

VEER SURENDRA SAI UNIVERSITY
OF TECHNOLOGY BURLA, ODISHA,

DEPARTMENT OF PRODUCTION
ENGINEERING

e-Lecture Notes on
PRECISION ENGINEERING (PE)
COURSE CODE: BMS 408
MODULE(I,II,III,IV)

8th Semester B. Tech
Production Engineering

Module-I

Precision Engineering: Micromilling and Microdrilling, MicroElectroMechanical Systems, Microelectronics fabrication methods, Principles of MEMS, mechanical MEMS, Thermal MEMS, Magnetic MEMS.

Nanotechnology- Carbon nanotubes and Structures, Processing system of nanometre accuracies, mechanism of material processing, Nano Physical processing of atomic bit-units, Nano-chemical and electrochemical atomic-bit processing.

1.2 Introduction

Manufacturing is the cornerstone of many industrial activities and significantly contributes toward the economic growth of a nation. Generally, the higher the level of manufacturing activity in a country, the better is the standard of living of its citizens. Manufacturing is the process of making large quantities of products by effectively utilizing the raw materials. It is a multidisciplinary design activity simply involving the synergistic integration of production and mechatronics engineering. The products vary greatly from application to application and are prepared through various processes. It encompasses the design and production of goods and systems, using various production principles, methodologies and techniques. The concept is hierarchical in nature in the sense that it inherits a cascade behaviour in which the manufactured product itself can be used to make other products or items. The manufacturing process may produce discrete or continuous products. In general, discrete products mean individual parts or pieces such as nails, gears, steel balls, beverage cans, and engine blocks, for example.

Conversely, examples of continuous products are spools of wires, hoses, metal sheets, plastic sheets, tubes, and pipes. Continuous products may be cut into individual pieces and become discrete parts. The scope of manufacturing technology includes the following broad topics:

- Precision engineering and ultra-precision engineering
- Micromanufacturing (Microelectronics and MEMS)
- Nanotechnology

1.2.1 Precision Engineering

The technical field of precision engineering has expanded over the past 25 years. In 1933, the Precision Engineering Society was established in Japan and soon thereafter the activities were accelerated due to new impetus from Europe. The first issue of the journal Precision Engineering appeared in 1979 and the first academic program began in 1982 (Source: American Society of Precision Engineering (ASPE)). According to ASPE, "...precision engineering is dedicated to the continual pursuit of the next decimal place." Precision engineering includes design methodology, uncertainty analysis, metrology, calibration, error compensation, controls, actuators and sensors design. A more complete list is given below

- Controls
- Dimensional metrology and surface metrology
- Instrument/machine design
- Interferometry
- Materials and materials processing
- Precision optics
- Scanning microscopes
- Semiconductor processing
- Standards

Frequently used terms within the domain of precision and ultra-precision engineering are precision processes, scaling, accuracy, resolution and repeatability. The precision process is a concept of design,

fabrication, and testing where variations in product parameters are caused by logical scientific occurrences. Identification of these logical phenomena and strategically controlling them is very fundamental to precision manufacturing. Scaling is a parameter that defines the ratio attributes with respect to the prototype model. It is also considered as a fundamental attribute for predicting the behaviour of structures and systems for analysis and synthesis of miniaturised systems. Accuracy defines the quality of nearness to the true value. In the context of machine or production systems, accuracy is the ability to move to a desired position. As an example, if the actual value is 1.123 units and it is recorded as 1.1 units, we are precise to the first decimal place but inaccurate by 0.023 units. Resolution is the fineness of position precision that is attainable by a motion system. The smallest increment that is produced by a servo system is the resolution. There are two types of resolutions, electrical and mechanical. With regard to mechanical resolution, it is defined as the smallest increment that can be controlled by a motion system, i.e., the minimum actual mechanical increment. One can note that mechanical resolution is significantly coarser than that due to the involvement of friction, stiction, deflections, and so on. Repeatability is the variation in measurements obtained when one person takes multiple measurements using the same instruments and techniques. Repeatability is typically specified as the expected deviation, i.e., a repeatability of 1 part in 10,000 or 1:10,000, for example.

1.2.2 Micromilling and Microdrilling

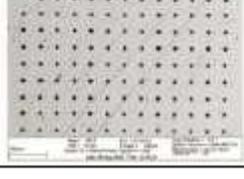
Micromilling and microdrilling are two important processes of precision engineering. The micromilling process is considered versatile and facilitates creating three-dimensional miniaturised structures. The process is characterized by milling tools that are usually in the order of hundreds of micrometers in diameter. These tools are designed by the use of focused-ion beam machining process and are used in a specially designed, high-precision milling machine. The focused-ion beam machining process uses a sharp tungsten needle wetted with gallium metal. The tip of the needle is subjected to a 5-10 kV (sometimes higher) so as to enable the field ionization effect on the gallium. The gallium ions are then accelerated by the use of another energy source and focused into a spot of sub-micrometer order. The kinetic energy acquired by the ions makes it possible to eject the atoms from the workpiece. This is referred to as a sputtering process. The sputtering yield varies inversely with the strength of the chemical bond in the materials. Either the movement of ions or the workpiece, depending upon the environmental conditions, can be controlled to obtain a wide variety of three-dimensional shapes and structures. It should be pointed out that the machining forces present in micromilling with tools of the order of micrometer diameters are dominated by contact pressure and friction between the tool cutting edges and the workpiece. As a rough calculation, one can note that in the focused-ion beam machining process, for a spot size of 0.45 μm with 2.5 nA of current, the required current density would be approximately 1.65 A/cm². The micromilling process is applied for making micromolds and masks to aid in the development of microcomponents. Typically, a high milling rate of 0.65 $\mu\text{m}^3/\text{nAs}$, corresponding to an average yield of 6.5 atoms/ion, can be obtained at 45 keV, 30° incidence, and 45 scans. Microdrilling is characterised by the drilling of ultrafine holes. Drilling in the micro ranges, using the special microdrills, requires a precision microdrilling instrument. The end of the microdrill is called the chisel edge, which is indeed removed material cutting at a negative rake angle. Microdrills are made of either micrograin tungsten carbide or cobalt steel. Some coarse microdrilling machines are available that drill holes from the size of 0.03 mm in diameter to 0.50 mm in diameter, with increments of 0.01 mm. However, the present demand is for drills capable of drilling in the order of micrometers. An example of this is a submicrodrilling technique utilising the phenomenon of ultrafast pulse laser interference. In this regard, for microdrilling and other delicate laser processing applications, Holo-Or Ltd. has released an optical element that creates an output spot in the form of a top hat circle with a diameter of 350 μm . The element accepts a collimated Gaussian incident beam with a diameter of 12 mm from a 10.6- μm CO₂ laser. Smooth 300 nm holes were successfully drilled on a 1000-Å thick gold film using the interfered laser beam, as compared to micrometer holes ablated using the conventional non-interfered laser beam. The most important parameters considered in microdrilling are: accuracy, sensitivity, quality and affordability. Some of the applications of microdrilling are given below:

- Air bearings and bushings
- EDM tooling
- Electronic components

- Gas and liquid flow
- Microwave components
- Nozzles
- Optical components

The major problem of conventional laser microdrilling is that the process has a short focal depth. It is known that this method typically achieves aspect ratios up to 100 in thick material, such as for a 15- μm hole in 1.5-mm-thick foil, for instance. This problem can be overcome by utilizing a Bessel beam. Deep high aspect ratio drilling is achieved due to the reason that the Bessel beam is nondiffracting and in practice they do not spread out. In the case of deep high-aspect ratio laser drilling, a pseudo-Bessel beam is generated using a pulsed laser. Some of the examples of microdrilling applications using a laser system developed by ATLASER di Andrea Tappi are presented in Table 1.1. The application of lasers to micromanufacturing has several advantages: noncontact processing, the capability of remote processing, automation, no tool wear and the possibility of machining hard and brittle materials.

Table 1.1. ATLASER di Andrea Tappi microdrilling system performance parameters

Si Wafer Thickness: 0.54 mm Hole diameter: 25 μm Hole pitch: 50 μm Process time: 0.65 s	
Silicon Carbide Wafer Thickness: 0.64 mm Through Hole: 130x500 μm In width: 130 μm Out width: 110 μm	
Aluminum Nitride Thickness: 425 μm In side width: 300 μm Out side width: 290 μm Drilling time: 33 s	
Cu-FR4 sandwich Thickness: 0.5mm Hole dimension: 200 μm Process time: 3.3 s	
Stainless Steel Sheet Thickness: 120 μm Hole diameter: 9 μm Hole pitch: 50 μm Matrix Process time: 0.15 s	

1.3 Microelectromechanical Systems (MEMS)

Microelectromechanical systems (MEMS) have already found significant applications in sectors that include, but are not limited to: automotive industry, aircraft industry, chemical industry, pharmaceuticals, manufacturing, defence, and environmental monitoring. The relative merit for MEM systems lies in the fact that these components are fabricated by batch manufacturing methods similar to microelectronics techniques, which fulfills the added advantage of miniaturization, performance and integrability. The topical areas under MEMS are micromachining methods, microsensors and actuators, magnetic MEMS, RFMEMS,

microfluidics, BioMEMS and MOEMS. The progress in microfabrication technologies is transforming the field of solid-state into MEMS. Micromachining is a process for the fabrication of MEMS devices and systems. Various energy transduction principles include thermal, magnetic, optical, electrical and mechanical. These are employed in designing the microsensors and actuators. Radio Frequency (RF) MEMS devices are mostly used in the field of wireless communication. Microfluidic MEMS devices handle and control small volumes of fluids in the order of nano and pico liter volumes. One popular application is a micro-nozzle for use in printing applications. MEMS technology has applications in the chemical industry, which gives rise to BioMEMS products. Surgical instruments, artificial organs, genomics, and drug discovery systems are based on BioMEMS products.

The miniaturised systems require less reagent, resulting in faster and more accurate systems. MEMS devices have better response times, faster analysis and diagnosis capabilities, better statistical results, and improved automation possibilities with a decreased risk and cost. Since most of the physical phenomenon and activations are to be measured and controlled precisely in a timely predictive manner so as to overcome real-time limitations, miniaturized components will have added advantages because of the inherent temporal behavior they possess. Moreover, prognostic measures in terms of sensor and actuator validation can be achieved through a system-on-a-chip (SOC) design approach. MEMS devices are useful for controlling micro mechanisms such as micro-manipulators, micro-handling equipment, micro-grippers, micro-robot, and others, which are primarily used for clinical, industrial and space applications.

1.4 Microelectronics Fabrication Methods

One of the major inventions in the last century is microelectronics, called microdevices. Micro devices can be integrated circuits, which are fabricated in sub-micron dimensions and form the basis of all electronic products. Microelectronics design entails the accommodation of essential attributes of modern manufacturing. Fabrication technology, starting from computer assisted off-line design to real fabrication, deals with the processes for producing electronic circuits, solid structures, printing circuits as well as various electronic components, sub-systems and systems of sub-miniature size.

The design of an IC with millions of transistors and even more interconnections is not a trivial task. Before the real design is manufactured, the circuit is prepared and tested by using EDA (electronic design automation) tools. These tools help in synthesizing and simulating the behavior of the desired circuit by arranging the placement of transistors and interconnections within the chip area. These computer-assisted tools can also verify and validate all defects and conditions, respectively. The technology has been driven by the demands of the computer industry, space technology, the car industry and telecommunications.

The first step in fabrication is always the preparation of a set of photographic masks. The mask represents the features of the various elements and layers of the chip to be manufactured. This procedure is repeated several times to replicate the circuit. The mask appears on the surface of a thin silicon crystal wafer. A single wafer can accommodate several identical chips. Hence, the IC fabrication process is a batch-processing scheme. The preparation of masks can be carried out by the use of a computer-controlled electron beam to expose the photographic mask material in accordance with the desired configuration. The information is supplied to the computer in terms of a design data file.

Then three important fabrication sequences are followed on the wafer surface. These are photography followed by chemical, and thermal operations. This phase is called masking. The mask features are transferred to the wafer by exposing a light-sensitive photoresist coating through the transparent areas of the mask. The material areas of the wafer unprotected by the hardened photoresist are then removed by etching. Etching techniques are characterised by their selectivity and degree of anisotropy. Etching can be either physical or chemical, or a combination of both. In order to develop active circuit elements such as transistors, n-type and p-type impurities are doped. Two commonly used doping methods are diffusion and ion implantation. Then a thin aluminum layer is deposited on the uppermost layers of the chip in order to allow metal to contact the device elements. The aluminum deposition is often achieved by using the chemical vapor deposition (CVD) method.

1.5 Nanotechnology

Nanoscale devices and equipment provide benefits in terms of an improved greener environment, miniaturization, efficiency and resource consciousness. Nanotechnology has accelerated research and

development in many disciplines. However, a key obstacle to its development remains in the need for cost-effective large-scale production methods. Nanotechnology has applications in many fields including automotive, aerospace, household appliances, sporting goods, telecommunication equipment and medical supplies.

1.6 Carbon Nanotubes and Structures

Besides the conventional forms of carbon, graphite and diamond, new forms of carbon such as fullerenes, carbon nanotubes and carbon onions have been discovered. A majority of current research focuses on the potential applications of carbon nanotubes (CNT). Carbon nanotubes are considered as ultra-fine unique devices, which can offer significant advantages over many existing materials due to their remarkable mechanical, electronic and chemical properties. With strong covalent bonding they possess unique one-dimensional structures. Nanotubes can be utilised as electronics devices, super-capacitors, lithium ion batteries, field emission displays, fuel cells, actuators, chemical and biological sensors and electron sources. Some of the R&D areas include: nanoscale phenomena, atomic and mesoscale modeling, carbon nano structures and devices, nano composites, biomaterials and systems, bio nano computational methods, fluidics and nanomedicine. CNTs are typically longer in length, usually measuring from about a few tens of nanometers to several micrometers, with a diameter up to 30 nanometers and as small as 2.5 nanometers. Apparently, they are hollow cylinders extremely thin with a diameter about 10,000 times smaller than a human hair. Each nanotube is a single molecule made up of a hexagonal network of covalently bonded carbon atoms. Freestanding carbon nanotubes can be grown by chemical vapor deposition (CVD) across the predefined trenches. The trenches can be fabricated lithographically in SiO₂ and then by depositing Pt over the sample to serve as the conducting substrate. Explicitly, they adhere to metallic semiconducting properties along with good thermal conductivity. Other essential properties they possess are:

- High tensile strength and high resilience
- High current densities

The ongoing research that is being carried out all over the world is based on the study of conductive and high-strength composites, energy storage and energy conversion devices, sensors for field emission displays and radiation sources, hydrogen storage media and nanometer-sized semiconductor devices such as probes and interfacing. Nanotube-based design scenarios anticipate the design of gears and bearings and hence the development of machines at the molecular level. The team at the Paul Pascal Research Center has been working on overcoming the obstacle of the formation of the carbon nanotubes and has developed a process for aligning them in the form of fibers and strips. Various structural constructions can be formed through appropriate methods. Continuous sputtering of carbon atoms from the nanotubes lead to a dimensional change, which facilitates surface reconstruction with annealing. An X-like junction with diverse angles between the branches can be formed. Under careful irradiation one of the branches of the X-junction can be removed, thereby creating Y- and T-like junctions. This new class of carbon junctions exhibits an intrinsic non-linear transport behavior, depending mainly on its pure geometrical configuration and on the kind of topological defects. Calculation and measurement of characteristic curves, like current versus voltage of different sets of Y junctions, show robust rectification properties, giving rise to the possibility of using these junctions as nanoscale three-point transistors. Advanced computational techniques, including large-scale parallelisable molecular dynamic simulations of the growth mechanism and first principles calculations of the electronic structure, are being applied to model the self-assembly and the electronic properties of nanostructures. Based on computational results, Han et al. at NASA Ames Research Center has suggested that nanotube-based gear (Fig.1.1(b)) can be made and operated and the gears can work well if the temperature is lower than 600-1000 K and the rotational energy is less than the teeth tilting energy at 20°.

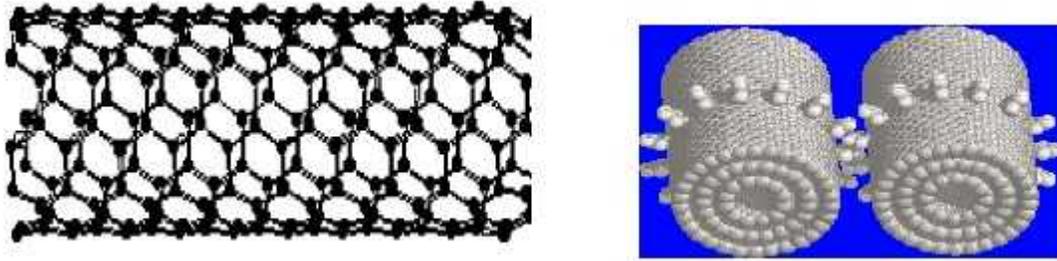


Fig. 1.1. (a) A carbon nanotube (b) A carbon nanotube-based gear

2 Principles of MEMS

A considerable amount of research is being carried out concerning the design and development of existing systems that reach down into micro- and nanometer scale levels. A technology that considers microscale sensors, actuators, valves, gears, and mirrors embedded in semiconductor chips is referred to as microelectromechanical systems, MEMS in short. In essence, MEMS are small, integrated devices that combine electronics, electrical as well as mechanical elements (Fig. 2.1). The size is in the order of a micrometer level. MEMS design technology is an extended form of traditional fabrication techniques used for IC (Integrated Circuit) manufacturing. MEMS add passive elements such as capacitors and inductors including mechanical elements such as springs, gears, beams, flexures, diaphragms, etc. MEMS are thus the integration of these elements on a single substrate (wafer) developed through more advanced microfabrication and micromachining technology. While the ICs are fabricated by the use of IC process, the mechanical micro components are fabricated using micromachining processes. This process helps in etching away the parts of the selected portions of the wafer. The process can also add new structural layers to form mechanical as well as electromechanical components. Thus, MEMS technology promises to revolutionise many products by combining microfabrication-based microelectronics with micromachining process sequences on silicon, making it possible for the realisation of a complete systems-on-a-chip (SoC). The technology allows for the development of smart systems and products inheriting increased computational capability, perception and control attributes. Smart systems can lead to expand the scope of possible solutions to diagnostics for target applications. It has been mentioned that microelectronic integrated circuits can be thought of as the brains of a system while MEMS augments the decision-making capability with eyes and arms, to allow microsystems to sense and control the environment. MEMS devices are manufactured by the use of batch fabrication techniques similar to those used for IC. Therefore, unparalleled levels of superiority, sophistication, functionality, reliability and availability can be achieved on a small silicon chip at a relatively low cost. Two important microsystems are microsensors and microactuators. Sensors gather information from the environment. The commonly used transduction principles are chemical, thermal, biological, optical, magnetic and mechanical phenomena. Accordingly, there are various types of microsensors. The integrated electronics process the information derived from the sensors. In many cases the decision-making logics are integrated into the devices. The decision is mostly transmitted to the actuator in order to achieve moving, positioning, regulating, pumping or filtering actions. In this way, the environment can be controlled depending on the desired purpose. The study of MEMS accommodates the topics listed below. These principles are presented in this chapter.

- Fabrication processes
- Mechanical sensors and actuators
- Thermal MEMS
- Magnetic MEMS
- Micro-opto-electromechanical systems (MOEMS)

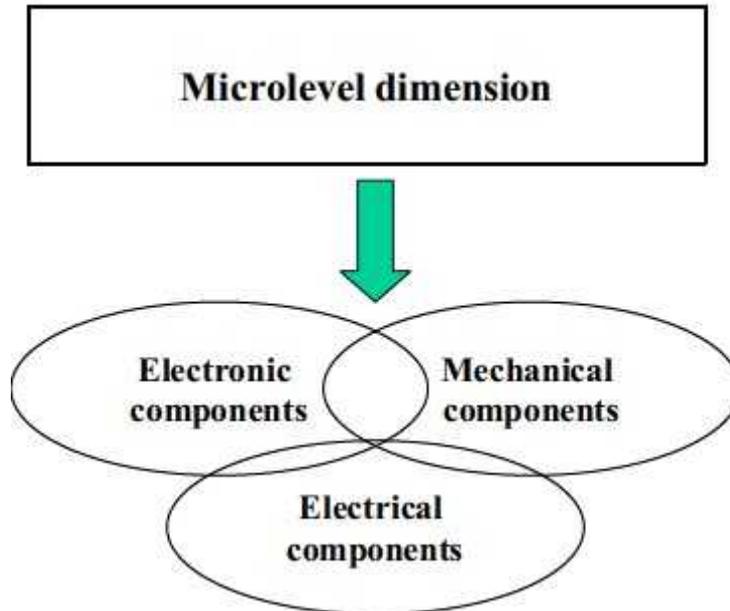


Fig. 2.1. Microelectromechanical systems

2.1 Driving Principles for Actuation

The driving principles used to drive various types of micromachined MEMS systems are primarily four types, namely,

- Electrostatic drive
- Magnetic drive
- Piezoelectric drive
- Electrothermal drive

Each driving principle has specific advantages and disadvantages with respect to deflection range, required force, environmental durability, and most importantly the response time. Furthermore, the required power supply mainly depends on the driving principles involved (Table 2.1).

Table 2.1 typical power requirements for respective driving principles

	Voltage	Current
Electrostatic Drive	tens of volts ~ hundreds of volts	nA ~ μ A
Piezoelectric Drive	tens of volts ~ hundreds of volts	nA ~ μ A
Electromagnetic Drive	about 1 V	hundreds of mAs
Electrothermal Drive	a few volts ~ tens of volts	mA ~ tens of mAs

Electrostatic drive is based on electrostatic forces between the microelectrodes. When an external voltage is applied between the electrodes, a potential energy is stored which enables the actuation. The electrostatic forces act perpendicular to the parallel electrode. Electromagnetic actuation is primarily a current controlled process. The driving mechanism again requires currents of the order of several hundreds of milliamps, and voltages in the range of less than one volt. Magnetic drive is an attractive driving principle and very suitable for applications like dustfilled environments and in environments where low driving voltages are acceptable or desired. Piezoelectric driving is based on the material properties of crystals, ceramics, polymers, and liquid crystals. In a piezoelectric material, the internal dielectric displacement is developed via an applied electric field and mechanical stress. Electrothermal devices use electrically generated heat as an energy source for actuation. The electrothermal effects can be divided

into three different categories: shape memory alloys, electrothermal bimorphs, and thermopneumatic actuators. Many electrostatic MEMS actuators that are investigated include micromotors, comb drive actuators and microvalves. Accelerometers, ink jet printer heads, color projection displays, scanning probe microscopes, pressure, temperature, chemical and vibration sensors, light reflectors, switches, vehicle control, pacemakers and data storage devices are examples of high-end applications of MEMS sensors and actuators.

2.2 Fabrication Process

MEMS support the principle of miniaturisation, multiplicity, and microelectronics. Miniaturisation is preferred in order to achieve faster response times and less space, whereas multiplicity is essential in order to advocate batch fabrication. Batch processing is a processing technique that allows for thousands of components to be simultaneously manufactured in order to significantly reduce the cost of the device. Microelectronics embeds the real part of the MEMS device. These can be sensors, actuators, signal conditioning, signal processing or even logic circuits. The design process involved for IC manufacturing is called microfabrication. The sequences of microfabrication include film growth, doping, lithography, etching, dicing and packaging. A polished silicon wafer is mainly used as the substrate. A thin film is grown on the substrate. Then the properties of the layer are modulated by appropriately introducing doped material in a controllable manner. The doping can be achieved by thermal diffusion. The subsequent process is called lithography, which refers to the creation of a masking pattern. The pattern on a mask is transferred to the film by means of a photoresist. A mask usually consists of a glass plate, which is coated with a patterned layer, which is usually chromium film. The subsequent process is called etching. Etching is a process of removing the portions of material from an insulating base by chemical or electrolytic means. The two types of etching are wet and dry etching. In wet etching the material is dissolved by immersing it in a chemical solution. On the other hand, in dry etching, the material is dissolved by using reactive ions or a vapor phase etchant (Vittorio 2001). The finished wafer has to be segmented into a small dice-like structure. Finally, the individual sections are packaged. Packaging is a complex process that involves physically locating, connecting, and protecting a component or whole device. MEMS design also considers all the process sequences employed for microfabrication, however in this case it is considered as an extension of the IC fabrication process. The following methods are common as far as manufacturing of MEMS designs are concerned:

- Bulk micromachining
- Surface micromachining
- Micromolding

Bulk micromachining makes micromechanical devices by etching deeply into the silicon wafer. It is a subtractive process that involves the selective removal of the wafer materials to form the microstructure, which may include cantilevers, holes, grooves, and membranes. The majority of currently used MEMS processes involve bulk etching. In light of newly introduced dry etching methods, which are compatible with complementary metal oxide semiconductors, it is unlikely that bulk micromachining will decrease in popularity in the near future (Kovacs 1998). The available etching methods fall into three categories in terms of the state of the etchant: wet, vapor, and plasma. The etching reactions rely on the oxidation of silicon to form compounds that can be physically removed from the substrate (Kovacs, 1998). Conversely, surface micromachining technology makes thin micromechanical devices on the surface of a silicon wafer. The surface micromachining sequences are as follows,

- Wafer cleaning
- Blanket n⁺ diffusion of Si substrate
- Passivation layer formation
- Opening up of the passivation layer for contacts
- Stripping of resist in piranha
- Removal of thin oxide through BHF etchant systems
- Deposition of a base, spacer or sacrificial layer using phosphosilicate glass (PSG)

- Densification at 950°C for 30-60 min in wet oxygen
- Base window etching in BHF for anchors
- Deposition of structural material deposition (e.g., poly-Si using CVD method at about 600°C, 100 Pa and 125 sccm at about 150 Å/min)
- Anneal of the poly-Si at 1050°C for 1 hour to reduce stress in the structure
- Doping: in-situ, PSG sandwich and ion implantation
- Release step, selective etching of spacer layer.

The micromolding process involves use of molds to define the deposition of the structural layer. In this case, the structural material is deposited only in those areas constituting the microdevice structure. This is apparently in contrast to both bulk and surface micromachining processes. Feature blanket deposition of the structural material followed by etching to realise the final device geometry is done in one step. Once the structural layer deposition is over the mold is dissolved by using a chemical etchant. Note that the etchant does not corrugate the structural material. One of the most widely used micromolding processes is the LIGA process. LIGA is a German acronym standing for lithographie, galvanofornung and abformung, or lithography, electroplating, and molding. Photosensitive polyimides are mostly used for fabricating plating molds.

2.4 Mechanical MEMS

2.4.1 Mechanical sensors

MEMS mechanical sensors are very popular because of their easy integration procedure. The sensing mechanism utilises the following methods and principles,

- Cantilever beam sensor
- Capacitive sensing
- Accelerometers
- Microphones
- Gyroscopes
- Piezoelectricity

2.4.2 Accelerometer, Cantilever and Capacitive Measurement

Cantilever sensors can be used for the detection of physical, chemical and biological analytes with relatively good sensitivities and selectivity. The vast application areas include: acoustic applications, vibration monitoring, detection of viscosity and density, infrared and UV radiation, magnetic and electric fields, detection of chemical vapours including medical and biological agents, contaminants in water, explosive vapours such as RDX, PETN and TNT, nuclear radiation and the detection of DNA, antibodies and pathogens.

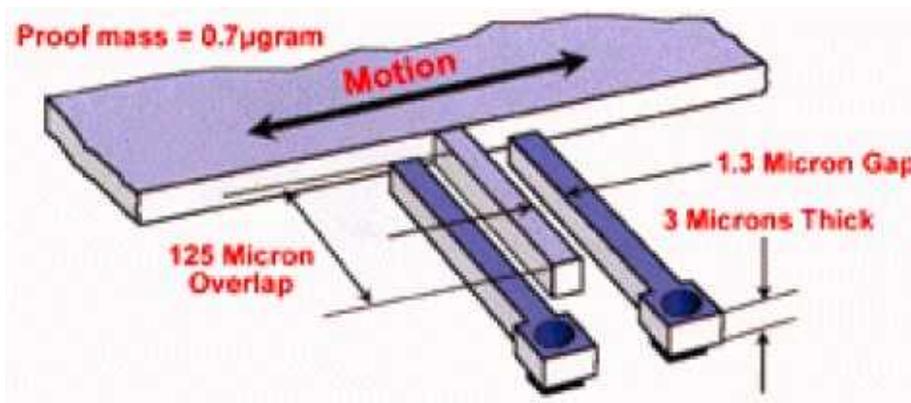


Fig. 2.2. Cantilever beam for the measurement of static and dynamic acceleration

Fig. 2.2 shows a typical cantilever sensor that can measure acceleration. The body of the sensor (proof mass) could be about 0.5-0.7 micrograms. The proof mass moves in the X- and Y-axes. Polysilicon springs suspend the MEMS structure above the substrate facilitating the proof mass to move freely.

Acceleration causes deflection of the proof mass from its centre position. There could be up to 32 sets of radial fingers around the four sides of the square proof mass. The fingers (middle one), shown in the figure, are positioned between two plates that are fixed to the substrate. Each finger and pair of fixed plates constitutes a differential capacitor, and the deflection of the proof mass is determined by measuring the differential capacitance. This sensing method has the ability of sensing both dynamic acceleration such as shock or vibration, as well as static acceleration such as inclination or gravity. Many accelerometers employ piezoelectric sensing techniques. Thin film piezoelectric materials such as lead zirconate titanate (PZT) are promising materials for MEMS applications due to their high piezoelectric properties. Piezoelectric polymers are now being used for sensor applications. Piezoelectric polymeric sensors offer the advantage of strains without fatigue, low acoustic impedance and operational flexibility. The PZT converts mechanical disturbances to electrical signals. The starting material for the front-side process (FSP) is a silicon wafer that has silicon dioxide, lower metal electrode (Ti/Pt), and deposited PZT films. Industrial applications of MEMS accelerometers include airbag release mechanisms, machinery failure diagnostics, and navigational systems.

2.4.3 Microphone

Acoustic MEMS are air-coupled, and can offer a wide range of applications such as detection, analysis and recognition of sound signals. The basic component is the microphone, called micro-microphone or simply MEMS microphone. A simple definition of a microphone is that it is an electromechanical acoustic transducer that transforms acoustical energy into electrical energy. It is an ultrasonic microsensor, which takes advantage of miniaturisation and also consumes low power. When multiple microphones are arranged in an array, the device is referred to as smart system as it can offer more reliable and intelligent operations. The basic challenge that is encountered in designing the MEMS microphone is the formulation, design, and implementation of signal processing circuitry that can adapt, control and utilise the signal in noisy environments. MEMS microphones are also very suitable for outdoor acoustic surveillance on robotic vehicles, wind noise flow turbulence sensing, platform vibration sensing, and so on.

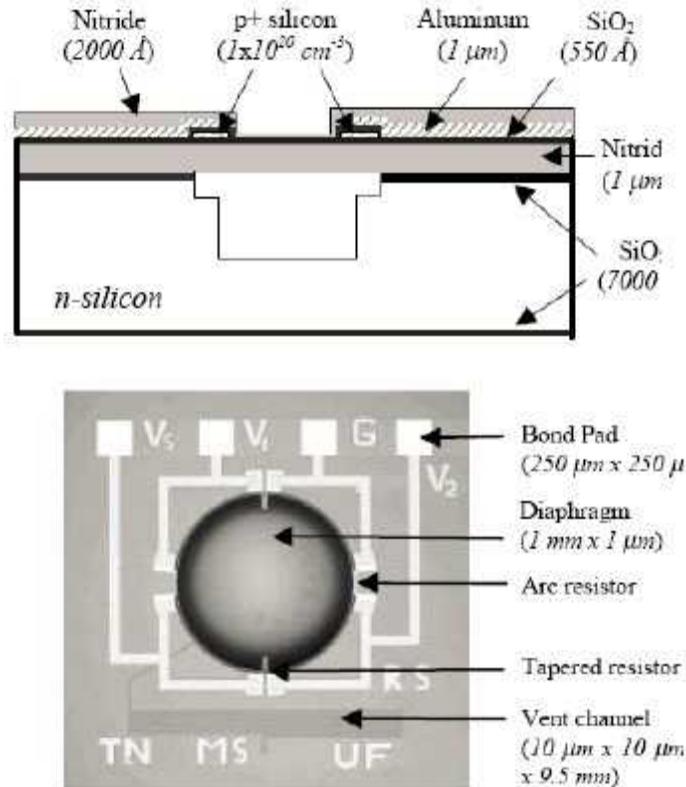


Fig. 2.3. Cross-sectional diagram of piezoelectric microphone; plan view

The ultrasonic sensor is based on the mechanical vibration of micro membrane or diaphragm realised in the silicon platform. The diaphragm is a thin, circular membrane held in tension and clamped at the edge. Many designs use a flat free plate that is held in proximity to the back plate by electrostatic attraction. The free plate makes up a variable capacitor with the back plate. Its value changes during the vibration caused by the sound signal. The deformation or deviation of the membrane (free-plate) from the normal values depends on the amplitude of the incident pressure. Fig. 2.3 shows the schematic as well as a micromachined SEM (Scanning Electron Microscope) picture of a MEMS microphone. The microphone can have a very low stray capacitance, is self-biasing, mass producible, arrayable, integrable with on-chip electronics, structurally simple and extremely stable overtime in an ordinary environment. The typical dynamic range is from 70 to 120 dB SPL and the sensitivity can be in the order of 0.2 mV/Pa over the frequency range 100-10 kHz.

2.4.4 Gyroscope

MEMS gyroscopes (Fig. 2.4) are typically designed to measure an angular rate of rotation. A measurement of the angle is useful in many applications. A very common application is the measurement of the orientation or tilt of a vehicle running in a curved path. The MEMS gyroscope design introduces sophisticated and advanced control techniques that can lead to measure absolute angles. Some design is based on the principle of measuring the angle of free vibration of a suspended mass with respect to the casing of the gyroscope. The gyroscope can accurately measure both the angle and angular rate for low bandwidth applications. The measurement of orientation, for instance, is very useful in the computer-controlled steering of vehicles as well as for differential braking systems for skid control in automobiles. A typical gyroscope consists of a single mass, oscillating longitudinally with rotation induced lateral deflections being sensed capacitively. The iMEMS ADXRS gyroscope from Analog Devices, Inc., integrates both an angular rate sensor and signal processing electronics onto a single piece of silicon. Mounted inside a 7x7x3 mm BGA package, the gyro consumes 5 mA at 5 V and delivers stable output in the presence of

mechanical noise up to 2000 g over a reasonably wide frequency range. A full mechanical and electronic self-test feature operates while the sensor is active. Mechanical sensors inherit many drawbacks as follows,

- The overall silicon area is generally larger
- Multi-chip modules require additional integration steps
- Existence of larger signals from the sensor output
- Stray capacitance of the interconnections
- Comparatively larger volume with respect to packages

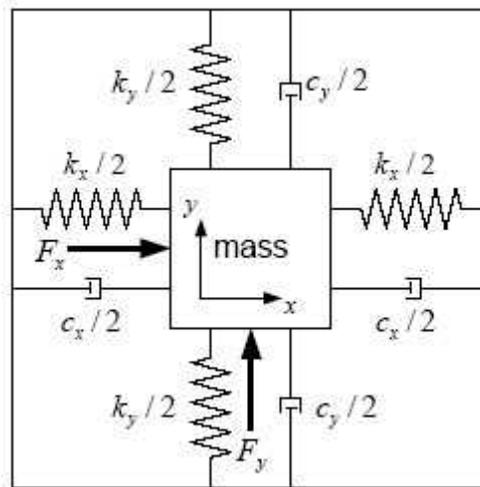
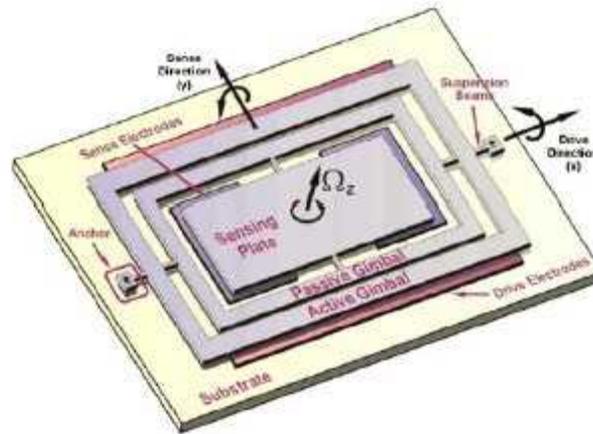


Fig. 2.4. (a) Conceptual schematic of a torsional micromachined gyroscope with nonresonant drive; (b) Schematic diagram of a vibratory gyroscope

2.4.5 Mechanical Actuators

One of the fundamental components in MEMS technology is the actuator. A mechanical actuator is a device that usually converts electrical signals into mechanical motion. For instance, if a voltage is applied to a quartz crystal, it will change its size in a very precise and predictable way. Cantilever structures are used for actuation purposes. Cantilevers bend when pressure is applied to them and oscillate in a way similar to a spring when properly put into place. The important point that is considered in electrostatically actuated cantilever based MEMS devices is their stability.

2.5 Thermal MEMS

Microsystems whose functionalities rely on heat transfer are called thermal MEMS. Thermal MEMS can be sensors and actuators. When a physical signal to be measured generates a temperature profile, the principle is called thermal sensing. The physical signals could be heat radiation, non-radiant heat flux or

areaction heat. Gas pressures, masses, volume fluxes, and fluidic thermalconductivities are measured using the principle of thermal sensing. As always,sensing or actuation is a process of energy conversion, called transduction. Thetransduction effect is realised in three ways: the thermoelectric effect (Seebeckeffect), the thermoresistive effect (bolometer effect) and the pyroelectric effect.

In 1821, Thomas Johann Seebeck discovered that a compass needle wasdeflected when it was placed in the vicinity of a closed loop formed from twodissimilar conductors and the junctions maintained different temperatures. This isas good as saying that a voltage (and hence the current) is developed in a loopcontaining two dissimilar metals, provided the two junctions are maintained atdifferent temperatures. The magnitude of the deflection is proportional to thetemperature difference and depends on the material, and does not depend on thetemperature distribution along the conductors. The effect is called thethermoelectric effect, and is the basis of the thermocouple, a real temperaturemeasuring device. The opposite of the Seebeck effect, in which current flowcauses a temperature difference between the junctions of different metals, is thePeltier effect. The reversed Seebeck effect is thus the Peltier effect. A thermopileis a serially interconnected array of thermocouples. Thermopiles are used forachieving better sensitivity.

Pyroelectricity is the migration of positive and negative charges to oppositeends of a crystal's polar axis as a result of a change in temperature. The cause isreferred to as electric polarisation. The polarisation phenomenon within thematerial further states that below a temperature, known as the Curie point,crystalline materials or ferroelectric materials exhibit a large spontaneouselectrical polarisation in response to a temperature change. Materials, whichpossess this property, are called pyroelectric materials. The change in polarization is observed as an electrical voltage signal and appears if electrodes are placed onopposite faces of a thin slice of such material. The design can be thought of as atypical form of a capacitor. Cooling or heating of a homogeneous conductorresulting from the flow of an electrical current in the presence of a temperaturegradient is known as the Thomson effect. It is hence defined as the rate of heat generated or absorbed in a single current carrying conductor, which is subjected toa temperature gradient. Arne Olander observed that some alloys such as NiTi,CuZnAl, CuAlNi, etc. change their solid-state phase. The property is reflected aspseudo-elasticity and the shape memory effect. After alloying and basicprocessing, the alloy can be formed into a desired shape, a coil for example, andthen set to that shape by a heat treatment. When the shape is cooled, it may bebent, stretched or deformed and then with subsequent re-heating (which should bethe heat setting temperature) the deformation can be recovered. Some of the main advantages of shape memory alloys (SMA) are that they are biocompatible with good mechanical properties such as strong and corrosion resistant and can beapplied to diverse actuator applications.

2.5.1 Thermometry

Rigorous thermal analyses and experiments that assist in designing MEMSstructures are in progress. Some effort has been made on liquid-crystalthermometry of micromachined silicon arrays for DNA replication. This work isbeing carried out at Perkin-Elmer Applied Biosystems. Replication is achievedthrough what is known as the polymerase chain reaction (PCR) process. PCRrequires accurate cycling of the liquid sample temperature with an operationalrange between 55 and 95oC. PCR that makes use of micromachined structures isshown in the Fig. 2.5(a). The structure assures uniformity as far as temperatureand cycle timing is concerned. It also utilises less reagent and samplevolumes. Thermal design requires measurement of the temperature distribution inthe reacting liquid. The measurement is possible by encapsulating the liquidcrystals suspended in the liquid. This in turn led to the measurement of thetemperature uniformity and the time constant for about 20 vessels in amicromachined silicon array. Two separate sets of crystals are used to imagetemperature variations near the two processing temperature thresholds with aresolution of 0.1oC. While the thermometry technique described above is usefulfor characterizingmicrofabricated PCR systems, it can also support thermal designs of a broad variety of MEMS fluidic devices.

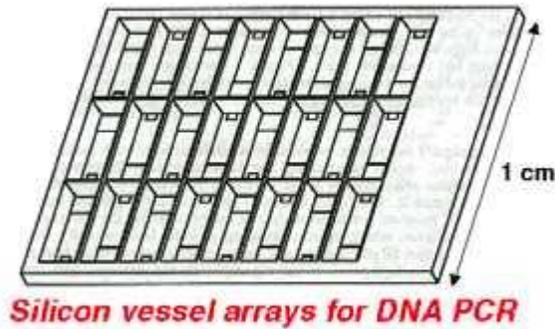


Fig. 2.5. (a) Liquid-crystal thermometry based micromachined vessel array for thermal processing of DNA using polymerase chain reaction (PCR)

Thermally activated systems are mostly employed in the bio-analytical microsystems or *lab-on-a-chip* (loac) devices. Several key issues in existing and emerging bio-analytical microsystems are directly related to thermal phenomena. Research and development on loac technology is essentially directed toward miniaturisation and the integration of chemical and biochemical analysis tools for manipulation, handling, processing and analysis of samples in a single integrated chip. In order to develop a highly sensitive and responsive thermal device, the thermal mass of the transducer element is kept as low as possible. This is only achieved by using thin film structures made by micromachining techniques.

2.5.2 Data Storage Applications

There are many other applications for thermal designs. One of the more important examples is the design of silicon cantilevers for high-density thermo-mechanical data storage applications. The design method can use the principle of the atomic force microscope (www.stanford.edu/group/microheat/seven.html). MEMS based scanning probe data storage devices are emerging as potential ultra-high-density, low-access-time, and low-power alternatives to conventional data storage devices. The implementation of a probe based storage system uses thermo-mechanical means, as described below, to store and retrieve information in thin films. The design and characterisation of a servomechanism to achieve precise positioning in a probe based storage device is extremely critical. The device includes a thermal position sensor that provides position information to the servo controller. The research work by Kenny and Chui at Stanford University in collaboration with IBM is based on the design of a microcantilever tip that exerts a constant force on a polycarbonate sample and induces localised softening and deformation during the heating phase. This is caused by a bias current along the cantilever. The resulting serrations serve as data bits, which are read by the use of another separate cantilever integrated with a piezoresistive displacement sensor (Fig. 2.5(b)) followed by measuring circuitry. The cooling time constant of the heated cantilever tip governs the rate at which it can achieve sub micrometer writing. One of the challenges in building such devices is the accuracy and the latency required in the navigation of the probe.

Data storage cantilever

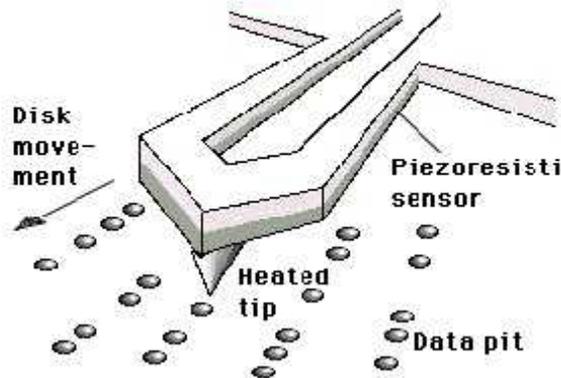


Fig. 2.5. (b) AFM based single-crystal silicon cantilevers tip for data storage

2.5.3 Microhotplate Gas Sensors

Semiconducting SnO_2 films when deposited on microhotplate platforms can detect gas species such as hydrogen and methanol. Microhotplates are thermally isolated micromachined platforms with integrated temperature sensing and actuation for closed-loop thermal control. Typical dimensions of a thermal platform have lateral dimensions of less than 100 micrometer, and are suspended over a bulk-etched cavity for thermal isolation. The physical architecture of each microhotplate consists of a polysilicon heater, a thermoresistive film for temperature measurement, and a semiconducting SnO_2 film, which exhibits a change in conductance with the adsorption of chemical species. With aluminum as the temperature sensing film, the microhotplate can operate up to 500°C . Thermal response time for a 100-micron wide microhotplate has been measured at around 0.6 ms, with a thermal efficiency of 8 C/mW.

2.5.4 Thermoactuators

Thermal actuation is another scenario with respect to its counterparts including electrostatic and piezoelectric types. Thermal actuation is based on electrothermal energy density transformation which is given by, $E = V^2 / \alpha L^2$, where V is the applied voltage, α is resistivity and L is the effective length of the actuator element. Microactuators based on electrothermal principles can offer significant energy density compared to electrostatic microactuators. It is worth mentioning that for some configurations, electrothermal actuation provides 1 to 2 orders of magnitude higher energy density than piezoelectric actuation, and 4 orders of magnitude higher energy density than electrostatic transduction. Several thermal microactuator geometries have been investigated, including U-beam and V-beam geometries, respectively. Piezoelectric microactuators usually provide a fraction of the energy density of electrothermal elements.

2.6 Magnetic MEMS

The use of magnetic materials in MEMS is a recent development in which particular emphasis is given on ferromagnetic materials. These magnetic materials could be soft or hard. The design of ferromagnetic MEMS and the methods of integrating both soft and hard magnetic materials with it are a current research and development field. Soft ferromagnetic materials have found their usefulness in microsensors, microactuators, and microsystems. However, hard magnetic materials have unique advantages that are driving their integration into other applications.

Patterns of hard magnetic materials in the micron scale are of interest for the novel design of magnetic recording media. Hard magnetic films with a thickness of several microns are grown by the sputtering technique, practically a difficult process. In order to overcome the problem, electrodeposited Co₈₀Pt₂₀ alloys are grown up to a several micron thickness while maintaining their hard magnetic properties. The batch fabrication process for micromachining thick films of high performance hard magnetic materials is improving. Subtractive etching processes are needed to define the patterns of the hard materials. Micron size patterns can also be obtained by optical lithography, thus allowing a better control of the magnetic

properties. Many researchers are currently trying to achieve a submicron size for the production of patterned media. Principles of magnetic MEMS are described below.

Anisotropic magnetoresistive (AMR) sensors allow for detecting the strength and direction of magnetic fields. They are used for the measurement of distance, proximity, position, angle and rotational speed. The AMR sensors undergo a change in resistance in response to an applied magnetic field vector generated by passing the current in a coil. When a magnetic field is applied, the magnetization rotates toward the field. The variation in resistance depends on the rotation of magnetisation relative to the direction of current flow. The change in resistance of the material used in the AMR sensor is highest if the magnetisation is parallel to the current and lowest if it is perpendicular to the current. The sensitivity to the direction of the magnetic field allows for the measurement of angle also. The development of micromachined magnetic devices has relied primarily on the use of nickel-iron permalloy. Permalloy has a low magnetic anisotropy. The anisotropy field for permalloy is between 3 and 5 oersted. Permalloy is used in a number of applications since it has good soft magnetic properties, high permeability, high magnetoresistive effect, low magnetostriction, stable high frequency operation, and excellent mechanical properties. In hard disk magnetic recording heads, permalloy is widely used for magnetoresistive sensors and flux guiding elements. Devices such as magnetic separators, micropumps, magnetic micromotors, inductors, switches, and microrelays have also been fabricated using permalloy as the magnetic material as well as in moving members.

