated by the normal, and since the binomial and Poisson distributions are closely related, the Poisson distribution can also be approximated by the normal. The normal approximation to the Poisson distribution with $\mu = \lambda$ and $\sigma^2 = \lambda$ is satisfactory for $\lambda \ge 15$.

4.2. SAMPLING FROM A NORMAL DISTRIBUTION

Statistics allow inferences to be made or conclusions to be drawn about a population based on a *sample* chosen from that population. *Random sampling* refers to any method of sample selection that lacks systematic direction or bias. A random sample of size n consists of observations $x_1, x_2, \ldots, x_n$ selected so that the observations x_i are *independently and identically distributed* (IID). In other words, random sampling allows every sample an equal likelihood of being selected. If it can be further assumed that the samples come from a normal distribution, then it is said that the samples are IID*N*.

Statistical inference procedures use quantities such as the sample mean ($\overline{x}$) and sample variance (s^2) to draw conclusions about the central tendency and dispersion, respectively, of a population based on a sample. If the probability distribution from which a sample was taken is known, then the distribution of statistics such as $\overline{x}$ and s^2 can be determined from the sample data. For example, suppose that a random variable x is normally distributed with mean μ and variance σ^2. If $x_1, x_2, \ldots, x_n$ is a random sample of size n from this population, then the distribution of $(\overline{x})$ is $N[\mu, (\sigma^2/n)]$, which follows directly from Eqs. (4.23) and (4.24). In general, the probability distribution of a statistic is called the *sampling distribution.*

4.2.1. Chi-Square Distribution

An important sampling distribution that originates from the normal distribution is the *chi-square* (χ^2) distribution. If $x_1, x_2, \ldots, x_n$ are normally distributed random variables with mean zero and variance one, then the random variable

$$\chi_n^2 = \chi_1^2 + \chi_2^2 + \cdots + \chi_n^2$$

is distributed as chi-square with *n degrees of freedom*. The probability density function of χ^2 is

$$f\left(\chi^2\right) = \frac{1}{2^{n/2}\Gamma(n/2)}\left(\chi^2\right)^{(n/2)-1} e^{-\chi^2/2} \tag{4.37}$$

where Γ is the gamma function. If a random sample of size n is collected from a $N(\mu, \sigma^2)$ distribution, and this sample yields a sample variance of s^2, it can be shown that

$$\frac{(n-1)s^2}{\sigma^2} \approx \chi_{n-1}^2 \tag{4.38}$$

that is, the sampling distribution of $(n-1)s^2/\sigma^2$ is χ_{n-1}^2. The chi-square distribution is used to make inferences about the variance of a normal distribution. A few chi-square distributions are shown in Figure 4.7. A table of values for the cumulative chi-square distribution function is given in Appendix C.

4.2.2. *t* Distribution

The t distribution is another useful sampling distribution based on the normal distribution. If x and χ_k^2 are standard normal and chi-square random variables,

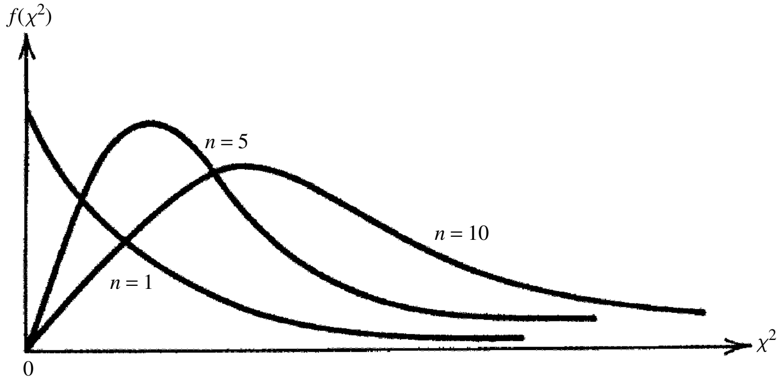


Figure 4.7. Several χ^2 distributions.

then the random variable

$$t_k \equiv \frac{x}{\sqrt{\chi_k^2/k}}$$

is distributed as t with k degrees of freedom. The probability density function of t is

$$f(t) = \frac{\Gamma[(k+1)/2]}{\sqrt{k\pi}(k/2)}\left(\frac{t^2}{k}+1\right)^{-(k+1)/2} \tag{4.39}$$

For a random sample of size n collected from a $N(\mu, \sigma^2)$ distribution with a sample mean $\overline{x}$ and a sample variance of s^2, it can be shown that

$$\frac{\overline{x}-\mu}{s/\sqrt{n}} \sim t_{n-1} \tag{4.40}$$

The t distribution is used to make inferences about the mean of a normal distribution. A few t distributions are shown in Figure 4.8. Note that as $k \to \infty$, the t distribution becomes the standard normal distribution. A table of values for the cumulative t distribution function is given in Appendix D.

4.2.3. *F* Distribution

The last sampling distribution to be considered that is based on the chi-square distribution is the F distribution. If χ_u^2 and χ_v^2 are chi-square random variables with u and v degrees of freedom, then the ratio

$$F_{u,v} \equiv \frac{\chi_u^2/u}{\chi_v^2/v}$$

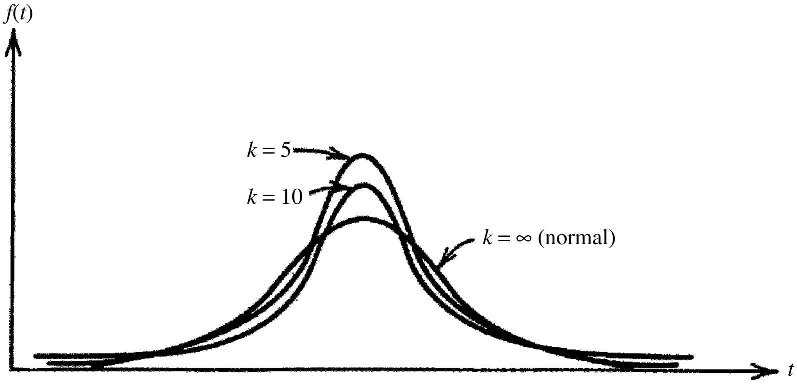


Figure 4.8. Several t distributions.

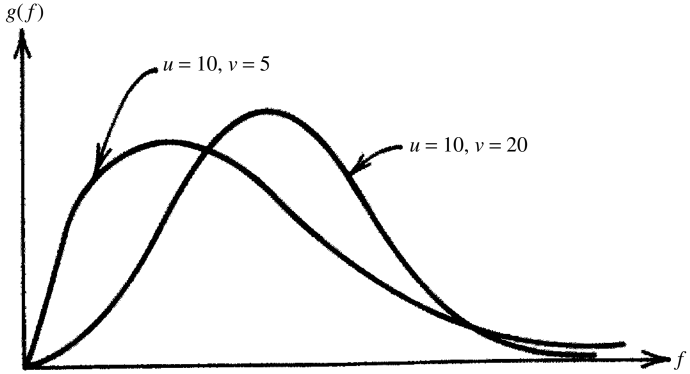


Figure 4.9. Several *F* distributions.

is distributed as F with u and v degrees of freedom. The probability density function of F is

$$g(F) = \frac{\Gamma\left(\dfrac{u+v}{2}\right)\left(\dfrac{u}{v}\right)^{u/2}}{\Gamma\left(\dfrac{u}{2}\right)\Gamma\left(\dfrac{v}{2}\right)} \frac{F^{u/2-1}}{\left[\left(\dfrac{u}{2}\right)F + 1\right]^{(u+v)/2}} \tag{4.41}$$

Consider two independent normal processes, $x_1 \sim N(\mu_1, \sigma_1^2)$ and $x_2 \sim N(\mu_2, \sigma_2^2)$. If random samples of sizes n_1 and n_2 yield sample variances s_1^2 and s_2^2, respectively, then it can be shown that

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1} \tag{4.42}$$

The F distribution can thus be used to make inferences in comparing the variances of two normal distributions. A few F distributions are shown in Figure 4.9. A table of values for the cumulative F distribution function is given in Appendix E.

4.3. ESTIMATION

Since the true values of the parameters of a distribution such as the mean (μ) or variance (σ^2) are generally unknown, procedures are required to estimate them from sample data. An estimator for such an unknown parameter may be defined as a statistic that approximates that parameter based on the sample data. A *point estimator* provides a single numerical value to estimate the unknown parameter. Examples of point estimators for the normal distribution are the sample mean ($\overline{x}$) and sample variance (s^2).

An *interval estimator*, on the other hand, provides a random interval in which the true value of the parameter being estimated falls with some probability. These intervals are called *confidence intervals*. A summary of some of the more useful confidence intervals for the normal distribution follows.

4.3.1. Confidence Interval for the Mean with Known Variance

Suppose that a sample of n independent observations $x_1, x_2, \ldots, x_n$ on a random variable x is taken. If $(\overline{x})$ is computed from the sample, then a $100(1 - \alpha)\%$ confidence interval on the mean μ of this population is defined as

$$\overline{x} - z_{(\alpha/2)}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x} + z_{(\alpha/2)}\frac{\sigma}{\sqrt{n}} \tag{4.43}$$

where $z_{\alpha/2}$ is the value of the $N(0, 1)$ distribution such that $P\{z \geq z_{\alpha/2}\} = \alpha/2$.

4.3.2. Confidence Interval for the Mean with Unknown Variance

Suppose that a sample of n independent observations $x_1, x_2, \ldots, x_n$ on a normally distributed random variable x is taken. If $\overline{x}$ and s^2 are computed from the sample, then a $100(1 - \alpha)\%$ confidence interval on the mean μ of this population is defined as

$$\overline{x} - t_{(\alpha/2),n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + t_{(\alpha/2),n-1}\frac{s}{\sqrt{n}} \tag{4.44}$$

where $t_{(\alpha/2),n-1}$ is the value of the t distribution with $n - 1$ degrees of freedom such that $P\{t_{n-1} \geq t_{(\alpha/2),n-1}\} = \alpha/2$.

Example 4.6. Suppose that the linewidth of $n = 16$ with supposedly identical interconnect traces is measured. The sample mean and sample standard deviation for these measurements are $\overline{x} = 49.86$ μm and $s = 1.66$ μm, respectively. What is the 95% confidence interval on this estimate of the mean?

Solution: Since $t_{0.025,15} = 2.132$, the 95% confidence interval on m can be found from Eq. (4.44) as follows:

$$49.86 - (2.132)1.66/\sqrt{16} \leq \mu \leq 49.86 + (2.132)1.66/\sqrt{16}$$

$$49.98 \leq \mu \leq 50.74$$

Thus, the estimate of the mean linewidth is 49.86 ± 0.88 μm with 95% confidence.

4.3.3. Confidence Interval for Variance

Suppose that a sample of n IIDN observations $x_1, x_2, \ldots, x_n$ on a random variable x is taken. If s^2 is computed from the sample, then a $100(1 - \alpha)\%$ confidence interval on the variance σ^2 of this population is defined as

$$\frac{(n-1)s^2}{\chi^2_{(\alpha/2),n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-(\alpha/2),n-1}} \tag{4.45}$$

where $\chi^2_{(\alpha/2),n-1}$ is the value of the χ^2 distribution with $n - 1$ degrees of freedom such that $P\{\chi^2_{n-1} \geq \chi^2_{\alpha/2,n-1}\} = \alpha/2$.

Example 4.7. For the dataset in Example 4.6, what is the 95% confidence interval on the estimate of the variance?

Solution: Since $\chi^2_{0.025,15} = 27.49$, $\chi^2_{0.975,15} = 6.27$, and $s^2 = 2.76$, the 95% confidence interval on σ^2 can be found from Eq. (4.45) as follows:

$$\frac{(15)(2.76)}{27.49} \leq \sigma^2 \leq \frac{(15)(2.76)}{6.27}$$

$$1.51 \leq \sigma^2 \leq 6.60$$

4.3.4. Confidence Interval for the Difference between Two Means, Known Variance

Consider two normal random variables from two different populations: x_1 with mean μ_1 and variance σ_1^2, and x_2 with mean μ_2 and variance σ_2^2. Suppose that samples of n_1 observations $x_{11}, x_{12}, \ldots, x_{1n_1}$ on random variable x_1 and n_2 observations $x_{21}, x_{22}, \ldots, x_{2n_2}$ on random variable x_2 are taken. If $\overline{x}_1$ and $\overline{x}_2$ are computed from the two samples and the variances are known, then a $100(1 - \alpha)\%$ confidence interval on the difference between the means of these two populations is defined as follows:

$$\left\{\overline{x}_1 - \overline{x}_2 - z_{(\alpha/2)}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right\} \leq (\mu_1 - \mu_2) \leq \left\{\overline{x}_1 - \overline{x}_2 + z_{(\alpha/2)}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right\} \tag{4.46}$$

4.3.5. Confidence Interval for the Difference between Two Means, Unknown Variances

Consider two normal random variables from two different populations: x_1 with mean μ_1 and variance σ_1^2, and x_2 with mean μ_2 and variance σ_2^2. Suppose that samples of n_1 observations $x_{11}, x_{12}, \ldots, x_{1n_1}$ on random variable x_1 and n_2 observations $x_{21}, x_{22}, \ldots, x_{2n_2}$ on random variable x_2 are taken. Assume that the means and variances are unknown, but the variances are equal; that is, $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

If $\overline{x}_1, \overline{x}_2, s_1^2$, and s_2^2 are computed from the two samples, then a *pooled estimate of the common variance* of the two populations is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \tag{4.47}$$

Under these conditions, a $100(1-\alpha)\%$ confidence interval on $\mu_1 - \mu_2$ is defined as

$$\left\{\overline{x}_1 - \overline{x}_2 - t_{(\alpha/2),\nu} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right\} \le (\mu_1 - \mu_2)$$

$$\le \left\{\overline{x}_1 - \overline{x}_2 + t_{(\alpha/2),\nu} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right\} \tag{4.48}$$

where $\nu = n_1 + n_2 - 2$.

Example 4.8. The average contact pad size for two different ICs is to be compared. $n_1 = n_2 = 10$ pads are selected at random, and their IIDN side dimensions are measured. For the first IC, $\overline{x}_1 = 90.70$ μm and $s_1^2 = 1.34$ μm^2; for the second IC, $\overline{x}_2 = 90.80$ μm and $s_2^2 = 1.07$ μm^2. What is the 99% confidence interval for the difference in pad size for the two ICs?

Solution: Assuming that the variances for pad size on each IC are the same, the pooled estimate of the common variance is found from Eq. (4.47) as follows:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(9)1.34 + (9)1.07}{10 + 10 - 2} = 1.21$$

The 99% confidence interval on $\mu_1 - \mu_2$ can be found from Eq. (4.48) as follows:

$$\left\{\overline{x}_1 - \overline{x}_2 - t_{0.005,18} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right\}$$

$$\le (\mu_1 - \mu_2) \le \left\{\overline{x}_1 - \overline{x}_2 + t_{0.005,18} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right\}$$

$$\left\{90.70 - 90.80 - (2.878)(1.1)\sqrt{\frac{1}{10} + \frac{1}{10}}\right\} \le (\mu_1 - \mu_2)$$

$$\le \left\{90.70 - 90.80 + (2.878)(1.1)\sqrt{\frac{1}{10} + \frac{1}{10}}\right\}$$

$$-1.51 \le \mu_1 - \mu_2 \le 1.31$$

4.3.6. Confidence Interval for the Ratio of Two Variances

Consider two normal random variables from two different populations: x_1 with mean μ_1 and variance σ_1^2, and x_2 with mean μ_2 and variance σ_2^2. Suppose that samples of n_1 observations $x_{11}, x_{22}, \ldots, x_{1n_1}$ on random variable x_1 and n_2 observations $x_{21}, x_{22}, \ldots, x_{2n_2}$ on random variable x_2 are taken. If $\overline{x}_1$, $\overline{x}_2$, s_1^2, s_2^2 are computed from the two samples, then a $100(1-\alpha)\%$ confidence interval on σ_1^2/σ_2^2 is defined as

$$\frac{s_1^2}{s_2^2}F_{1-(\alpha/2),\nu_2,\nu_1} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2}{s_2^2}F_{(\alpha/2),\nu_2,\nu_1} \tag{4.49}$$

where $\nu_1 = n_1 - 1$, $\nu_2 = n_2 - 1$, and $F_{\alpha/2,u,v}$ is the value of the F distribution with u and v degrees of freedom such that $P\{F_{u,v} \geq F_{\alpha/2,u,v}\} = \alpha/2$.

Example 4.9. Consider the dataset in Example 4.8. What is the 95% confidence interval for the ratio of the variances of contact pad size for the two ICs?

Solution: From Appendix E, $F_{0.025,9,9} = 4.03$ and $F_{0.975,9,9} = 0.248$. Using Eq. (4.49), the required confidence interval is

$$\frac{1.34}{1.07}(0.248) \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{1.34}{107}(4.03)$$

$$0.31 \leq \frac{\sigma_1^2}{\sigma_2^2} \leq 5.05$$

4.4. HYPOTHESIS TESTING

A *statistical hypothesis* is a statement about the values about the parameters of a probability distribution. A *hypothesis test* is an evaluation of the validity of the hypothesis according to some criterion. Hypotheses are expressed in the following manner:

$$\begin{aligned} &\mathrm{H}_0\colon\ \mu = \mu_0 \\ &\mathrm{H}_1\colon\ \mu \neq \mu_0 \end{aligned} \tag{4.50}$$

where μ is the unknown mean of the distribution and μ_0 is a hypothesized value of μ. The statement H_0: $\mu = \mu_0$ is called the *null hypothesis*, and H_1: $\mu \neq \mu_0$ is called the *alternative hypothesis*. Hypothesis testing procedures form the basis for many of the statistical process control techniques described in Chapter 6. To perform a hypothesis test, select a random sample from a population, compute an appropriate test statistic, and then either accept or reject the null hypothesis H_0.

Two types of error may result when performing such a test. If the null hypothesis is rejected when it is actually true, then a *type I error* has occurred. On the other hand, if the null hypothesis is accepted when it is actually false, this is called a *type II error*. The probabilities for each of these errors are denoted as follows:

$$\alpha = P(\text{type I error}) = P(\text{reject } \mathrm{H}_0|\mathrm{H}_0 \text{ is true})$$

$$\beta = P(\text{type II error}) = P(\text{accept } \mathrm{H}_0|\mathrm{H}_0 \text{ is false})$$

For statistical process control applications, α is considered the probability of a *false alarm* and β is the probability of a *missed alarm*. The statistical power of a test is defined as follows:

$$\text{Power} = 1 - \beta = P(\text{reject } \mathrm{H}_0 | \mathrm{H}_0 \text{ is false})$$

The power, therefore, represents the probability of correctly rejecting H_0. The basic procedure required for hypothesis testing involves specifying a desired value of α, and then designing a test that produces a small value of β. A few common test scenarios are illustrated in the following sections.

4.4.1. Tests on Means with Known Variance

Let x be a normally distributed random variable with unknown mean μ and known variance σ^2. Suppose that the hypothesis that the mean is equal to some constant value μ_0 must be tested. This hypothesis is described by Eq. (4.50). The procedure to perform the test requires taking a random sample of n independent observations and computing the following test statistic:

$$z_0 = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \tag{4.51}$$

The null hypothesis H_0 is rejected if $|z_0| > z_{\alpha/2}$, where $z_{\alpha/2}$ is the value of the standard normal distribution such that $P\{z \geq z_{\alpha/2}\} = \alpha/2$. In some cases, it may be necessary to test the hypothesis that the mean is larger than μ_0. Under these circumstances, the *one-sided alternative hypothesis* is $\mathrm{H}_1\colon\ \mu > \mu_0$, and H_0 is rejected only if $z_0 > z_\alpha$. To test the hypothesis that the mean is smaller than μ_0, the one-sided alternative hypothesis is $\mathrm{H}_1\colon\ \mu < \mu_0$, and H_0 is rejected if $z_0 < -z_\alpha$.

Suppose now that there are two populations with unknown means (μ_1 and μ_2) that must be compared. Assume that the two populations have known variances σ_1^2 and σ_2^2. To compare the two means, test the following hypothesis:

$$\begin{aligned} &\mathrm{H}_0\colon\ \mu_1 = \mu_2 \\ &\mathrm{H}_1\colon\ \mu_1 \neq \mu_2 \end{aligned} \tag{4.52}$$

To perform this test, n_1 and n_2 sample observations from each population are collected and then the test statistic

$$z_0 = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \tag{4.53}$$

is computed. H_0 is rejected if $|z_0| > z_{\alpha/2}$. The one-sided tests are similar to those described above.

Example 4.10. Suppose that it must be determined whether the mean thickness of a film exceeds 175 Å. The standard deviation of this thickness is known to be

10 Å. A random sample of 25 locations on a wafer yields an average thickness of $\overline{x} = 182$ Å.

Solution: The following hypothesis test is of interest:

$$\mathrm{H_0}: \ \mu = 175$$
$$\mathrm{H_1}: \ \mu > 175$$

The value of the test statistic is

$$z_0 = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{182 - 175}{10/\sqrt{25}} = 3.50$$

If a type I error of $\alpha = 0.05$ is specified, then, from Appendix B, $z_\alpha = z_{0.05} = 1.645$. Therefore, $\mathrm{H_0}$ is rejected, and the mean thickness does exceed 175 Å.

4.4.2. Tests on Means with Unknown Variance

Let x be a normal random variable with unknown mean μ and unknown variance σ^2. Suppose that the hypothesis that the mean is equal to some constant value μ_0 must be tested. Since the variance is unknown, it must be estimated by the sample variance s^2. The procedure to perform the test then requires taking a random sample of n observations and computing the following test statistic:

$$t_0 = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \tag{4.54}$$

$\mathrm{H_0}$ is rejected if $|t_0| > t_{\alpha/2,n-1}$, where $t_{\alpha/2,n-1}$ is the value of the t distribution with $n-1$ degrees of freedom such that $P\{t \geq t_{\alpha/2,n-1}\} = \alpha/2$. In some cases, the hypothesis that the mean is larger than μ_0 must be tested. Under these circumstances, the one-sided alternative hypothesis is $\mathrm{H_1}$: $\mu > \mu_0$, and $\mathrm{H_0}$ is rejected only if $t_0 > t_{\alpha,n-1}$. To test the hypothesis that the mean is smaller than μ_0, the one-sided alternative hypothesis is $\mathrm{H_1}$: $\mu < \mu_0$, and $\mathrm{H_0}$ is rejected if $t_0 < -t_{\alpha,n-1}$.

Suppose now that there are two normal populations with unknown means (μ_1 and μ_2) that must be compared. Assume that the two populations have unknown variances σ_1^2 and σ_2^2. To compare the two means, the hypothesis given by Eq. (4.52) is tested. The test procedure depends on whether the two variances can reasonably be assumed to be equal. If they are equal, and n_1 and n_2 sample observations are collected from each population, then a "pooled" estimate of the common variance of the two populations is given by Eq. (4.47). The appropriate test statistic is then

$$t_0 = \frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \tag{4.55}$$

H_0 is rejected if $|t_0| > t_{\alpha/2,n_1+n_2-2}$. The one-sided tests are similar to those described above. If the variances are *not equal*, then the appropriate test statistic is

$$t_0 = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} \tag{4.56}$$

and the number of degrees of freedom for t_0 are

$$v = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{(s_1^2/n_1)^2}{n_1+1} + \dfrac{(s_2^2/n_2)}{n_2+1}} - 2 \tag{4.57}$$

Once again, H_0 is rejected if $|t_0| > t_{\alpha/2,\nu}$, and the one-sided tests are similar to those described above.

Example 4.11. Consider the data in Example 4.8. Suppose that the hypothesis that the mean pad size for the first IC is equal to the mean pad size for the second IC must be tested, or

$$H_0\colon\ \mu_1 = \mu_2$$
$$H_1\colon\ \mu_1 \neq \mu_2$$

Solution: Assuming, $\sigma_1^2 = \sigma_2^2$, which is reasonable if the ICs have undergone the same manufacturing process, then $s_p = 1.10$. The test statistic is then

$$t_0 = \frac{\overline{x}_1 - \overline{x}_2}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} = -0.20$$

If a type I error of $\alpha = 0.01$ is specified, then, from Appendix D, $t_{0.005,18} = 2.878$. Since $|t_0| < t_{(\alpha/2),n-1}$, H_0 must be accepted, and there is no strong evidence that the two means are different.

4.4.3. Tests on Variance

Suppose that we want to test the hypothesis that the variance of a normal distribution is equal to some constant value σ_0^2. The hypotheses are expressed as follows:

$$H_0\colon\ \sigma^2 = \sigma_0^2$$
$$H_1\colon\ \sigma^2 \neq \sigma_0^2 \tag{4.58}$$

The appropriate test statistic is

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} \tag{4.59}$$

where s^2 is the sample variance computed from a random sample of n observations. Hypothesis H_0 is rejected if $\chi_0^2 > \chi^2_{(\alpha/2),n-1}$ or if $\chi_0^2 < \chi^2_{1-(\alpha/2),n-1}$, where $\chi^2_{(\alpha/2),n-1}$ and $\chi^2_{1-(\alpha/2),n-1}$ are the upper $\alpha/2$ and lower $1-(\alpha/2)$ percentage points of the χ^2 distribution with $n-1$ degrees of freedom. For the one-sided alternative hypothesis H_1: $\sigma^2 > \sigma_0^2$, we reject H_0 if $\chi_0^2 > \chi^2_{\alpha,n-1}$. To test the hypothesis that the variance is smaller than σ_0^2, the one-sided alternative hypothesis is H_1: $\sigma^2 < \sigma_0^2$, and we reject H_0 if $\chi_0^2 < \chi^2_{1-\alpha,n-1}$.

Now consider two normal populations with variances σ_1^2 and σ_2^2. To compare these populations, n_1 and n_2 sample observations from each are collected, and the hypothesis

$$\begin{aligned} H_0&: \sigma_1^2 = \sigma_2^2 \\ H_1&: \sigma_1^2 \neq \sigma_2^2 \end{aligned} \tag{4.60}$$

is tested. The test statistic is

$$F_0 = \frac{s_1^2}{s_2^2} \tag{4.61}$$

Hypothesis H_0 is rejected if $F_0 > F_{(\alpha/2),n_1-1,n_2-1}$ or if $F_0 < F_{1-(\alpha/2),n_1-1,n_2-1}$, where $F_{(\alpha/2),n_1-1,n_2-1}$ and $F_{1-(\alpha/2),n_1-1,n_2-1}$ are the upper $\alpha/2$ and lower $1-(\alpha/2)$ percentage points of the F distribution with n_1-1 and n_2-1 degrees of freedom. For the one-sided alternative hypothesis H_1: $\sigma_1^2 > \sigma_2^2$, H_0 is rejected if $F_0 > F_{\alpha,n_1-1,n_2-1}$. For the one-sided alternative hypothesis H_1: $\sigma_1^2 < \sigma_2^2$, H_0 is rejected if $F_0 > F_{\alpha,n_2-1,n_1-1}$.

Example 4.12. Consider once again the data in Example 4.8. Suppose that the hypothesis that the variances of the pad sizes are equal is to be tested, or

$$\begin{aligned} H_0&: \sigma_1^2 = \sigma_2^2 \\ H_1&: \sigma_1^2 \neq \sigma_2^2 \end{aligned}$$

Solution: Given that $s_1^2 = 1.34$ and ${s_2}^2 = 1.07$, the test statistic is

$$F_0 = \frac{s_1^2}{s_2^2} = 1.25$$

If a type I error of $\alpha = 0.05$ is specified, then, from Appendix E, $F_{0.025,9,9} = 4.03$. Since $F_0 < F_{\alpha/2,n_1-1,n_2-1}$, H_0 is accepted, and there is no strong evidence that the variances are different.

SUMMARY

This chapter provided an introduction to the concept of process variability and a brief survey of the statistical tools used to analyze semiconductor manufacturing processes. An understanding of these statistical fundamentals is essential for describing, analyzing, modeling, and controlling these processes, all of which are the subject of subsequent chapters.

PROBLEMS

4.1. A random sample of 50 dies is collected from each lot in a given processes. Calculate the probability that we will find less than three defective dies in this sample if the yield of the process is 98%.

4.2. An IC manufacturing process is subject to defects that obey a Poisson distribution with a mean of four defects per wafer.

(a) Assuming that a single defect will destroy a wafer, calculate the functional yield of the process.

(b) Suppose that we can add extra redundant dies to account for the defects. If one redundant die is needed to replace exactly one defective die, how many dies are required to ensure a yield of at least 50%?

4.3. Suppose the concentration of particles produced in an etching operation on any given day is normally distributed with a mean of 15.08 particles/ft^3 and a standard deviation of 0.05 particles/ft^3. The specifications on the process call for a concentration of 15.00 +/− 0.1 particles/ft^3. What fraction of etching systems conform to specifications?

4.4. The time to failure of printed circuit boards is modeled by the following exponential distribution probability density function:

$$f(t) = 0.125e^{-0.125t} \quad \text{for} \quad t > 0$$

where t is the time in years. What percentage of the circuit boards will fail within one year?

4.5. A new process has been developed for spin coating photoresist. Ten wafers have been tested with the new process, and the results of thickness measurements (in μm) are shown below and are assumed to be IIDN. Find a 99% confidence interval on the mean photoresist thickness.

13.3946	13.4002
13.3987	13.3957
13.3902	13.4015
13.4001	13.3918
13.3965	13.3925

4.6. Suppose that we are interested in calibrating a chemical vapor deposition furnace. The furnace will be shut down for repairs if significant difference is found between the thermocouples that are measuring the deposition temperature at the two ends of the furnace tube. The following temperatures have been measured during several test runs:

Thermocouple 1 (°C)	Thermocouple 2 (°C)
606.5	604.0
605.0	604.5
605.5	605.5
605.5	605.7
606.2	605.5
606.5	605.2
603.7	606.0
607.7	606.5
607.7	607.7
604.2	604.2

(a) Using the appropriate hypothesis test, determine whether we can be 95% confident that these temperatures are the same at both ends of the tube.

(b) Find the 90% confidence interval for the ratio of the two variances $(\sigma^2_{T1}/\sigma^2_{T2})$.

REFERENCE

1. D. Montgomery, *Introduction to Statistical Quality Control*, Wiley, New York, 1993.

5

YIELD MODELING

OBJECTIVES

- Provide a general definition of yield.
- Differentiate between functional and parametric yield.
- Introduce various yield models and simulators.
- Address financial aspects of yield.

INTRODUCTION

The primary objective of any semiconductor manufacturing operation is to produce outputs that meet required performance specifications. However, the variability inherent in manufacturing processes can lead to deformations or nonconformities in semiconductor products. Such process disturbances often result in faults, or unintentional changes in the performance or conformance of the finished integrated circuits. The presence of such faults is quantified by the *yield*.

Yield is in many ways the most important financial factor in producing ICs. This is because yield is inversely proportional to the total manufacturing cost—the higher the yield, the lower the cost.

Fundamentals of Semiconductor Manufacturing and Process Control,
By Gary S. May and Costas J. Spanos

5.1. DEFINITIONS OF YIELD COMPONENTS

Yield can be defined in many different ways. The first, and perhaps most basic definition, is that of *manufacturing yield*. This figure simply measures the proportion of successfully fabricated products compared to the number that have started the process. This definition applies to integrated circuits, which are batch-fabricated on semiconductor wafers, as well as to printed circuit boards, which are processed as individual parts.

Wafers that for one reason or another get scrapped along the way contribute to *wafer yield* losses. These losses can occur as a result of equipment malfunctions, wafer transport problems, or other difficulties. Clearly, identifying and removing problematic wafers as early as possible is an important objective, as it preserves processing resources. While factories implement early tests for that purpose, frequently wafers have to be rejected near the end, when they fail the various electrical or "probe" tests that are performed to confirm the overall electrical properties along the way, or the final electrical test done on the various devices and simple circuit structures. Further refining these definitions, production engineers distinguish the following three manufacturing yield components:

Wafer yield—the percentage of wafers that make it to final probing

Probe testing yield—the percentage of wafers that make it through the probe testing steps

Final testing yield—the percentage of wafers that make it through the final electrical testing step

Once a wafer has been successfully completed to the point that the product die can be electrically tested, then the figure of interest is the *design yield*, or *die yield*. There are two basic *die yield* components:

Functional yield (also known as "hard" or "catastrophic" yield)—the proportion of fully functional ICs

Parametric yield (also known as "soft" yield)—the overall performance achieved by the functional ICs

The one that can be determined first is the functional yield, which is usually limited by processing defects (such as particles), or artifacts that in general destroy the functionality of a circuit. These artifacts might cause short circuits, open circuits, or other types of "binary" failures. *Functional yield* is typically measured with high but finite precision, by running a series of functionality tests before individual ICs are diced from the wafer. These functionality tests are designed to balance the test coverage and the testing cost, and the overall objective is to avoid packaging, or worse, shipping nonfunctional ICs. Functional yield depends not only on process and material cleanliness but also on IC design practices. The issue of understanding, modeling, and improving functional yield will be discussed in some depth in this chapter.

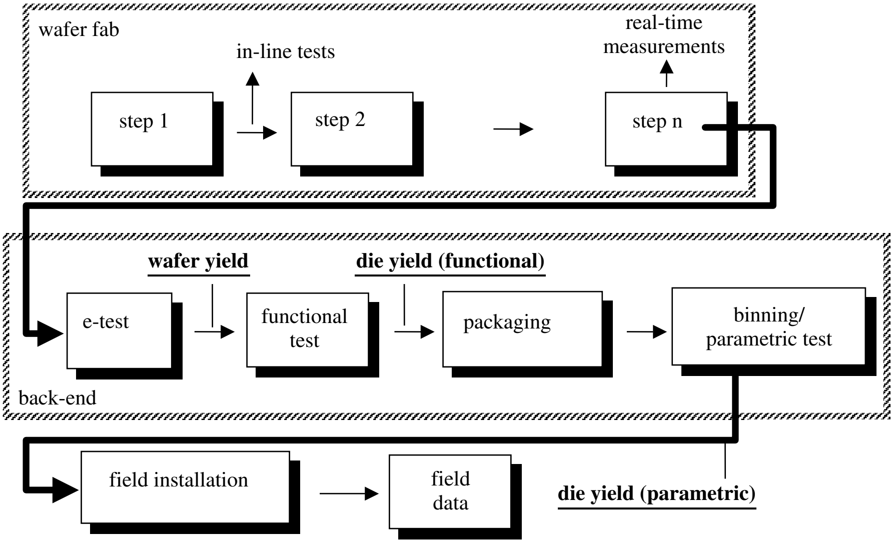


Figure 5.1. Manufacturing process flow from the perspective of yield monitoring and control.

It is usually after the individual ICs have been diced from the wafer that the latter yield component, the *parametric yield,* is determined. Performance might be quantified by various metrics, such as speed of execution or power consumption. The natural variability of the process, as well as the non-catastrophic impact of some types of defects, will lead to a statistical spread of the various device parameters, and this spread will in turn result in a spread of IC performances. During this last stage of testing, IC products are typically separated into various performance "bins" and parametric variation determines the percentage of ICs that end up into each bin. The issues that determine parametric spread relate to process control practices, as well as process and material variations. The impact of these variations on the parametric yield can be further controlled by the appropriate design practices.

All yield components are subject to intense scrutiny, by means of material and process studies and IC failure analysis. Once the *assignable causes* of yield loss have been eliminated, the emphasis shifts to understanding and quantifying the systematic causes, leading to a body of work focusing on yield modeling and simulation. The concept of *design for manufacturability* then comes into play, in an attempt to mitigate the impact of these causes by means of appropriate circuit and process design consideration. Figure 5.1 outlines the overall process flow, from the perspective of yield monitoring and control.

5.2. FUNCTIONAL YIELD MODELS

The development of models to estimate the functional yield of microelectronic circuits and packages is fundamental to manufacturing. A model that provides accurate estimates of manufacturing yield can help predict product cost, determine optimum equipment utilization, or be used as a metric against which actual measured manufacturing yields can be evaluated. Yield models are also used to support decisions involving new technologies and the identification of problematic products or processes.

As mentioned previously, functional yield is significantly impacted by the presence of *defects*. Defects can result from many random sources, including contamination from equipment, processes, or handling; mask imperfections; and airborne particles. Physically, these defects include shorts and opens (short and open circuits), misalignment, photoresist splatters and flakes, pinholes, scratches, and crystallographic flaws. This is illustrated by Figure 5.2.

Yield models are usually presented as a function of the average number of defects per unit area (D_0) and the *critical area* (A_c) of the electronic system. In other words,

$$Y = f(A_c, D_0) \tag{5.1}$$

where Y is the functional yield. The *critical area* is the area in which a defect occurring has a high probability of resulting in a fault. For example, if the particles in Figure 5.2 (which repeats Figure 2.8) are conductive, only particle 3 has fallen into an area in which it causes a short between the two metal lines that it bridges. The relationship between the yield, defect density, and critical area is complex. It depends on the circuit geometry, the density of photolithographic patterns, the number of photolithography steps used in the manufacturing process, and other factors. A few of the more prevalent models that attempt to quantify this relationship are described in the following sections.

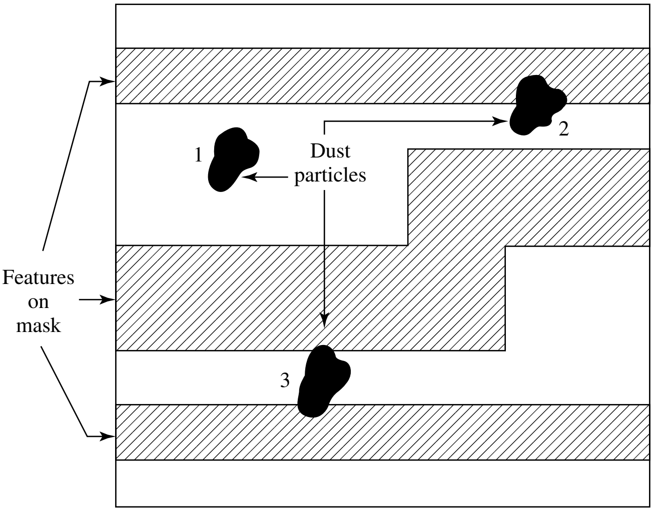


Figure 5.2. Various ways in which foreign particles can interfere with interconnect patterns.

5.2.1. Poisson Model

The Poisson yield model requires that defects be considered as perfect points that are spatially uncorrelated and uniformly distributed (with a defect density D_0) across a substrate. The Poisson model further requires that each defect result in a fault. J. Piñeda de Gyvez provides an excellent derivation of this model [2]. Let C be the number of circuits on a substrate (the number of ICs, modules, etc.), and let M be the number of possible defect types. Under these conditions, there are C^M unique ways in which the M defects can be distributed on the C circuits. For example, if there are three circuits (C1, C2, and C3) and three defect types (e.g., M1 = metal open, M2 = metal short, and M3 = metal 1–metal 2 short), then there are

$$C^M = 3^3 = 27 \tag{5.2}$$

possible ways in which these three defects can be distributed over three chips. These combinations are illustrated in Table 5.1.

If one circuit is removed (i.e., is found to contain no defects), the number of ways to distribute the M defects among the remaining circuits is

$$(C - 1)^M \tag{5.3}$$

Thus, the probability that a circuit will contain zero defects of any type is

$$\frac{(C-1)^M}{C^M} = \left(1 - \frac{1}{C}\right)^M \tag{5.4}$$

Substituting $M = CA_cD_0$, the yield is the number of circuits with zero defects, or

$$Y = \lim_{C\to\infty}\left(1 - \frac{1}{C}\right)^{CA_cD} = \exp(-A_cD_0) \tag{5.5}$$

Table 5.1. Table of unique fault combinations.

	C1	C2	C3		C1	C2	C3
1	M1 M2 M3			15	M3		M2 M1
2		M1 M2 M3		16		M1 M2	M3
3			M1 M2 M3	17		M1 M3	M2
4	M1 M2	M3		18		M2 M3	M1
5	M1 M3	M2		19		M1	M2 M3
6	M2 M3	M1		20		M2	M1 M3
7	M1 M2		M3	21		M3	M2 M1
8	M1 M3		M2	22	M1	M2	M3
9	M2 M3		M1	23	M1	M3	M2
10	M1	M2 M3		24	M2	M1	M3
11	M2	M1 M3		25	M2	M3	M1
12	M3	M2 M1		26	M3	M1	M2
13	M1		M2 M3	27	M3	M2	M1
14	M2		M1 M3				

For N circuits to have zero defects, this becomes

$$Y = \exp(-A_c D_0)^N = \exp(-N A_c D_0) \tag{5.6}$$

The same result can be obtained using Poisson statistics directly. Poisson statistics represent an approximation of the Maxwell–Boltzmann (or *binomial*) distribution when large sample sizes are used. Recall the Poisson probability distribution given by Eq. (4.12). If x is the number of faults per circuit and $\lambda = N A_c D_0$ is the fault density, the yield is defined at $x = 0$, or

$$Y = P(x = 0) = \exp(-N A_c D_0) \tag{5.7}$$

We thus achieve an equivalent expression to that given by [Eq. (5.6).]

The Poisson model is simple and relatively easy to derive. It provides a reasonably good estimate of yield when the critical area is small. However, if D_0 is calculated based on small-area circuits, using the same D_0 for large-area yield computations results in a yield estimate that is overly pessimistic compared to actual measured data.

5.2.2. Murphy's Yield Integral

B. T. Murphy first proposed that the value of the defect density (D) should not be constant [1]. Instead, he reasoned that D must be summed over all circuits and substrates using a normalized probability density function $f(D)$. The yield can then be calculated using the integral

$$Y = \int_0^\infty e^{-A_c D} f(D) dD \tag{5.8}$$

Various forms of $f(D)$ form the basis for the differences between many analytical yield models. The Poisson model described in the previous section assumes that $f(D)$ is a delta function, that is

$$f(D) = \delta(D - D_0) \tag{5.9}$$

where D_0 is the average defect density as before (see Figure 5.3a). Using this density function, the yield is determined from Eq. (5.8) as

$$Y_{\text{Poisson}} = \int_0^\infty e^{-A_c D} f(D) dD = \exp(-A_c D_0) \tag{5.10}$$

as shown before.

Murphy initially investigated a uniform density function as shown in Figure 5.3b. The evaluation of the yield integral for the uniform density function gives

$$Y_{\text{uniform}} = \frac{1 - e^{-2D_0 A_c}}{2D_0 A_c} \tag{5.11}$$

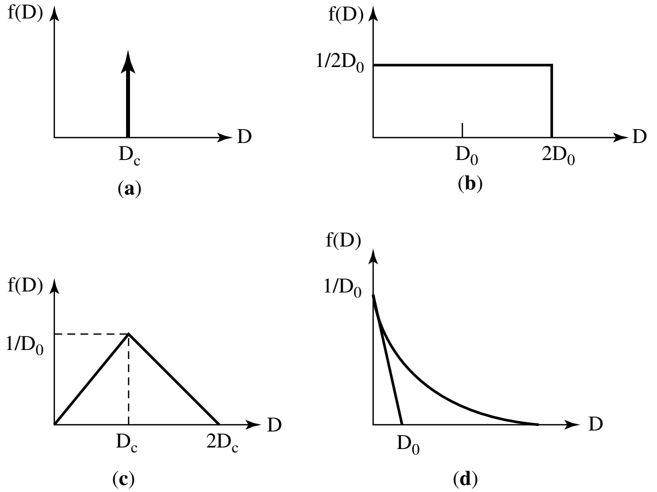


Figure 5.3. (a) Probability density function (pdf) for the Poisson model; (b) pdf for the uniform Murphy model; (c) pdf for the triangular Murphy model; (d) pdf for the exponential Seeds model [2].

Murphy later believed that a Gaussian distribution would be a better reflection of the true defect density distribution than the delta function. However, since evaluating the yield integral with a Gaussian function substituted for $f(D)$ would not have resulted in a closed-form solution, he approximated it using the triangular function in Figure 5.3c. This function results in the yield expression

$$Y_{\text{triangular}} = \left(\frac{1 - e^{-D_0 A_c}}{D_0 A_c} \right)^2 \tag{5.12}$$

The triangular Murphy yield model is widely used today in industry to determine the effect of manufacturing process defect density.

R. B. Seeds was the first to verify Murphy's predictions [3]. However, Seeds theorized that high yields were caused by a large population of low defect densities (which are not high enough to cause faults) and a small proportion of high defect densities (i.e., high enough to cause faults). He therefore proposed the exponential density function given by

$$f(D) = \frac{1}{D_0} \exp\left(\frac{-D}{D_0} \right) \tag{5.13}$$

and shown in Figure 5.3d. This function implies that the probability of observing a low defect density is significantly higher than that of observing a high defect density. Substituting this exponential function in the Murphy integral and integrating yields

$$Y_{\text{exponential}} = \frac{1}{1 + D_0 A_c} \tag{5.14}$$

It should be noted that the Seeds model may also be derived in an alternate manner using Bose–Einstein statistics. This was accomplished independently by Price [4]. The Bose–Einstein distribution is relevant for indistinguishable particles in which there is no constraint on the number of particles that can occupy a given state. Recall Table 5.1, and assume that the three defects (M1–M3) are now indistinguishable. Under these new conditions, there are only 10 combinations of defects that are uniquely identifiable, and there are

$$Z_1 = \frac{(C+M-1)!}{M!(C-1)!} \tag{5.15}$$

unique ways of identifying the M defects on C chips. If one chip has no defects, then the number of unique ways to distribute the M defects in the rest of the $(C-1)$ chips is

$$Z_2 = \frac{(C+M-2)!}{M!(C-2)!} \tag{5.16}$$

The yield in this case is Z_2/Z_1, or

$$Y = \left[\frac{(C-1)!}{(C-2)!}\right]\left[\frac{(C+M-2)!}{(C+M-1)!}\right] = \frac{C-1}{C+M-1} = \frac{\left(1-\dfrac{1}{C}\right)}{\left(1+\dfrac{M}{C}+\dfrac{1}{C}\right)} \tag{5.17}$$

If we now substitute $M = CA_cD_0$, taking the limit as C tends to infinity gives

$$Y = \lim_{C\to\infty} \frac{\left(1+\dfrac{1}{C}\right)}{\left(1+\dfrac{M}{C}+\dfrac{1}{C}\right)} = \frac{1}{1+A_cD_0} \tag{5.18}$$

which is the same as the model given in Eq. (5.14).