Permanent magnets are used for sensors and actuator applications that can provide the desired constant magnetic field without the consumption of electrical energy. Another important characteristic is that they do not generate heat. Furthermore, the energy stored in a permanent magnet does not deteriorate when the magnet is properly handled and micromachined. This feature is due to the fact that it does not interact with its surroundings. Moreover, permanent magnets can achieve relatively high energy density in microstructures, as compared to other energy storage devices. This is why there has been a growing interest in the realisation of permanent magnets in MEMS devices recently. Apparently, magnetic MEMS devices have several advantages over electrostatic types. One of the important advantages is the generation of long-range force and deflection with low driving voltage in harsh environments. MEMS devices with a permanent magnet have benefits of low power consumption, favorable scaling and simple electronic circuitry. Lagorce et al. introduced screen-printing technology to integrate a permanent magnet in a microactuator. Actuators are capable of generating large bi-directional forces with long working lengths. These permanent magnets were screen printed on a copper cantilever beam using a magnetic paste composed of epoxy resin and strontium ferrite particles. Coercivity of 350 kA/m and residual induction of 65 mT have been reported in a disk type permanent magnet with a millimeter range diameter and 100 micron in thickness.

A number of micromachined magnetic actuators require materials with desirable magnetic properties such as high permeability and a thickness in the range of several micrometers. Investigations on the use of rare-earth magnetic powders in microsystems are in progress. The powders are essentially deposited by screen printing or other methods to fabricate strong permanent magnets with wafer level processes. Deflections of over 900 micron have been reported with surface micromachined cantilevers, magnets and pancake coils. Magnets with typical volumes measuring in the range of 0.11 mm³ can produce forces in the mN range. A typical magnetically actuated microcantilever incorporating a screen printed magnet measures 2000x1000x100 microns. Deflection at the tip of the cantilever is about 1000 micron, subjected to a driving current of 200 mA which is passed through a 40 turn pancake coil. Large deflection causes nonlinear performance similar to electrostatic systems, however, small deflection shows a better linear current-deflection relationship.

A magnetic actuator consists of a magnetic field source and a magnetically susceptible unit. The actuation force is proportional to the magnetic field intensity, the magnetic susceptibility, and the mass of magnetic susceptible material. A large-force, fully integrated, electromagnetic actuator for microrelay applications is reported by Wright (Fig. 2.6). The actuator has a footprint of less than 8mm² and its fabrication is potentially compatible with CMOS processing technology. It is designed for high efficiency actuation applications. The actuator integrates a cantilever beam and planar electromagnetic coil into a low-reluctance magnetic circuit using a combined surface and bulk micromachining process. Test results show that a coil current of 80 mA generates a 200 μ N actuation force. Theoretical extrapolation of the data, however, indicates that a coil current of 800mA can produce an actuation force in the millinewton range.

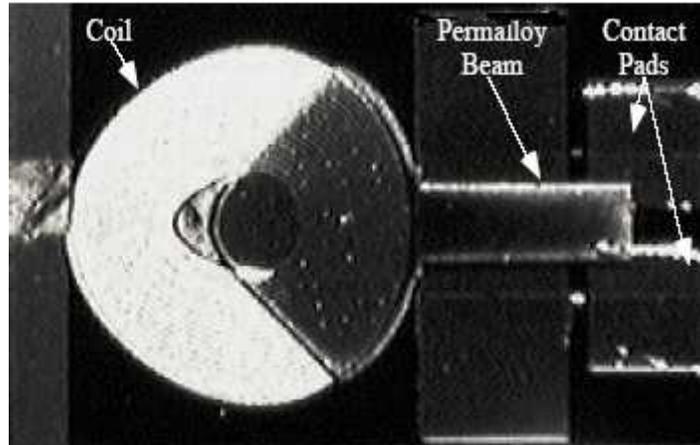


Fig. 2.6. Top view of a magnetic actuator developed by Write, et.al

An example of a bi-directional magnetic actuator used for optical scanning applications can be given. It is composed of a silicon cantilever beam and an electromagnet. At the tip of the cantilever beam, a permanent magnet array is electroplated in order to achieve the bi-directional actuation. Below the cantilever beam, the permanent magnet array is placed along the axis of the electromagnet. For a large bi-directional deflection and dynamic scanning capability, Cho has designed an optical scanner supported by two serpentine torsion bars. A schematic diagram of the scanner is illustrated in Fig. 2.7. The scanner is designed to have a silicon mirror supported by two serpentine torsion bars (Fig. 2.8), carrying a permanent magnet made by bumper filling at its tip on the opposite side.

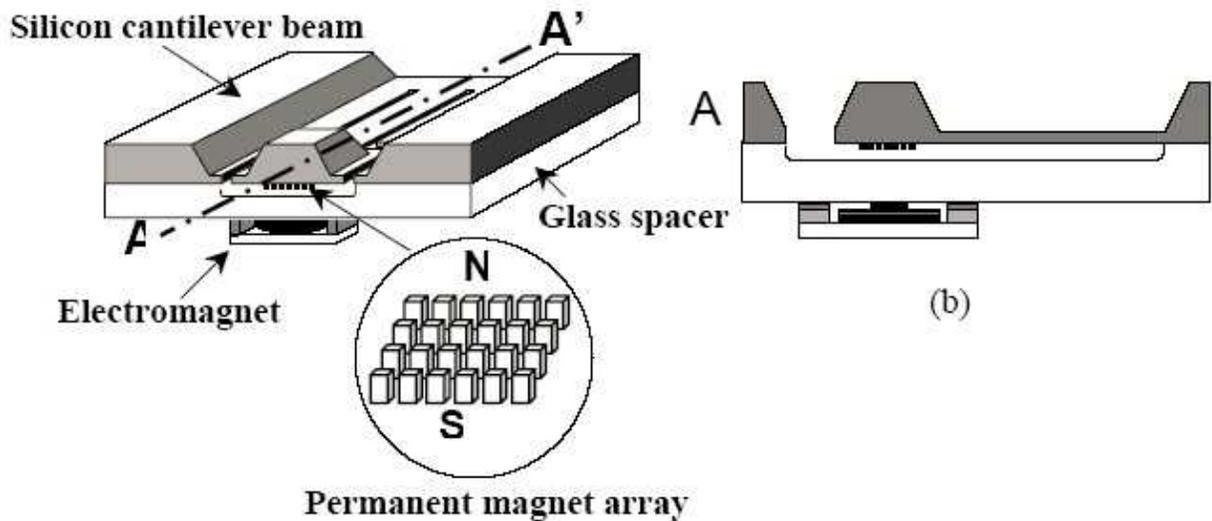


Fig. 2.7. Cantilever beam magnetic microactuator. (a) Schematic view and (b) Another view

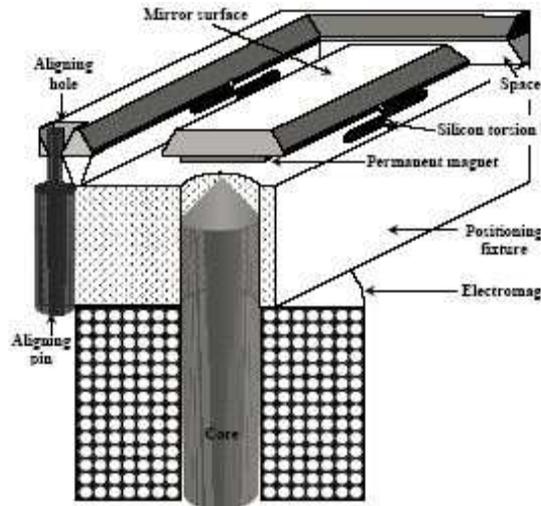


Fig. 2.8. Magnetically driven optical scanner

Son and Lal reported a novel magnetic actuator that allows remote magnetic actuation with piezoresistive feedback for microsurgery applications. The actuator consists of an electromagnet, a ferromagnetic mass, and a cantilever. Conventional actuator fabrication lithography combined with electroplating is required to make high aspect ratio structures. However, high aspect ratio columns with a suspension of ferromagnetic nanoparticles and epoxy using magnetic extrusion were fabricated. The fabrication process did not require lithography. The remote magnetic actuator with feedback consists of two basic units. They are the actuation unit and feedback unit. The ferromagnetic mass is first placed on the tip of the cantilever. The electromagnet is placed above the ferromagnetic mass. If AC current is applied to the electromagnet, the magnet repeats pulling and releasing the cantilever. If the cantilever is released, the bent cantilever returns to the initial position by its own spring force. One implied factor is that if the frequency of current applied to the electromagnet equals the resonance frequency of the cantilever, the actuation is amplified by the Q factor of the mechanical resonance. The resonant frequency of the cantilever with mass can be varied due to the change of its mechanical boundary conditions. The feedback unit consists of three sub-units, namely the strain gauge, amplifier and voltage controlled oscillator (VCO). When the cantilever vibrates at different frequencies, the value at the strain gauge decreases. The drop in the value of the strain gauge triggers the VCO to adjust the frequency using linear feedback to find the new f_r (Fig. 2.9).

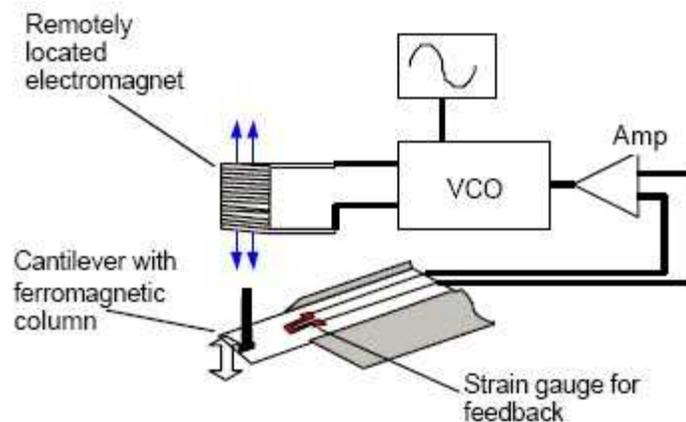


Fig. 2.9. Remotely actuated magnetic actuator

Nano-machining or processing systems with nanometre accuracies

1.1 Processing unit, breaking stress, and processing energy density

Lately it has become necessary to manufacture intelligent precision products with extremely high precision and fine construction, of the order of nanometre accuracies. Clearly, to produce such high-precision products, ultra-precision processing systems must be used to ensure extremely small deviational and scattering errors, below 1 nm. To ensure such high precision and fine resolution, processing units in the sub-nanometre range or atomic-bit size must be applied. The processing unit corresponds in size to one bit of chip in removal processes, one step in deforming processes and one molecular-cell cluster in consolidation processes.

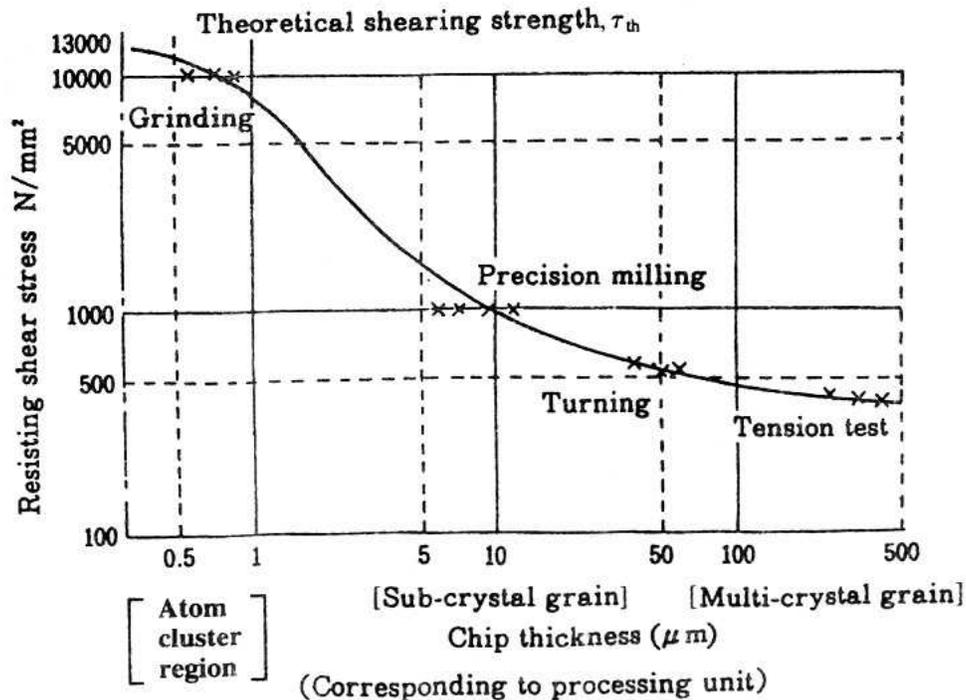


Fig. 1.1.1. Relation between chip thick-ness and resisting shear stress for carbonsteel SAE 1112.

1.1.1 Breaking stress in the atomic lattice range

When processing units of atomic-bit size are used, a serious problem arises: the resisting shear or breaking stress τ_s (N mm⁻²), or the specific shearing energy δ_s (J cm⁻³), becomes extremely large. An example of the dependence of the resisting shear stress on chip thickness in carbon steel is shown in Fig. 1.1.1. The curve shows that as the chip thickness becomes smaller, the resisting shear stress at the cutting edge of a solid-bite tool or grinding abrasive grain becomes extremely large, approaching the theoretical shear stress τ_{th} in the defect-free material or the atomic bonding strength of carbon steel:

$$\tau_{th} = G / 2\pi = 1.3 \times 10^4 \text{ N mm}^{-2}$$

where $G = 8.2 \times 10^4 \text{ N mm}^{-2}$: the modulus of rigidity of carbon steel (see subsection 1.3.6). The reason why the resisting shear stress becomes so large at the solid tool edge at atomic-bit size is that there are only point defects to initiate the breakage of the atomic bonding structure. However, as shown schematically in Fig. 1.1.2, in ductile metals, breakdown slip for processing units between 0.1 and $10 \mu\text{m}$ originates in relatively easily movable dislocations in the metal's crystal grain, where the mean distributed interval of movable dislocation is about $1 \mu\text{m}$; in its crystal grain of brittle ceramics, breakdown occurs due to microcrack defects which are also distributed at a mean interval of about $1 \mu\text{m}$.

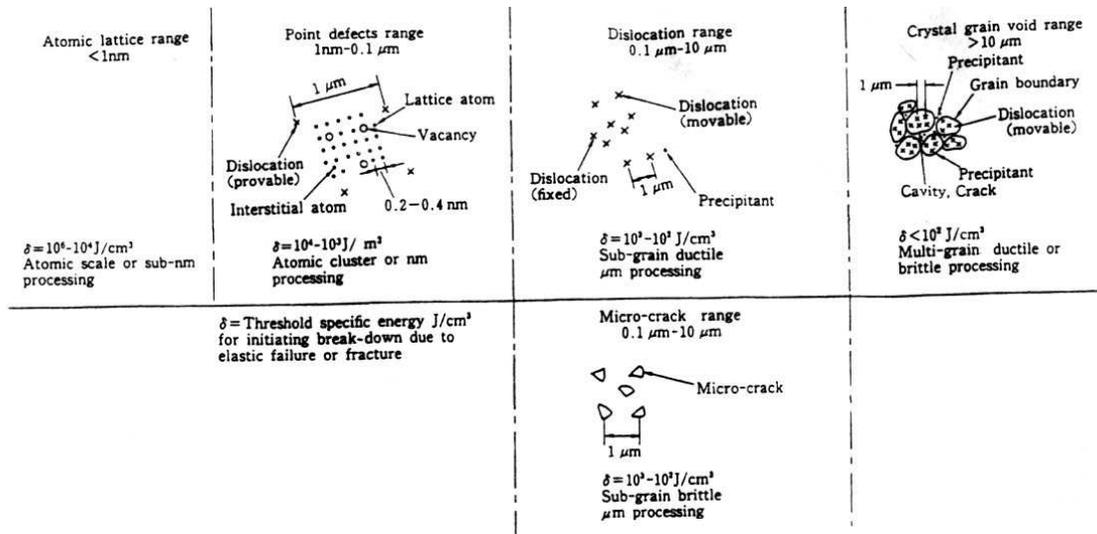


Fig. 1.1.2. Distribution of defects in materials: failure due to movable dislocations in ductile materials and fracture due to microcracks in brittle materials.

For processing units larger than $10 \mu\text{m}$, the breakdown of ductile metals due to shear slip begins at a weak point at a grain boundary or cavity; in brittle ceramics, the breakdown due to brittle fracture occurs mainly from cracks around the grain boundaries. Therefore machining with the sharp edge of an ordinary solid tool or the fixed abrasive of a grinding wheel cannot produce chips of fine atomic-bit sizes, because the cutting edges wear quickly due to high resistive stress, but diamond tools and abrasives can be used for mirror cutting or grinding because of their greater wear resistance. Moreover, lapping and polishing using replenishable free abrasives can be used to realize atomic-bit processing of materials.

1.1.3 Processing methods with atomic-bit and atom-cluster processing units

In order to obtain ultrahigh — precision and fine parts with nanometre accuracies, it is generally necessary to use an atomic-bit or atom-cluster processing unit. Atomic-bit processing is realized by atom-by-atom treatment of materials and hence achieves resolutions of sub-nanometre order, whereas atom-cluster processing involves clusters of atoms and so the resolution is in the region

of several nanometres. As already indicated in Table 1.1.1 and Fig. 1.1.3 the processing energy density δ for atomic bits at the processing point must reach $10^4 - 10^6 \text{ J cm}^{-3}$,

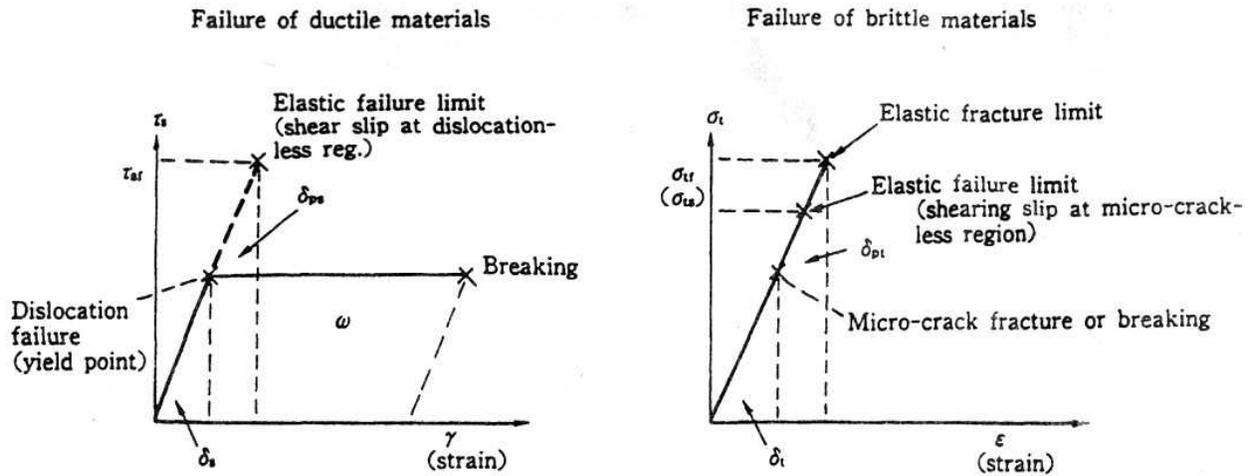


Fig. 1.1.3. Elastic failure and fracture limit, ω : specific stock removal energy (J cm^{-3}) $\gg \delta_s$; : elastic failure limit at dislocationless region = τ_{th} ; $\delta_s = \delta_{ps}$: specific energy for shear slip; σ_{tf} : elastic fracture limit without microcrack = $E / 2\pi = \sigma_{th}$; σ_{ts} : equivalent elastic failure limit in microcrackless region: shear slip; $\delta_t = \delta_{pt}$: specific energy for tensile breakage.

Which corresponds at the microscopical level to the specific volumetric lattice bonding energy U_b (MJ m^{-3} or J cm^{-3}) or atomic bonding energy E_b (J/atom) as given in table 1.1.2. The values for the abrasives Al_2O_3 , SiC and diamond are two or three orders of magnitude higher than for Fe; for atom-cluster processing, the values may be one to two orders of magnitude lower.

(a) Atom-cluster processing with free fine abrasives

Ordinary solid cutting tools and abrasive grinding wheels cannot be used for cutting and grinding with processing units of atom cluster, because wear of the cutting edge becomes extremely high for tough stock material such as steel. For ductile light metals such as aluminum, however, it is possible to realize micro-machining based on atom-cluster processing using diamond tools or diamond powder (see Fig. 1. 1.4).

Table 1.1.2 Specific volumetric lattice bonding energy U_b and atomic bonding energy E_b , and hardness values of materials*

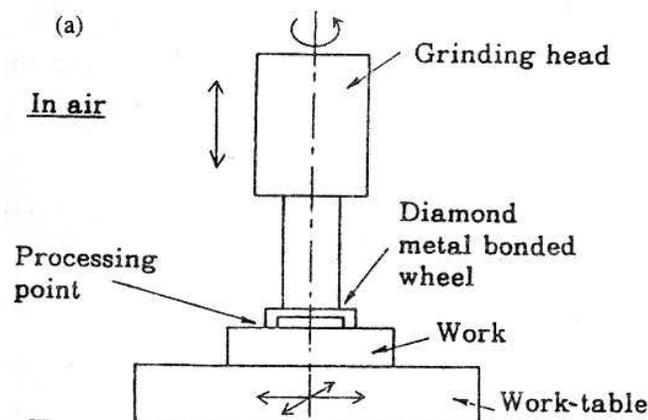
Material	U_b (MJ m^{-3})	E_b (J/atom)	(eV/atom)	Remarks	Knoop indentation hardness (10 MPa)	Scratch hardness Morse Woodwell
Fe	2.6×10^3	1.6×10^{-20}	0.1	for tension	(200 hardened 700-800)	

	(1.03×10^3)	(8×10^{-21})	0.05	for shear		
SiO ₂	5×10^2	4.24×10^{-19}	2.65	for shear	820	
Al	3.34×10^2	2.06×10^{-21}	0.013	for shear		
Al ₂ O ₃	6.2×10^5	5.26×10^{-18}	32	for tension	1600-2050	9
Si	7.5×10^5	1.59×10^{-17}	36	for tension	2400-2550	11
SiC	1.38×10^6	1.1×10^{-17}	67	for tension	2400-2550	14.0
cBN	2.09×10^6	1.07×10^{-17}	106	for tension	3000-3200	19.0
B ₄ C	2.26×10^6	1.8×10^{-17}	111	for tension	2700-2800	19.7
Diamond I (natural)	5.64×10^6	4.5×10^{-17}	274	abundant N	8000-8500	71.0
Diamond II	1.02×10^7	8.2×10^{-17}	513	N-free	(5700-10 400)	42.5

In addition to micro-machining, lapping using fine free abrasives of diamond, Al₂O₃, SiC, etc., and polishing using fine free abrasives of Fe₂O₃, Cr₂O₃, CeO₂, etc., are widely used in atom-cluster processing. Lapping abrasives are refreshed and resharpened by crushing during operation to achieve continuous removal of material, while polishing abrasives burnish under extremely large shear stresses based on point defects in the atom-cluster range. Thus polishing is performed with tough and heat-resistant abrasives.

Geometrical surface contours can be precisely shaped by lapping using preformed medium-hard plate and subsequent polishing using a small soft plate, because the surfaces of these plates wear little during processing, because of the processing mechanism. Continuous surface processing with nearnanometre accuracies can therefore be performed.

So far it has been very difficult to generate aspherical and other curved surfaces, because highlyprecise relative motion between the workpiece and plate has been difficult to achieve. Recently however, numerical control of fine steps of 0.1 μm has been achieved and as a result, new fabrication technologies for precision curved surfaces using generating systems with lapping and polishing tools will soon be developed (see Fig. 1.1.5).



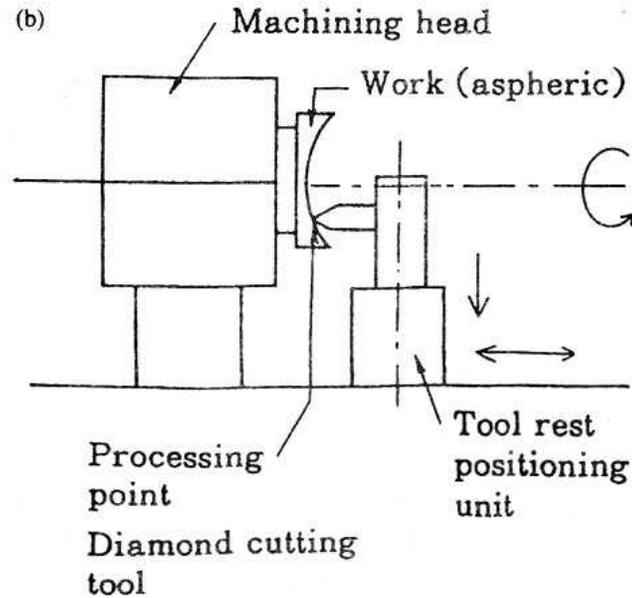
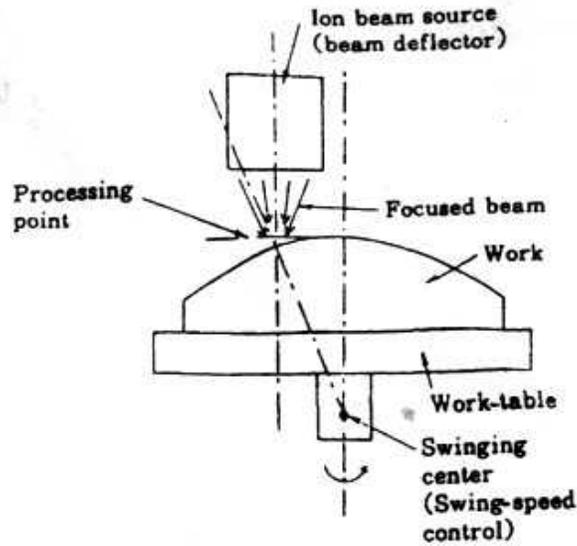


Fig. 1.1.4. Typical processing systems: cutting, (a) Grinding (smoothing), (b) Turning (mirror-cutting).

(b) Atomic-bit processing using elementary high-energy particles or a concentrated electric field

In order to perform atomic-bit removal, for which a high-density processing energy of 10^4 - 10^6Jcm^{-3} is necessary, processing methods using high-energy particles have been developed, in which a beam of elementary particles such as photons, electrons or ions, chemically and electrochemically reactive atoms (reactants) or neutral atoms is applied to the processing point (see Figs 1.1.6 and 1.1.7 and Tables 1.1.3 and 1.1.4).

Processing methods using high-energy particles have pressing resolutions on the atomic or sub-nanometre scale. However, it is very difficult to position the processing point with nanometre accuracies, because unlike the situation with solid tools on machine tools, there is no reference geometrical surface or axis to control the position of the particle beam at such highprecision.



(b)

Fig. 1.1.6. Typical processing systems: ion-beam processing, (a) Fine patterning, (b) Aspherical lens processing.

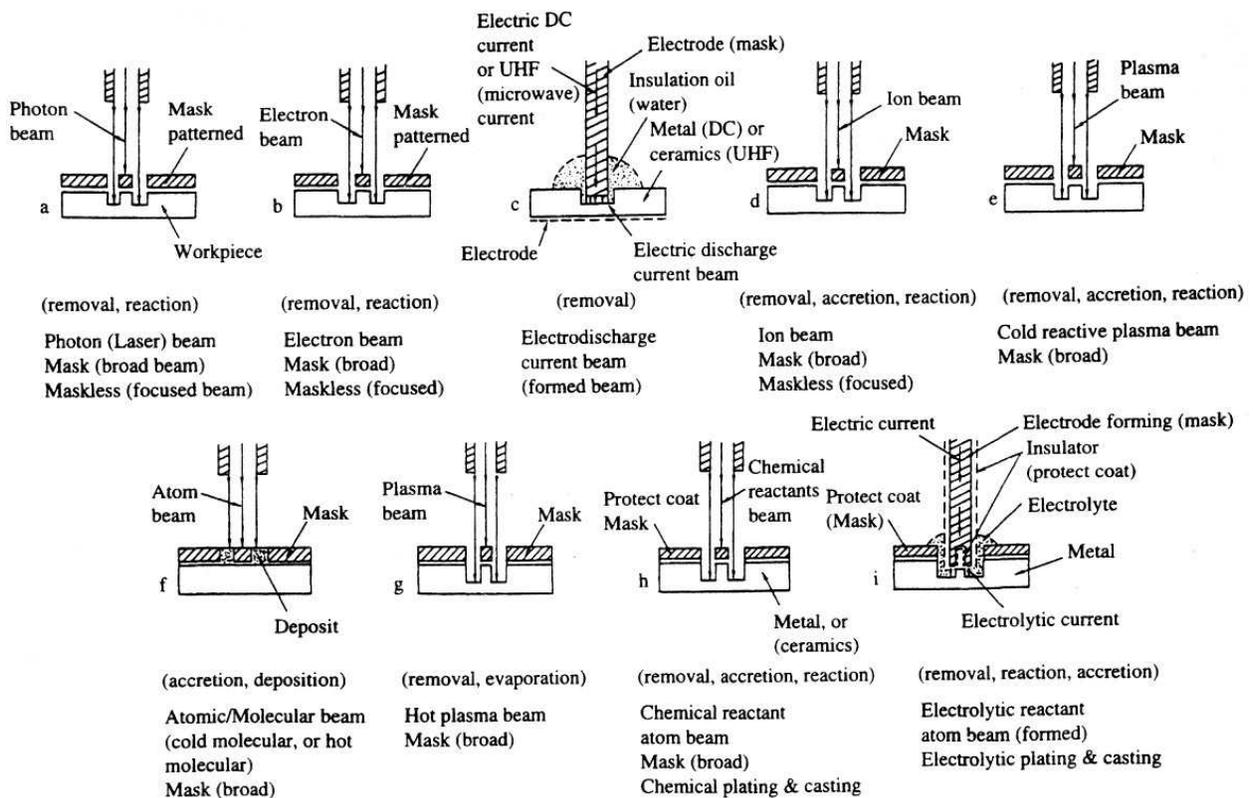


Fig. 1.1.7. Schematic models for energy-beam processing.

Accordingly, high-energy particle beam processing is presently performed only with open-loop control to position the energy-beam axis, without any feedback of in-process positioning information. At present, two-dimensional patterning is effected with a patterned mask using broad beams, or with vector and raster scanning beam systems using focused beams. For three-dimensional energy-beam processing of curved surfaces, the projection time is controlled. In the near future, closed-loop position control of processing points with ultrahigh-precision sensing and actuating systems operating in-process must be developed.

Recently the use of STM (scanning tunnelling microscope) systems has made it possible to remove specified atoms by means of a high-potential electric field. In this system, closed-loop control has been used successfully to position the processing point of specified atoms to sub-nanometre accuracies.

1.2 Mechanism of materials processing based on imperfections or defects in materials

1.2.1 Failure and fracture behaviour of materials under uniform and localized loading

The behaviour of material breakdown under uniform loading is shown graphically in Fig. 1.2.1. Brittle materials break down as a result of tensile fractures, based on microcrack defects existing in the plane of maximum tensile stress, whereas ductile materials break down due to shear failure or slip, based on dislocation defects existing in the plane of maximum shear stress.

The behaviour of material breakdown under localized loading is shown in Fig. 1.2.2 and is quite different from that for uniform loading. For localized loading using indenter tools, the failure and fracture behaviour of ceramics and amorphous glasses varies widely depending on the size of the indentation area. For an indenter with a radius of several millimetres, ring cracks occur at the periphery of the indented contact area, where the maximum tensile stress acts, and the crack fracture begins at microcracks existing in the stressed area. For an indenter with a smaller radius of several micrometres, a very small indentation mark due to plastic deformation or a degenerate zone remains in the area of contact with the workpiece.

If ceramics and glass are scratched with a load of several tens of newtons, using an indenter of relatively large radius, a ploughing fracture occurs, as shown in the figure, but for a sharp indenter with a radius of several micrometres, under a load of a few hundredths of a newton, only a scratch mark a few micrometres in width will remain as a result of plastic deformation. When highly brittle materials such as Si and Ge crystals are scratched, a similar phenomenon can be observed, but the mark is sharper. However, for plastic or ductile materials such as metals and synthetic polymers, plastic deformation will always occur, whether by indentation or scratching, as shown in the figure.

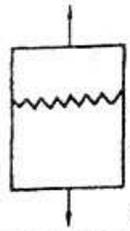
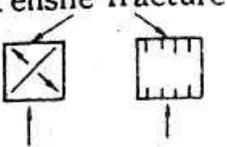
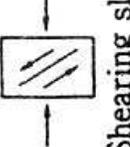
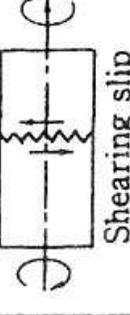
	Brittle ceramics	Ductile metal
	Brittle fracture (micro-crack)	Plastic failure (dislocation)
Tension	 <p>Tensile fracture</p>	 <p>Shearing slip</p>
Compression	 <p>Tensile fracture</p>	 <p>Shearing slip</p>
Twisting	 <p>Tensile fracture</p>	 <p>Shearing slip</p>
Bending	 <p>Tensile fracture</p>	 <p>Shearing slip</p>

Fig. 1.2.1. Material breakdown behaviour due to uniform loading (room temperature).

The reason for this complex behaviour is that materials are not uniform but invariably possess various defects such as point defects, dislocations, microcracks, boundary cracks, layers surrounding the crystal grains, etc., which initiate failure or fracture under load. Accordingly, defects in the work material play a very important role in materials processing such as removal, deformation and consolidation.

Furthermore, the failure or fracture behaviour of materials is affected by environmental conditions such as temperature, atmospheric pressure and humidity and the number of repeated loading cycles (fatigue limit), as will be discussed later.

In subsequent sections, processing mechanisms are discussed according to the type and presence of imperfections or defects in the materials, in the following order: processing units of atomic bits in the no-defect region; processing units of atom clusters based on point defects; processing units of sub-grain size based on dislocations or microcracks; and processing units of multiple grains based on crystal grain boundaries.

1.2.2 Atomic-bit processing of materials in the atomic-lattice no-defect range or sub-nanometre region

In general, the atomic lattice distance or effective radius in crystals is 0.1 to 1 nm. Materials processing in this region concerns the treatment of atomic bits, or atom-by-atom. Examples include atomic-bit removal by evaporation, diffusion and dissolution, based on localized thermal energy supplied by photon, electron or plasma energy particle beams; the sputtering of surface atoms by the transferred kinetic energy of ions; and the removal of reacting atoms by chemical and electrochemical reactions. Recently, electric-field evaporation of atomic bits, called ‘atom craft’, has been proposed by Aono, as discussed later (subsection 1.4.1).

The processing energy necessary to remove stock material by atomic bits is the atomic lattice bonding energy, of the solid surface layer, which is shown in Fig. 1.2.3, obtained from modified Morse free potential energy between two atoms considering the effect of solid surface. Namely to remove an atom of the stock material from the workpiece surface, surplus energy is necessary to overcome the surface barrier potential energy, due to the surface energy based on discontinuous structure of the atomic arrangement at the surface.

Macroscopically, the amount of energy necessary to remove stock material by atomic bits is estimated to be 6 to 25 eV ($1 \text{ eV} = 1.6 \times 10^{-19} \text{ J}$) in thermal evaporation processes.

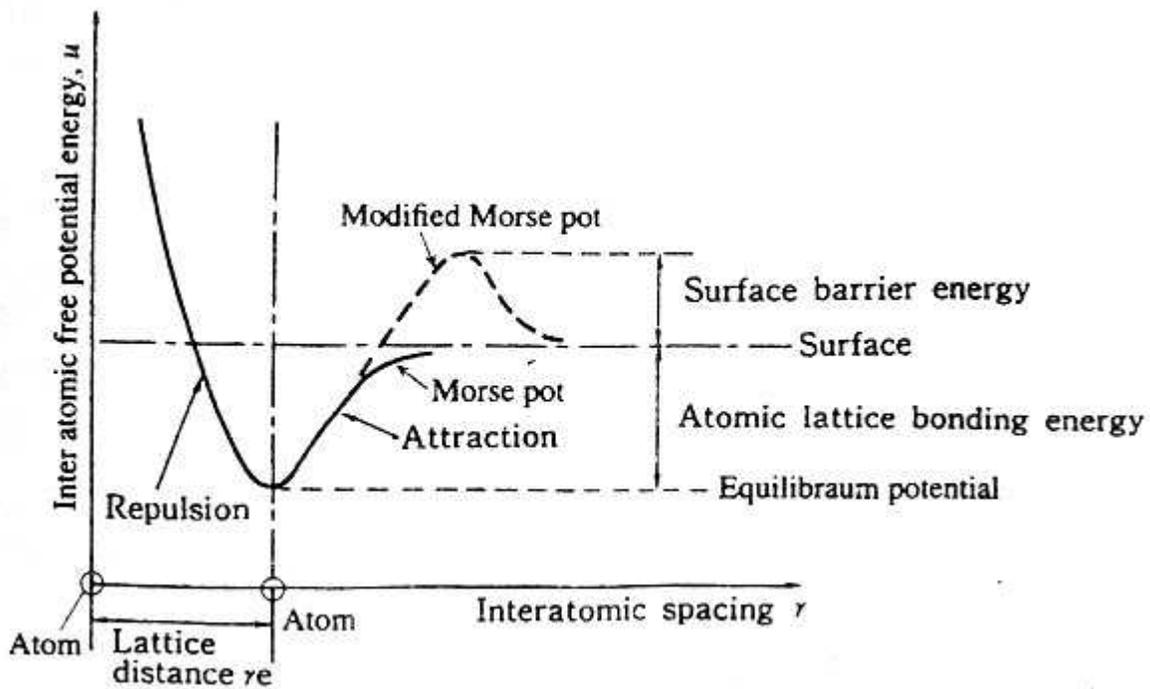


Fig. 1.2.3. Modified Morse free potential energy between two atoms considering the effect of solid surface. Morse free potential energy $U = D \{1 - \exp(-a(r-r_e))\}^2$ where r is the interatomic spacing of two separator atoms, r_e is the interatomic spacing of two separate atoms at minimum free potential energy D is the dissolution potential energy or lattice bonding energy, and a^2 is the constant referring to displacing force near the equilibrium position r_e . Although this is the net energy required for removal, in practice the efficiency of the available supplied energy is considerably lower and varies with the type of energy-beam processing system used. Consolidation processes involving atomic bits include evaporation deposition, ion-sputter deposition, chemical and electrochemical deposition, and direct ion deposition. The minimum energy required for atomic-bit consolidation is estimated to be almost the same as the surface barrier potential energy. The minimum energy required for atomic-bit displacement is estimated to be smaller than that for atomic-bit separation, because the former involves the shearing energy for slippage between crystal atoms.

Figure 1.2.4 and Table 1.1.2 (section 1.1.) present values of the specific volumetric lattice bonding energy as calculated by Plendle. The lattice structures of metals, ceramics and organic polymers are shown schematically in Figs 1.2.5-1.2.8.

1.4 Nano-physical processing of atomic-bit units

1.4.1 Electric-field evaporation of specified atoms

This method was developed from the concept of the STM (scanning tunnelling microscope). A schematic diagram of electric-field evaporation is shown in Fig. 1.4.1. The system consists of a sharp tip electrode made of tungsten which is controlled by a 3D positioning unit based on a piezoelectric stack with sub-nanometre accuracy and resolution and a d.c. power source of several volts. A workpiece of conductive material is placed below the electrode with a gap of 0.1 nm or so. The electric field intensity in the gap is in the region of $5 \times 10^{10} \text{ V m}^{-1}$, creating a traction or pull-out force on an ionized surface atom of $5 \times 10^{10} \text{ V m}^{-1} \times 1e = 8 \times 10^{-8} \text{ N}$, ($1 \text{ eV} = 1.6 \times 10^{-19} \text{ J}$). Against this, the bonding force on an atom in its lattice site position is estimated to be $\sim 5 \times 10^{-8} \text{ N}$ for Si, calculated from the atomic bonding energy as shown in Table 1.1.2. Therefore a specific surface atom can be expected to be removed by such an electric field.

1.4.2 Directional photon beam processing

A photon is a massless elementary energy particle, i.e. the energy quantum in Planck's law. A photon has an energy equal to $h\nu$, where h is the Planck constant, $6.626 \times 10^{-34} \text{ J s}$, and ν is the equivalent frequency (Hz). Since the photon may be at the same time considered as an electromagnetic wave of frequency ν , the equivalent wavelength of a photon (m) is given by

$$\ell = c / \nu, \quad (1.4.1)$$

where c is the speed of light, $2.998 \times 10^8 \text{ m s}^{-1}$.

Hence the wavelength of photons in a light beam or laser beam is of the order of 10 to $0.1 \mu\text{m}$ for photon energies in the range 0.1 to 10 eV. On the other hand, the wavelength of photons in

radiation such as X-rays and SOR (synchrotron orbit radiation) is of the order of 0.1 to 10 nm for photon energies of 0.1 to 10 keV, and 0.01 to 0.001 nm for photon energies of 0.1 to 1 MeV, respectively. A photon's wavelength can be considered as the region in which it exists.

Table 1.4.1 Types of laser source

Laser source	Element	Optically active ion	Wavelength / (μm) (Photon energy, eV)	*
Solid	ruby	Al_2O_3	Cr^{3+}	pulse, 1 J, mean max, 5 W
	YAG	$\text{Y}_3\text{Al}_5\text{O}_{12}$	Nd^{3+}	0.694 3 (1.78) 1.065 pulse, cont. 40-50 W
	glass	glass	Nd^{3+}	(1.15) 1.065 (1.15) pulse, cont.
Semi conductor	GaAs	GaAs	-	~ 0.8 (1.53-) pulse, cont.
Gas	CO_2	$\text{CO}_2 + \text{He} + \text{Ne}$	CO_2 ion	10.63 (1.15) cont. < 5 kW pulse, 40-100 W-TEA CO_2 (10-100 Hz)
	Ar	Ar	Ar^+	0.488 (2.51), 0.545 (2.24) cont. < 5 kW
	He-Ne excimer	He-N_2 (Xe, Kr, Ar) + (F, Cl, Br)	$\text{Xe}^+, \text{Kr}^+, \text{Ar}^+$	0.633 (1.93) 0.5-0.2 (2.44-6.12) cont. < 5 mW pulse, 1 kHz 10 W direct photon sputtering

Several examples of photon beam processing using light, laser and radiation beams are shown in Table 1.4.1 and Fig. 1.4.2. The photon beam projected locally on to the surface of the workpiece transmits its energy to the outer-shell electron(s) of the atoms. This energy is transformed into thermal vibration energy or chemical activation energy of the atom that is hit, as shown in Fig. 1.4.3(a) and (b).

However, the thermal energy transferred from a laser photon, as shown in Table 1.4.1, is smaller than the atomic bonding energy of materials as shown in Table 1.1.2. Accordingly, atomic-bit processing of an atom by a single photon cannot be performed with ordinary lasers; instead, photon beam thermal processing is performed, by locally accumulated heat at the processing point, conducted from the photon beam's projected area. The saturation temperature T_s (K) due to concentrated heat at the centre of the projected beam is derived from heat conduction theory as (see Fig. 1.4.4)

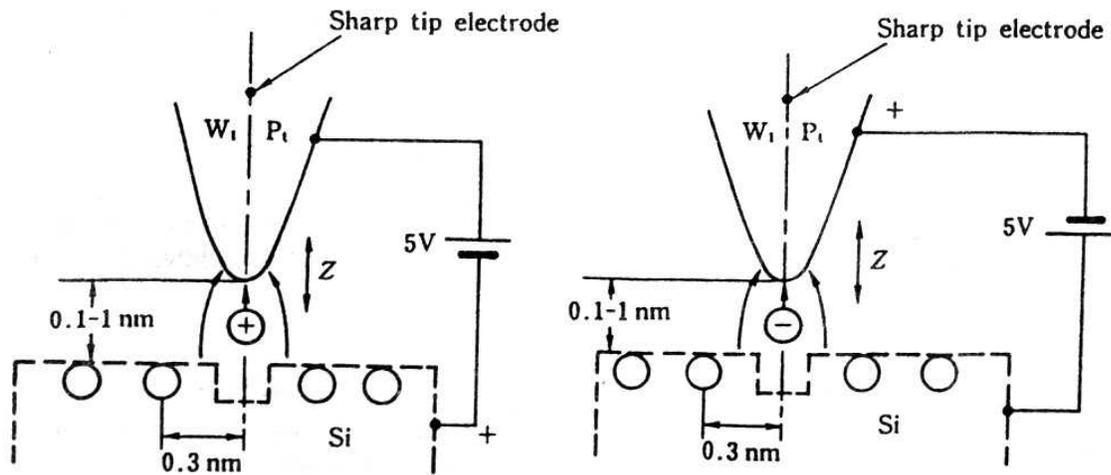
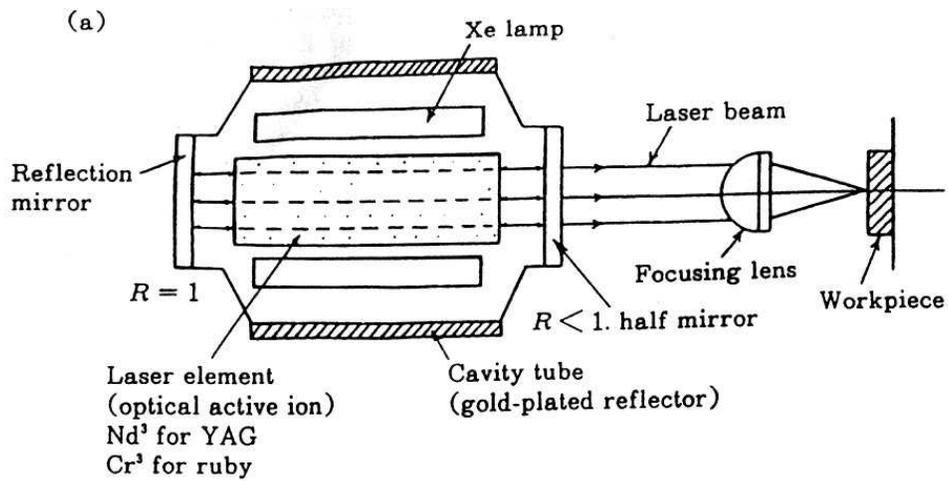


Fig. 1.4.1. Models for electric field evaporation: sharp tip electrode controlled by piezoelectric stack positioning device with sub-nanometre resolution. P_t denotes traction force due to concentrated electric field on the sharp tip.