Although the Seeds model is simple, its yield predictions for large-area chips are too optimistic. This is because the assumption of indistinguishable defects is seldom valid for IC fabrication processes, where defects are often visually distinguishable from one another. Therefore, this model has not been widely used in industry.

5.2.3. Negative Binomial Model

Okabe et al. recognized the physical nature of defect distributions and proposed the gamma probability density function [5]. C. H. Stapper has likewise written several papers on the development and applications of yield models using the gamma density function [6]. The gamma probability density function is given by

$$f(D) = [\Gamma(\alpha)\beta^{\alpha}]^{-1} D^{\alpha-1} e^{-D/\beta} \tag{5.19}$$

where α and β are two parameters of the distribution and $\Gamma(\alpha)$ is the gamma function. The shape of $\Gamma(\alpha)$ is shown for several values of α in Figure 5.4. In

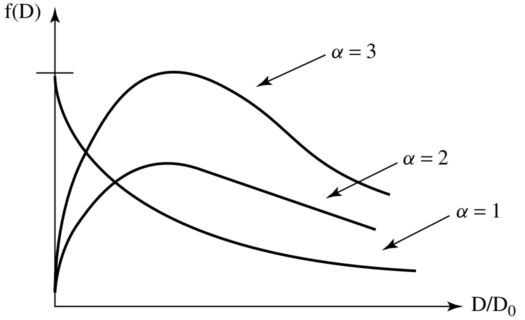


Figure 5.4. Probability density function for the gamma distribution.

this distribution, the average defect density is $D_0 = \alpha\beta$, and the variance of D_0 is $\alpha\beta^2$. The yield model derived by substituting Eq. (5.19) into Murphy's integral is

$$Y_{\text{gamma}} = \left(1 + \frac{A_c D_0}{\alpha}\right)^{-\alpha} \tag{5.20}$$

This model is commonly referred to as the *negative binomial* model. The parameter α is generally called the "cluster" parameter since it increases with decreasing variance in the distribution of defects.

If α is high, that means that the variability of defect density is low (little clustering). Under these conditions, the gamma density function approaches a delta function, and then the negative binomial model reduces to the Poisson model. Mathematically, this means

$$Y = \lim_{\alpha \to \infty} \left(1 + \frac{A_c D_0}{\alpha}\right)^{-\alpha} = \exp(-A_c D_0) \tag{5.21}$$

If α is low, on the other hand, the variability of defect density across the substrate is significant (much clustering), and the gamma model reduces to the Seeds exponential model, or

$$Y = \lim_{\alpha \to 0} \left(1 + \frac{A_c D_0}{\alpha}\right)^{-\alpha} = \frac{1}{1 + A_c D_0} \tag{5.22}$$

The parameter α must be determined empirically. Methods for doing so that involve particle counting using laser reflectometry exist [7], but several authors have found that values of $\alpha = 2$ provide a good approximation for a variety of logic and memory circuits [8]. Therefore, if the critical area and defect density are known (or can be accurately measured), the negative binomial model is an excellent general-purpose yield predictor that can be used for a variety of electronics manufacturing processes.

5.3. FUNCTIONAL YIELD MODEL COMPONENTS

The functional yield models that we have discussed thus far have been defined in terms of independent parameters such as D_0 and A_c. These parameters are statistically independent and can be measured directly. The following sections describe these and other critical parameters in greater detail.

5.3.1. Defect Density

Defect density is clearly a critical parameter in yield modeling and production yield planning. A "defect" is an unintended pattern on the wafer surface. It can consist of either extra material or missing material. In order to properly describe defect density, a few other terms must be defined:

Contamination—any foreign material on a wafer surface or embedded in a thin film. Sources of contamination include human skin, dirt, dust, or particles resulting from an oxidized gas, residual chemicals, or sputtering.

Defect—any alteration in the desired physical pattern intended to be printed. Typical defects include metal stringers, open and short circuits, notches, splotches, bridges, or hillocks.

Fault—an electrical circuit failure caused by a defect.

On the basis of these definitions, we observe that contamination is a random physical event that may or may not lead to a defect. Similarly, a defect may or may not result in a fault. The correlation between contamination, defects, and faults is weak. Mapping contamination to defects or defects to faults is difficult and time-consuming.

A physical interpretation of defect density should incorporate the size distribution of defects, as well as the probability that a defect will cause a failure. Typically, defects smaller than the minimum feature size will not cause failures. However, if a defect of a particular size causes a fault, then a larger defect at the same location will also cause a fault. An example of the effect of defects of different sizes at the same location is shown in Figure 5.5. The two adjacent metal lines in this figure will be shorted (short-circuited) by a defect of greater size than the spacing between them.

In general, the defect density is defined mathematically as the area under the defect size distribution curve for specific size limits. For a mature manufacturing process, defect density has been shown experimentally to follow an inverse power-law relationship with respect to size [9]. In other words

$$D(x) = \frac{N}{x^p} \tag{5.23}$$

where x is the defect diameter (assuming spherical defects), N is a technology-dependent parameter, and p must be determined empirically. This power law also

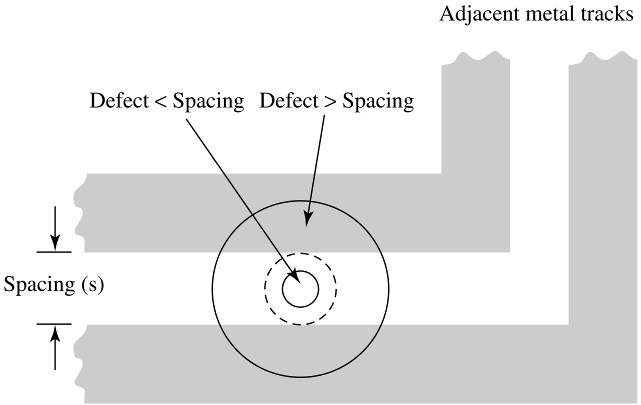


Figure 5.5. Illustration of the effect of defect size distribution on critical area [2].

assumes that defects are located randomly across the wafer surface. The average defect density may thus be determined by integrating this expression, or

$$D_0 = N \int_{x_0}^{\infty} D(x)dx = N \int_{x_0}^{\infty} \frac{1}{x^p} dx = \frac{N}{1-p}(1 - x_0^{1-p}) \tag{5.24}$$

where x_0 is the minimum defect diameter, which is usually the minimum feature size for a given technology. Neither the defect density nor the critical area can be determined without the defect size distribution.

A simpler method of extracting the defect density involves using a particular yield model to solve for D_0 mathematically. For example, using the negative binomial model for a single chip gives

$$D_0 = \frac{\alpha(\sqrt[\alpha]{1/Y} - 1)}{A_c} \tag{5.25}$$

This approach works best for similar products fabricated using the same mature technology with chip areas within a factor of 2–3, but should be applied cautiously otherwise. The most useful aspect of this approach is in using the calculated value of D_0 as a metric of manufacturing process performance.

5.3.2. Critical Area

The concept of critical area is used to account for the fact that not all parts of a chip layout are equally likely to fail because of the presence of defects. This allows greater accuracy when calculating the defect sensitivity of a chip layout. Consider Figure 5.6, in which the dark areas represent the first metal layer for a given circuit. The crosshatched area represents the sensitive regions at the minimum spacing for this technology. The critical area is a measure of such sensitive regions for the entire chip.

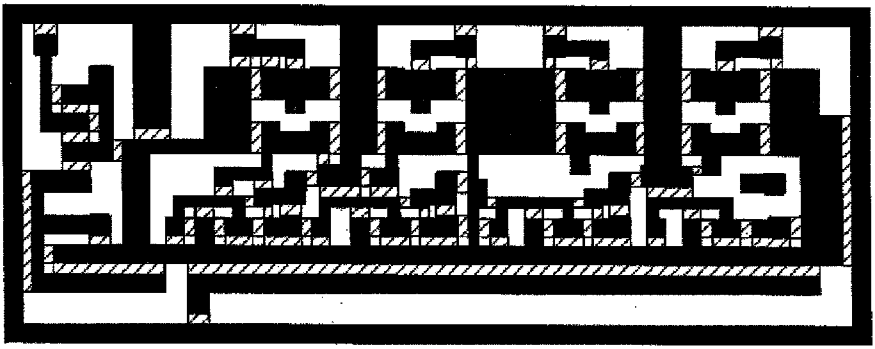


Figure 5.6. Subcircuit metal layer in which shaded region indicates critical area [2].

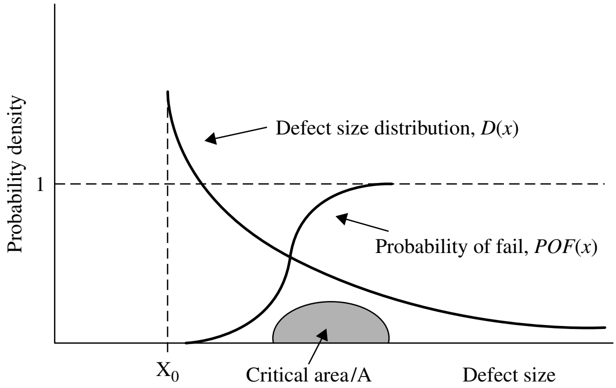


Figure 5.7. Graphical representation of the critical area integral [2].

Critical area is defined mathematically by the following relationship:

$$A_c = A \int_{x_0}^{\infty} \mathrm{PoF}(x) D(x) dx \tag{5.26}$$

where A is the chip area, x_0 is the minimum defect size, $D(x)$ is the defect size distribution, and PoF(x) is the probability of failure, which is a strong function of the defect size. A graphical interpretation of this relationship is shown in Figure 5.7. Several methods have been reported to determine critical area, and in each case, the calculation of PoF is crucial. More detail on these techniques is provided in Section 5.4.

5.3.3. Global Yield Loss

The yield models discussed thus far have focused solely on yield loss due to the presence of *local* defects. However, *global defects* can also be present. Global yield loss is usually spatially correlated and often manifests itself as a consequence of variability in electrical parameters (such as transistor threshold voltage) caused by process fluctuations (such as temperature or film thickness

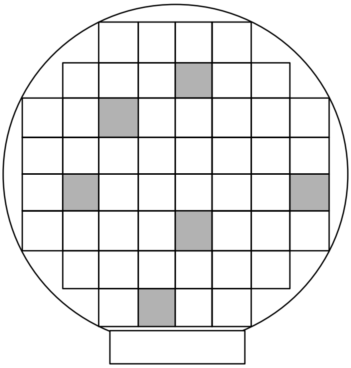


Figure 5.8. Example of a wafer map [2].

variations). Global yield loss, which is quantified *by parametric yield* models (see Section 5.4), is often identifiable as an anomalous spatial pattern on a wafer, such as an annual ring or cluster of failing chips. Yield models can take global defects into account by incorporating a factor Y_0. For example, the negative binomial model that includes global yield loss effects is

$$Y = Y_0 \left(1 + \frac{A_c D_0}{\alpha}\right)^{-\alpha} \tag{5.27}$$

The Y_0 factor is not related to the defect density or the critical area.

Although it is difficult to determine global yield loss analytically, spatial analysis techniques can be used to evaluate whether the measured yield loss is consistent with this model. Spatial analysis typically requires wafer maps generated from automated test equipment. In these maps, failing chips are categorized by the similarity of failures (e.g., function fail, speed fail). An example of a typical wafer map appears in Figure 5.8.

One way to estimate the value of Y_0 is to use a "windowing" technique in which individual chips are grouped together into windows of increasing size. The effective yield of each window size is then plotted against the effective chip size. The y intercept of this plot is the yield with an area of zero. Thus, Y_0 is equal to one minus this intercept. If the intercept is at 100% yield, there is no global yield loss.

5.4. PARAMETRIC YIELD

Even in a defect-free manufacturing environment, random processing variations can lead to varying levels of system performance. These variations result from global defects that cause the fluctuation of numerous physical and environmental parameters (linewidths, film thicknesses, ambient humidity, etc.), which in turn manifest themselves as variations in final system performance (such as speed or noise level). These performance variations lead to "soft" faults and are characterized by the parametric yield of the manufacturing process.

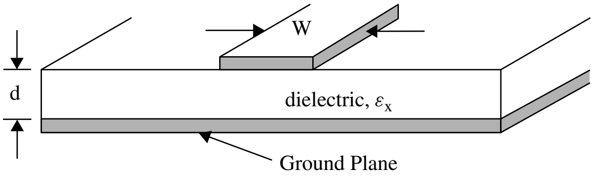


Figure 5.9. A microstrip transmission line of width *W* on top of an insulating dielectric with thickness *d*.

Parametric yield is a measure of the quality of functioning systems, whereas *functional yield* measures the proportion of functioning units produced by the manufacturing process.

A common method used to evaluate parametric yield is *Monte Carlo simulation*. In the Monte Carlo approach, a large number of pseudorandom sets of values for circuit or system parameters are generated according to an assumed probability distribution (usually the normal distribution) based on sample means and standard deviations extracted from measured data. For each set of parameters, a simulation is performed to obtain information about the predicted behavior of a circuit or system, and the overall performance distribution is then extracted from the set of simulation results.

To illustrate the Monte Carlo technique, consider as a performance metric the characteristic impedance (Z_0) of a microstrip transmission line of width W on top of an insulating dielectric with thickness d (refer to Figure 5.9). Under the condition that $W/d \ll 1$, it can be shown that

$$Z_0 = \frac{60}{\sqrt{\varepsilon_e}} \ln\left(\frac{8d}{W} + \frac{W}{4d}\right) \tag{5.28}$$

where ε_e is the effective dielectric constant of the insulator, which is given by

$$\varepsilon_e = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2\sqrt{1 + \dfrac{12d}{W}}} \tag{5.29}$$

where ε_r is the relative permittivity of the insulator [10]. From these equations, it is clear that Z_0 is a function of the physical dimensions, d and W, or $Z_0 = f(d, W)$. Both of these dimensions are subject to manufacturing process variations. They can thus be characterized as varying according to normal distributions with means μ_d and μ_W and standard deviations σ_d and σ_W, respectively (see Figure 5.10a).

Using the Monte Carlo approach, we can estimate the parametric yield of microstrips produced by a given manufacturing process within a certain range of characteristic impedances by computing the value of Z_0 for every possible combination of d and W. The result of these computations is a final performance distribution such as the one shown in Figure 5.10b. This probability density function can then be used to compute the proportion of microstrips having a given

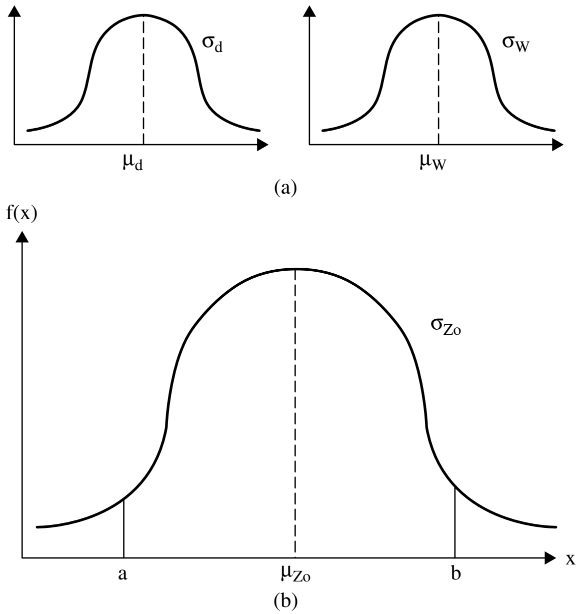


Figure 5.10. (a) Normal probability density functions for *W* and *d*; (b) overall pdf for characteristic impedance.

range of impedances. For example, if we wanted to compute the percentage of microstrips manufactured that would have the a value of Z_0 between two limits a and b, we would evaluate the integral

$$\text{Yield(microstrips with } a < Z_0 < b) = \int_a^b f(x)dx \tag{5.30}$$

Thus, once the overall distribution of a given output metric is known, it is possible to estimate the fraction of manufactured parts with any range of performance. Estimation of parametric yield is useful for system designers since it helps identify the limits of the manufacturing process to facilitate and encourage *design for manufacturability*.

5.5. YIELD SIMULATION

It is highly desirable for IC manufacturers to be able to predict yield loss prior to circuit fabrication. This enables corrective action to be taken before production starts and can prevent misprocessing. Yield simulation software tools are the primary means for facilitating yield prediction.

Local and global defects are the two basic sources of yield loss. The effects of global defects, which result in parametric yield loss, have been modeled in statistical process simulators such as the FABRication of Integrated Circuits

Simulator (FABRICS) [11].[1] Local defects, on the other hand, which can cause catastrophic failures that impact functional yield, have been modeled using Monte Carlo–based yield simulators such as the VLSI LAyout Simulator for Integrated Circuits (VLASIC) [12]. In this section, we briefly explore the capabilities of these two yield simulation tools.

5.5.1. Functional Yield Simulation

The effect of local defects on yield can be determined by generating a population of chip samples that has a distribution that closely approximates the distribution of circuit faults observed in fabrication. This circuit fault distribution may be obtained using a Monte Carlo simulation in which defects are repeatedly generated, placed on the chip layout, and then analyzed to identify what circuit faults have occurred. This procedure is implemented by the VLASIC simulator.

The VLASIC simulation algorithm is illustrated by the block diagram in Figure 5.11. A control loop generates as many chip samples as desired for a given simulation. Defect random-number generators are used to determine the number and location of defects on the chip layout with the appropriate statistical distributions. These statistics are derived from fabrication line measurements.

Once the defects have been placed on the layout, fault analysis is used to determine what, if any, circuit faults have occurred. The resulting faults are then filtered so that those faults that do not affect functional yield are ignored. The output is a chip sample containing the list of faults that have occurred during simulated fabrication. When simulation is complete, the list of unique chip faults

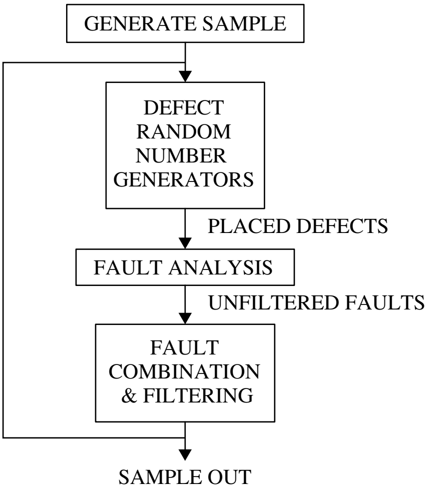


Figure 5.11. VLASIC main loop [12].

[1] The FABRICS parametric yield simulator was developed by Maly and Strojwas of Carnegie Mellon University in 1982.

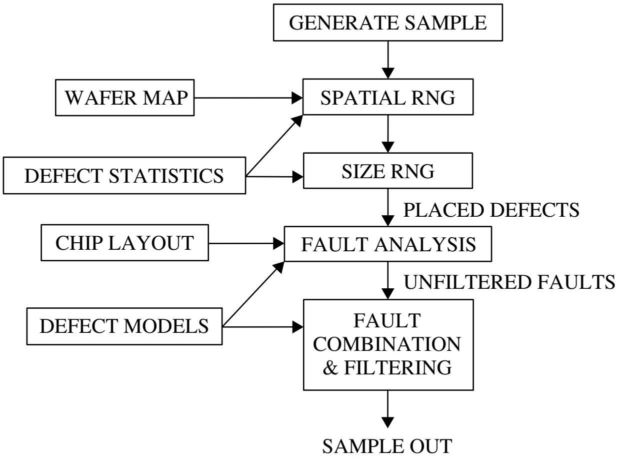


Figure 5.12. VLASIC system structure [12].

and their frequency of occurrence is passed to postprocessors designed to predict yield, optimize design rules, generate test vectors, or evaluate process sensitivity.

A detailed view of the VLASIC system structure is shown in Figure 5.12. Defects in VLASIC have both size and spatial distributions. A wafer map is used to place defects on a wafer. The output of the defect random-number generators is a list of defect types, locations within the chips, and defect diameters. The defect statistics are similar to those described in Section 5.3.

For each defect, the fault analysis phase calls a series of procedures to examine the layout geometry in the immediate vicinity of the defects to determine whether any circuit faults have occurred. A separate procedure is used for each fault type (shorts, opens, etc.). Since the defects and layout features are represented as polygons, the analysis procedures manipulate layout geometry using general-purpose polygon operations. The resulting output of fault analysis is a list of unfiltered circuit faults caused by the defects. The unfiltered faults then pass through a filtering–and combination phase in which faults that do not cause a change in DC circuit operation are ignored, and some faults are combined together to form a composite fault, respectively.

Fault analysis and filtering operations both depend on input from defect models. These models describe a fabrication process as a number of patterned layers in which defects are represented as modifications to the layout of each layer (i.e., extra or missing material). The models also specify the circuit faults that can be caused by each defect type, which layers are affected by the defect, and the manner in which layers are electrically connected. After filtering, the resulting output is a list of circuit faults that have occurred during simulated fabrication. Each fault is specified by its type, size, location, type of defect that caused it, location of the fault in the circuit graph, and number of times the fault occurred. The fault list is then ready for postprocessing.

To illustrate the use of VLASIC, consider the simulation of a simplified chip containing only a single three-transistor dynamic RAM cell (see Figure 5.13).

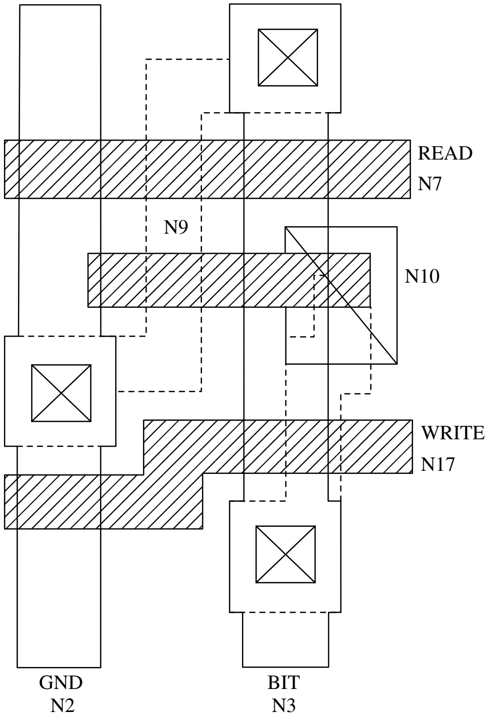


Figure 5.13. Three-transistor DRAM cell [12].

Suppose that the chip is placed at 100 locations on the wafer as shown in Figure 5.14. The process conditions used in the simulation are given in Table 5.2. Note that relatively high defect densities are used in order to obtain an average of 2.5 defects per sample. The α parameter for the negative binomial model is also high, indicating very little spatial clustering of defects.

The result of simulating 1000 chip samples is shown in Figure 5.15. This output represents a list of unique chip fabrication outcomes. Each unique outcome has a frequency count and a list of fault groups (i.e., sets of faults caused by a single defect). For each fault, the fault type, defect type that caused it, defect location, defect diameter, and fault description are provided. For the single instance where several defects of the same type have caused the same fault to occur (i.e., several oxide pinholes shorting the same nets together), defect size and defect location are meaningless, since only the values for the first defect causing the fault are recorded.

The simulated fabrication of the 1000 samples results in 25 unique chip faults, the distribution of which appears in Table 5.3. Despite an average of 2.5 defects per sample, only 5.9% of the simulated chips had a circuit fault. This result is typical of yield simulations. To explore reasons for this, note that only 3.3% of the DRAM cell area contains a gate oxide. Thus, a gate oxide pinhole defect has only a 1 in 30 chance of causing a gate-to-channel short (circuit). Consider also the case of extra metal defects, recalling that extra material defects must have a

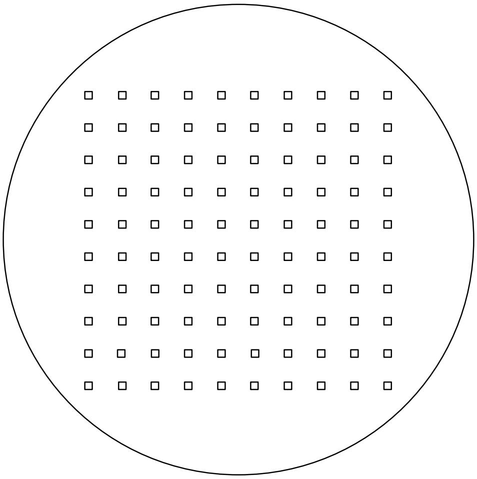


Figure 5.14. Wafer map.

Table 5.2. DRAM cell process conditions.

Defect density	
Extra/missing metal	20,000 cm^{-2}
Extra/missing polysilicon	20,000 cm^{-2}
Extra/missing active	20,000 cm^{-2}
First-level pinholes	20,000 cm^{-2}
Gate oxide pinholes	20,000 cm^{-2}
Design rules	
Metal width/space	6 μm
Metal contact width	4 μm
Polysilicon width/space	4 μm
Active width/space	4 μm
Polysilicon/active space	2 μm
Diameter of peak density	
Extra/missing metal	2 μm
Extra/missing polysilicon	2 μm
Extra/missing active	2 μm
Maximum defect diameter	18 μm
Between-lot alpha	100
Between-wafer alpha	100
Wafers per lot	1
Radial distribution	None
Minimum line spacing	0
Minimum linewidth	0

minimum diameter to cause a circuit fault. In this simulation, the combination of the peak defect size of 2 μm (using a defect size distribution similar to that shown in Figure 5.6) and minimum metal width/space of 6 μm leads to only a 1 in 18 chance for a defect of this type causing a short.

```
SAM> vlasic −t omcellproc.dat omcell.pack
Reading wirelist omcell.pack
Writing faults to stdout
Parsing wafer file omcellwaf.cif
Initializing random number generators
left: −27.00 right: 3.00 bottom: −51.00 top: −3.00
NumXBins: 1 NumYBins 2
Allocating bins
Putting wirelist polygons into bins
Generating intermediate layers
Place and analyze defects
Results of fault testing:
Sample Count: 1000
Trial Count:
type POSM1: 234
type NEGM1: 237
type POSP: 283
type NEGP: 271
type POSD: 298
type NEGD: 335
type PIN1: 274
type PIN2: 276
type PING: 318
Total Trials: 2526
Total number of distinct chiplists: 25
Distinct Circuit Faults and Counts:
941 NIL
9 SHORT PIN1 X −10 Y −37 Diam 0 N3 N17
6 SHORT PIN1 X −25 Y −15 Diam 0 N2 N7
5 NEWVIA PING X −15 Y −16 Diam 0 NCHAN N7
5 SHORT PIN1 X −9 Y −17 Diam 0 N3 N7
4 SHORT PIN1 X −5 Y −21 Diam 0 N3 N10
4 NEWVIA PING X −13 Y −25 Diam 0 NCHAN N10
3 SHORT PIN1 X −20 Y −25 Diam 0 N2 N10
3 NEWVIA PING X −3 Y −37 Diam 0 NCHAN N17
3 SHORT PIN1 X −20 Y −37 Diam 0 N2 N17
2 SHORT POSP X 1 Y −27 Diam 16.09 N10 N17
2 SHORT POSP X −4 Y −12 Diam 7.43 N3 N7
1 SHORT POSM1 X −16 Y −14 Diam 17.19 N2 N3
1 SHORT POSP X −10 Y −43 Diam 9.04 N3 N17
1 SHORT PIN1 X −20 Y −37 Diam 0 N2 N17
  SHORT PIN1 X −25 Y −15 Diam 0 N2 N7
1 SHORT PIN1 X −10 Y −37 Diam 0 N3 N17
  SHORT PIN1 X −5 Y −21 Diam 0 N3 N10
1 NEWVIA PING X −15 Y −16 Diam 0 NCHAN N7
  SHORT PIN1 X −5 Y −21 Diam 0 N3 N10
  SHORT PIN1 X −20 Y −25 Diam 0 N2 N10
1 OPEN NEGP X −21 Y −41 Diam 12.25 N17/1 LEFT NP N17/2 Tran DO G
1 NEWVIA PING X −3 Y −37 Diam 0 NCHAN N17
  OPEN NEGM1 X −7 Y −43 Diam 7.58 N3/1 Tran DO SD N3/2 BOTTOM NM1 TOP NM1
1 OPEN NEGP X −7 Y −23 Diam 4.89 N10/1 Tran DO SD N10/2 Tran D1 G
1 OPEND NEGD X −14 Y −21 Diam 16.04 Tran D2
    OPEND NEGD X −14 Y −21 Diam 16.04 Tran D1
1 NEWGD POSP X −10 Y −19 Diam 8.43 CVTMULTI Tran D1 D2 SD: N2/0 N3/0 N9/0 G: N7/0 N10/0
    SHORT POSP X −10 Y −19 Diam 8.43 N7 N10
1 OPEND NEGD X −5 Y −35 Diam 4.91 Tran DO
1 SHORTD NEGP X −17 Y −15 Diam 8.10 Tran D2
    OPEN NEGP X −17 Y −15 Diam 8.10 N7/1 Tran D2 G N7/2 LEFT NP
1 OPEN POSP X −24 Y −28 Diam 18.00 N2/1 Tran D1 SD LEFT ND N2/2 LEFT NM1 BOTTOM NM1
    SHORT POSP X −24 Y −28 Diam 18.00 N2 N10 N17
```

Figure 5.15. VLASIC DRAM example [12].

Another noteworthy aspect of this simulation that is typical of all yield simulations is that chips with a single simple fault are much more common than those with multiple fault groups. This is directly attributable to the fact that single defects are more common than multiple defects. Complex multiple fault groups are rare because large extra or missing material defects must be present to cause

Table 5.3. DRAM chip sample distribution.

94.1%	No faults
4.2%	One oxide pinhole short
0.6%	One extra material short
0.2%	Two oxide pinhole shorts
0.2%	One missing material open
0.1%	Three oxide pinhole shorts
0.1%	Two-open device
0.1%	One-open device
0.1%	One oxide pinhole short and one missing material open
0.1%	One new gate device and one extra material short
0.1%	One shorted device and one missing material open
0.1%	One extra material open and one extra material short

them. Of the four fault groups that occurred in this example, the smallest defect causing one was 8.1 μm in diameter. Only about 1 in 33 defects is this large.

5.5.2. Parametric Yield Simulation

The FABRICS parametric yield simulator embodies an approach to modeling the IC fabrication process that accounts for the statistical fluctuations that occur during manufacturing. This simulator is capable of generating values for the parameters of IC circuit elements (resistances, capacitances, transconductances, etc.), as well as estimates of inline measurements typically made during fabrication (junction depths, sheet resistances, oxide thicknesses, etc.). These quantities are described statistically as random variables characterized by a joint probability density function.

FABRICS accounts for the dependence of IC elements on both layout and process parameters. Each process step is modeled individually, with its outcome dependent on a set of control parameters, a set of process disturbances, and the outcome of the previous process step (see Figure 5.16). To formally describe

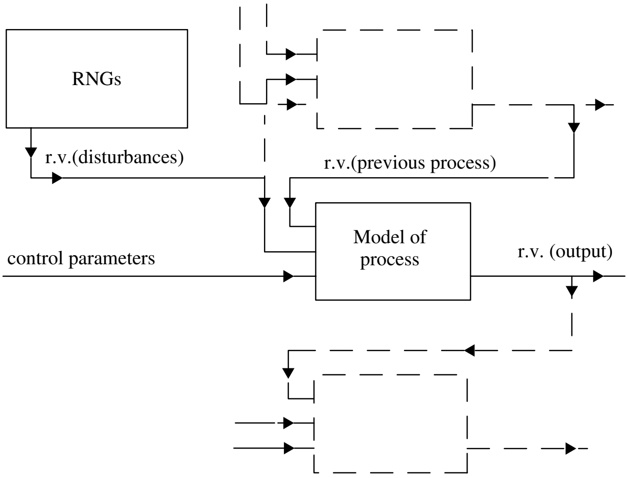


Figure 5.16. Model of a single process step (rv = "random varaible") [11].

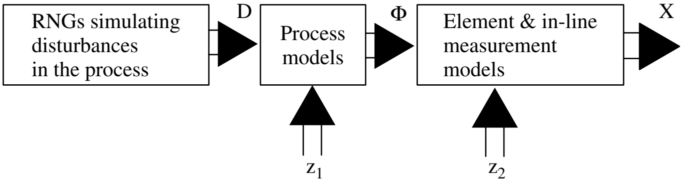

RNGs simulating disturbances in the process
D
Process models
Φ
Element & in-line measurement models
X
z1
z2

Figure 5.17. Basic FABRICS structure [11].

the FABRICS simulation procedure, let X be a vector of random variables that denotes the parameters of the IC elements or values of inline measurements. In addition, let the vector z_1 represent the process control parameters (temperatures, times, gas flows, etc.), and let the vector z_2 represent the layout dimensions. Finally, let D be a vector of random variables representing uncontrollable process disturbances. These disturbances are simulated in FABRICS as appropriately defined random-number generators (RNGs).

A basic flowchart describing the structure of the *FABRICS* simulator is shown in Figure 5.17. *FABRICS* uses analytical models for each manufacturing process step and circuit element of the form:

$$\mathbf{\Phi}_j = g_j(\hat{\mathbf{\Phi}}^j, \mathbf{z}_1^j, \mathbf{D}^j) \qquad j = 1, \ldots, m \tag{5.31}$$

where $\mathbf{\Phi}_j$ is a component of the m-dimensional vector $\mathbf{\Phi}$ of physical parameters which describes the outcome of a given step (i.e., oxide thicknesses, doping profile parameters, misalignment, etc.), $\hat{\mathbf{\Phi}}^j$ is a vector of physical parameters obtained from previous steps, and z_1^j and D^j are vectors containing those components of z_1 and D affecting the jth physical parameter. Models of the IC circuit elements are of the form

$$\mathbf{X}_i = h_i(\hat{\mathbf{\Phi}}^i, \mathbf{z}_2^i) \qquad i = 1, \ldots, n \tag{5.32}$$

where X_i is an electrical parameter associated with a given circuit element (such as the β of a bipolar transistor) and $\hat{\mathbf{\Phi}}^i$ and z_2^j are subsets of the vectors $\mathbf{\Phi}$ and z_2 that affect this element. Simulation of the random variable X consists of generating samples of D, evaluating the components of $\mathbf{\Phi}$ for subsequent steps, and calculating X using the appropriate model. The analytical functions g_j and h_i can be regarded as approximations to the solutions of the differential equations used in numerical process and circuit models, respectively. For example, a commonly known analytical model of the diffusion process is the *erfc* model (see Chapter 2).

The statistical parameters of the probability density function (pdf) of X resulting from the simulated samples should be in good agreement with measured parameters from the real process. Since z_1 and z_2 are known, achieving such good agreement requires determination of the pdf of the process disturbances, f_D. This is accomplished by collecting data from inline and test pattern measurements and using statistical optimization techniques to estimate the parameters of f_D that provide a good fit. Although this rather computationally intensive identification task is valid for only a particular manufacturing process, the results can be used to simulate the manufacture for a variety of ICs, irrespective of the IC layout. Therefore, once f_D is known, the simulator may be used instead of the actual fabrication process to optimize layout or fine-tune process control parameters.

The random variable of process disturbances affecting one chip, D_i, is simulated with a RNG that generates data with a mean equal to μD_i and a standard deviation of σD_i. Assuming that μD_i and σD_i change randomly from one chip to another, each disturbance must be simulated by a two-level RNG that accounts for local (within-chip) and global (chip-to-chip) variations. This approach is illustrated in Figure 5.18, which shows a two-level structure consisting of three RNGs. RNG1 simulates a disturbance within a chip by generating a normally distributed random variable D_i. RNG2 and RNG3 provide RNG1 with μD_i and σD_i for the chip, respectively. The inputs to RNG2 are the mean of means (μ_μ) and standard deviation of means (σ_μ) of the chips in the wafer. Similarly, inputs to RNG3 are the mean of standard deviations (μ_σ) and standard deviation of standard deviations (σ_σ) of the chips in the wafer.

A more detailed data flow diagram for FABRICS is shown in Figure 5.19. Data entered into the simulator include process parameters, IC layout dimensions, and control parameters used to activate the RNGs and models in the correct sequence.

To illustrate the operation of FABRICS, consider the production of the MC1530 operational amplifier (shown schematically in Figure 5.20) as a typical bipolar manufacturing process. Suppose that we want to ascertain the effect modifying the surface concentration of phosphorus during the predeposition of the emitter layer of the transistors in this circuit. Since the modification of the surface concentration will result in a change in the sheet resistance of the emitter layer (R_{SE}), we will use FABRICS to examine the relationship between the parametric yield and R_{SE}.

Using FABRICS Monte Carlo simulation in conjunction with a circuit simulator [such as SPICE (a simulation program with integrated circuit emphasis)], the yield of the amplifier for six different phosphorus surface concentrations is

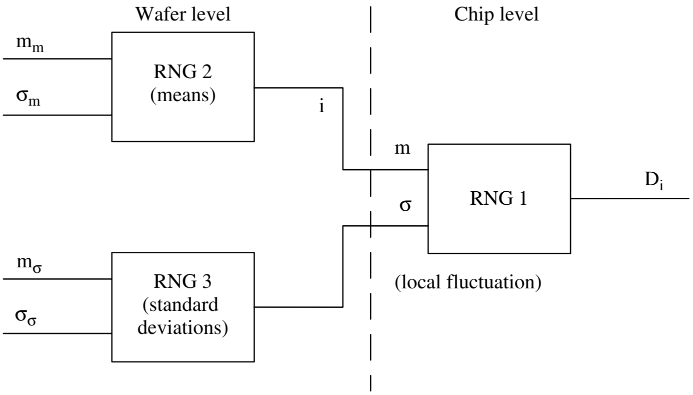


Figure 5.18. Illustration of two-level RNG architecture that simulates within-chip and chip-to-chip process variations [11].

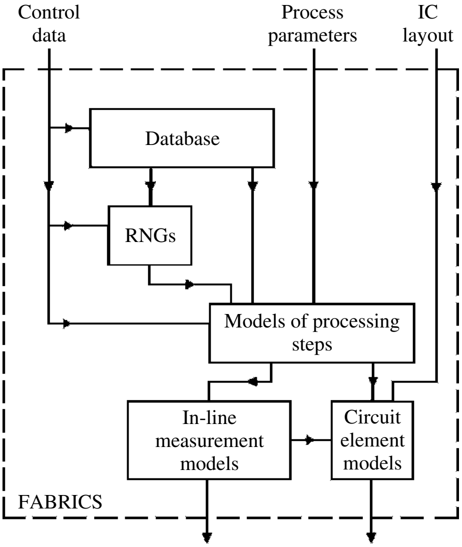


Figure 5.19. Detailed FABRICS data flow [11].

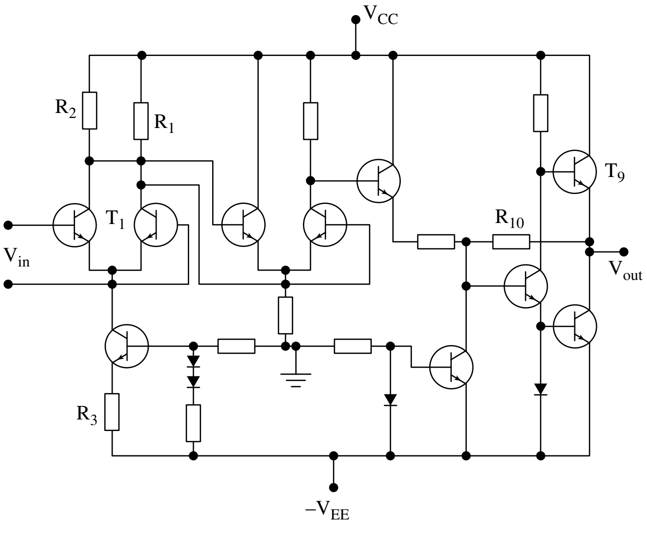


Figure 5.20. Schematic of MC1530 operational amplifier [11].

computed. For each concentration, data simulating 100 chips are generated, and the mean R_{SE} is determined. Amplifier performance is then evaluated in terms of differential gain (A_d), input offset voltage ($V_{\text{in,off}}$), and input bias current (I_{bias}). The results are shown in Figure 5.21. Performance is considered acceptable if $A_d > 8000$, $-1.5\text{ mV} < V_{\text{in,off}} < 1.5\text{ mV}$, and $-1.2\ \mu\text{A} < I_{\text{bias}} < 1.2\ \mu\text{A}$. The best yield is obtained when R_{SE} is near 4 Ω/square, and cannot be increased very much without a significant drop in yield.

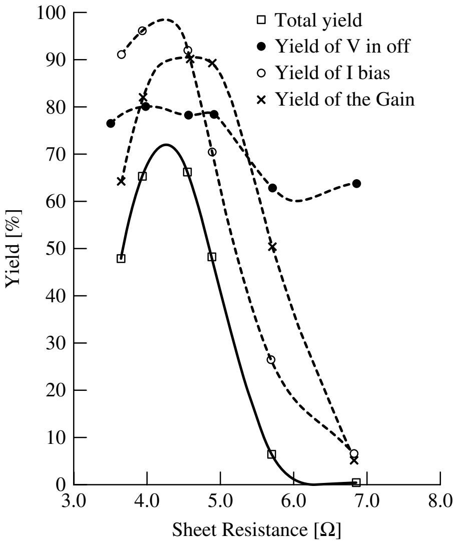


Figure 5.21. Yield of MC1530 operational amplifier versus emitter sheet resistance [11].

FABRICS is a powerful tool for computing parametric yield that can easily be tuned to the random variations of a real manufacturing process. It can be used for bipolar, MOS, or any other process technology, so long as appropriate data and analytical physical models are available.

5.6. DESIGN CENTERING

Yield simulation tools provide a mechanism for yield optimization and quality enhancement through accounting for manufacturing variations in IC components during the design phase. The objective of such efforts is to minimize circuit performance sensitivity with respect to potential component and parameter fluctuations. The objective of *design centering* is to maximize yield by identifying an optimal set of design parameters (x_{opt}) such that yield is optimized.

This concept is illustrated graphically for a simple two-parameter design space in Figure 5.22. In this figure, the region labeled *A* represents the region of acceptable circuit performance. The oval centered at the coordinates of the design

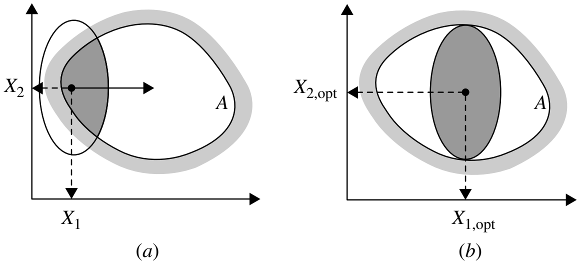


Figure 5.22. Illustration of design centering: (a) initial low yield; (b) optimized yield [2].

parameters x_1 and x_2 represents the area over which these two parameters may vary during manufacturing. The yield is therefore the shaded area represented by the overlap of these two regions. Figure 5.22a corresponds to a situation in which a poor choice of x_1 and x_2 results in low initial yield. In contrast, by centering the design with an optimal choice in the two parameters ($x_{1,\text{opt}}$ and $x_{2,\text{opt}}$), yield is maximized. Another term for the process of design centering is *design for manufacturability*.

5.6.1. Acceptability Regions

The *acceptability region* is defined as the part of the space of performance parameters in which all constraints imposed on circuit performance are fulfilled. For the two-dimensional example represented by Figure 5.22, the area A represents the acceptability region. In general, A is an m-dimensional hypersurface defined by the inequality

$$S_j^L \leq y_j \leq S_j^U \qquad j = 1, \ldots, m \tag{5.33}$$

where y_j is one of m performance parameters and S_j^L and S_j^U are the (usually designer-defined) lower and upper bounds imposed on these parameters.

Since complicated relationships between performance parameters and bounds can be defined, acceptability regions can also be very complicated, including nonconvex regions or internal unacceptababilty regions ("holes"). In order to determine whether a given point in the circuit parameter space belongs to A or its complement, an indicator function $I(\boldsymbol{x})$ is used, where

$$I(x) = \begin{cases} 1 & \mathbf{x} \in A \\ 0 & \mathbf{x} \notin A \end{cases} \tag{5.34}$$

The points for which $I(x) = 1$ are called successful, or "pass" points, and those for which $I(x) = 0$ are called "fail" points.

Except for some simple cases, the shape of the acceptability region in the performance space is unknown and nearly impossible to define completely. However, for yield optimization purposes, either implicit or explicit knowledge of A and its boundaries is required. Therefore, it is necessary to approximate the shape of A. Several methods are available to do so [2]. A few of these are illustrated in Figure 5.23.