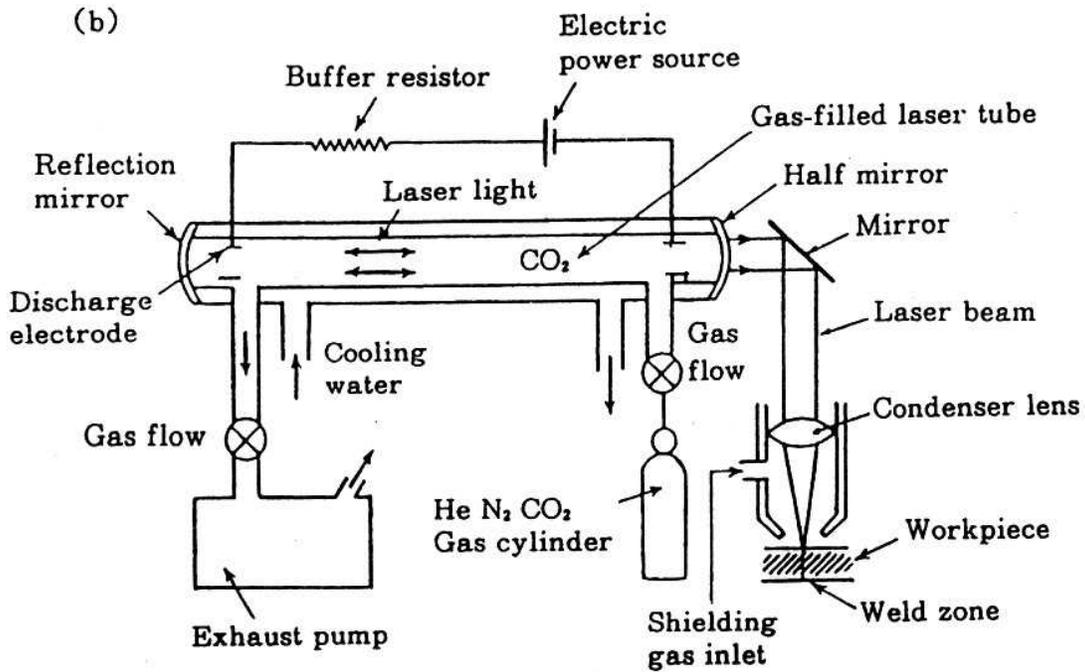


Fig. 1.4.2. Laser sources: (a) solid; (b) CO₂ gas (coaxial).

$$T_s = Q / (\pi \lambda a) = qa / \lambda \quad (1.4.2)$$

where Q is the input power (J s^{-1}), λ is the thermal conductivity ($\text{J m}^{-1}\text{K}^{-1}\text{s}^{-1}$), a is the radius of the projected energy beam (m), and q is the input power density, $Q / (\pi a^2)$ (W m^{-2}). For example, the thermal power density q of a photon beam of radius $0.2 \mu\text{m}$ necessary to achieve the evaporation temperature of 1500°C for stainless steel is $\sim 2.5 \times 10^{11} \text{W m}^{-2}$ with $\lambda = 0.14 \times 10^2 \text{J m}^{-1}\text{K}^{-1}\text{s}^{-1}$. Temperature analysis of the evaporation process due to a projected photon or electron beam will be discussed in a later section.

Unlike those of lasers, photon beams of X-ray and SOR radiation can achieve direct evaporation of one atom by one photon, owing to the high energy of the photons, as shown in Table 1.4.1.

Besides thermal evaporation due to photon energy, there are several reactive photochemical processes such as those used for photosensitive dry plates and photosensitive plastics, in which reactions are caused by the projected photons, as shown in Fig. 1.4.3(c). The projected photons penetrate through the surface of a transparent photosensitive material and react directly with photoreactive atoms in their path. This creates a centre for a latent image, for example an activated silver atom in silver halide photography or polymerized active atoms in photoresist plastics.

Of course, the positioning accuracy of the photo-reacted area is limited to the range of the equivalent wavelength along the photon beam path. Hence, to increase the resolution for positioning, the argon laser and excimer laser with short wavelengths, e.g. $0.2 \mu\text{m}$, are used.

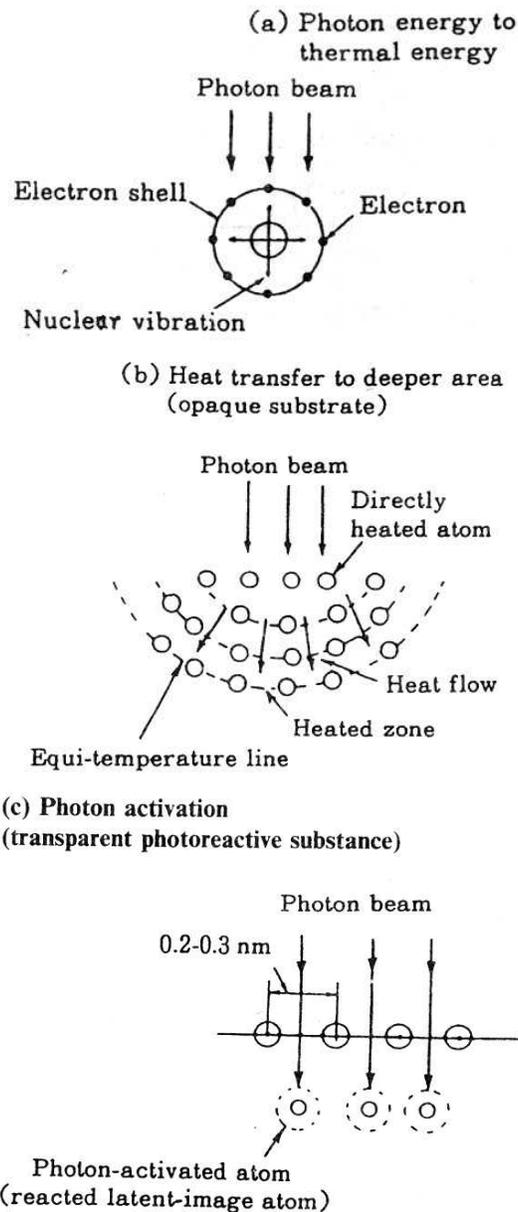


Fig. 1.4.3. Photon energy transfer mechanisms.

Applications for this kind of photolithography are found in the manufacture of IC wafers with very fine patterns in the sub-micrometre range, and videodisc master plates with very fine channels and spots of $\sim 1 \mu\text{m}$. Details are given in Chapter 4. Recently, X-ray and SOR beams have also been used for photoreactive processes. These beams consist of comparatively high-energy photons and consequently have very small equivalent wavelengths, of the order of nanometres. As a result, these beams are used successfully for nano-patterning of LSIs. Using such radiation beams, the pattern resolution can be improved to the order of 10 nm, but to achieve high-resolution patterning in the sub-nanometre range, the technology for post-chemical etching to develop latent images must be improved.

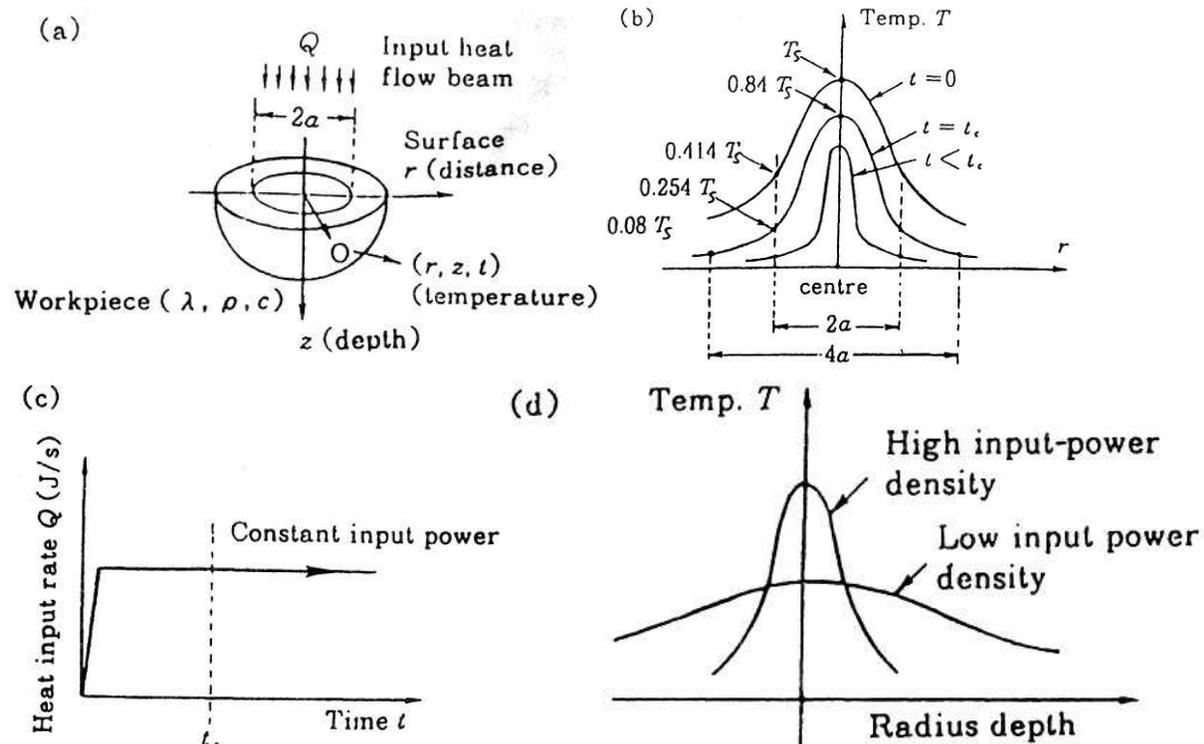
1.4.3 Directional electron beam processing

A focused electron beam processing apparatus is shown schematically in Fig. 1.4.5; the mechanism of energy transfer from electron to workpiece atom is depicted in Figs. 1.4.6 and 1.4.7. The basic process performed by an energized electron is thermal evaporation of atomic bits. The accelerated electrons projected on to the workpiece generally transfer their energy to the outer shell of the atom and increase the thermal vibration energy of the nucleus. Hence energized electrons can effectively supply the processing energy necessary for thermal evaporation in a very finely localized portion of the workpiece. However, it is important to recognize that the projected electron is absorbed mainly in the region of the penetration depth as shown in Fig. 1.4.7 and not at the workpiece surface. Electron beam processing was initially developed to form fine patterns on semiconductor wafers and fine holes or textured surfaces on diamond and other gemstones. This is because the electron has a very small diameter of 2.8×10^{-6} nm from the classical viewpoint and a small mass of 9×10^{-31} kg estimated from classical theory, so it can be easily focused into a very narrow beam a few micrometres in diameter and can be highly energized to several hundred kilovolts ($1 \text{ keV} = 1.602 \times 10^{-16} \text{ J}$).

The incident high-energy electron is able to penetrate the workpiece surface through the network of the atomic lattice structure because its effective diameter is much smaller than the atomic lattice distance of 0.2-0.4 nm. The incident electrons penetrate through the surface layer to a depth R_p (see Fig. 1.4.7) where most are absorbed. The experimental and theoretical penetration depth is

$$R_p = 2.2 \times 10^{-12} V^2 / \rho \text{ (cm)}, \quad (1.4.3)$$

where V is the acceleration voltage (V) and ρ the mass



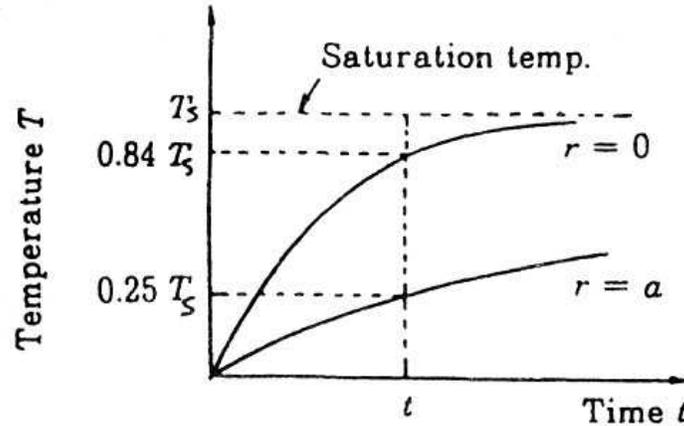


Fig. 1.4.4. Input of thermal energy and distribution and rate of change of temperature, (a) Model for heating: Q , energy input rate; λ , thermal conductivity; p , mass density; c , specific heat capacity; t , time; a , radius of energy beam, (b) Temperature distribution and gradient: T_s , saturation temperature as defined by eqn. (1.4.2); temperature gradient = $(0.84 - 0.254)T_s / a \approx 0.6q / \lambda$. (c) Rate of temperature change at constant input power: t_c , characteristic response time = $\pi a^2 / k$, where k is the thermal diffusivity = $\lambda / \rho c$. (d) Change in temperature distribution with input power density.

density of the workpiece (g cm^{-3}). For example, the penetration depth in steel for an electron of energy 50 keV is $\sim 7 \mu\text{m}$ and in aluminium $\sim 10 \mu\text{m}$.

As shown in Fig. 1.4.7, the number of electrons absorbed at the workpiece surface is very small; the depth of maximum absorption rate X_e and the centre of the absorption depth x_d are indicated. Consequently, electron beam processing by thermal evaporation is not suitable for fine patterning and other fine machining requiring nanometre accuracy on the workpiece surface.

In contrast to the thermal evaporation process, electron beam processing based on reactive activation by a focused electron beam, as shown in Fig. 1.4.8, is widely used in electron-beam lithography with nanometre accuracy and resolution. This process is based on the activation of electron-sensitive materials such as polymers. Atoms of such materials are activated by electrons passing near the nucleus. In other words, the incident electrons activate the electron-sensitive atoms and cause polymerization or depolymerization along the electron path. The process has been used to achieve very high dimensional resolutions in the sub-nanometre range, because no thermal diffusion and little secondary scattering occur. Therefore a finely focused electron beam can produce patterning with sub-nanometre accuracy using atomic-bit processing based on reactive activation.

1.4.4 Directional ion beam processing

A fundamental method of processing materials that uses energized ions is ion sputter machining. The basic equipment for ion sputter machining is shown schematically in Fig. 1.4.9.

The mechanism of sputter machining is shown v diagrammatically in Fig. 1.4.10. Electrically accelerated inert-gas ions such as Ar ions with an average energy of 10 keV (corresponding to a speed of $\sim 200 \text{ km s}^{-1}$) are unidirectionally orientated and projected on to the workpiece surface in a high vacuum ($1.3 \times 10^{-4} \text{ Pa}$).

Unlike electron beam processing, most of the projected ions interfere with the surface atoms of the workpiece, because the mean diameter of an Ar ion, 0.1 nm, and the mean lattice distance between surface atoms, $\sim 0.3 \text{ nm}$, are comparable. Consequently the projected energized ions frequently collide with the nuclei of atoms of the workpiece and knock out or sputter the surface atoms. Hence processing is performed mainly by atom-by-atom removal; this is called ion sputter etching or ion sputter machining.

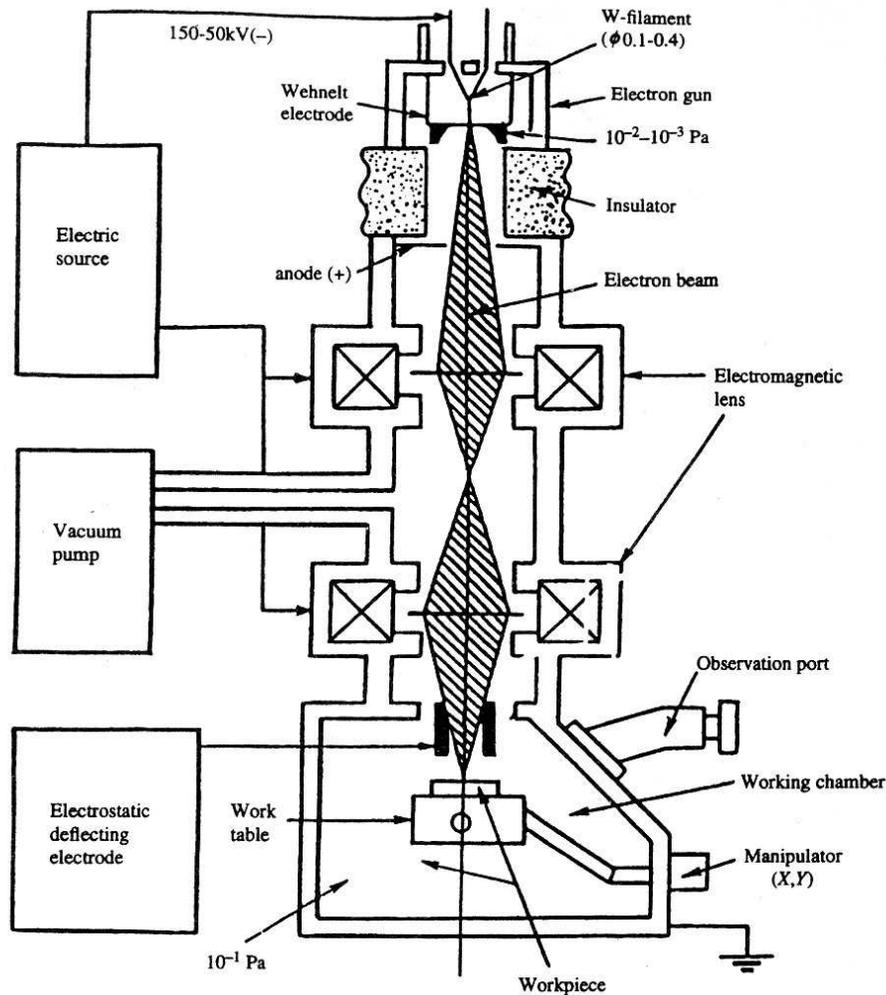


Fig. 1.4.5. Electron beam processing apparatus.

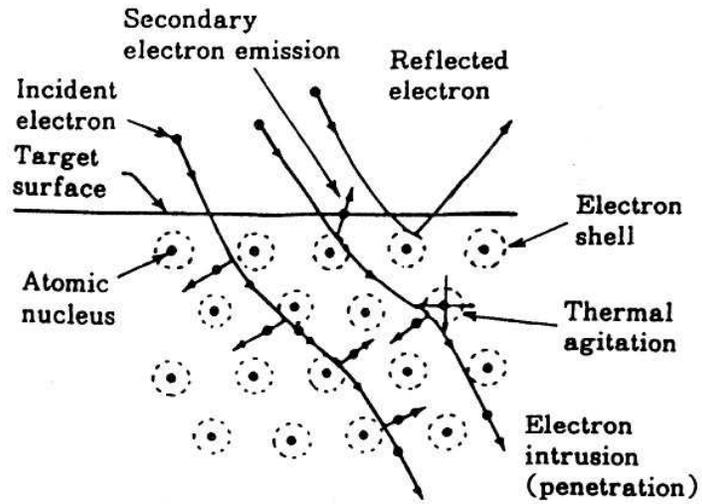


Fig. 1.4.6. Mechanism of energy conversion of electron beam in a material.

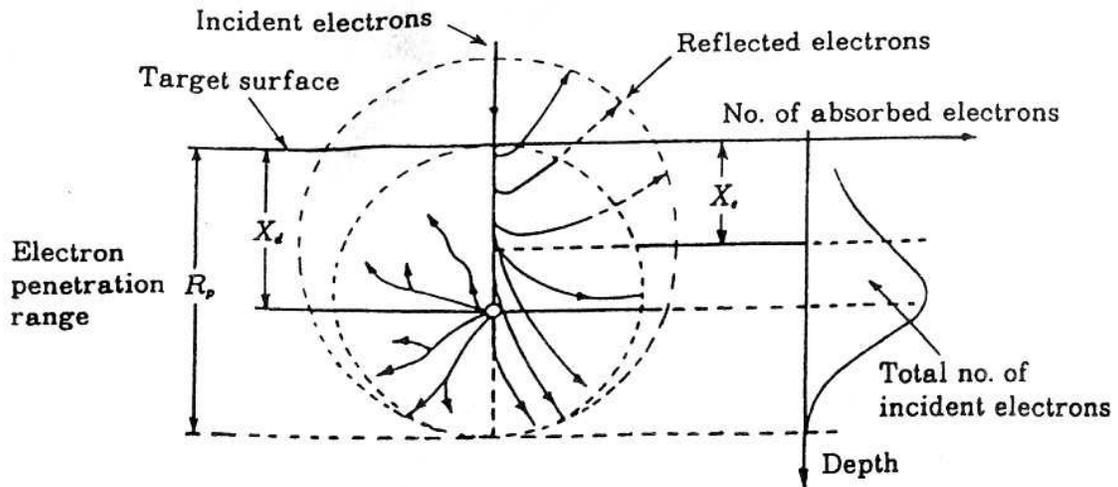


Fig. 1.4.7. Electron penetration range

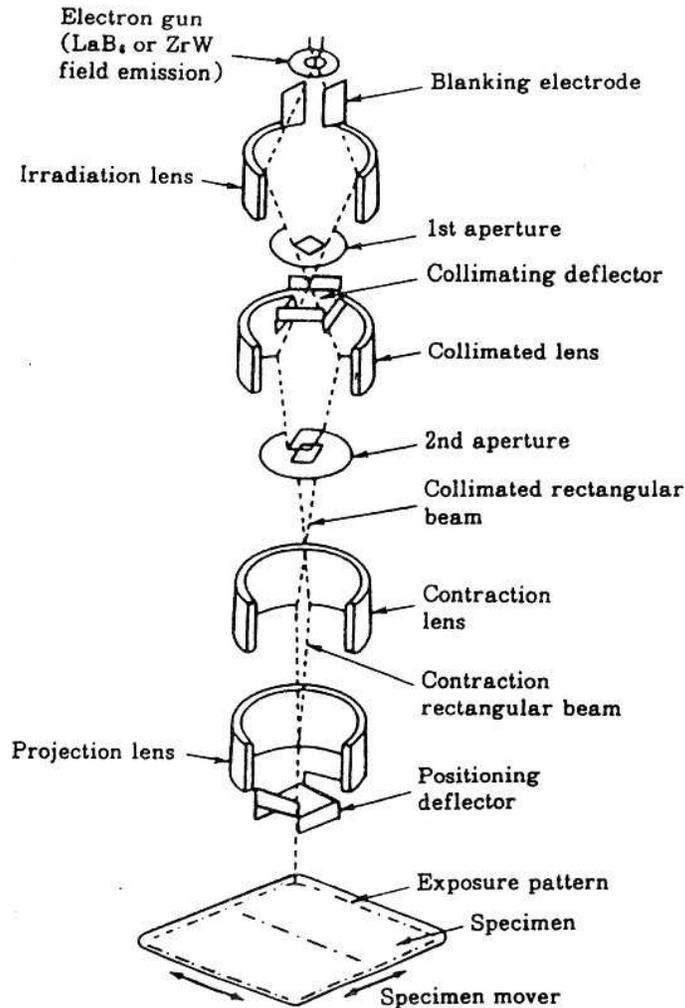


Fig. 1.4.8. Electron beam lithography: variable-area electron beam exposure device JBX-6A (vector scanning type). Furthermore, when the sputtered atom has a kinetic energy of several tens of electron volts, higher than that of an ordinary evaporated atom, ion sputterdeposition may be performed; the sputtered atom with an energy beyond the surface barrier potential collides with a target located opposite and adheres more firmly than in ordinary vapour deposition.

The penetration depth of an impinging 10 keV Ar ion is estimated to be several nanometres or about ten atomic layers, as shown in Fig. 1.4.10. On the other hand, ions with higher energies, e.g. 100 keV, can penetrate further through the atomic lattice and become interstitial or substitutional atoms in the surface layer. This kind of deep penetration process is widely used for ion implantation, in which impurities of atomic size are injected in semiconductor wafer processing. A recent development is ion mixing, which uses high-energy ions of various elements; it is widely used for the surface treatment or surface modification of workpieces. Another development is, reactive processing using energized chemically reactive ions. Here, reactive ions derived from CCL_4 are used to remove Al, or CF_4 to remove Si; a much higher reaction rate is obtained with these energized ions. Details are given in Section 1.5 and Chapter 6. Ion beam sources of the high-frequency plasma type and ion-shower type, using a d.c.

discharge or a 2.45 GHz microwave resonator, shown respectively in Figs 1.4.11 and 1.4.12, have also been widely used.

1.4.5 Plasma surface processing

Plasma is defined as an electrically conductive state of gases in which approximately equal numbers of electrons and ions are concurrently present. Generally, at atmospheric pressure, plasma appears in arc discharge gas at temperatures of 10 000-20 000 K, and is caused by the recombination of dissociated electrons and ions. However, at lower pressures, plasma due to d.c. discharge or microwave excitation appears at comparatively low temperatures; the temperature is determined by the kinetic energy of electrons and ions in that state. Nano-processing of materials can be performed using plasma at low temperatures and reduced pressures; the processing mechanism is similar to that using electron beams and ion beams.

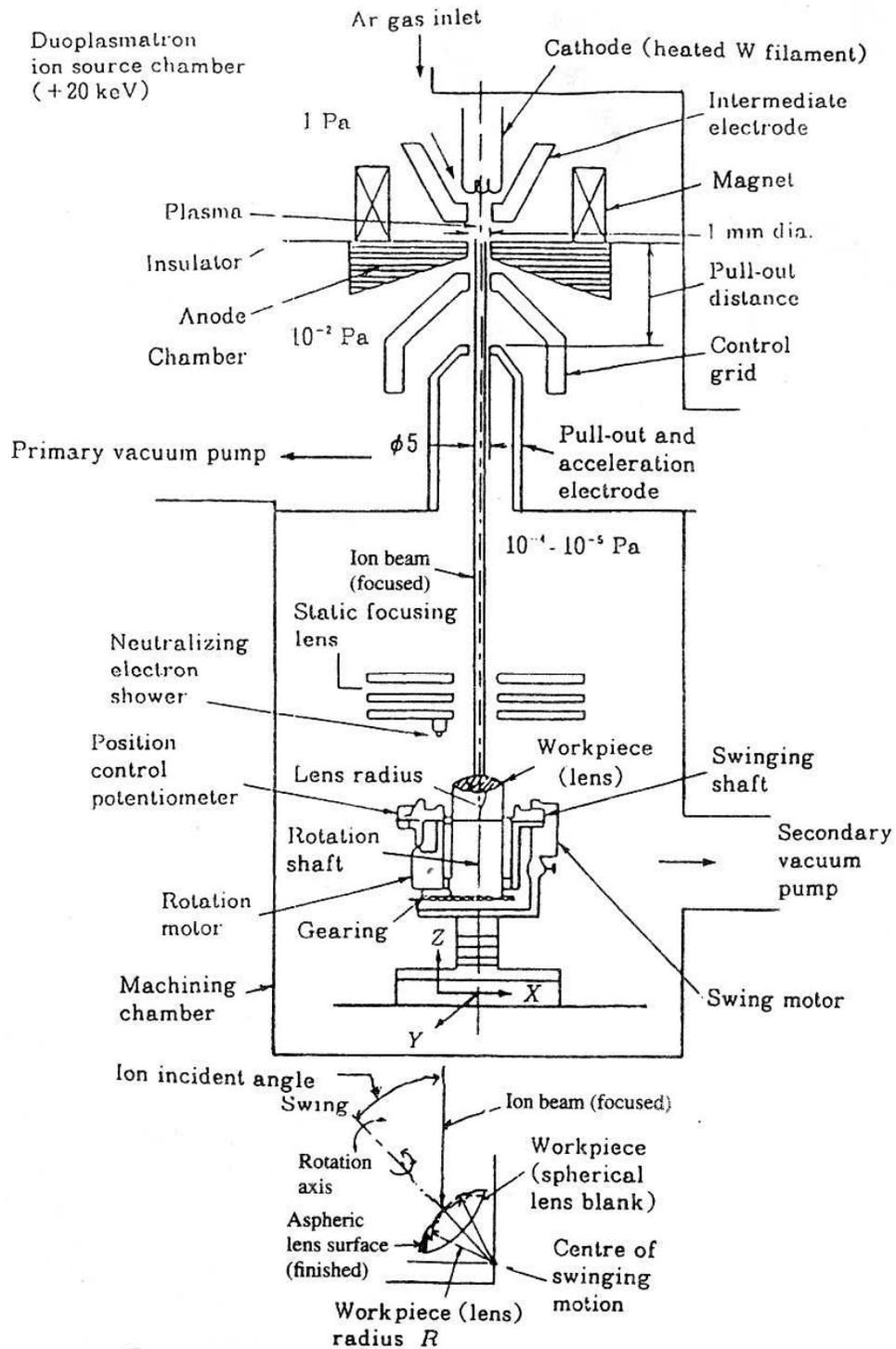


Fig. 1.4.9. Ion beam source apparatus: focused ion beam type for fabricating aspheric lens.

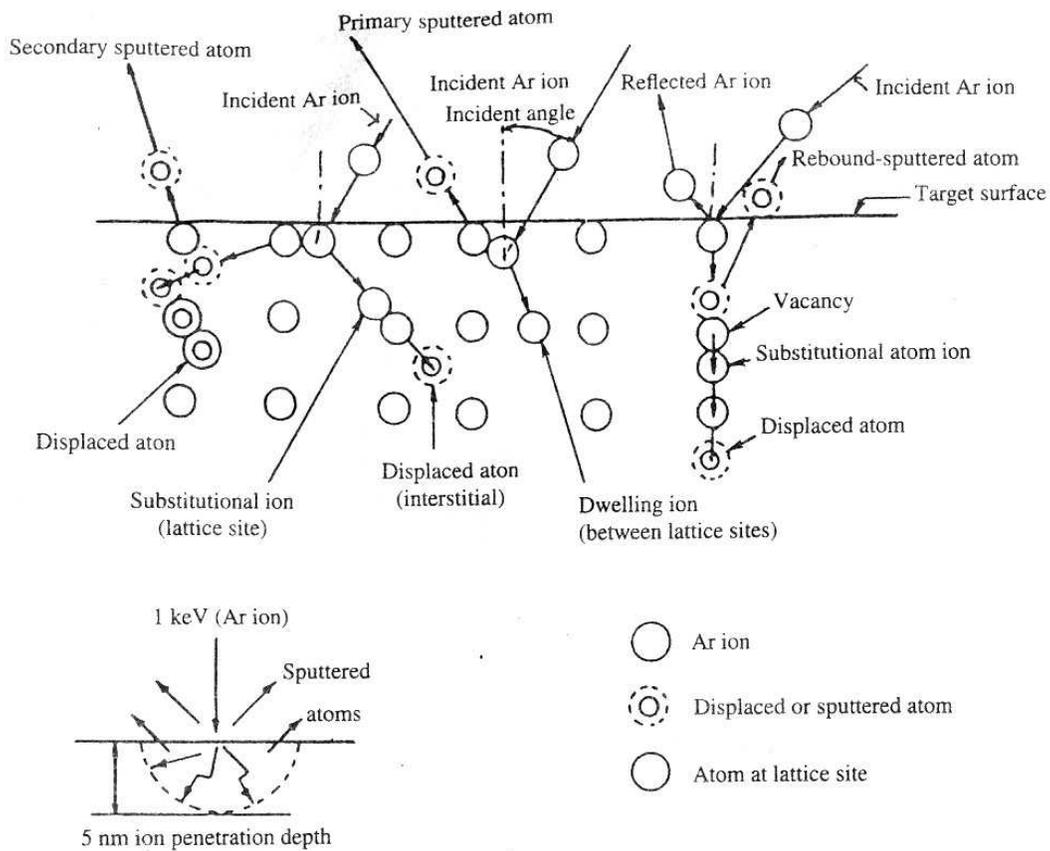


Fig. 1.4.10. Diagrammatic model of ion sputter machining.

1.5 Nano-chemical and electrochemical atomic-bit processing

Owing to the nature of chemical reactions, chemical processing is inherently atomic-bit processing of materials. A chemical reaction is a change in the atomic combination of reacting molecules, in which the atomic bonding of a reacting molecule is broken and a new molecule is generated. For example, molecules with H-H and O-O atomic bonds can be decomposed to form a new molecule with atomic bonding H-O-H. In the process, a chemically reactive gas or liquid is applied to a specified position on the solid surface of the workpiece. The reactive molecules then react with the surface molecules and the reacted soluble or

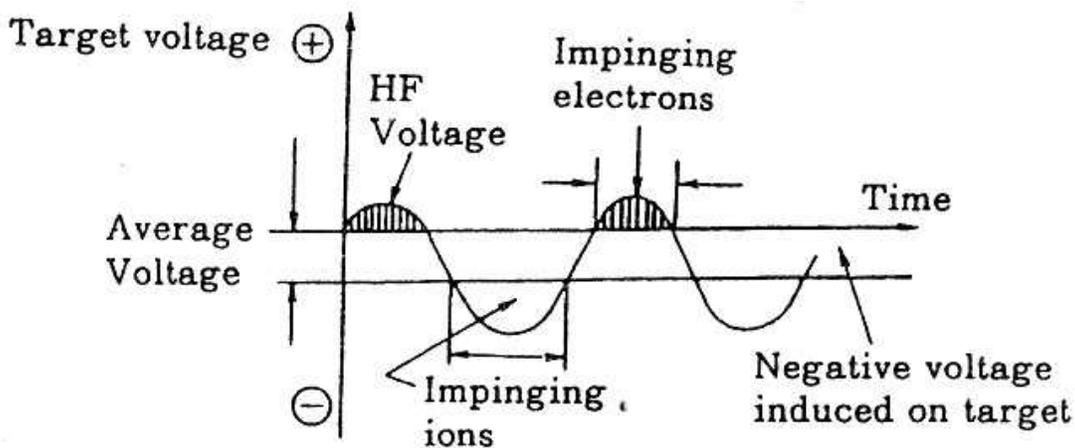
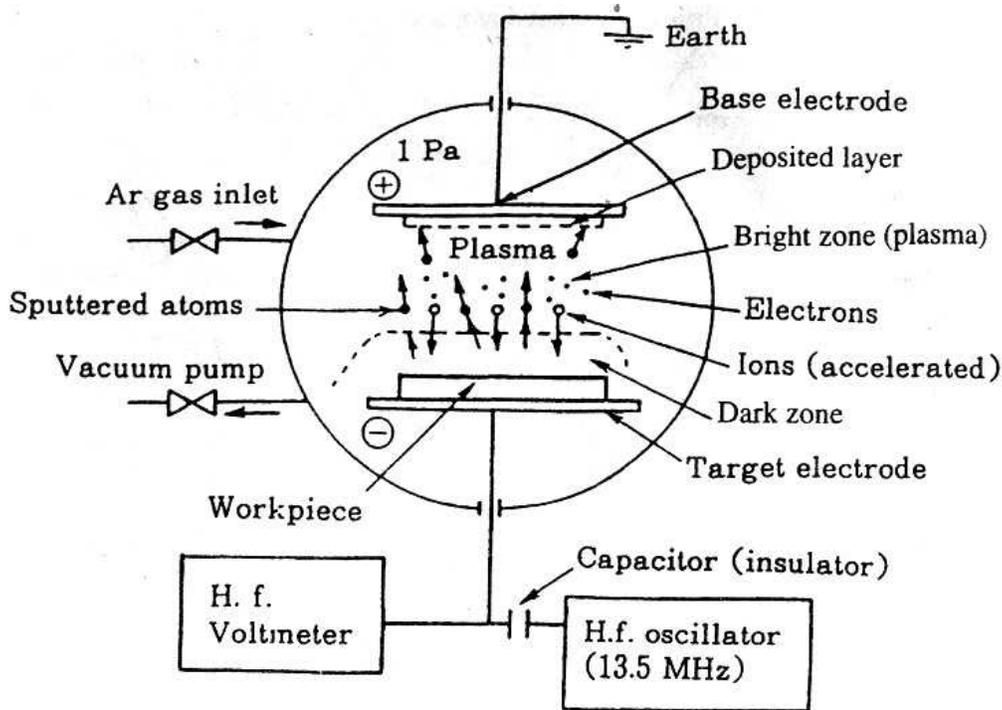


Fig. 1.4.11. Ion beam source: high-frequency plasma type.

vapour molecules are removed or diffused into the surrounding reacting gas or liquid. Such chemical processing of atomic bits occurs uniformly and at random on the workpiece surface to create a flat, smooth surface.

If the reacted molecules are insoluble or not in vapour form, chemically reactive deposition occurs on the workpiece surface, but if the reacted or reagent molecules diffuse into the surface layers of the workpiece and react with the atoms or molecules there, then chemically reactive surface treatment is performed. The dimensional accuracy obtainable (due to scattering errors) in chemically reactive processing is in the nanometre range, with stable processing conditions (temperature and state of turbulence in reagent and liquid flow). To obtain ultra-precision products with nanometre accuracy (due to deviational errors) by means of chemical processing

or etching, in-process measurement and feedback control of the position of the processing point and processed volume (area and depth) are necessary. However, this is difficult to realize in practice.

In general, control of the processing-point position or area in patterning is achieved with a patterned mask made by the photoresist method. However, control of the processed volume or depth can be done only by adjusting the processing time and flow rate of etchants. Although this kind of chemical etching process uses open-loop control without in-process measurement and control, the dimensional accuracy or precision is fairly high. This is because for chemical reactions the processing rate is very slow and the processing resolution very fine. To obtain high processing accuracy, in-process measurement and feedback control with nanometre accuracy should be developed and applied to chemically reactive processing. Although several in process measurement and control systems which use secondary radiation emissions at the point of chemical reaction are being developed, these are not yet satisfactory.

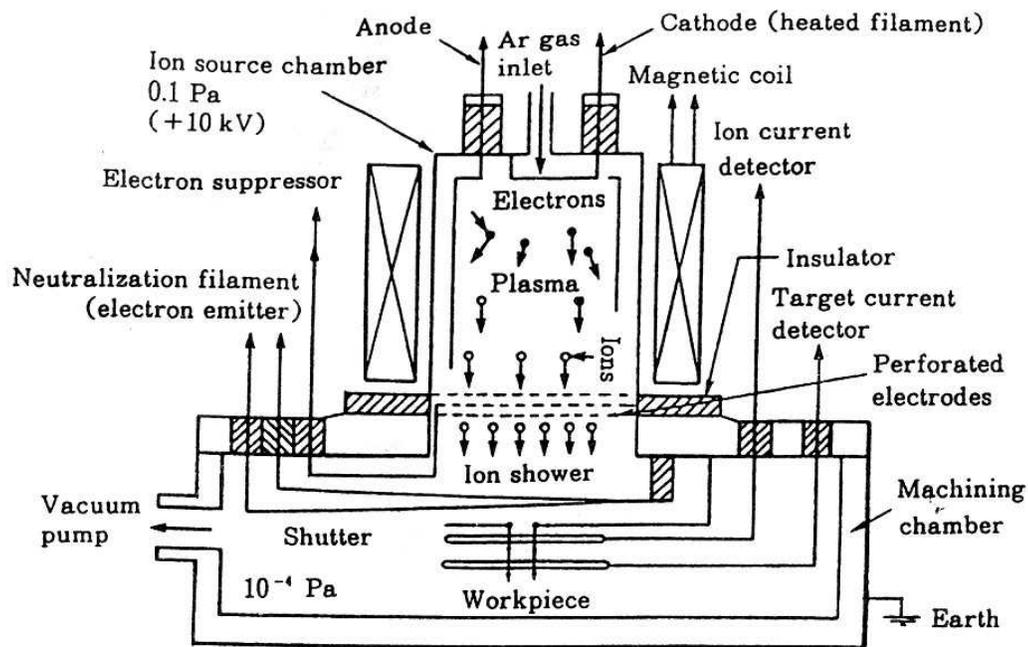


Fig. 1.4.12. Ion beam source: ion shower type (d.c. discharge and 2.45 GHz microwave resonator).

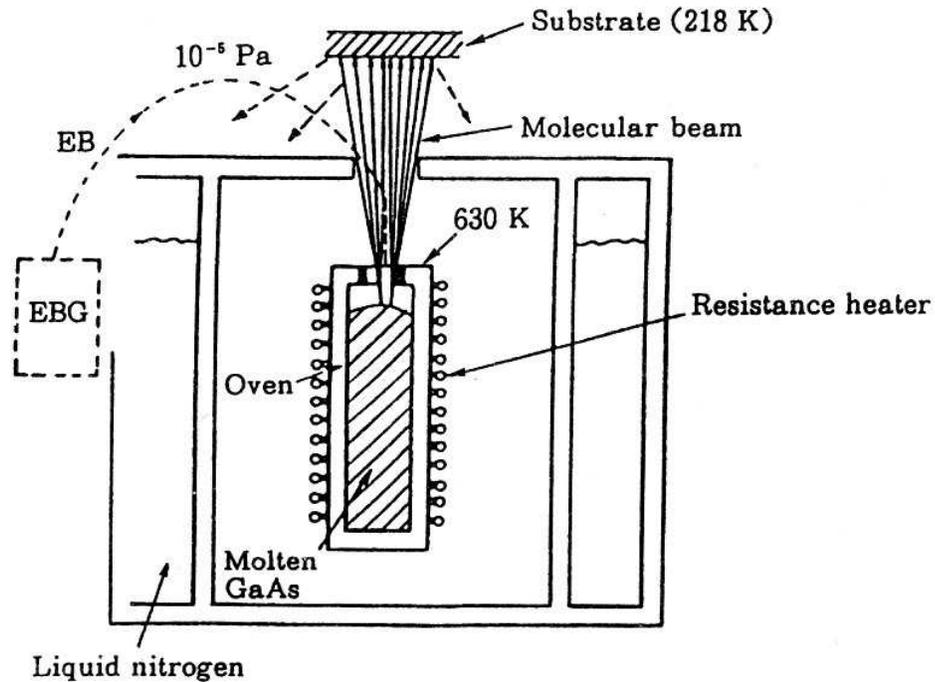


Fig. 1.4.13. Molecular beam epitaxy (MBE).

In this section, chemical reactions are treated only at the macroscopic level, dealing with chemical equilibrium and reaction rates, based on the chemical potential energy due to the quantum state of electrons in the Bohr model. Microscopic treatment of chemical reactions using quantum theory, based on the wave function or atomic orbit, is not discussed here, because atomic-bit processing of materials does not directly concern quantum-mechanical atomic structure. This section also covers electrochemical processing, because the basic reaction is the same as in other ordinary chemical processes.

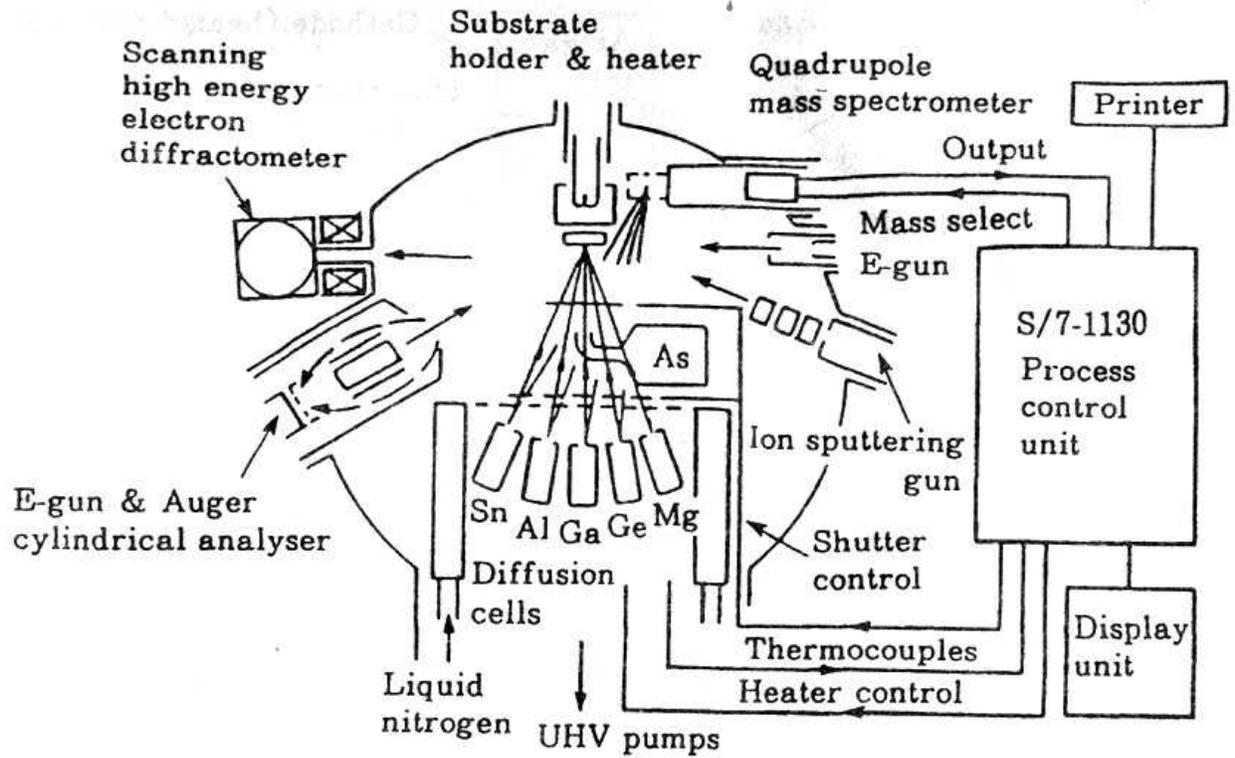


Fig. 1.4.14. Crystal growth by molecular beam controlled by digital computer.

However, the activation energy for the reactions is given by the electric field potential and differs from the ordinary activation energy in chemical reactions based on thermal potential energy.

Module-II

Nano-Measuring Systems of Sub-Nanometre Accuracy and Resolution: In process or in situ measurement of position of processing point, Post process and on machine measurement of dimensional features and surface, Mechanical measuring systems, Optical measuring systems, Electron beam measuring systems, Pattern recognition and inspection systems.

Nano-measuring systems of sub-nanometre accuracy and resolution

2.1 In-process or in situ measurement of position of processing point

In order to process a workpiece to nanometre-order accuracy, measuring and control techniques of nanometre accuracy or better are essential. Nanotechnology must therefore be considered as an integrated technology of processing, measurement, and control. The measurement techniques used in connection with precision processing measure many characteristics of machined workpieces and machine tool components. The quantities measured are length, displacement, vibration, runout, figure, surface roughness, etc. The phenomena detected are sound, temperature, mechanical force, light, electromagnetic field, etc. Because of its widespread applications and present status, optical detection is the main theme here.

This section discusses some advantages, the present potential, and the future of nanometre-order measurement techniques for in-process or in-situ mode. In process measurement will be desirable at the processing point if possible.

2.1.1 In-process measurement

Measurements for production processes can be classified as follows:

- (1) pre-process measurement
- (2) on-machine measurement (or process intermittent measurement, stop-and-measure, in-situ measurement, etc.)
- (3) in-process measurement (or real-time measurement, on-line measurement, during-process measurement, etc.)
- (4) post-process measurement.

Today's measurement or inspection processes on production lines usually occupy locations between the machining processes or the assembly ones. Rejected parts or products have to be processed again for correction, or sometimes discarded. These inspection processes may be rigorous or brief, total or by sampling, and sometimes omitted. However, large losses may be suffered if rejection of a product happens in the later part of the production line. So an appropriate arrangement of processes should be designed and operated.

In spite of the desire to omit the inspection processes from the viewpoint of simple productivity, they are becoming essential as the requirement for machining accuracy increases. Process control with information from on-machine or in-process measurements will be essential for nanometre-

order production in the near future, to achieve better accuracy and to increase productivity. On-machine measurement means spatial unification of machine tool and measuring instrument, and in-process measurement means their temporal unification or simultaneous operation. These techniques will make the processing line simple and rational.

2.1.2 Examples of in-process measurement and X machining control

In-process measurement means measurement during processing and allows real-time control of the process from the measurement signal. For single-point diamond turning, the application of in-process measurement should be easier than for grinding or polishing, because of the simplicity of its form generation mechanism. Seven laser interferometer position sensors, five differential capacitance gauges and one spindle encoder are used to control the famous LODTM (large optics diamond turning machine)⁽¹⁾. These sensors measure the positions of carriage, workpiece and tool during processing, as well as the angle of the spindle. As the LODTM requires accurate positioning during machining for attaining a machining accuracy of 25 nm, the positions of the machine tool elements are sensed with reference to the metrology frame with a resolution of up to 0.6 nm. Although errors exist in all machine tool elements, the turning accuracy may be determined eventually by the relative, not the absolute, positions of the workpiece and the cutting tool point. ‘Workpiece referred form accuracy control’ (WORFAC)² is a simple method to control the relative positioning of workpiece and tool instead of trying to control all the mechanical elements, to increase their individual accuracies, and to improve environmental conditions such as temperature, humidity, and vibration. Figure 2.1.1 shows schematically an experiment to demonstrate the feasibility of WORFAC. A non-contact surface sensor such as a HIPOSS optical stylus (see section 2.5) or a capacitance gauge detects the machined surface of a flat workpiece and feeds the information to a microtool servodrive which controls the tool position by means of a piezoelectric element. The effects of the control in the elimination of disturbances are shown in Fig. 2.1.2. When the control is off, the machined surface has a profile with the same amplitude as a disturbance wave imposed artificially. This profile is suppressed completely when the servo loop is closed.

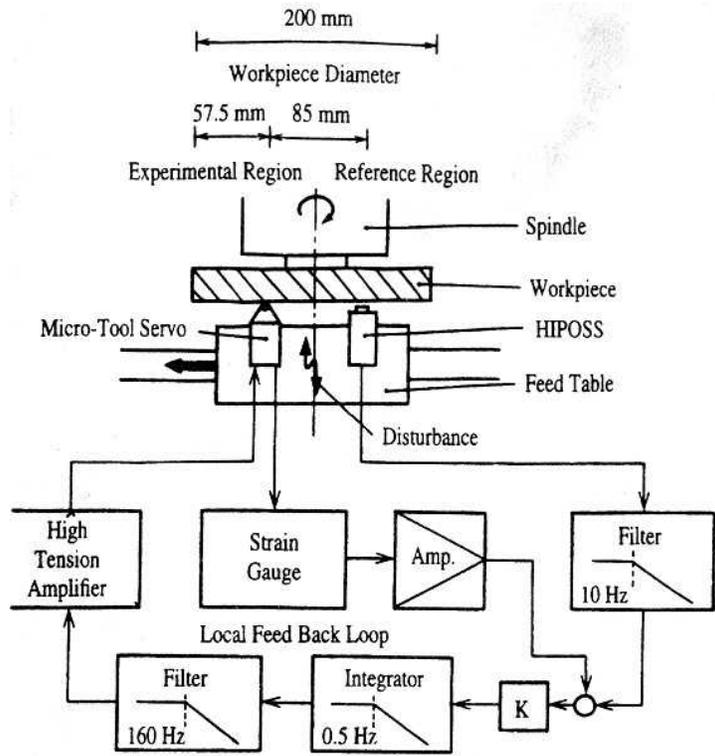


Fig. 2.1.1. Block diagram of WORFAC experiment.

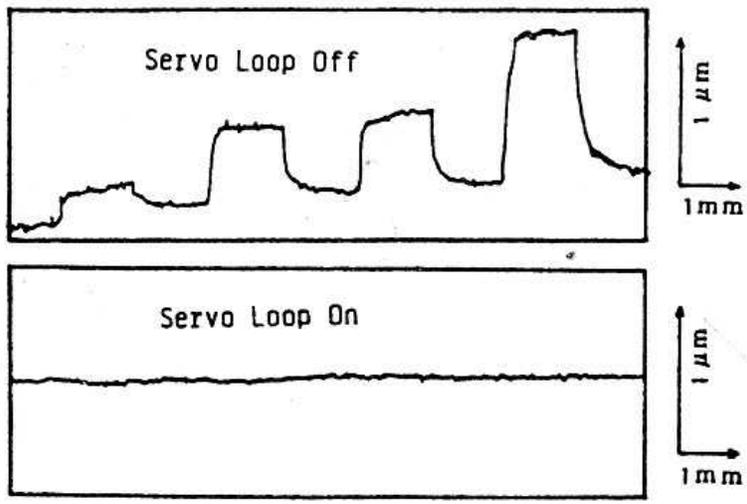


Fig. 2.1.2. Results of WORFAC experiment: elimination of disturbances.

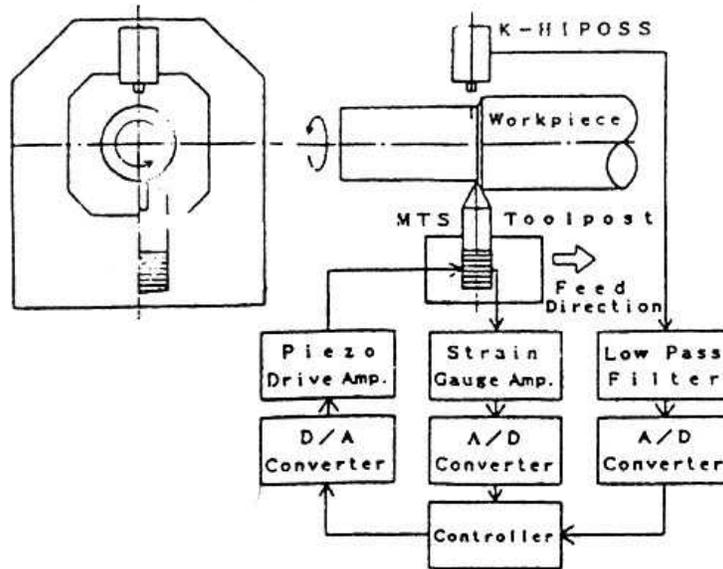


Fig. 2.1.3. WORFAC experimental set-up for cylindrical turning.

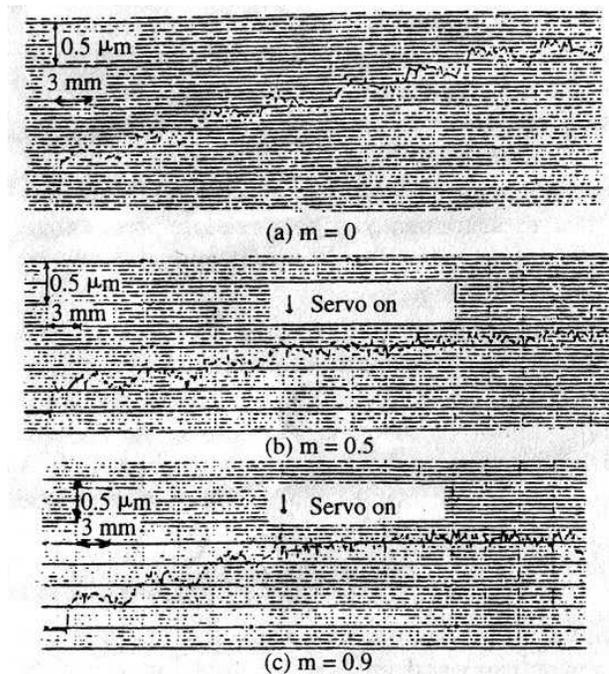


Fig. 2.1.4. Correction of waviness and lack of straightness by WORFAC experiment.

Besides disturbances, some inherent errors of a machine tool have been shown to be suppressed in another WORFAC experiment, for cylindrical turning³. The in-process measurement is carried out at the opposite side from the cutting point, as shown in Fig. 2.1.3. The errors of the slide motion, such as waviness and lack of straightness, are found to be reduced and converged by

computer simulation as well as a turning experiment using a particular feedback level which is the detected error multiplied by a certain coefficient (< 1). The corrected contour of cylindricity is shown to the right of the arrow in Fig. 2.1.4, where the slide motion errors are corrected by the WORFAC control, while some waviness and inclination can be recognised on the left, the no-control side of the figure.

2.2 Post-process and on-machine measurement of dimensional features and surface integrity of processed work

Post-process measurement is the quality testing of a machined workpiece, usually dimensional measurement such as length, outer and/or inner diameter, hole distance, surface roughness, and surface contour. Such measurements of nanometre accuracy have often been performed using high-accuracy measuring instruments on a vibration-isolated table in an air-conditioned room. However, on a mass production line, this is not a popular processing system and is time-consuming. Some stand-alone optical measuring systems for practical nanotechnology with elimination of various disturbances are described in Section 2.5.

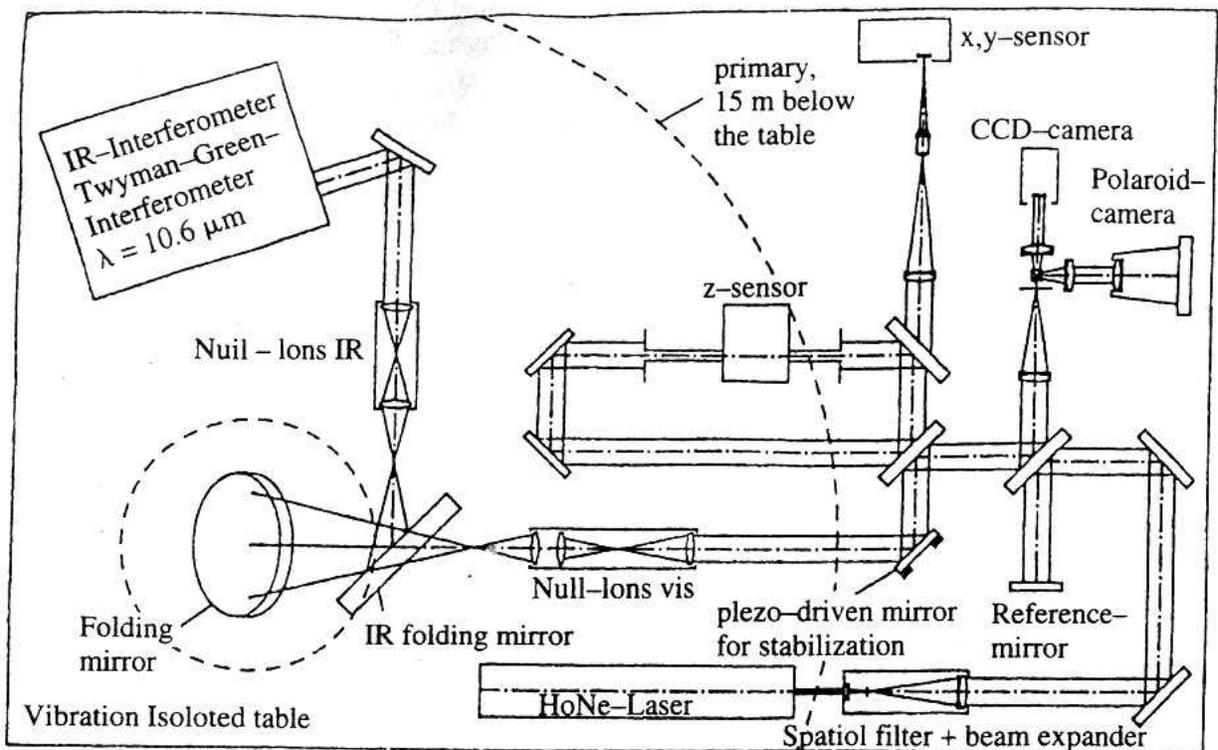


Fig. 2.2.1. Zeiss system of stabilization of the interferogram for large optics measurement.

On-machine measurement is intermittent or stop-and-measure, in which the measuring instrument is placed on or unified with the machine tool. It is important to keep the workpiece held with a chuck during the measurement so as to be able to reprocess immediately if necessary.

2.2.1 Examples of on-machine measurement

On-machine measurement has been widely practised in the manufacture of optical parts. In the polishing of relatively small optical parts, Newton interference fringes formed between a test plate and a workpiece surface are sometimes observed to check the residual error in shape. For large optics, such as telescope primary mirrors, the stop-and-measure method is used from a tall tower. Movements of the measuring instrument relative to the mirror and air turbulence are the main problems to be solved in order to perform measurements of nanometre-order accuracy, especially for interferometers with temporal scanning.

Figure 2.2.1 shows a Zeiss solution to the problem(1). To solve the 10 μm -range vibration problem, adaptive optics and a new type of interferometer with DMI (direct measuring interferometry) are used simultaneously. The adaptive optics compensates three-axis displacement up to 1000 Hz by a piezo-driven mirror. This may be the first successful application of adaptive optics ever to compensate for air turbulence for practical precision measurement.

Unlike the well-known interferometers with fringe scanning, in which several interferograms are detected and processed to map a wavefront, the DMI interferometer uses only one interferogram with narrow-spaced fringes to calculate a phase map in real time. The phase is determined by calculating a video signal of fringes multiplied with sine and cosine signals at the carrier frequency and subsequent application of a low-pass filter. Digital pipeline processing gives DMI some advantages, e.g. insensitivity to vibrations and ability to average a large number of wavefronts so as to reduce random errors such as errors from air turbulence. Profile correction for accurate aspherical lens or mirror surfaces is achieved by repeating a figuring cycle of geometrical measurement and corrective polishing.

The system effects corrective polishing by on-machine measurement. The CSSP omits the re-chucking process by adopting a figuring system having both the measurement and polishing units on a single baseplate. Figure 2.2.3 shows the measurement system for the Z-coordinate. The metrology frame from which all displacements, including the X- and Y-coordinates, are

measured with interferometric sensors is placed above the measurement area of the baseplate. A simple experiment has been conducted to make on-machine measurements of a metal mirror surface immediately after machining with an ultra-precision lathe. The cutting tool was removed and the HIPOSS optical stylus was set to make surface measurements. An example is shown in Fig. 2.2.4, measured with the spindle stopped and the feed slide in operation. Because the slide motion of an ultra-precision lathe is now as accurate as that of the measuring instrument, reliable surface characteristics can be determined by on-machine measurement.

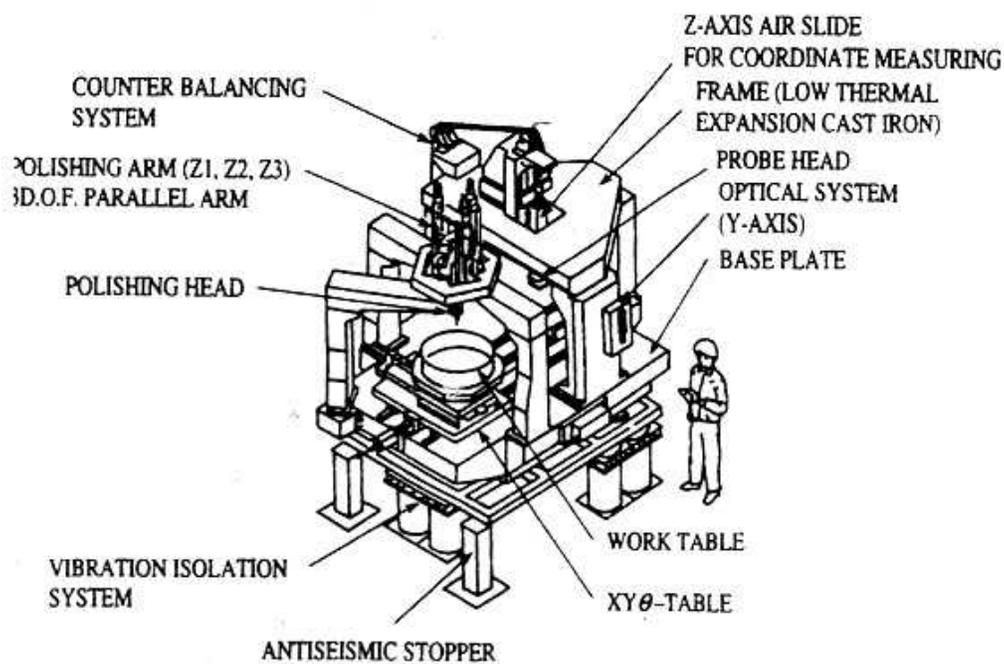


Fig. 2.2.2. Diagram of the Canon super smooth polisher (CSSP).

To save time and avoid some errors, the CSSP (Canon super smooth polisher) system has been developed®, as shown in Fig. 2.2.2.

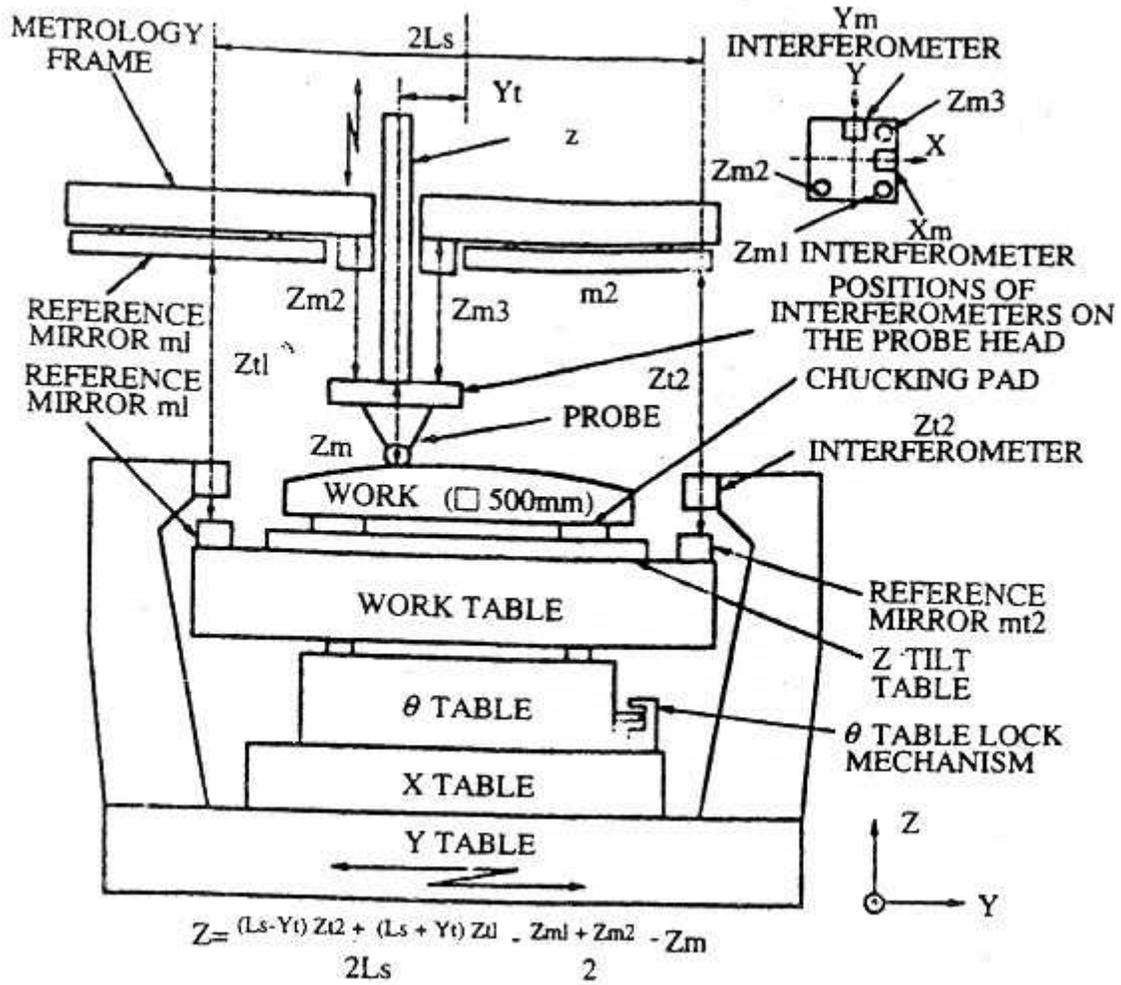


Fig. 2.2.3. Z-coordinate measurement system of the CSSP.

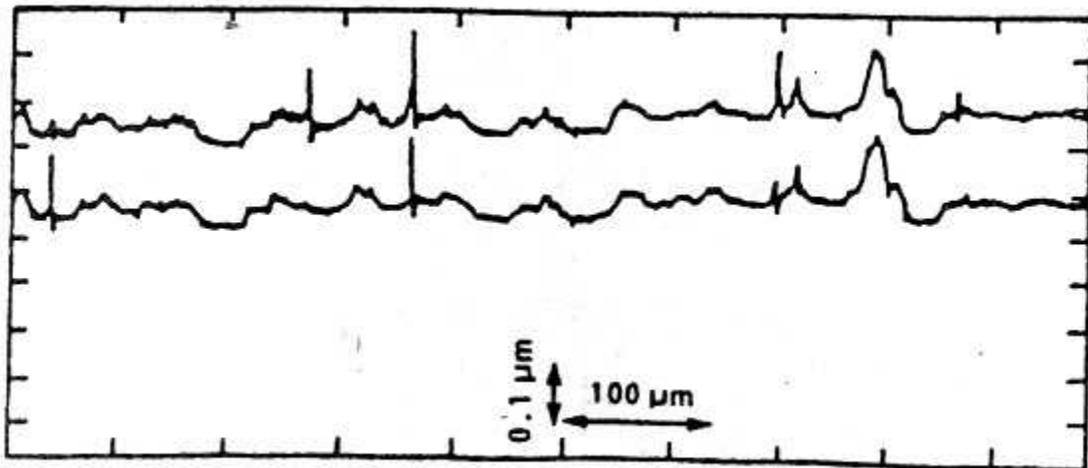


Fig. 2.2.4. Repeatability of on-machine measurement with the HIPOSS.

2.4 Mechanical measuring systems

2.4.1 Introduction

A mechanical measuring system is one that uses mechanical contact with a solid probe to measure the change in displacement or position (i.e. the mechanical quantity) of the part of a sample that we wish to measure (i.e. the measurement point). A mechanical measuring instrument, therefore, consists of a probe, a device to support the probe while converting its motion into an electrical quantity, and a device to analyse the measured results and display the evaluated value. A typical instrument is the high-magnification profile instrument, which successively measures the surface profile from the motion of a solid stylus. The displacement and position of the measurement point are given by the vertical movement and horizontal position, respectively, of the stylus.

In this section we describe the profile instrument in some detail, to highlight the typical features of mechanical measuring systems. In addition, mechanical displacement measuring instruments and capacitance-type sensors with maximum resolutions < 10 nm are mentioned in the discussion on conversion methods for the mechanical quantity.

2.4.2 Features of mechanical measuring systems

Because the mechanical quantity at the measurement point is directly measured in a mechanical measuring system, such a system has the advantages that the measurement procedure is straightforward and the measurement accuracy and resolution can be directly inspected. On the other hand, if the mechanical quantity at the measurement point is measured directly as an electrical or optical quantity, a mechanical-to-electrical/optical conversion takes place; hence the measurement is affected by the electrical or optical characteristics of the surface material of the measured specimen.

In mechanical systems, a measurement force is needed to maintain contact between the probe and measurement point to transfer the mechanical quantity accurately. This force causes deformation at the measurement point. Although this deformation is usually an elastic one, plastic deformation can occur in soft metals, resulting in surface damage. Nevertheless, the causes of measurement errors are relatively easy to identify, in comparison with other measuring methods, and measurements are affected less by environmental factors. Therefore as long as measurement conditions pertaining to the problematic factors have been spelled out, one can

obtain highly stable and reliable measurement results with good repeatability using mechanical measuring systems. It is also relatively easy to manufacture instruments that have identical performances. Since the probe's minute movements must ultimately be converted into an electrical quantity, the detection and conversion system must not disturb those movements. For this purpose, compact and lightweight units that respond linearly by non-contact methods have been developed. Thus in modern mechanical measuring systems, the change in mechanical quantity at the measurement point is mechanically extracted and then measured by a non-contact measuring system.

2.4.3 Features of the high-magnification profile instrument

In profile measurement, a record of the magnified surface profile is obtained by allowing the detector unit which houses the stylus to scan mechanically over the measured surface in one direction to measure the coordinates of consecutive points on the surface profile. Other surface properties are also evaluated. Thus the horizontal scanning line provides the measurement standard, the distance of which from each measured point is measured by the stylus to give the height (i.e. vertical) dimension such as surface roughness. At the same time, the position of the measured point is measured by the feed distance (i.e. horizontal direction). A high-precision feed device is necessary, since the accuracy (i.e. straightness) of the feeding motion directly affects that of the vertical measurement, while the readout accuracy of the feed distance directly affects the horizontal accuracy. Thus there is a vertical and a horizontal measurement resolution. There are other instruments that also mechanically scan the detector unit to determine coordinates of consecutive surface points, such as a device that uses an optical stylus, or the scanning tunnelling microscope; these are basically similar in structure to the above profile instrument.

2.4.4 Vertical resolution of profile instrument

(a) Historical advances in vertical resolution

A method that uses optical levers to magnify microscopic movements of the stylus was developed in the 1920s and used until the early 1950s. It achieved a vertical resolution of ~ 0.1 μm . Research on ways to accomplish magnification by electrical means began in the 1930s, and ever since 1935, when Rank Taylor Hobson Ltd announced the Talysurf Model 1, electrical magnification has become the more usual method. In electrical magnification, a transducer is

used to convert the movement of the stylus into electrical signals, which are then amplified. Several countries undertook research in this field, with the result that a vertical resolution of 0.5 nm was achieved in the 1960s, improving 0.1 nm by the mid-1980s and 0.03 nm by the late 1980s.

(b) Measurement requirements and functions of the instrument

The major requirements for making measurements of ~ 1 nm are discussed below.

(1) Radius of curvature of stylus tip and measurement force In general, if F_s is the static load or measurement force at the mean position when the stylus and sample surface are in contact, the maximum dynamic measurement force of the stylus continually tracing the surface profile can be approximated⁽¹⁾ by $2F_s$. This measurement force causes elastic deformation at the contact point. If the tip radius of the stylus is sufficiently small, however, the maximum pressure at the contact point will be large enough to cause plastic deformation in soft materials such as aluminium. Moving the stylus horizontally in this state will result in a scratch. To prevent this, the stylus assembly must have a small mass, and a coil- or plate-spring mechanism must be installed to exert an upward force to reduce the measurement force within the measurement range. Generally a stylus with 2 μm tip radius is used with a measurement force of 0.7 mN; however, for measurements of surface features of ≤ 10 nm, a stylus with a tip radius of 0.5 or 1.0 μm is used with a measurement force of 0.01-0.3 mN.

(2) Tracking characteristic of stylus When the measurement point makes vertical up/down motions, the limiting frequency at which the contacting stylus is able to track the measurement point faithfully is called the tracking characteristic. This characteristic is inversely proportional to the square root of the amplitude of the measurement point's vertical motion, and roughly proportional to the square root of the measurement force. In general, when the surface profile curve has small amplitudes (i.e. height of protuberances), its wavelengths are also small, giving a large number of waves per unit length. Hence when measuring a surface with a low surface roughness, the feed velocity should be kept low so as not to exceed the stylus's tracking characteristic. For measurements of profile features of < 10 nm, a feedspeed of 2-100 $\mu\text{m s}^{-1}$ is used; the accuracy of the feed motion achieved in these cases is normally a straightness of 50 nm per 25 mm to 3 nm per 3 mm.

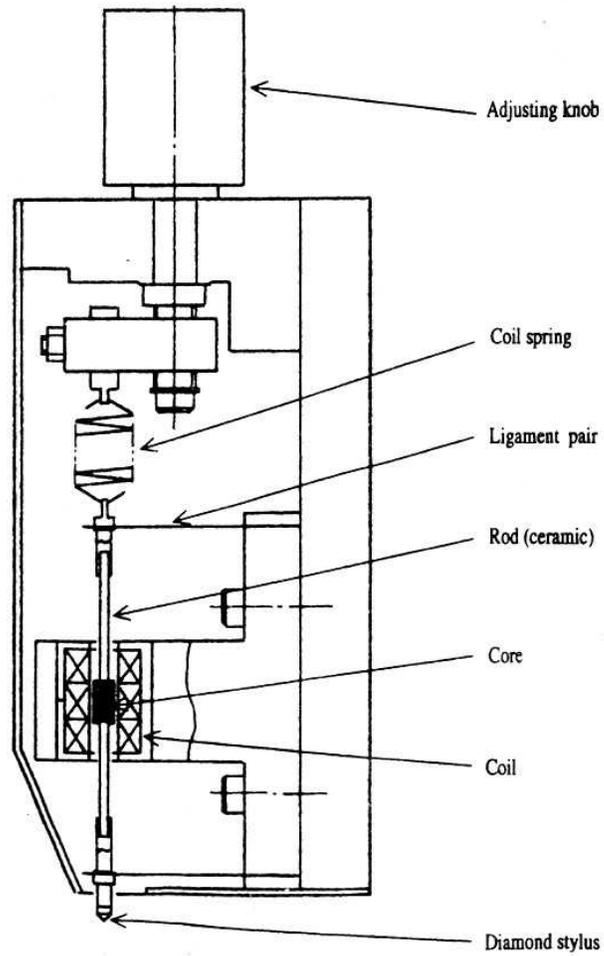


Fig. 2.4.1. Surf-corder ET (Kosaka Laboratory Ltd). Stylus—transducer assembly, showing stylus support mechanism.

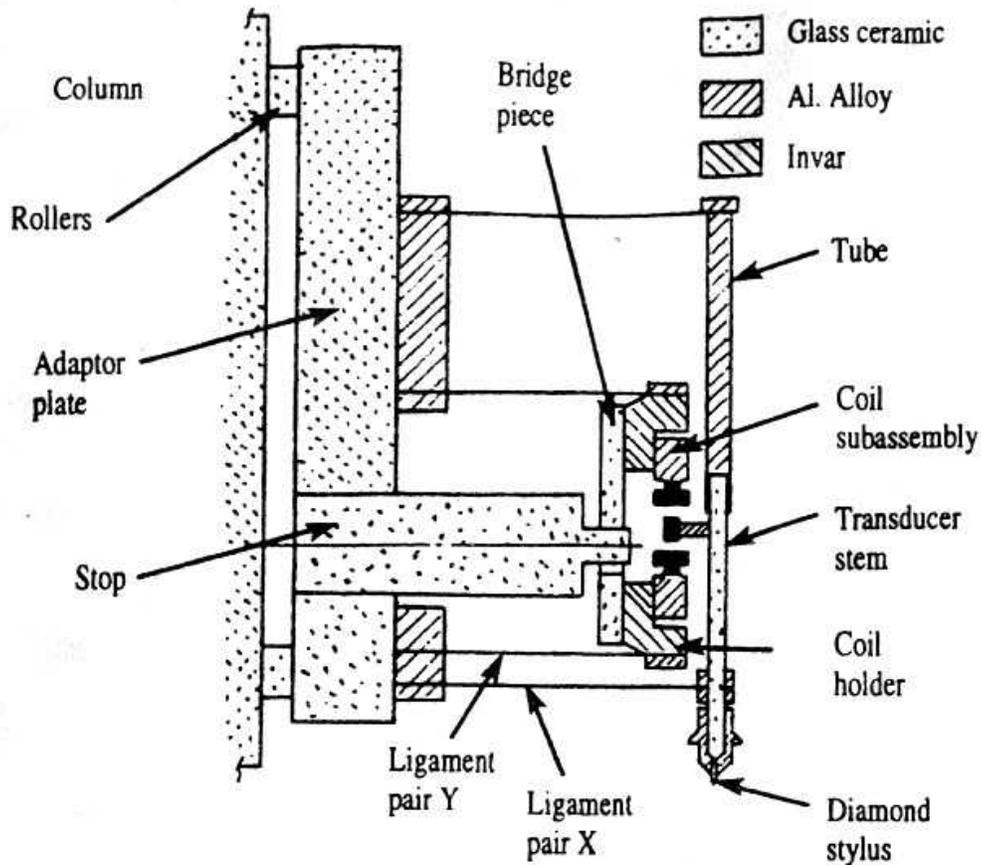


Fig. 2.4.2. Nanostep (Rank Taylor Hobson Ltd). Transducer assembly, showing materials used in measurement loop.

(3) Examples of detector units Figure 2.4.1 shows a high-magnification detector unit (Kosaka Laboratory, Japan), which has a measurement force of 0.2 mN for the mean stylus position, and a maximum tracking characteristic of ~ 100 Hz for a surface roughness of 3 μm . The smallest vertical resolution is 0.5 nm. Figure 2.4.2 shows the construction of the detector unit for the Nanostep (Rank Taylor Hobson, UK). Its newest features include the use of a glass ceramic with a low thermal expansion coefficient for its major parts to minimize the effects of temperature changes. The Nanostep which incorporates this unit achieves a maximum vertical resolution of 0.03 nm, with a system accuracy of $\pm 3-4\%$ of the measured result.

In both of these units, a transducer is used to convert the stylus movement (i.e. the mechanical quantity) to an electrical quantity. To minimize errors of transmission from the stylus to the transducer core, they are assembled as a single unit, and to minimize the measurement force the transducer is made as compact as possible and the mount kept lightweight.

2.4.5 Horizontal resolution of profile instrument

The feed unit consists of a high-precision linear guide mechanism (or circular guide for out-of-roundness measurements), which provides the measurement standard for the vertical direction, and the readout unit for the horizontal position of the stylus. When the horizontal range of measurement is large or when the specimen is large in volume and mass, a feed unit which horizontally moves the detector unit itself is commonly used. Conversely, when measuring vertical displacements of ~ 1 nm, the horizontal measurement range tends to be small and the specimen itself small and lightweight, so it is better, in terms of both construction and accuracy, to move the sample

(a) *Feed unit for detector unit and horizontal resolution*

An example of a feed (or traverse) unit is given in Fig. 2.4.3. The guide mechanism consists of a precision-machined sleeve and guide rod; the sleeve, which holds the detector unit, slides along the fixed guide rod and reads the feed distance on a digital scale. The feed motion has a straightness given by the formula $0.05 + 1.5L/1000$ μm , where L (mm) is the feed distance (i.e. measurement). The maximum horizontal resolution is 0.1-0.05 μm .

Several types of digital scale are used in the feed unit: glass with gratings, metal tape, diffraction grating, magnetic, etc. The scale reads by dividing signals obtained by scanning either a photocell or a magnetic head. The resolution is 0.1-0.05 μm . Since the minimum radius of curvature for the stylus tip that can be achieved today is $\sim 0.1\mu\text{m}$, the maximum horizontal resolution for this measurement system is also 0.1 μm . In contrast, scanning tunnelling microscopes, in which the extremely lightweight probe is moved by a piezo-actuator, can achieve a horizontal resolution of 0.2-0.4 nm.

(b) *Feed unit for sample and horizontal resolution*

The advantage of this kind of feed unit is that it can be designed and fabricated as a separate unit. It is connected to the instrument's main body by a three- point mounting so as to minimize force and heat deformations. A special material is used for the guide plane to provide smoother sliding, and measures have been taken to prevent vibrations that accompany the feed motions. Since these units have a small range of movement, sometimes an encoder is used to determine the feed distance through the feed rotation angle, instead of taking direct measurements of the feed distance. In these units too the horizontal resolution of the feed distance is 0.1 μm , limited by the resolution of position detection.

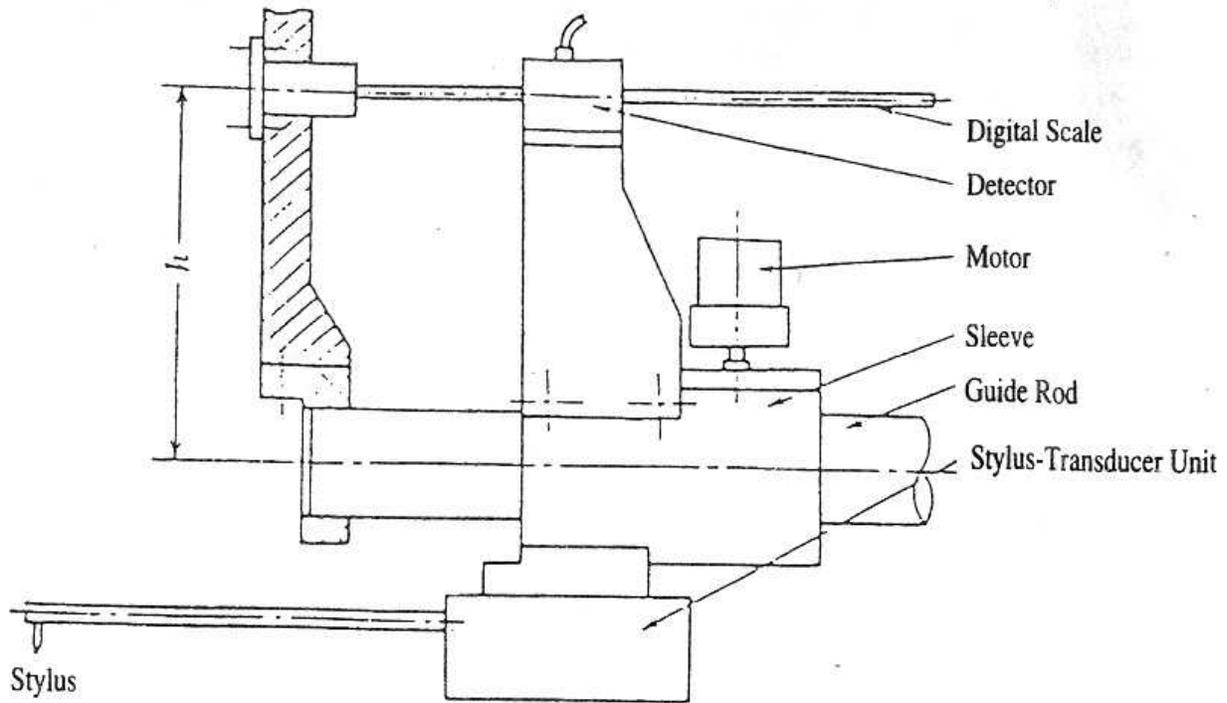


Fig. 2.4.3. Surfcoorder SE (Kosaka Laboratory Ltd). Mechanism of traverse unit.

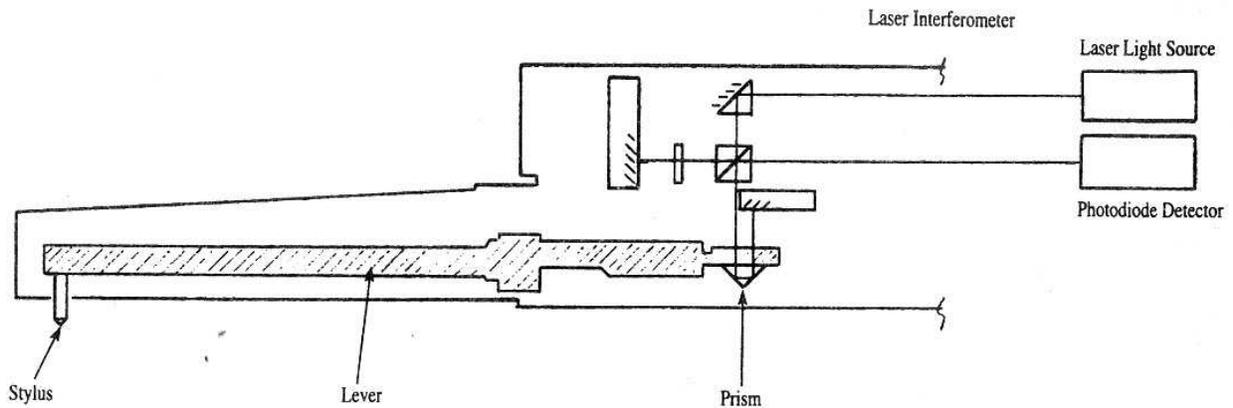


Fig. 2.4.4. From Talysurf S5 (Rank Taylor Hobson Ltd.) Schematic diagram of laser pickup

2.4.6 Other methods of converting the mechanical quantity

(a) Vertical displacements

With a small transducer, it is possible to measure only a narrow range. To overcome this limitation, an instrument (Form Talysurf Model S5) that converts the vertical movement of the stylus into motion of interference fringes has been developed. It can measure a range of 6 mm with a vertical resolution of 10 nm. As shown in Fig. 2.4.4, when the stylus, which is connected to a small prism, moves vertically, the optical path length in the laser interferometer changes,

causing the interference fringes to move. This movement is measured to determine the distance travelled by the stylus.

(b) Detection of horizontal position

One commercial product makes measurements with 8 nm resolution by allowing a semiconductor laser to impinge on a hologram grating and then utilizing the resulting interference of the diffracted light.

(c) Capacitance-type sensors

Non-contact capacitance-type sensors are now often used for position measurements in precision transfer devices and measurements of the dynamic characteristics of rotating samples.

A schematic drawing of such an active-probe sensor (ADE Corporation, USA) is shown in Fig. 2.4.5. It houses a sensor electrode and a balance electrode which provides the reference capacitance; the capacitance difference is used to make measurements. This probe yields an output inversely proportional to the distance between the sensor electrode and specimen surface (mean surface). This output is linearized (with a standard linearity of $\pm 0.2\%$) by a linearizer circuit and then displayed. Although the static resolution is generally considered to be 10 nm at the most, in one instance in which the transfer position for a precision transfer device was measured, a resolution of 1 nm was obtained.

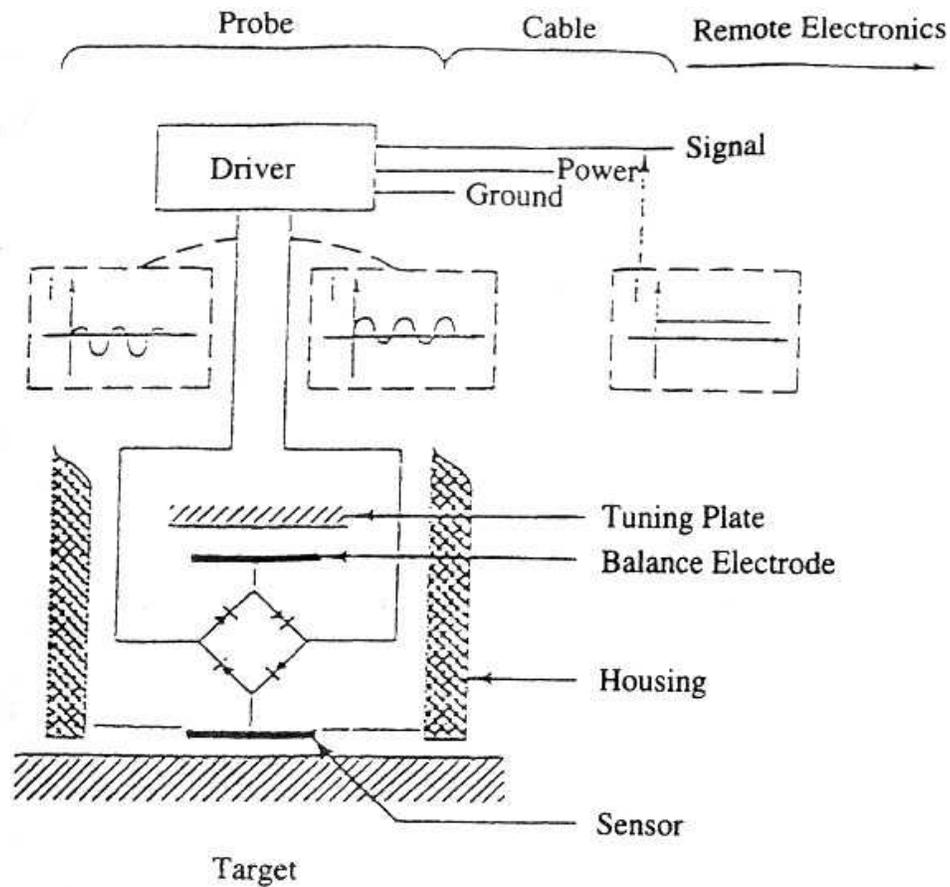


Fig. 2.4.5. Micro Sense-Active Probe (ADE Corporation). Simplified schematic diagram of 'active probe'.

References

1. Ajioka, S. (1966). Dynamic response of stylus, *Journal of the Japan Society of Precision Engineering*, 1, 228.
2. Garratt, J.D. and Bottomly, S.C. (1990). Technology transfer in the development of a nanotopographic instrument, *Nanotechnology*, 1, 38.
3. McRae, D.J. (1987). Non-contact dimensional gaging using capacitive sensing. Paper presented at Sensors Expo.

2.5 Optical measuring systems

Precision measurement is essential for almost all nanometre-order processing. Non-contact, especially optical, measuring systems are now greatly needed in this particular field.

2.5.1 Laser interferometer

Most ultra-precision processing systems have their own high-precision scales integrated with the machine tool or stand-alone measuring instruments in the inspection room. Some types of laser interferometer are used widely as typical precision scales with nanometre-order resolution. In-process or on-machine positioning control of machine tools, semiconductor production systems and three-dimensional measuring instruments are typical applications. The coherence of laser sources permits fringe counting systems with a range up to 50 m. Over the lifetime of the laser tube, the laser wavelength can remain stable to one part in 10^8 or better. The most popular type of laser interferometer uses a heterodyne method with a two-wavelength Zeeman laser or an AO (acousto-optical) element to obtain accurate phase information. Another length-measuring interferometer using a $\frac{1}{8}$ phase plate in one arm and a polarization beam splitter detects three or four signals with phases mutually differing by 90 or 180 degrees⁽¹⁾. The length displacement is calculated from these signals as accurately as it is by a heterodyne interferometer. The main drawback of interferometer systems of nanometre accuracy used in the free atmosphere is the necessity to correct for the refractive index of the air, which can change by one part in 10^5 under the usual conditions. Calculation of and correction for the refractive index by simultaneous measurement of air temperature, pressure, humidity, etc. may be an effective means of overcoming this problem to some extent, but a better way of reducing this source of uncertainty is to use a gas refractometer. If the atmosphere is uniform along the light path, the uncertainty will be reduced to one part in 10^8 by the correction.

2.5.2 Optical figure-measuring instruments

For measurement the figure of ultra-precision machined surfaces, optical interference wavefront detectors are widely used. They are mainly based on the Michelson, Fizeau, Mach-Zehnder, or Young interferometers or holography. The Fizeau-type interferometer with a laser source as shown in Fig. 2.5.1 gives a relatively stable interferogram and is one of the most useful instruments in optical and precision machining workshops. Besides the heterodyne technique mentioned in connection with the length-measuring interferometer, the figure-measuring interferometer uses another technique, fringe scanning⁽²⁾, in which the reference surface is scanned by means of piezoelectric drives. All the interferograms in every scanning position are

detected by a CCD camera and processed in a digital computer to obtain a figure map of nanometre-order resolution.

For accurate phase measurements of large optics over long optical paths, as explained in Section 2.2, the temporal fringe scanning interferometer has certain weak points arising from mutual displacement, vibration and air turbulence between the interferometer and the surface under test. Like the DMI interferometer, the simultaneous phase shift interferometer (SPSI) of Michelson type⁽³⁾, shown in Fig. 2.5.2, has the ability to make accurate phase measurements in dynamic environments where fringe patterns are rapidly changing. The SPSI uses a stabilized single-frequency He—Ne laser and four CCD cameras. Interference fringes at each of the four cameras are phase-shifted 90 degrees relative to one another using polarization techniques. Pixels of the four CCD cameras are aligned with each other exactly. A shutter synchronized to 0.1 ms creates four high-contrast fringes simultaneously, even with severe vibration. The Micro PMI compact Fizeau interferometer with the same principle as the SPSI can measure surface height to better than 10 nm accuracy and is now commercially available.

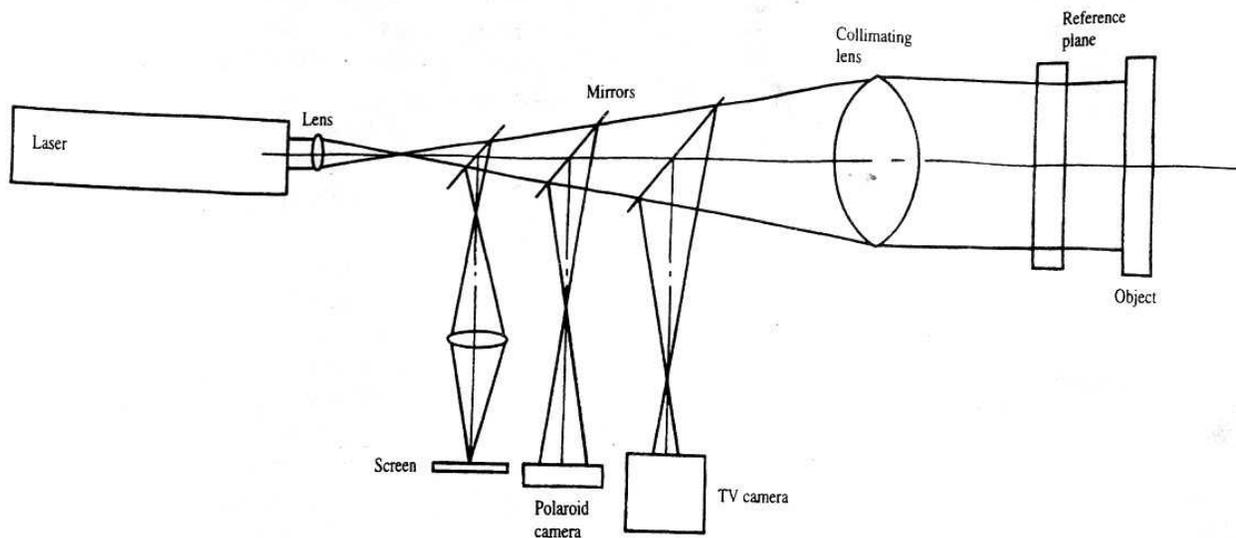


Fig. 2.5.1. Laser interferometer for figure measurement.