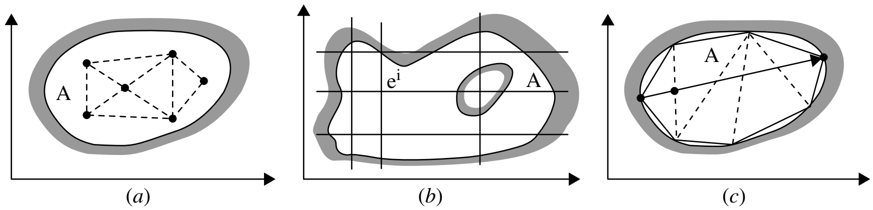


Figure 5.23. Various methods of acceptability region approximation: (a) point-based; (b) ODOS; (c) simplicial [2].

For example, in the *point-based* approximation shown in Figure 5.23a, subsequent approximations to A are generated using Monte Carlo simulation [13]. After each new point is generated, it is determined whether it belongs to A_{i-l} (i.e., the latest approximation to A). If it does, the sampled point is considered successful. If it does not belong to A_{i-l} and the next circuit simulation reveals that it belongs to A, the polyhedron is expanded to include the new point. In the method of approximation called *one-dimensional orthogonal search* (ODOS) shown in Figure 5.23b, line segments passing through the points e_i are randomly sampled in the performance space parallel to the coordinate axes and used for the approximation of A [14]. ODOS is very efficient for large linear circuits, since the intersections with A can be *directly* found from analytical formulas. The *simplicial approximation* is based on approximating the boundary of A in the performance space by a polyhedron [15]. The boundary of A is assumed to be convex (see Figure 5.23c). The simplicial approximation is obtained by locating points on the boundary of A by a systematic expansion of the polyhedron. The search for the next vertex is always performed in the direction passing through the center of the largest face of the polyhedron already existing and perpendicular to that face.

5.6.2. Parametric Yield Optimization

After estimation of the acceptability region, yield optimization is the objective of design centering techniques. In so doing, design centering attempts to inscribe the largest hypersphere of input parameter variation (also called the *norm body*; see Figure 5.22) into the approximation of the acceptability region A. The center of the largest norm body is taken as the optimal vector of input parameters $\boldsymbol{x}$. Consider, for example, a simplicial approximation to the acceptability region. Under these conditions, several yield optimization schemes have been proposed. The most typical is as follows.

After a nominal point x belonging to the acceptability region is identified, line searches via circuit simulation are performed from that point to obtain some points located on the boundary of A. Several simulations may be required to find one boundary point. To form a polyhedron in an n-dimensional space, at least $(n + 1)$ boundary points need to be found. Once the first polyhedron approximation to A is obtained, the largest possible norm body is inscribed into it (using linear programming techniques), and its center is assumed as the first approximation to the center of A. The next steps involve improvements to the current approximation, $\tilde{A}$, by expanding the simplex. The center of the largest polyhedron face is found, and a line search is performed from the center along the line passing through it in a direction orthogonal to the face considered, to obtain another vertex point on the boundary of A. The polyhedron is then inflated to include the new point generated. This process is repeated until no further improvement is obtained.

Intuitively, this method should improve yield but, because of the approximation used, will not necessarily maximize it. For example, the simplicial approximation will not be accurate if A is nonconvex, and it will fail if A is not simply connected, since some parts of the approximation A will be outside the actual acceptability region. Notably, the computational cost of obtaining the simplicial approximation quickly increases for high-dimensional performance spaces. Thus, this method is most suitable for problems with a small number of designable parameters.

5.7. PROCESS INTRODUCTION AND TIME-TO-YIELD

While the final or "steady state" yield is important, one does not achieve the final yield instantaneously. Indeed, extensive field studies show that new processes arrive in the manufacturing line with limited initial yield [16]. As shown graphically in Figure 5.24, the initial process introduction period is followed by a period of intense learning where the various key yield detractors are identified and removed. The length of this "rapid learning" period is of paramount importance, as it often limits the amount of time it takes to bring a new product to market. Time-to-market is critical, if one is to capture significant market share and avoid the rapid price erosion that follows the introduction of high-end IC products. The final period, in which yield levels off and approaches a maximum, is one characterized by small gains due to the removal of the last few yield detractors. During this period, the investment of further effort and expense for marginal returns is questionable.

The discussion above underscores the importance of a "dynamic" study of yield. The objective of studying the metric known as *time-to-yield* is to identify key methods, tools, and actions that can accelerate the initial learning period following the introduction of a new process. As one might suspect, many factors affect time-to-yield, and some do so in ways that are not easily quantifiable. For example, it has been shown that time-to-yield can be accelerated by simply accelerating the processing cycle, as this allows for more '*work-in-progress*' (WIP) turns and more rapid acquisition of the required process understanding. Another factor that accelerates time-to-yield appears to be the systematic (and

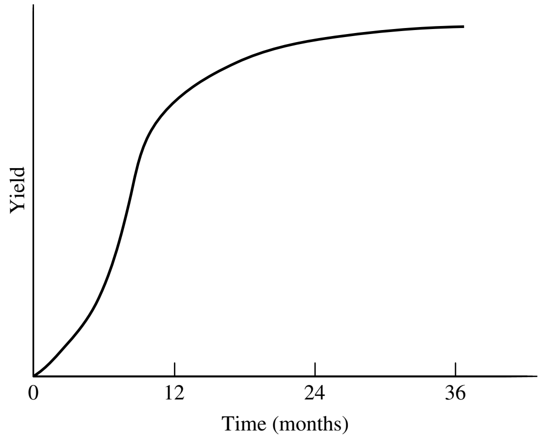


Figure 5.24. Yield learning curve.

often automated) collection of data, especially if this is done in the context of a well-structured quality control program.

While these aspects are difficult to quantify, one can draw interesting conclusions from carefully organized field studies. One such study was done in the context of the Competitive Semiconductor Manufacturing program, run at the University of California at Berkeley in 1994. This program involved studying a large number of IC production facilities, and it included detailed questionnaires, as well as site visits by multidisciplinary groups of experts. In this study, researchers recorded yield–time data for a variety of facilities, covering products ranging from memories to high-end logic ICs. While the main objective of the study was to capture the main reasons behind achieving high yield numbers, the data also offered a rare opportunity to examine the time-to-yield figure on a qualitative basis. Some of the yield data are shown in Figure 5.25.

Graphs like Figure 5.24 plot a "normalized" total yield figure that is appropriately adjusted for die size, minimum feature size, and other properties. Using several such datasets for various IC technologies, the authors of the study created and calibrated an empirical model of the following form

$$W_j = \alpha_o j + \alpha_{1j}(\text{die Size}) + \alpha_{2j}\log(\text{process age}) \qquad (5.35)$$

where j is the index defining the various semiconductor manufacturing facilities participating in the study, (die size) is in cm^2, and (process age) is the time in months from the oldest to the most recent die yield data point. This model implies that the yield increases logarithmically as processes mature, and the coefficient α_{2j} is a quantitative figure of merit that captures the unique ability of each facility to rapidly improve the yield of a new process. Table 5.4 attempts to capture the impact of various factors on the α_{2j} "yield learning coefficient." In this table, because of the limited sample size, the facilities are divided into three

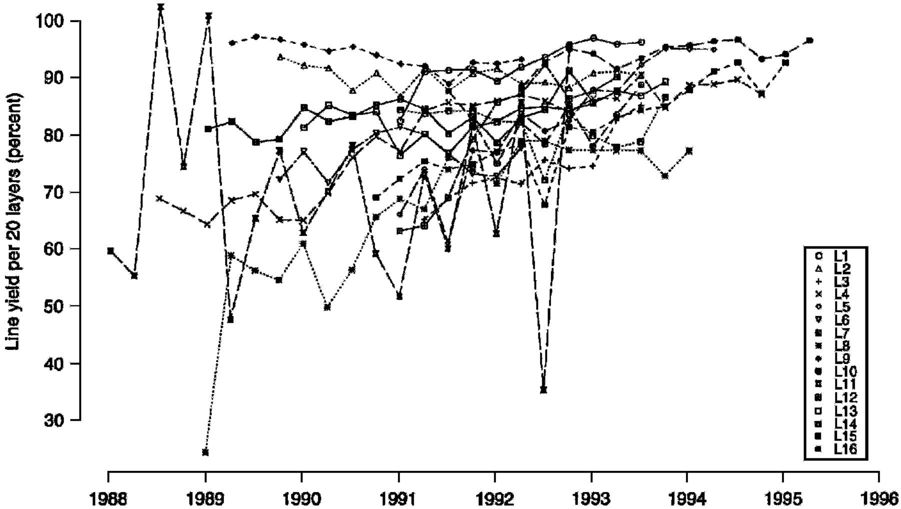


Figure 5.25. Line yield per IC layer versus year for several major IC manufacturers.

Table 5.4. Impact of various factors on yield learning.

	SPC AUTOMATION				EXTENT OF SPC USE		
	Yes	No			High	Med	Low
High	3	0		High	1	2	0
Med	6	0		Med	2	3	1
Low	2	4		Low	1	3	2
	PAPERLESS WAFER TRACKING				EXTENT OF CAM AUTOMATION		
	Yes	No			Full	Semi	Manual
High	1	2		High	1	2	0
Med	3	3		Med	3	2	1
Low	1	5		Low	1	1	4
	YIELD MODEL IN USE				YIELD GROUP PRESENT IN FACTORY		
	Homegrown		Away		Yes	No	
High	2		1	High	3	0	
Med	1		5	Med	3	3	
Low	2		4	Low	4	2	
	GEOGRAPHIC FAB REGION						
	USA	Asia	Europe				
High	3	0	0				
Med	1	4	1				
Low	4	2	0				

categories relating the yield learning speed (low, medium, high). Facilities with $\alpha_2 > 1.00$ received a high yield improvement rating, and facilities with $\alpha_2 < 0.30$ were given a low improvement rating. The number of facilities falling in each category is also noted below.

Here, the term "yield model" refers the specific formula used by that organization to predict the yield of the process as a function of a measure of defect density. "Paperless" is an indicator of the extent of computer-aided manufacturing in the facility. Only three of the facilities under study were fully paperless (i.e., had no lot travelers or run cards accompanying production lots). "SPC automation" is an indicator of whether the SPC control charting function is automated. "Extent of SPC" practice is a subjective rating of each facility's commitment to and execution of SPC. "Yield group" refers to the existence of a yield engineering group at the facility. In nearly every case where there is a yield group, their efforts are supplemented by product engineering and other entities within the fab. This wide spread in "yield learning" rates indicates that there are still many unknown factors that control this very important figure.

SUMMARY

In this chapter, we have provided a general overview of the concept of manufacturing yield for semiconductor products. We have done so by differentiating between functional and parametric yield, and by describing various quantitative models and simulation tools for each. Finally, we have discussed yield learning in the context of its financial implications with regard to product time-to-market. In subsequent chapters, we will discuss how high yield is maintained via statistical process control techniques.

PROBLEMS

5.1. Assuming a Poisson model, calculate the maximum defect density allowable on 100,000 NMOS transistors in order to achieve a functional yield of 95%. Assume that the gate of each device is 10 μm wide and 1 μm long.

5.2. Use Murphy's yield integral to derive Eqs. (5.11), (5.12), and (5.14).

5.3. Suppose that the probability density function of the defect density for a given IC manufacturing process is given by

$$f(D) = -100D + 10$$
$$0 \leq D \leq 0.1$$

If D is measured in cm^{-2} and the critical area for this IC is 100 cm^2, what functional yield can we expect for the process over the range of defect densities from 0.05 to 0.1 cm^{-27}.

5.4. Consider the effect of defects on IC interconnect. Figure P5.4 illustrates the impact of defect size on critical area for circular defects of diameter x. The area in which the center of such defects must fall to cause a failure increases linearly as a function of defect size. It can be expressed as

$$A_c(x) = L(x + w - 2R)$$
$$R \leq x \leq \infty$$

where L is the interconnect length and R is the allowable gap in the interconnect line. Suppose that the normalized probability density function of defect sizes is given by

$$g(x) = \frac{X_U^2 X_L^2}{x^3(X_U^2 - X_L^2)}$$

where X_L and X_U are the lower and upper limits of the range of defect sizes, respectively.

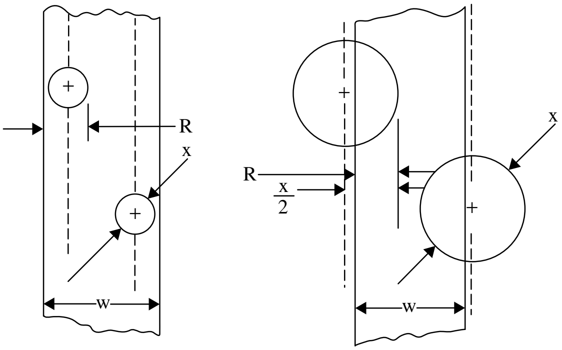


Figure P5.4

(a) Given $R \geq X_L$, find an expression for the average critical area (A_{av}) by evaluating the integral

$$A_{\mathrm{av}} = \int_{X_L}^{X_U} A_c(x) g(x) dx$$

(b) Show that as the upper limit on defect size approaches infinity, then

$$A_{\mathrm{av}} = \frac{L X_L^2 w}{2R^2}$$

5.5. Assume that 10,000 units of a product with area 0.5 cm^2 and 200 chips per wafer are to be produced in three manufacturing areas, with a D_0 of 0.9, 1.1, and 1.3 cm^{-2}, respectively. How many wafers need to be ordered? Assume a negative binomial model with $\alpha = 2$, and also assume that these steps will be followed by assembly and test. The combined assembly and test steps have a yield of 95%.

5.6. A new product with a critical area of 0.45 cm^2 is to be produced using a technology with a defect density of 0.5 cm^{-2}. Three similar products are already being produced using this technology, and their critical areas and yield data appear in Table P5.6. Analyze the data and calculate the short-term and long-term yield expectations using the Poisson model.

Table P5.6

A$_C$(cm^2)	Measured Yield (%)
0.1	81
0.2	78
0.4	70

5.7. Suppose we are given the joint distribution of several parameters which vary in an IC manufacturing process, and we would like to evaluate the impact of these variations on the overall performance of the IC by evaluating its parametric yield. For example, the drive current in mA of a MOSFET in saturation (I_{Dsat}) is given by

$$I_{Dsat} = \frac{k}{2}(V_{GS} - V_T)^2$$

where k is the device transconductance parameter, V_{GS} is the gate-source voltage, and V_T is the threshold voltage. Suppose that V_T is subject to variation, and that variation ultimately impacts I_{Dsat}. There is a way to find (analytically) the pdf of y, $f_y(y)$, if $y = g(x)$ and if the pdf of x is known.

To do so, we first solve the equation:

$$y = g(x)$$

for x in terms of y. If $x_1, x_2, \ldots, x_n$ are the real roots of this expression, then

$$f_y(y) = \frac{f_x(x_1)}{|g'(x_1)|} + \cdots + \frac{f_x(x_n)}{|g'(x_n)|}$$

where $g'(x) = \dfrac{dg(x)}{dx}$

(a) Find the analytical expression for the pdf for the drive current, if V_T is uniformly distributed between 0.3 and 0.8 V, and if all other parameters are deterministic.

(b) If $k = 1$ mA/V^2, determine the parametric yield for a large population of transistors that achieve drive currents between 1.5 and 2.0 mA, if V_{GS} is 2.5 V.

5.8. An oncoming 150-mm wafer has 10 randomly spaced point defects on it. The chip size is 1.0 cm^2, and the final target yield is 75%. If there are eight additional processing levels, what is the maximum number of defects that we can afford to accumulate on each level? (Use the Poisson yield model.)

5.9. A manufacturing facility has a yield that is controlled purely by random defects. The density of these random defects depends on the design rule used. More specifically, for a 1-μm design rule, the density is 0.5/cm^2, while for a 0.5-μm design rule, the density is 2.0/cm^2. (Use the Poisson yield model.)

(a) A given product takes 1.0 cm^2. Further, 90% of this area is using 1 μm design rules, while the rest 10% is using the 0.5 μm design rules. Estimate the yield of this product.

(b) This product can be redesigned (shrunk) to take only 0.5 cm^2, but now 50% of the chip is using the 0.5 μm design rules. Estimate the yield of the redesigned product.

(c) What would be the ratio of good die per wafer of the redesigned product to that of the original product?

5.10. Suppose that you use 200-mm wafers, and also assume that you can get functional dies only within the inner 190-mm diameter (outer 5-mm margin is full of defects). On the one product that you have run so far, a chip with area 5 × 5 mm, the yield is 80%.

(a) Using the simple Poisson model, find the defect density (in the good area of the wafer) and plot the yield as a function of S, where S is the square root of the area of the die in production. Plot the total and the good die per wafer as a function of S on the same graph.

(b) Repeat the calculations and plots in (a) using the negative binomial model ($\alpha = 1.5$).

(c) Suppose that an alternative explanation for the data were that some fraction f of the wafer were perfect and the rest were totally dead.

This is the "black–white" model that assumes a perfect deterministic clustering of defects. What is f? Plot the "good die" per wafer for this model on the same graph as in (a)–(b).

(d) What defect density reduction would you have to achieve to yield 50% of the available die at $S = 15$ mm according to models (a), (b), and (c)?

REFERENCES

1. B. Murphy, "Cost-Size Optima of Monolithic Integrated Circuits," *Proc. IEEE* **52**(12), 1537–1545 (Dec. 1964).
2. J. Piñeda de Gyvez and D. Pradhan, *Integrated Circuit Manufacturability*, IEEE Press, 1999.
3. R. Seeds, "Yield and Cost Analysis of Bipolar LSI," *IEEE Intl. Electron Devices Meet.*, Washington, DC, Oct., 1967.
4. J. Price, "A New Look at Yield of Integrated Circuits," *Proc. IEEE* (Lett.) **58**, 1290–1291 (Aug. 1970).
5. T. Okabe, Nagata, and Shimada, "Analysis of Yield of Integrated Circuits and a New Expression for the Yield," in *Defect and Fault Tolerance in VLSI Systems*, C. Stapper, ed., Vol. 2, Plenum Press, 1990, pp. 47–61.
6. C. Stapper, "Fact and Fiction in Yield Modeling," *Microelectron. J.* **210**(1–2), 129–151 (May 1989).
7. J. Cunningham, "The Use and Evaluation of Yield Models in Integrated Circuit Manufacturing," *IEEE Trans. Semiconduct. Manuf.* **3**(2), 60–71 (May 1990).
8. C. Stapper and R. Rossner, "A Simplified Method for Modeling VLSI Yields," *Solid State Electron.* **25**(6), 655–657 (July 1973).
9. C. Stapper, "Modeling Defects in Integrated Circuit Photolithographic Patterns," *IBM J. Res. Devel.* **28**(4), 461–475 (July 1984).
10. W. Brown, *Advanced Electronic Packaging*, IEEE Press, Piscataway, NJ, 1999.
11. W. Maly and A. Strojwas, "Statistical Simulation of the IC Manufacturing Process," *IEEE Trans. CAD ICs Syst.* **1**(3) (July 1982).
12. D. Walker, *Yield Simulation for Integrated Circuits*, Kluwer, Boston, 1987.
13. S. Director and G. Hatchel, "A Point Basis for Statistical Design," *Proc. IEEE Intl. Symp. Circuits and Systems*, New York, May, 1978.
14. J. Ogrodzki, L. Opalski, and M. Styblinski, "Acceptability Regions for a Class of Linear Networks," *Proc. IEEE Intl. Symp. Circuits and Systems*, Houston, May 1980.
15. S. Director and G. Hatchel, "The Simpilical Approximation Approach to Design Centering," *IEEE Trans. Circuits Syst.* **CAS-24**(7) (July 1977).
16. S. Cunningham, C. J. Spanos, and K. Voros, "Semiconductor Yield Improvement: Results and Best Practices", *IEEE Trans. Semiconduct. Manuf.* **8**(2), 103–109 (May 1995).

6

STATISTICAL PROCESS CONTROL

OBJECTIVES

- Provide an overview of statistical process control (SPC) techniques.
- Define and describe various types of control charts.
- Differentiate between control charts for attributes and control charts for variables.
- Introduce a few advanced SPC concepts.

INTRODUCTION

Manufacturing processes must be stable, repeatable, and of high quality to yield products with acceptable performance. This implies that all individuals involved in manufacturing a product (including operators, engineers, and management) must continuously seek to improve manufacturing process output and reduce variability. Variability reduction is accomplished in a large part by strict process control. The application of process control in manufacturing continues to expand in the semiconductor industry. In this chapter we will focus on statistical process control techniques a as means to achieve high-quality products.

Statistical process control (SPC) refers to a powerful collection of problem-solving tools used to achieve process stability and reduce variability. Perhaps the primary and most technically sophisticated of these tools is the control chart. The

Fundamentals of Semiconductor Manufacturing and Process Control,
By Gary S. May and Costas J. Spanos

control chart was developed by Dr. Walter Shewhart of Bell Telephone Laboratories in the 1920s. For this reason, control charts are also often referred to as *Shewhart control charts*.

6.1. CONTROL CHART BASICS

A control chart is used to detect the occurrence of shifts in process performance so that investigation and corrective action may be undertaken to bring an incorrectly behaving manufacturing process back under control. A typical control chart is shown in Figure 6.1. This chart is a graphical display of a quality characteristic that has been measured from a sample versus the sample number or time. The chart consists of: (1) a *centerline*, which represents the average value of the characteristic corresponding to an in-control state; (2) an *upper control limit* (UCL); and (3) a *lower control limit* (LCL). The control limits are selected such that if the process is under statistical control, nearly all the sample points will plot between them. Points that plot outside the control limits are interpreted as evidence that the process is out of control.

There is a close connection between control charts and the concept of hypothesis testing, which was discussed in Chapter 4. Essentially, the control chart represents a continuous series of tests of the hypothesis that the process is under control. A point that plots within the control limits is equivalent to accepting the hypothesis of statistical control, and a point outside the limits is equivalent to rejecting this hypothesis. We can think of the probability of a type I error (a "false alarm") as the probability of concluding that the process is out of statistical control when it really is under control, and of the probability of a type II error (a "missed alarm") as the probability of concluding that the process is under control when it really is not.

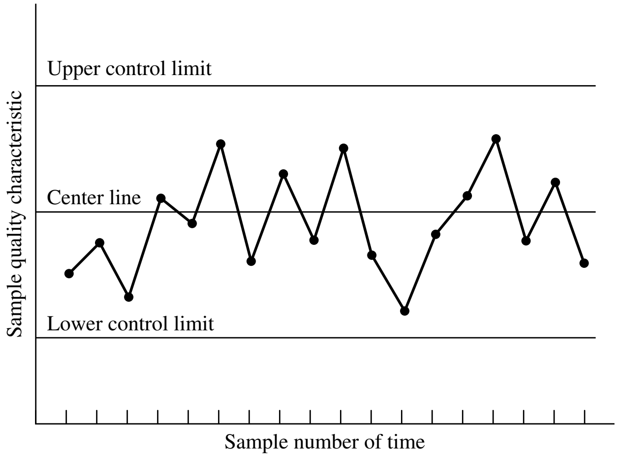


Figure 6.1. Typical control chart [1].

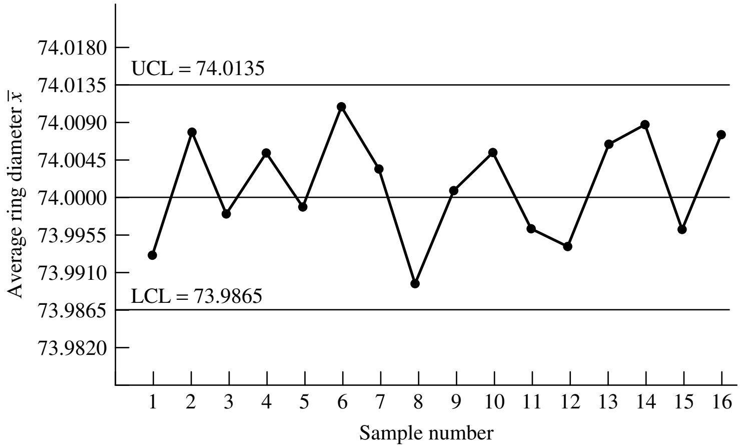


Figure 6.2. $\bar{x}$ chart for via diameter [1].

To illustrate, consider an example pertaining to the formation of vias in a dielectric layer. Suppose that this process can be controlled at mean via diameter of 74 μm, and the standard deviation of the diameter is 0.01 μm. A control chart for via diameter is shown in Figure 6.2. For every product wafer, a sample of five via diameters are measured, and that sample average, $\bar{x}$, is plotted on the chart. Note that all the points fall within the control limits, indicating that the via formation process is under statistical control.

Let's examine how the control limits in this example were determined. For a sample size of $n = 5$ vias, the standard deviation of the sample average is

$$\frac{\sigma}{\sqrt{n}} = \frac{0.01}{\sqrt{5}} = 0.0045\ \mu\text{m} \tag{6.1}$$

If we assume that x is normally distributed, we would expect $100(1 - \alpha)\%$ of the sample mean diameters to fall within $74 + z_{\alpha/2}(0.0045)$ and $74 - z_{\alpha/2}(0.0045)$. If the constant $z_{\alpha/2}$ is selected to be 3, the upper and lower control limits become

$$\text{UCL} = 74 + 3(0.0045) = 74.0135\ \mu\text{m}$$

$$\text{LCL} = 74 - 3(0.0045) = 73.9865\ \mu\text{m}$$

These are typically called "3-sigma" (3σ) control charts, where "sigma" refers to the standard deviation of the sample average computed in Eq. (6.1). Note that the selection of the control limits is equivalent to testing the hypothesis

$$\text{H}_0\text{:}\ \ \mu = 74$$

$$\text{H}_1\text{:}\ \ \mu \neq 74$$

where $\sigma = 0.01$ is known. Essentially, the control chart just tests this hypothesis repeatedly for each sample. This is illustrated graphically in Figure 6.3.

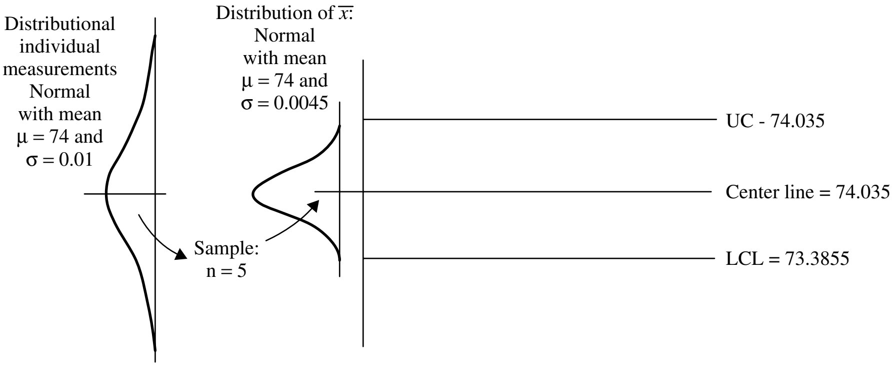


Figure 6.3. Illustration of how a control chart works [1].

An important parameter for any control chart is the *average runlength* (ARL), which is defined as the average number of samples taken before the control limits are exceeded. Mathematically, the ARL is $1/P$ (*a sample point plots out of control*). Thus, if the process is *in control*, the ARL is

$$\text{ARL} = 1/\alpha \tag{6.2}$$

where α is the probability of a type I error. If the process is *out of control*, then the ARL is

$$\text{ARL} = \frac{1}{1 - \beta} \tag{6.3}$$

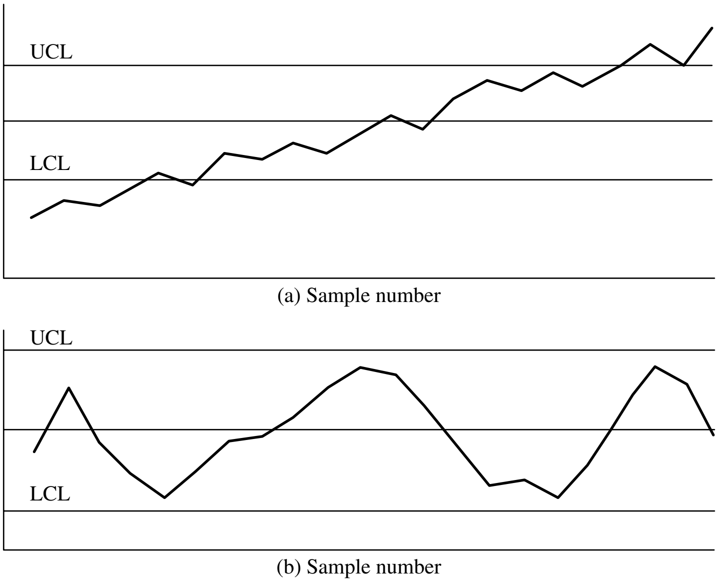
where β is the probability of a type II error.

6.2. PATTERNS IN CONTROL CHARTS

A control chart may indicate an out-of control condition when a point plots beyond the control limits or when a sequence of points exhibit nonrandom behavior. For example, consider the charts shown in Figure 6.4. The pattern in Figure 6.4a is called a "trend" (or "run"). Although most of the points in this chart are within the control limits, they are not indicative of statistical control because their pattern is very nonrandom. A pattern of several consecutive points on the same side of the centerline is also called a "run." A run of several points has a very low probability of occurrence in a truly random sample.

Other types of patterns may also indicate an out-of-control state. For example, the chart in Figure 6.4b exhibits cyclic (or periodic) behavior, even though all the points are within the control limits. This type of pattern might result from operator fatigue, raw-materials depletion, or other periodic problems. Several other special patterns in control charts might be suspicious, including

- *Mixtures*—points from two or more source distributions
- *Shifts*—abrupt changes
- *Stratification*—charts that exhibit unusually small variability



(a) Sample number

(b) Sample number

Figure 6.4. Examples of patterns in control charts: (a) trend; (b) cyclic.

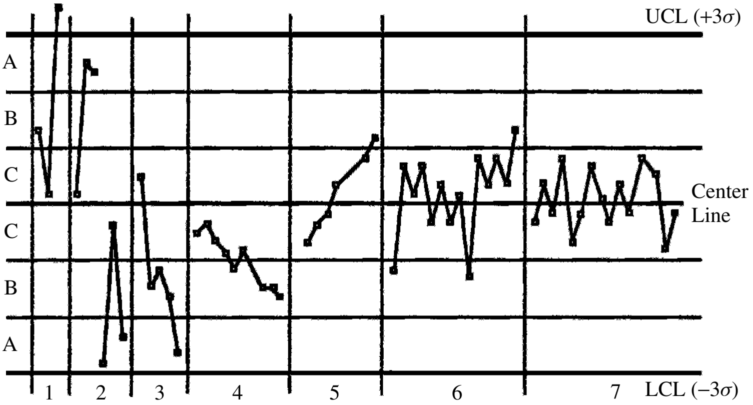


Figure 6.5. Illustration of Western Electric rules.

In order to detect patterns such as these, special rules must be applied. The Western Electric *Statistical Quality Control Handbook* [3] provides a set of rules for detecting nonrandom patterns in control charts. Referring to Figure 6.5, these rules state that a process is out of control if either:

1. Any single point plots beyond the 3σ control limits.
2. Two out of three consecutive points plot beyond the 2σ warning limits (zone A).
3. Four out of five consecutive points plot beyond 1σ (zone B).

4. Nine consecutive points plot on the same side of the centerline.
5. Six consecutive points increase or decrease.
6. Fourteen consecutive points alternate up and down.
7. Fifteen consecutive points plot on either side in zone C.

Each of these rules describes events with a low natural probability of occurrence, thereby making them indicative of potential out-of-control behavior. To illustrate, consider rule 1. The probability of one point being outside the 3σ control limits is given by

$$P(z > 3 \text{ or } z < -3) = P(z > 3) + P(z < -3) = 2(0.00135) = 0.0027$$

where z is the standard normal random variable presented in Chapter 4, and the probabilities are calculated using Appendix B. These so-called Western Electric rules have been shown to be very effective in enhancing the sensitivity of control charts and identifying troublesome patterns.

Example 6.1. Consider Western Electric rule 2. What is the probability of two out of three consecutive points plotting beyond the 2σ warning limits?

Solution: Using Appendix B, we can find the probability of one point plotting beyond the 2σ limits as

$$P(z > 2 \text{ or } z < -2) = P(z > 2) + P(z < -2) = 2(0.02275) = 0.0455$$

Therefore, assuming that each point represents an independent event, the probability of two out of three points plotting beyond the 2σ limits is

$$P(2 \text{ out of } 3) = (0.0455)(0.0455)(1 - 0.0455) = 0.00198$$

where the $(1 - 0.0455)$ factor is the probability of the third point plotting inside the 2σ limits.

6.3. CONTROL CHARTS FOR ATTRIBUTES

Some quality characteristics cannot be easily represented numerically. For example, we may be concerned with whether a contact is defective. In this case, the contact is classified as either "defective" or "nondefective" (or equivalently, "conforming" or "nonconforming"), and there is no numerical value associated with its quality. Quality characteristics of this type are referred to as *attributes*. In this section, three commonly used control charts for attributes are presented: (1) the fraction nonconforming chart (*p chart*), (2) the defect chart (*c chart*), and (3) the defect density chart (*u chart*).

6.3.1. Control Chart for Fraction Nonconforming

The fraction nonconforming is defined as the number of nonconforming items in a population divided by the total number of items in the population. The control chart for fraction nonconforming is called the *p chart*, which is based on the binomial distribution (see Section 4.1.1.1). Suppose that the probability that any product in a manufacturing process will not conform is p. If each unit is produced independently, and a random sample of n products yields D units that are nonconforming, then D has a binomial distribution. In other words

$$P(D = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, \ldots n \tag{6.4}$$

The sample fraction nonconforming ($\hat{p}$) is defined as

$$\hat{p} = \frac{D}{n} \tag{6.5}$$

As noted in Section 4.1.1.1, the mean and variance of $\hat{p}$ are $\mu_{\hat{p}} = p$ and $\sigma^2_{\hat{p}} = p(1-p)/n$, respectively. On the basis of these relationships, we can set up the centerline and ±3σ control limits for the p chart as follows:

$$\begin{aligned} \text{UCL} &= p + 3\sqrt{\frac{p(1-p)}{n}} \\ \text{Centerline} &= p \\ \text{LCL} &= p - \sqrt{\frac{p(1-p)}{n}} \end{aligned} \tag{6.6}$$

This above implementation of the p chart assumes that p is known (or given). If p is not known, it must be computed from the observed data. The usual procedure is to select m preliminary samples, each of size n. If there are D_i nonconforming units in the ith sample, then the fraction nonconforming is

$$\hat{p}_i = \frac{D_i}{n} \qquad i = 1, 2, \ldots, m \tag{6.7}$$

and the average of the individual fractions nonconforming is

$$\overline{p} = \frac{1}{mn}\sum_{i=1}^{mn} D_i = \frac{1}{m}\sum_{i=1}^{mn} \hat{p}_i \tag{6.8}$$

The centerline and control limits for the p chart under these conditions are

$$\begin{aligned} \text{UCL} &= \overline{p} + 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} \\ \text{Centerline} &= \overline{p} \\ \text{LCL} &= \overline{p} - 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} \end{aligned} \tag{6.9}$$

Example 6.2. Consider a wire bonding operation. Suppose that 30 samples of size $n = 50$ have been collected from 30 chips. Given a total of 347 defective bonds found, set up the $\pm 3\sigma$ p chart for this process.

Solution: Using Eq. (6.8), we have

$$\overline{p} = \frac{1}{mn} \sum_{i=1}^{m} D_i = \frac{347}{(30)(50)} = 0.2313$$

This is the centerline for the p chart. The upper and lower control limits can be found from Eq. (6.9) as

$$\text{UCL} = \overline{p} + 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} = 0.4102$$

$$\text{LCL} = \overline{p} - 3\sqrt{\frac{\overline{p}(1-\overline{p})}{n}} = 0.0524$$

It should be pointed out that the limits defined by Eq. (6.9) are actually just *trial* control limits. They permit the determination of whether the process was in control when the m samples were collected. To test the hypothesis that the process was in fact under control during this period, the sample fraction nonconforming from each sample on the chart must be plotted and analyzed. If all points are inside the control limits and no systematic trends are evident, then it may be concluded that the process was indeed under control, and the trial limits are reasonable.

If, on the other hand, one or more of the $\hat{p}_i$ statistics plots out of control when compared to the trial control limits, then the hypothesis of past control must be rejected, and the trial limits are no longer valid. It then becomes necessary to revise the trial control limits by first examining each out-of-control point in an effort to identify an assignable cause. If a cause can be found, the point in question is discarded and the control limits are recalculated using the remaining points. The remaining points are then reexamined, and this process is repeated until all points plot in control, at which point the trial limits may be adopted as valid.

6.3.1.1. Chart Design

Constructing a p chart requires that the sample size, frequency of sampling, and width of the control limits all be specified. Obviously, the sample size and sampling frequency are interrelated. Assuming 100% inspection for a given production rate, selecting a sampling frequency fixes the sample size.

Various rules have been suggested for the choice of sample size (n). If p is very small, n must be sufficiently large that we have a high probability of finding at least one nonconforming unit in a sample in order for the p chart to be effective. Otherwise, the control limits might end up being so narrow that the presence of only a single nonconforming unit in a sample might indicate an out-of-control condition. For example, if $p = 0.01$ and $n = 8$, then the 3σ upper control limit is

$$\text{UCL} = p + \sqrt{\frac{p(1-p)}{n}} = 0.1155$$

With only one nonconforming unit, $\hat{p} = 0.125$, and the process appears to be out of control.

To avoid this problem, Duncan has suggested that the sample size be large enough to ensure an approximately 50% chance of detecting a process shift of some specified amount [2]. For example, let $p = 0.01$, and suppose that we want the probability of detecting a shift from $p = 0.01$ to $p = 0.05$ to be 0.5. Assuming that the normal approximation to the binomial distribution applies, this implies that n must be selected such that the UCL exactly coincides with the fraction nonconforming in the out-of-control state. In general, if δ is the magnitude of this process shift, then n is given by

$$\delta = k\sqrt{\frac{p(1-p)}{n}} \tag{6.10}$$

In our example, $\delta = 0.05 - 0.01 = 0.04$, and if 3σ limits are used (i.e., $k = 3$), then

$$n = \left(\frac{k}{\delta}\right)^2 p(1-p) = \left(\frac{3}{0.04}\right)^2 (0.01)(0.99) = 56$$

If the in-control value of the fraction nonconforming is small, it is also desirable to choose n large enough so that the p chart will have a positive lower control limit. This will allow us to detect samples that have an unusually small number of nonconforming items. In other words, we want

$$\text{LCL} = p - k\sqrt{\frac{p(1-p)}{n}} > 0 \tag{6.11}$$

or

$$n > \frac{(1-p)}{p}k^2 \tag{6.12}$$

Note that this is not always practical. If we want the chart in our example to have a positive LCL, this will require that $n \geq 891$.

6.3.1.2. Variable Sample Size

In some applications, the sample size for the fraction nonconforming control chart is not fixed. In these cases, there are several approaches to constructing the p chart. The first, and probably the simplest, approach is to determine control limits according to the specific size of each sample. In other words, if the ith sample is of size n_i, the upper and lower 3σ control limits are placed at $p \pm 3\sqrt{p(1-p)/n_i}$. However, this results in control limits that vary for each sample, as shown in Figure 6.6.

The approach described above is somewhat unappealing. A second approach is to use the *average* sample size to compute the control limits. This assumes that the

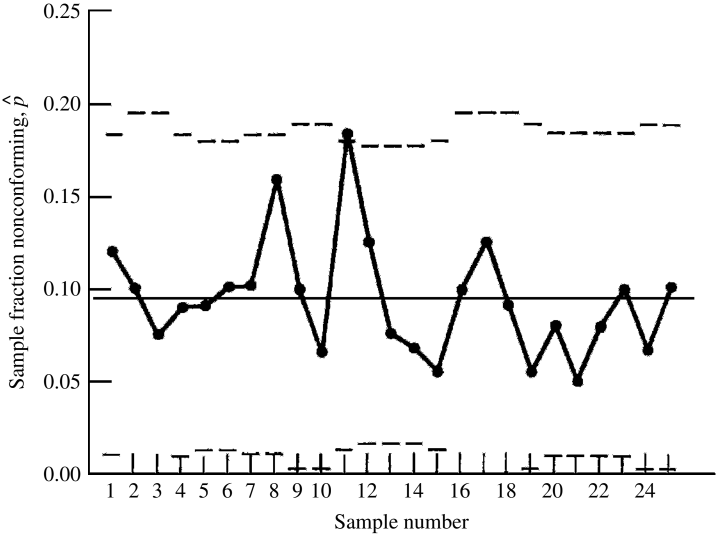


Figure 6.6. Example of control chart for fraction nonconforming with variable sample size [1].

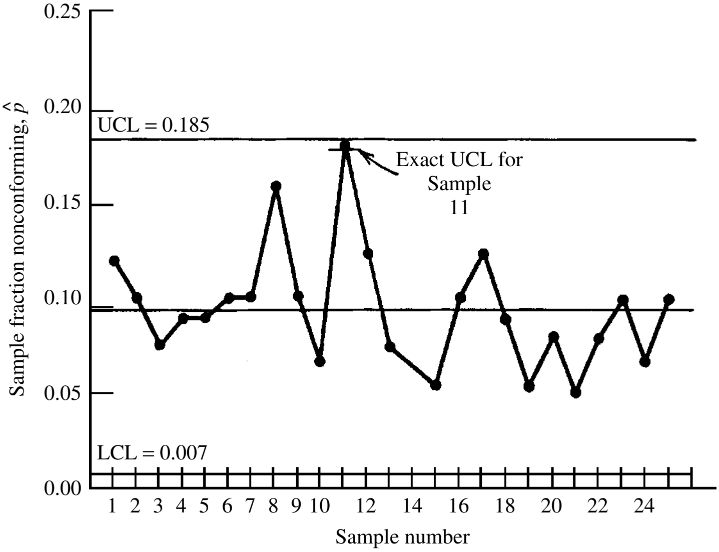


Figure 6.7. Control chart for fraction nonconforming based on average sample size [1].

sample sizes will not differ appreciably over the duration of the chart. The result is a set of control limits that are approximate, but constant, and therefore more satisfying and easier to interpret. Applying this approach to the same dataset used in Figure 6.6 results in the chart shown in Figure 6.7. Care must be exercised in the interpretation of points near the approximate control limits, however. Notice that $\hat{p}$ for sample 11 in Figure 6.7 is close to the upper control limit, but appears to be in control. When compared the exact limits used in Figure 6.6, though, this point appears to be out of control. Similarly, points outside the approximate limits may indeed be inside their exact limits.

Using the second approach, care must also be taken in analyzing patterns such as those indicated in the Western Electric rules. Since the sample size actually changes from run to run, such analyses are practically meaningless. A solution to this problem is to use a "standardized" control chart where all points are plotted using standard deviation units. This type of chart has a centerline at zero and upper and lower control limits at ±3, respectively, for 3σ control. The variable plotted on the chart is

$$Z_i = \frac{\hat{p}_i - p}{\sqrt{\dfrac{p(1-p)}{n_i}}} \tag{6.13}$$

where p (or $\overline{p}$) is the process nonconforming in the in-control state. The standardized chart for the same dataset as in Figures 6.6 and 6.7 is shown in Figure 6.8. Tests for patterns can be safely applied to this chart since the relative changes from one point to another are all expressed in the same units.

6.3.1.3. Operating Characteristic and Average Runlength

The operating characteristic (OC) curve of a control chart is a graph of the probability of incorrectly accepting the hypothesis of statistical control (i.e., a type II error) versus the fraction nonconforming. The OC provides a measure of the sensitivity of the chart to a given process shift. In the case of the p chart, the OC provides the graphical display of its ability to detect a shift from the nominal

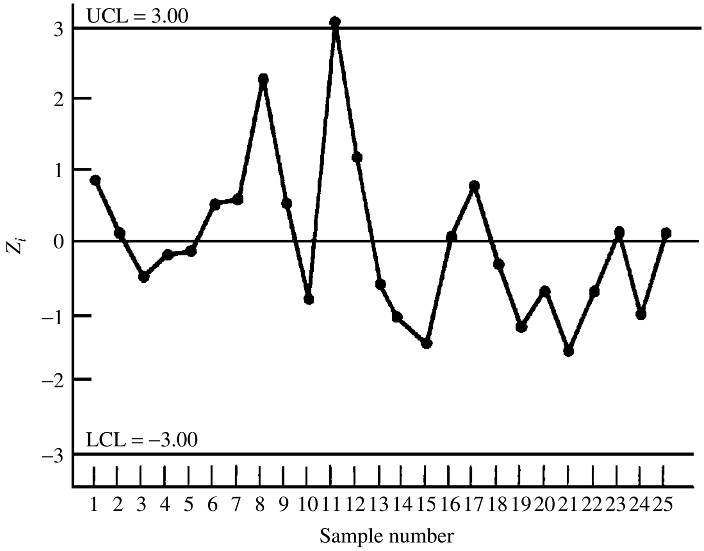


Figure 6.8. Standardized control chart for fraction nonconforming [1].

value of $\overline{p}$ to some new value. The probability of a type II error for this chart is given by

$$\begin{aligned} \beta &= P\{\hat{p} < \text{UCL}|p\} - P\{\hat{p} \leq \text{LCL}|p\} \qquad (6.14) \\ &= P\{D < n\text{UCL}|p\} - P\{D \leq n\text{LCL}|p\} \end{aligned}$$

where D is a binomial random variable with parameters n and p. The probability defined by Eq. (6.14) can be obtained from the cumulative binomial distribution. A typical OC curve for the fraction nonconforming chart is shown in Figure 6.9.