2.5.3 Optical surface roughness-measuring instruments

Hitherto, surface roughness has been measured primarily by stylus instruments, as a convenient and versatile method. However, it has been criticized because, it is relatively slow and may damage the surface of soft materials or the function of semiconductors. Several non-contact methods are now being developed and used extensively. Optical stylus focus error detection and

optical interferometers seem the most promising techniques for practical applications. Astigmatism, critical-angle, and knife-edge methods are typical examples of the optical stylus technique, and micro-Fizeau, Mirau, and Michelson instruments are typical interference profilometers.

The principle of the optical stylus using the critical-angle method is shown in Fig. 2.5.3. If the surface under test is at the focus of the lens at B, the laser light passing through the objective lens is converted into parallel flux. A total-reflection prism is positioned to reflect the light at the critical angle and thus the same level of light is incident on the two photodiodes. Consequently the out-of-focus signal becomes zero. When the object surface is at position A close to the lens, the light diverges after passing through the lens. The light on the upper side of the optical axis shown in the figure strikes the prism at an angle smaller than the critical angle. This causes the light to be refracted and pass out of the prism, whereas the light on the lower side of the optical axis is totally reflected at the large incident angle. As a result, a difference in the output of the photodiodes is created, thereby producing an out-of-focus signal. At position C, far from the lens, the opposite phenomenon to that at A occurs and a signal with the reverse sign is obtained. As shown in Fig. 2.5.4, the high-precision optical surface sensor (HIPOSS)⁽⁴⁾ has a half-mirror to split the optical path into two total-reflection prisms and split detectors so as to avoid any effects of object surface inclination on the measured result. The resolution of the HIPOSS with a 0.6 NA (numerical aperture) objective is better than 0.2 nm r.m.s over a linear range of 2 gm. Besides its use as a profilometer, the HIPOSS has practical applications in inprocess and on-machine measurement as described in Sections 2.1 and 2.2.

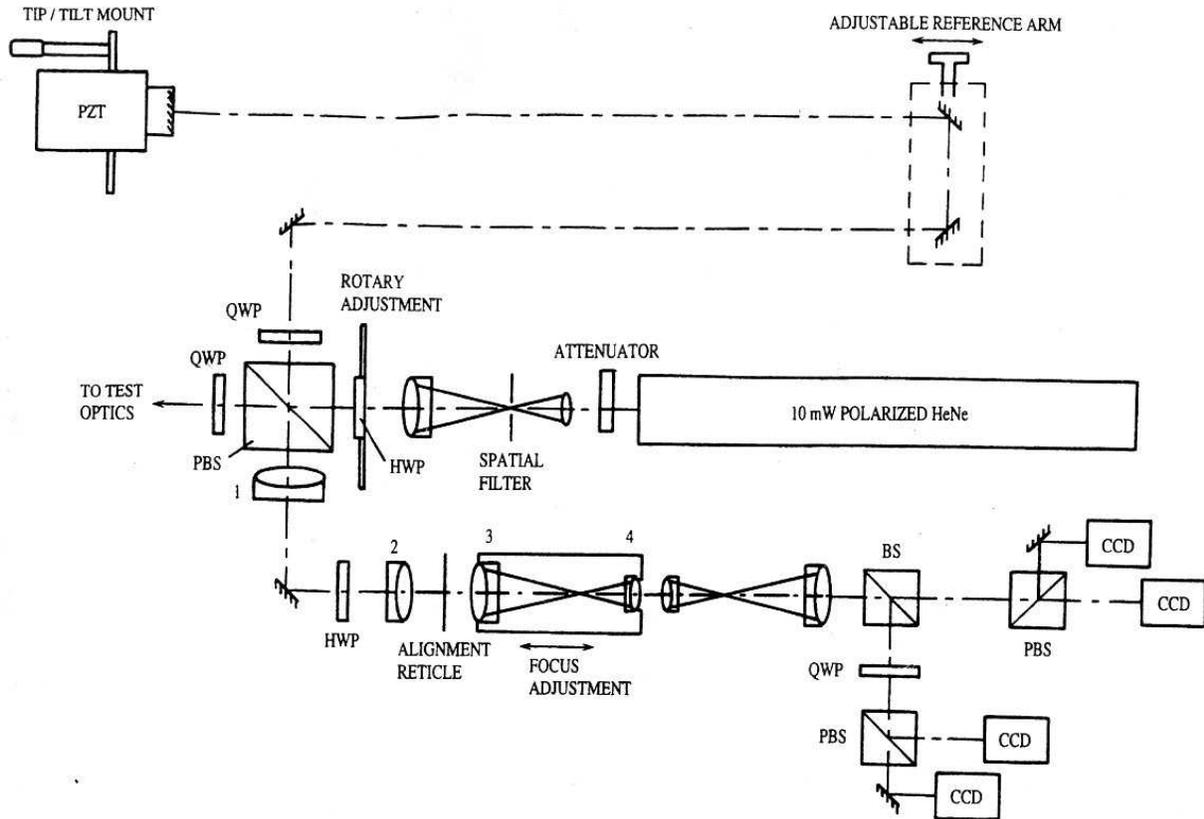


Fig. 2.5.2. Optical layout of the simultaneous phase shift interferometer (SPSI).

Interference profilometers have been gaining wide use in ultra-precision machining shops, the semiconductor industry and academic laboratories. Although the optical principles and the mechanical structures differ somewhat, almost all the characteristics and sophisticated computer algorithms are very similar for both the Mirau⁽⁵⁾ and micro-Fizeau⁽⁶⁾ interference profilometers. Typical resolutions in the vertical and horizontal directions are 0.3 nm and 1.5 μm respectively. A differential measurement method can eliminate common errors and provides a stable signal of extremely high resolution. Figure 2.5.5 shows a circular scanning heterodyne laser interferometer[^] with better than 0.1 nm vertical resolution. A Wollaston prism divides a Zeeman split two-wavelength laser beam by polarization into two beams. The measuring beam is focused on a point on a circular rotating surface, while the reference beam is fixed on the centre of rotation. The reflected beams are recombined at the prism and create an interference signal arising from the beam path difference.

The Nomarski-type microscope is recognized as an excellent optical instrument for observing surface microtexture. An automatic profilometer with the Nomarski prism (8) is shown in Fig.

2.5.6. Two spots focused on the surface are separated by less than the spot size. Since the interference signal from these spots shows the local inclination of the surface, integration of the signal produces the surface micro figure. Because of the cancellation of common errors and little influence of environmental factors, the profilometer achieves a resolution as high as 0.2 nm over along scanning length, 20 cm.

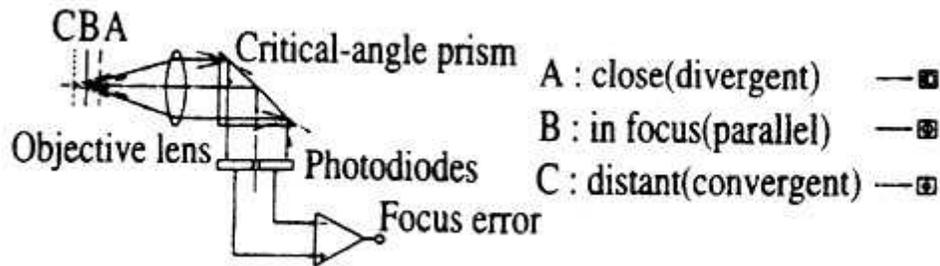


Fig. 2.5.3. Principle of the critical-angle method.

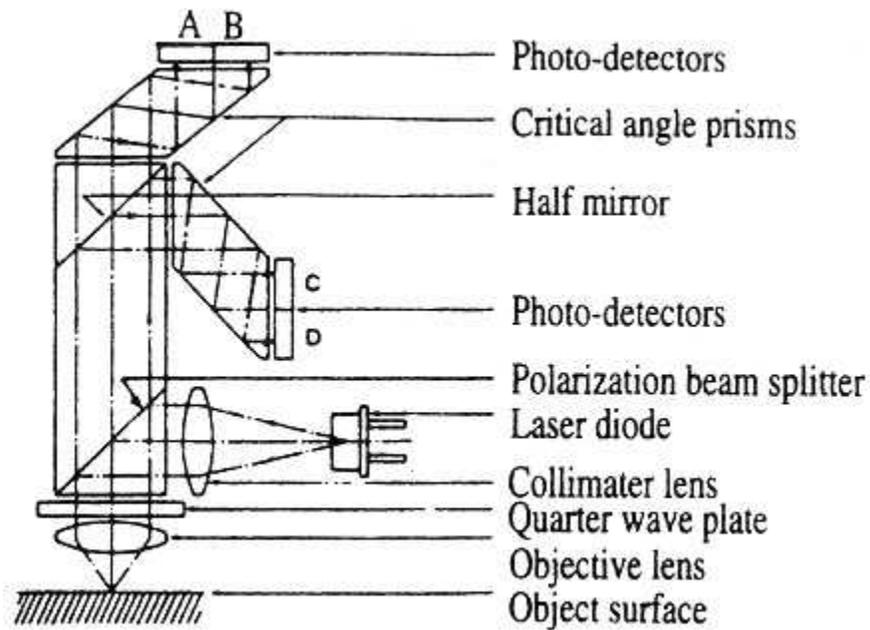


Fig. 2.5.4. Optical path of the high-precision optical surface sensor (HIPOSS).

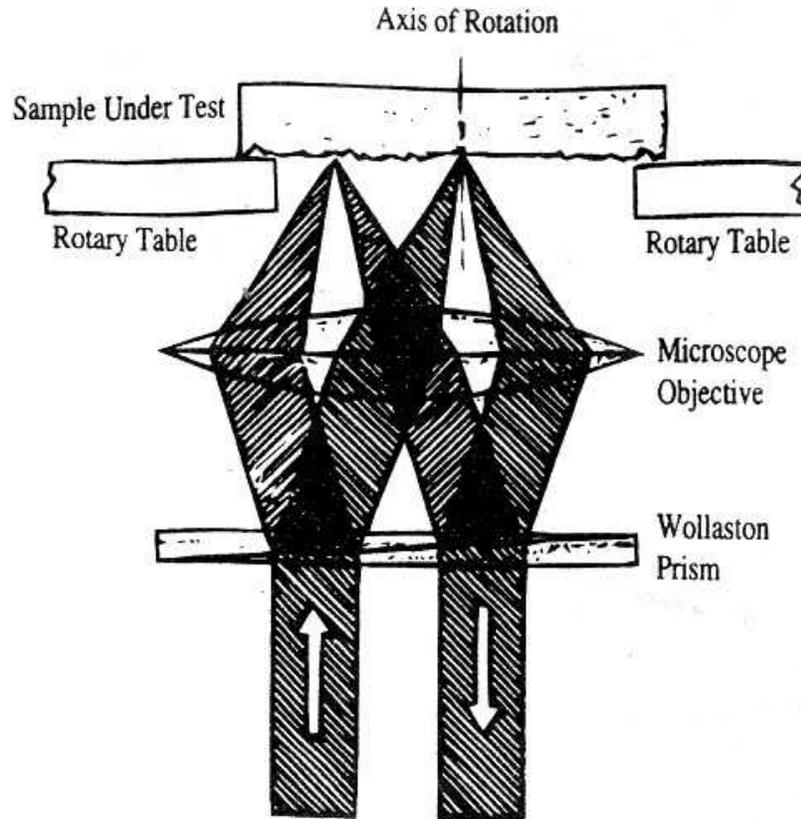


Fig. 2.5.5. Schematic diagram of circular-path inter-ferometer.

References

1. Downs, M.J. and Reine, K.W. (1979). A modulated bi-directional fringe-counting interferometer system for measuring displacement. *Precision Engineering*, 1, 85-8.
2. Bruning, J.H. (1978). Fringe scanning interferometers. In *Optical shop testing*, (ed. D. Malacara), pp. 409-37. Wiley, New York.
3. Koliopoulos, C.L. (1991). Simultaneous phase shift interferometer. *Phase Shift Technology leaflet*, 1-9.
4. Kohno, T., Ozawa, T., Miyamoto, N., and Musha, T. (1988). High precision optical surface sensor. *Applied Optics*, 27, 103-8.
5. Bhushan, B., Wyant, J.C., and Koliopoulos, C.L. (1985). Measurement of surface topography of magnetic tapes by Mirau interferometry. *Applied Optics*, 24, 1489-97.
6. Biegen, J.F. and Smithe, R.A. (1988). High resolution phase measuring laser interferometric microscope for engineering surface metrology. *Proceedings of the SPIE*, 1009, 35-44.
7. Sommargren, G.E. (1981). Optical heterodyne profilometry. *Applied Optics*, 20, 610-18.

8. Bristow, T.C. (1992). Surface roughness and waviness measurements for optical parts. Proceedings of the SPIE, 1720, 119-21.
9. Maeda, S., Hiroi, T., Makihira, H. and Kubota, H. (1991). Automated visual inspection of LSI wafer patterns using a delivative-polarimetry comparison algorithm. Proceedings of the SPIE, 1567, 100-9.

2.6 Electron beam measuring systems: / SEM and TEM

2.6.1 Introduction

Since the electron beam has a short wavelength, it provides a much higher resolution than light when used in a microscope. This excellent spatial resolution is taken advantage of in the use of the scanning electron microscope (SEM) and transmission electron microscope (TEM) for measurements. The line-width measurement system and the electron beam tester are established measuring systems using electron beams, both of which were developed for the semiconductor industry. When ‘measurement system’ is defined to include surface analysis systems and microscopes, EPMA (electron probe microanalyser), SAM (scanning auger microscope), SEM and TEM are all measurement systems effectively used in the field of industrial measurement.

This section describes the line-width measurement system and the morphology inspection system for wafers (here after called ‘wafer inspection systems’) as examples of measurement systems using electron beams. It describes some electron optical technologies adopted to improve resolution, which directly influences the measurement accuracy. It also describes some applications of the TEM as an inspection system for the internal structures of semiconductor devices. This section mostly describes applications in the semiconductor field, which is only natural in view of the fact that semiconductor devices were the first nanotechnological products into which nanometre- order structures were incorporated practically.

2.6.2 Wafer inspection systems

(a) Required resolution

As is well known, large-scale integration of devices and improvement of their functions and performance have taken place rapidly in the semiconductor industry. Along with this progress, the minimum-sized unit for device design (called the ‘design rule’) is becoming smaller year by year. For example, with the 256M DRAM now under development, this size is $\sim 0.2\mu\text{m}$. The size

of defects or errors that influence the electrical properties of devices is said to be $\frac{1}{10}$ of the design rule. Therefore it is necessary to control the dimension of each part of a device to an accuracy of $\frac{1}{10}$ of the design rule. Moreover, wafer inspection systems require a measuring accuracy of $\frac{1}{4}$ to $\frac{1}{6}$ of the fabrication accuracy. Hence for a 256M DRAM a system with a resolution of < 10 nm is required.

(b) Need for low accelerating voltage and its influence on resolution

Wafer inspection systems carry out dimensional measurements and morphological observations on resist patterns and etching patterns formed on Si wafers. Since the wafer surface is mostly covered with insulators, it will be negatively charged during the observation with high accelerating voltages that are used by ordinary SEM (i.e. 10-20 kV). Moreover, when a high-energy electron beam is targeted on an MOS transistor, a major component in today's semiconductor devices, the electrical properties of the transistor deteriorate or are destroyed®. To alleviate these problems, wafer inspection systems are used exclusively at low accelerating voltages, ~ 1 kV.

The SEM's probe size d (which determines the resolution) can be approximately expressed by the following equation⁽²⁾:

$$d^2 = \left(\frac{2}{\pi\alpha} \sqrt{\frac{1}{B}} \right)^2 + \left(\frac{1.22\lambda}{\alpha} \right)^2 + \left(C_c \alpha \frac{\delta V}{V} \right)^2 + \left(\frac{1}{2} C_s \alpha^3 \right)^2 \quad (2.6.1)$$

where α is the illumination angle of the probe, δV is the full-width half-maximum of the initial energy of the electrons (V) at the electron source, B is the brightness of the electron gun, which is proportional to the accelerating voltage V , I is the probe current, λ is the wavelength of the electron beam, which is proportional to $1/\sqrt{V}$ and C_s , C_c are the spherical and chromatic aberration coefficients respectively. As is clear from the dependence of the above equation on V , lowering V increases all the terms except for spherical aberration. With the SEM using a conventional thermal electron gun (TEG), if the accelerating voltage is lowered to 1 kV, d is ~ 40 nm. To reduce d at low accelerating voltages, it is necessary to use an electron gun with large B and small δV values and to reduce the C_s and C_c of the objective lens.

(c) Field emission electron gun (FEG)

The field emission electron gun is characterised by high brightness and small energy spread. The brightness is as high as $10^8 \text{ A cm}^{-2} \text{ sr}^{-1}$ even at 1 kV (in contrast to about $\sim 10^3$ for a TEG). Therefore, under typical operating conditions ($I = 10 \text{ pA}$, $\alpha = 5 \text{ mrad}$), the first term of eqn (2.6.1) is 0.4 nm, which can be neglected compared with the other terms. δV is $\sim 0.3 \text{ V}$ for a cold FEG and $\sim 0.5 \text{ V}$ for a thermal FEG. These values are $\frac{1}{7}$ and $\frac{1}{4}$ of that of a TEG. Hence the use of the FEG makes it possible to improve considerably the resolution at low accelerating voltages.

(d) *Improvement of objective lens aberrations*

Since the conventional SEM is designed for multipurpose use, its objective lens is not optimal at low accelerating voltages. If the objective lens is specifically designed for use at low accelerating voltages, this creates more room for design improvements. For the magnetic lens it is known that the larger the excitation parameter (J^2/V), where J is the number of ampere-turns, the higher is the lens performance. At low accelerating voltages, a large J^2/V value can be obtained with a relatively small J . Thus the magnetic circuit can be made thin and small, with a simpler cooling method or without cooling. Furthermore, an extreme modification of lens shape becomes possible.

Since the line-width measurement system is mainly intended for measuring the line width automatically, the wafer is placed horizontally. With this arrangement, it is possible to reduce drastically the 'working distance' (i.e. the distance between the wafer and objective lens), thus allowing the C_s and C_c values to be reduced to millimetre order. Figure 2.6.1 shows a cross-section of an objective lens that can be used up to 3 kV. As seen in the figure, the wafer is immersed in the lens field and the electron beam is focused by the pre-field on to the specimen. With this lens, $J^2/V = 269 \text{ A}^2 \text{ V}^{-1}$, $C_s = 3.2 \text{ mm}$ and $C_c = 3.4 \text{ mm}$ were obtained. The calculated probe diameter at 1 kV was 5 nm. In this example, secondary electrons are detected by a detector placed above the lens. Because of their low energy, the secondary electrons are trapped by the magnetic force line and effectively transported upward. At the point where the magnetic field disappears, however, they become dispersed. To prevent this dispersion and guide the secondary electrons to the detector with no loss, a group of properly designed electrodes is necessary. The practical use of a high-performance lens such as shown in Fig. 2.6.1 was made possible by the development of a computer program that permitted optimal design of the transport electrodes®.

Figure 2.6.2(a) shows a high-resolution image of a contact hole obtained with a line-width measurement system that incorporates the above lens and a thermal FEG. Resist residues on the bottom surface of the hole are clearly seen. An example of line-width measurement is shown in Fig. 2.6.2(b).

The morphology inspection system is mainly intended for observing fine structures formed on a wafer surface. With this system, therefore, it is important that the wafer can be tilted as much as possible. Although the objective lens of a conventional multipurpose SEM (Fig. 2.6.3) has a specimen-tilting mechanism, the tilt angle of a large-diameter wafer is limited to $\sim 45^\circ$ at a working distance of 15 mm. In this case, the C_s and C_c values were 61 mm and 26 mm respectively. If the working distance is made longer, the tilt angle can be made larger, but this rapidly increases both C_s and C_c . To solve this dilemma and obtain an objective lens which allows the specimen to be tilted up to 60° and still have small C_s and C_c values, the ‘conical lens’ proposed by Bassett and Mulvey⁽⁴⁾ was put to practical use. Figure 2.6.4 shows a conical lens with an apical angle of 60° . This lens eliminates the inner polepiece and uses the magnetic field produced by the coil directly as the lens field. The outer shroud of magnetic material, acting as the return circuit for the generated magnetic flux, serves to reduce the total ampere-turns and eliminate the magnetic field leakage out of the lens. As shown in the figure, the profile of the axial magnetic field distribution has a peak in the vicinity of the coil’s lower end and decreases gradually towards the upper end. The rate of this decrease can be adjusted by controlling the coil length along the conic generator and thickness distribution. The wafer is placed slightly below the virtual apex of the cone, so that it can be tilted up to 60° with a relatively small working distance. The C_s and C_c values obtained were 33 mm and 15 mm respectively, at a working distance of 6.5 mm — about half those of a conventional objective lens. The heat generated by the coil is transmitted to the specimen chamber wall by means of an inner coil former of copper. The maximum temperature rise of the coil was 30°C at 512 ampere-turns (the focusing ampere-turns for a 5 kV electron beam).

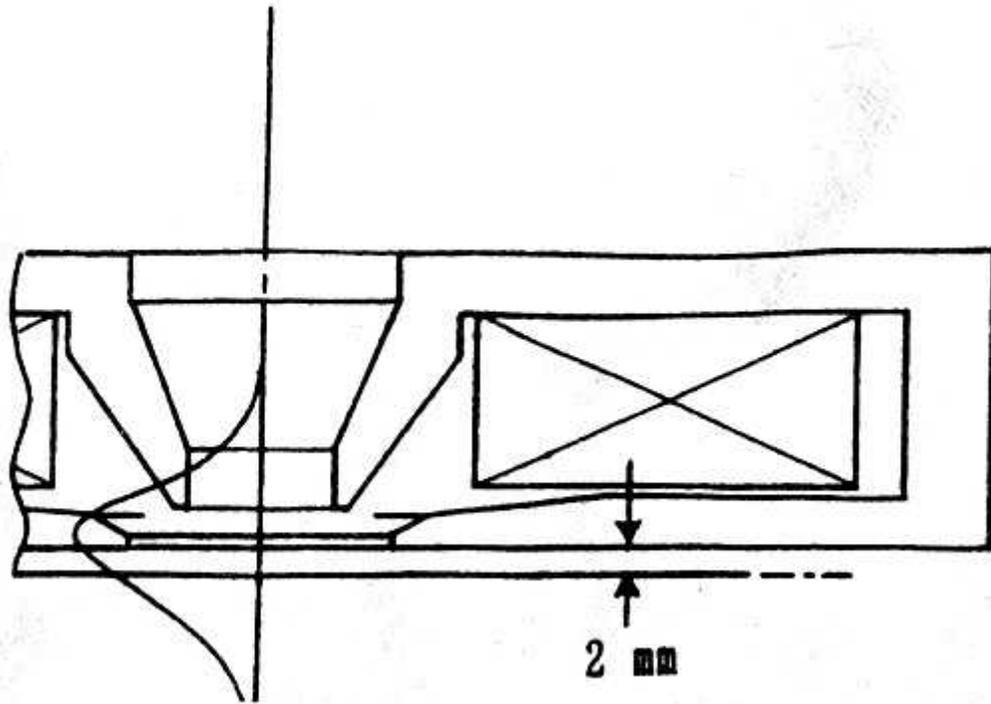


Fig. 2.6.1. Objective lens for line-width measurement system. $C_s = 3.2$ mm, $C_c = 3.4$ mm (at working distance 2 mm), $J_{\max} = 986$ (at 3 kV).

Figure 2.6.5 shows an image of the resist pattern of a 60° tilted wafer obtained with the morphological inspection system incorporating this lens and a thermal FEG. The side wall of the resist pattern, which cannot be seen by observation from above, is observed three-dimensionally.

2.6.3 Transmission electron microscope (TEM)

The TEM is the only instrument that allows observations of the internal structure of a specimen obtained with an electron beam line-width measurement system incorporating the lens shown in Fig. 2.6.1 and thermal FEG ($V = 0.8$ kV). Average line width of five measurements marked + is $1.173 \mu\text{m}$. Edge detection was done automatically by the threshold method with atomic resolution. TEM resolutions, require thin-sectioning of the specimen, so the instrument has rarely been used for the inspection of the internal structures of semiconductor devices. With the recent advent of focused ion beam (FIB) equipment, however, a selective thinning technique⁽⁵⁾ has been developed for TEM observation. Hence major efforts are now being made to use the TEM to inspect the internal structures of semiconductor devices. Another factor in the heightened interest in the TEM is that it now allows elemental analysis of microareas, $< 1\text{nm}$. Even with the TEM, the amount of current that can be applied with a fixed probe size is markedly increased by

fitting it with an FEG. Figure 2.6.6 shows the relation between the probe current and probe size in a 200 kV TEM. When equipped with the FEG, the TEM provided a probe current of 100 pA at a probe diameter of 0.5 nm, allowing X-ray analysis with an EDS (energy-dispersive X-ray spectrometer). Figure 2.6.7 shows a TEM image of the insulation layer (~ 10 nm) of a trench capacitor, and the results of X-ray analysis of the three thin layers ($\text{SiO}_2\text{-Si}_3\text{N}_4\text{-SiO}_2$) constituting the insulation layer. It is clearly seen that nitrogen is concentrated at point B in the figure.

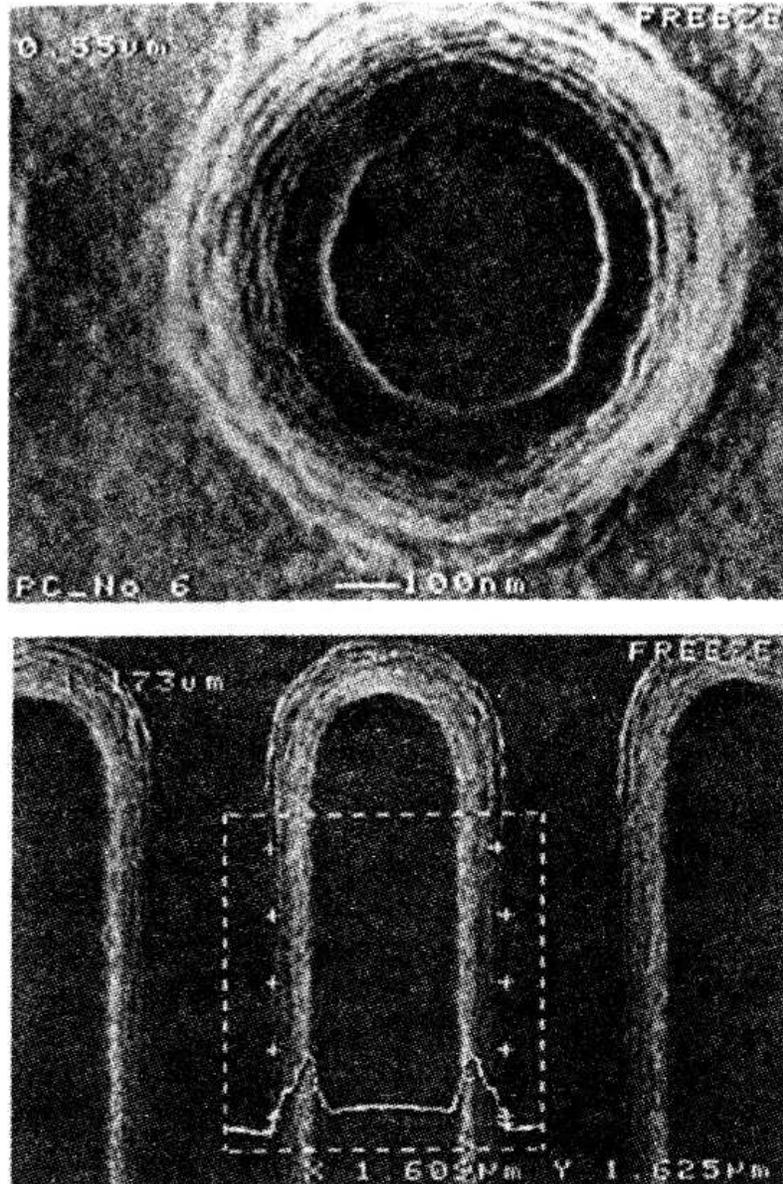


Fig. 2.6.2. (a) High-resolution image of contact hole, (b) Example of line-width measurement,

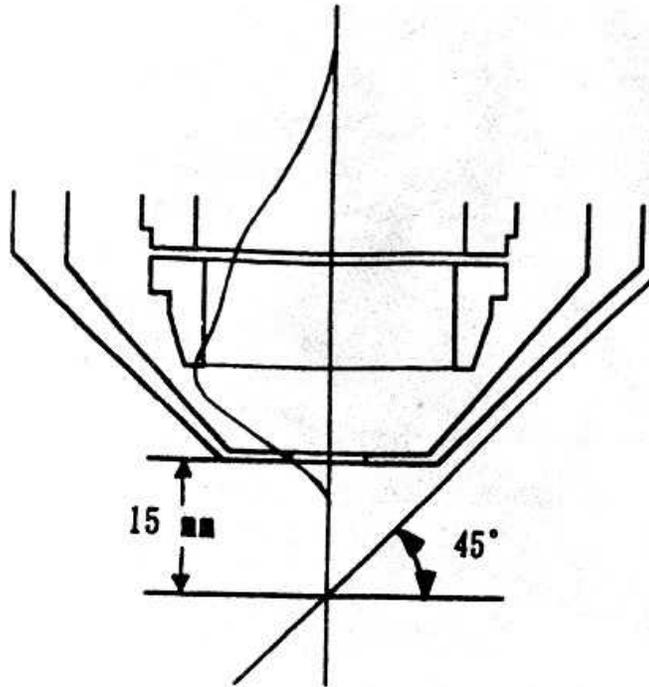


Fig. 2.6.3. Objective lens of a conventional SEM. $C_s = 61$ mm, $C_c = 26$ mm, maximum tilt angle 45° (at working distance 15 mm).

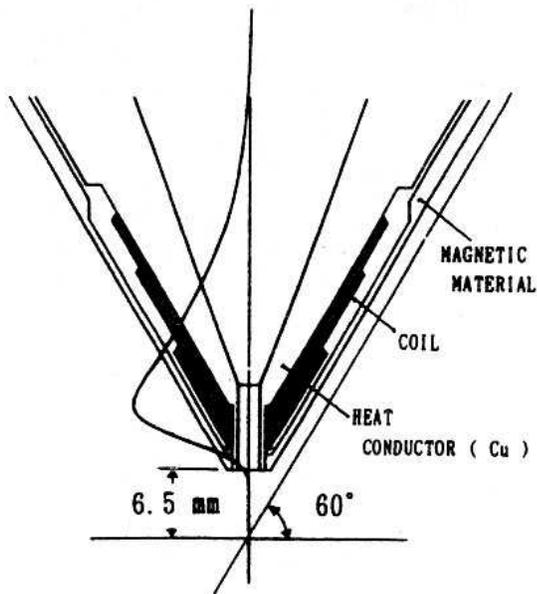


Fig. 2.6.4. Conical lens optimized for use with 60° tilting. $C_s = 33$ mm, $C_c = 15$ mm (at working distance 6.5 mm), $J_{\max} = 512$ (at 5 kV).

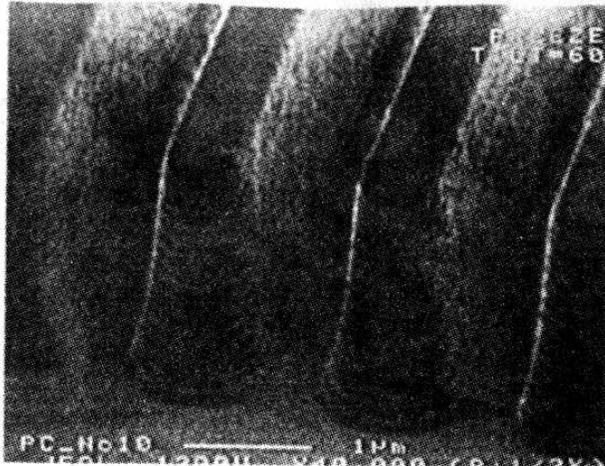


Fig. 2.6.5. Line and space pattern of resist on 60°-tilted wafer, obtained with a wafer inspection system incorporating the conical lens shown in Fig. 2.6.4 and thermal FEG ($V = 1$ kV, working distance 6.5 mm).

References

1. Ura, K. and Fujioka, H. (1989). *Advances in Electronics and Electron Physics*, 73, 260.
2. Wells, O. (1974). *Scanning Electron Microscopy*, p. 75.
3. Munro, E. and Rouse, J. (1989). *Journal of Vacuum Science & Technology*, B7, 1891.
4. Bassett, R. and Mulvey T. (1972). US Patent No. 3, 707, 628.
5. Szot, J., Hornsey R., Ohnishi, T., and Minagawa, S. (1992). *Journal of Vacuum Science & Technology*, B10, 575.

2.9 Pattern recognition and inspection systems

2.9.1 Basic concept of pattern recognition systems

Pattern recognition is the technology of analysing pictorial information using digital computers. Although it was once a very specialized and expensive technology, with rapid advances in digital computers, pattern recognition technology has emerged from the research laboratory and is being used in a wide array of applications such as FA (factory automation), OA (office automation), CG (computer graphics), medical systems, publishing, security, remote sensing, and the arts.

A basic pattern recognition system is shown in Fig. 2.9.1. In general, such a system consists of an illumination source, a sensor, an image processor and a display unit. How the image processing unit handles the data is shown in Fig. 2.9.2. This system consists of an A—D (analogue—digital) converter, image memory, image processor, D—A converter, CPU,

program memory and keyboard. Image data observed by a camera are digitized by the A—D converter and stored in the image memory. The data are then transferred to the image processor to be processed by an algorithm, and are reconverted to analogue form by the D—A converter and displayed.

2.9.2 Image data

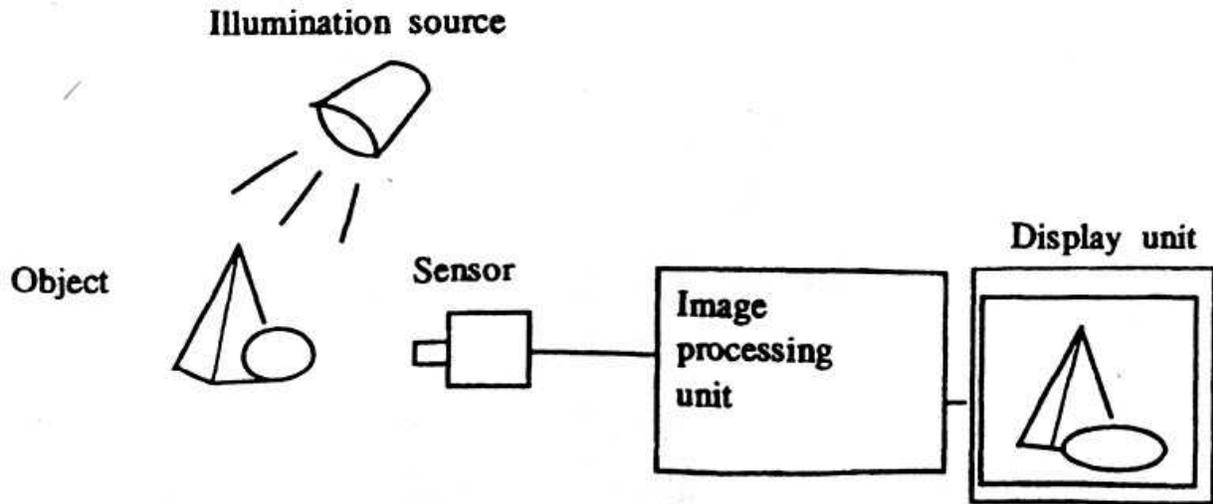


Fig. 2.9.1. Pattern recognition system.

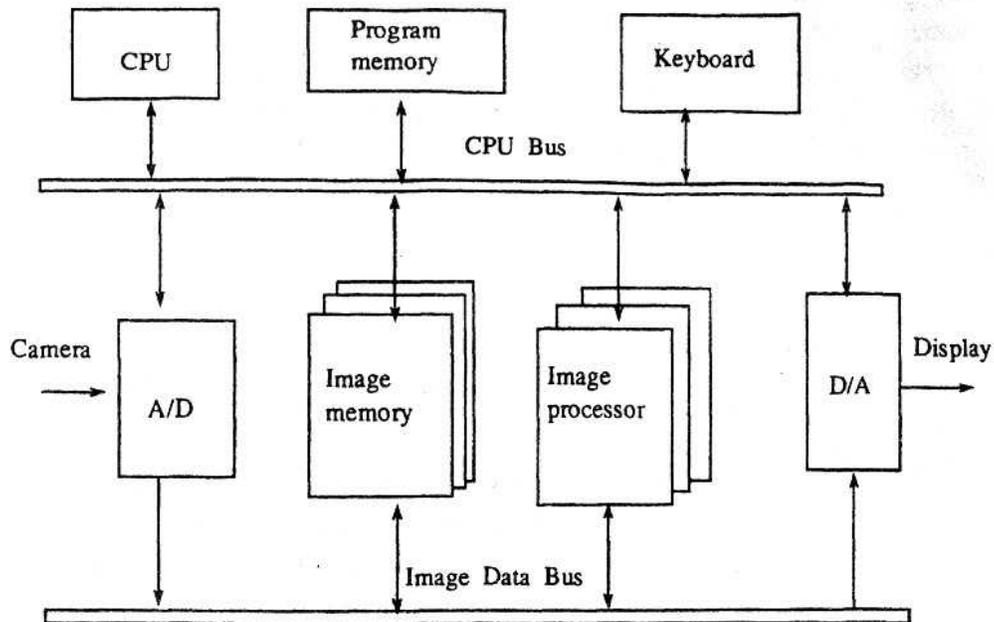


Fig. 2.9.2. Block diagram of image processing unit.

The image data consist of two-dimensional numerical data corresponding to an observed image. The pixel is the minimum sampling unit for digitization; one pixel has two dimensional position data and one numerical datum which represents the observed brightness, as shown in Fig. 2.9.3. In the computer, the image data are represented by a two-dimensional $M \times N$ matrix as shown in Fig. 2.9.4. The elements in the matrix represent the brightness data. Generally the image processing algorithm can be divided into pre-processing and main processing algorithms. The pre-processing algorithm accomplishes normalization by geometrical transformations, sharpening by filtering, and contrast enhancement by grey-level transformations, and elimination of noise by smoothing. The main algorithm analyses or evaluates the pre-processed input image for feature extraction including edge detection, clustering, segmentation, texture analysis, and pattern matching. The main algorithm also carries out basic operations such as binarization, affine transformation, grey-level transformation, filtering, two-dimensional Fourier transformation, and pixel operations for addition, subtraction, multiplication, and division.

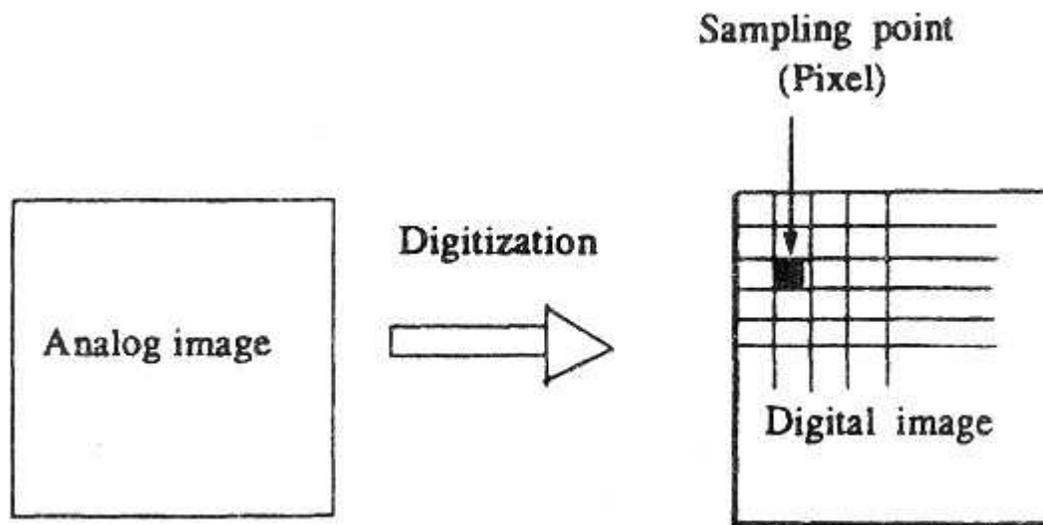


Fig. 2.9.3. Representation of image data.

2.9.3 Examples of image processing algorithms

Figure 2.9.5 shows the visual images of an ARI (assembly robot with intelligence)⁽¹⁾. The ARI has two vision systems, corresponding to the left and right eyes. The pair of stereo images from the two vision systems are analysed together. The edges and vertices of an object are extracted and a pair of two-dimensional object descriptions is generated. The three-dimensional

information is then created by combining the two-dimensional descriptions. In this example, white lines represent the perceived edge positions.

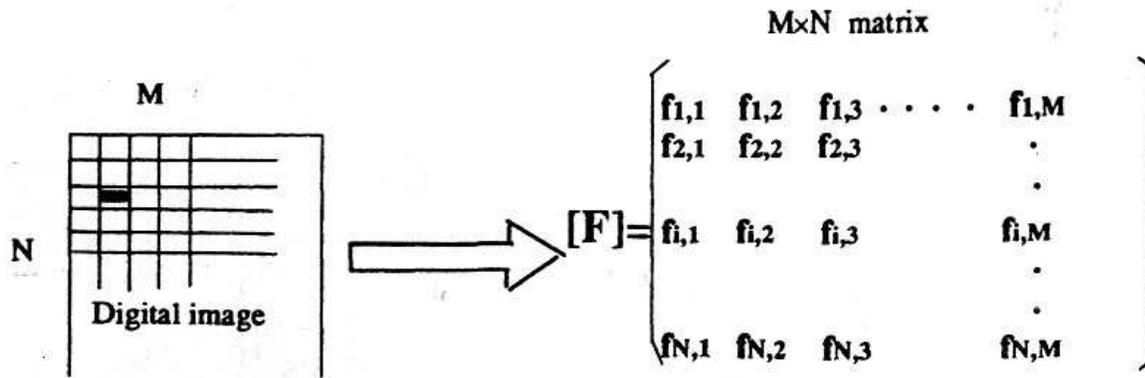


Fig. 2.9.4. Image data representation by an $M \times N$ matrix.

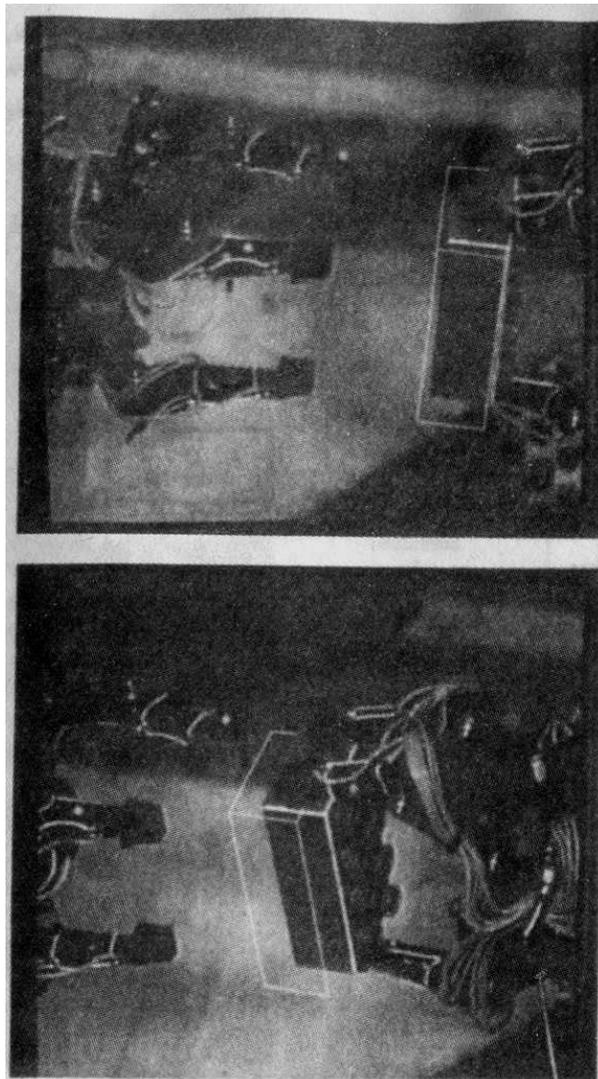


Fig. 2.9.5. Vision images for an ARI robot top, from left camera; bottom, from right camera.

A second example shows how a combination of image processing algorithms is used to detect small defects on printed-circuit boards. As shown in Fig. 2.9.6, after repeated expansion and contraction of the image, small defects are detected and eliminated. In this method, the input pattern creates a 'pseudo standard' pattern by itself, which is then compared with the original input pattern⁽²⁾.

A third example shows an automatic inspection system using a shadow image method for inspecting chip components on a printed-circuit board⁽³⁾. As shown in Fig. 2.9.7, the printed-circuit board is illuminated alternately from two different oblique directions and the corresponding pictures are stored in separate memories. Next, the two pictures in the memories are superimposed and one is subtracted from the other, leaving only the shadow image which corresponds to the chip mounting error. Using this method, the vast amount of unnecessary image data of printed patterns and characters on the board is removed, making possible very rapid inspection as compared with pixel-to-pixel calculation. In this algorithm, grey-level transformations and pixel subtractions are used effectively.

2.9.4 Industrial applications of pattern recognition

Today, many skilled manual operations of assembly, inspection, and adjustment are carried out on production lines. Examples include the magnetic alignment of an electron gun and the electron beam landing control for colour picture tubes (CPT) and colour display tubes (CDT), or the optical alignment of pickup heads for compact disc players and so on. There are many factors governing production quality, such as parts accuracies, assembled accuracies, level of operator's skill, and the setting of references in manual operations. Thus these manual operations require highly skilled work and are difficult to automate.

The problems concerning skilled work can be summarized as follows:

1. Skilled and trained personnel are needed for these operations.
2. Variability of product quality is basically high, even if operators are highly skilled.
3. Training and education of operators involve costs.
4. These skilled operations can become a bottleneck in flexible manufacturing systems.

In various industrial applications, pattern recognition technology can be a very powerful and versatile tool for streamlining or automating production lines. It can be used not only to reduce

costs but also to automate dangerous, tedious, inexact, or time-consuming production processes done by operators.