The OC curve may also be used to compute the average runlength (ARL) for the fraction nonconforming chart. Recall that the ARL is given by Eqs. (6.2) and (6.3). From the OC in Figure 6.9, for $p = 0.2$, the process is in control, and the probability that a point plots within the control limits is 0.9973. The in-control ARL is therefore

$$\text{ARL} = \frac{1}{\alpha} = \frac{1}{0.0027} = 370$$

This implies that if the process is in control, there will be a "false alarm" about every 370 samples. Suppose that the process shifts out of control to $p = 0.3$. From Figure 6.9, a value of $p = 0.3$ corresponds to $\beta = 0.8594$. The out-of-control ARL is then

$$\text{ARL} = \frac{1}{1-\beta} = \frac{1}{1-0.8594} = 7$$

This means that it will take seven samples, on average, for the p chart to detect this shift.

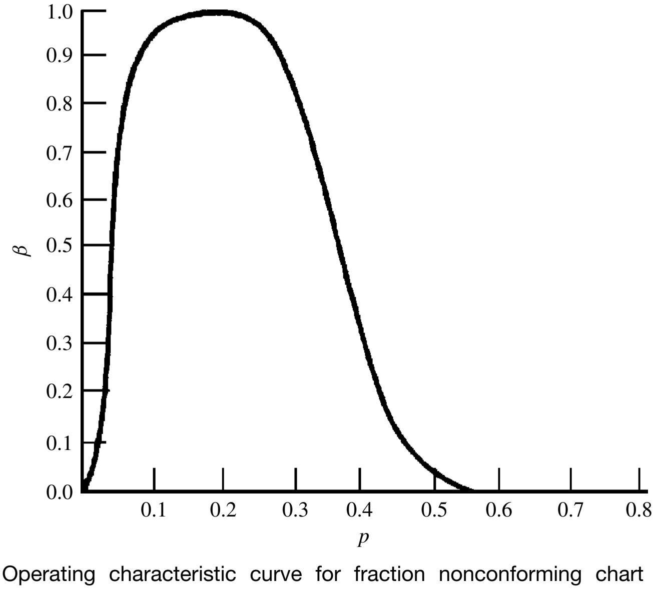


Figure 6.9. Operating characteristic curve for fraction nonconforming chart with $n = 50$, $\overline{p} = 0.2$, LCL = 0.0303, and UCL = 0.3697 [1].

6.3.2. Control Chart for Defects

When a specification is not satisfied in a product, a defect or nonconformity may result. In many cases, it is preferable to directly control the actual number of defects rather than the fraction nonconforming. In such cases, it is possible to develop control charts for either the total number of defects or the defect density. These charts assume that the presence of defects in samples of constant size is appropriately modeled by the Poisson distribution; that is

$$P(x) = \frac{e^{-c}c^x}{x!} \tag{6.15}$$

where x is the number of defects and $c > 0$ is the parameter of the Poisson distribution. Since c is both the mean and variance of the Poisson distribution, the control chart for defects (c chart) with 3σ limits is given by

$$\begin{aligned} \text{UCL} &= c + 3\sqrt{c} \\ \text{Centerline} &= c \\ \text{LCL} &= c - 3\sqrt{c} \end{aligned} \tag{6.16}$$

assuming that c is known. (*Note*: If these calculations yield a negative value for the LCL, the standard practice is to set the LCL $= 0$.) If c is not known, it may be estimated from an observed average number of defects in a sample ($\overline{c}$). In this case, the control chart becomes

$$\begin{aligned} \text{UCL} &= \overline{c} + 3\sqrt{\overline{c}} \\ \text{Centerline} &= \overline{c} \\ \text{LCL} &= \overline{c} - 3\sqrt{\overline{c}} \end{aligned} \tag{6.17}$$

Example 6.3. Suppose that the inspection of 26 silicon wafers yields 516 defects. Set up a c chart for this situation.

Solution: We estimate $\overline{c}$ using

$$\overline{c} = \frac{516}{26} = 19.85$$

This is the centerline for the c chart. The upper and lower control limits can be found from Eq. (6.17) as

$$\begin{aligned} \text{UCL} &= \overline{c} + 3\sqrt{\overline{c}} = 33.22 \\ \text{LCL} &= \overline{c} - 3\sqrt{\overline{c}} = 6.484 \end{aligned}$$

6.3.3. Control Chart for Defect Density

Suppose that we would like to set up a control chart for the *average* number of defects over a sample size of n products. If there were c total defects among the

n samples, then the average number of defects per sample is

$$u = \frac{c}{n} \tag{6.18}$$

The parameters of a 3σ defect density chart (u chart) are then given by

$$\begin{aligned} \text{UCL} &= \overline{u} + 3\sqrt{\frac{\overline{u}}{n}} \\ \text{Centerline} &= \overline{u} \\ \text{LCL} &= \overline{u} - 3\sqrt{\frac{\overline{u}}{n}} \end{aligned} \tag{6.19}$$

where $\overline{u}$ is the average number of defects over m groups of sample size n.

Example 6.4. Suppose that a manufacturer wants to establish a defect density chart. Twenty different samples of size $n = 5$ wafers are inspected, and a total of 193 defects are found. Set up the u chart for this situation.

Solution: We estimate u using

$$\overline{u} = \frac{u}{m} = \frac{c}{mn} = \frac{193}{(20)(5)} = 1.93$$

This is the centerline for the u chart. The upper and lower control limits can be found from Eq. (6.19) as

$$\text{UCL} = \overline{u} + 3\sqrt{\frac{\overline{u}}{n}} = 3.79$$

$$\text{LCL} = \overline{u} - 3\sqrt{\frac{\overline{u}}{n}} = 0.07$$

The operating characteristic (OC) curves for both the c and u charts are derived from the Poisson distribution. For the c chart, the OC represents the probability of type II error (β) as a function of the true mean number of defects. The expression for β is

$$\beta = P\{x < \text{UCL}|c\} - P\{x \le \text{LCL}|c\} \tag{6.20}$$

where x is a Poisson random variable with parameter c. A typical OC for a c chart is shown in Figure 6.10.

For the u chart, the OC is generated from

$$\begin{aligned} \beta &= P\{x < \text{UCL}|u\} - P\{x \le \text{LCL}|u\} \\ &= P\{c < n\text{UCL}|u\} - P\{c \le n\text{LCL}|u\} \\ &= P\{n\text{LCL} < c \le n\text{UCL}|u\} \\ &= \sum_{c=\langle n\text{LCL}\rangle}^{[n\text{UCL}]} \frac{e^{-nu}(nu)^c}{c!} \end{aligned} \tag{6.21}$$

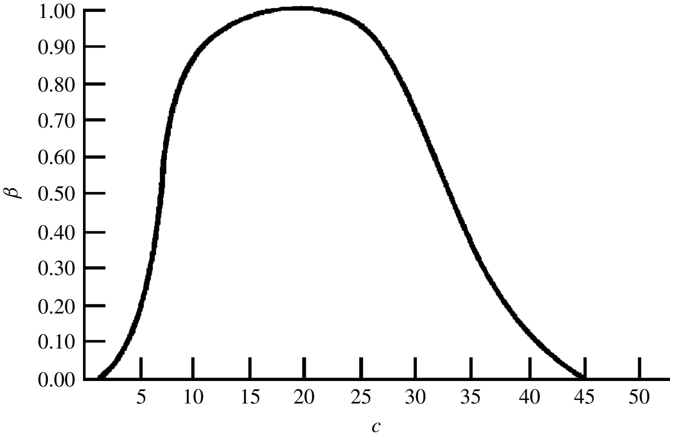


Figure 6.10. OC curve for fraction *c* chart with LCL = 6.48 and UCL = 33.22 [1].

where $\langle n\text{LCL}\rangle$ represents the smallest integer greater than or equal to nLCL and $[n\text{UCL}]$ is the largest integer less than or equal to nUCL. These summation limits occur because the total number of defects observed must be an integer.

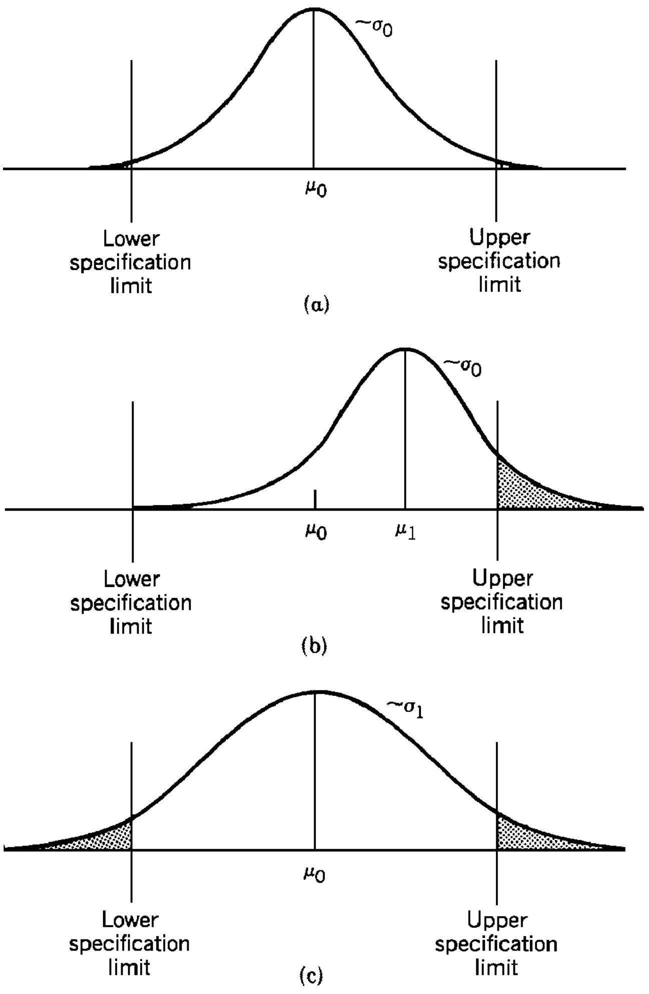
6.4. CONTROL CHARTS FOR VARIABLES

In many cases, quality characteristics are expressed as specific numerical measurements, rather than assessing the probability or presence of defects. For example, the thickness of an oxide layer is an important characteristic to be measured and controlled. Control charts for continuous variables such as this can provide more information regarding manufacturing process performance than attribute control charts like the p, c, and u charts.

When attempting to control continuous variables, it is important to control both the mean and the variance of the quality characteristic. This is true because shifts or drifts in either of these parameters can result in significant misprocessing. Consider a process represented by Figure 6.11. In Figure 6.11a, both the mean and the standard deviation are in control at their nominal values (μ_0 and σ_0). Under these conditions, most of the process output falls within the specification limits. However, in Figure 6.11b, the mean has shifted to a value $\mu_1 > \mu_0$, leading to a higher fraction of nonconforming product. Similarly, in Figure 6.11c, the standard deviation has shifted to a value $\sigma_1 > \sigma_0$, also resulting in more nonconforming products (even though the mean remains at its nominal value). Control of the mean is achieved using the $\overline{x}$ chart, and variance can be monitored using either the standard deviation (as in the s chart) or the range (as in the R chart). The x and R (or s) are among the most important and useful SPC tools.

6.4.1. Control Charts for $\overline{x}$ and R

We showed in Section 4.2.2 that if a quality characteristic is normally distributed with a known mean μ and standard deviation σ, then the sample mean ($\overline{x}$) for a



~σ0
μ0
Lower specification limit
Upper specification limit
(a)
~σ0
μ0
μ1
Lower specification limit
Upper specification limit
(b)
~σ1
μ0
Lower specification limit
Upper specification limit
(c)

Figure 6.11. Illustration of the need to control both process mean and standard deviation: (a) nominal mean and standard deviation; (b) mean shifted to $\mu_1 > \mu_0$; (c) standard deviation shifted to $\sigma_1 > \sigma_0$ [1].

sample of size n is also normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. Under these conditions, the probability that a sample mean will be between

$$\mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{6.22}$$

and

$$\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{6.23}$$

is $1-\alpha$. As a result, Eqs. (6.22) and (6.23) can be used as upper and lower control limits for a control chart for the sample mean. For 3σ control, we replace $z_{\alpha/2}$ by 3. This chart is called the $\overline{x}$ *chart.*

In practice, μ and σ rarely will be known. They must therefore be estimated from sample data. Suppose that m samples of size n are collected. If $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_m$ are the sample means, the best estimator for μ is the grand average ($\overline{\overline{x}}$), which is given by

$$\overline{\overline{x}} = \frac{\overline{x}_1 + \overline{x}_2 + \cdots + \overline{x}_m}{m} \tag{6.24}$$

Since $\overline{\overline{x}}$ estimates μ, $\overline{\overline{x}}$ is used as the centerline of the $\overline{x}$ chart.

To estimate σ, we can use the ranges of the m samples. The *range* (R) is defined as the difference between the maximum and minimum observation:

$$R = x_{\text{max}} - x_{\text{min}} \tag{6.25}$$

Another random variable $W = R/\sigma$ is called the *relative range*. The mean of W is a parameter called d_2, which is a function of the sample size n. (Values of d_2 for various sample sizes are given in Appendix F). Consequently, an estimator for σ is R/d_2. Let $R_1, R_2, \ldots, R_m$ be the ranges of the samples. The average range is then given by

$$\overline{R} = \frac{R_1 + R_2 + \cdots + R_m}{m} \tag{6.26}$$

and an estimate of σ is then

$$\hat{\sigma} = \frac{\overline{R}}{d_2} \tag{6.27}$$

If the sample size is small (i.e., $n < 10$), then the range is nearly as good an estimate of σ as the sample standard deviation (s).

If $\overline{\overline{x}}$ is used as an estimate of μ and $\overline{R}/d_2$ is used to estimate σ, then the parameters if the $\overline{x}$ chart are

$$\begin{aligned} \text{UCL} &= \overline{\overline{x}} + \frac{3\overline{R}}{d_2\sqrt{n}} \\ \text{Centerline} &= \overline{\overline{x}} \\ \text{LCL} &= \overline{\overline{x}} - \frac{3\overline{R}}{d_2\sqrt{n}} \end{aligned} \tag{6.28}$$

Note that the quantity $3/d_2\sqrt{n}$ is a constant that depends only on sample size. It is therefore possible to rewrite Eq. (6.28) as

$$\begin{aligned} \text{UCL} &= \overline{\overline{x}} + A_2\overline{R} \\ \text{Centerline} &= \overline{\overline{x}} \\ \text{LCL} &= \overline{\overline{x}} - A_2\overline{R} \end{aligned} \tag{6.29}$$

where the constant $A_2 = 3/d_2\sqrt{n}$ can be found tabulated for various sample sizes in Appendix F.

To control the range, the R chart is used. The centerline of the R chart is clearly $\overline{R}$, but to set up ±3σ control limits for the R chart, we must first derive an estimate of the standard deviation of R ($\hat{\sigma}_R$). To do so, we again use the relative range. The standard deviation of W is d_3, which is a known function of n (see Appendix F). Since $R = W\sigma$, the true standard deviation of R is

$$\sigma_R = d_3\sigma \tag{6.30}$$

Since σ is unknown, σ_R can be estimated from

$$\hat{\sigma}_R = d_3\frac{\overline{R}}{d_2} \tag{6.31}$$

Therefore, the parameters of the R chart assuming 3σ control limits are

$$\begin{aligned} \text{UCL} &= \overline{R} + 3d_3\frac{\overline{R}}{d_2} \\ \text{Centerline} &= \overline{R} \\ \text{LCL} &= \overline{R} - 3d_3\frac{\overline{R}}{d_2} \end{aligned} \tag{6.32}$$

If we let

$$D_3 = 1 - 3\frac{d_3}{d_2}$$

and

$$D_4 = 1 + 3\frac{d_3}{d_2}$$

then the parameters of the R chart may be defined as

$$\begin{aligned} \text{UCL} &= \overline{R}D_4 \\ \text{Centerline} &= \overline{R} \\ \text{LCL} &= \overline{R}D_3 \end{aligned} \tag{6.33}$$

The constants D_3 and D_4 may also be found in Appendix F.

Example 6.5. Suppose that we want to establish an $\overline{x}$ chart to control linewidth for a lithography process. Twenty-five different samples of size $n = 5$ linewidths are measured. Suppose that the grand average for the 125 total lines measured is 74.001 μm and the average range for the 25 samples is 0.023 μm. What are the control limits for the $\overline{x}$ chart?

Solution: The value for d_2 for $n = 5$ (found in Appendix F) is 2.326. The upper and lower control limits for the $\overline{x}$ chart can therefore be found from

Eq. (6.28) as

$$\mathrm{UCL} = \overline{\overline{x}} + \frac{3\overline{R}}{d_2\sqrt{n}} == 74.014\ \mu\mathrm{m}$$

$$\mathrm{LCL} = \overline{\overline{x}} - \frac{3\overline{R}}{d_2\sqrt{n}} == 73.988\ \mu\mathrm{m}$$

6.4.1.1. Rational Subgroups

A fundamental idea in the use of control charts is the collection of sample data according to the *rational subgroup* concept. In general, this means that subgroups (i.e., samples of size n) should be selected so that if assignable causes for misprocessing are present, the chance for differences *between* subgroups will be maximized, whereas the chance for differences *within* a subgroup will be minimized. In other words, only random variation should be allowed within a subgroup.

The rational subgroup concept plays a particularly important role in the use of $\overline{x}$ and R control charts. The $\overline{x}$ chart monitors the average level of quality in a process, and the R chart measures the variability *within* a sample. In other words, the $\overline{x}$ chart monitors *between-sample* variabilty (variability in the process over time), and the R chart measures *within-sample* variability (instantaneous process variability for a given sample at a given time).

In semiconductor manufacturing, intuitive categories for rational subgroups include devices within a die, die within a wafer, or wafers in a lot. The following inequality represents the expected level of variation in these groupings:

$$\begin{aligned}(\text{Within-die variation}) &< (\text{within-wafer variation})\\ &< (\text{within-lot variation}) < (\text{lot-to-lot variation})\end{aligned}$$

Care must be exercised when establishing such groupings. For example, grouping wafers within a quartz boat in a CVD furnace operation is inappropriate since reactant gas depletion effects down the length of the tube cause systematic variations in the deposition reaction [4].

Note from Eqs. (6.28) and (6.29) that the range is used to compute the control limits for the $\overline{x}$ chart. The range of a subgroup is used to estimate the standard deviation (σ) of that subgroup. This implies that the range across a lot, for example, should not be used to estimate the standard deviation between lots (i.e., the lot-to-lot variation); thus, *within-lot* statistics are different from *between-lot* statistics. The same is true for other rational subgroups. Since the within-lot variation is less than the between-lot variation, the wrong choice of the rational subgroup used to compute the range can bias the estimation and result in misleading interpretations of SPC data. Consider Figure 6.12, which shows two different $\overline{x}$ charts for monitoring linewidth in the same manufacturing process. In Figure 6.12a, the within-lot range has been used to compute the control limits, and the linewidth appears to be out of statistical control. However, when the between-lot range is used to compute the control limits (Figure 6.12b), there is apparently no problem.

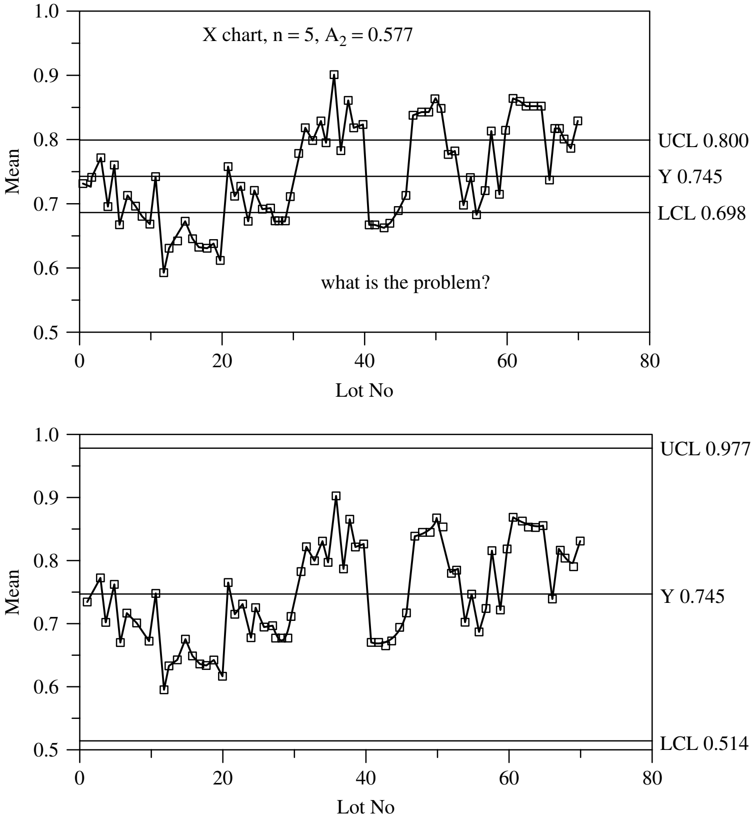


Figure 6.12. (a) $\overline{x}$ chart for linewidth control using within-lot range to compute control limits; (b) $\overline{x}$ chart for linewidth control using between-lot range to compute control limits.

6.4.1.2. Operating Characteristic and Average Runlength

Consider the operating characteristic (OC) curve for an $\overline{x}$ chart with a known standard deviation. If the process mean shifts from an in-control value (μ_0) to a new mean $\mu_1 = \mu_0 + k\sigma$, the probability of missing this shift on the next subsequent sample (i.e., the probability of type II error) is

$$\beta = P\{\text{LCL} \le \hat{x} \le \text{UCL} | \mu = \mu_0 + k\sigma\} \tag{6.34}$$

Since $\overline{x} \sim N(\mu, \sigma^2/n)$, and the control limits are $\text{UCL} = \mu_0 + 3\sigma/\sqrt{n}$ and $\text{LCL} = \mu_0 - 3\sigma/\sqrt{n}$, Eq. (6.34) can be rewritten as

$$\begin{aligned}\beta &= \Phi\left[\frac{\text{UCL} - (\mu_0 + k\sigma)}{\sigma/\sqrt{n}}\right] - \Phi\left[\frac{\text{LCL} - (\mu_0 + k\sigma)}{\sigma/\sqrt{n}}\right] \\ &= \Phi\left[\frac{\mu_0 + 3\sigma/\sqrt{n} - (\mu_0 + k\sigma)}{\sigma/\sqrt{n}}\right] - \Phi\left[\frac{\mu_0 - 3\sigma/\sqrt{n} - (\mu_0 + k\sigma)}{\sigma/\sqrt{n}}\right] \\ &= \Phi(3 - k\sqrt{n}) - \Phi(-3 - k\sqrt{n}) \end{aligned} \tag{6.35}$$

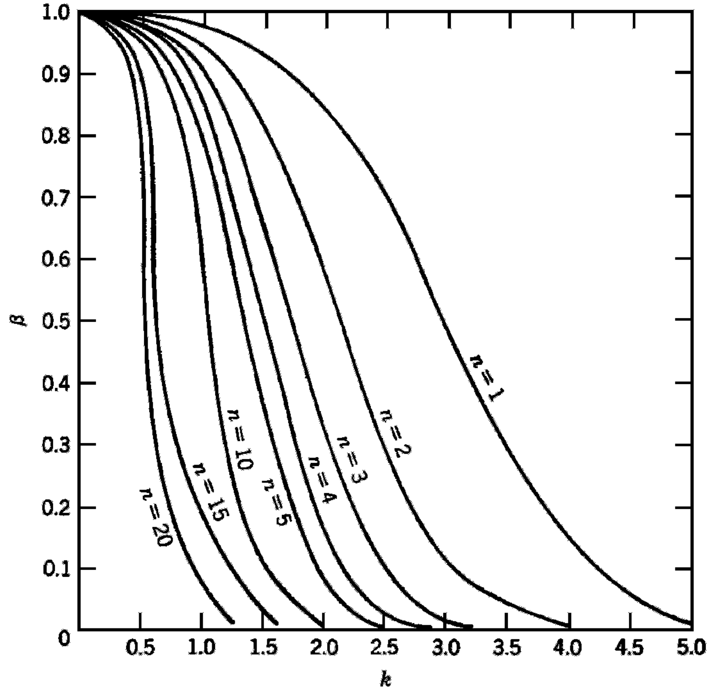


Figure 6.13. OC curve for $\bar{x}$-chart with 3-σ limits [1].

To construct the OC for the $\bar{x}$ chart, β is plotted versus k (the magnitude of the shift to be detected) for various sample sizes (see Figure 6.13). This figure shows that for small sample sizes ($n = 4$–6), the $\bar{x}$ chart is not particularly effective for detecting small shifts (i.e., shifts on the order of 1.5σ or less).

If the probability that a shift will be missed on the first sample after it occurs is β, then the probability that the shift will be detected in the first sample is $1 - \beta$. It then follows that the probability that the shift is detected on the second sample is $\beta(1 - \beta)$. Thus, the probability that a shift will be detected on the ith subsequent sample is

$$\beta^{i-1}(1 - \beta)$$

In general, the expected number of samples collected before the shift is detected is just the average runlength, so for the $\bar{x}$ chart, the ARL is

$$\text{ARL} = \sum_{i=1}^{\infty} i\beta^{i-1}(1 - \beta) = \frac{1}{1 - \beta} \tag{6.36}$$

This relationship suggests the advantage of using small sample sizes for the $\bar{x}$ chart. Even though small sample sizes result in a relatively high β, there is a good chance that a shift will be detected reasonably quickly in subsequent samples.

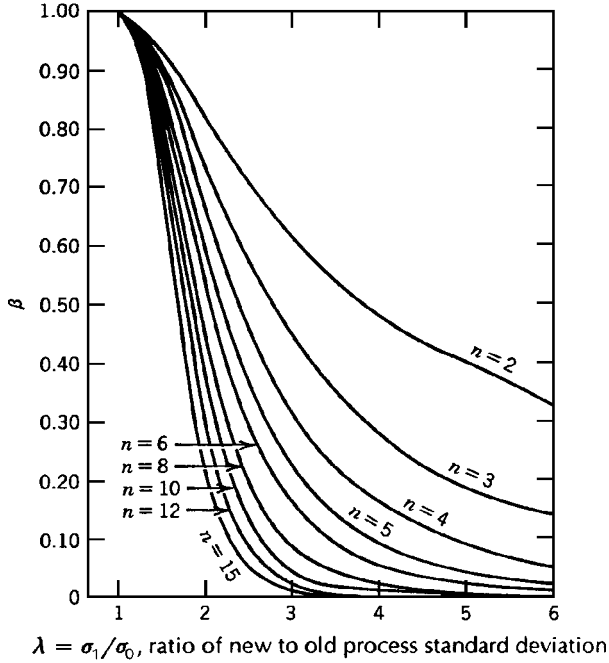


Figure 6.14. OC curve for R-chart with 3σ limits [1].

To construct the OC for the R chart, the distribution of the relative range ($W = R/\sigma$) is used. Let the in-control value of the standard deviation be σ_0. The OC curve for the R chart then plots the probability of not detecting a shift to a new value (σ_1). Figure 6.14 shows the OC curve for b versus $\lambda = \sigma_1/\sigma_0$ for various values of n.

6.4.2. Control Charts for $\overline{x}$ and s

Although the range chart is quite popular, when the sample size is large (i.e., $n > 10$), it is desirable to estimate and control the standard deviation directly. This leads to control charts for $\overline{x}$ and s, where s is the sample standard deviation, which is computed using Eq. (4.2). Setting up these charts is similar to setting up $\overline{x}$ and R charts, except that for each sample, s is calculated rather than R.

The only caution that must be applied in this situation is that s cannot be used directly as the centerline of the s chart. This is due to the fact that s is *not* an unbiased estimator of s. (The term "unbiased" refers to the situation where the expected value of estimator is equal to the parameter being estimated.) Instead s actually estimates $c_4\sigma$, where c_4 is a statistical parameter that is dependent on the sample size (see Appendix F). In addition, the standard deviation of s is $\sigma\sqrt{1 - c_4^2}$. Using this information the control limits for the s chart can be set up as follows:

$$\begin{aligned} \text{UCL} &= c_4\sigma + 3\sigma\sqrt{1 - c_4^2} \\ \text{Centerline} &= c_4\sigma \\ \text{LCL} &= c_4\sigma - 3\sigma\sqrt{1 - c_4^2} \end{aligned} \tag{6.37}$$

It is customary to define two constants

$$\begin{aligned} B_5 &= c_4 - 3\sqrt{1 - c_4^2} \\ B_6 &= c_4 + 3\sqrt{1 - c_4^2} \end{aligned} \tag{6.38}$$

As a result, the parameters of the s chart become

$$\begin{aligned} \text{UCL} &= B_6\sigma \\ \text{Centerline} &= c_4\sigma \\ \text{LCL} &= B_5\sigma \end{aligned} \tag{6.39}$$

If σ is unknown, then it must be estimated by analyzing past data. For m preliminary samples of size n, the average sample standard deviation is

$$\bar{s} = \frac{1}{m}\sum_{i=1}^{m} s_i \tag{6.40}$$

The statistic $\bar{s}/c_4$ is an unbiased estimator of σ. The parameters for the s chart then become

$$\begin{aligned} \text{UCL} &= \bar{s} + 3\frac{\bar{s}}{c_4}\sqrt{1 - c_4^2} \\ \text{Centerline} &= \bar{s} \\ \text{LCL} &= \bar{s} - 3\frac{\bar{s}}{c_4}\sqrt{1 - c_4^2} \end{aligned} \tag{6.41}$$

Once again, it is customary to define two constants:

$$\begin{aligned} B_3 &= 1 - \frac{3}{c_4}\sqrt{1 - c_4^2} \\ B_4 &= 1 + \frac{3}{c_4}\sqrt{1 - c_4^2} \end{aligned} \tag{6.42}$$

Consequently, the parameters of the s chart become

$$\begin{aligned} \text{UCL} &= B_4\bar{s} \\ \text{Centerline} &= \bar{s} \\ \text{LCL} &= B_3\bar{s} \end{aligned} \tag{6.43}$$

Note that $B_4 = B_6/c_4$ and $B_3 = B_5/c_4$.

When $\overline{s}/c_4$ is used to estimate σ, the limits on the corresponding $\overline{x}$ chart may be defined as

$$\begin{aligned} \text{UCL} &= \overline{\overline{x}} + \frac{3\overline{s}}{c_4\sqrt{n}} \\ \text{Centerline} &= \overline{\overline{x}} \\ \text{LCL} &= \overline{\overline{x}} - \frac{3\overline{s}}{c_4\sqrt{n}} \end{aligned} \tag{6.44}$$

Let the constant $A_3 = 3/c_4\sqrt{n}$. It is therefore possible to rewrite Eq. (6.44) as

$$\begin{aligned} \text{UCL} &= \overline{\overline{x}} + A_3\overline{s} \\ \text{Centerline} &= \overline{\overline{x}} \\ \text{LCL} &= \overline{\overline{x}} - A_3\overline{s} \end{aligned} \tag{6.45}$$

Example 6.6. Consider the lithography process in Example 6.5. If $s =$ 0.009 mm, what are the control limits for the s chart?

Solution: The value for c_4 for $n = 5$ (found in Appendix F) is 0.94. The upper and lower control limits can therefore be found from Eq. (6.41) as

$$\begin{aligned} \text{UCL} &= \overline{s} + 3\frac{\overline{s}}{c_4}\sqrt{1 - c_4^2} = 0.019\ \mu\text{m} \\ \text{LCL} &= \overline{s} - 3\frac{\overline{s}}{c_4}\sqrt{1 - c_4^2} = 0\ \mu\text{m}^{1} \end{aligned}$$

6.4.3. Process Capability

Process capability quantifies what a process can accomplish when in control. Shewhart control charts are useful for estimating process capability. For example, suppose that the interconnect being defined by the lithography process described in Examples 6.5 and 6.6 must have a linewidth of 74.000 ± 0.05 μm. If these tolerances are not met, then some loss in product quality results. Tolerances such as this are called *specification limits*. Specification limits (SLs) differ from control limits in that they are externally imposed on the manufacturing process, whereas control limits are derived from the natural variability inherent in the process.

Control chart data can be used to investigate the capability of the process to produce linewidths according to the specification limits. Recall that our estimates for the process mean and standard deviation were

$$\begin{aligned} \overline{x} &= 74.001\ \mu\text{m} \\ \hat{\sigma} &= \frac{R}{d_2} = 0.0099\ \mu\text{m} \end{aligned}$$

[1] Since the LCL is actually (slightly) negative in this case, we automatically set it to zero.

Assuming that the linewidth is normally distributed, we can estimate the fraction of nonconforming lines as

$$\begin{aligned}\hat{p} &= P\{x < 73.95\} + P\{x > 74.05\}\\ &= \Phi\left(\frac{73.95 - 74.001}{0.0099}\right) + 1 - \Phi\left(\frac{74.05 - 74.001}{0.0099}\right) \cong 0.00002\end{aligned}$$

In other words, about 0.002% of the lines produced will be outside the specification limits. This means that the process is capable of achieving the specification limits 99.998% of the time. The remaining 0.002% of the lines will not meet the specifications no matter what steps are taken to improve the process.

Another way to express the process capability is in terms of the process capability ratio (PCR, or C_p). The PCR is defined as

$$C_p = \text{PCR} = \frac{\text{USL} - \text{LSL}}{6\sigma} \tag{6.46}$$

where USL and LSL are the upper and lower specification limits, respectively. Since σ is usually unknown, it frequently replaced by $\hat{\sigma} = R/d_2$. For the interconnect linewidth process, we can compute the PCR as

$$C_p = \text{PCR} = \frac{\text{USL} - \text{LSL}}{6\sigma} = \frac{74.05 - 73.95}{6(0.0099)} = 1.68$$

A PCR $>$ 1 implies that the "natural" tolerance limits (NTLs) inherent in the process (as quantified by the $\pm 3\sigma$ control limits) are well inside the specification limits. This results in a relatively low number of nonconforming lines being produced. A common variation of the C_p parameter is C_{pk}, where

$$C_{pk} = \min\left\{\left(\frac{\text{USL} - \mu}{3\sigma}\right), \left(\frac{\mu - \text{LSL}}{3\sigma}\right)\right\} \tag{6.47}$$

The C_{pk} parameter is a measure of the capability of the process to achieve control chart values that lie in the center of the specification range. This metric is useful when the specification limits are not symmetric about the centerline.

The PCR can also be interpreted using the quantity

$$P = \left(\frac{1}{\text{PCR}}\right) \times 100\% \tag{6.48}$$

This is just the percentage of the specification band that the process under consideration "uses up." For the interconnect linewidth example, we compute $P = 59.5\%$, which means that this process uses 59.5% of the specification band. Figure 6.15 illustrates the relationship between the PCR and the specification limits. In Figure 6.15a, the PCR is greater than one, which means that the process uses up much less that 100% of the tolerance band. In this case, few nonconforming products are produced. In Figure 6.15b, PCR = 1, which means that the process uses up all of the tolerance band. Finally, in Figure 6.15c, PCR $<$ 1, and

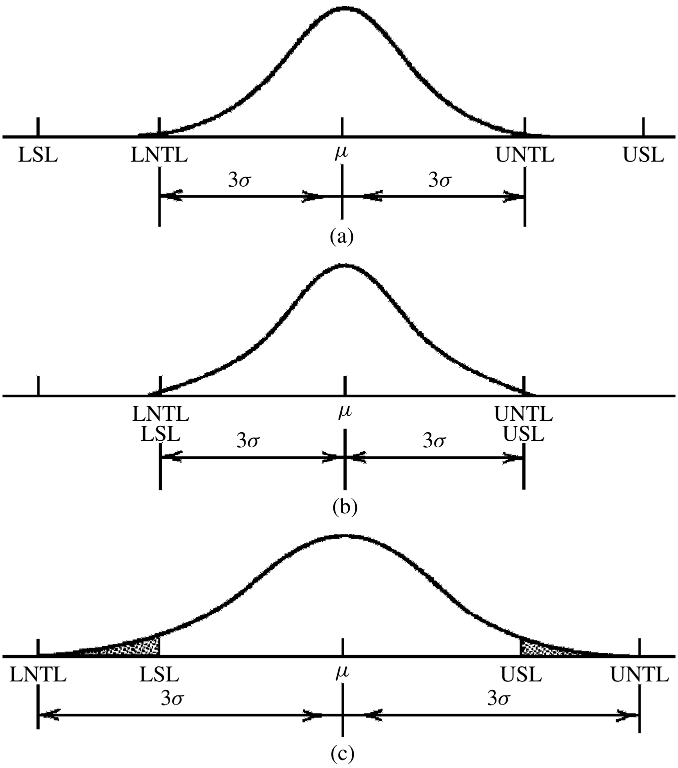


Figure 6.15. Illustration of relationship between specification limits, natural tolerance limits, and process capability ratio [1].

the process uses more than 100% of the tolerance band. In the latter case, a large number of nonconforming products will be produced.

6.4.4. Modified and Acceptance Charts

When $\bar{x}$ charts are used to control the fraction of nonconforming products, two important variations to standard SPC charts can be employed: the *modified* chart and the *acceptance* chart. Modified control limits are generally used when the natural tolerance limits of the process are smaller than the specification limits (i.e., PCR $>$ 1). This occurs frequently in practice, particularly when a quality improvement program exists. In these situations, the modified control chart is designed to detect whether the true process mean (μ) is located such that the process yields a fraction nonconforming in excess of some specified value δ. Essentially, μ is allowed to vary over an interval $\mu_L \leq \mu \leq \mu_U$, where μ_L and μ_U represent lower and upper bounds on μ, respectively, that are consistent with producing a fraction nonconforming of at most δ. This scenario is represented graphically in Figure 6.16.

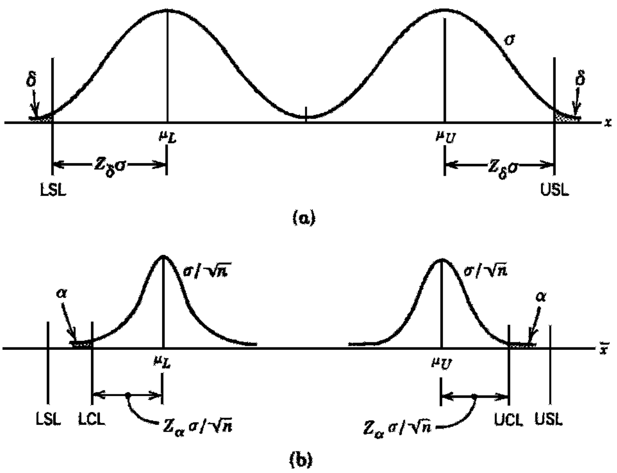


Figure 6.16. Control limits for modified control char: (a) distribution of process output; (b) distribution of the sample mean $\bar{x}$ [1].

To specify control limits for the modified chart (assuming a normally distributed process), for the fraction nonconforming to be less than δ, we must have

$$\mu_L = \mathrm{LSL} + Z_\delta \sigma$$
$$\mu_U = \mathrm{USL} - Z_\delta \sigma \tag{6.49}$$

where Z_δ is the upper 100(1-δ) percentage point of the standard normal distribution. If the specified probability of type I error is α, the upper and lower control limits are then

$$\mathrm{UCL} = \mu_U + \frac{Z_\alpha \sigma}{\sqrt{n}} = \mathrm{USL} - \left(Z_\delta - \frac{Z_\alpha}{\sqrt{n}} \right) \sigma$$
$$\mathrm{LCL} = \mu_L - \frac{Z_\alpha \sigma}{\sqrt{n}} = \mathrm{LSL} + \left(Z_\delta - \frac{Z_\alpha}{\sqrt{n}} \right) \sigma \tag{6.50}$$

Note that using the modified chart is equivalent to testing the hypothesis that the process mean lies in the interval $\mu_L \leq \mu \leq \mu_U$.

Another approach to using the $\bar{x}$ chart to control the fraction nonconforming accounts for both the risk of rejecting a process operating at a satisfactory level (probability of type I error, or α) and the risk of accepting a process that is unsatisfactory (probability of type II error, or β). This second approach is called the *acceptance chart*. The design of this chart is based on a specified sample size (n) and a process fraction nonconforming (γ) that should be rejected with probability 1 − β. In this case, the control limits for the acceptance chart are

$$\mathrm{UCL} = \mu_U - \frac{Z_\beta \sigma}{\sqrt{n}} = \mathrm{USL} - \left(Z_\gamma + \frac{Z_\beta}{\sqrt{n}} \right) \sigma$$
$$\mathrm{LCL} = \mu_L + \frac{Z_\beta \sigma}{\sqrt{n}} = \mathrm{LSL} + \left(Z_\gamma + \frac{Z_\beta}{\sqrt{n}} \right) \sigma \tag{6.51}$$

Note that when n, γ, and β are specified, the control limits are inside the μ_L and μ_U values that yield the fraction nonconforming γ. On the other hand, when n, δ, and α are specified in the modified chart, the lower control limit falls between μ_L and the LSL, and the upper control limit is between μ_U and the USL.

It is also possible to select a sample size for an acceptance chart such that desired values of α, β, γ, and δ are obtained. Equating the expressions for the control limits in Eqs. (6.50) and (6.51) yields

$$n = \left(\frac{Z_\alpha + Z_\beta}{Z_\delta - Z_\gamma}\right)^2 \tag{6.52}$$

Clearly, values of $\delta = \gamma$ are prohibited to achieve a finite sample size.

6.4.5. Cusum Chart

Consider the Shewhart control chart shown in Figure 6.17. This chart corresponds to a normally distributed process with a mean $\mu = 10$ and a standard deviation $\sigma = 1$. Note that all of the first 20 observations appear to be under statistical control. The last 10 observations in this chart were drawn from the same process after the mean has shifted to a new value $\mu = 11$. We can think of these latter observations as having been taken from the process after the mean has shifted out of statistical control by an amount 1σ. However, none of the last 10 points plots outside the control limits, so there is no strong evidence that the process is truly out of control. Even applying the *Western Electric rules*, the Shewhart chart has failed to detect the mean shift.

The reason for this failure is the relatively small magnitude of the shift. Shewhart charts are generally effective for detecting shifts on the order of $1.5–2\sigma$ or larger. For smaller shifts the "cumulative sum" (or *cusum*) control chart is preferred. The cusum chart incorporates historical information from a sequence of

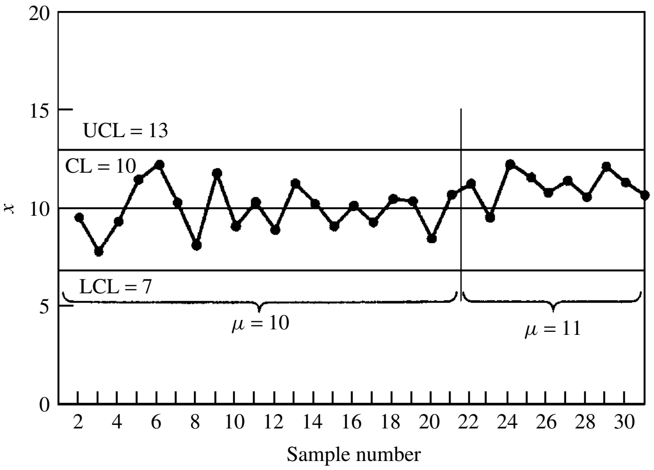


Figure 6.17. Shewhart chart for before and after a mean shift from $\mu = 10$ to $\mu = 11$ [1].

samples by plotting the cumulative sums of the sample deviations from a target value. If samples of size $n \geq 1$ are collected and μ_0 is the target for the process mean, the cusum chart is formed by plotting the quantity

$$C_i = \sum_{j=1}^{i} (\overline{x}_j - \mu_0) \tag{6.53}$$

versus sample i, where $\overline{x}_j$ is the average of the jth sample. Because they combine information from several samples, cusum charts are sensitive to smaller process shifts than are Shewhart charts.

If the process remains in control at the target value μ_0, the sum defined by Eq. (6.53) is a random variable with mean zero. If the mean shifts upward to some value $\mu_1 > \mu_0$, then a positive drift will develop in the cusum chart. Conversely, if the mean shifts downward to some value $\mu_1 < \mu_0$, then a negative drift will be manifested in C_i. This effect is demonstrated in Figure 6.18, which depicts the cusum chart for the same dataset used in Figure 6.17. The upward trend after the first 20 samples is indicative of the mean shift to $\mu = 11$ described previously. This figure, however, does not represent a control chart because it lacks control limits. The methodology for establishing such limits is described in the following subsections.

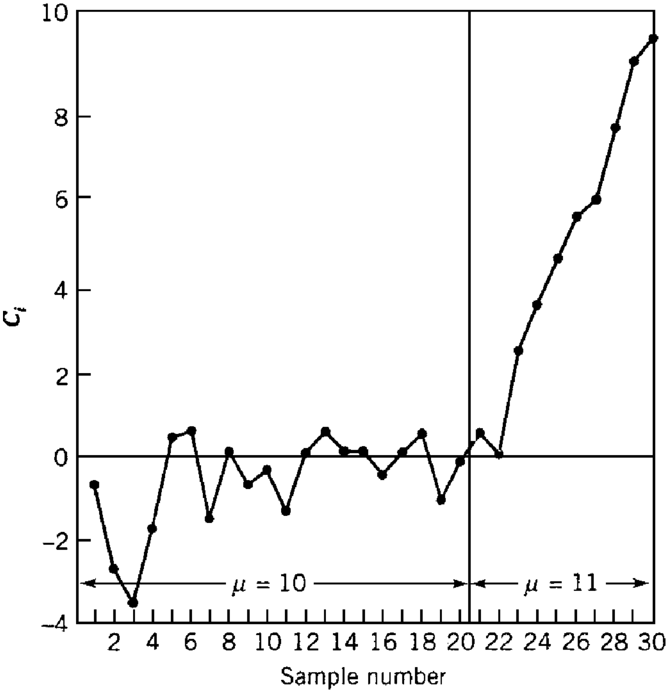


Figure 6.18. Cusum chart for before and after a mean shift from $\mu = 10$ to $\mu = 11$ [1].

6.4.5.1. Tabular Cusum Chart

The tabular form of the cusum chart may be constructed for both individual observations and for averages of rational subgroups. Let x_i be the ith observation of a normally distributed process with mean μ_0 and standard deviation σ. We can think of μ_0 as a "target" value for quality characteristic x. If the process shifts or drifts from this target value, the cusum chart should generate an alarm signal.

The tabular cusum accumulates deviations from μ_0 that are above the target with a statistic C^+ and deviations that are below the target with another statistic C^-. The quantities C^+ and C^- are called the *upper and lower cusums*, respectively. The are computed using the relations

$$C_i^+ = \max\left[0, x_i - (\mu_0 + K) + C_{i-1}^+\right] \tag{6.54}$$

$$C_i^- = \max\left[0, (\mu_0 + K) - x_i + C_{i-1}^-\right] \tag{6.55}$$

where the starting values are $C_0^+ = C_0^- = 0$. In these equations, K is called the *reference value*, and it is usually chosen to be about halfway between the target mean (μ_0) and the shifted mean that we are interested in detecting (μ_1). If the shift is expressed in terms of the standard deviation as $\mu_1 = \mu_0 + \delta\sigma$, the K is given by

$$K = \frac{\delta}{2}\sigma = \frac{|\mu_1 - \mu_0|}{2} \tag{6.56}$$

Both C^+ and C^- accumulate deviations from μ_0 that are greater than K, and both quantities reset to zero on becoming negative. If either quantity exceeds the *decision interval* (H), the process is considered to be out of control. A reasonable value for H is $H = 5\sigma$.

The tabular cusum is particularly useful for determining when a shift has occurred. This can be accomplished by simply counting backward from the out-of-control signal to the time period when the cusum was greater than zero to identify the first period following the shift. To assist in this process, we can define the counters N^+ and N^-, where N^+ represents the number of consecutive periods since C_i^+ rose above zero, and N^- is the number of consecutive periods since C_i^- rose above zero. These quantities may also be used to estimate the new process mean following a shift. This can be computed from

$$\begin{aligned} \hat{\mu} &= \mu_0 + K + \frac{C_i^+}{N^+} \quad \text{if } C_i^+ > H \\ &= \mu_0 - K - \frac{C_i^-}{N^-}, \quad \text{if } C_i^- > H \end{aligned} \tag{6.57}$$

6.4.5.2. Average Runlength

The format of the tabular cusum depends on the values selected for the reference value (K) and decision interval (H). These parameters are usually selected to provide a certain average runlength. Let $H = h\sigma$ and $K = k\sigma$. Choosing $h = 4$–5 and $k = 0.5$ generally results in a cusum that has a reasonable ARL. Table 6.1

Table 6.1. ARL performance of tabular cusum with $k = 0.5$ and $h = 4$ or $h = 5$.

Shift in Mean (Multiple of σ)	$h = 4$	$h = 5$
0	168	465
0.25	74.2	139
0.50	26.6	38
0.75	13.3	17
1.00	8.38	10.4
1.50	4.75	5.75
2.00	3.34	4.01
2.50	2.62	3.11
3.00	2.19	2.57
4.00	1.71	2.01

provides the ARL performance of the tabular cusum for various shifts in the process mean under these conditions.

Generally, k should be chosen relative to the size of the shift to be detected. In other words, $k = 0.5\delta$, where δ is the size of the shift to be detected (in standard deviation units). This approach comes very close to minimizing the out-of-control ARL value for detecting a shift of size δ for a fixed in-control ARL. Once k is chosen, h is then selected to give the desired in-control ARL.

For a *one-sided cusum* (i.e., for either C_i^+ or C_i^-), the ARL may generally be approximated as [1]

$$\text{ARL} = \frac{\exp(-2\Delta b) + 2\Delta b - 1}{2\Delta^2} \tag{6.58}$$

for $\Delta \neq 0$, where $\Delta = \delta^* - k$, $b = h + 1.166$, and

$$\delta^* = \frac{\mu_1 - \mu_0}{\sigma} \tag{6.59}$$

where μ_0 and μ_1 are the target and shifted mean, respectively. If $\Delta = 0$, then the ARL $= b^2$. The quantity δ^* represents the shift in the mean (in standard deviation units) for which the ARL is calculated. The ARL of the two-sided cusum can be derived from the ARLs of the two one-sided statistics (ARL$^+$ and ARL$^-$) as

$$\frac{1}{\text{ARL}} = \frac{1}{\text{ARL}^+} + \frac{1}{\text{ARL}^-} \tag{6.60}$$

6.4.5.3. Cusum for Variance

It is also possible to use the cusum technique to monitor process variability. Again, let x_i be a normally distributed process measurement with mean μ_0 and standard deviation σ. Further, let y_i be the standardized value of x_i, or

$$y_i = \frac{x_i - \mu_0}{\sigma} \tag{6.61}$$

Hawkins [5] suggests creating a new standard quantity, v_i, which is sensitive to both mean and variance changes. This parameter is given by

$$v_i = \frac{\sqrt{|y_i|} - 0.822}{0.349} \tag{6.62}$$

Since the in-control distribution of v_i is approximately $N(0,1)$, the two-sided cusums can be written as

$$S_i^+ = \max[0, v_i - k + S_{i-1}^+] \tag{6.63}$$

$$S_i^- = \max[0, -k - v_i + S_{i-1}^-] \tag{6.64}$$

where $S_0^+ = S_0^- = 0$, and the values of k and h are selected in the same way as the values for controlling the process mean. The interpretation of this cusum is also the same as that of the cusum for controlling the mean. If the process standard deviation increases, the values of S_i^+ will increase and eventually exceed h, and if the standard deviation decreases, the values of S_i^- will increase and eventually exceed h.

6.4.6. Moving-Average Charts

6.4.6.1. Basic Moving-Average Chart

Suppose that $x_1, x_2, \ldots, x_n$ individual observations of a process have been collected. The moving average of span w at time i is defined as

$$M_i = \frac{x_i + x_{i-1} + \cdots + x_{i-w+1}}{w} \tag{6.65}$$

At time period i, the oldest observation is dropped and the newest one is added to the set. The variance of the moving average is

$$V(M_i) = \frac{1}{w^2} \sum_{j=i-w+1}^{i} V(x_j) = \frac{1}{w^2} \sum_{j=i-w+1}^{i} \sigma^2 = \frac{\sigma^2}{w} \tag{6.66}$$

Thus, if μ_0 is the target mean used as the centerline of the moving-average control chart, the 3σ control limits for M_i are

$$\text{UCL} = \mu_0 + \frac{3\sigma}{\sqrt{w}} \tag{6.67}$$

$$\text{LCL} = \mu_0 - \frac{3\sigma}{\sqrt{w}} \tag{6.68}$$

The control procedure then consists of calculating a new value for M_i as each new observation becomes available and plotting M_i on a control chart with limits given by Eqs. (6.67) and (6.68). Note that for samples in which $i < w$, i replaces w in these equations. This causes the control limits for the first few samples to

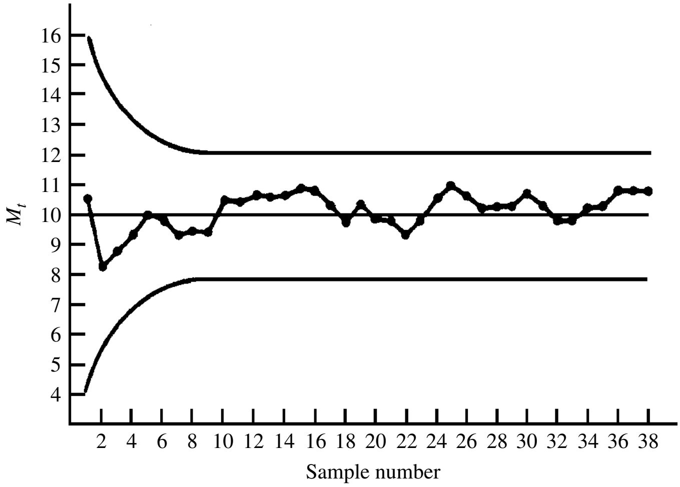


Figure 6.19. Control limits for moving-average chart for a sample dataset [1].

become variable, as is depicted in Figure 6.19. In general, the moving-average control chart is more sensitive than Shewhart charts for detecting small process shifts. However, it is not as effective in that regard as either the cusum or the EWMA (see discussion below).

6.4.6.2. Exponentially Weighted Moving-Average Chart

The *exponentially weighted moving-average* (EWMA) *control chart*, sometimes referred to as the *geometric moving average* (GMA) *chart*, is another alternative to Shewhart charts when it is desirable to detect small process shifts. The performance of the EWMA chart is comparable to that of the cusum chart. The exponentially weighted moving average is defined as

$$z_i = \lambda x_i + (1 - \lambda) z_{i-1} \tag{6.69}$$

where $0 < \lambda \leq 1$ is a constant and the starting value is the process target (i.e., $z_0 = \mu_0$).

To show that the parameter z_i is a weighted average of all previous sample means, we can substitute z_{i-1} on the right side of Eq. (6.69) to obtain

$$\begin{aligned} z_i &= \lambda x_i + (1 - \lambda)[\lambda x_{i-1} + (1 - \lambda) z_{i-2}] \\ &= \lambda x_i + \lambda(1 - \lambda) x_{i-1} + (1 - \lambda)^2 z_{i-2} \end{aligned}$$

If we continue to substitute recursively for z_{i-j} for $j = 2, 3, \ldots, t$, we obtain

$$z_i = \lambda \sum_{j=0}^{i-1} (1 - \lambda)^j x_{i-j} + (1 - \lambda)^i z_0 \tag{6.70}$$

The weights $\lambda(1-\lambda)^j$ thus decrease geometrically with the age of the sample mean. Since the EWMA is a weighted average of all previous observations, it is insensitive to the assumption of normality and can therefore be used for individual process measurements.

If the observations (x_i) are random variables with variance σ^2, then the variance of z_i is

$$\sigma_{zi}^2 = \sigma^2\left(\frac{\lambda}{2-\lambda}\right)[1-(1-\lambda)^{2i}] \qquad (6.71)$$

The EWMA control chart can then be constructed by plotting z_i versus i (or time). The centerline and 3σ control limits for this chart are

$$\begin{aligned} \text{UCL} &= \mu_0 + 3\sigma\sqrt{\frac{\lambda}{(2-\lambda)}[1-(1-\lambda)^{2i}]} \\ \text{CL} &= \mu_0 \\ \text{LCL} &= \mu_0 - 3\sigma\sqrt{\frac{\lambda}{(2-\lambda)}[1-(1-\lambda)^{2i}]} \end{aligned} \qquad (6.72)$$

Notice that the term $[1-(1-\lambda)^{2i}]$ in these equations approaches unity as i gets larger. The control limits therefore reach steady-state values of

$$\begin{aligned} \text{UCL} &= \mu_0 + 3\sigma\sqrt{\frac{\lambda}{(2-\lambda)}} \\ \text{CL} &= \mu_0 \\ \text{LCL} &= \mu_0 - 3\sigma\sqrt{\frac{\lambda}{(2-\lambda)}} \end{aligned} \qquad (6.73)$$

This variation in control limits with i is depicted in Figure 6.20.

The EWMA method is related to the *proportional–integral–differential* (PID) approach often used in classical control problems. Note that the EWMA parameter z_i in Eq. (6.69) can be manipulated algebraically and rewritten as

$$z_i = z_{i-1} + \lambda(x_i - z_{i-1}) \qquad (6.74)$$

If z_{i-1} is viewed as a forecast of the process mean in sample period i, then we can think of $x_i - z_{i-1}$ as the forecast error (e_i) for period i, or

$$z_i = z_{i-1} + \lambda e_i \qquad (6.75)$$

In other words, the forecast for period i is the forecast from the previous period plus a fraction of the forecast error. The second term in Eq. (6.75) is therefore known as the *proportional* term.

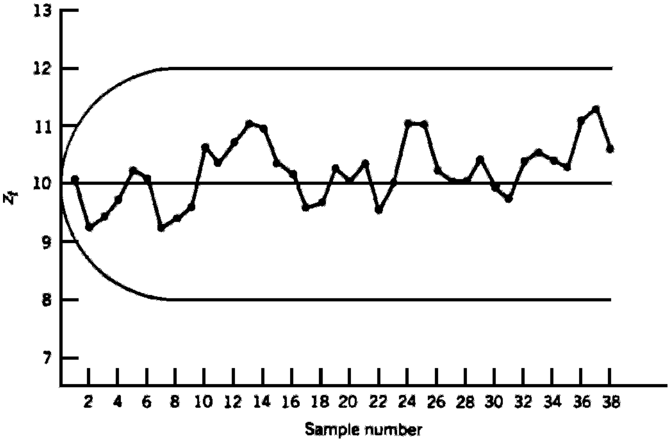


Figure 6.20. Control limits for EWMA chart with λ = 0.2 for a sample dataset [1].

We can add a second *integral* term to Eq. (6.75) to get

$$z_i = z_{i-1} + \lambda_1 e_i + \lambda_2 \sum_{j=1}^{i} e_j \tag{6.76}$$

where λ_1 and λ_2 are coefficients that weight the error at period i and the sum of the errors accumulated up to period i, respectively. If we let $\nabla e = e_i - e_{i-1}$ represent the difference between the errors in periods i and $i - 1$, then can add a third *differential* term to Eq. (6.76) to yield

$$z_i = z_{i-1} + \lambda_1 e_i + \lambda_2 \sum_{j=1}^{i} e_j + \lambda_3 \nabla e_i \tag{6.77}$$

In summary, the empirical control equation represented by Eq. (6.77) states that the EWMA in period i (which is a forecast of the process mean in period $i + 1$) is the sum of the current estimate of the mean (z_{i-1}), a term proportional to the forecast error, a term related to the sum of the forecast errors, and a term related to the difference between the two most recent forecast errors. The latter three terms can be thought of as *proportional, integral*, and *differential* adjustments, and the parameters λ_1, λ_2, and λ_3 are selected to provide the best forecasting performance.

6.5. MULTIVARIATE CONTROL

In many situations, it is desirable to control two or more quality characteristics simultaneously. For example, we may be interested in controlling the linewidth of a test structure on two different product wafers. Suppose that these two characteristics are represented by the random variables x_1 and x_2, which have a bivariate

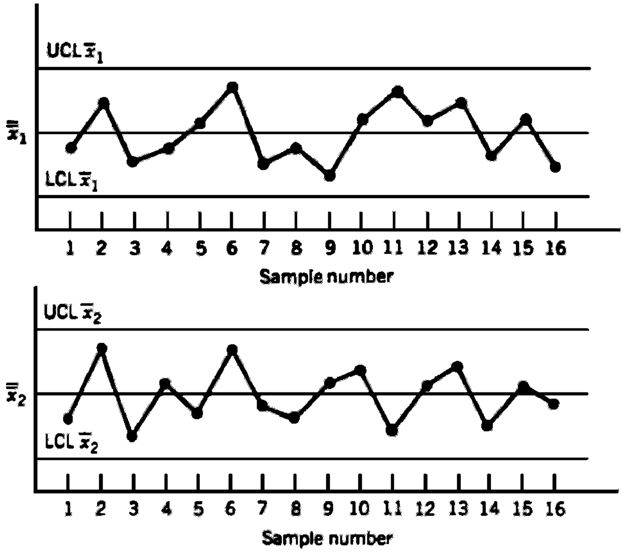


Figure 6.21. $\bar{x}$ control charts for bivariate normal process variables [1].

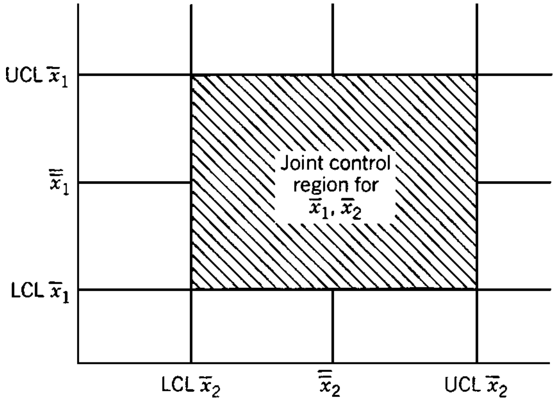


Figure 6.22. Control region using independent control limits for $\bar{x}_1$ and $\bar{x}_2$ [1].

normal distribution. Suppose also that each variable is controlled by an $\bar{x}$ chart (see Figure 6.21). Since the process is under control only if both sample means ($\bar{x}_1$ and $\bar{x}_2$) fall within their respective control limits, the joint control region for both variables is as shown in Figure 6.22.

Controlling these two process variables in this manner can be misleading. Since the probability that *either* $\bar{x}_1$ or $\bar{x}_2$ exceeds its control limits when in control is 0.0027, the joint probability that *both* variables simultaneously exceed their control limits when both are in fact in control is $(0.0027)^2 = 0.00000729$, which is significantly less than 0.0027. Moreover, the probability that both variables will plot inside the control limits when under control is $(0.9973)^2 = 0.99460729$. The use of independent $\overline{x}$ charts to control both variables simultaneously thus distorts the probability of type I error as compared to individual control charts.

Such distortion increases with the number of quality characteristics. If there are p independent quality characteristics, each with $P\{\text{type I error}\} = \alpha$, then the overall probability for type I error is

$$\alpha' = 1 - (1 - \alpha)^p \tag{6.78}$$

and the probability that all p process means will simultaneously plot inside their control limits when the process is in control is

$$P\{\text{all means plot in control}\} = (1 - \alpha)^p \tag{6.79}$$

In addition, if the p quality characteristics are not all mutually independent (which would be the case if they were related to the same product), then Eqs. (6.78) and (6.79) would not be valid, and there would be no easy way to measure the distortion in this control procedure.

6.5.1. Control of Means

Let μ_1 and μ_2 represent the mean values of x_1 and x_2, and let σ_1 and σ_2 be their respective standard deviations. The covariance between x_1 and x_2, a measure of their dependence, is denoted by σ_{12}. If $\overline{x}_1$ and $\overline{x}_2$ are the sample averages computed from a sample of size n, then the statistic

$$\chi_0^2 = \frac{n}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}\left[\sigma_2^2(\overline{x}_1 - \mu_1)^2 + \sigma_1^2(\overline{x}_2 - \mu_2)^2 - 2\sigma_{12}(\overline{x}_1 - \mu_1)(\overline{x}_2 - \mu_2)\right] \tag{6.80}$$

has a chi-square distribution with 2 degrees of freedom. This equation can be used to develop a control chart for the process means. If the means remain under control (i.e., have not shifted), then $\chi_0^2 < \chi_{\alpha,2}^2$, where $\chi_{\alpha,2}^2$ is the upper percentage point of the chi-square distribution with 2 degrees of freedom. If, on the other hand, one of the means shifts to an out-of-control value, then $\chi_0^2 > \chi_{\alpha,2}^2$.

This control procedure can be represented graphically, as shown in Figures 6.23 and 6.24. Figure 6.23 depicts the case where x_1 and x_2 are independent ($\sigma_{12} = 0$), and the principal axes of the "control ellipse" are parallel to the $\overline{x}_1$ and $\overline{x}_2$ axes. Figure 6.24 shows the control ellipse when $\sigma_{12} \neq 0$. In both cases, sample averages yielding χ_0^2 points plotting inside the ellipse are indicative of statistical control, and points plotting outside represent out-of-control conditions.

There are two primary disadvantages of the control ellipse approach. The first is that the time sequence of the sample measurements is completely lost. The second shortcoming is the difficulty in graphically depicting the control region for more than two variables. To avoid these difficulties, it is customary to plot the values of χ_0^2 computed from Eq. (6.80) on a control chart with only an upper control limit at $\chi_{\alpha,2}^2$ (see Figure 6.25).

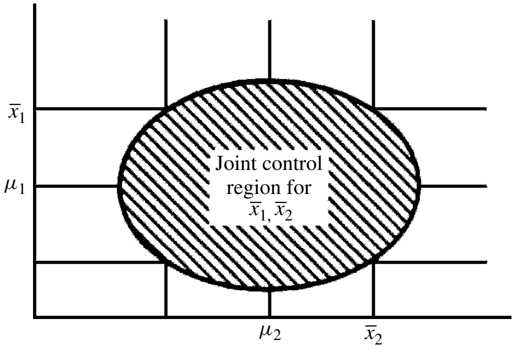


Figure 6.23. Control ellipse for two independent variables [1].

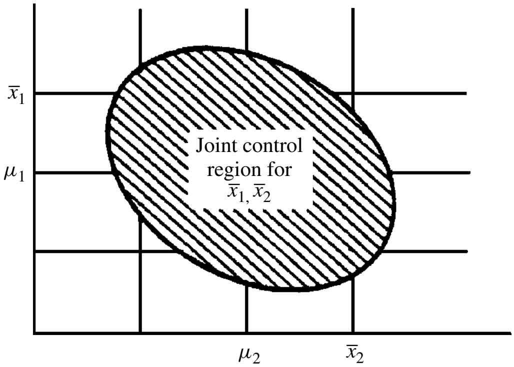


Figure 6.24. Control ellipse for two dependent variables [1].

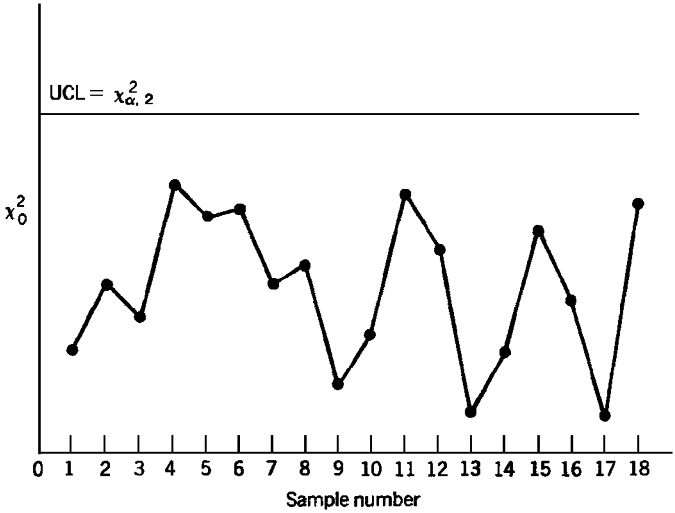


Figure 6.25. χ^2 control chart for $p = 2$ quality characteristics [1].

It is possible to extend this approach to situations where $p > 2$. Assuming again that the p random variables are multivariate normal, this procedure requires computing the sample mean of each quality characteristic from a sample of size n. This set of sample means is represented by the vector

$$\overline{\mathbf{x}} = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_p \end{bmatrix}$$

The test statistic to be plotted on a control chart like that of Figure 6.25 is then

$$\chi_0^2 = n(\overline{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu}) \tag{6.81}$$

where $\boldsymbol{\mu}' = [\mu_1, \mu_2, \ldots, \mu_p]$ and $\boldsymbol{\Sigma}$ is the covariance matrix. The upper control limit for this chart is then $\chi^2_{\alpha,p}$.

In practice, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ must be estimated from samples taken when the process is under control. Suppose that m such samples are taken. The sample means and variances are

$$\overline{x}_{jk} = \frac{1}{n}\sum_{i=1}^{n} x_{ijk} \qquad \begin{cases} j = 1, 2, \ldots, p \\ k = 1, 2, \ldots, m \end{cases} \tag{6.82}$$

$$S_{jk}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ijk} - \overline{x}_{jk})^2 \qquad \begin{cases} j = 1, 2, \ldots, p \\ k = 1, 2, \ldots, m \end{cases} \tag{6.83}$$

where x_{ijk} is the ith observation on the jth quality characteristic in the kth sample. The covariance between characteristic j and h in the kth sample is

$$S_{jhk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ijk} - \overline{x}_{jk})(x_{ihk} - \overline{x}_{hk}) \qquad \begin{cases} k = 1, 2, \ldots, m \\ j \neq h \end{cases} \tag{6.84}$$

The statistics $\overline{x}jk$, S_{jk}^2, and S_{jhk} are averaged over all m samples to obtain

$$\overline{\overline{x}}_j = \frac{1}{m}m\sum_{k=1}\overline{x}_{jk} \qquad j = 1, 2, \ldots, p \tag{6.85}$$

$$S_j^2 = \frac{1}{m}\sum_{k=1}^{m} S_{jk}^2 \qquad j = 1, 2, \ldots, p \tag{6.86}$$

$$S_{jh} = \frac{1}{m}\sum_{k=1}^{m} S_{jhk} \qquad j \neq h \tag{6.87}$$

The $\{\overline{\overline{x}}_j\}$ are elements of a vector $\overline{\overline{\mathbf{x}}}$, and the $p \times p$ sample covariance matrix S is

$$\mathbf{S} = \begin{bmatrix} S_1^2 & S_{12} & S_{13} & \cdots & S_{1p} \\ & S_2^2 & S_{23} & \cdots & S_{2p} \\ & & \ddots & & \vdots \\ & & & & S_p^2 \end{bmatrix} \tag{6.88}$$

The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in Eq. (6.81) are then replaced by $\overline{\overline{\mathbf{x}}}$ and $\mathbf{S}$, respectively. The test statistic is now

$$T^2 = n(\overline{\mathbf{x}} - \overline{\overline{\mathbf{x}}})'\mathbf{S}^{-1}(\overline{\mathbf{x}} - \overline{\overline{\mathbf{x}}}) \tag{6.89}$$

The T^2 statistic is know as *Hotelling's* T^2. The distribution of this T^2 statistic is related to the F distribution by the expression

$$\frac{n-p}{p(n-1)}T^2 \sim F(p, n-p) \tag{6.90}$$

The T^2 statistic can be plotted on a control chart with a UCL $= \chi^2_{\alpha,p}$. However, one difficulty that arises in the use of either the χ_0^2 or T^2 statistics in control charts is the interpretation of an out-of-control signal. Specifically, it is difficult to determine which subset of the p variables is responsible for the signal.

6.5.2. Control of Variability

Multivariate process variability is summarized by the $p \times p$ covariance matrix $\boldsymbol{\Sigma}$. One approach to controlling variability is based on the determinant of the sample covariance ($|\mathbf{S}|$), which is known as the *generalized sample variance*. Let $E(|\mathbf{S}|)$ and $V(|\mathbf{S}|)$ be the mean and variance of $|\mathbf{S}|$, respectively. It can be shown that [1]

$$E(|\mathbf{S}|) = b_1|\boldsymbol{\Sigma}| \tag{6.91}$$

$$V(|\mathbf{S}|) = b_2|\boldsymbol{\Sigma}|^2 \tag{6.92}$$

where

$$b_1 = \frac{1}{(n-1)^p}\prod_{i=1}^{p}(n-i) \tag{6.93}$$

$$b_2 = \frac{1}{(n-1)^{2p}}\prod_{i=1}^{p}(n-i)\left[\prod_{i=1}^{p}(n-j+2) - \prod_{i=1}^{p}(n-j)\right] \tag{6.94}$$

The parameters of the control chart for $|\mathbf{S}|$ are then

$$\begin{aligned} \text{UCL} &= |\boldsymbol{\Sigma}|\left(b_1 + 3\sqrt{b_2}\right) \\ \text{CL} &= b_1|\boldsymbol{\Sigma}| \\ \text{LCL} &= |\boldsymbol{\Sigma}|\left(b_1 - 3\sqrt{b_2}\right) \end{aligned} \tag{6.95}$$

The LCL in Eq. (6.95) is set to zero if the calculated value is less than zero. Also, in practice, $\mathbf{\Sigma}$ is usually estimated by $\mathbf{S}$. In that case, $|\mathbf{\Sigma}|$ in Eq. (6.95) is replaced by $|\mathbf{S}|/b_1$, which is an unbiased estimator of $|\mathbf{\Sigma}|$.

6.6. SPC WITH CORRELATED PROCESS DATA

The standard assumptions for using Shewhart control charts are that the data generated by the monitored process while it is under control are normally and independently distributed. Both the process mean (μ) and standard deviation (σ) are considered fixed and unknown. Therefore, when the process is under control, it can be represented by the model

$$x_t = \mu + \varepsilon_t \qquad t = 1, 2, \ldots \tag{6.96}$$

where $\varepsilon_t \sim \mathrm{N}(0, \sigma^2)$. When these assumptions hold, one may apply conventional SPC techniques to such charts.

The most critical of these assumptions is the independence of the observations. Shewhart charts do not work well if the process measurements exhibit any level of correlation over time. They will give misleading results under these conditions, usually in the form of too many false alarms. Unfortunately, the assumption of uncorrelated (or independent) observations is not satisfied for many semiconductor manufacturing processes. For example, in a CVD process, consecutive temperature measurements are often highly correlated. Automated test and inspection procedures are also examples of processes that yield measurements that are correlated in time. Alternative SPC methods must therefore be applied to these situations.

6.6.1. Time-Series Modeling

When a process measurement taken at time t depends on the value measured at time $t - 1$, the measurements are said to be *autocorrelated*. A sequence of time-oriented observations such as this is referred to as a *time series*. It is possible to measure the level of autocorrelation in a time series analytically using the *autocorrelation function*

$$\rho_k = \frac{\operatorname{cov}(x_t, x_{t-k})}{V(x_t)} \tag{6.97}$$

where $\operatorname{cov}(x_t, x_{t-k})$ is the covariance of observations that are k time periods apart and $V(x_t)$ is the variance of the observations (which is assumed to be constant). Autocorrelation is usually estimated using the sample autocorrelation function

$$r_k = \frac{\sum_{t=1}^{n-k}(x_t - \overline{x})(x_{t-k} - \overline{x})}{\sum_{t=1}^{n}(x_t - \overline{x})^2} \tag{6.98}$$

Autocorrelated data can be modeled using *time-series models*. For example, the quality characteristic x_t could be modeled using the expression

$$x_t = \xi + \phi x_{t-1} + \varepsilon_t \tag{6.99}$$

where ξ and $\phi(-1 < \phi < 1)$ are unknown constants, and $\varepsilon_t \sim N(0, \sigma^2)$. Equation (6.99) is called a *first-order autoregressive model*. The observations (x_t) from this model have mean $\xi(1 - \phi)$ and standard deviation $\sigma/\sqrt{1 - \phi^2}$. Observations that are k time periods apart (x_t and x_{t-k}) have a correlation coefficient ϕ^k.

The first-order autoregressive model is clearly not the only possible model for correlated time-series data. A natural extension to Eq. (6.99) is

$$x_t = \xi + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \varepsilon_t \tag{6.100}$$

which is a *second-order autoregressive model*. In autoregressive models, the variable x_t is directly dependent on previous observations. Another possibility is to model this dependence through ε_t. The simplest way to do so is

$$x_t = \mu + \varepsilon_t - \theta\varepsilon_{t-1} \tag{6.101}$$

which is called a *first-order moving-average model*. In this model, the correlation between x_t and x_{t-1} is

$$\rho_1 = -\theta/(1 + \theta^2) \tag{6.102}$$

Combinations of autoregressive and moving-average models are often useful as well. A first-order *autoregressive, moving-average* (ARMA) model is

$$x_t = \xi + \phi x_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1} \tag{6.103}$$

More generally, this ARMA model may be extended to arbitrary order using

$$\hat{x}_t = \sum_{i=1}^{p} \phi_i x_{t-i} - \sum_{j=1}^{q} \theta_j e_{t-j} \tag{6.104}$$

where $\hat{x}_t$ is the model prediction of the time-series data, e_t is the residual for each timepoint (i.e., $e_t = x_t - \hat{x}_t$), ϕ_i are the autoregressive coefficients of order p, and θ_j are the moving-average coefficients of order q. The parameters in the ARMA or other time-series models may be estimated by the method of least squares (see Chapter 8). Using that technique, values of ξ, ϕ_i, or θ_j are selected that minimize the sum of squared errors (e_t^2).

A further extension of the basic ARMA model is the autoregressive *integrated* moving-average (ARIMA) model. The ARIMA model originates from the *first-order integrated moving-average model*

$$x_t = x_{t-1} + \varepsilon_t - \theta\varepsilon_{t-1} \tag{6.105}$$

While the previous models are used to describe stationary behavior (i.e., x_t wanders around a fixed mean), the model in Eq. (6.106) describes nonstationary behavior in which the process mean drifts. This situation often arises in processes in which no control actions are taken to keep the mean close to a target value. Occasionally, the original data may also show seasonal, periodic patterns. These seasonal patterns can be modeled by creating ARIMA models for seasonal means. This composite model is known as the *seasonal* ARIMA (SARIMA).

6.6.2. Model-Based SPC

One approach for dealing with autocorrelated data is to directly model the correlation with an appropriate time-series model, use that model to remove the autocorrelation from the data, and apply control charts to the residuals. This approach is known as *model-based SPC*.

Consider the first-order autoregressive model described by Eq. (6.99). Suppose that $\hat{\phi}$ is an estimator of ϕ obtained from the analysis of sample data obtained from the process. Then $\hat{x}_t$ is an estimate of x_t and the corresponding residuals ($e_t = x_t - \hat{x}_t$) are approximately normally and independently distributed with zero mean and constant variance. Conventional Shewhart or other control charts may now be legitimately applied to the sequence of residuals. Points out of control or exhibiting unusual patterns would then be indicative of a shift in ϕ, implying that the original variable x_t was out of control.

This approach is equally valid for more complex time-series models such as the ARMA family. As an example, Figure 6.26 shows time-series data collected from a Lam Rainbow 4400 reactive-ion etching system [6]. This particular dataset

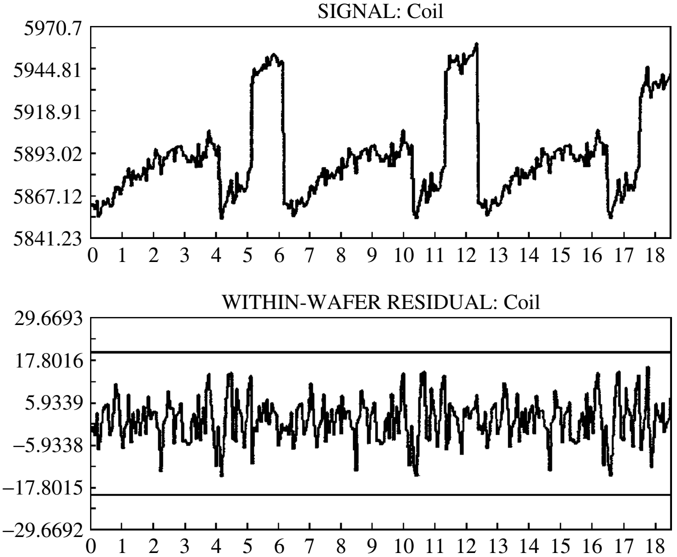


Figure 6.26. RIE RF coil position time series: (a) raw data and (b) control chart for residuals [6].

represents the signal for the coil position in the RF matching network of the RIE. Figure 6.26a shows the raw data for this signal, and Figure 6.26b shows the $\overline{x}$ control chart for the residuals after an ARMA model for the raw data has been constructed. In this case, the time-series data are under statistical control, as the residuals do not exceed the upper and lower control limits for the interval under observation.

SUMMARY

In this chapter, we have provided an overview of statistical process control, from basic control charts to advanced techniques. This overview has focused on the use of SPC to analyze quality issues and improve the performance of semiconductor manufacturing processes. As ever, the overall goal is to reduce variability and improve yield. In the next chapter, we turn our attention to statistical experimental design, an essential method for identifying the key variables influencing the quality characteristics that are monitored by SPC.

PROBLEMS

6.1. Consider Western Electric rule 3. What is the probability of four out of five consecutive points plotting beyond the 2σ warning limits?

6.2. A normally distributed quality characteristic is monitored by a control chart with 3σ limits. Derive a general expression for the probability that a point will plot outside the control limits when the process is in fact in control.

6.3. A control chart is designed to monitor the threshold voltage of NMOS transistors. Assume that the process is under control for some time before an abrupt process shift occurs. Suppose that the chart is set so that it signals an alarm with probability $1 - \beta$ the first time a sample arrives from the shifted process. Find

(a) The probability of signaling an alarm on the second sample after the shift.

(b) The probability that the alarm will be *missed* for K samples following the shift.

(c) The expected number of samples needed after the shift in order to generate an alarm.

6.4. Suppose that out of a group of 10 coins, 9 of them are “fair” (i.e., they turn up heads 50% of the time). One of them is “unfair”—it gives tails only 35% of the time. Assume that each of the coins is thrown n times and the outcome is plotted on a p chart. Calculate the control limits and n so that the unfair coin will be caught 90% of the time, while the chance of rejecting a fair coin will be at most 1%.

6.5. A particle counting device monitors wafers emerging from a plasma etcher. From previous experience, it is known that the machine generates an average of two defects per wafer. Establish a control procedure that will generate false alarms only 1% of the time (there is no lower control limit). What is the best type of control chart, and what is the necessary UCL?

6.6. Control charts for $\overline{x}$, R, and s are to be maintained for the threshold voltage of short-channel MOSFETs using sample sizes of $n = 10$. It is known that the process is normally distributed with $\mu = 0.75$ V and $\sigma = 0.10$ V. Find the centerline and control limits for each of these control charts.

6.7. Repeat the previous problem assuming that μ and σ are unknown and that we have collected 50 observations of sample size 10. These samples yielded a grand average of 0.734 V, an average s_i of 0.125 V, and an average R_i of 0.365 V.

6.8. A fabrication line used for the manufacture of analog ICs requires tight control on the relative sizes of small geometric features. In a particular case, it is required that transistors on either side of a differential pair differ less than 0.1 μm in their effective gate lengths. If that difference is normally distributed with $\mu = 0$ and $\sigma = 0.05$ μm:

(a) Calculate the process capability (C_p) and the fraction of nonconforming product when the process is in control.

(b) Suppose the mean of the process shifts and this shift doubles the fraction of nonconforming product. Calculate the sample size needed to implement a p chart that will detect this shift on the first subsequent sample with 50% certainty.

(c) Design a 3σ $\overline{x}$ and R charts that will detect the previous shift with the same 50% certainty.

6.9. The following values of saturation drain current ($I_{D,\text{sat}}$) were collected from several test wafers with a sample size of $n = 5$.

$\overline{x}$ (mA)	R (mA)
1.03	0.04
1.02	0.05
1.04	0.02
1.05	0.11
1.04	0.04
1.06	0.03
1.02	0.07
1.05	0.02
1.06	0.04
1.04	0.03

(a) Calculate the centerlines and control limits.

(b) Assuming $I_{D,\mathrm{sat}}$ to be normally distributed, compute the standard deviation of the process.

(c) Give an estimate of the fraction nonconforming if the specification limits are 1.03 ± 0.04 mA.

(d) Suggest ways to reduce the fraction of nonconforming product.

6.10. The $\overline{x}$ and R values for 20 samples of size $n = 5$ are shown below. The specification limits of this product are 530–570.

$\overline{x}$	549	548	548	551	553	552	550	551	553	556
R	2.5	2.1	2.3	2.9	1.8	1.7	2.0	2.4	2.2	2.8
$\overline{x}$	547	545	549	552	550	548	556	546	550	551
R	2.0	3.0	3.1	2.2	2.3	2.1	1.9	1.8	2.1	2.2

(a) Construct a modified control chart with 3σ limits. Assume that if the true fraction nonconforming is as large as 1%, the process is acceptable. Is this process satisfactory?

(b) Suppose that if the true fraction nonconforming is as large as 1%, the modified control chart should detect this out-of-control condition with probability 0.9. Construct the modified chart and compare it to the chart obtained in part (a).

(c) Is this process in statistical control?

6.11. The following data represent temperature measurements from a CVD process. The target temperature is 1050°C and the standard deviation is $\sigma = 25$°C. Set up the cusum chart for the mean of this process. Design the cusum to quickly detect a shift of 1σ in the process mean.

Observation	T (°C)	Observation	T (°C)
1	1045	11	1139
2	1055	12	1169
3	1037	13	1151
4	1064	14	1128
5	1095	15	1238
6	1008	16	1125
7	1050	17	1163
8	1087	18	1188
9	1125	19	1146
10	1146	20	1167

6.12. Consider a process with $\mu_0 = 10$ and $\sigma = 1$. Set up 3σ EWMA control charts for $\lambda = 0.1$, 0.2, and 0.4. Discuss the effect of λ on the behavior of the control limits.

6.13. The data below come from a process with two observable quality characteristics. The data are the means of each characteristic, based on a sample size of $n = 25$. The nominal values and covariance matrix are

$$\bar{\bar{\mathbf{x}}} = \begin{bmatrix}55\\30\end{bmatrix} \qquad \mathbf{S} = \begin{bmatrix}200 & 130\\130 & 120\end{bmatrix}$$

Construct the T^2 control chart using these data.

Sample	$\bar{\mathbf{x}}_1$	$\bar{\mathbf{x}}_1$
1	58	32
2	60	33
3	50	27
4	54	31
5	63	38
6	53	30
7	42	20
8	55	31
9	46	25
10	50	29
11	49	27
12	57	30
13	58	33
14	75	45
15	55	27

REFERENCES

1. D. Montgomery, *Introduction to Statistical Quality Control*, Wiley, New York, 1993.
2. A. Duncan, *Quality Control and Industrial Statistics*, Irwin, Homewood, IL, 1974.
3. Western Electric, *Statistical Quality Control Handbook*, Western Electric Corp., Indianapolis, IN, 1956.
4. S. Campbell, *The Science and Engineering of Microelectronic Fabrication*, Oxford Univ. Press, New York, 2001.
5. D. Hawkins, "Cumulative Sum Control Charting: An Underutilized SPC Tool," *Qual. Engi.* **5** (1993).
6. S. Lee, E. Boskin, H. Liu, E. Wen, and C. Spanos, "RTSPC: A Software Utility for Real-Time SPC and Tool Analysis," *IEEE Trans. Semiconduct. Manuf.* **8**(1) (Feb. 1995).

7

STATISTICAL EXPERIMENTAL DESIGN

OBJECTIVES

- Provide an overview of statistical experimental design techniques.
- Introduce the concept of analysis of variance (ANOVA).
- Define and describe various types factorial designs.
- Discuss the Taguchi method of experimental design.

INTRODUCTION

Experiments allow investigators to determine the effects of several variables on a given process or product. A *designed experiment* is a test or series of tests that involve purposeful changes to these variables in order to observe the effect of the changes on that process or product. *Statistical experimental design* is an efficient approach for systematically varying these controllable process variables and ultimately determining their impact on process or product quality. This approach is useful for comparing methods, deducing dependences, and creating models to predict effects.

Statistical process control and experimental design are closely interrelated. Both techniques can be used to reduce variability. However, SPC is a passive approach in which a process is monitored and data are collected, whereas experimental design requires active intervention in performing tests on the process under

Fundamentals of Semiconductor Manufacturing and Process Control,
By Gary S. May and Costas J. Spanos

different conditions. Experimental design can also be beneficial in implementing SPC, since designed experiments may help identify the most influential process variables, as well as their optimum settings.

Overall, experimental design is a powerful engineering tool for improving a manufacturing process. Application of experimental design techniques can lead to

- Improved yield
- Reduced variability
- Reduced development time
- Reduced cost

Ultimately, the result is enhanced manufacturability, performance, and product reliability. This chapter illustrates the use of experimental design methods in semiconductor manufacturing.

7.1. COMPARING DISTRIBUTIONS

In the method of statistical inference known as *hypothesis testing* (see Chapter 4), an investigator must evaluate a result produced by making some experimental modification of a system. The investigator must determine whether the result is explainable by mere chance or whether it is due to the effectiveness of the modification. In order to make this determination, the experimenter must identify a relevant reference set that represents a characteristic set of outcomes that could occur if the modification were completely without effect. The actual experimental outcome can then be compared with this reference set. If the experimental results are found to be exceptional, the results are considered *statistically significant*.

Consider the yield data in Table 7.1 obtained from a semiconductor manufacturing process in which two batches of 10 wafers each were fabricated using a standard method (method A) and a modified method (method B). The question to be answered from the experiment is what evidence (if any) do the data collected provide that method B is really better than method A?

To answer this question, we examine the average yields for each process. The modified method (method B) gave an average yield that was 1.30% higher than the standard method. However, because of the considerable variability in the individual test results, it might not be correct to immediately conclude that method B is superior to method A. In fact, it is conceivable that the difference observed could be due to experimental error, operator error, or even pure chance.

One approach to determining the significance of the differences between method A and method B is to use an *external reference distribution*. Suppose in this instance that additional data were available in the form of 210 past process records. These 210 past observations, plotted in Figure 7.1, were made using the standard process, method A. The key question now becomes: How often have the yield differences between the averages of successive groups of 10 wafers been at as large as 1.30%? If the answer is “frequently,” we conclude that the

Table 7.1. Yield data from a hypothetical semiconductor manufacturing process [1].