Table 2.9.1 Specification for colour display tube (CDT)

Colour purity (μm)	5
Beam convergency (mm)	
Centre position	0.05
Comer position	0.45

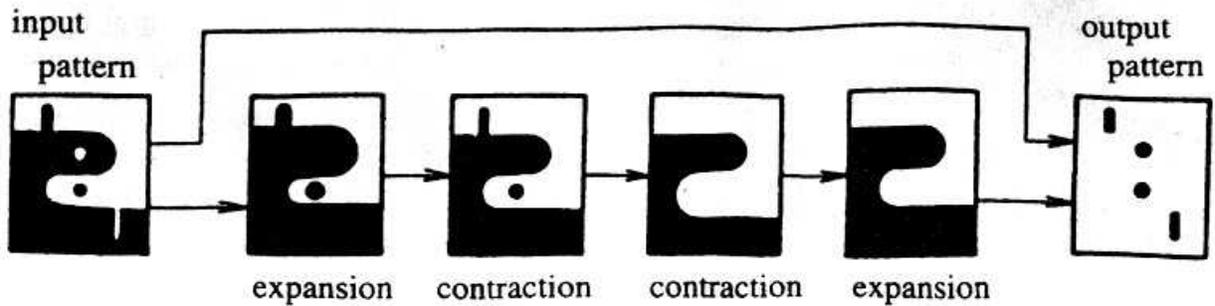


Fig. 2.9.6. Principle of the expansion—contraction method⁽²⁾

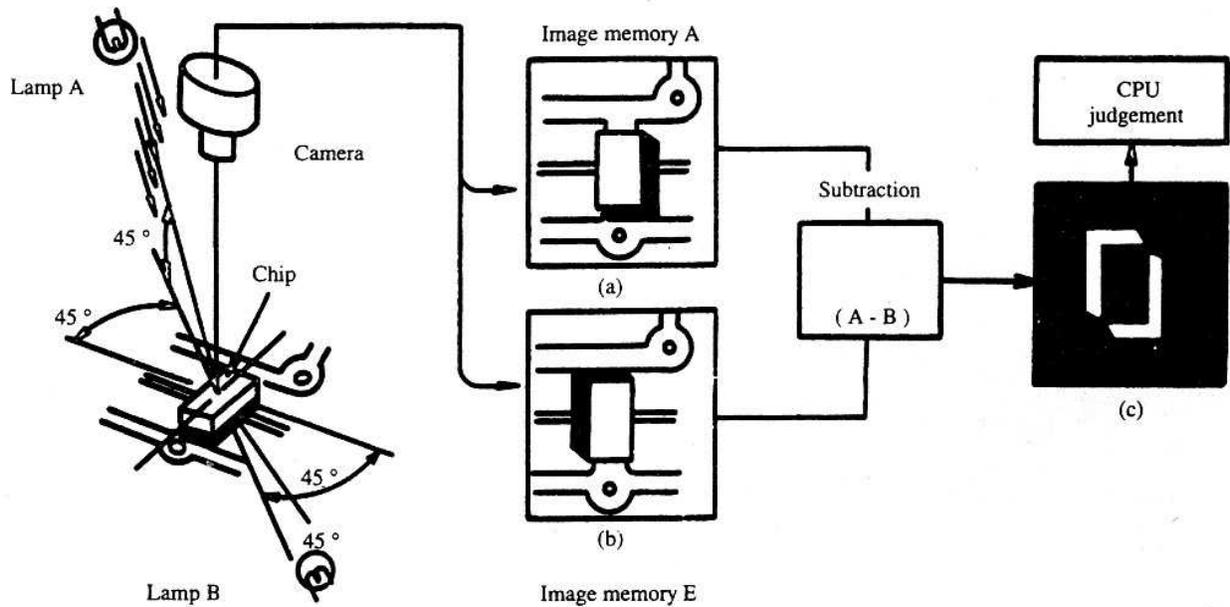


Fig. 2.9.7. Principle of the shadow image method⁽³⁾.

There three main application fields for pattern recognition technology in factory automation:

(1) printed-circuit boards and electrical parts assembly

- (a) defect detection in printed-circuit boards
- (b) automatic position recognition for bonding or mounting machines
- (c) inspection of parts on printed circuit boards (e.g. positioning, missing parts)
- (d) inspection of soldered parts (e.g. bridge or short circuit, excess or lack of soldering)

(2) displays

- (a) automatic inspection of CPTs, CDTs, and LCDs
- (b) inspection of shadow masks
- (c) inspection of fluorescent screens

(3) semiconductors

- (a) alignment or positioning for steppers
- (b) inspection of mask patterns or wafer patterns
- (c) monitoring of etching or photolithographic processes
- (d) inspection of moulding, lead frame or marking

The automatic adjustment of CDT picture quality is one example of micrometre-order adjustments using pattern recognition technology. The construction and specifications of the CDT are shown in Fig. 2.9.8 and Table 2.9.1 respectively. The electron beam emitted by an electron gun passes through the shadow mask and strikes the fluorescent screen. To control the electron beam, the positions and rotations of a deflection yoke and ring magnets, positioned around the neck of the CDT, are adjusted to obtain the best picture quality.

On a conventional production line, adjustments to obtain the required picture quality are made by skilled workers using tool microscopes; that is, the deflection yoke and ring magnets are adjusted manually. There are four adjustment items: colour balancing, picture inclination, beam convergency, and colour purity. These items are also interrelated, making manual adjustments tedious, slow, and unreliable. To solve this problem, an automatic CDT adjustment system has been developed. The system configuration is shown in Fig. 2.9.9. Using CCD cameras, standard patterns on the CDT display are observed and the image data are sent to an image memory. After calculation of the beam landing error, calculated parameters are sent to the feedback system to adjust the position of the deflection yoke and ring magnet. For beam convergence, the landing error for each colour (red, green, blue) is calculated individually.

Another example of micrometre-order adjustments is found in automatic positioning systems for VCR magnetic heads on cylinder units (see Section 7.7). As an example of nanometre-order applications, an image processing algorithm is installed in a high-precision interferometer to measure the surface roughness or flatness of optical mirrors or wafers^(4,5). The phase shift of light is detected by the fringe scanning method. Image processing technology is also used in STM (scanning tunnelling microscopy) or AFM (atomic force microscopy) to obtain smooth and noise-free object images.

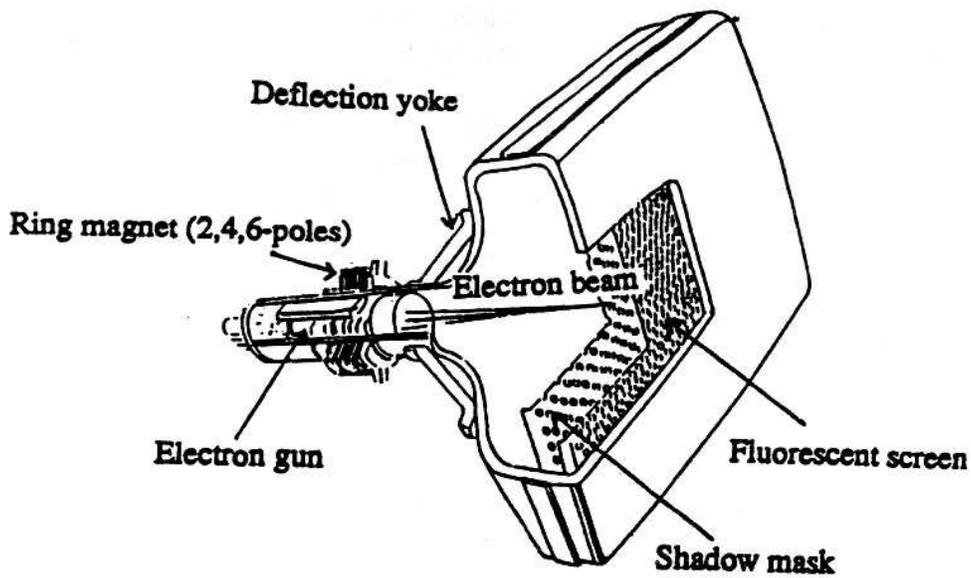


Fig. 2.9.8. Construction of colour display tube (CDT).

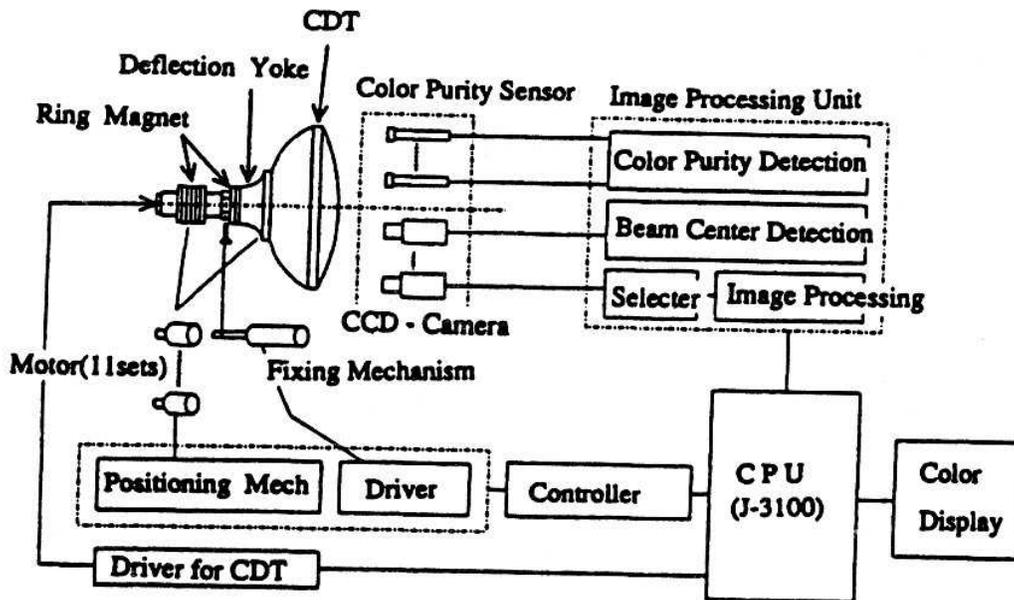


Fig. 2.9.9. Configuration of CDT automatic adjustment system.

2.9.5 Automatic measurement, adjustment, and inspection by pattern recognition technology

In general, the automation of skilled manual operations is very difficult to achieve. However, once it is done, quality and productivity are higher than on manual lines. Although pattern recognition technology can be a very powerful tool in the automation of these skilled operations, there are some key points when this technology is applied to automatic measurement, automatic adjustment, or automatic inspection.

(a) Automatic measurement by pattern recognition technology

If the definition for measurement is clear, automatic measurement using pattern recognition is comparatively easy. For example, to automate distance measurements between points A and B, these points must be clearly defined.

(b) Automatic adjustment by pattern recognition

Automatic adjustment can be broken down into the adjustment reference, adjustment mechanism, and the algorithm used. For the adjustment reference, selection of the appropriate reference as well as the method of detecting work-reference differences is important. The adjustment mechanism must be adaptable and flexible, while the adjustment algorithm must possess convergency and robustness.

(c) Automatic inspection by pattern recognition

There is great demand in industry today to automate visual inspection using pattern recognition technology. However, there are many obstacles in realizing this goal. It is often said that inspection does not create added value in products. Obviously if the production process is perfect, then inspection is not required at all. However, processes are never perfect in reality, and accuracies required have become extremely high. Hence an optimum investment cost for inspection must be sought within the total production process.

Another problem concerns real defects and dummy (pseudo-) defects. If the inspection machine is unable to find some real defect, the tendency is to develop more powerful algorithms with higher resolutions. Yet this will result in an increase in dummy defects. Thus it is sometimes more effective to develop a new algorithm which suppresses dummy defects, rather than to develop an algorithm with higher resolution.

A third problem is how to automate visual inspections using limit samples. Limit samples are actual product samples with defects that are at the acceptable limit, and provide a definition of a

defect for operators. Unfortunately, limit samples cannot be interpreted by computers directly, so a way must be found to define the defects as numerical data.

In general, defects occur in diverse modes. So, even if an algorithm can inspect a certain defect, to eliminate all defects would require an infinite number of algorithms. In order to achieve a high overall production quality, therefore, rather than concentrating on counting the number of defective products, the distribution of the on-line products must be examined and analysed to establish an optimum quality control production line.

References

1. Fujita, N. (1988). Assembly of blocks by autonomous assembly robot with intelligence (ARI). *Annals of CIRP*, 37, 33-5.
 2. Ejiri, M., Uno, T., Mese, M, and Ikeda, S. (1973). A process for detecting defects in complicated patterns. *Computer Graphics and Image Processing*, 2, 326-39.
 3. Komatsu, T. (1987). An automatic inspection system for chip electronic parts on a printed circuit board. *Annals of CIRP*, 36, 399-402.
 4. WYKO Corp. (n.d.). WKYO 400D catalogue. USA.
- ZYGO Corp. (n.d.). ZYGO MAXIM • 3D catalogue. US

Module-III

Nano-Positioning System of Nanometre Accuracy and Repeatability: Guide systems for moving elements, Servo control systems for tool positioning, Computer aided digital ultra precision position control, Future development of micro actuators.

3.1 Guide systems for moving elements: tool rest and workpiece

For positioning of moving elements at nanometre accuracy, it is necessary to provide guide systems with sub-nanometre scattering error for the transfer mechanism.

3.1.1 Elastic hinge or spring guide

This system is composed of an elastic element lying between two solid moving elements. The guiding action is performed by deformation of the elastic element, which is called an elastic hinge or spring guide. One of the most accurate single-block spring guides for nanometre accuracy is shown in Fig. 3.1.1, used for the STM (scanning tunnelling microscope).

In this system, the friction accompanying relative transfer motion of the moving elements is removed, except for the internal friction of the elastic material. Also there is no backlash in the mechanism joining the elastic element and the two moving elements. As a result, elastic hinge systems are most favourable for guiding two moving elements at nanometre accuracy. Of course, wide-range transfer cannot be realized by this system.

Actual examples of spring guides for linear, torsional, and combined motion are shown in Fig. 3.1.2.

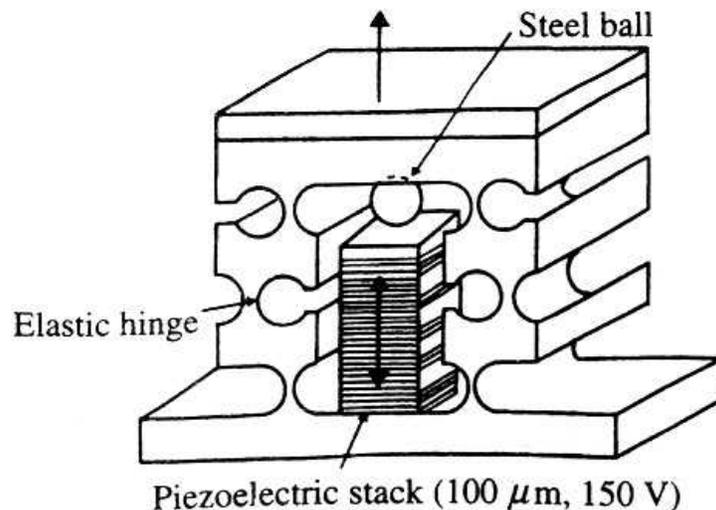


Fig. 3.1.1. Single-block elastic hinge or linear guide.

3.1.2 Linear slide guide and linear roller bearing

These systems consist of lubricants or rolling elements applied between two moving elements

(a) Linear slide guide

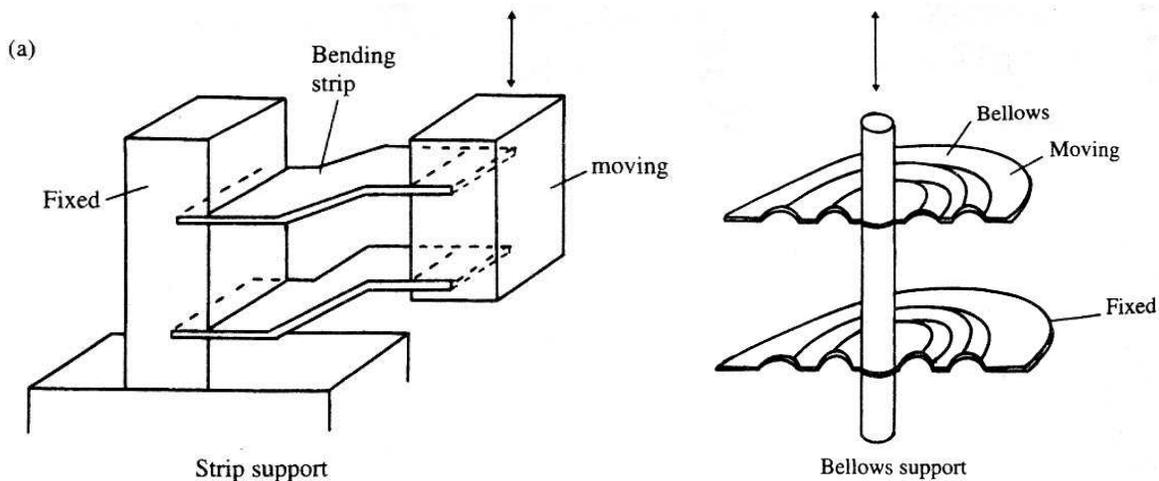
The system is basically sliding of two solid surfaces in contact, but there is always stick-and-slip motion due to friction arising from contact bonding between the two countering surfaces, causing erratic motion between the two elements.

To avoid friction effects, lubricants are used in general, but in microfine motion, boundary lubrication is always used. Hydrodynamic lubrication cannot be realized, because the relative speed of the two moving elements is very low. Therefore, in this kind of very slow relative motion under considerable pressure, a lubricant of high ‘oiliness’ should be used, oiliness being defined as the adhesive force between metal and oil due to van der Waals molecular bonding forces. Accordingly, in the relative motion of metal on metal, a lubricant of high oiliness performs a very important role: it always adheres firmly to the metal surfaces, but relative motion is achieved by shear slip in the lubricant with its relatively low resistive force. In practice, the lubricants of high oiliness used can be castor oil (a vegetable oil) and animal lard or grease, but in general a mixture of mineral and synthetic oils is used, because these vegetable and animal oils are easily oxidized. The lubricants used in watches and other instruments are all of this high-oiliness type.

Furthermore, the guide surface should be finished to sub-nanometre order because the oil may be degraded by a rough surface.

Solid powder lubricants as MoS_2 and graphite cannot be used for such guide systems, because shear slip of these powders is not very stable.

Slide guide systems of nanometre accuracy are effectively realized by lubrication with oil of high oiliness and guide surfaces of sub-nanometre roughness. Of course, the geometry of the plane and cylindrical guide parts should be maintained to nanometre accuracy.



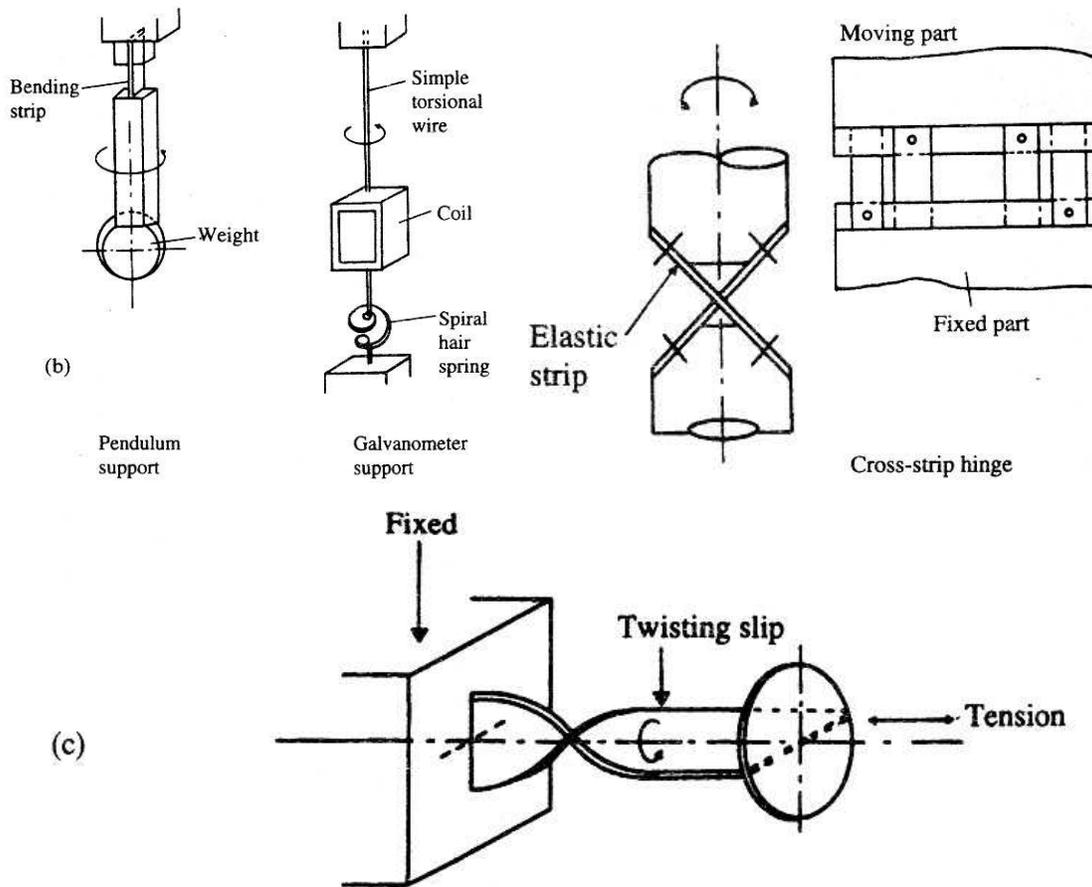


Fig. 3.1.2. Elastic hinges: (a) linear; (b) torsional; (c) combined (tensional and torsional).

(b) Linear roller bearing

The relative motion of moving elements through rolling elements such as balls, rollers and needles is effectively achieved in many accurate mechanisms. The material of the rolling elements may be hardened steel, but recently ceramic rollers of alumina and silicon nitride have been widely used to avoid direct adhesive wear due to direct contact between rolling elements

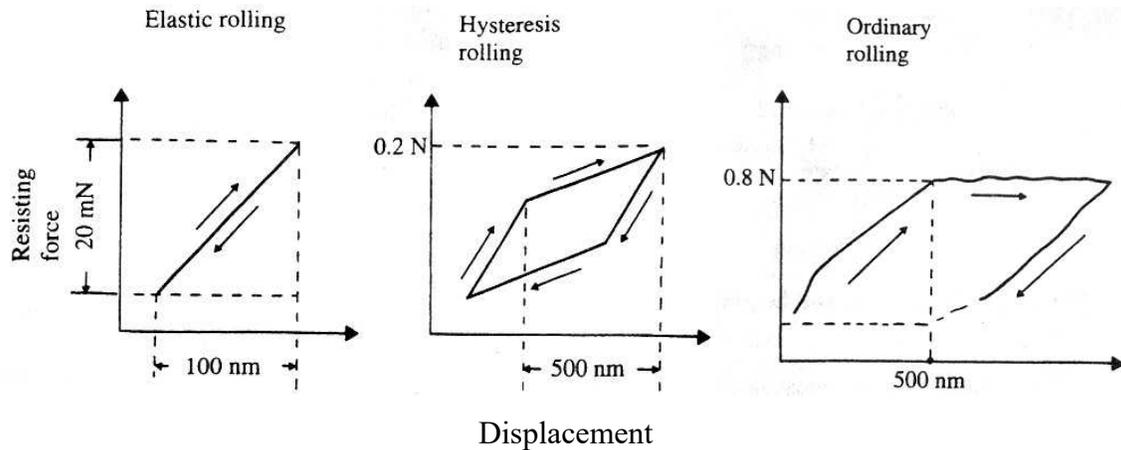


Fig. 3.1.3. Force-displacement relations in micro-rolling motion.

and steel guide races. However, in order to realize motion of nanometre accuracy, it is necessary to maintain the geometrical accuracy of rolling elements: the maximum eccentricities and diametral accuracies of the balls and rollers must be 10 nm and 50 nm respectively. Moreover, prestressing should be always applied between the rolling elements and guide races, to remove backlash between the two components. Of course, in the prestressed condition, the rolling resistance and also friction increase, but the accuracy of relative linear motion can be maintained by moving elements of suitable construction for prestressing.

On the microfine motion of rolling contact, experiments reported by the Yoshida Nanomechanism Organization reveal that, as shown in Fig. 3.1.3, if the rolling displacement is < 100 nm, a linear relation exists between the transfer resistance and the displacement, but in the case of displacement between 100 and 400 nm, hysteresis occurs, whereas above 500 nm, ordinary rolling friction takes place.

These phenomena can be understood as follows. On the basis of the Hertz analysis, under perfect elastic deformation, the rolling resistance becomes zero, apart from the inertial force. Accordingly, the linear relation at < 100 nm may represent the difference between the resistances of the loaded and restoring sides of the roller, based on internal friction of elastic materials. Therefore, even at < 100 nm very slight hysteresis should be expected. At larger displacements, microslip and also internal friction losses are always to be expected. However, it is confirmed that in the region < 100 nm, a linear relation is maintained between transfer resistance and displacement.

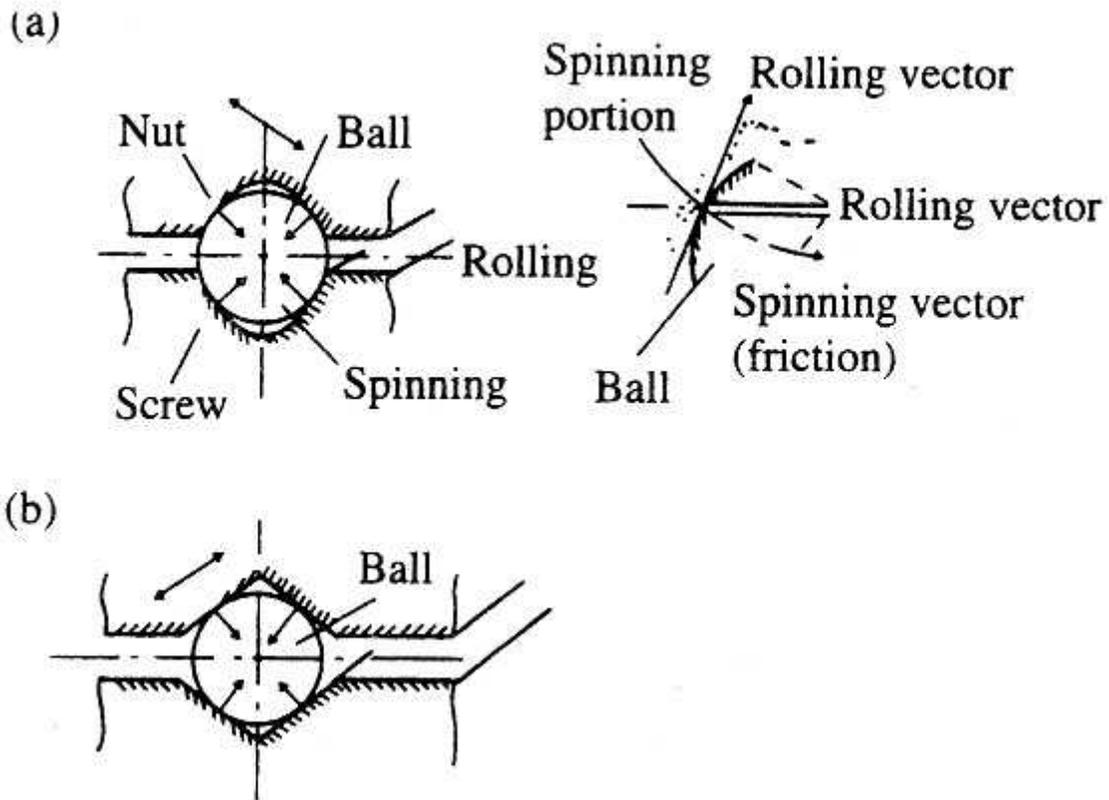


Fig. 3.1.4. Guide mechanisms: (a) ball and screw groove; (b) linear V-groove.

The ball screw V groove guide mechanisms shown in Fig. 3.1.4 are of great use, but there is always a relatively large spinning friction between ball and guide race. Therefore they always show stick-and-slip phenomena due to driving force and transfer, and also a relatively large backlash due to erratic geometrical form. As a result, these devices cannot be used for nanometre-accuracy mechanisms. However, if the geometrical forms of the ball and races are made to higher accuracy, nanometre-accuracy transfer with these systems will be possible using lubricants of high oiliness.

Another mechanism, using a knife edge with sharp edges interfaced between two moving elements, is shown in Fig. 3.1.5. This mechanism, with a hardened knife edge and agate bearing, has been developed as a very accurate supporting system for microbalances. This kind of knife edge interface allows very fine linear transfer of moving elements with a smaller range.

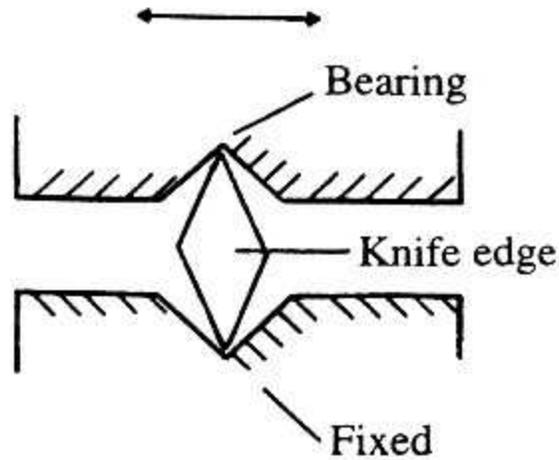


Fig. 3.1.5. Knife-edge and bearing guide.

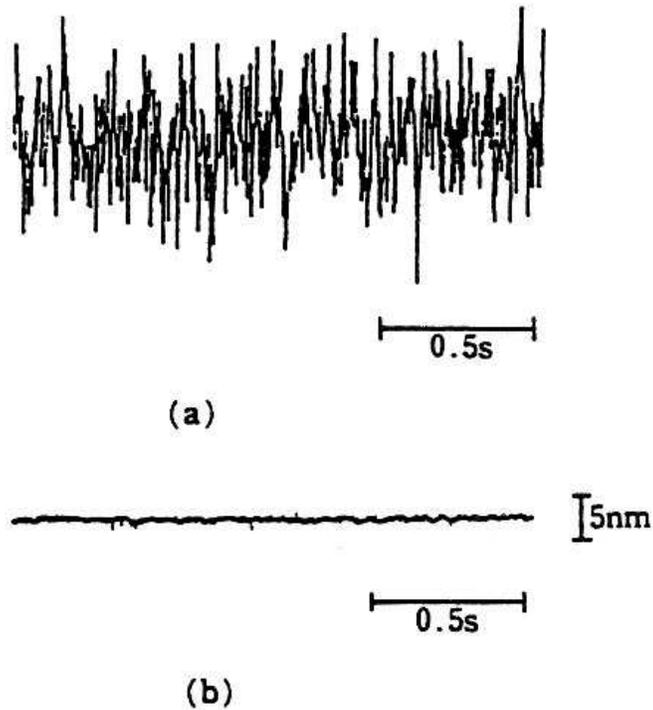


Fig. 3.1.6. Vibration of tables supported on (a) aerostatic and (b) hydrostatic slides.

3.1.3 Hydrostatic slide guide and journal bearing

(a) Hydrostatic slide guide

Hydrostatic slides have been successfully applied to ultra-precision machine tools because of some attractive characteristics, such as high accuracy, high stiffness and damping, low friction torque, and absence of wear.

Figure 3.1.6 demonstrates the excellent damping effect of a hydrostatic slide in comparison with an aerostatic slide⁽¹⁾. A table supported on an aerostatic slide vibrates with ~ 20 nm amplitude even in the stationary condition, whereas the magnitude of vibration of the table supported on a hydrostatic slide is very small and equivalent to that of a noise signal of the displacement measuring equipment. This excellent damping effect is one reason why hydrostatic slides have often been used in ultra-precision machine tools.

In the design of hydrostatic slides having the many features mentioned above, various configurations can be considered. Figure 3.1.7 shows typical configurations of a hydrostatic slide-table system for ultra precision machine tools. There is usually only one degree of freedom of motion in this kind of table, and the other five degrees of freedom of motion have to be strictly constrained. Opposed-pad hydrostatic slides are therefore used to reduce the displacements in the other five directions as much as possible. In addition, the table or the guideway may be elastically deformed by the high pressure applied to hydrostatic slides, but the configuration shown in Fig. 3.1.7(b) prevents the table from deforming by balancing the upper and lower forces acting on it.

In the configurations of hydrostatic slides shown in Fig. 3.1.7(a) and (b), pressurized liquid has to be fed to the table through pipes, and these pipes can sometimes disturb the precise motion of the table. In order to avoid this effect, the hydrostatic slide shown in Fig. 3.1.8 has been proposed^(2,3). On the surface of this slide are a deep long groove for liquid feed and many T-shaped shallow grooves. This kind of slide does not require a capillary or orifice restrictor to obtain bearing stiffness, and is called a groove compensation bearing. In this case, liquid is fed to the table through the feed hole in the guideway, and pipes do not disturb the motion of the table.

In conventional hydrostatic slides using capillary or orifice restrictors, the liquid film thickness is varied according to changes in external load. However, in order to achieve a higher accuracy of motion of the table, a higher stiffness is required for the supporting hydrostatic slide. Figures 3.1.9 and 3.1.10 show new types of hydrostatic slides which can achieve a very high stiffness using self-controlled restrictors. The self- controlled restrictor shown in Fig. 3.1.9(a)⁽⁴⁾ uses the elastic deformation of a diaphragm to control liquid flow entering the slide, while the one shown in Fig. 3.1.10(a)⁽⁵⁾ uses a floating disk and controls the liquid flow by using the force balance acting on the upper and lower sides of the disk. Figures 3.1.9(b) and 3.1.10(b) show the relations between liquid film thickness and applied load. It is clearly seen that the liquid film thicknesses

of these slides change very little even when the applied load is varied over a wide range, and that these slides can have a very high stiffness.

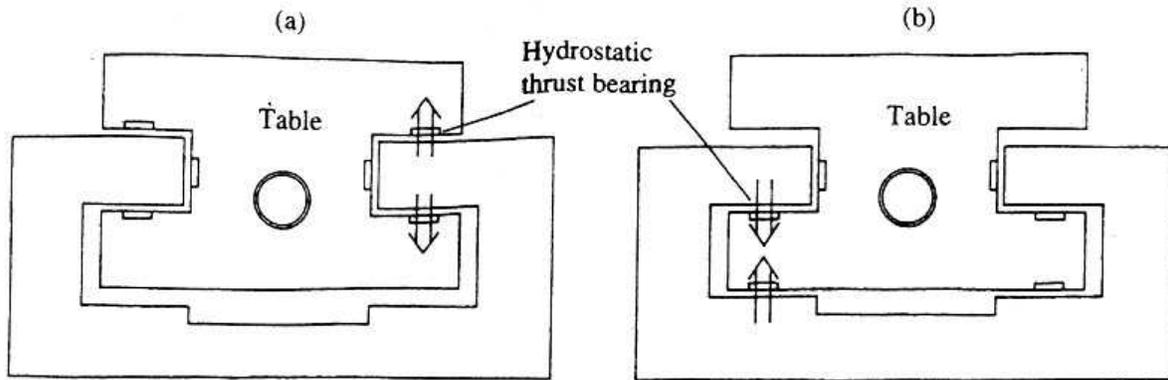


Fig. 3.1.7. Table-slide guide system, (a) Opposed-pad hydrostatic slide, (b) Arrangement of hydrostatic pad considering the elastic deformation of the table.

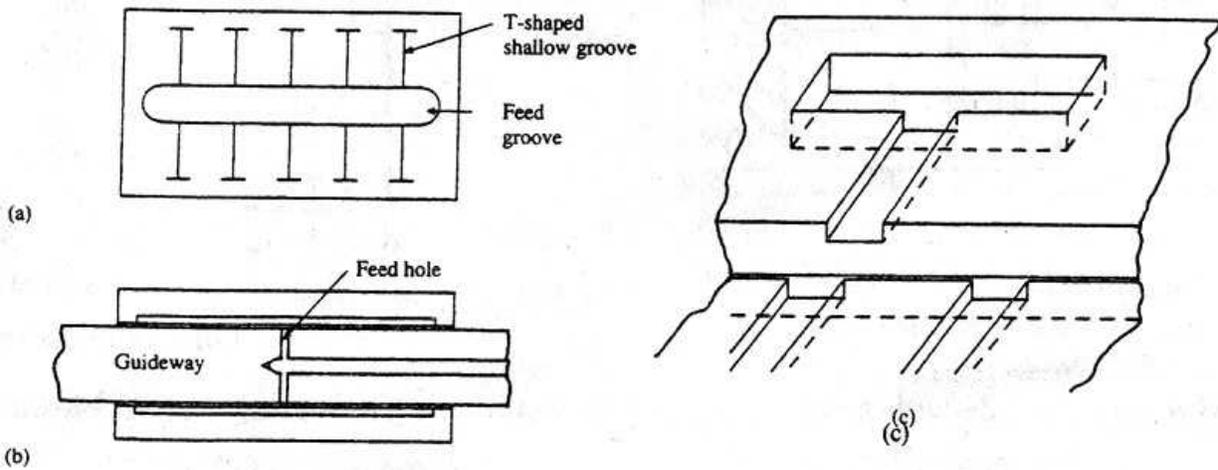


Fig. 3.1.8. Groove compensation hydrostatic slide, (a) Arrangement of grooves, (b) Guideway feed system, (c) Detail of T-shaped groove.

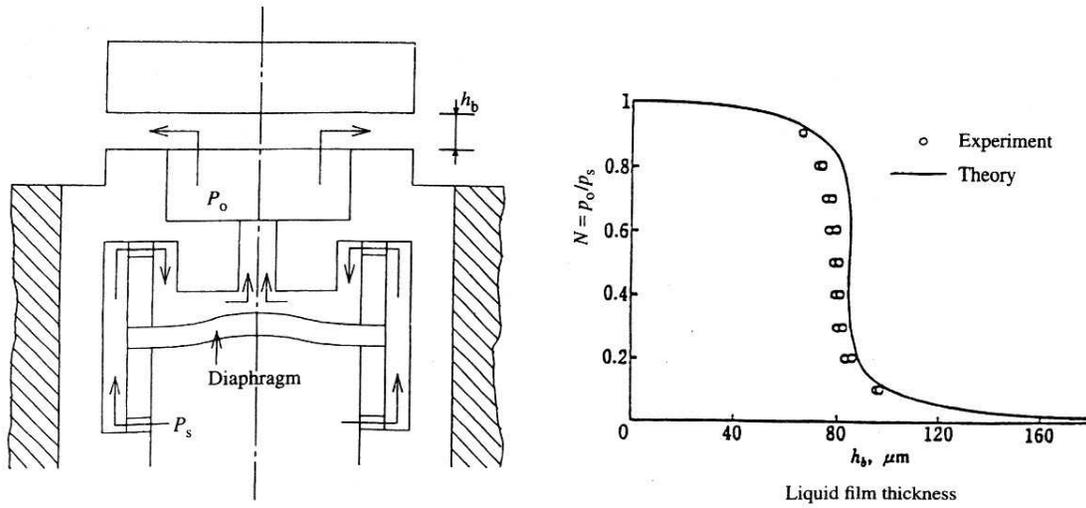
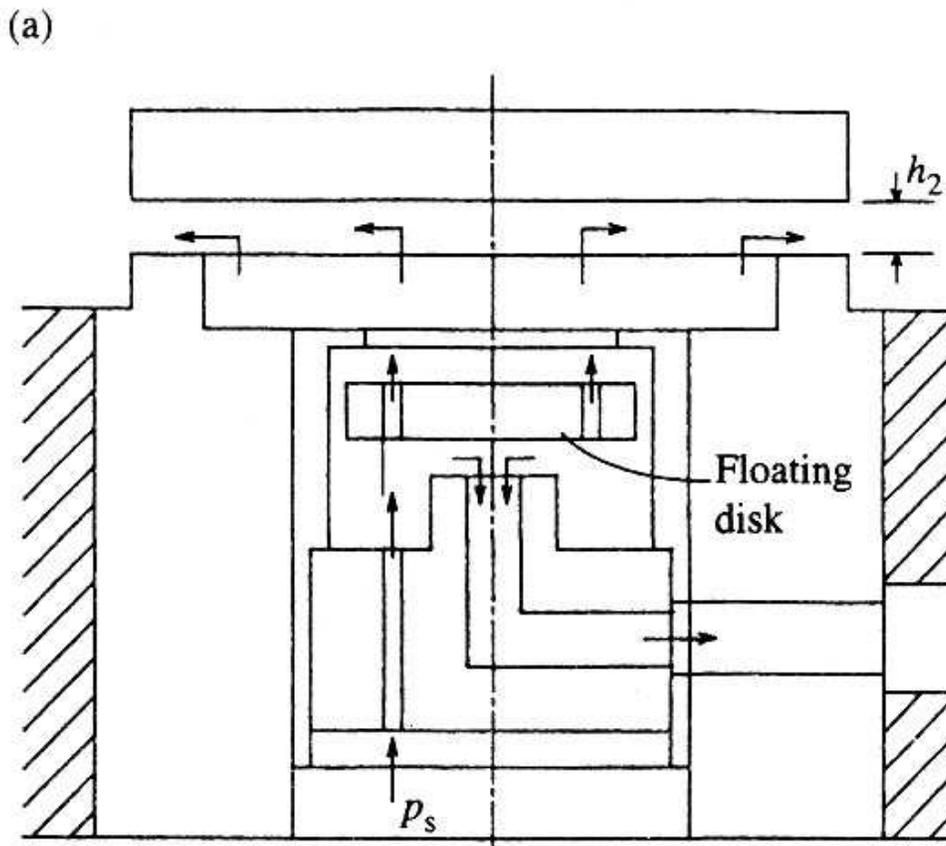


Fig. 3.1.9. (a) Hydrostatic slide with a diaphragm restrictor, (b) Static characteristic.



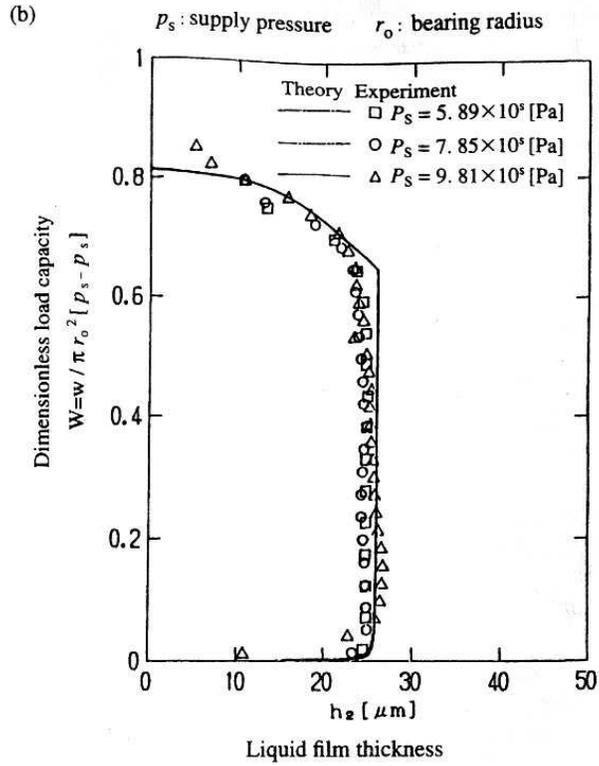
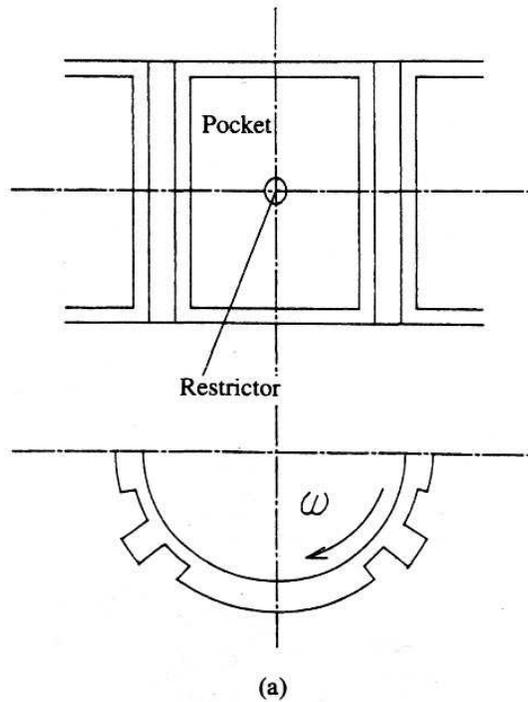


Fig. 3.1.10. (a) Hydrostatic slide with a self-controlled restrictor using a floating disk, (b) Static characteristic.



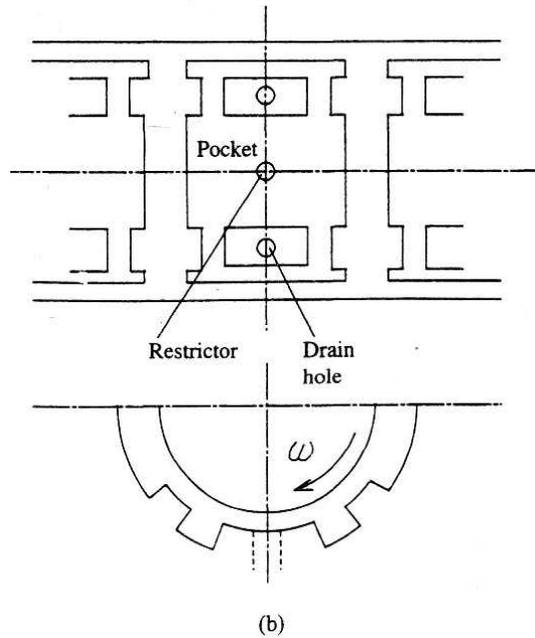
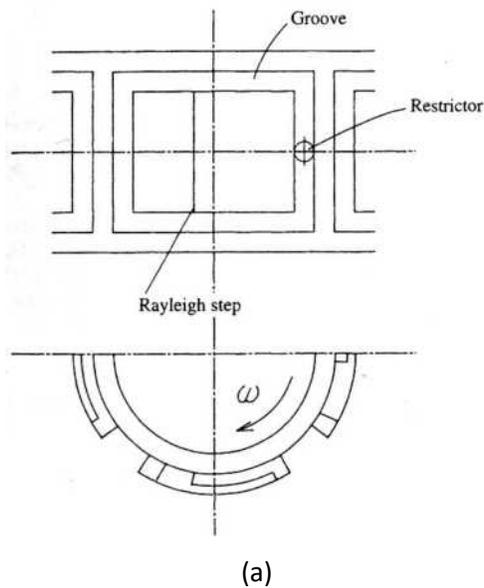


Fig. 3.1.11. Hydrostatic journal bearings: (a), (b), see text.

(b) Hydrostatic journal bearing

Like hydrostatic slides, hydrostatic journal bearings have various advantages for ultra-precision machine tools.

Figure 3.1.11 shows typical configurations of hydrostatic journal bearings. The bearing shown in Fig. 3.1.11(a) is an ordinary bearing which has several pockets and restrictors inside the pockets. The bearing shown in Fig. 3.1.11(b)⁽⁶⁾ has drain holes inside the pockets, and a wider range of design conditions can be selected by changing the diameter of the drain holes.



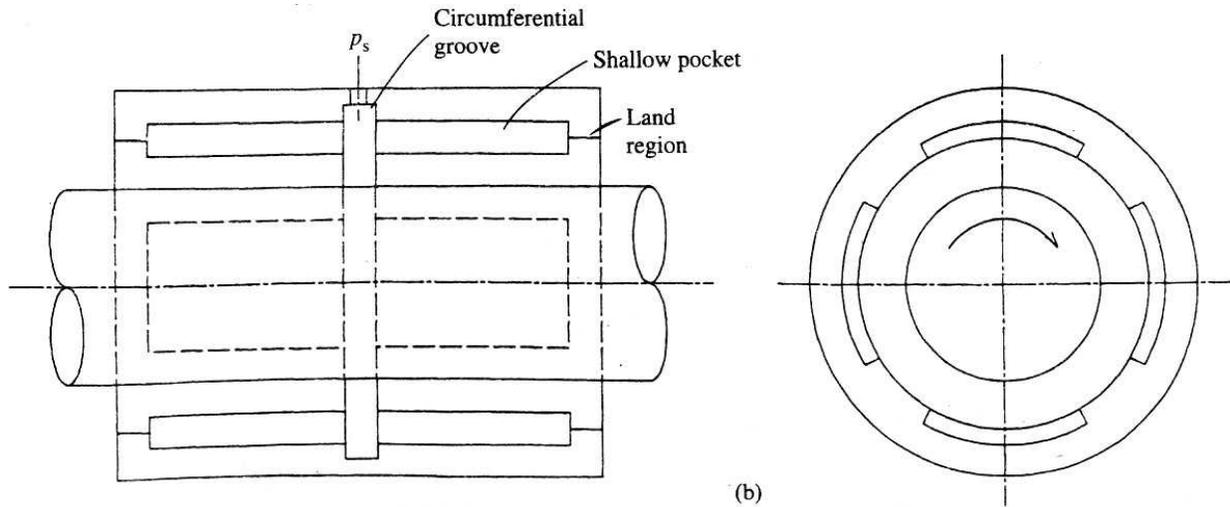


Fig. 3.1.12. Hybrid journal bearings with (a) Rayleigh step, (b) inherent compensation.

However, since in these bearings no hydrodynamic effect exerted by shaft rotation can be expected in the concentric condition, the stiffness of these bearings depends mainly on the supply pressure. A higher supply pressure is therefore required to obtain a higher stiffness.