Wafer	Method A Yield (%)	Method B Yield (%)
1	89.7	84.7
2	81.4	86.1
3	84.5	83.2
4	84.8	91.9
5	87.3	86.3
6	79.7	79.3
7	85.1	86.2
8	81.7	89.1
9	83.7	83.7
10	84.5	88.5
Average	84.24	85.54

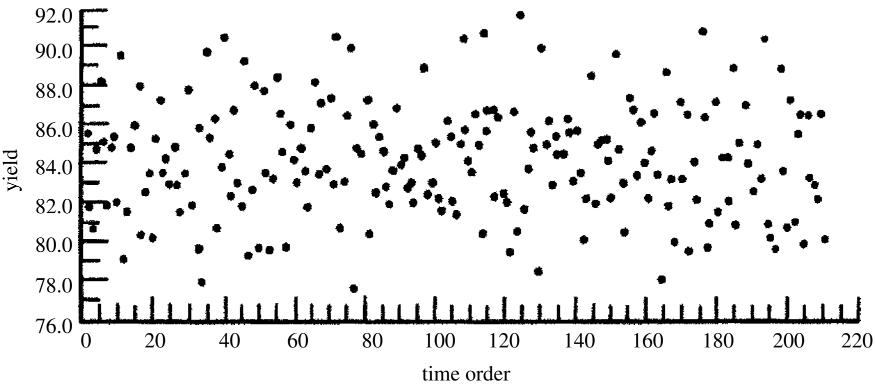


Figure 7.1. Plot of 210 prior observations of method A yield [1].

observed difference can be readily explained by the purely chance variations in the process. However, if the answer is "rarely," a better explanation is that the modification in method B has truly produced an increase in the mean yield.

Figure 7.2 shows the 191 differences between yield averages of adjacent groups of 10 observations in the database of 210 past process records. These 191 differences were obtained by comparing the averages of wafers 1–10, 2–11, and so on. They provide a *relevant reference set* with which the observed difference of 1.30% may be legitimately compared. Doing so, it is seen that rarely, in fact, do the differences in the reference set exceed 1.30% (specifically, in only nine cases). Using statistical terminology, we can say that in relation to this reference set, the observed difference of 1.30% is statistically significant at the $9/191 = 0.047$ level. In other words, less than 5 times in 100 would a difference as large as 1.30% be found in the reference set. Thus, it is likely that an actual difference does exist, and method B is truly better than method A.

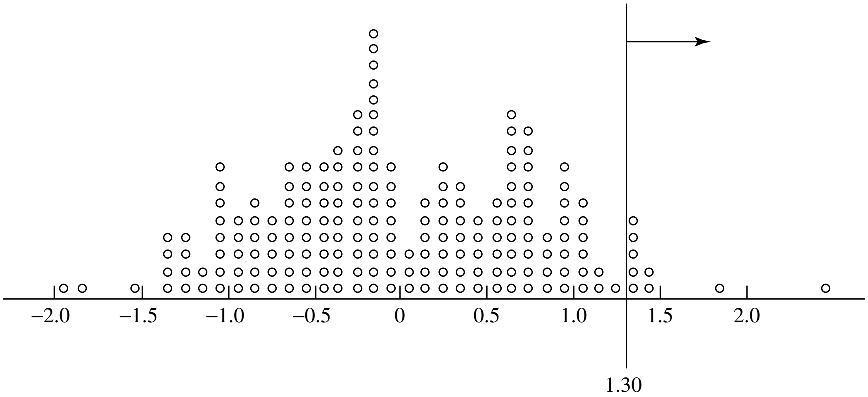


Figure 7.2. Reference distribution for historical method A yield data [1].

The external reference distribution technique can be problematic. Suppose that there is no historical database of yield obtained using method A. In this case, the proper approach to determine whether the difference between the two manufacturing processes is significant is a statistical *hypothesis test* (Section 4.5). In this case the hypothesis can be represented as

$$\begin{aligned} &\mathrm{H}_0\colon\ \mu_A = \mu_B \\ &\mathrm{H}_1\colon\ \mu_A \neq \mu_B \end{aligned} \tag{7.1}$$

where μ_A and μ_B represent the mean yields for the two methods. Since the variance for this process is not known, the test statistic for this hypothesis is

$$t_0 = \frac{(\overline{y}_A - \overline{y}_B)}{s_p\sqrt{\dfrac{1}{n_A} + \dfrac{1}{n_B}}} \tag{7.2}$$

where y_A and y_B are the sample means, n_A and n_B are the number of trials in each sample (10 each in this case), and

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2} \tag{7.3}$$

We use the pooled estimate of the common variance since although the variance for the process is unknown, there is no reason to suspect that the application of method A or method B will produce a different variance. The values of the sample variances [calculated using Eq. (4.2)] are $s_A = 2.90$ and $s_B = 3.65$. Using Eqs. (7.3) and (7.2) then gives values of $s_p = 3.30$ and $t_0 = 0.88$, respectively. Interpolating from Appendix D, we find that the likelihood of computing a t statistic with $n = n_A + n_B - 2 = 18$ degrees of freedom equal to 0.88 is 0.195. The value 0.195 is the *statistical significance* of the hypothesis test. This means that there is only an 19.5% chance that the observed difference between the mean yields is due to pure chance. In other words, we can be 80.5% confident that method B is really superior to method A.

7.2. ANALYSIS OF VARIANCE

The scenario described above is a useful example of how we might use hypothesis testing to compare two distributions. However, in many cases, we would like to go even further; it is often important in manufacturing applications to be able to compare several distributions simultaneously. Moreover, we might also be interested in determining which process conditions in particular have a significant impact on process quality. *Analysis of variance* (ANOVA) is an excellent technique for accomplishing these objectives. ANOVA builds on the idea of hypothesis testing and allows us to compare different sets of process conditions (i.e., "treatments"), as well as to determine whether a given treatment results in a statistically significant variation in quality.

The ANOVA procedure is best illustrated by example. Consider the data in Table 7.2, which represents hypothetical yield data measured for four different sets of process recipes (labeled A–D). Through the use of ANOVA, we will determine whether the discrepancies *between* recipes (i.e., treatments) is truly greater than the variation of the yield *within* the individual groups processed with the same recipe. We assume that the data can be treated as independent random samples from four normal populations having the same variance and differing only in their means (if at all).

Let k be the number of treatments ($k = 4$ in this case). Note that the sample size (n) for each treatment varies ($n_1 = 4$, $n_2 = n_3 = 6$, and $n_4 = 8$). The treatment means (in %) are $\overline{y}_1 = 61$, $\overline{y}_2 = 66$, $\overline{y}_3 = 68$, and $\overline{y}_4 = 61$. The total number of samples (N) is 24, and the *grand average* of all 24 samples, which is sum of all observations divided by the total number of observations, is $\overline{y} = 64\%$.

7.2.1. Sums of Squares

To perform ANOVA, several key parameters must be computed. These parameters, called *sums of squares*, serve to quantify deviations within and between different treatments. Let y_{ti} represent the ith observation for the tth treatment.

Table 7.2. Hypothetical yield (in %) for four different process recipes [1].

Recipe A	Recipe B	Recipe C	Recipe D
62	63	68	56
60	67	66	62
63	71	71	60
59	64	67	61
	65	68	63
	66	68	64
			63
			59

The sum of squares within the tth treatment is given by

$$S_t = \sum_{i=1}^{n_t} (y_{ti} - \overline{y}_t)^2 \tag{7.4}$$

where n_t is the sample size for the treatment in question and $\overline{y}_t$ is the treatment mean. The *within-treatment sum of squares* for all treatments is

$$S_R = S_1 + S_2 + \cdots + S_k = \sum_{t=1}^{k} \sum_{i=1}^{n_t} (y_{ti} - \overline{y}_t)^2 \tag{7.5}$$

In order to quantify the deviations of the treatment averages about the grand average, we use the *between-treatment sum of squares*, which is given by

$$S_T = \sum_{t=1}^{k} n_t (\overline{y}_t - \overline{y})^2 \tag{7.6}$$

Finally, the total sum of squares for all the data about the grand average is

$$S_D = \sum_{t=1}^{k} \sum_{i=1}^{n_t} (y_{ti} - \overline{y})^2 \tag{7.7}$$

Each sum of squares has an associated number of *degrees of freedom* required for its computation. The degrees of freedom for the within-treatment, between-treatment, and total sums of squares, respectively, are

$$\begin{aligned} \nu_R &= N - k \\ \nu_T &= k - 1 \\ \nu_D &= N - 1 \end{aligned} \tag{7.8}$$

The final quantity needed to carry out analysis of variance is the pooled estimate of the variance quantified by each sum of squares. This quantity, known as the *mean square*, is equal to the ratio of the sum of squares to its associated number of degrees of freedom. The within-treatment, between-treatment, and total mean squares are therefore

$$\begin{aligned} s_R^2 &= \frac{S_R}{\nu_R} = \frac{\displaystyle\sum_{t=1}^{k} \sum_{i=1}^{n_t} (y_{ti} - \overline{y}_t)^2}{N - k} \\ s_T^2 &= \frac{S_T}{\nu_T} = \frac{\displaystyle\sum_{t=1}^{k} n_t (\overline{y}_t - \overline{y})^2}{k - 1} \\ s_D^2 &= \frac{S_D}{\nu_D} = \frac{\displaystyle\sum_{t=1}^{k} \sum_{i=1}^{n_t} (y_{ti} - \overline{y})^2}{N - 1} \end{aligned} \tag{7.9}$$

For a null hypothesis that there are no differences between the treatment means, the within-treatment mean square (s_R^2) and the between-treatment mean square (s_T^2) provide two estimates of the true process variance (σ^2). For the dataset in Table 7.2, using Eqs. (7.9), we obtain $s_R^2 = 5.6$ and $s_T^2 = 76.0$. The fact that the between-treatment estimate of σ^2 is much larger than the within-treatment estimate tends to discredit the null hypothesis. We are thus led to suspect that some of the between-treatment variation must be caused by real differences in the treatment means. In the following section, we show how the necessary calculations to draw this conclusion may be conveniently arranged in tabular form.

7.2.2. ANOVA Table

Once the sums of squares and mean squares have been computed, it is customary to arrange them in a tabular format called and *ANOVA table*. The general form of the ANOVA table is depicted in Table 7.3. The ANOVA table that corresponds to the via diameter data in Table 7.2 is shown in Table 7.4.

The astute reader will note that in both the "sum of squares" and "degrees of freedom" columns, the values for between and within treatments add up to give the corresponding total value. This additive property of the sum of squares arises from the algebraic identity

$$\sum_{t=1}^{k}\sum_{i=1}^{n_t}(y_{ti}-\overline{y})^2 = \sum_{t=1}^{k} n_t(\overline{y}_t-\overline{y})^2 + \sum_{t=1}^{k}\sum_{i=1}^{n_t}(y_{ti}-\overline{y}_t)^2 \tag{7.10}$$

or equivalently, $S_D = S_T + S_R$.

The complete ANOVA table provides a mechanism for testing the hypothesis that all of the treatment means are equal. The null hypothesis in this case is thus

$$H_0\colon\ \mu_1 = \mu_2 = \mu_3 = \mu_4$$

Table 7.3. General format of the ANOVA table.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F* Ratio
Between treatments	S_T	$\nu_T = k - 1$	s_T^2	s_{T}^2/s_R^2
Within treatments	S_R	$\nu_R = N - k$	s_R^2	
Total about the grand average	S_D	$\nu_D = N - 1$	s_D^2	

Table 7.4. ANOVA table for yield data.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F* Ratio
Between treatments	$S_T = 228$	$\nu_T = 3$	$s_T^2 = 76.0$	$s_T^2/s_R^2 = 13.6$
Within treatments	$S_R = 112$	$\nu_R = 20$	$s_R^2 = 5.6$	
Total about the grand average	$S_D = 340$	$\nu_D = 23$	$s_D^2 = 14.8$	

If the null hypothesis were true, the ratio s_T^2/s_R^2 would follow the F distribution with n_T and n_R degrees of freedom. According to Appendix E, the significance level for the observed F ratio of 13.6 with 3 and 30 degrees of freedom is 0.000046. This means that there is only a 0.0046% chance that the means are in fact equal, and the null hypothesis is discredited. In other words, we can be 99.9954% sure that real differences exist among the four different processes used in our example.

An alternative format for the ANOVA table exists. The quantity S_D, the total sum of squares about the grand average, can also be written as

$$S_D = \sum_{t=1}^{k} \sum_{i=1}^{n_t} y_{ti}^2 - N\overline{y}^2 \tag{7.11}$$

In this expression, the latter term ($N\overline{y}^2$) is the sum of squares due to the grand average, which is often called the *correction factor for the average*. It is denoted by S_A (i.e., $S_A = N\overline{y}^2$). The first term in the expression $\left(\sum_{t=1}^{k} \sum_{i=1}^{n_t} y_{ti}^2\right)$ is called the *total sum of squares*, and it is denoted by S. Combining Eqs. (7.10) and (7.11), we can thus decompose the sum of squares of the original N observations into three additive terms:

$$\sum_{t=1}^{k} \sum_{i=1}^{n_t} y_{ti}^2 = N\overline{y}^2 + \sum_{t=1}^{k} n_t(\overline{y}_t - \overline{y})^2 + \sum_{t=1}^{k} \sum_{i=1}^{n_t} (y_{ti} - \overline{y}_t)^2 \tag{7.12}$$

or equivalently, $S = S_A + S_T + S_R$. The associated degrees of freedom are

$$N = 1 + (k-1) + (N-k) \tag{7.13}$$

This representation leads to the "full" ANOVA table (Table 7.5), which specifically includes the contributions from the grand average. However, this contribution is of limited practical interest, so the ANOVA table of the form shown in Table 7.3 is usually preferred.

7.2.2.1. Geometric Interpretation

Equation (7.12) can be further explained by breaking up the yield data from Table 7.2 in the manner shown in Table 7.6. This table shows that each individual observation is composed of the following components: the grand average ($\overline{y}$); the between-treatment deviation ($\overline{y}_t - \overline{y}$); and the within-treatment deviation, or residual ($y_{ti} - \overline{y}_t$). Each of the four entries in Table 7.6 can be considered a

Table 7.5. Full ANOVA table.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Average	S_A	$\nu_A = 1$	$s_A^2 = S_A/\nu_A$
Between treatments	S_T	$\nu_T = k-1$	$s_T^2 = S_T/\nu_T$
Within treatments	S_R	$\nu_R = N-k$	$s_R^2 = S_R/\nu_R$
Total	S	N	

Table 7.6. Arithmetic decomposition of yield data in Table 7.2 [1].

	Observations		Grand Average		Treatment Deviations		Residuals
	y_{ti}		$\overline{y}$		$\overline{y}_t - \overline{y}$		$y_{ti} - \overline{y}_t$
	$\left[\begin{array}{cccc} 62 & 63 & 68 & 56 \\ 60 & 67 & 66 & 62 \\ 63 & 71 & 71 & 60 \\ 59 & 64 & 67 & 61 \\ & 65 & 68 & 63 \\ & 66 & 68 & 64 \\ & & & 63 \\ & & & 59 \end{array}\right]$	$=$	$\left[\begin{array}{cccc} 64 & 64 & 64 & 64 \\ 64 & 64 & 64 & 64 \\ 64 & 64 & 64 & 64 \\ 64 & 64 & 64 & 64 \\ & 64 & 64 & 64 \\ & 64 & 64 & 64 \\ & & & 64 \\ & & & 64 \end{array}\right]$	$+$	$\left[\begin{array}{cccc} -3 & 2 & 4 & -3 \\ -3 & 2 & 4 & -3 \\ -3 & 2 & 4 & -3 \\ -3 & 2 & 4 & -3 \\ & 2 & 4 & -3 \\ & 2 & 4 & -3 \\ & & & -3 \\ & & & -3 \end{array}\right]$	$+$	$\left[\begin{array}{cccc} 1 & -3 & 0 & -5 \\ -1 & 1 & -2 & 1 \\ 2 & 5 & 3 & -1 \\ -2 & -2 & -1 & 0 \\ & -1 & 0 & 2 \\ & 0 & 0 & 3 \\ & & & 2 \\ & & & -2 \end{array}\right]$
Vector	**Y**	=	**A**	+	**T**	+	**R**
Sum of squares	98,644	=	93,304	+	228	+	112
Degrees of freedom	24	=	1	+	3	+	20

vector. Let **Y** represent the vector of observations, **A** represent the grand average, **T** represent the between-treatment deviations, and **R** represent the residuals. Using the rules of vector addition, we can write

$$\mathbf{Y} = \mathbf{A} + \mathbf{T} + \mathbf{R} \tag{7.14}$$

The sums of squares in the ANOVA table, therefore, are merely the squares of the individual vector elements summed. In other words, the sums of squares are the squared lengths of the vectors **Y**, **A**, **T**, and **R**.

The geometry of this example is illustrated graphically in Figures 7.3–7.5. In Figure 7.3, the vector **Y** is resolved into two components: **A**, which corresponds to the grand average; and **D**, whose elements are the deviations from the grand average. The vector **D** is orthogonal to **A** since $\sum_{j=1}^{N} \overline{y}(y_j - \overline{y}) = 0$. In Figure 7.4, the vector **D** is likewise resolved into two components: **T**, associated with the treatment deviations; and **R**, which corresponds to the residuals. Finally, in Figure 7.5, the observation vector **Y** is resolved into its three orthogonal components, as indicated in Eq. (7.14). The fact that these three vectors are mutually orthogonal is easily confirmed by noting that their inner products are equal to zero.

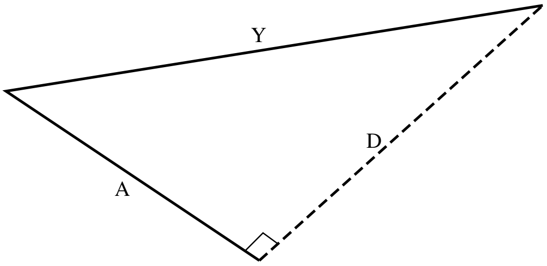


Figure 7.3. Geometric representation of the decomposition of **Y** in terms of **A** and **D** [1].

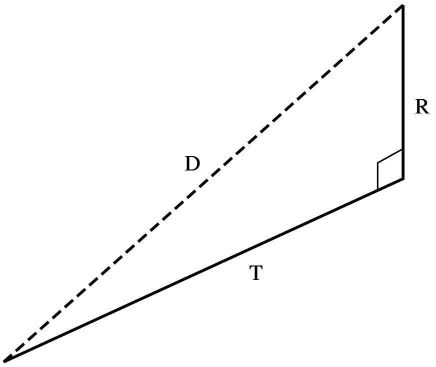


Figure 7.4. Geometric representation of the decomposition of **D** in terms of **T** and **R** [1].

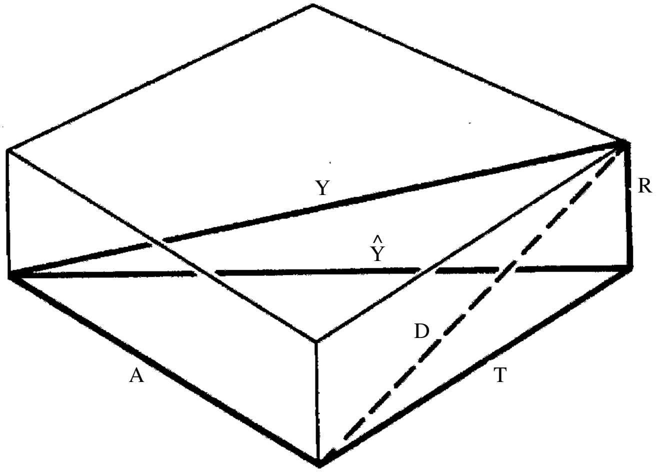


Figure 7.5. Geometric representation of ANOVA in terms of an orthogonal decomposition of **Y** in terms of **A**, **T**, and **R** [1].

The additive relationship $S = S_A + S_T + S_R$ arises from the Pythagorean theorem, which relates the square of the length of the "hypotenuse" **Y** to the sum of squares of the lengths of the three other sides: **A**, **T**, and **R**. The estimated values in the ANOVA technique are the elements of the vector $\widehat{\mathbf{Y}}$, where

$$\widehat{\mathbf{Y}} = \mathbf{A} + \mathbf{T} \tag{7.15}$$

As mentioned previously, the ANOVA technique is frequently applied after "elimination" of the grand average. Table 7.7 shows this approach to the analysis. The vector **D** represents the deviations from the grand average $(y_{ti} - \overline{y})$ after **A** has been subtracted from **Y**.

7.2.2.2. ANOVA Diagnostics

The ANOVA technique is appropriate for a specific implied model that links the experimental observations and the various decompositions with the underlying

Table 7.7. Arithmetic decomposition of deviations from the grand average [1].

	Deviations from Grand Average		Treatment Deviations		Residuals
	$y_{ti}-\overline{y}$		$\overline{y}_t-\overline{y}$		$y_{ti}-\overline{y}_t$
	$\begin{bmatrix} -2 & -1 & 4 & -8 \\ -4 & 3 & 2 & -2 \\ -1 & 7 & 7 & -4 \\ -5 & 0 & 3 & -3 \\ & 1 & 4 & -1 \\ & 2 & 4 & 0 \\ & & & -1 \\ & & & -5 \end{bmatrix}$	=	$\begin{bmatrix} -3 & 2 & 4 & -3 \\ -3 & 2 & 4 & -3 \\ -3 & 2 & 4 & -3 \\ -3 & 2 & 4 & -3 \\ & 2 & 4 & -3 \\ & 2 & 4 & -3 \\ & & & -3 \\ & & & -3 \end{bmatrix}$	+	$\begin{bmatrix} 1 & -3 & 0 & -5 \\ -1 & 1 & -2 & 1 \\ 2 & 5 & 3 & -1 \\ -2 & -2 & -1 & 0 \\ & -1 & 0 & 2 \\ & 0 & 0 & 3 \\ & & & 2 \\ & & & -2 \end{bmatrix}$
Vector	$\mathbf{D}=\mathbf{Y}-\mathbf{A}$	=	**T**	+	**R**
Sum of squares	340	=	228	+	112
Degrees of freedom	23	=	3	+	20

parameters of the sampled population. Specifically, if the data are uncorrelated random samples from *normal* populations having the same variance, but possibly different means, then the implied model is

$$y_{ti} = \eta_t + \varepsilon_{ti} \tag{7.16}$$

where η_t is the mean for the t^{th} treatment, and the errors $\varepsilon_{ti} \sim N(0,\sigma^2)$. If this normality assumption for the errors is appropriate, then all the relevant information about $\eta_1, \eta_2, \ldots, \eta_k$ and σ^2 is supplied by the k treatment averages $(\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_k)$ and s_R^2, respectively. If the assumption is *exact*, then after all of these statistics have been calculated, no further relevant information remains in the original data. Under these conditions, the residuals and original observations can be ignored, and interpretation of the experimental results rests solely with the interpretation of the statistics.

However, in practice, it is unwise to proceed in this manner without further checks. The data may in fact contain information not accounted for by the model in Eq. (7.16) and therefore not revealed by the ANOVA methodology. Discrepancies of this type may be detected by studying the residuals $(y_{ti} - \hat{y}_{ti})$, which are the elements of the vector **R**. These residuals are the quantities that remain after the systematic contributions from the treatment averages have been removed. When the assumptions regarding the adequacy of the model in Eq. (7.16) are true, these residuals should vary randomly. If, however, the residuals display unexplained systematic tendencies, then the model becomes suspicious.

One type of residual inspection that must be carried out is plotting an overall *dot diagram*. The dot diagram for the yield data in Table 7.2 is shown in Figure 7.6. If the normality assumption for the model errors is true, then this diagram should essentially have the appearance of a sample from a normal distribution with mean zero. (Note that considerable fluctuation in appearance will occur if the number of observations is too small.) A common discrepancy revealed by an abnormal dot diagram occurs when one or more of the residuals is much larger or smaller than the others. The plot in Figure 7.6 gives no indication of such an abnormality.

Abnormal residual behavior may also be associated with a particular treatment. To detect problems of this sort, individual dot diagrams for each treatment are prepared, as shown in Figure 7.7. Again, these plots should appear as samples from a normal distribution. The plots in Figure 7.7 do not suggest any anomalous behavior.

If the model in Eq. (7.16) is appropriate, then the residuals should also be unrelated to the levels of any known variable. In particular, they should be unrelated to the level of the response itself. This can be investigated by plotting the residuals versus the estimated response $\hat{y}_{ti}$, as shown in Figure 7.8. This plot should also appear random. If the variance increased with the value of the response, then this plot would have a "funnel-like" appearance. No such behavior is apparent in Figure 7.8.

Finally, sometimes a process may drift or the skill of the experimenter may change with time. Tendencies such as this are revealed by plotting the residuals

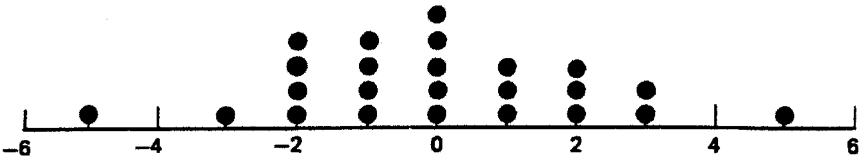


Figure 7.6. Overall dot diagram for all residuals [1].

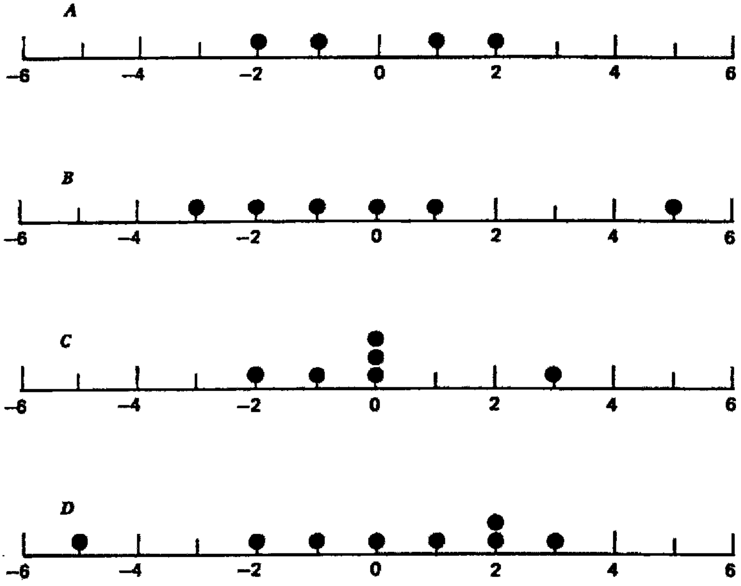


Figure 7.7. Plots of residuals for each treatment [1].

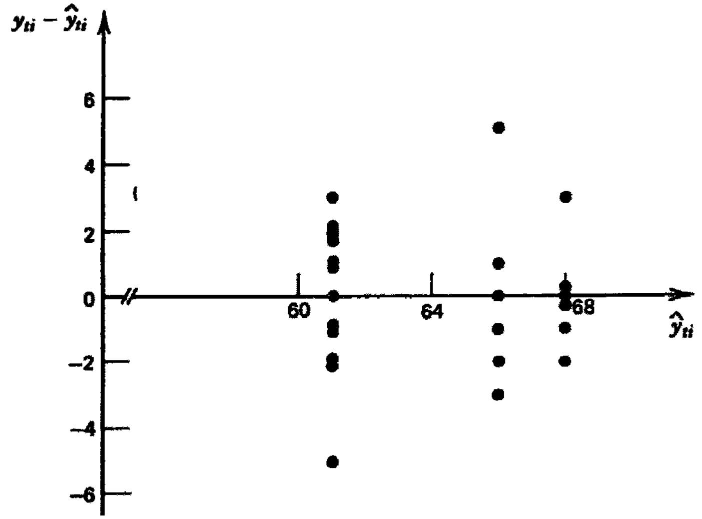


Figure 7.8. Plot of residuals versus estimated values [1].

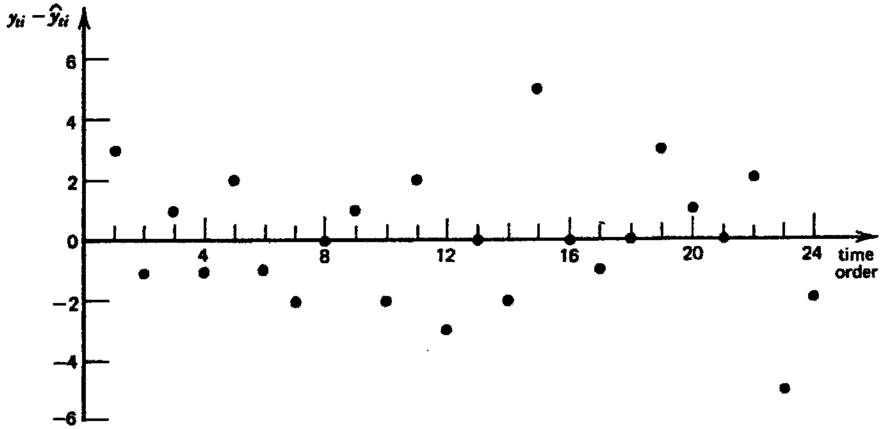


Figure 7.9. Plot of residuals versus time [1].

against their time order, as shown in Figure 7.9. Since the plot appears random in this case, there seems to be no reason to suspect any such effect for this dataset.

7.2.3. Randomized Block Experiments

We now extend the comparison of *k* treatments using ANOVA to examining experimental designs with *blocking*. Blocks might represent, for example, different batches of manufactured products (such as semiconductor wafers) or different contiguous periods of time. In blocked designs, the goal is to quantify both the effects of the treatments and the effect of the blocking arrangement.

As an example of a blocked experiment, consider the yield data in Table 7.8 obtained from a manufacturing process in which five batches of silicon wafers were fabricated using various methods (labeled A–D). In this case, there are $k = 4$ treatments and $n = 5$ blocks. A randomized block design of this kind serves to eliminate variations between blocks (i.e., the batches) from the comparison of treatments. It also provides a broader inductive basis than an experiment with only a single batch.

Analysis of data of this type is undertaken using the ANOVA table with the format shown in Table 7.9. In this table, $\overline{y}$ is the grand average, $\overline{y}_i$ are the block averages, and $\overline{y}_t$ are the treatment averages. The ANOVA table computed for the yield data in Table 7.8 is Table 7.10.

We are now ready to test the hypothesis that all the treatment means are equal (i.e., H$_0$: $\mu_A = \mu_B = \mu_C = \mu_D$). If the null hypothesis were true, the ratio s_T^2/s_R^2 would follow the F distribution with ν_T and ν_R degrees of freedom. According to Appendix E, the significance level for the observed F ratio of 1.24 with 3 and 12 degrees of freedom is 0.33. This means that there is a 33% chance that the means

Table 7.8. Yield data from a hypothetical semiconductor manufacturing process [1].

Block	method A Yield (%)	Method B Yield (%)	Method C Yield (%)	Method D Yield (%)	*Block Average*
Batch 1	89	88	97	94	*92*
Batch 2	84	77	92	79	*83*
Batch 3	81	87	87	85	*85*
Batch 4	87	92	89	84	*88*
Batch 5	79	81	80	88	*82*
Treatment Average	*84*	*85*	*89*	*86*	$\overline{y} = 86$

Table 7.9. Format for two-way ANOVA table with blocking.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio
Average	$S_A = nk\overline{y}^2$	$\nu_A = 1$	$s_A^2 = S_A/\nu_A$	
Between blocks	$S_B = k\sum_{i=1}^{n}(\overline{y}_i - \overline{y})^2$	$\nu_B = n - 1$	$s_B^2 = S_B/\nu_B$	s_B^2/s_R^2
Between treatments	$S_T = n\sum_{t=1}^{k}(\overline{y}_t - \overline{y})^2$	$\nu_T = k - 1$	$s_T^2 = S_T/\nu_T$	s_T^2/s_R^2
Residuals	$S_R = \sum_{t=1}^{k}\sum_{i=1}^{n} \times (y_{ti} - \overline{y}_i - \overline{y}_t + \overline{y})^2$	$\nu_R = (n-1)(k-1)$	$s_R^2 = S_R/\nu_R$	
Total	$S = \sum_{t=1}^{k}\sum_{i=1}^{n} y_{ti}^2$	$\nu = nk$		

Table 7.10. Two-way ANOVA table for yield data.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio
Average	$S_A = 147,920$	$\nu_A = 1$	$s_A^2 = 147{,}920$	
Between blocks	$S_B = 264$	$\nu_B = 4$	$s_B^2 = 66.0$	3.51
Between treatments	$S_T = 70$	$\nu_T = 3$	$s_T^2 = 23.3$	1.24
Residuals	$S_R = 226$	$\nu_R = 12$	$s_R^2 = 18.8$	
Total	$S = 148,480$	$\nu = 20$		

are in fact equal. In other words, we can be only 67% sure that real differences exist among the four different methods used to manufacture the wafers in this example. Thus, the four methods have not been conclusively demonstrated to give different yields.

The blocking arrangement of this experiment also allows us to test the hypothesis that the block means are equal. If this null hypothesis were true, the ratio s_B^2/s_R^2 would follow the F distribution with ν_B and ν_R degrees of freedom. According to Appendix E, the significance level for the observed F ratio of 3.51 with 4 and 12 degrees of freedom is 0.04. This means that there is only a 4% chance that the means are in fact equal. Thus, there exists a 96% chance that there are in fact differences between the batches.

7.2.3.1. Mathematical Model

The mathematical model implicit in randomized block experiments is

$$y_{ti} = \eta + \beta_i + \tau_t + \varepsilon_{ti} \tag{7.17}$$

where η is the general mean, β_i is the block effect, τ_t is the treatment effect, and ε_{ti} is the experimental error. It is assumed that $\varepsilon_{ti} \sim \mathrm{N}(0, \sigma^2)$. Associated with this additive model is the following decomposition of the observations:

$$y_{ti} = \overline{y} + (\overline{y}_i - \overline{y}) + (\overline{y}_t - \overline{y}) + (y_{ti} - \overline{y}_i - \overline{y}_t + \overline{y}) \tag{7.18}$$

The last term, $(y_{ti} - \overline{y}_i - \overline{y}_t + \overline{y})$, is known as the *residual* because it represents what remains after the grand average, block effects, and treatment effects have all been accounted for. The model is called *additive* since, for example, if treatment τ_3 caused an increase of five units in the response and the influence of block β_4 increased the response by seven units, then the cumulative increase caused by both acting together would be $5 + 7 = 12$ units.

In vector notation, the decomposition in Eq. (7.18) can be written as follows:

$$\mathbf{Y} = \mathbf{A} + \mathbf{B} + \mathbf{T} + \mathbf{R} \tag{7.19}$$

In this equation, each of the symbols represents a vector containing $N = nk$ elements of the corresponding two-way ANOVA table. The sums of squares in the ANOVA table are once again the squares of the individual vector elements

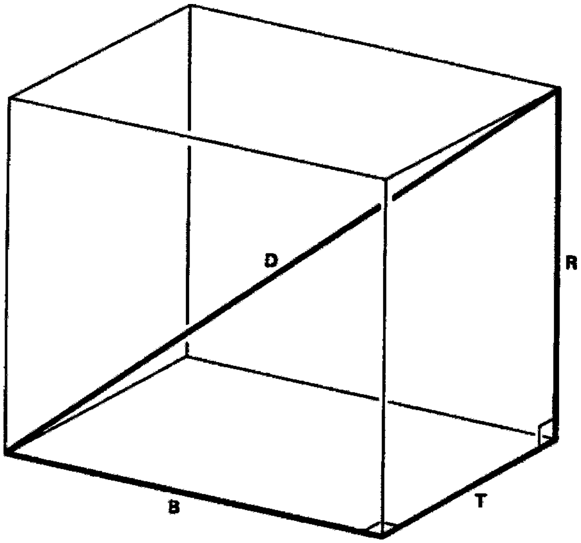


Figure 7.10. Vector decomposition for randomized block ANOVA [1].

summed. In other words, the sums of squares are the squared lengths of the vectors **Y**, **A**, **B**, **T**, and **R**, or

$$S = S_A + S_B + S_T + S_R \tag{7.20}$$

The vectors **A, B, T**, and **R**, are all mutually perpendicular, as illustrated in Figure 7.10. This figure also illustrates the relationship

$$\mathbf{D} = \mathbf{B} + \mathbf{T} + \mathbf{R} \tag{7.21}$$

where $\mathbf{D} = \mathrm{Y} - \mathbf{A}$ is a vector of deviations of the data from the grand average. Since **B, T**, and **R** are mutually orthogonal, we also have

$$S_D = S_B + S_T + S_R \tag{7.22}$$

In other words, the sum of squares of the deviations from the grand average equals the sum of squares for the blocks plus the sum of squares for the treatments plus the sum of squares of the residuals.

7.2.3.2. Diagnostic Checking

Any potential inadequacies in the model proposed in Eq. (7.17) and analyzed using the randomized block ANOVA technique must be investigated by diagnostic methods similar to those discussed in Section 7.2.2.2. The residual plots for the model of the yield data in Table 7.8 are shown in Figure 7.11. The plots in (a) and (b) of this figure reveal nothing of special concern, but the plot of residuals versus predicted values in (c) shows a possible problem in its "funnel" shape, suggesting a possible relationship between the mean and variance. Such discrepancies can

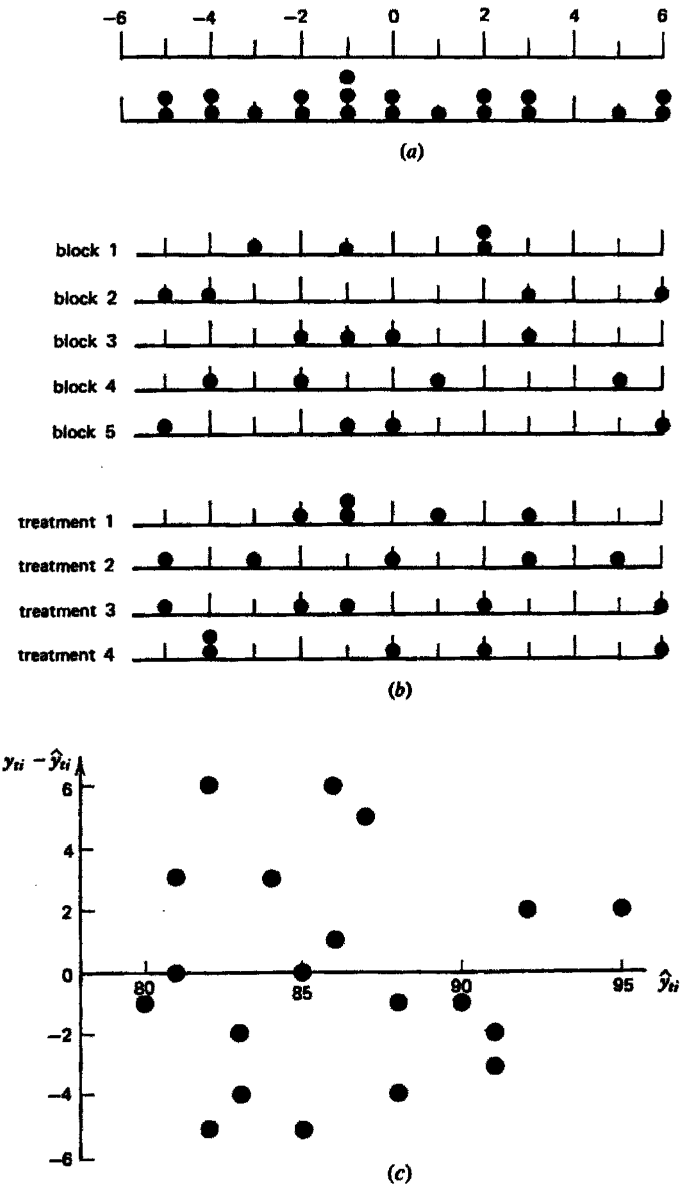


Figure 7.11. Plots of residuals, yield example: (a) overall plot; (b) plots by block and treatment; (c) $y_{ti} - \hat{y}_{ti}$ versus $\hat{y}_{ti}$ [1].

also be indicative of *nonadditivity* between the block and treatment effects. Such discrepancies can sometimes be eliminated using a suitable *transformation* of the response variable. Data transformations are discussed in greater detail in Section 7.2.4.

7.2.4. Two-Way Designs

Experimental designs with two sets of treatments (or *factors*) are referred to as *two-way designs*, and their corresponding analysis is accomplished by *two-way ANOVA*. As an example, consider the data in Table 7.11, which corresponds to the film uniformity (in %) achieved in a CVD experiment. Assume that treatments A, B, C, and D represent different gas compositions, and treatments 1, 2, and 3 represent different temperatures. This arrangement, which has been replicated four times, is also known as a 3 × 4 *factorial design* (see Section 7.3). There is no blocking. Both factors are of equal interest, and there is a possibility that these factors interact.

7.2.4.1. Analysis

Let y_{tij} be the nonuniformity of the jth wafer deposited at the ith temperature using the tth gas composition. The corresponding model and estimate for the nonuniformity response are given respectively by

$$y_{tij} = \eta_{ti} + \varepsilon_{ti} \tag{7.23}$$

$$= \overline{y}_{ti} + (y_{tij} - \overline{y}_{ti}) \tag{7.24}$$

If the temperature and gas composition effects do not behave additively, then

$$\eta_{ti} = \eta + \tau_t + \beta_i + \omega_{ti} \tag{7.25}$$

where τ_t is the incremental nonuniformity associated with the tth gas composition, and β_i is the increment associated with the ith temperature. Their non-additive behavior requires an additional term, ω_{ti}, which represents the *interaction effect* between the two factors. The τ_t and β_i terms are known as the *main effects*. The estimate corresponding to Eq. (7.25) is given by

$$\overline{y}_{ti} = \overline{y} + (\overline{y}_t - \overline{y}) + (\overline{y}_i - \overline{y}) + (\overline{y}_{ti} - \overline{y}_t - \overline{y}_i + \overline{y}) \tag{7.26}$$

Table 7.11. Nonuniformity (in %) of films grown by CVD [1].

Treatments	A	B	C	D
1	0.31	0.82	0.43	0.45
	0.45	1.10	0.45	0.71
	0.46	0.88	0.63	0.66
	0.43	0.72	0.76	0.62
2	0.36	0.92	0.44	0.56
	0.29	0.61	0.35	1.02
	0.40	0.49	0.31	0.71
	0.23	1.24	0.40	0.38
3	0.22	0.30	0.23	0.30
	0.21	0.37	0.25	0.36
	0.18	0.38	0.24	0.31
	0.23	0.29	0.22	0.33

The arithmetic for carrying out the data analysis closely parallels that used for the randomized block design. Group averages ($\overline{y}_{ti}$) replace the basic data, and an interaction sum of squares replaces the residual sum of squares in the randomized block analysis. In general, for n levels of some factor P ($n = 3$ temperatures in this example), k levels of another factor T ($k = 4$ gas compositions), and m replications ($m = 4$ wafers per group), the following sums of squares may be defined:

$$S_P = mk \sum_i (\overline{y}_i - \overline{y})^2 \tag{7.27}$$

$$S_T = mn \sum_t (\overline{y}_t - \overline{y})^2 \tag{7.28}$$

$$S_I = m \sum_t \sum_i (\overline{y}_{ti} - \overline{y}_t - \overline{y}_i + \overline{y})^2 \tag{7.29}$$

$$S_e = \sum_t \sum_i \sum_j (y_{tij} - \overline{y}_{ti})^2 \tag{7.30}$$

$$S = \sum_t \sum_i \sum_j (y_{tij} - \overline{y})^2 \tag{7.31}$$

Given these parameters, the ANOVA table for the nonuniformity data in Table 7.11 is shown in Table 7.12.

7.2.4.2. Data Transformation

If the model described by Eq. (7.23) is accurate and the model errors are independently and normally distributed with a constant variance [i.e., $\varepsilon_{ti} \sim N(0, \sigma^2)$], then the significance of the factors can be evaluated using the F distribution. In the CVD example, examination of Table 7.12 reveals that the effects of both the temperature and the gas composition are highly significant. For example, in the case of temperature, an F ratio of 23.27 for an F distribution with $\nu_p = 2$ and $\nu_e = 36$ degrees of freedom has less than a 0.001 significance level. This analysis of variance also indicates some suggestion of interaction between temperature and gas composition. The F ratio of 1.88 for an F distribution with $\nu_I = 6$ and $\nu_e = 36$ degrees of freedom has an approximate 0.01 significance level.

Table 7.12. ANOVA table for two-way factorial experiment.