As another means of obtaining a higher stiffness in hydrostatic bearings, the use of the hydrodynamic effect by shaft rotation is very effective. Figure 3.1.12(a) and (b) shows hydrostatic journal bearings using the hydrodynamic effect to increase stiffness; these are called hybrid bearings. The hybrid journal bearing in Fig. 3.1.12(a)⁽⁷⁾ has Rayleigh step pads inside the pockets to use the hydrodynamic effect. The one in Fig. 3.1.12(b)⁽⁸⁾ has no restrictors such as capillaries or orifices but has several shallow pockets whose depths are usually several times the bearing clearance. In this bearing, stiffness in the non-rotating condition is generated by using the difference in viscous resistance between the shallow pockets and the lands. Moreover a pair of these pockets and lands forms a Rayleigh step bearing in the rotating condition and produces hydrodynamic pressures.

Hybrid bearings can achieve a higher stiffness by using the hydrodynamic effect, but they also have some disadvantages such as higher power consumption and higher temperature rise under the rotating condition. To overcome these disadvantages in hybrid bearings, hydrostatic journal bearings with self- controlled restrictors have therefore been proposed. Since this type of bearing can achieve a very high stiffness (almost infinite) without using the hydrodynamic effect and a higher supply pressure, a low power consumption and low temperature rise can be achieved even in the rotating condition.

Figure 3.1.13⁽⁹⁾ shows an example of a hydrostatic journal bearing with a self-controlled restrictor. The narrow gap h_i between the floating bush and the inner bearing sleeve is varied according to changes in applied load and controls the liquid flow entering the bearing clearance. This bearing can thus achieve a very high stiffness by adjustment of the pressure P^*s . However, when the applied loads is dynamically changed, this bearing does not always achieve a high stiffness, and the dynamic characteristics should be taken into account.

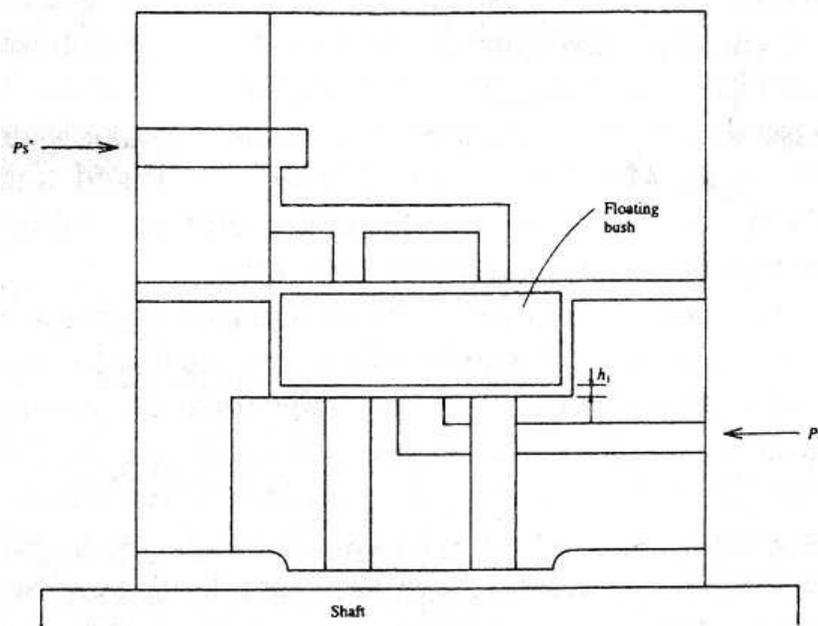


Fig. 3.1.13. Hydrostatic journal bearing with a self- controlled restrictor.

3.1.4 Aerostatic slide guide and journal bearing (spherical)

(a) Aerostatic slide guide

Like hydrostatic slides, aerostatic slides have various excellent features as ultra-precision machine elements. Especially when compared with hydrostatic slides, aerostatic slides are non-polluting because extremely clean pressurized air is used. They also have such features as less friction and lower temperature rise. However, they also possess some disadvantages such as a low damping coefficient and the need for higher manufacturing accuracy to obtain higher stiffness.

Figure 3.1.14 shows typical restrictors used in aerostatic slides. The supply pressure in aerostatic slides is usually set at < 1 MPa for safety reasons, so there is some limit to obtaining a higher stiffness by raising the supply pressure. To achieve higher stiffness, a very effective means is to

design the slide so that the maximum stiffness is obtained by making the slide clearance $< 10 \mu\text{m}$. Orifice and inherently compensated feed hole restrictors, shown in Fig. 3.1.14(a) and (b), are very simple and easy to manufacture, but they are not always able to achieve sufficient stiffness in real precision devices. This is because it is difficult in these restrictors to obtain the maximum stiffness in the $< 10 \mu\text{m}$ slide clearance. Other types of restrictor are therefore usually selected when a higher bearing stiffness is required. In particular, groove compensation, Fig. 3.1.14(e)⁽¹⁰⁾, and hybrid, Fig. 3.1.15, restrictors are often used in actual devices. A hybrid restrictor combines a groove compensation restrictor with a feed-hole restrictor, and various configurations have been proposed, as shown in the figure.

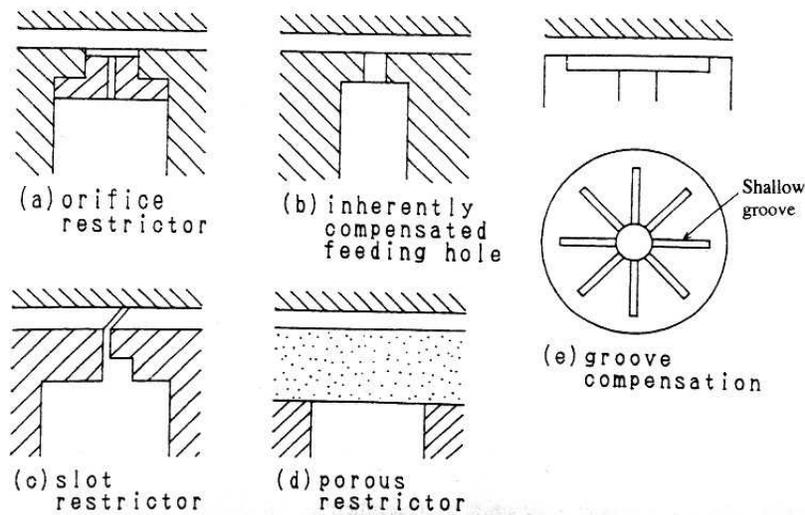


Fig. 3.1.14. Restrictors used in aerostatic slides.

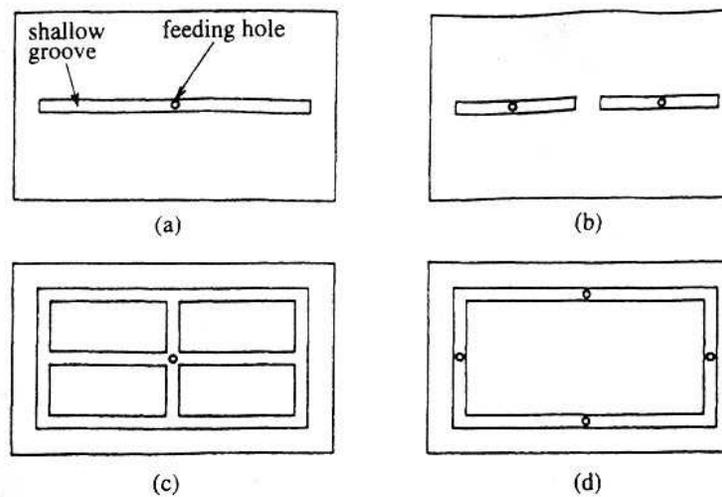


Fig. 3.1.15. Groove configurations of aerostatic slides with hybrid restrictors (a, b from ref. 11).

Figure 3.1.16 shows pressure distributions within the slides for different slide clearances. The variation of pressure distribution in hybrid restrictors is considerably larger than in groove compensation restrictors. Accordingly, slides with hybrid restrictors can achieve a higher bearing stiffness. However, since slides with this type of restrictor have unstable regions owing to pneumatic instability, as shown in Fig. 3.1.17, the design parameters should be carefully selected.

To obtain a very high bearing stiffness, aerostatic slides with self-controlled restrictors have been proposed. In fact, some types have almost infinite stiffness. Figure 3.1.18 shows three types of slide with diaphragm-type self-controlled restrictors. Slides (a)⁽¹²⁾ and (b)⁽¹³⁾ control the air pressure distribution within the slide clearance using elastic deformation of

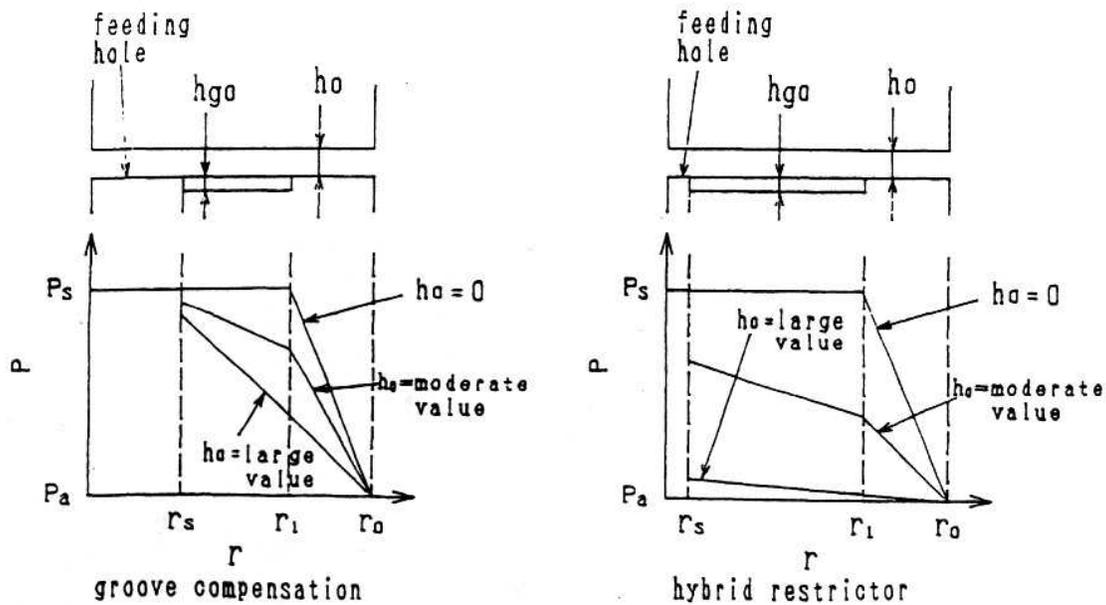


Fig. 3.1.16. Pressure distributions within slide clearance.

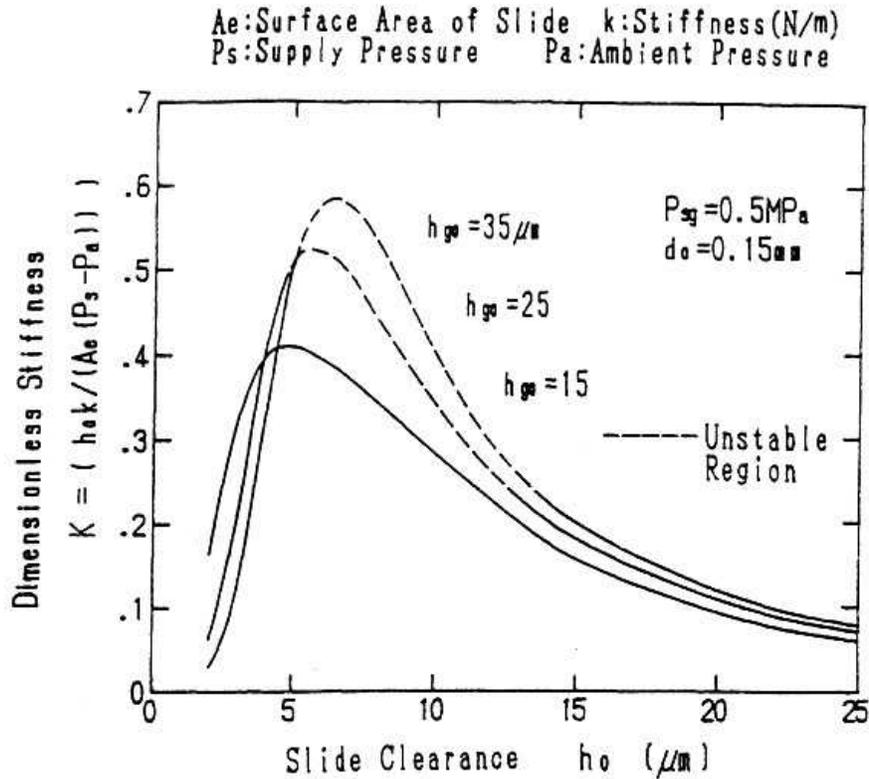


Fig. 3.1.17. Dimensionless stiffness of aerostatic slides with hybrid restrictors (type c in Fig. 3.1.15).

the diaphragm, although they cannot provide infinite stiffness. Slide (c)⁽¹⁴⁾ controls both the back-pressure of the diaphragm and the air pressure distribution within the slide clearance, so the top surface of this slide does not move even when the applied load is changed, achieving infinite stiffness.

Recently, aerostatic slides which can actively control the slide position have been proposed to improve the accuracy of table motion. These active slides are shown in Fig. 3.1.19. Slide (a)⁽¹⁵⁾ uses a piezoelectric actuator to control directly the air flow entering the slide clearance through the narrow gap above the

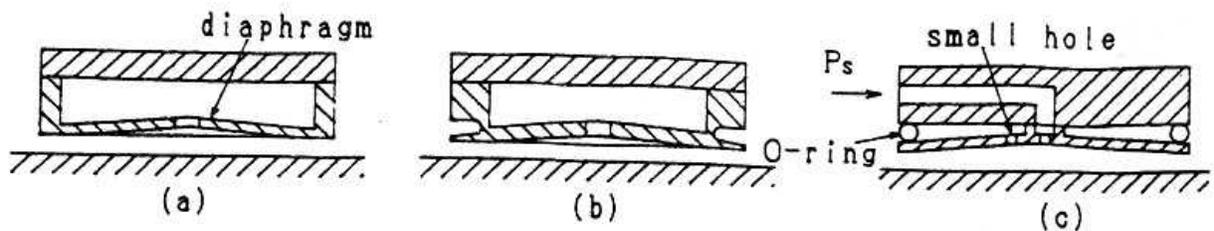


Fig. 3.1.18. Aerostatic slides with diaphragm restrictors.

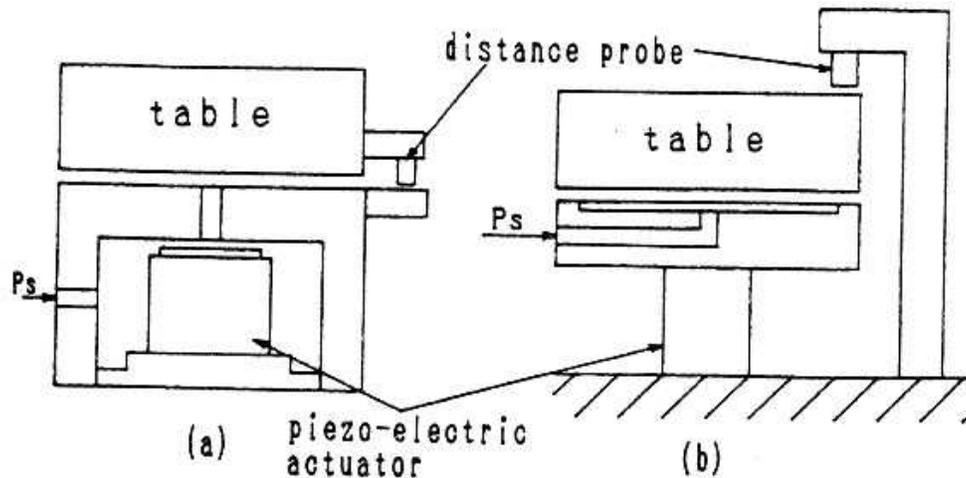


Fig. 3.1.19. Actively controlled aerostatic slides.

actuator. Design (b)⁽¹⁶⁾ uses a piezoelectric actuator to control directly the position of the slide. Therefore, when the applied load is varied in this slide, the slide clearance is adjusted to keep the table at the same position. Design (b) can control the table position much faster than (a) can.

(b) Aerostatic journal bearing and spherical bearing

Aerostatic journal bearings have been successfully applied to ultra-precision machine tools because of

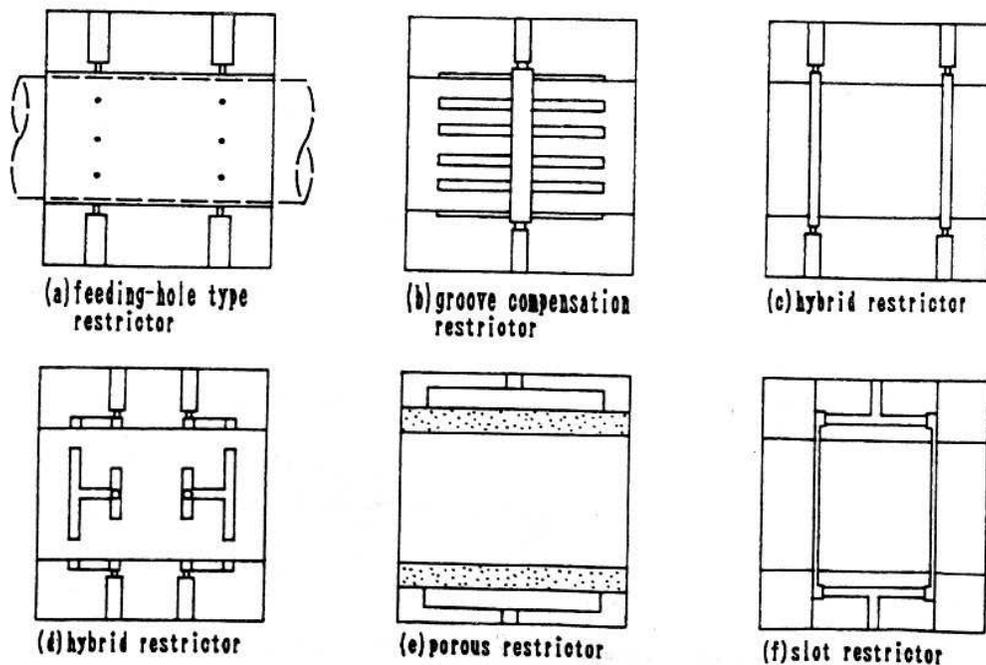


Fig. 3.1.20. Aerostatic journal bearings (b From ref. 17, c From ref. 18, d From ref. 19, f From refs 20, 21).

their low friction and high rotational accuracy of motion. Figure 3.1.20 shows aerostatic journal bearings which are often used in actual precision machine tools. Figure 3.1.20(a) shows a journal bearing with feed-hole type restrictors such as orifices and inherently compensated feed holes. Figure 3.1.21 compare the dimensionless stiffnesses that can be achieved in these bearings. It should be noted that the bearing stiffness largely depends on the value of the bearing clearance, since the dimensionless stiffness includes the clearance.

The rotational accuracy of a shaft supported by aerostatic bearings is greatly affected by the accuracy of shape of the shaft. Thus a spherical shaft manufactured to a very accurate shape will have a high rotational accuracy. Figure 3.1.22 shows a spherical-bearing shaft system⁽²²⁾. The rotational accuracy of this system is $< 0.05 \mu\text{m}$.

(c) Aerostatic lead screw

Aerostatic lead screws have the same kind of superior characteristics as aerostatic bearings: low friction, low torque fluctuation and high accuracy of motion. But since they are difficult to manufacture, they have rarely been applied to precision devices. Recently, aerostatic

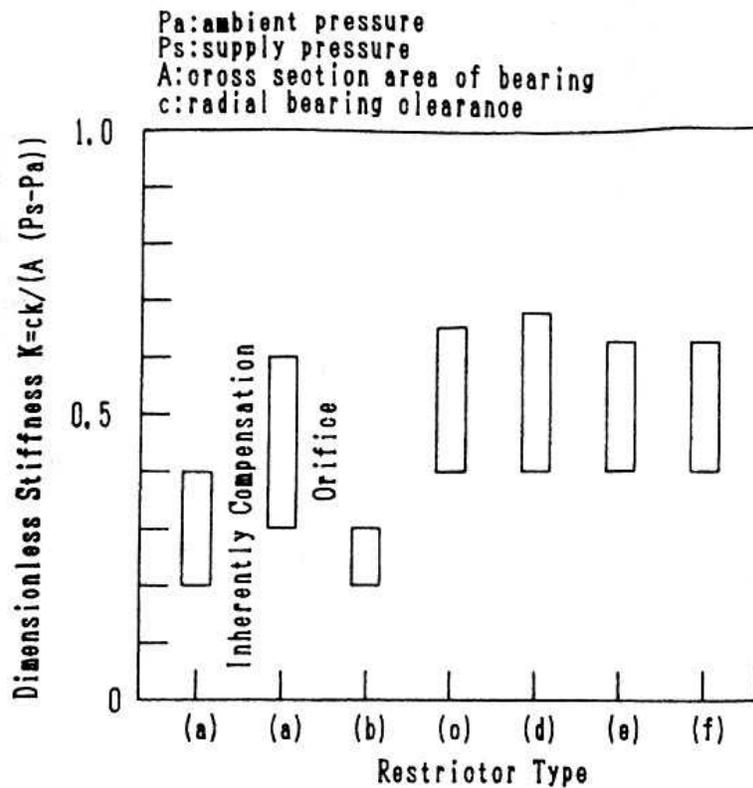


Fig. 3.1.21. Comparison of dimensionless stiffness of aerostatic journal bearings.

lead screws have been fabricated for application to certain precision devices. Figure 3.1.23 shows an aerostatic lead screw made of a porous ceramic material⁽²³⁾, planned for use in an X-ray stepper. This lead screw, which has an axial stiffness of $32 \text{ N}/\mu\text{m}^{-1}$, achieves a positioning accuracy of 4.5 nm. An aerostatic lead screw applied to a nanometre positioning system is described in reference⁽²⁴⁾.

3.1.5 Magnetic linear guide and rotational bearing

A guide mechanism using magnetic force working between two moving elements can be realized in passive or active form. Passive magnetic force appears

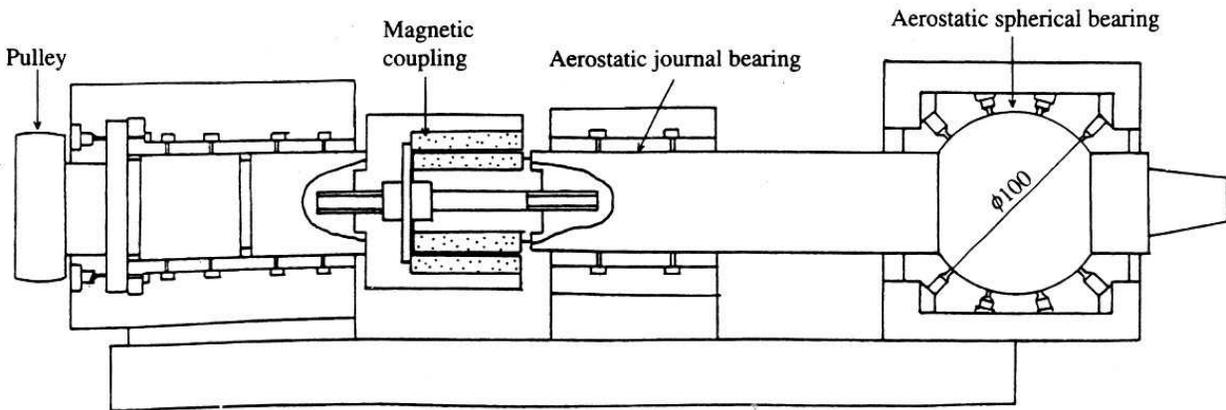


Fig. 3.1.22. Spherical-bearing shaft system.

in relative motion between magnetic material and a magnetic field. However, in relative motion of nanometre accuracy, the active system is applied, because the relative transfer range is very small.

Concrete examples are shown in Fig. 3.1.24 for rotational and linear guide systems. The systems are based on the magnetic attractive force due to an electro magnetic coil and ferrous material. In practice, as shown in the figure, two magnetizing coils and positioning sensor coils are provided to control the gap distance between the electromagnetic coil and moving elements. In these systems, the attractive forces between two opposed parts are balanced and the relative gaps can be maintained constant. However, these controlled gaps are always affected by mass and load variations. Accordingly it appears to be very difficult to maintain nanometre accuracy. Hybrid systems using ball and roller and magnetic systems as well as pneumatic bearing systems are used in scanners of A—D converters.

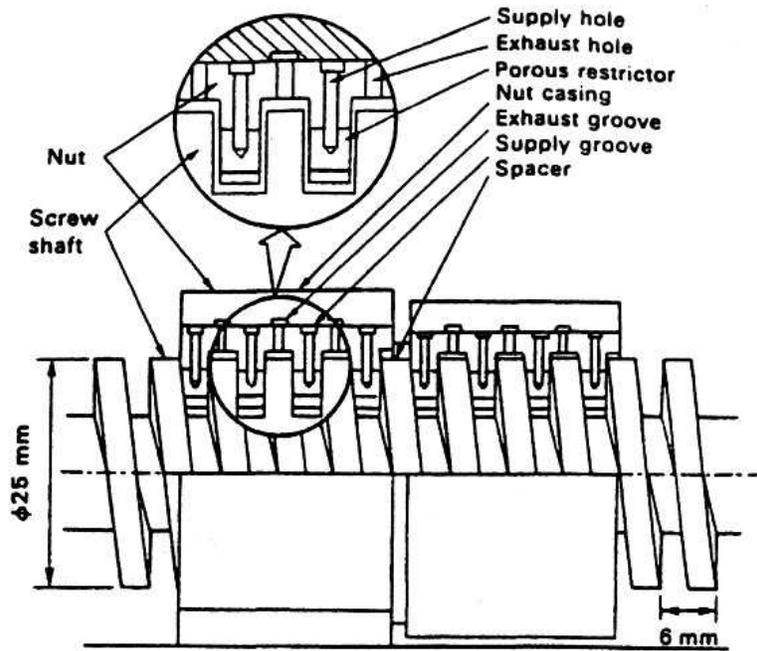


Fig. 3.1.23. Aerostatic lead screw

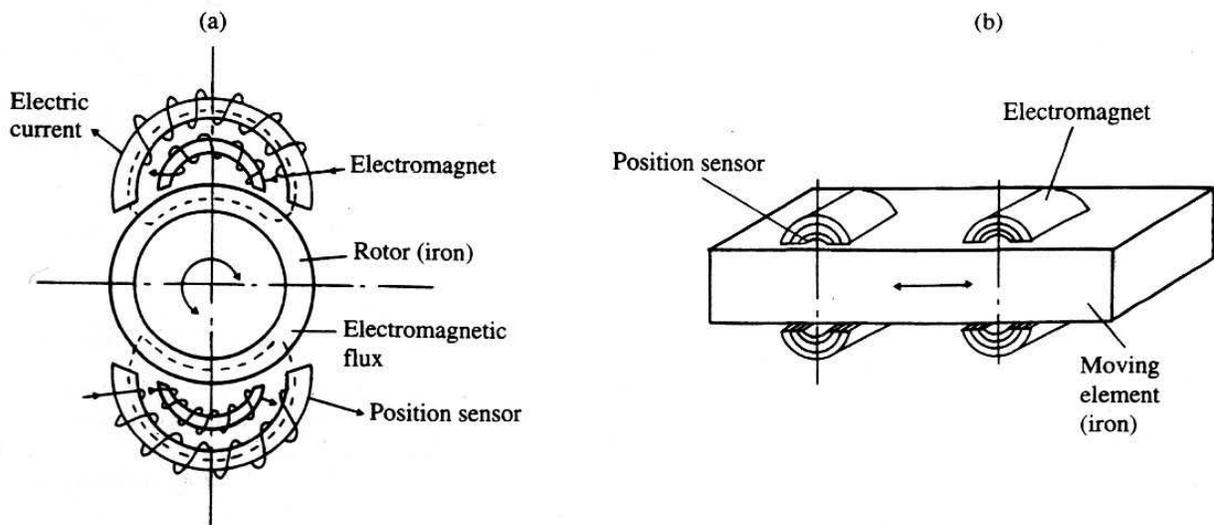


Fig. 3.1.24. Magnetic guides (active systems): (a) rotary; (b) linear.

References

1. Ikawa, N., and Shimada, S. (1986). Accuracy limitation in ultra-precision metal cutting. *Journal of the Japan Society of Precision Engineering*, 52, 2000.
2. Kogure, K., Kaneko, R., and Otani, K. (1982). Characteristics of T-shaped groove compensation bearing, *JSME Journal*, 48, 583.
3. Osaka, T., Unno, K., Tsubo, A., Maeda, Y., and Takeuchi, K. (1991). Development of high-precision aspheric grinding/turning machine. In *Progress in precision engineering*, pp. Springer-Verlag, Berlin.
4. Ono, K. and Hirose, S. (1985). Static characteristics of self-controlled externally pressurized thrust bearing with diaphragm restrictor, *JSLE*, 30, 652.
5. Yoshimoto, S., Anno, Y., and Fujimura, M. (1993). Static characteristics of a rectangular hydrostatic thrust bearing with a self-controlled restrictor employing a floating disk. *Transactions of the ASME, Journal of Tribology*, 115, 307.
- Sugita, K. (1984). Ultra-precision bearing. *Journal of Japan Society of Precision Engineering*, 50, 796.
5. Harada, M., Suda, M., and Miyaji, R. (1985). The improvement of load capacity of hydrostatic journal bearing using a pad added in the recess (2nd report). *Journal of the Japan Society of Precision Engineering*, 51, 841.
6. Qian, D.Z., and Yan, J.F. (1983). 'DYNASTAT' bearing-a new hydrodynamic and hydrostatic hybrid bearing. In 24th MTDR, 361
7. Mizumoto, H., Kubo, M., Yoshimochi, S., Okamura, S., and Matsubara, T. (1987). A hydrostatically-controlled restrictor for infinite stiffness hydrostatic journal bearing. *Bulletin of the Japan Society of Precision Engineering*, 21, 49.
8. Yabe, H., Mori, H., Asakawa, H., and Kihara, M. (1979). Characteristics of surface restriction compensation aerostatic thrust bearings, *JSME Journal*, 45, 100-7.
9. Boffey, D.A., Waddell, M., and Deardent, J.K. (1985). A theoretical and experimental study into the steady-state performance characteristics of industrial air lubricated thrust bearings. *Tribology International*, 18, 229-33.
1. 12. Hayashi, K. (1981). Investigation on externally pressurized gas-lubricated circular thrust bearing with flexible surface. In 8th Gas Bearing Symposium, paper 1. ,

10. Holster, P. and Jacobs, J. (1986). Theoretical analysis and experimental verification on the static properties of externally pressurized air-bearing pads with load compensation, *Tribology International*, 20, 276.
11. Blondeel, E. and Snoeys, R. (1976). Externally- pressurized bearings with variable gap geometries. In *Gas Bearing Symposium*, paper E2.
12. Hongo, T., Harada, M., and Miyaji, R. (1987). The static characteristics of externally pressurized gas thrust bearings with an electrically controlled restrictor. *Journal of the Japan Society of Lubrication Engineers*, 32, 894-9.
13. Watanabe, I., Aoyama, N., and Shimokobe, A. (1988). An active air bearing — ultra precision control of floating position and vibration. *Journal of the Japan Society of Precision Engineering*, 54, 329-34.
14. Yabe, H., Mori, H., Aoki, S., and Osame, M. (1985). Characteristics of surface restriction compensation aerostatic journal bearings. *JSME Journal*, 51, 3275-80.
15. Ono, K. (1984). Circumferentially grooved hydrostatic gas journal bearing. *Bulletin of the Japan Society of Mechanical Engineers*, 27, 95-101.
16. Fujikawa, Y., Yamazaki, S., and Suzuki, M. (1988). Development of high stiffness air bearing spindle, *NTN Technical Review*, 54, 65-74.
17. Dee, C.W. and Shires, G.L. (1971). The current state of the art of fluid bearings with discrete slot inlets. *Transactions of the ASME, Series F*, 93, 441.
18. Yoshimoto, S., Nakano, Y., and Kakubari, T. (1984). Static characteristics of externally pressurized gas journal bearings with circular slot restrictors. *Tribology International*, 17, 199-203.
19. Kobayashi, A., Hoshina, N., Tsukada, T., and Ueda, K., *Annals of the CIRP*, 27
20. Ishihara, S., Kanai, M., Une, A. and Suzuki, M. (1990). An X-ray stepper for SOR lithography *NTT Review*, 2, 92-100.
21. Sakano, T. et al. (1992). Development of nano servo system. *FAN AC Technical Review*, 5(2), 1-14.

3.2 Servo-control systems for tool-positioning of nanometre accuracy

3.2.1 Introduction

The concept of position servo-control systems is shown in Fig. 3.2.1. The system consists of a position command unit, a servo actuating unit, and a material processing unit within nanometre accuracy, and also in-process position and speed sensing devices with sub nanometre resolution for feedback control networks with fully closed and semi-closed loops.

In general, tool-positioning on hardware machines tools is carried out by the servo-control systems based on a software command signal: dynamic motion of hardware of heavy mass and large braking load is servo-controlled by software information on receipt of a position command signal. However, position-sensing on 3D measuring machines can now be done in the moving state using a touch sensor, so the positioning systems are quite different from rigid fixing of the tool position.

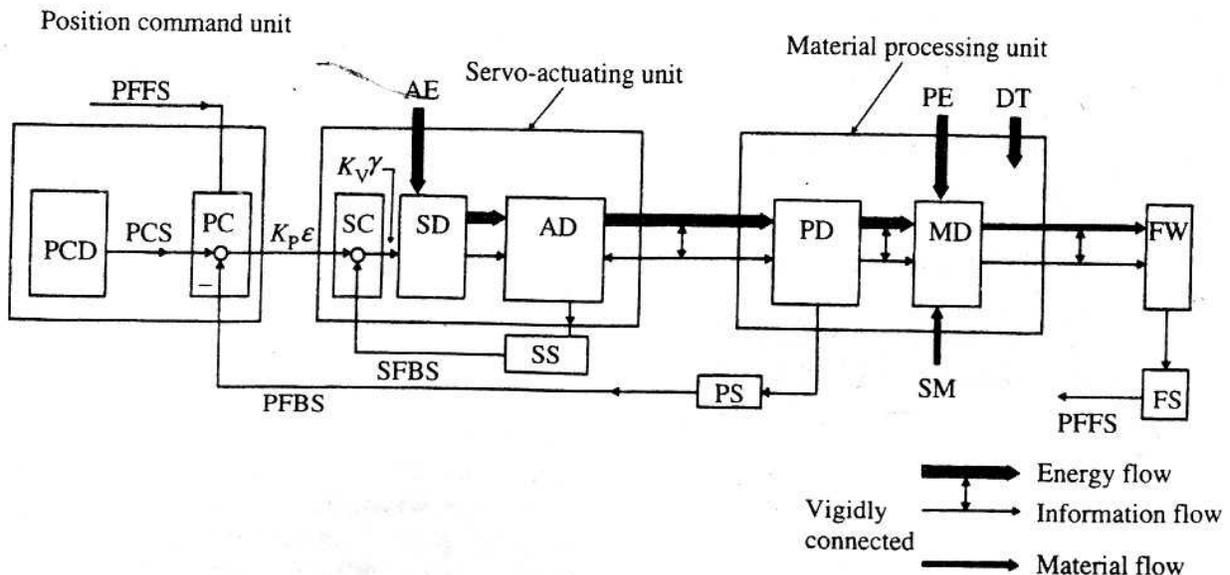


Fig. 3.2.1. Block diagram of servo-control system for tool-positioning. PCD: position command device PCS: position command signal PC: position compensator $k_p \epsilon$: positioning speed signal SC: positioning speed compensator ($0,7$: actuator speed signal SD: servo pilot device AD: actuator for tool-positioning PD: positioning device for tool MD: machining device FW: finished work SS: speed sensor of actuator, in situ PS: position sensor of tool or workpiece, in situ FS: form sensor of finished work, post-process AE: actuator driving energy PE: material-processing

energy DT: disturbance SM: stock material PFFS: position feedforward signal, post-process
PFBS: position feedback signal, in process (semi-closed loop) SFBS: speed feedback signal, in-
process (fully closed loop)

1. The position command unit is composed of a position command device PCD generating a position command signal, and position compensator generating an actuator positioning speed signal $k_p \epsilon$, proportional to the difference ϵ between the position command signal and the *in-situ* position of the tool.

2. The servo actuating unit is composed of a speed compensator SC, servo pilot device SD and actuator device AD. The speed compensator SC sends an actuator positioning speed signal $k_v \gamma$ to SD, where γ is the difference between the command actuator speed signal and the *in-situ* actuator speed, performing semi- closed loop control of actuator positioning speed. The servo pilot device SD provides the drive for the servo- actuator, corresponding to the actuator positioning speed signal $k_v \gamma$.

The elementary servo-control actuating unit may be one of the following types:

- (a) mechanical device using friction wheel and control roller systems
- (b) hydraulic device
- (c) pneumatic device
- (d) electric d.c. or a.c. servomotors for stepping and continuous moving systems (rotary and linear, and direct drive motors)
- (e) electromagnetic voice coil device using electric current and magnetic field
- (f) electrostatic motor using electrostatic attractive or repulsive force
- (g) piezoelectric stack actuator using electric potential control (see Table 3.2.1)
- (h) magnetostriction stack actuator using electric current control
- (i) ultrasonic driving motor using friction between travelling elliptic wave motion of an ultrasonically vibrating stack and moving elements
- (j) thermal expansion actuator using temperature control
- (k) shape memory alloy actuator using temperature control
- (l) anisotropic polymer actuator with piezoelectric ceramic stack.

Table 3.2.1 Comparison of actuator characteristics^a

	Amplitude	Accuracy	Working pressure/force	Response time
Pneumatic cylinder	100 mm	100 μm	0.1 N mm ⁻²	10 s
Hydraulic cylinder	1000 mm	10 μm	100 N mm ⁻²	1 s
Voice coil motor	1 mm	0.1 μm	300 N	1 ms
Piezoelectric actuator	0.1 mm	0.1 nm	30 N mm ⁻²	0.1 ms

^a From *Piezoelectric actuators*, by K. Uchino, Morikita, 1986.

Recently several kinds of actuators have been developed.

3. The material-processing unit comprises positioning device (PD) for the tool or workpiece and a machining or processing device (MD). The PD is rigidly connected to the actuator AD, and the information on tool position is fed back to the position compensator PD for closed-loop control.

Positioning errors occurring in the processing unit and servo actuating unit produce deviational and random errors in tool position. The deviational or systematic error can be compensated by this feedback closed-loop control system, but the random error, or scatter cannot be corrected by the control system and persists in the positioning of the tool. To obtain higher accuracy therefore, the random error of the processing system should be fixed in the processing unit itself. This kind of machine with an extremely limited nanometre range of random errors may be called a machine of nanometre accuracy.

The detailed construction and function of these units and devices are described in the following sections.

3.2.2 Mechanical servo-actuator positioning systems

(a) The most typical mechanical servo-actuator system is the friction wheel mechanism as shown in Fig. 3.2.2. The system consist of a servo pilot friction roller FR guided by a nut mating with a servo pilot screw SS, a driving plane friction wheel PW of constant angular speed, and a driven cylindrical friction wheel CW with a load of considerable inertial and braking torque.

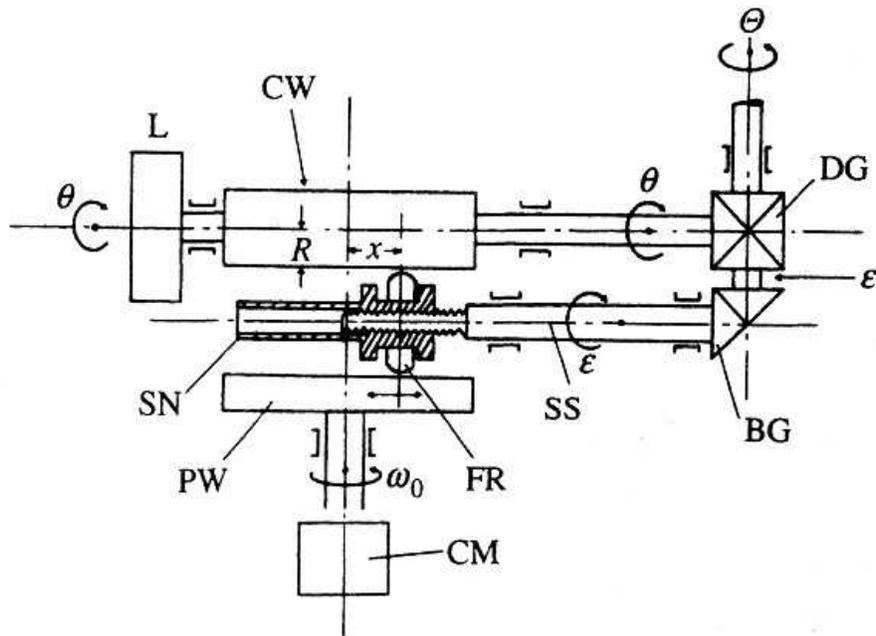
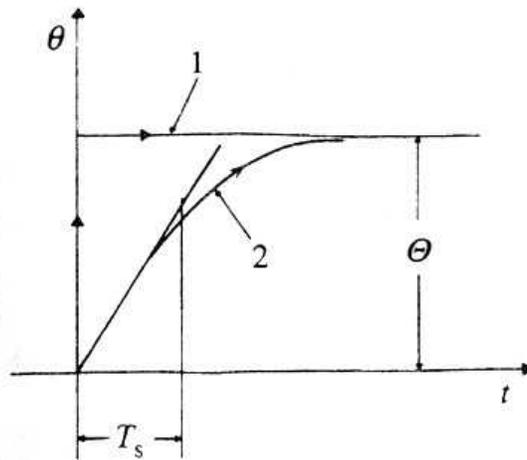


Fig. 3.2.2. Servo-actuator system using friction wheel, x : position of servo pilot roller (m) SN: servo pilot screw nut (non-rotating) Θ : input command angle (rad) θ : output angle of load (rad) DG: differential gearing ε : differential angle $\Theta - \theta$ (rad) BG: bevel gear, gear ratio 1:1 SS: servo pilot screw thread with lead p (m) FR: servo pilot friction roller PW: driving plane friction wheel of constant angular speed of ω_0 (rads^{-1}) CW: Driven cylindrical friction wheel of radius R (m) L: load with considerable inertia and braking torque CM: constant angular speed motor



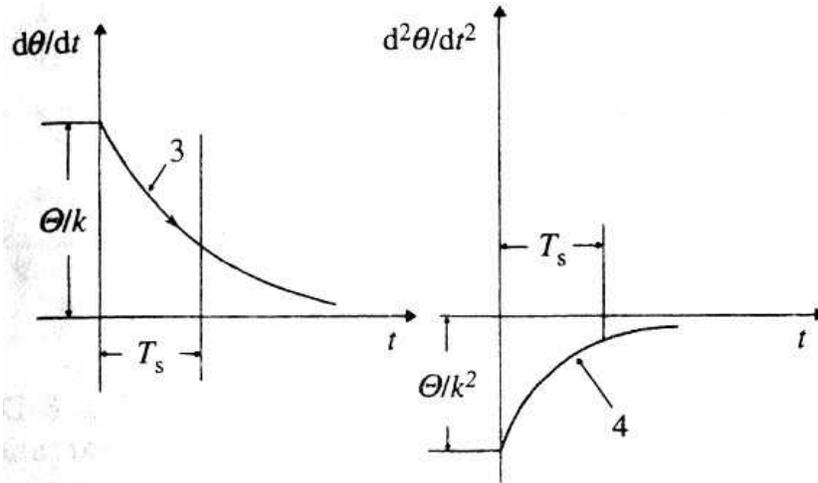


Fig. 3.2.3. Position of angular motion, change of angular speed, and change of angular acceleration. 1: Step input of angular position. 2: Controlled angular position. 3: Controlled angular speed. 4: Controlled angular acceleration

This servo actuating system using a friction wheel mechanism operates as follows. When the friction roller FR is located at a distance x from the centre of plane wheel PW, the cylindrical wheel CW is driven through FR by PW. The angular speed of CW is given by

$$d\theta / dt = (\omega_0 / R)x, \quad (3.2.1)$$

where θ is the drive angle of CW, ω_0 is the constant angular speed of PW, and R is the radius of CW. Of course, eqn (3.2.1) assumes that the friction forces acting at the contact points of CW, FR, and PW, under the no-slip condition, are sufficiently greater than the necessary transitional dynamic forces to accelerate the load L.

The position x of FR is set by the rotation angle ε of the servo pilot screw SS, as follows:

$$x = p\varepsilon / 2\pi, \quad (3.2.2)$$

where p is the length of screw SS. From eqns (3.2.1)

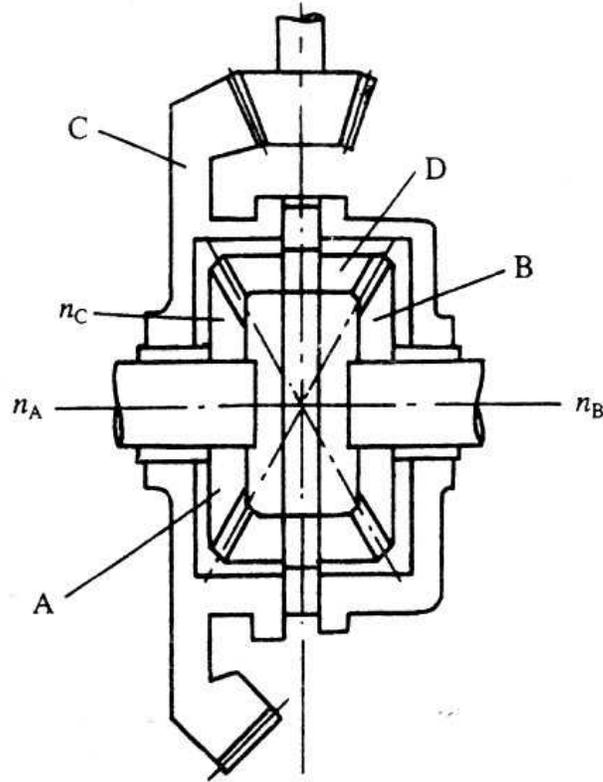


Fig. 3.2.4. Differential planetary bevel gearing. n_A : input rotation angle of gear A, Θ n_B : output rotation angle of gear B, θ n_C : differential rotation of gear C, $\varepsilon/2$, where $e = \Theta - \theta$ D: planetary bevel gear z_A : no. of teeth of gear A z_B : no. of teeth of gear B

and (3.2.2) the system gain for the control circuit, given by the ratio of angular speed to rotation angle of the servo pilot screw, is

$$\left(\frac{d\theta}{dt}\right) / \varepsilon = (\omega_0 / R) \cdot (p / 2\pi). \quad (3.2.3)$$

In the servo-actuating positioning system, ε is given by the compensator unit DG as follows:

$$\varepsilon = \Theta - \theta \quad (3.2.4)$$

where Θ is the input command angle and $\Theta - \theta$ is given by the differential-gear compensator device, details of which are given later.