Source of Variation	Sum of Squares (× 1000)	Degrees of Freedom	Mean Square	F Ratio
Temperatures	$S_P = 1033.0$	$\nu_P = n - 1 = 2$	$s_P^2 = S_P/\nu_P = 516.5$	$s_P^2/s_e^2 = 23.27$
Gas compositions	$S_T = 922.4$	$\nu_T = k - 1 = 3$	$s_T^2 = S_T/\nu_T = 307.5$	$s_T^2/s_e^2 = 13.85$
Interaction	$S_I = 250.1$	$\nu_I = (n-1)(k-1) = 6$	$s_I^2 = S_I/\nu_I = 41.7$	$s_I^2/s_e^2 = 1.88$
Error	$S_e = 800.7$	$\nu_e = nk(m-1) = 36$	$s_e^2 = S_e/\nu_e = 22.2$	
Total	$S = 3006.2$	$\nu = nkm - 1 = 47$		

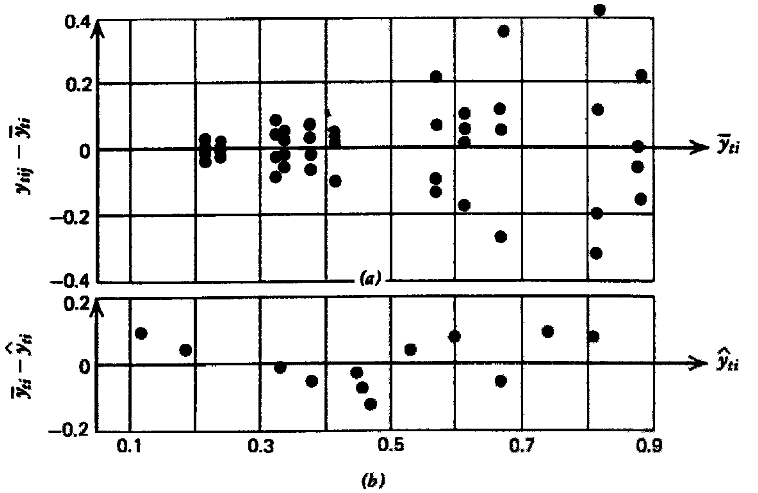


Figure 7.12. Residual diagnostics, CVD experiment: (a) plot of $y_{tij} - \bar{y}_{ti}$ versus $\bar{y}_{ti}$; (b) plot of $\bar{y}_{ti} - \hat{y}_{ti}$ versus $\hat{y}_{ti}$ [1].

Diagnostic checking of the residuals for these data, however, leads to suspicion that this model is inadequate. Figure 7.12 shows a plot of the residuals versus $\bar{y}_{ti}$. The funnel shape in Figure 7.12a suggests that the standard deviation of the data is not constant as previously assumed, but instead increases with the mean. Furthermore, ignoring the interaction term by letting $\hat{y}_{ti} = \bar{y}_t + \bar{y}_i - \bar{y}$, and plotting $\bar{y}_{ti} - \hat{y}_{ti}$ against $\hat{y}_{ti}$ reveals a curvilinear relationship, which contradicts the linearity assumption (Figure 7.12b).

In cases such as this, when σ_y is actually a function of the mean (η), it may be possible to find a convenient data transformation $Y = f(y)$ that does have a constant variance. If so, the data are said to possess a *transformable nonadditivity*. For example, suppose that σ_y is proportional to some power of η, or

$$\sigma_y \propto \eta^y \tag{7.32}$$

and the following power transformation of the data is made:

$$Y = y^\lambda \tag{7.33}$$

Then

$$\sigma_y = \theta\sigma_y \propto \theta\eta^\alpha \tag{7.34}$$

where θ is the gradient of the graph of Y versus y (see Figure 7.13). It can be shown that if Eq. (7.33) is true, then $\theta \propto \eta^{\lambda-1}$. Thus

$$\sigma_Y \propto \eta^{\lambda-1}\eta^\alpha = \eta^{\lambda+\alpha-1} \tag{7.35}$$

Therefore, Y is chosen so that σ_y does not depend on η if $\lambda = 1 - \alpha$. Some values of α with appropriate variance stabilizing transformations are presented in Table 7.13.

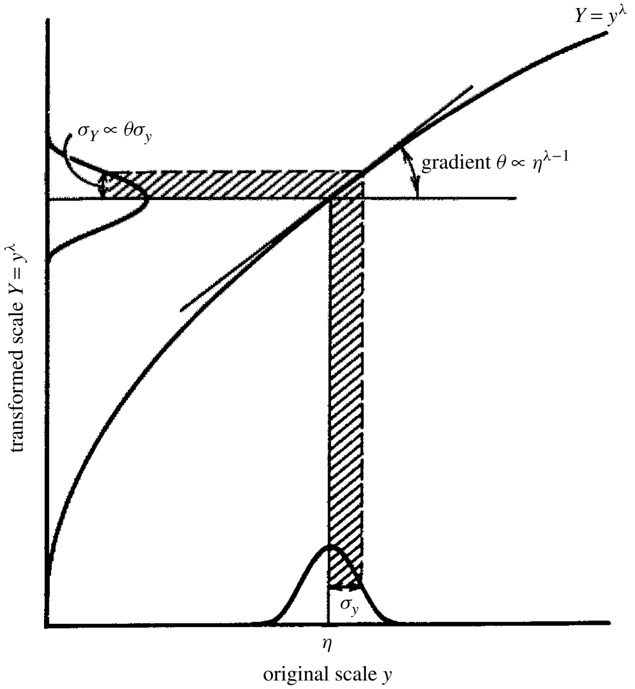

$Y = y^{\lambda}$
$\sigma_Y \propto \theta\sigma_y$
gradient $\theta \propto \eta^{\lambda-1}$
transformed scale $Y = y^{\lambda}$
σ_y
η
original scale y

Figure 7.13. Data transformation from y to $Y = y^{\lambda}$[1].

Table 7.13. Variance stabilizing data transformations when $\sigma_y \propto \eta^{\alpha}$ [1].

Dependence of σ_y on η	α	$\lambda = 1 - \alpha$	Variance Stabilizing Transformation
$\sigma \propto \eta^2$	2	-1	Reciprocal
$\sigma \propto \eta^{\frac{3}{2}}$	$\frac{3}{2}$	$-\frac{1}{2}$	Reciprocal square root
$\sigma \propto \eta$	1	0	Log
$\sigma \propto \eta^{\frac{1}{2}}$	$\frac{1}{2}$	$\frac{1}{2}$	Square root
$\sigma \propto$ constant	0	1	None

In order to identify an appropriate transformation for the data in the CVD experiment, α must first be determined empirically. Since $\sigma_y \propto \eta^{\alpha}$, it is also true that $\log \sigma_y = \text{constant} + \alpha \log \eta$. Thus, if we plot σ_y versus $\log \eta$, we obtain a straight line with a slope of α. Although σ_y and η are not known in practice, they may be estimated using s and $\overline{y}$, respectively (where s is the sample standard deviation of the data). Carrying out this procedure for the CVD data yields a slope of $\alpha \cong 2$. From Table 7.13, this implies that a reciprocal transformation is appropriate in this case. We therefore convert the entire dataset in Table 7.11 into reciprocals and repeat the analysis of variance.

Table 7.14. ANOVA for transformed and untransformed data, CVD experiment [1].

Source of Variation	Untransformed Degrees of Freedom	Untransformed Mean Square ($\times$ 1000)	Transformed ($Y = y^{-1}$) Degrees of Freedom	Transformed ($Y = y^{-1}$) Mean Square ($\times$ 1000)
Temperatures	2	516.5	2	1743.9
Gas compositions	3	307.5	3	680.5
Interaction	6	41.7	6	26.2
Error	36	22.2	35	24.7

A comparison of the ANOVA for the untransformed and transformed datasets is provided in Table 7.14. The fact that the data themselves have been used to choose the transformation is accounted for by reducing the number of degrees of freedom in the mean square of the error (from 36 to 35) [2]. The effects of the transformation are noteworthy. The mean squares for the transformed data are now much larger relative to the error, indicating an increase in sensitivity of the experiment. In addition, the interaction mean square, which previously gave a slight indication of statistical significance, is now closer in size to the error, contradicting that assertion. Verification of the improvement in the residual diagnostics for the transformed data is left as an exercise.

7.3. FACTORIAL DESIGNS

Experimental design is essentially an organized method of conducting experiments in order to extract the maximum amount of information from a limited number of experiments. Experimental design techniques are employed in semiconductor manufacturing applications to systematically and efficiently explore the effects of a set of input variables, or *factors* (such as processing temperature), on *responses* (such as yield). The unifying feature in statistically designed experiments is that all factors are varied simultaneously, as opposed to the more traditional "one variable at a time" technique. A properly designed experiment can minimize the number of experimental runs that would otherwise be required if this approach or random sampling was used.

Factorial experimental designs are of great practical importance for manufacturing applications. To perform a factorial experiment, an investigator selects a fixed number of *levels* for each of a number of variables (factors) and runs experiments at all possible combinations of the levels. If there are i levels for the first variable, j levels for the second, ..., and k levels for the kth, the complete set of $i \times j \times \cdots \times k$ experimental trials is called an $i \times j \times \cdots \times k$ factorial design. As mentioned previously, the data presented in Table 7.11 represent a 3×4 factorial design. In general, an $l \times m \times n$ design requires lmn runs. For example, a $2 \times 3 \times 4$ design requires 24 runs.

Two of the most important issues in factorial experimental designs are choosing the set of factors to be varied in the experiment and specifying the ranges over which variation will take place. The choice of the number of factors directly impacts the number of experimental runs (and therefore the overall cost of the experiment). The most common approach in factorial designs is the two-level factorial, which is described in Section 7.3.1.

7.3.1. Two-Level Factorials

The ranges of the process variables investigated in factorial experiments can be discretized into minimum, maximum, and "center" levels. In a *two-level factorial design*, the minimum and maximum levels of each factor (normalized to take on values -1 and $+1$, respectively) are used together in every possible combination. Thus, a full two-level factorial experiment with n factors requires 2^n experimental runs. The various factor level combinations of a three-factor experiment can be represented pictorially as the vertices of a cube, as shown in Figure 7.14.

Two-level factorial designs are important for several reasons. First, although they require relatively few trials per factor, they allow an experimenter to identify major trends and promising directions for future experimentation. These designs are also easily augmented to form more advanced designs (see Section 7.3.4). Furthermore, two-level factorials form the basis for two-level fractional factorial designs (see Section 7.3.2), which are useful for screening large numbers of factors at an early stage of experimentation. Finally, analysis of these designs facilitates the systematic analysis of the impact of *interactions* between factors. Such interactions can be obscured if the traditional "change one variable at a time" approach to experimentation, in which factors are varied individually while the remaining factors are held constant, is used. The traditional approach assumes that all of the factors act on the response additively, which is often not the case in complex processes. In addition, the factorial approach is more economical, since a n-factor traditional experiment requires a n-fold increase in the number of trials as compared to a 2^n factorial experiment.

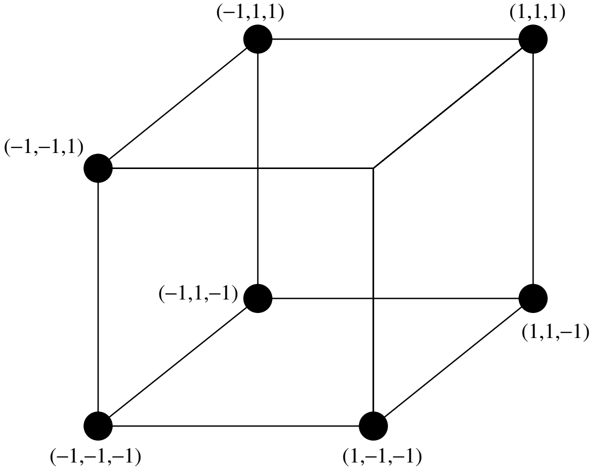


Figure 7.14. Factor combinations for a three-factor experiment represented as vertices of a cube.

7.3.1.1. Main Effects

To illustrate the use of two-level factorials, Table 7.15 shows a 2^3 factorial experiment for another CVD process. The three factors are temperature (T), pressure (P), and gas flowrate (F). The response being measured is the deposition rate (D) in angstroms per minute (Å/min). The highest and lowest levels of each factor are represented by the "+" and "−" signs, respectively. The display of levels depicted in the first three columns of this table is called a *design matrix.*

The relevant issue is what we can determine from this factorial design. For example, what do the data collected tell us about the effect of pressure on deposition rate? The effect of any single variable on the response is called a *main effect.* The method used to compute such a main effect is to find the difference between the average deposition rate when the pressure is high (i.e., runs 2, 4, 6, 8) and the average deposition rate when the pressure is low (runs 1, 3, 5, 7). Mathematically, this is expressed as

$$P = d_{p+} - d_{p-} = \tfrac{1}{4}[(d_2 + d_4 + d_6 + d_8) - (d_1 + d_3 + d_5 + d_7)] = 40.86 \tag{7.36}$$

where P is the main effect for pressure, d_{p+} is the average deposition rate when the pressure is high, and d_{p-} is the average deposition rate when the pressure is low. The manner in which we interpret this result is that the average effect of increasing pressure from its lowest to its highest level is to increase the deposition rate by 40.86 Å/min. The other main effects for temperature and flowrate are computed in a similar manner. In general, the main effect for each variable in a two-level factorial experiment is the difference between the two averages of the response (y), or

$$\text{(Main effect)} = y_+ - y_- \tag{7.37}$$

7.3.1.2. Interaction Effects

We might also be interested in quantifying how two or more factors interact. For example, suppose that the pressure effect is much greater at high temperatures than it is at low temperatures. A measure of this interaction is provided by the difference between the average pressure effect with temperature high and the average pressure effect with temperature low. By convention, *half* of

Table 7.15. 2-Level Factorial CVD Experiment.

Run	P	T	F	D (Å/min)
1	−	−	−	$d_1 = 94.8$
2	+	−	−	$d_2 = 110.96$
3	−	+	−	$d_3 = 214.12$
4	+	+	−	$d_4 = 255.82$
5	−	−	+	$d_5 = 94.14$
6	+	−	+	$d_6 = 145.92$
7	−	+	+	$d_7 = 286.71$
8	+	+	+	$d_8 = 340.52$

this difference is called the *pressure by temperature interaction,* or symbolically, the $P \times T$ interaction. This interaction may also be thought of as one-half the difference in the average temperature effects at the two levels of pressure. Mathematically, this is

$$P \times T = d_{PT+} - d_{PT-} = \tfrac{1}{4}[(d_1 + d_4 + d_5 + d_8) - (d_2 + d_3 + d_6 + d_7)] = 6.89 \tag{7.38}$$

The $P \times F$ and $T \times F$ interactions are computed in a similar fashion. Just as main effects can be viewed as a *contrast* between observations on faces of a cube like the one in Figure 7.14 (see Figure 7.15a), an interaction is a contrast between results on two diagonal planes (Figure 7.15b).

Finally, we might also be interested in the interaction of all three factors, denoted as the *pressure by temperature by flowrate* or the $P \times T \times F$ interaction. This interaction defines the average difference between any two-factor interaction at the high and low levels of the third factor. It is given by

$$P \times T \times F = d_{PTF+} - d_{PTF-} = -5.88 \tag{7.39}$$

This interaction is depicted graphically in Figure 7.15c. It is important to note that the main effect of any factor can be individually interpreted only if there is no evidence that the factor interacts with other factors.

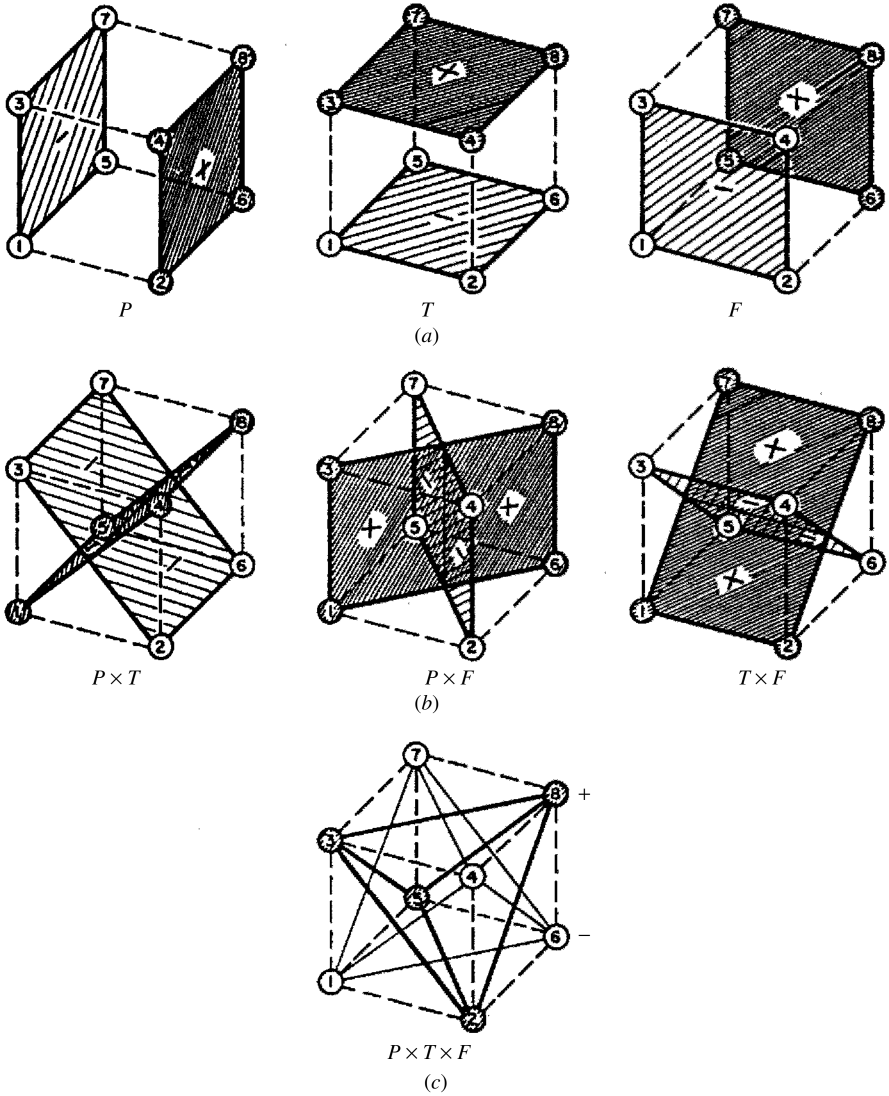
7.3.1.3. Standard Error

When valid run replicates are made under a given set of experimental conditions, the variation between associated observations can be used to estimate the standard deviation of a single observation, and hence the standard deviation (or *standard error*) of the effects. A comparison of the size of an effect to its standard error allows one to determine the significance of the effect relative to experimental error or noise. In other words, an effect that is much larger than its standard error (in an absolute sense) is more likely to be significant, as opposed to an effect that is less than or equal to its standard error. The notion of "validity" in the context is usually accomplished by randomization of the run order. Randomization helps ensure that the variation between runs made at the same experimental conditions reflects the total variability that can be ascribed to runs made under different experimental conditions.

If there are r sets of experimental conditions replicated, and the n_i replicate runs made at the ith set provide an estimate s_i^2 of the true variance (σ^2) having $\nu_i = n_n - 1$ degrees of freedom, then the pooled estimate of the run variance is

$$s^2 = \frac{\nu_1 s_1^2 + \nu_2 s_2^2 + \cdots + \nu_r s_r^2}{\nu_1 + \nu_2 + \cdots + \nu_r} \tag{7.40}$$

with $\nu = \nu_1 + \nu_2 + \cdots + \nu_r$ degrees of freedom. If there are only $n_i = 2$ replicates at each of the r sets of conditions, then the formula for the ith variance reduces to $s_i^2 = d_i^2/2$ with $\nu_i = 1$, where d_i is the difference between the duplicate observations at the ith set of conditions. From (7.40), this implies $s^2 = \sum d_i^2/2r$.



P *T* *F*

(*a*)

P × *T* *P* × *F* *T* × *F*

(*b*)

P × *T* × *F*

(*c*)

Figure 7.15. Geometric representation of contrasts corresponding to main effects (a) and two- (b) and three-factor (c) interactions [1].

Since each main effect and interaction in a two-level factorial experiment are statistics of the form $= y_+ - y_-$a, the overall variance of each effect (assuming independent errors) is given by

$$V(\text{effect}) = V(\overline{y}_+ - \overline{y}_-) = \frac{4}{N}\sigma^2 \tag{7.41}$$

where N is the total number of runs made in conducting the factorial design or replicated factorial design and σ^2 is estimated using s^2. The standard error may be computed by taking the square root of V(effect). Equation (7.41) implies that conducting larger numbers of experiments can reduce the variance in our estimates of the effects.

7.3.1.4. Blocking

The term "blocking" refers to a systematic methodology used to eliminate the effects of parameters that the experimenter cannot control. As an example, consider once again the 2^3 factorial design used in the CVD experiment discussed in Section 7.3.1.1. Suppose that the CVD reactor needed to be cleaned every four runs. This means that each group of four runs occurs under a different set of experimental conditions. Table 7.16 shows how the 2^3 factorial design can be arranged in two blocks of four runs to neutralize the effect of reactor cleaning.

The design is blocked in this way by placing all runs in which the "product" of columns P, T, and F is minus in block 1, and all other runs are placed in block 2. This arrangement eliminates the spurious effect of cleaning since if the deposition rate of all the runs in block 2 were higher by some amount Δd than they would have been if they had been performed in block 1, then no matter what the value of Δd is, it will cancel out in the calculation of effects P, T, F, *PT, PF*, and *TF*.

Note that a tradeoff in the information that can be derived from this experiment has occurred under this blocking arrangement. The three-factor interaction effect *PTF* has now been *confounded* (i.e., "confused") with the block effect. Therefore, using this blocking scheme, we are now unable to independently estimate this interaction. However, it is usually assumed that higher-order interactions such as this can be neglected. In exchange, this design ensures that main effects and two-factor interactions can be more precisely measured than would be the case in the absence of blocking.

It is common practice for such a design to assign a numerical symbol to each column. In other words, $P = \mathbf{1}$, $T = \mathbf{2}$, and $F = \mathbf{3}$. Using this terminology, we can assign the block variable the numerical identifier **4**. Then we can think of the experiment as having four variables, the latter of which does not interact with the other three. If the new variable is produced by having its plus and minus signs correspond to the signs of the **123** interaction, then the blocking is said to be *generated* by the relationship $\mathbf{4} = \mathbf{123}$.

Table 7.16. A 2^3 factorial design in blocks of size 2.

Run	*P*	*T*	*F*	*PT*	*PF*	*TF*	*PTF*	Block
1	−	−	−	+	+	+	−	1
2	+	−	−	−	−	+	+	2
3	−	+	−	−	+	−	+	2
4	+	+	−	+	−	−	−	1
5	−	−	+	+	−	−	+	2
6	+	−	+	−	+	−	−	1
7	−	+	+	−	−	+	−	1
8	+	+	+	+	+	+	+	2

Table 7.17. Trial blocking scheme for 2^3 CVD experiment for blocks of size 2.

4 = 123	5 = 23	Block	45
−	+	2	−
+	+	4	+
+	−	3	−
−	−	1	+
+	−	3	−
−	−	1	+
−	+	2	−
+	+	4	+

Suppose instead in the 2^3 factorial CVD experiment that the reactor had to be cleaned every two runs. This would require four blocks of two runs, rather than two blocks of four, which means that two block generators are also required. Suppose that we initially selected the block generators $\mathbf{4} = \mathbf{123}$ and $\mathbf{5} = \mathbf{23}$. The resulting design is shown in Table 7.17. As a consequence of this arrangement, the *PTF* and *TF* effects are clearly confounded. However, there is an additional unintended consequence as well. Note in Table 7.17 that the column that represents the product of the two block generators is identical to the column for the *P* main effect! This means that this blocking scheme prevents us from identifying this main effect, which is clearly unacceptable.

Fortunately, there is a simple method to identify confounding patterns and evaluate the consequences of any proposed blocking scheme so that situations like this can be avoided. Let I be a column consisting entirely of plus signs. Thus, we can write

$$I = \mathbf{11} = \mathbf{22} = \mathbf{33} = \cdots \text{etc.} \tag{7.42}$$

where **11, 22**, and **33** represent the product of the elements in columns **1, 2**, and **3**, respectively, with themselves. The effect of multiplying the elements of any column with I is to leave those elements unchanged. Now in the blocking arrangement just considered, the product of the block generators is

$$\mathbf{45} = \mathbf{123} \times \mathbf{23} = \mathbf{12233} = \mathbf{1}II = \mathbf{1} \tag{7.43}$$

which indicates that **45** is identical to column **1**, thereby clarifying the confounding inherent in this scheme.

This suggests a better blocking scheme to achieve four blocks of size two in this experiment. If we let $\mathbf{4} = \mathbf{12}$ and $\mathbf{5} = \mathbf{13}$, the *PT* and *PF* interactions are clearly confounded. Also, since $\mathbf{45} = \mathbf{12} \times \mathbf{13} = \mathbf{1123} = I\mathbf{23} = \mathbf{23}$, the *TF* interaction is also confounded. However, this new blocking arrangement has the advantage of not confounding any main effect, which is much more desirable than the previous scheme.

7.3.2. Fractional Factorials

A major disadvantage of the two-level factorial design is that the number of experimental runs increases exponentially with the number of factors. To alleviate this concern, *fractional factorial* designs are often constructed by systematically eliminating some of the runs in a full factorial design. For example, a half fractional design with n factors requires only 2^{n-1} runs. Full or fractional two-level factorial designs can be used to estimate the main effects of individual factors as well as the interaction effects between factors. However, they cannot be used to estimate quadratic or higher-order effects. This is not a serious shortcoming, since higher order effects and interactions tend to be smaller than low-order effects (main effects tend to be larger than two-factor interactions, which tend to be larger than three-factor interactions, etc.). Ignoring high-order effects is conceptually similar to ignoring higher-order terms in a Taylor series expansion.

7.3.2.1. Construction of Fractional Factorials

To illustrate the use of fractional factorial designs, let $n = 5$ and consider a 2^5 factorial design. The full factorial implementation of this design would require 32 experimental runs. However, a 2^{5-1} fractional factorial design requires only 16 runs. This 2^{5-1} design is generated by first writing the design matrix for a 2^4 full factorial design in standard order. Then plus and minus signs in the four columns of the 2^4 design matrix are each "multiplied" together to form a fifth column (i.e., **5** = **1234**).

However, just as in the case of blocking, some information is lost in this fractional factorial arrangement. A 2^{5-1} design allows the estimation of 16 quantities: the mean, the 5 main effects, and the 10 two-factor interactions. The higher-order effects (the 10 three-factor interactions, 5 four-factor interactions, and single five-factor interaction) are now confounded with one of the first 16 effects. To illustrate, consider the **45** and **123** interactions. These yield the identical design sequences:

$$\mathbf{45} = - + + - + - - + - + + - + - - +$$

$$\mathbf{123} = - + + - + - - + - + + - + - - +$$

thereby indicating that these two interactions are confounded. If the 16 outputs of this 2^{5-1} fractional factorial experiment are labeled $y_1, \ldots, y_{16}$, then the symbol l_{45} denotes the linear function of these observations used estimate the **45** interaction, or

$$l_{45} = \tfrac{1}{8}(-y_1 + y_2 + y_3 - y_4 + y_5 - y_6 - y_7 + y_8 - y_9 + y_{10} + y_{11} - y_{12} + y_{13} - y_{14} - y_{15} + y_{16}) \quad (7.44)$$

The symbol l_{45} is called a *contrast*, since it is the difference between two averages of eight results. Since interactions **45** and **123** are confounded, the contrast l_{45} actually estimates the sum of the mean values of their effects, which is indicated by the notation $l_{45} \rightarrow 45 + 123$. However, if we accept the convention that higher-order interactions are generally less significant than lower order interactions, we would attribute the numerical value of contrast l_{45} primarily to the **45** interaction.

Recall that the 2^{5-1} design was constructed by setting $\mathbf{5} = \mathbf{1234}$. This relation is called the generator of the design. Multiplying both sides of the relation by **5**, we obtain

$$\mathbf{5} \times \mathbf{5} = \mathbf{1234} \times \mathbf{5} \tag{7.45}$$

or equivalently, $I = \mathbf{12345}$. The latter relation is called the *defining relation* of the fractional factorial design. The defining relation is the key to determining the confounding pattern of the design. For example, multiplying the defining relation on both sides by **1** yields $\mathbf{1} = \mathbf{2345}$, which indicates that main effect **1** is confounded with the 4-factor interaction **2345**.

Let's take another look at our CVD experiment. Suppose we only have the time and/or resources available to perform four deposition experiments, rather than the eight required for a 2^3 full factorial design. This calls for a 2^{3-1} fractional factorial alternative. This new design could be generated by writing the full 2^2 design for the pressure and temperature variables, and then multiplying those columns to obtain a third column for flowrate. This procedure is illustrated in Table 7.18. The only drawback in using this procedure is that since we have used the PT relation to define column F, we can no longer distinguish between the effects of the $P \times T$ interaction and the F main effect. These effects are therefore confounded.

7.3.2.2. Resolution

A fractional factorial design of *resolution* R is one in which no p-factor interaction is confounded with any other effect containing less than $R - p$ factors. The resolution of a design is denoted by a Roman numeral and appended as a subscript. For example, the 2^{5-1} fractional factorial discussed in the previous section is called a *resolution V* design, and is denoted as 2_V^{5-1}. In this case, main effects are confounded with four-factor interactions, and two-factor interactions are confounded with three-factor interactions. In general, the resolution of a two-level fractional factorial design is just the length of the defining relation.

7.3.3. Analyzing Factorials

Although various methods for analysis of factorial experiments that are based on simple hand calculations exist, it should be pointed out that modern analysis

Table 7.18. Illustration of 2^{3-1} fractional factorial design for CVD example.

Run	P	T	F
1	−	−	+
2	+	−	−
3	−	+	−
4	+	+	+

of statistical experiments is accomplished almost exclusively by commercially available statistical software packages. A few of the more common packages include *RS/1, SAS*, and *Minitab*. These packages completely alleviate the necessity of performing any tedious hand calculations. Nevertheless, a few of the more well-known hand methods are presented below.

7.3.3.1. The Yates Algorithm

It is quite tedious to calculate the effects and interactions for two-level factorial experiments using the methods described Section 7.3.1, particularly if there are more than three factors involved. Fortunately, the *Yates algorithm* provides a quicker method of computation that is also relatively easily programmed via computer. To implement this algorithm, the experimental design matrix is first arranged in what is called *standard order*. A 2^n factorial design is in standard order when the first column of the design matrix consists of alternating minus and plus signs, the second column of successive pairs of minus and plus signs, the third column of four minus signs followed by four plus signs, and so on. In general, the kth column consists of 2^{k-l} minus signs followed by 2^{k-1} plus signs.

The Yates calculations for the deposition rate data are shown in Table 7.19. Column y contains the deposition rates for each run. These are considered in successive pairs. The first four entries in column 1 are obtained by adding the pairs together, and the next four are obtained by subtracting the *top number from the bottom number* of each pair. Column 2 is obtained from column 1 in the same way, and column 3 is obtained from column 2. To obtain the experimental *effects*, one only needs to divide the column 3 entries by the *divisor*. In general, the first divisor will be 2^n, and the remaining divisors will be 2^{n-1}. The first element in the *identification* (ID) column is the grand average of all of the observations, and the remaining identifications are derived by locating the plus signs in the design matrix. The Yates algorithm provides a relatively straightforward methodology for computing experimental effects in two-level factorial designs.

7.3.3.2. Normal Probability Plots

One problem that can arise when analyzing the effects of unreplicated factorial experiments is that real and meaningful higher-order interactions do occasionally

Table 7.19. Illustration of the Yates algorithm.

P	T	F	y	(1)	(2)	(3)	Divisor	Effect	ID
−	−	−	94.8	205.76	675.70	1543.0	8	192.87	Average
+	−	−	110.96	469.94	867.29	163.45	4	40.86	P
−	+	−	214.12	240.06	57.86	651.35	4	162.84	T
+	+	−	255.82	627.23	105.59	27.57	4	6.89	PT
−	−	+	94.14	16.16	264.18	191.59	4	47.90	F
+	−	+	145.92	41.70	387.17	47.73	4	11.93	PF
−	+	+	286.71	51.78	25.54	122.99	4	30.75	TF
+	+	+	340.52	53.81	2.03	−23.51	4	−5.88	PTF

occur. In such cases, methods are needed to evaluate these effects. One way to do so is to plot the effects on *normal probability paper*.

A normal distribution is shown in Figure 7.16a. The probability of the occurrence of some value less than X is given by the shaded area P. Plotting P versus X results in the sigmoidal cumulative normal distribution curve shown in Figure 7.16b. Normal probability paper simply adjusts the vertical scale of this plot in the manner shown in Figure 7.16c, so that P versus X becomes a straight line.

Suppose that the dots in Figure 7.16 represent a random sample of 10 observations from a normal distribution. Since $n = 10$, the leftmost observation can be interpreted as representing the first 10% of the cumulative distribution. In Figure 7.16b, this observation is therefore plotted midway between zero and 10% (i.e., at 5%). Similarly, the second observation represents the next 10% of the cumulative distribution and is plotted at 15%, and so on. In general, we have

$$P_i = 100(i - 1/2)/m \tag{7.46}$$

for $i = 1, 2, \ldots, m$.

When all the sample points are plotted on normal paper, they should ideally form a straight line. However, this is only true if the effects represented by the points are *not* significant. To illustrate, consider the effects computed from a hypothetical 2^4 factorial experiment shown in Table 7.20. The $m = 15$ main effects plus interactions in this experiment represent 15 contrasts between pairs of averages containing eight observations each. If these effects are not significant,

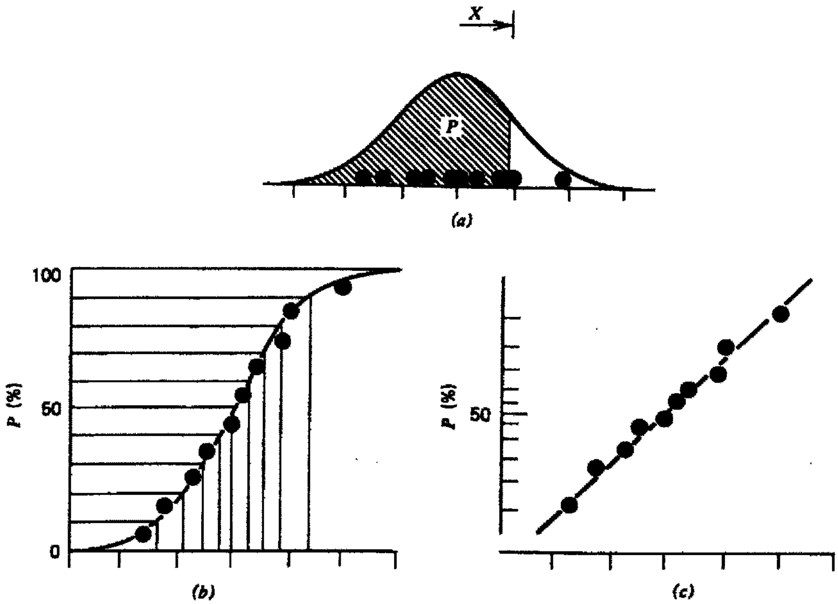


Figure 7.16. Normal probability plot concepts: (a) normal distribution; (b) ordinary graph paper; (c) normal probability paper [1].

Table 7.20. Effects and probability points for normal probability plot example [1].

i	Value of Effect	Identity of Effect	$P = 100(i - \frac{1}{2})/15$
1	−8.0	**1**	3.3
2	−5.5	**4**	10.0
3	−2.25	**3**	16.7
4	−1.25	**23**	23.3
5	−0.75	**123**	30.0
6	−0.75	**234**	36.7
7	−0.25	**34**	43.3
8	−0.25	**134**	50.0
9	−0.25	**1234**	56.7
10	0	**14**	63.3
11	0.5	**124**	70.0
12	0.75	**13**	76.7
13	1.0	**12**	83.3
14	4.5	**24**	90.0
15	24.0	**2**	96.7

they should be roughly normally distributed about zero, and they would plot on normal probability paper as a straight line. To see whether they do, we put the effects in order and plot them on normal paper, as shown in Figure 7.17. As it turns out 11 of the 15 effects fit reasonably well on a straight line, but those representing effects **1**, **4**, **23**, and **2** do not. We therefore conclude that these effects cannot be explained by chance and are in fact significant.

7.3.4. Advanced Designs

Factorial and fractional factorial designs are used for fitting either linear response models or models based on factor interactions to the experimental data (see Chapter 8). When higher-order models are necessary, more advanced experimental designs are required. One example of such a design is the *central composite design* (CCD), which is used for fitting second-order models. These designs are widely used because of their relative efficiency with respect to the number of trials required.

In a CCD, the standard two-level factorial "box" is enhanced by replicated experiments at the center of the design space (called *centerpoints*), as well as by symmetrically located *axial points*. Thus, a complete CCD with k factors requires 2^k factorial runs, $2k$ axial runs, and 3–5 centerpoints. The centerpoints provide a direct measure of the experimental replication error, and the axial points facilitate fitting of the second-order responses. Designs for $k = 2$ and $k = 3$ are shown in Figure 7.18.

The CCD can be made *rotatable* by the proper choice of the axial spacing (α in Figure 7.18). Rotatability implies that the standard deviation of the predicted response is constant at all points equidistant from the center of the design. To

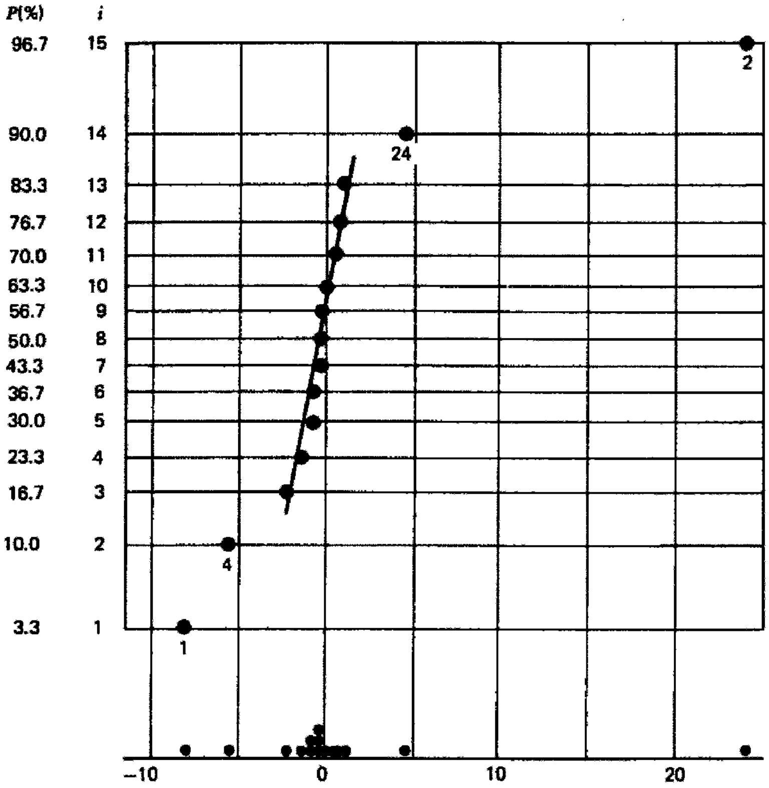


Figure 7.17. Normal probability plot example [1].

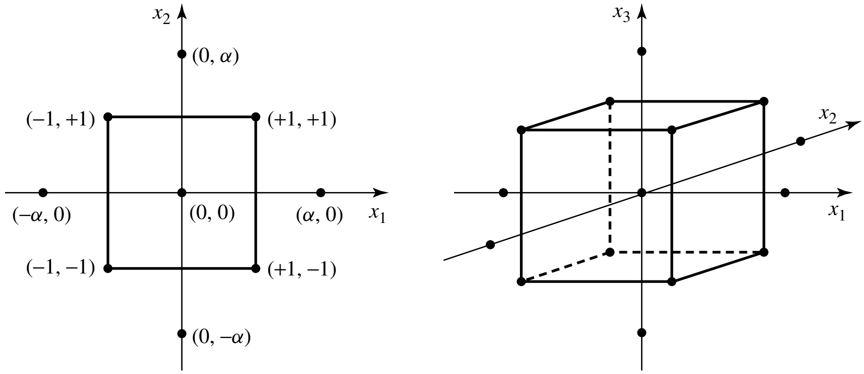


Figure 7.18. Central composite designs for $k = 2$ and $k = 3$ [3].

ensure rotatability, we select

$$\alpha = (2^k)^{1/4} \tag{7.47}$$

For the case of $k = 2$, $\alpha = 1.414$. This is the case represented on the left in Figure 7.18.

7.4. TAGUCHI METHOD

Until relatively recently, the use of statistical experimental design has not been as prominent in the West as in Japan. The widespread use of statistical methods in Japanese manufacturing can be traced directly to the contributions of Professor Genichi Taguchi. In the early 1980s, Taguchi introduced an approach to using experimental design to develop products that are robust to environmental conditions and process variation [4].

Taguchi outlines three critical stages in process development: system design, parameter design, and tolerance design. *System design* essentially refers to establishing the basic configuration of the manufacturing sequence and equipment. In *parameter design*, specific values of process recipe parameters are determined, with the overall objective of minimizing the variability generated by uncontrollable (or *noise*) variables. Finally, *tolerance design* is used to identify the tolerances of the manufacturing parameters. Variables without much effect on product performance can be specified with a wide tolerance.

Taguchi advocates the use of experimental design to facilitate quality improvement primarily during the parameter design and tolerance design stages. Experimental design methods are used to identify a process that is *robust* (i.e., insensitive) to uncontrollable environmental factors. Thus, a key component of the Taguchi approach is reduction of variability. The objective is to reduce the variability of a quality characteristic around a target, or *nominal*, value. Differences between actual and nominal values are described by a *loss function*. The loss function quantifies the cost incurred by society when a consumer uses a product whose quality characteristics differ from nominal values. Taguchi defines a quadratic loss function of the form

$$L(y) = k(y - T)^2 \tag{7.48}$$

which is shown in Figure 7.19. In this function, y represents the measured value of the quality characteristic, T is the target value, and k is a constant. This function penalizes even small excursions from the target value, as opposed to the traditional control chart-oriented approach, which attaches penalties only when y is outside of specification limits.

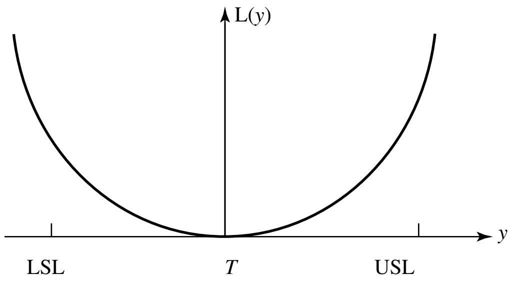


Figure 7.19. Quadratic loss function.

Taguchi's overall philosophy can be summarized by three central ideas:

1. Products and processes should be robust to variability.
2. Experimental design can be used to accomplish this.
3. Operation on target is more important to conformance to specifications.

However, a word of caution is appropriate. Although his philosophy is sound, some of the methods of statistical analysis and some of the approaches to experimental design he advocates have been shown to be unnecessarily complicated, inefficient, and even ineffective. Thus, care should be exercised in applying Taguchi's methods.

The Taguchi methodology is best illustrated by example. In the following sections, we use an example first published in the *Bell System Technical Journal* [5]. In this example, Taguchi's approach is used to optimize the photolithographic process used to form square contact windows in a CMOS microprocessor fabrication process. The purpose of the contact windows is to facilitate the interconnection between transistors. The goal is to produce windows of a size near the target dimension.

7.4.1. Categorizing Process Variables

The first step in applying the Taguchi methodology is to identify the important process variables that can be manipulated, as well as their potential working levels. These variables are categorized as either controllable or uncontrollable. The controllable factors are also referred to as either *control factors* or *signal factors*, whereas the uncontrollable factors are called *noise factors*.

For the contact window formation example, the key process steps are (1) applying the photoresist by spin coating, (2) prebaking the photoresist, (3) exposing the photoresist, (4) developing the resist, and (5) etching the windows using plasma etching. The controllable factors associated with each step are as follows:

- *Applying photoresist*—resist viscosity (B) and spin speed (C)
- *Baking*—bake temperature (D) and bake time (E)
- *Exposure*—mask dimension (A), aperture (F), and exposure time (G)
- *Development*—developing time (H)
- *Plasma etch*—etch time (I)

The operating levels of these nine factors are shown in Table 7.21. Six factors have three levels each, and three factors have only two levels. The levels of spin speed in this table are dependent on the levels of viscosity. For the 204 photoresist viscosity, the low, normal, and high levels of the spin speed are 2000, 3000, and

Table 7.21. Factors and Levels for Taguchi Example.

Label	Factor Name	Low Level	Medium Level	High Level
A	Mask dimension (μm)	—	2	2.5
B	Viscosity	—	204	206
C	Spin speed (rpm)	Low	Normal	High
D	Bake temperature (°C)	90	105	
E	Bake time (min)	20	30	40
F	Aperture	1	2	3
G	Exposure time	20% over	normal	20% under
H	Developing time (s)	30	45	60
I	Etch time (min)	14.5	13.2	15.8

4000 rpm, respectively. For the 206 viscosity, those levels are 3000, 4000, and 5000 rpm.

Examples of potential noise factors in this study include the relative humidity or the number of particles in the cleanroom. The objective of this procedure is to determine the levels of the controllable factors that lead to windows closest to the target dimension of 3.5 μm.

7.4.2. Signal-to-Noise Ratio

Taguchi recommends analyzing the results from designed experiments using the mean response and the appropriately selected *signal-to-noise ratio* (*SN*). Signal-to-noise ratios are derived from the quadratic loss function given in Eq. (7.48). The three standard *SN*s are

$$\text{Nominal the best:} \qquad SN_N = 10\log(\overline{y}/s) \tag{7.49}$$

$$\text{Larger the better:} \qquad SN_L = -10\log\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{y_i^2}\right) \tag{7.50}$$

$$\text{Smaller the better:} \qquad SN_S = -10\log\left(\frac{1}{n}\sum_{i=1}^{n}y_i^2\right) \tag{7.51}$$

where s is the sample standard deviation. Each of these ratios is expressed in a decibel scale. SN_N is used if the objective is to reduce variability around a specific target, SN_L is appropriate if the system is optimized when the response is as large as possible, and SN_S is selected to optimize a system by making the response as small as possible. Factor levels that maximize the appropriate *SN* ratio are considered optimal. Clearly, SN_N is the right choice for the contact window formation experiment.

7.4.3. Orthogonal Arrays

A full factorial experiment to explore all possible interactions of the factors in Table 7.21 would require $3^6 \times 2^3 = 5832$ trials. Clearly, when cost of material, time, and availability of facilities are considered, the full factorial approach is prohibitively large. Taguchi recommends an alternative fractional factorial design known as the *orthogonal array*. The columns of such an array are pairwise orthogonal, meaning that for every pair of columns, all combinations of levels occur and they occur an equal number of times.

Table 7.22 shows the L_{18} orthogonal array design for the contact window formation study [5]. In this table, factors B and D are treated as a joint factor BD with levels 1, 2, and 3 representing the combinations B_1D_1, B_2D_1, and B_2D_2, respectively. This was done to accommodate the L_{18} array, which can be used to evaluate a maximum of eight factors.

The L_{18} array is a "main effects only" design that assumes that the response(s) can be approximated by a separable function. In other words, it is assumed that the response(s) can be written in terms of a sum of terms where each term is a function of a single independent variable. This type of model can yield misleading conclusions in the presence of factor interactions. However, Taguchi claims that the use of the *SN* ratio generally eliminates the need to examine interactions.

For estimating the main effects, there are 2 degrees of freedom associated with each three-level factor, one degree of freedom associated with each two-level factor, and one degree of freedom associated with the mean. Since we need at least one experiment for each degree of freedom, the minimum number of experiments required for optimizing the contact window formation process is 16. The L_{18} array has 18 trials, which provides additional precision in estimating the effects.

Table 7.22. Factor levels for the L_{18} orthogonal array.

Experiment	A	BD	C	E	F	G	H	I
1	1	1	1	1	1	1	1	1
2	1	1	2	2	2	2	2	2
3	1	1	3	3	3	3	3	3
4	1	2	1	1	2	2	3	3
5	1	2	2	2	3	3	1	1
6	1	2	3	3	1	1	2	2
7	1	3	1	2	1	3	2	3
8	1	3	2	3	2	1	3	1
9	1	3	3	1	3	2	1	2
10	2	1	1	3	3	2	2	1
11	2	1	2	1	1	3	3	2
12	2	1	3	2	2	1	1	3
13	2	2	1	2	3	1	3	2
14	2	2	2	3	2	2	1	3
15	2	2	3	1	2	3	2	1
16	2	3	1	3	2	3	1	2
17	2	3	2	1	3	1	2	3
18	2	3	3	2	1	2	3	1

7.4.4. Data Analysis

The postetch window size is the most appropriate quality measure for this experiment. Unfortunately, because of the size and proximity of the windows, the existing equipment in the paper by Phadke et al. [5] was unable to provide reproducible window size measurements. As a result, a linewidth test pattern on each chip was used to characterize the window size. The postetch line width was used as a window size metric.

Five chips were selected from each of the wafers used in the L_{18} design in Table 7.22. These five chips correspond to the top, bottom, left, right, and center of the wafers. Once again, the optimization problem posed by this experiment is to determine the optimum factor levels such that SN_N is maximum while keeping the mean on target. This problem is solved in two stages:

1. Use ANOVA techniques to determine which factors have a significant effect on SN_N. These factors are called the *control factors*. For each control factor, we choose the level with the highest SN_N as the optimum level, thereby maximizing the overall SN_N (under the separability assumption).
2. Select a factor that has the smallest effect on SN_N among the control factors. Such a factor is called a *signal factor*. Set the levels of the remaining factors (i.e., those that are neither control nor signal factors) to their nominal levels prior to the optimization experiment. Then, set the level of the signal factor so that the mean response is on target.

In cases where multiple responses exist, engineering judgment is used to resolve conflicts when different response variables suggest different levels for any single factor.

For each trial in Table 7.21, the mean, standard deviation, and SN_N for the postetch line width were computed. The following linear model was used to analyze these data

$$y_i = \mu + x_i + e_i \tag{7.52}$$

where y_i is the SN_N for experiment i, μ is the overall mean, x_i is the sum of the main effects of all eight factors in experiment i, and e_i is the random error in experiment i. The ANOVA table for SN_N and the mean postetch line width are shown in Tables 7.23 and 7.24, respectively. In Table 7.24, a *pooled ANOVA* was derived by pooling the sum of squares for those factors whose sums of squares were smaller than the error sum of squares (D, E, F, and I) with the error sum of squares. The "percent contribution" in the last column of Table 7.24 is a Taguchi metric that is equal to the total sum of squares explained by a factor after an appropriate estimate of the error sum of squares has been removed from it. A larger percent contribution implies that more can be expected to be achieved by changing the level of that factor.

Table 7.23 indicates that none of the nine process factors has a significant effect on SN_N for postetch linewidth. Thus, none of these may be considered

Table 7.23. Postetch linewidth ANOVA for SN_N.

Factor	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio
A	1	0.005	0.005	0.02
B	1	0.134	0.134	0.60
C	1	0.003	0.003	0.01
D	2	0.053	0.027	0.12
E	2	0.057	0.028	0.13
F	2	0.085	0.043	0.19
G	2	0.312	0.156	0.70
H	2	0.156	0.078	0.35
I	2	0.008	0.004	0.02
Error	2	0.444	0.222	
Total	17	1.257		

Table 7.24. Pooled ANOVA for mean postetch linewidth.

Factor	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio	% Contribution
A	1	0.677	0.677	16.92	8.5
B	1	2.512	2.512	63.51	32.9
C	2	1.424	0.712	17.80	17.9
G	2	1.558	0.779	19.48	19.6
H	2	0.997	0.499	12.48	12.2
Error	9	0.356	0.040		8.9
Total	17	7.524			100.0

control factors in this experiment. However, all the factors in Table 7.24 (viscosity, exposure, spin speed, mask dimension, and developing time) were significant at a 95% confidence level for the mean value of this response. The mean linewidth for each factor is shown in Figure 7.20.

To keep the process on target, a signal factor must be selected that has a significant effect on the mean, but little effect on SN_N. Changing the signal factor then affects only the mean. In this experiment, exposure time (G) was selected as the signal factor. This factor was adjusted to obtain the optimum linewidth and therefore, window size. This adjustment resulted in a factor of 2 decrease in window size variation and a factor of three decrease in the number of windows not printed. Thus the Taguchi methodology was proven to be effective in this example.

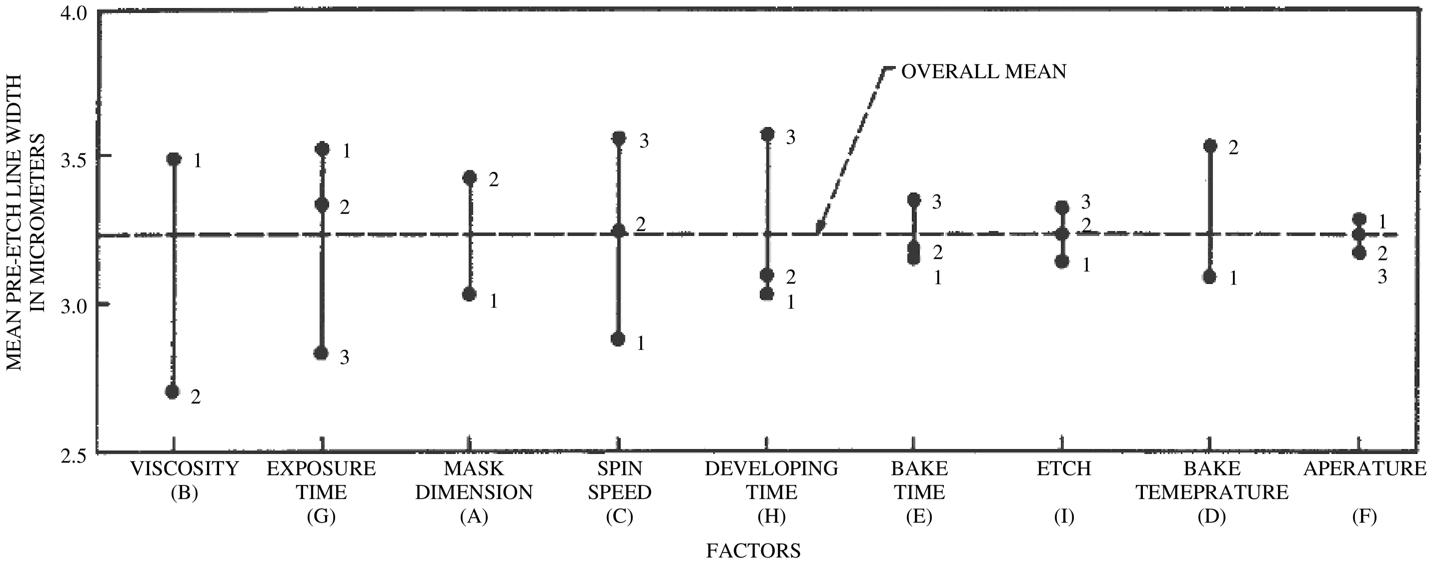


Figure 7.20. Mean postetch linewidth; The mean for each factor is indicated by a dot, and the number next to the dot indicates the factor level [5].

SUMMARY

In this chapter, we have provided an overview of statistical experimental design by introducing the concept of analysis of variance and describing various types designs, including two-level factorial designs and the Taguchi methodology. Important topics also included the analysis of such experiments and the use of various analytical and graphical methods to interpret experimental results. In the next chapter, we will examine how data generated from design experiments may be used to construct models that predict process behavior.

PROBLEMS

7.1. To compare two photolithography processes (A and B), 4 of 8 wafers were randomly assigned to each. The electrically measured linewidth of several NMOS transistors gave the following averages (in μm):

A:	1.176	1.230	1.146	1.672
B:	1.279	1.000	1.146	1.176

Assuming that the processes have the same standard deviation, calculate the significance for the comparison of means.

7.2. Suppose that there are now four photolithography processes to compare (A, B, C, and D). Using 15 wafers, the measurements are as follows (in μm):

	I	II	III	IV
A	1.176	1.230	1.146	1.672
B	1.279	1.000	1.146	1.176
C	0.954	1.079	1.204	—
D	0.699	1.114	1.114	—

Calculate the full ANOVA table and find the level of significance for rejecting the hypothesis of equality. Explain any assumptions and perform the necessary diagnostics on the residuals.

7.3. The following data are for the throughput, as measured by the number of wafer lots produced per day by different operators (A, B, C, and D) on different machines (each operator used each machine on two different days):

Machine	A	B	C	D
1	18(9), 17(76)	16(11), 18(77)	17(22), 20(72)	27(3), 27(73)
2	17(1), 13(71)	18(3), 18(73)	20(57), 16(70)	28(2), 23(78)
3	16(3), 17(77)	17(7), 19(70)	20(25), 16(73)	31(33), 30(72)
4	15(2), 17(72)	21(4), 22(74)	16(5), 16(71)	31(6), 24(75)
5	17(17), 18(84)	16(10), 18(72)	14(39), 13(74)	28(7), 22(82)

Eighty-four working days were needed to collect the data. The numbers in parentheses refer to the days on which the results were obtained. For example, on the first day, operator A produced 17 lots using machine 2, and on the 84th day, operator A produced 18 lots using machine 5. On some days (such as the third day), more than one item of data was collected, and on other days (such as day 40), no data was collected. Analyze the data, stating all assumptions and conclusions.

7.4. Consider the data in Table 7.11. Carry out analysis of variance using the data transformation $Y = y^{-1}$. Consider whether in the new response metric there is evidence of model inadequacy. Compare the treatment averages for the two different representations of the response.

7.5. The following single-replicate 2^3 factorial design was used to develop a nitride etch process. State any assumptions you make, and analyze this experiment.

Temperature (°F)	Concentration (%)	Catalyst	Yield (%)
160	20	1	60
180	20	1	77
160	40	1	59
180	40	1	68
160	20	2	57
180	20	2	83
160	40	2	45
180	40	2	85

7.6. **(a)** Why do we "block" experimental designs?

(b) Write a 2^3 factorial design.

(c) Write a 2^3 factorial design in four blocks of two runs each such that the main effects are not confounded with the blocks.

7.7. Consider a 2^{8-4} fractional factorial design.

(a) How many variables does the design have?

(b) How many runs are involved in the design?

(c) How many levels are used for each variable?

(d) How many independent block generators are there?

(e) How many words are there in the defining relations (counting I)?

7.8. Construct a 2^{7-1} fractional factorial design. Show how the design may be divided into eight blocks of eight runs each so that no main effect or two-factor interaction is confounded with any block effect.

7.9. A 2^3 factorial design on a CVD system was replicated 20 times with the following results:

P	T	F	$\overline{y}$
$-$	$-$	$-$	7.76 ± 0.53
$+$	$-$	$-$	10.13 ± 0.74
$-$	$+$	$-$	5.86 ± 0.47
$+$	$+$	$-$	8.76 ± 1.24
$-$	$-$	$+$	9.03 ± 1.12
$+$	$-$	$+$	14.59 ± 3.22
$-$	$+$	$+$	9.18 ± 1.80
$+$	$+$	$+$	13.04 ± 2.58

In this tabulation, $\overline{y}$ is the deposition rate in μm/min and the number following the ± sign is the standard deviation of $\overline{y}$. The variables P, T, and F represent pressure, temperature, and flowrate. Analyze the results. Should a data transformation be made?

REFERENCES

1. G. Box, W. Hunter, and J. Hunter, *Statistics for Experimenters*, Wiley, New York, 1978.
2. G. Box and D. Cox, "An Analysis of Transformations," *J. Roy. Stat. Soc. B.* **26**, 211 (1964).
3. D. Montgomery, *Introduction to Statistical Quality Control*, Wiley, New York, 1993.
4. G. Taguchi and Y. Wu, Control Japan Quality Control Organization, Nagoya, Japan, 1980.
5. M. Phadke, R. Kackar, D. Speeney, and M. Grieco, "Off-Line Quality Control in Integrated Circuit Fabrication Using Experimental Design," *Bell Syst. Tech. J.*, (May–June 1983).

8

PROCESS MODELING

OBJECTIVES

- Provide an overview of statistical modeling techniques such as regression and response surface methods.
- Introduce the concept of principal-component analysis (PCA).
- Discuss new modeling methods based on artificial intelligence techniques.
- Describe methods of model-based process optimization.

INTRODUCTION

As discussed in Chapter 7, a designed experiment is an extremely useful tool for discovering key variables that influence quality characteristics. Statistical experimental design is a powerful approach for systematically varying process conditions and determining their impact on output parameters that measure quality. Data derived from such experiments can then be used to construct *process models* of various types that enable the analysis and prediction of manufacturing process behavior.

The models so derived may be used to visualize process behavior in the form of a *response surface*. The proper fit is obtained using statistical regression techniques such as the method of *least squares* (also known as *linear regression analysis*). The goal of regression analysis is to develop a quantitative model

Fundamentals of Semiconductor Manufacturing and Process Control,
By Gary S. May and Costas J. Spanos

(usually in the form of a polynomial) that predicts a relationship between input factors and a given response. An accurate model should minimize the difference between the observed values of the response and its own predictions.

Over time, several novel methods have been developed to augment regression modeling. For example, *principal component analysis* (PCA) is a useful statistical technique for streamlining a multidimensional dataset to facilitate subsequent modeling. Dimensionality reduction through PCA is achieved by transforming a data to a new set of variables (i.e., the *principal components*), which are uncorrelated and ordered such that the first few retain most of the variation present in the original dataset.

Approaches that utilize *artificial intelligence* (AI) methods such as *neural networks* or *fuzzy logic* are capable of performing highly complex mappings on noisy and/or nonlinear experimental data, thereby inferring very subtle relationships between diverse sets of input and output parameters. Moreover, these techniques can also generalize well enough to learn overall trends in functional relationships from limited training data.

Process modeling permits an engineer to manipulate and optimize the process efficiency with a minimum amount of experimentation. A well-developed process model can in turn be used to generate a recipe of the process deposition conditions to obtain particular desired responses. In effect, this required that the neural process model be used "in reverse" to predict the necessary operating conditions to achieve the desired film characteristic. This chapter explores various process modeling methodologies, from traditional regression analysis to more contemporary AI-based approaches for deriving predictive models in semiconductor manufacturing applications. We then explore various optimization (or recipe synthesis) procedures.

8.1. REGRESSION MODELING

Raw experimental data have limited meaning in and of themselves; they are most useful in relation to some conceptual model of the process being studied. Once such data have been obtained from a designed experiment (see Chapter 7), the results may be summarized in the form of a *response surface*. The proper fit for a response surface is obtained using statistical regression techniques. When the formulation of the response surface is such that the outcome is a linear function of the unknown parameters, these parameters can be estimated by the method of *least squares* (also known as *linear regression analysis*). Linear regression analysis is a statistical technique for modeling and investigating the relationship between two or more variables. The goal of regression analysis is to develop a quantitative model that predicts this relationship between controllable input factors and a given response.

In general, suppose that there is a single *dependent variable* or *response* y that is related to k *independent variables*, say, $x_1, x_2, \ldots, x_k$. Assume that the dependent variable y is a random variable, and the independent variables $x_1, x_2, \ldots, x_k$

are exactly known or can be measured with negligible error. The independent variables are controllable by the experimenter. The relationship between these variables is characterized by a mathematical model called a *regression equation*. More precisely, we speak of the regression of y on $x_1, x_2, \ldots, x_k$. This regression model is fitted to a set of data. In some instances, the experimenter will know the exact form of the true functional relationship between y and $x_1, x_2, \ldots, x_k$, say, $y = f(x_1, x_2, \ldots, x_k)$. However, in most cases, the true functional relationship is unknown, and the experimenter must derive an appropriate function to approximate the function f. A polynomial model is often employed as the approximating function. An accurate model should minimize the difference between the observed values of the response and its own predictions. In addition to predicting the response, such a model can also be used for process optimization or process control purposes.

8.1.1. Single-Parameter Model

The simplest polynomial response surface is merely a straight line. Models fit to a straight line are derived using *linear regression*.[1] Consider fitting experimental data to a straight line that passes through the origin. Although rather elementary, this example illustrates the basic principles of least squares.

Suppose that we are studying the etch rate of a wet etchant, and we collect $n = 9$ observations of the data shown in Table 8.1, where x is the time in minutes and y is the thickness of film etched away. Physical considerations indicate that a simple proportional relationship between x and y is reasonable; that is, the relationship between x and y should be described by a straight line through the origin, or

$$y_u = \beta x_u + \varepsilon_u \qquad u = 1, 2, \ldots, n \tag{8.1}$$

Table 8.1. Hypothetical etching data.

Observation (u)	Time [x_u (min)]	Thickness [y_u (μm)]
1	8	6.16
2	22	9.88
3	35	14.35
4	40	24.06
5	57	30.34
6	73	32.17
7	78	42.18
8	87	43.23
9	98	48.76

[1] The term “linear” refers to the fact that the regression equation is linear to the unknown parameters. In this sense, as we will see, linear regression is also capable of deriving models that are non-linear to the regressors.

where β is a constant of proportionality (i.e., the slope of the "best fit" line) and the ε_u are random, independent experimental errors with zero mean and constant variance [i.e., $\varepsilon_u \sim N(0, \sigma^2)$]. The response or output variable y is the dependent variable, and the input variable x is the independent variable or the *regressor*. The objective of regression analysis is to find an estimate of b that minimizes the difference between the measured values of y and the predictions of Eq. (8.1).

According to the method of least squares, the best-fit model is the one that minimizes the quantity

$$S(\beta) = \sum_{u=1}^{n} (y_u - \beta x_u)^2 = \sum (y - \hat{y})^2 \tag{8.2}$$

where $\hat{y} = \beta x$ is an estimate of y and the subscripts have been dropped to simplify the notation. The curve represented by this equation is a parabola, so the goal is to find the value of β at the minimum of the parabola. Let b be the value of β at the minimum point. Using the rules of calculus, we can find b by simply taking the derivative of S with respect to β and setting the derivative equal to zero, or

$$\frac{dS}{d\beta} = 2\sum (y - \hat{y})x = 2\sum (y - bx)x = 0 \tag{8.3}$$

since $\hat{y} = bx$ at the minimum point. Solving (8.3) for b yields

$$b = \frac{\sum xy}{\sum x^2} \tag{8.4}$$

Using the etching data in Table 8.1, we compute $b = 0.501$ μm/min. This value has been substituted into $\hat{y} = bx$ and plotted in Figure 8.1.

8.1.1.1. Residuals

Once the least-squares estimate (b) of the unknown coefficient (β) has been obtained, the estimated response $\hat{y}_u = bx_u$ can be computed for each x_u. These estimated responses can be compared with the observed values (y_u). The differences between the estimated and observed values ($y_u - \hat{y}_u$) are known as

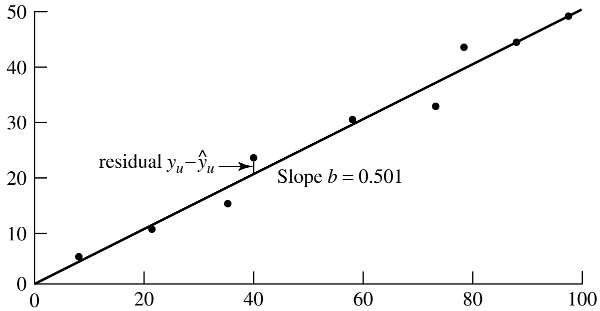


Figure 8.1. Plot of data fitted to least-squares line, etching example [1].

residuals. The *sum of squares* of the residuals is given by

$$S_R = S(b) = \sum_{u=1}^{n} (y_u - \hat{y}_u)^2 \tag{8.5}$$

In this example, $S_R = 64.67\ \mu\text{m}^2$. As discussed in Chapter 7, it is important to examine the residuals individually and collectively for inadequacies in the model.

8.1.1.2. Standard Error

If the one-parameter linear model is adequate, then an estimate (s^2) of the experimental error variance (σ^2) can be obtained by dividing the residual sum of squares by its number of degrees of freedom. The number of degrees of freedom will generally equal the number observations less the number of parameters estimated. Since only a single parameter is estimated in this model, there are $n - 1 = 8$ degrees of freedom in this example. An estimate of σ^2 is therefore

$$s^2 = \frac{S_R}{n-1} = 8.08 \tag{8.6}$$

The corresponding estimated variance for b is then [1]

$$V(b) = \frac{s^2}{\displaystyle\sum_{u=1}^{n} x_u^2} = 0.00023 \tag{8.7}$$

Thus, the *standard error* of b is $\text{SE}(b) = \sqrt{V(b)} = 0.015\ \mu\text{m/min}$. This metric can be used to perform a hypothesis test (see Chapter 4) to determine whether the true value of β is equal to some specific value β^* using the test statistic

$$t_0 = \frac{b - \beta^*}{\text{SE}(b)} \tag{8.8}$$

which is distributed according to the t distribution with $n - 1 = 8$ degrees of freedom. The $1 - \alpha$ confidence interval for β is bounded by

$$b \pm [t_{\alpha/2} \times \text{SE}(b)] \tag{8.9}$$

8.1.1.3. Analysis of Variance

For linear least-squares problems such as the one considered in the preceding sections, the following relationships exist among the sums of squares and their corresponding degrees of freedom

$$\sum y_u^2 = \sum \hat{y}_u^2 + \sum (y_u - \hat{y}_u)^2 \tag{8.10}$$

$$n = p + (n - p) \tag{8.11}$$

where p is the number of parameters estimated by least squares. For the etching problem, $p = 1$, and Eqs. (8.10) and (8.11) yield the *analysis of variance* shown in Table 8.2 (see Chapter 7).

This ANOVA table is appropriate for testing the null hypothesis that $\beta^* = 0$. To do so, the ratio of the mean squares ($s_M^2/s_R^2 = 1094$) is compared with the

Table 8.2. ANOVA table for etch data.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio
Model	$S_M = 8836.64$	1	$s_M^2 = 8836.64$	$s_M^2/s_R^2 = 1094$
Residual	$S_R = 64.67$	8	$s_R^2 = 8.08$	
Total	$S_T = 8901.31$	9		

value of the F distribution with 1 and 8 degrees of freedom. The ratio for this example is overwhelmingly significant, indicating that there is little probability that β^* is in fact zero. This test is exactly equivalent to applying the t test implied by Eq. (8.8).

8.1.2. Two-Parameter Model

Many modeling situations require more than a single parameter. Consider the data in Table 8.3 representing the level of impurities in a polymer dielectric layer as a function of the concentration of a certain monomer and a certain dimer. Here, the appropriate model is

$$y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{8.12}$$

where y is the percent impurity concentration, x_1 is the percent concentration of the monomer, x_2 is the percent concentration of the dimer, and $\varepsilon \sim N(0, \sigma^2)$.

The best-fit model in this case is the one that minimizes the quantity

$$S(\beta) = \sum (y - \beta_1 x_1 - \beta_2 x_2)^2 \tag{8.13}$$

Since there are two parameters, this equation now represents a plane rather than a line. We could find the values of β_1 and β_2 that minimize $S(\beta)$ (i.e., b_1 and b_2, respectively) using the same calculus-based approach as we used for the single-parameter model. Alternatively, we can also use what are called the *normal equations* to compute these values. If we let $\hat{y} = b_1 x_1 + b_2 x_2$, this approach utilizes the fact that the vector of *residuals* (i.e., the vector composed of the

Table 8.3. Hypothetical polymer impurity data.

Observ-ation	Monomer Concentration [x_1 (%)]	Dimer Concentration [x_2 (%)]	Impurity Concentration [y (%)]
1	0.34	0.73	5.75
2	0.34	0.73	4.79
3	0.58	0.69	5.44
4	1.26	0.97	9.09
5	1.26	0.97	8.59
6	1.82	0.46	5.09

values of $y - \hat{y}$ for each of the n observations) has the property of being normal (at right angles) to each vector of x values when the least squares estimate is used.

In this model, there are two regressors, x_1 and x_2. The normal equations in this case are

$$\sum(y - \hat{y})x_1 = 0 \qquad \sum(y - \hat{y})x_2 = 0 \tag{8.14}$$

or

$$\sum(y - b_1x_1 - b_2x_2)x_1 = 0 \qquad \sum(y - b_1x_1 - b_2x_2)x_2 = 0 \tag{8.15}$$

Simplifying further gives

$$\sum yx_1 - b_1\sum x_1^2 - b_2\sum x_1x_2 = 0 \qquad \sum yx_1 - b_1\sum x_1x_2 - b_2\sum x_2^2 = 0 \tag{8.16}$$

Solving these two equations simultaneously using the data in Table 8.3 yields $b_1 = 1.21$ and $b_2 = 7.12$. The fitted surface appears in Figure 8.2. In general,

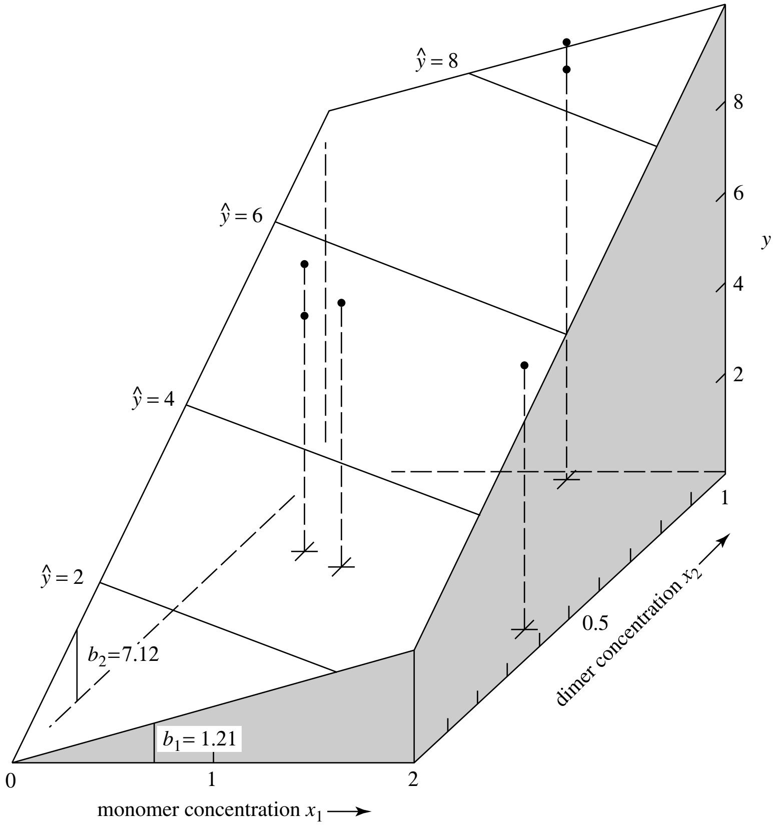


Figure 8.2. Fitted plane $\hat{y} = 1.21x_1 + 7.12x_2$ for polymer impurity example [1].

this method can be applied to a linear equation of the form given by Eq. (8.12) with an arbitrary number of regressor variables.

8.1.2.1. Analysis of Variance

The first step in evaluating the adequacy of the model presented above is inspection of the residuals (i.e., $y - \hat{y}$). However, with such a small dataset, only gross discrepancies would be revealed by such an analysis. In this case, no such discrepancies are evident.

Another type of analysis is also appropriate when some of the experimental runs have been replicated. In this case, runs 1 and 2 are replicates, as are runs 4 and 5. The sum of squares associated with these replicate runs is

$$S_E = \frac{(y_1 - y_2)^2}{2} + \frac{(y_4 - y_5)^2}{2} \tag{8.17}$$

This sum of squares, which has 2 degrees of freedom, is part of the overall residual sum of squares (S_R) and is a measure of the "pure" experimental error. The remaining part of the residual sum of squares is given by

$$S_L = S_R - S_E \tag{8.18}$$

This quantity measures the experimental error plus any contribution from possible lack of fit of the model. A comparison of the mean squares derived from S_E and S_L can therefore be used to check the lack of fit. These concepts are summarized in the ANOVA given in Table 8.4.

In this example, the close agreement between the two mean squares (as indicated by the F ratio near unity) gives no reason to suspect a significant lack of fit. An examination of Appendix E reveals that a mean-square ratio greater than 1.2 can be expected about 45% of the time with this small number of degrees of freedom. It can therefore be concluded that the fit for this model is adequate.

8.1.2.2. Precision of Estimates

According to the assumption that the model is adequate, an estimate of the error variance of the model is

$$s^2 = \frac{S_R}{n - p} = 0.33 \tag{8.19}$$

Table 8.4. ANOVA table for polymer impurity data.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio
Model	$S_M = 266.59$	2		
Residual	$S_R = 1.33$			
Lack of fit	$S_L = 0.74$	2	$s_L^2 = 0.37$	$s_L^2/s_E^2 = 1.2$
Pure error	$S_E = 0.59$	2	$s_E^2 = 0.30$	
Total	$S_T = 267.92$	6		

To estimate the variances of b_1 and b_2, the correlation (ρ) between these parameters must first be computed using

$$\rho = \frac{-\sum x_1 x_2}{\sqrt{\sum x_1^2 x_2^2}} = -0.825 \tag{8.20}$$

The variances are then given by

$$V(b_1) = \frac{1}{(1-\rho)^2} \frac{s^2}{\sum x_1^2} = 0.147 \tag{8.21}$$

$$V(b_2) = \frac{1}{(1-\rho)^2} \frac{s^2}{\sum x_2^2} = 0.285$$

Since the standard error for each parameter is just the square root of its variance, $\mathrm{SE}(b_1) = 0.383$, and $\mathrm{SE}(b_2) = 0.534$. Given these values for standard error, it is possible to define $(1 - \alpha)$ confidence limits for each parameter using Eq. (8.9).

8.1.2.3. Linear Model with Nonzero Intercept

Consider the problem of fitting data to a linear model that does not pass through the origin. The equation of such a line is given by

$$y = \beta_0 + \beta x + \varepsilon \tag{8.22}$$

where the intercept $\beta_0 \neq 0$. This model is just a special case of the model given in Eq. (8.12), with the following substitutions:

$$\beta_1 = \beta_0, \quad x_1 = 1, \quad \beta_2 = \beta, \quad \text{and} \quad x_2 = x$$

The "variable" $x_1 = 1$ is referred to as an *indicator variable*. For this model, the normal equations [Eqs. (8.14)–(8.16)] simplify to

$$b_0 n + b \sum x = \sum y \tag{8.23}$$

$$b_0 \sum x + b \sum x^2 = \sum xy$$

and the solutions are

$$b = \frac{\sum (x - \overline{x})(y - \overline{y})}{\sum (x - \overline{x})^2} \tag{8.24}$$

$$b_0 = \overline{y} - b\overline{x}$$

where n is the number of points to be fitted, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

To illustrate this situation, consider the hypothetical data in Table 8.5, which represents particle counts in a class 100 cleanroom as a function of equipment utilization. Applying Eq. (8.24) to these data yields $b = 11.66$ and $b_0 = 65.34$. Therefore, the appropriate linear model in this case is $\hat{y} = 65.34 + 11.66x$. This

Table 8.5. Hypothetical particle data.

Observation	Equipment Utilization [x (Arbitrary Units)]	Particle Count [y (ft^{-3})]
1	2.00	89
2	2.50	97
3	2.50	91
4	2.75	98
5	3.00	100
6	3.00	104
7	3.00	97

line is plotted in Figure 8.3. Given the expression for b_0 in Eq. (8.24), this line can also be written as

$$\hat{y} = b_0 + bx = \overline{y} - b\overline{x} + bx = \overline{y} + b(x - \overline{x}) = a + b(x - \overline{x}) \qquad (8.25)$$

where $a = \overline{y}$. For the current example, $\hat{y} = 96.57 + 11.66(x - 2.28)$.

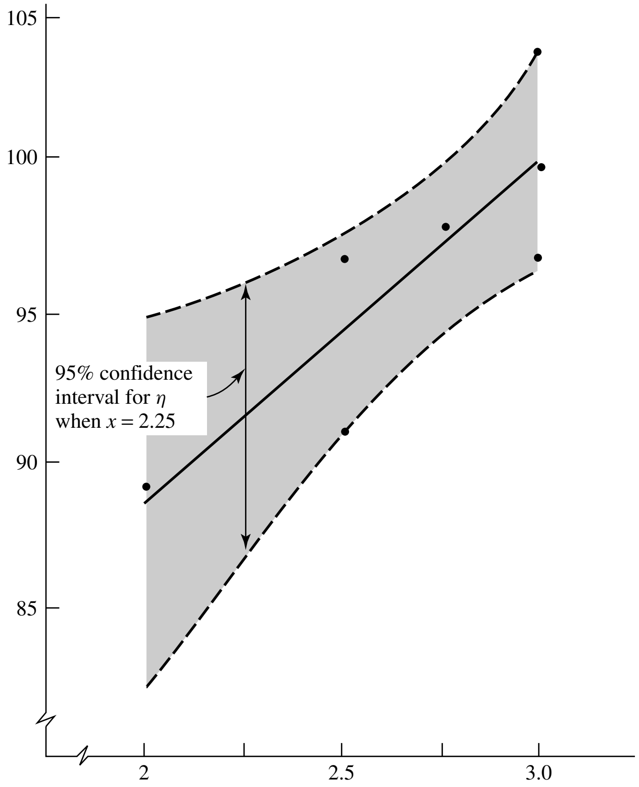


Figure 8.3. Fitted line $\hat{y} = 65.34 + 11.66x$ for cleanroom particle example [1].

The precision of the estimates for this model can be evaluated using an approach similar to that outlined in Section 8.1.2.2. Assuming that the model is adequate, an estimate of the error variance of the model is

$$s^2 = \frac{S_R}{n-p} = 8.72 \tag{8.26}$$

where $S_R = \sum (y - \hat{y})^2 = 43.62$, $n = 7$, and $p = 2$. The standard errors for the coefficients are

$$\begin{aligned}
\mathrm{SE}(b_0) &= s\left[\frac{1}{n} + \frac{\overline{x}^2}{\sum (x - \overline{x})^2}\right]^{1/2} = 8.64 \\
\mathrm{SE}(b) &= \frac{s}{\sqrt{\sum (x - \overline{x})^2}} = 3.22 \\
\mathrm{SE}(a) &= \frac{s}{\sqrt{n}} = 1.12
\end{aligned} \tag{8.27}$$

Using the form of the model given in Eq. (8.25), the variance at a given point (x_0, y_0) is

$$V(\hat{y}_0) = V(\overline{y}) + (x_0 - \overline{x})^2 V(b) = \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum (x - \overline{x})^2}\right] s^2 \tag{8.28}$$

A $(1 - \alpha)$ confidence interval for $\hat{y}_0$ is then

$$\hat{y}_0 \pm t_{\alpha/2}\sqrt{V(\hat{y}_0)} \tag{8.29}$$

A 95% confidence interval computed for this example is indicated by the dotted lines in Figure 8.3.

Example 8.1. Perform analysis of variance and test the goodness of fit for the linear model resulting from Table 8.5.

Solution:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F Ratio
Model	$S_M = 65{,}396.38$	2		
Residual	$S_R = 43.62$			
Lack of fit	$S_L = 0.953$	2	$s_L^2 = 0.477$	$s_L^2/s_E^2 = 0.034$
Pure error	$S_E = 42.67$	3	$s_E^2 = 14.213$	
Total	$S_T = 65{,}440$	7		

The ANOVA table is shown above. Note that since observations 2 and 3, as well as 5, 6, and 7 are replicates:

$$S_E = \frac{(y_2 - y_3)^2}{2} + \frac{(y_5 - y_6)^2}{3} + \frac{(y_6 - y_7)^2}{3} + \frac{(y_5 - y_7)^2}{3} = 42.67$$

The ratio of lack of fit to pure error mean squares is only 0.034. A true lack of fit would be indicated by a much larger value of this ratio (see Appendix E). We can be more than 99% confident that this model fits the data.

8.1.3. Multivariate Models

The method of least squares described above can be used in general for modeling any process in which the estimated parameters of the model (β_1, β_2, etc.) are *linear.* A model is linear in its parameters if it can be written in the form

$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{8.30}$$

where the x terms are the quantities known for each experimental run and are not functions of the β terms. The models discussed in Sections 8.1.1 and 8.1.2 are clearly of the linear type. Another example of a model that is linear in its parameters is the polynomial model:

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p \tag{8.31}$$

A polynomial model with $p \geq 2$ would be used when the process has been observed to exhibit higher order effects that are inadequately captured by a straight-line model. Another example is the sinusoidal model

$$\hat{y} = \beta_0 + \beta_1 \sin\theta + \beta_2 \cos\theta \tag{8.32}$$

where θ is varied in the different experimental runs. This type of model might be appropriate for a process known to exhibit periodic or cyclical behavior. In general, one could develop any functional relationship between the independent and dependent variables in a set of experimental data by simply substituting an arbitrary function for the x values in Eq. (8.30). For example, if we let $x_1 = \log \xi_1$ and $x_2 = e^{\xi_2}\xi_3$, then we obtain the model

$$\hat{y} = \beta_0 + \beta_1 \log \xi_1 + \beta_2 \frac{e^{\xi_2}}{\xi_3} \tag{8.33}$$

where ξ_1, ξ_2, and ξ_3 are known for each experimental trial.

When limited to models that are linear in their estimated coefficients, standard matrix algebra provides a convenient approach to solving least-squares regression problems. For example, in matrix notation, Eq. (8.30) or (8.31) can be rewritten as

$$\hat{\mathbf{y}} = \mathbf{Xb} \tag{8.34}$$

where $\hat{\mathbf{y}}$ is the $n \times 1$ vector of predicted values for the response, $\mathbf{X}$ is the $n \times p$ matrix of independent variables, and $\mathbf{b}$ is the $p \times 1$ vector of parameters to be estimated. Under these circumstances, the normal equations can be written as

$$\mathbf{X}^T(\mathbf{y} - \hat{\mathbf{y}}) = 0 \tag{8.35}$$

where T represents the transpose operation. Substituting Eq. (8.34) yields

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{b}) = 0 \tag{8.36}$$

If we assume that $\mathbf{X^T X}$ has an inverse, solving Eq. (8.36) for $\mathbf{b}$ yields

$$b = [\mathbf{X}^T\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{y} \tag{8.37}$$

In general, the variance–covariance matrix for the estimates is

$$V(\mathbf{b}) = [\mathbf{X}^T\mathbf{X}]^{-1}\sigma^2 \tag{8.38}$$

if the experimental variance σ^2 is known. Otherwise, assuming the form of the model is appropriate, σ^2 can be estimated using $s^2 = S_R/(n - p)$, where S_R is the residual sum of squares.

Although the β values in these models can be found by calculus-based methods or using the normal equations, computer programs are now widely available for this purpose. Such programs have become virtually indispensable for model building, as well as for use in model validation and verification. This is especially true when the model form is not known, and several functional forms must be analyzed and compared in terms of prediction and lack-of-fit characteristics.

Example 8.2. Assume that the yield (y) of a given process varies according to process condition x according to the relationship in Table 8.6.
Use Eq. (8.37) to fit these yield data to the quadratic model

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2$$

Table 8.6. Hypothetical yield data.

Observation	Process Condition [x (Arbitrary Units)]	Yield [y (%)]
1	10	73
2	10	78
3	15	85
4	20	90
5	20	91
6	25	87
7	25	86
8	25	91
9	30	75
10	35	65

Solution: The matrices needed are

$$\mathbf{X} = \begin{array}{c} \begin{array}{ccc} x_0 & x & x^2 \end{array} \\ \begin{bmatrix} 1 & 10 & 100 \\ 1 & 10 & 100 \\ 1 & 15 & 225 \\ 1 & 20 & 400 \\ 1 & 20 & 400 \\ 1 & 25 & 625 \\ 1 & 25 & 625 \\ 1 & 25 & 625 \\ 1 & 30 & 900 \\ 1 & 35 & 1225 \end{bmatrix} \end{array} \qquad \mathbf{y} = \begin{bmatrix} 73 \\ 78 \\ 85 \\ 90 \\ 91 \\ 87 \\ 86 \\ 91 \\ 75 \\ 65 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 10 & 215 & 5225 \\ 215 & 5225 & 138{,}125 \\ 5225 & 138{,}125 & 3{,}873{,}125 \end{bmatrix} \qquad \mathbf{X}^T\mathbf{y} = \begin{bmatrix} 821 \\ 17{,}530 \\ 418{,}750 \end{bmatrix}$$

Solving for **b** yields

$$\mathbf{b} = \begin{bmatrix} 35.66 \\ 5.26 \\ -0.128 \end{bmatrix}$$

so the appropriate quadratic equation is $\hat{y} = 35.66 + 5.26x - 0.128x^2$. This curve is plotted in Figure 8.4.

8.1.4. Nonlinear Regression

While a vast array of regression problems can be approximated by linear regression models (i.e., models that are linear in the parameters to be estimated), there are also models that must be nonlinear to their estimated parameters. Consider, for example, the exponential model

$$\hat{y} = \beta_1(1 - e^{-\beta_2 x}) \tag{8.39}$$

where x is known for each experimental trial. This model clearly cannot be written in the form of Eq. (8.30). It is, therefore, an example of a model that is nonlinear in its parameters. Fortunately, however, the general concept of least squares can still be applied to fit such models. However, while in linear regression we have an exact, closed-form solution, for most nonlinear regression problems we have an approximate, iterative solution. Further, some of the statistical

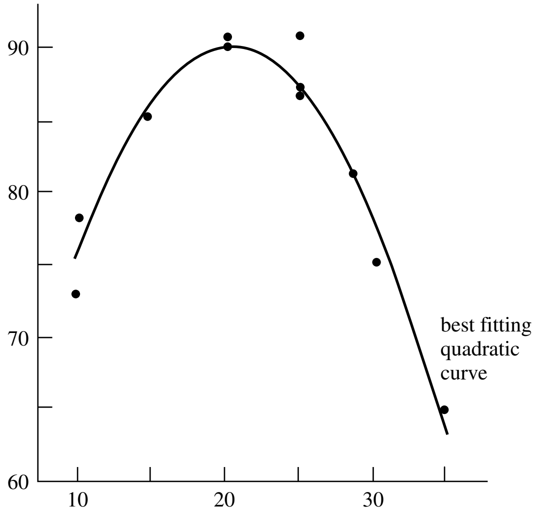


Figure 8.4. Fitted curve $\hat{y} = 35.66 + 5.26x - 0.128x^2$ for yield example [1].

assumptions that allowed us to use ANOVA in selecting model forms and in validating the results may be weak and subject to some speculation.

As an example, suppose that the number of particles generated by a particular process over time is given according to Table 8.7. Furthermore, assume that physical considerations suggest that the exponential model given by Eq. (8.39) should describe the phenomenon. The sum of squares in this case is given by

$$S = \sum_{u=1}^{n} [y_u - \beta_1(1 - e^{-\beta_2 x_u})]^2 \tag{8.40}$$

The estimated values of β_1 and β_2 that minimize S are $b_1 = 213.8$ and $b_2 = 0.5473$. Substituting these values into Eq. (8.39) gives the fitted least-squares curve shown in Figure 8.5.

Today, such curve fitting is not typically done by hand. On the contrary, modern computer software packages (such as *RS/Explore* [2]) exist that are capable

Table 8.7. Hypothetical particle data.

Observation	Particle Count [y (ft^{-3})]	Day (x)
1	109	1
2	149	2
3	149	3
4	191	5
5	213	7
6	224	10