Using eqns (3.2.1), (3.2.2), and (3.2.4) the dynamic kinematic equation of the system is

$$\theta + k \frac{d\theta}{dt} = \Theta \quad (3.2.5)$$

where $k = 2 / p\omega_0$ corresponds to the time constant T_s of the system. The solution of eqn (3.2.5) at the initial condition $\theta = 0$ or input command angle of step form Θ is

$$\theta = \Theta [1 - \exp(-t/k)] \quad (3.2.6)$$

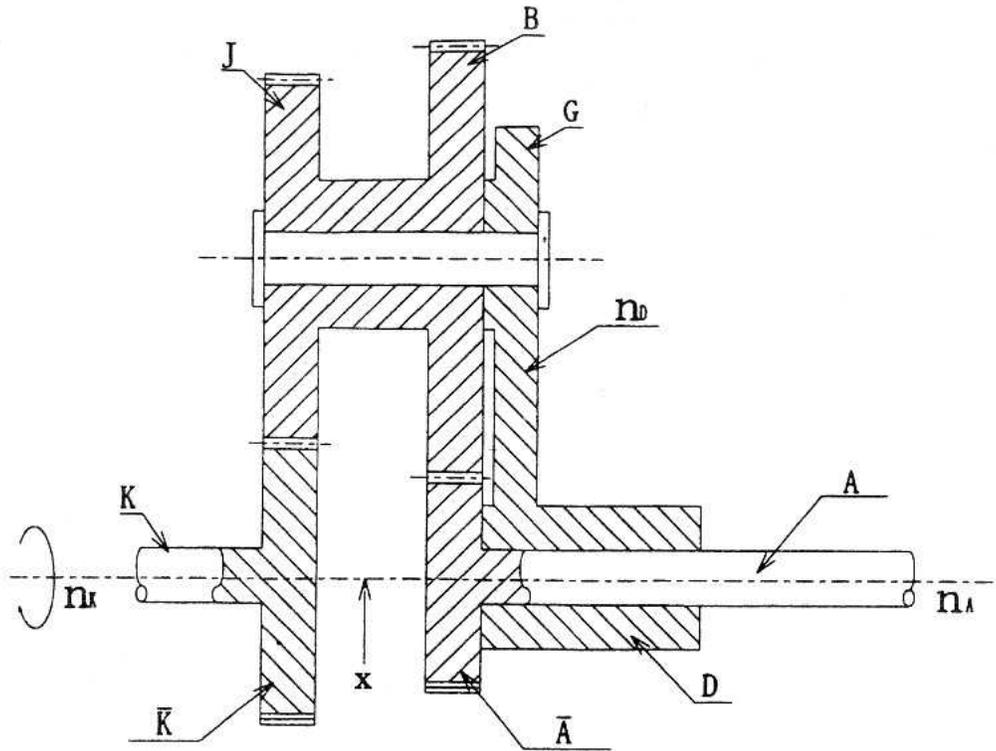


Fig. 3.2.5. Differential planetary plane gearing. A: main coaxial shaft with gear A K: main coaxial shaft with gear K D: arm shaft with arm G J, B: planetary gears z_A, z_K, z_J, z_B : no. of teeth of gears A, K, J, B i_1 : gear ratio z_K / z_J i_2 : gear ratio z_B / z_A x : spatial revolution axis n_K, n_A, n_D : revolution angle of shafts K, A, D; clockwise rotation with positive sign

Therefore if we can set the time constant k or T_s sufficiently small, then the control equation (3.2.6) tends to the following form in practical use:

$$\theta = \Theta \quad (3.2.7)$$

Accordingly, with this servo-control system, a load of high inertia and large braking torque can be moved quickly to the command position.

The transient performance of this system is shown graphically in Fig. 3.2.3. From eqn (3.2.6), the speed of control angle variation is given by

$$d\theta / dt = \Theta / k \cdot \exp(-t/k) \quad (3.2.8)$$

Therefore, at the starting point $t = 0$, the speed of the control angle becomes instantaneously Θ / k and decreases to zero. The acceleration of the control angle variation is

$$d^2\theta / dt^2 = -\Theta / k^2 \cdot \exp(-t / k) \quad (3.2.9)$$

Therefore, at the starting point, the acceleration becomes instantaneously $-\Theta / k^2$ and then tends to zero.

In practice a step input of angle Θ cannot be performed instantaneously, so the initial speed and acceleration of the system increase gradually as the Θ increases.

(b) The mechanisms of mechanical compensators using differential bevel and plane planetary gear are as follows.

The construction of the differential bevel planetary gear is shown in Fig. 3.2.4. The relations between input and output angles and the differential angle are obtained as follows. When gear B is fixed,

$$n_B = 0, \quad n_A / n_C = 1 - i_0, \quad (3.2.10)$$

and when gear A is fixed,

$$n_A = 0, \quad n_B / n_C = (i_0 - 1) / i_0, \quad (3.2.11)$$

where $i_0 = z_A / z_B$. Then the algebraic sum of eqns (3.2.10) and (3.2.11) becomes

$$n_C = (i_0 n_B - n_A) / (i_0 - 1) \quad (3.2.12)$$

Therefore, putting $i_0 = z_A / z_B = -1$, due to planetary gearing, then we obtain a difference between the rotation angles of A and B as a rotation angle of C:

$$n_C = (n_A - n_B) / 2 \quad (3.2.13)$$

The construction of the differential plane planetary gear is shown in Fig. 3.2.5. The relation between input and output angles and the differential angle are obtained by the following procedures.

With shaft A fixed:

Arm D, n_D Shaft K, n_K Shaft A, n_A
 1st Revolution 0 1 (clockwise) $i_2 \cdot i_1$
 2nd Revolution $-i_2 \cdot i_1$ $-i_2 \cdot i_1$ $-i_2 \cdot i_1$
 (anticlockwise around space axis X at 2nd revolution)
 Sum of 1st and 2nd $-i_2 \cdot i_1$ $1-i_2$ i_1 0

Therefore,

$$n_K / n_D = 1 - i_2 \cdot i_1 / -i_2 \cdot i_1 = (1 - k / -k), \quad (3.2.14)$$

Where $k = i_2 \cdot i_1$, at $n_A = 0$

With shaft K fixed:

Arm D, n_D Shaft K, n_K Shaft A, n_A
 1st Revolution 0 1 / ($i_1 \cdot i_2$) 1 (clockwise)
 2nd Revolution $-1 / (i_1 \cdot i_2)$ $-1 / (i_1 \cdot i_2)$ $-1 / (i_1 \cdot i_2)$
 (anticlockwise around space axis X at 2nd revolution)
 Sum of 1st and 2nd $-1 / (i_1 \cdot i_2)$ 0 $1 - 1 / (i_1 \cdot i_2)$

Putting $k = i_1 \cdot i_2$ as before, then

$$n_A / n_D = k - 1 \quad (3.2.15)$$

at $n_K = 0$. Therefore, if n_A and n_K are given independently, then

$$n_D = n_A / (k - 1) + n_K \cdot k / (k - 1), \quad (3.2.16)$$

c) For mechanical positioning, there are several rigid feedback systems or positively connecting units using various kinds of mechanical mechanism, for example crank and lever, cam and roller, and screw and nut, as well as several kinds of gearing, etc. These systems are arranged to transmit driving power to follower parts and simultaneously to provide positioning information. Here we describe two kinds of special important positioning systems: indexing mechanisms and harmonic gearing.

Several kinds of indexing positioning mechanism using a cam and follower are shown in Fig. 3.2.6. These mechanisms are used for intermittent positioning of the worktable. However,

nanometre positioning accuracy cannot be obtained, because neither the mechanical linkage between cam and follower nor the cam features can be obtained to such accuracy.

A harmonic gearing mechanism to obtain a high reduction ratio is shown in Fig. 3.2.7. A harmonic gearing consists of three basic elements: wave generator, flex spline, and circular spline. The wave generator is an elliptical cam with a thin layer of ball bearings on its exterior and is in general used for input. The flex spline is a very thin deformable cup of high-tensile steel with fine gear teeth on its exterior and serves in general for output. The circular spline fixed to the gear case is a rigid hollow ring with fine gear teeth on the inner side, mating with the flex spline; it has two more teeth than the flex spline has.

The principle of gearing performance is that when the flex spline is deformed elastically by the elliptical cam of the wave generator, the teeth of the flex spline at the extremities of the cam mate with those of the circular spline. Accordingly, when the wave generator cam rotates clockwise by 360° , the flex spline teeth mate in turn with those of the circular spline, and as a net result the flex spline has rotated anticlockwise by two teeth. Therefore if the total number of circular spline teeth is 200, then the flex spline rotates by $1/100$ per revolution of the wave generator cam.

This mechanism is very efficacious, because the effective gear reduction is achieved by a plane gear mechanism with small gear surface slip and as a result, in spite of the very large reduction ratio, the transmission efficiency is very high. However, although the accuracy of angle transmission can be expected to be good, at present it is not yet of nanometre-accuracy standard.

3.2.3 Hydraulic servo-actuator systems

(a) A hydraulic servo-actuating system is shown in Fig. 3.2.8. It consists of three elements: hydraulic actuator, servo pilot valve, and position feedback lever.

When the lap gap width between servo pilot piston PP and corresponding valve port VP is e , as shown in the figure, the actuating fluid flows through the gaps into and out of actuator cylinder AC. If the pressures on both sides of piston AP are denoted as p_1 and p_2 .

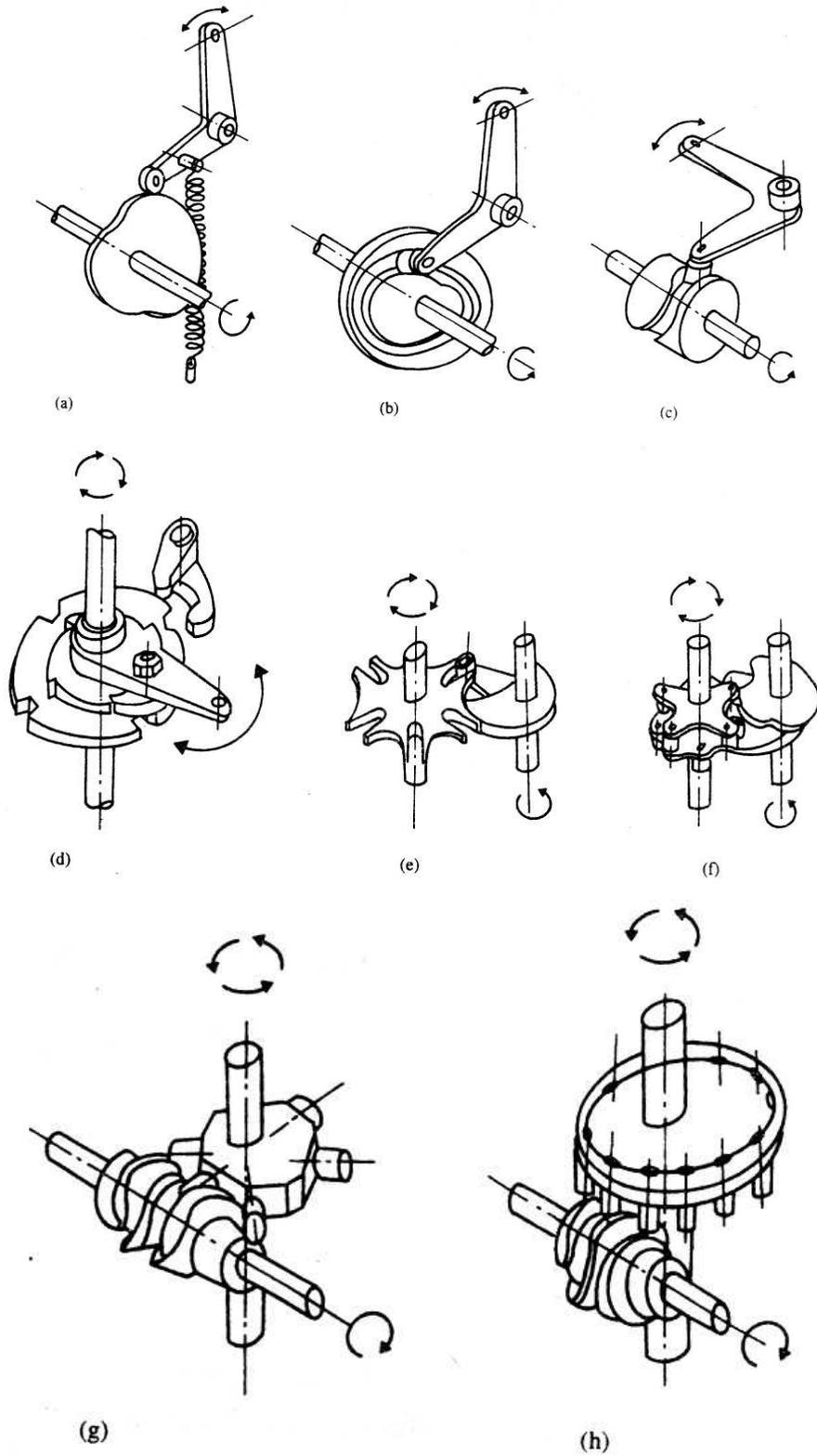
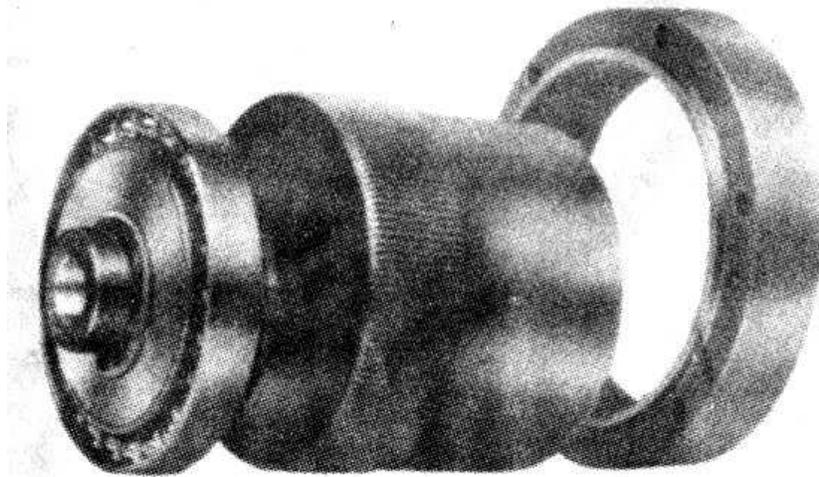
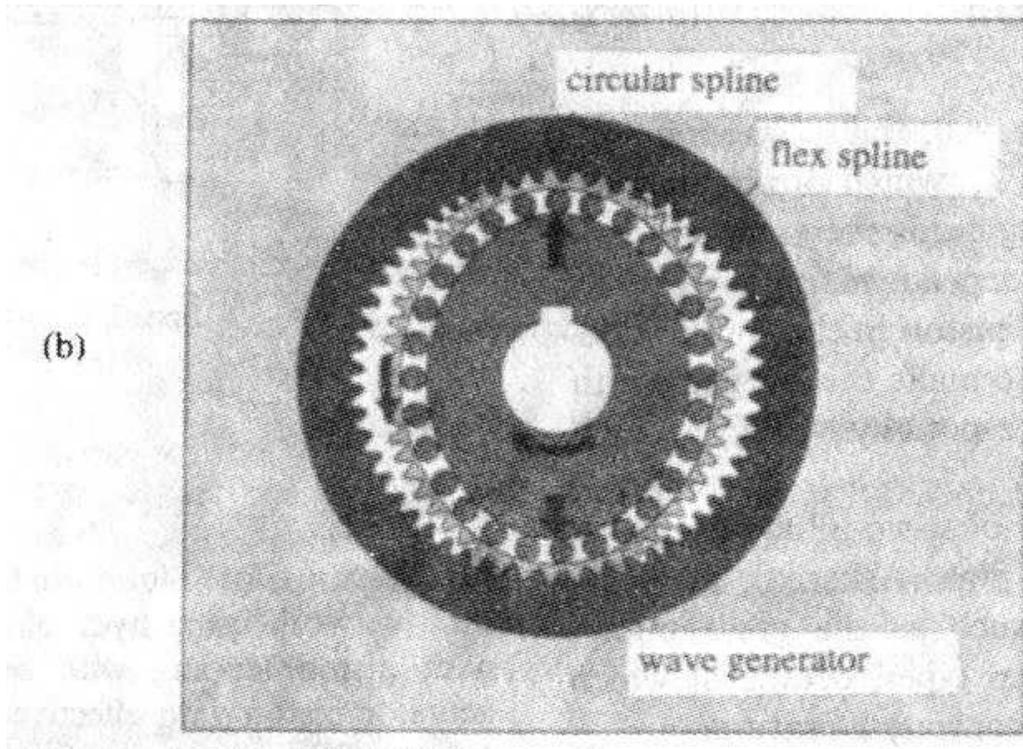


Fig. 3.2.6. Indexing cam mechanisms. (a) Outer cam. (b) Groove cam (c) Barrel (cylindrical) cam. (d) Ratchet wheel. (e) Geneva cam. (f) Parallel cam. (g) Roller gear. (h) Barrel cam.



(a)



(b)

Fig. 3.2.7. Harmonic reduction gearing. (a) Exploded view. (b) principle.

the kinematic equation for the displacement y of the actuator piston rod with load is

$$(p_1 - p_2)A = m \cdot d^2 y / dt^2 \pm L, \quad (3.2.17)$$

where m is the mass of the operating actuator piston rod system with braking load L , and A is the effective area of the actuator piston. Moreover, the acting fluid flow rate is given by

$$A \cdot dy / dt = c(p_i - p_1) \epsilon = c(p_2 - p_d) \epsilon, \quad (3.2.18)$$

where c is a coefficient of fluid flow under the linear flow assumption.

Therefore we have the following equations:

$$\begin{aligned} p_1 &= p_i - A \cdot dy/dt / (c \in) \\ p_2 &= p_d + A \cdot dy/dt / (c \in) \end{aligned} \quad (3.2.19)$$

Using eqn. (3.2.17), we obtain

$$(p_i - p_d) - 2A \cdot dy/dt / (c \in) = (m \cdot d^2y/dt^2 \pm L) / A, \quad (3.2.20)$$

If we set the value of total working pressure very high, $p_i - p_d \gg (m \cdot d^2y/dt^2 \pm L) / A$, then the equation reduces to

$$\in (p_i - p_d) = 2A(dy/dt) / c \quad (3.2.21)$$

Therefore the speed of the actuator piston rod becomes

$$dy/dt = [c(p_i - p_d) / 2A] \cdot \in, \quad (3.2.22)$$

and if we put $T_s = 2A / c(p_i - p_d)$, then T_s corresponds to the starting time or response time constant of the control system. The system gain $(dy/dt) / \in = 1 / T_s$ is used to characterize the system, in general.

Moreover, if for the displacement or position of the actuator rod we form a closed loop by a feedback lever system,

$$\in = Y - \beta y, \quad (3.2.23)$$

then the performance of the closed-loop control system becomes

$$T_s(dy/dt) + \beta y = Y \quad (3.2.24)$$

Accordingly, the position of the actuator rod tends to

$$y = Y / B \quad (3.2.25)$$

at the stable condition, $dy/dt = 0$, and its starting time or response time constant is T_s , which can be set very small under a large actuating fluids pressure $(p_i - p_d)$.

The positioning accuracy depends mainly on the lap gap width of the servo pilot piston in the stable state. It is therefore very difficult to make the lap gap width to nanometre accuracy, especially with a high actuating pressure.

There are also other types of servo pilot valve, such as the D slide and the jet pipe (Askania), and also several types of actuator such as the great pump, rotary vane, and axial pump types, details of which may be obtained in the appropriate literature.

(b) A typical hydraulic servo-actuator system applied to an NC (numerical control) machine is shown in Fig. 3.2.9. It consists of three basic elements: electric stepping motor, hydraulic servo pilot valve with feedback compensator, and hydraulic actuator of axial piston pump type.

The angular positioning signal delivered from an NC pulse generator is supplied to the stepping motor SM. Then the analogue output as angular displacement of the stepping motor is transferred to the piston of the hydraulic servo pilot valve PV through the reduction gear RG. From the other side, the pilot valve piston is pushed back by the screw thread mechanism of feedback compensator FC, depending on the angular displacement of output shaft OS.

As shown in section AA of the figure, the output shaft OS is driven by the axial piston pump-type or swash plate hydraulic actuator SPA, controlled by the servo-pilot valve.

Accordingly, the acting principle of this system seems to be the same as that of the previous system,

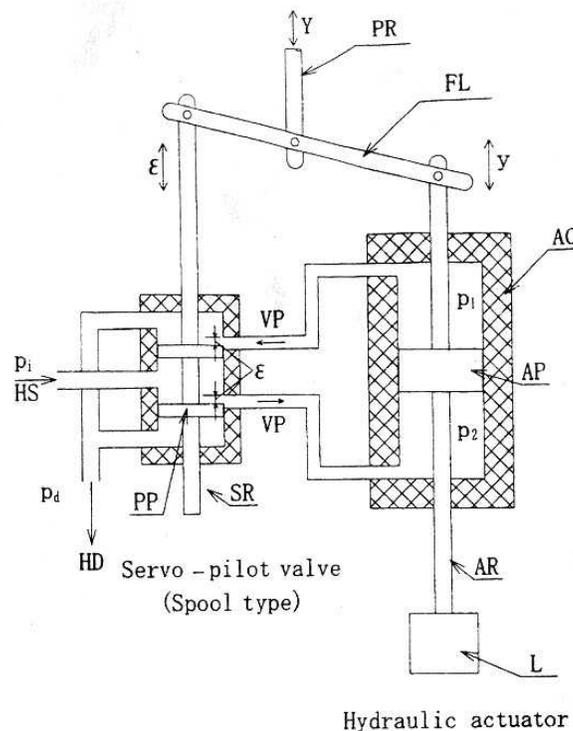


Fig. 3.2.8. Closed-loop control system with rigid feedback network using hydraulic servo-actuator. PP: servo pilot piston lapping with servo pilot valve port AP: actuator piston with

effective area A (m^2) AC: actuator cylinder PR: position commanding rod with displacement Y (m) AR: positioning rod of actuator with displacement Y (m) FL: position feedback lever of lever ratio β SR: servo pilot valve rod of displacement $\varepsilon = Y - \beta y$ (m) VP: servo pilot valve port HS, HR: hydraulic source and drain respectively p_i, p_d . pressures of source and drain p_1, p_2 : pressures in actuator cylinder at both sides of piston.

but the positioning angular resolution attains the order of $10 \mu\text{m}$, the limit depending mainly on the accuracy of the lap gap width.

In addition, the response time constant T_s of the system can be set to the order of 1 ms or the system gain to several hundred reciprocal seconds. Therefore the driving frequency of the NC signals can be set to the order of several hundred hertz.

By means of this electro-hydraulic stepping actuator system, the positioning of heavy workpieces and rigid tools for large cutting forces has been controlled by high-frequency NC signals since 1963. However, recently, this system has been replaced by the d.c. servo motor with high-resolution encoder, and a.c. inverter servo motor systems are also going to be used.

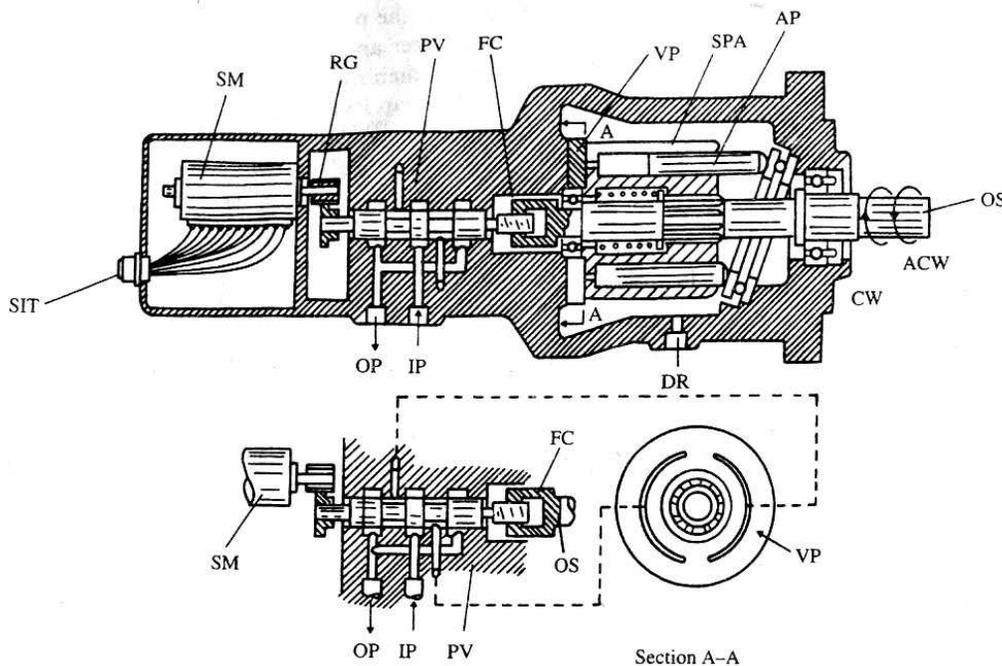
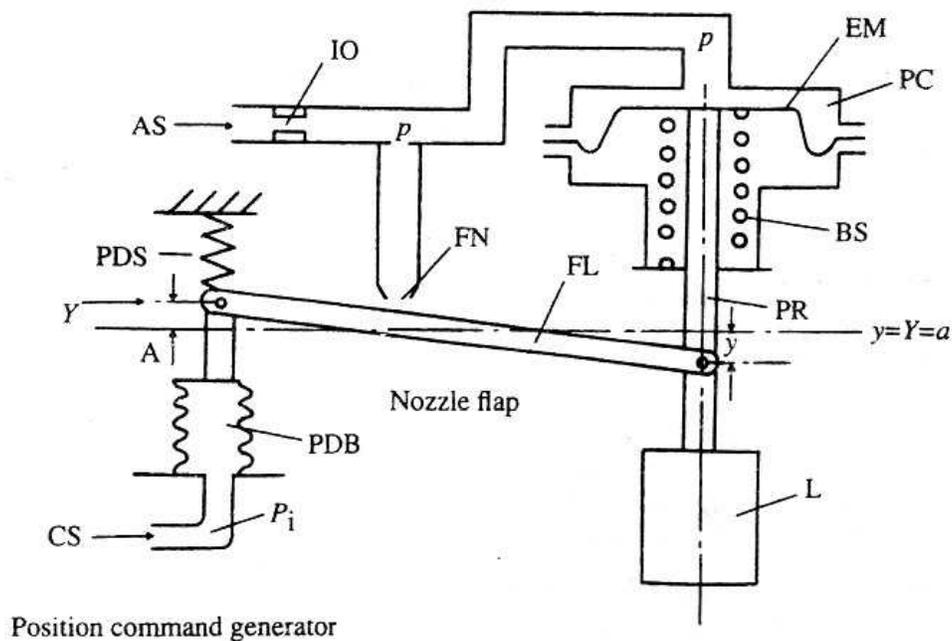
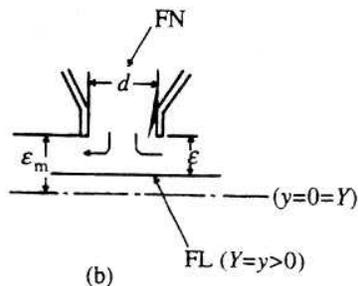
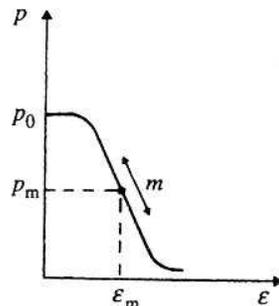


Fig. 3.2.9. Electro-hydraulic stepping servomotor. SM: electric stepping motor RG: reduction gear PV: pilot servo-valve FC: feedback compensator with thread mechanism VP: valve plate

AP: axial drive piston OS: output shaft OP: outlet oil port IP: inlet oil port CW: clockwise ACW: anticlockwise DR: drain SIT: position command signal input terminal SPA: swash plate actuator



(a)



(b)

Fig. 3.2.10. Pneumatic servo-positioner system with nozzle flap, (a), (b) See text. AS: air source IO: inlet orifice PR: positioner rod PC: positioner air chamber EM: elastic membrane of positioner BS: balancing spring of positioner CP: position command pressure p_i FN: flap nozzle

FL: feedback lever with flapper PDB: command displacement bellows PDS: command displacement spring p : positioner pressure equal to flap nozzle p_0 . air source pressure ε : gap between flap and nozzle Y : position command displacement y : positioner rod displacement L : load.

3.2.4 Pneumatic servo-actuator systems

(a) The construction of a typical pneumatic servo- positioner system with nozzle flapper is shown in Fig. 3.2.10a. The system consists of three basic elements: servo-positioner, nozzle flap, and position command generator.

w i

The position command generator comprises position command signal pressure generator CS, command position displacement bellows PDB, and spring PDS. The input position quantity Y is given by the displacement of the end point A of the bellows, according to the pressure p_i , of the input position command.

The nozzle flap mechanism is a highly sensitive pressure transformer, because very fine change of nozzle gap ε causes a considerable change in air pressure p in the nozzle. The details of the mechanism are as follows.

The volumetric flow rate v at the flap nozzle is determined on the assumption that the air flow in the nozzle is throttled by the area generated by the nozzle periphery πd and nozzle gap ε as shown in Fig. 3.2.10b:

$$v = c_1 \pi d \varepsilon p^{1/2} \quad (3.2.26)$$

where c_1 is a flow coefficient and p is the air pressure in the nozzle. However, v is determined by the throttling due to inlet orifice IO:

$$v = c_2 a (p_0 - p)^{1/2} \quad (3.2.27)$$

where c_2 is a flow coefficient, a is the flow area of the orifice and p_0 is the source pressure.

Therefore the relation between the gap ε of the flap nozzle and the pressure within the nozzle is

$$(P_0 - P) / P = (c_1 \pi d / c_2 a)^2 \cdot \varepsilon^2 \quad (3.2.28)$$

Accordingly the general $\varepsilon - p$ characteristic is as shown in Fig. 3.2.10b. As a result, the differential increment of pressure p with gap ε at the midpoint m of the characteristic curve is

$$(dp / d\varepsilon)_m = -2(c_1 \pi d / c_2 a)^2 \cdot \varepsilon_m \cdot (P_m^2 / P_0) \quad (3.2.29)$$

Then, putting

$$K_m = 2(c_1 \pi d / c_2 a)^2 \cdot (P_m^2 / P_0) \quad (3.2.30)$$

$$dp = -K_m d\varepsilon$$

around the point m, where k_m becomes very large at the larger and smaller values of a, which means that the mechanism can be used as a highly sensitive sensor of fine gap length.

The integrated form in the region of m is

$$(p - p_m) = -K_m (\varepsilon - \varepsilon_m) \quad (3.2.31)$$

and the gap distance ε of flap nozzle is set by the flap lever as follows:

$$\varepsilon - \varepsilon_m = Y - (1 - \Delta)y \quad (3.2.32)$$

where Δ is a characteristic constant of the mechanism, determined as shown below, and y is the output displacement determined by the servo-positioner as

$$y = k(p - p_m) \quad (3.2.33)$$

where k is the characteristic constant of the servo-positioner.

Eliminating p from eqns. (3.2.31)—(3.2.33), we obtain the following relation:

$$y/k = -k_m [Y - (1 - \Delta)y] \quad (3.2.34)$$

Putting $Y = y$ in eqn (3.2.34), we have

$$\Delta = -1/(k/K_m) \quad (3.2.35)$$

By means of this mechanism having this constant Δ , an input command signal Y generates an output y equal to Y . In other words, the position of load y is always maintained equal to the command signal Y , but with a response time lag.

(b) An electro-pneumatic servo-actuator as shown in Fig. 3.2.11 has been proposed. The actuator consists of three basic elements: position command stepping motor SM, positioning screw thread rod PS with a small lead angle, and actuator piston nut PN of restricted rotation and mating with PS, holding the braking load.

The operating mechanism of the servo-actuator is as follows. With the self-locking condition occurring in the screw thread mating with a small lead angle θ the moving element of the system holds its initial position. In the self-locking state as shown in Fig. 3.2.12a, the rotation-restricted piston PN cannot be moved by the air pressure acting on the piston nut, because the latter cannot effect a rotation of the mating screw thread. As

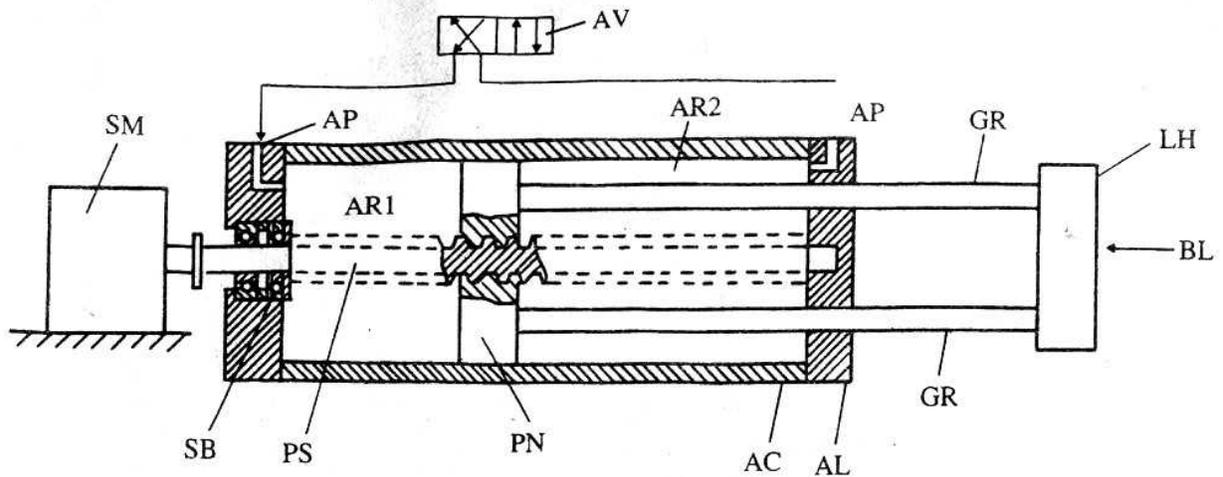


Fig. 3.2.11. Electro-pneumatic servo-actuator SM: position command stepping motor PS: positioning screw rod PN: positioning piston nut AR1, AR2: air chambers BL: braking load LH: load holder GR: guide rod AV: air changeover valve AP: air inlet or outlet SB: bearing for PS AC: air cylinder AL: air cylinder lid

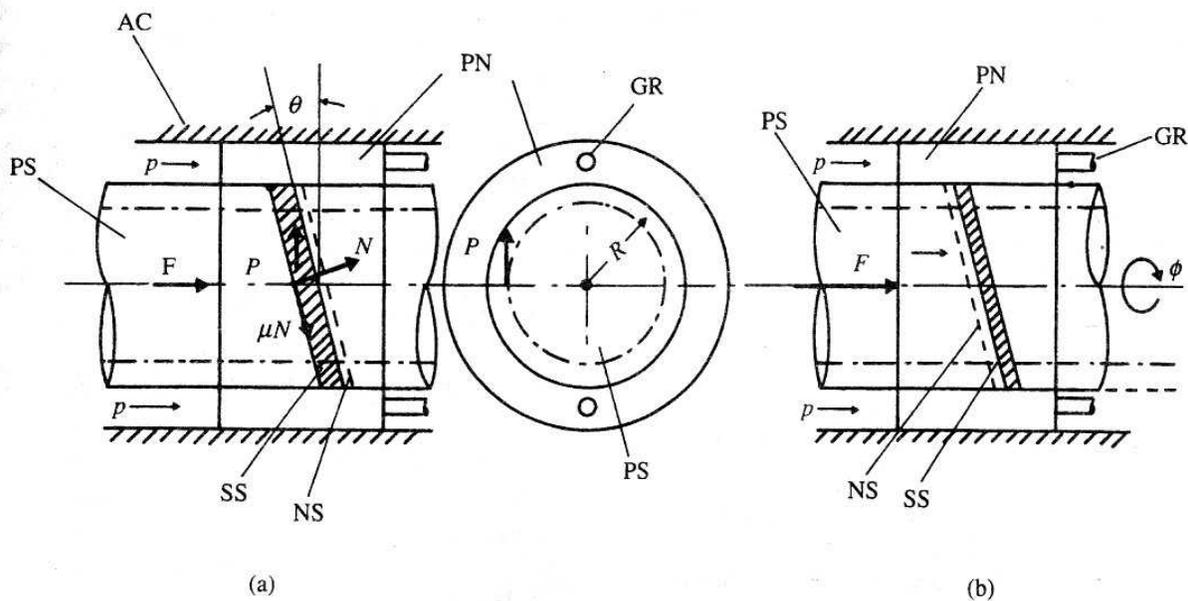


Fig. 3.2.12. Principle of electro-pneumatic servo-actuator, (a) Self-locking of piston nut. (b) Positioning of piston nut.

PS: positioning screw rod PN: positioning piston nut SS: screw and nut thread surface section AC: air cylinder GR: guide rod for piston nut p : air pressure in air chamber θ : lead angle of screw thread F : force acting on piston nut N : normal force acting on surface of PS πN : frictional

force acting on surface of PS against screw sliding direction P : radial driving force on the surface of PS due to F R : radius of screw thread pitch ϕ : angle of rotation of positioning screw rod PS

NS: assumed surface section of screw thread

shown in the figure, the normal force N acting on the piston nut PN generates a rotation force P acting along the screw thread surface, given by

$$P = N \sin \theta - \mu N \cos \theta = N \cos \theta (\tan \theta - \mu) \quad (3.2.36)$$

Accordingly, if $\tan \theta < \mu$, then $P < 0$ (negative). In this case, no rotation of the screw thread PS can occur i.e. self-locking occurs.

For example, if the friction coefficient $\mu = 0.2$, then the critical self-locking lead angle of the rectangular thread $\theta_c = 10^\circ$; at $\mu = 0.25$, $\theta_c = 14^\circ$. For a trapezoidal thread, both μ and θ_c are greater than for a rectangular thread.

However, if the positioning screw nut PN rotates positively clockwise by ϕ due to the rotation of stepping motor, as shown in Fig. 3.2.12b, then the piston nut PN moves to the right, corresponding to the rotation of the screw thread PS. The reason is that the screw surface and nut surface are separated by the mating clearance between the two surfaces, owing to the force pushing on the piston nut PN. Therefore a small rotation of the stepping motor is easily communicated.

By means of this servo-actuator system, a piston nut with a heavy braking load can easily be moved, aided by pneumatic power, by the command amount of displacement.

3.2.5 Electric servomotor systems

At present the positioning systems of robot hands and tools and workplaces on machine tools are usually operated by an electric servomotor unit, d.c. or a.c. using a numerical input signal.

(a) *Basic electrokinetic motion analysis of d.c. servomotor.*

Referring to Fig. 3.2.13a, the basic electromechanical characteristics of a d.c. servomotor are given by two fundamental equations:

$$V_M = N_M \cdot \omega_M \quad (3.2.37)$$

$$T_M = N_M \cdot i_M \quad (3.2.28)$$

where

V_M = induced voltage due to armature winding (V) ω_M = angular speed of armature rotor (rad

ω_M = electromechanical coupling factor of electric motor (N m A⁻¹ or V s rad⁻¹)

T_M = output torque of armature shaft (N m)

i_M = electric current of servomotor (A).

Accordingly, when the electrical source V_0 or input voltage V_a is applied to the servomotor with a rotating load, the following current equations are obtained:

$$V_a = V_0 - R_0 i_M = i_M R_i + V_M \quad (3.2.39)$$

where R_0 and R_i are the internal resistances (Ω) of the applied electrical source and servomotor respectively.

Therefore the output torque of the armature shaft is expressed by

$$T_M = \left[N_M / (R_i + R_0) \right] \cdot V_0 - \left[N_M^2 / (R_i + R_0) \right] \cdot \omega_M \quad (3.2.40)$$

Or

$$T_M = N_M / R_i \cdot V_a - N_M^2 / R_i \cdot \omega_M \quad (3.2.40)$$

Moreover, eqn. (3.2.39) can be transformed to

$$\begin{aligned} V_0 &= i_M (R_i + R_0) + N_M \omega_M \\ &= i_M \left[R_i + R_0 + \left(N_M^2 \omega_M / T_M \right) \right] \\ &= i_M (R_i + R_0 + R_e) \end{aligned} \quad (3.2.39)$$

where $R_e = N_M^2 \omega_M / T_M$ is called the equivalent resistance of the mechanical load.

The output power or consumed power P (W) of the servomotor becomes

$$P = (R_i + R_e) i_M^2 = (R_i + R_e) V_0^2 / (R_0 + R_i + R_e)^2 \quad (3.2.41)$$

Therefore the maximum output power is obtained at matched internal and external resistances of the cell $(R_i + R_e) = R_0$, because under the resistance- or load matching condition, $dP / d(R_i + R_e) = 0$ is realized.

Accordingly, when the angular speed of the device is predetermined and the necessary torque is also limited, a servomotor of suitable electromechanical coupling constant must be selected.

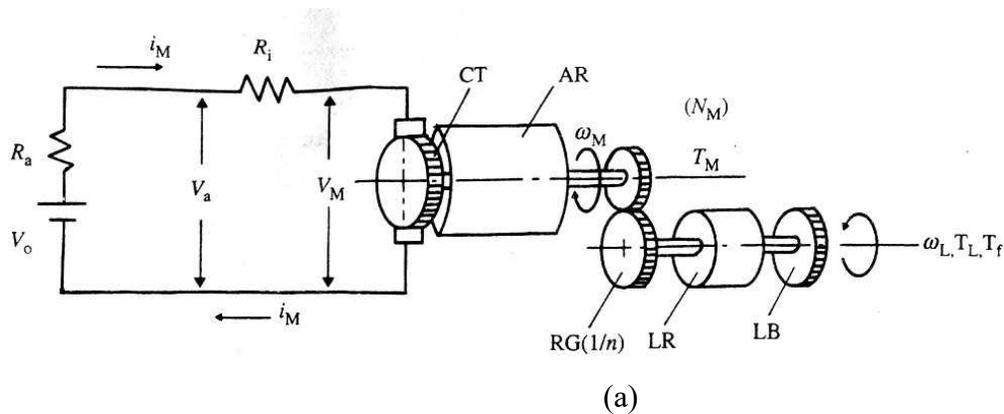
The dynamic kinetic equation of the servomotor unit with load is

$$J_{ML} (d\omega_M / dt) + \mu \omega_M + T_L / n \pm T_f / n = T_M \quad (3.2.42)$$

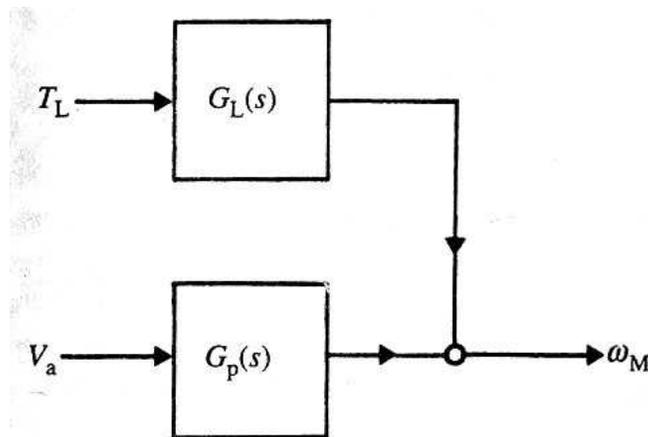
where $J_{ML} = J_M + J_L / n^2$ is the total moment of inertia around the armature axis (kg m^2), and J_M is the moment of inertia of the armature itself, J_L the moment of inertia of the load itself, and n the reduction ratio of the gearing, i.e. J_L / n^2 is the equivalent moment of inertia of the rotating load around the armature axis; $\mu\omega_m$ is the viscous resistance proportional to the angular speed of armature and other parts, and T_f and T_L (N m) are the friction torque and braking torque of the rotating load respectively.

Then the transfer function of working unit WU, $G_p(s)$, on the speed of the combined devices of servomotor and tool or worktable for the applied driving voltage is expressed as follows, using the Laplace notation $s = d / dt$:

$$(sJ_{ML} + \mu)[\omega_M] - [T_L] / n - [T_f] / n = N_M / R_i \cdot [V_a] - N_M^2 / R_i \cdot [\omega_M] - [T_L] / n - [T_f] / n \quad (3.2.43)$$



(a)



(b)

Fig. 3.2.13. Electric servomotor systems, (a) Circuit network of d.c. servomotor, (b) Transfer functions. See text for meaning of symbols

As a result, putting $T_f = 0$

$$G_p(s) = [\omega_M / V_a] = 1 / \{ (sJ_{ML} + \mu) \cdot R_i / N_M + N_M \} \quad (3.2.44)$$

$$G_L(s) = [\omega_M / T_L] = - (R_i / nN_M) / \{ (sJ_{ML} + \mu) \cdot R_i / N_M + N_M \} \quad (3.2.45)$$

Then

$$[V_a]G_p(s) + [T_L]G_L(s) = (\omega_M) \quad (3.2.46)$$

as shown in Fig. 3.2.13b.

(b) Machine-tool numerical control or positioning system

As shown in Figs 3.2.14 and 3.2.15, the basic system consists of positioning pulse signal deliverer PD, position pack or positioning compensator PP, servo- pack or speed compensator SP, and working unit WU which contains the driving servomotor and operating machine tool set, together with feedback devices consisting *in situ* speed meter or tachometer TG and *in situ* position meter or position pulse generator PG or encoder EC.

The numerical position control system is basically the same as an ordinary feedback control system. The difference ε_d between the digital command position Y_d based on pulse rate $F_c(S^{-1})$ and *in situ* digital position y_d is obtained at the position compensator PP,

$$\varepsilon_d = Y_d - y_d \quad (3.2.47)$$

Then the analogue speed command signal $K_p \varepsilon$ is sent to the speed pack SP, which produces the amplified driving power for the servomotor with voltage V_a proportional to the difference between command speed signal $K_p \varepsilon$ and *in situ* speed feedback signal $\beta \omega_M$. That is,

$$V_a = K_v (K_p \varepsilon - \beta \omega_M) \cdot K_a \quad (3.2.48)$$

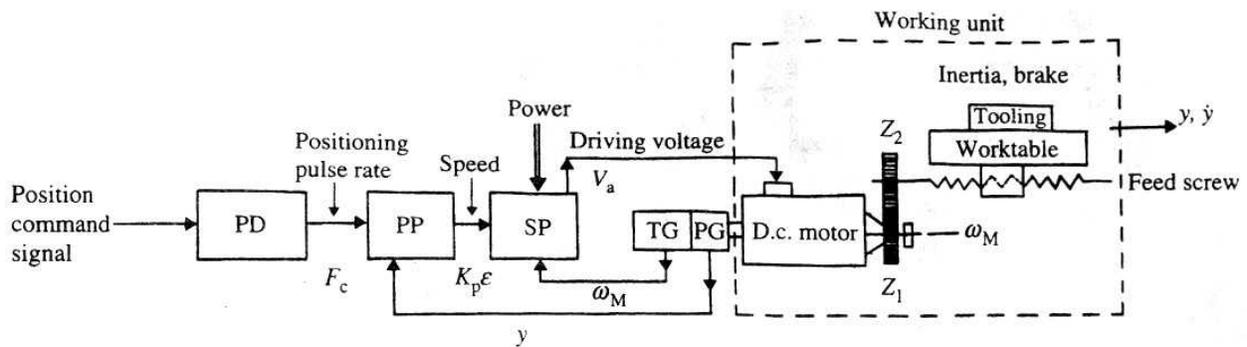


Fig. 3.2.14. Numerical positioning system (NC system) for machine tools PD: position command pulse rate deliverer F_c : position command pulse rate (10^{-2}mm s^{-1}) PP: position pack FF: feed

forward signal processor K_p : co-efficient of positioning speed proportional to ε ($V \text{ mm}^{-1}$) ε : difference between command position Y and *in situ* position y (mm) Y_d : digital command position (mm) y_d : digital *in situ* position (mm) SP: speed pack K_v : speed signal amplifying factor K_a : power amplifying factor WU: working unit (servomotor and working device) T_M : output torque of servomotor (Nm) T_L : braking torque of load (Nm) V_a : driving voltage for servomotor (V) ω_M : angular speed of servomotor (rad s^{-1}) $G_p(s)$: transfer function for angular speed of servomotor.

From eqn. (3.2.44),

$$[V_a] = [\omega_M] / G_p(s),$$

then the system gain transfer function $G_s(S) = [(dy/dt)/\varepsilon]$, the ratio of worktable speed (dy/dt) or $c\omega_M$ to *in situ* position difference ε , is obtained, where c is a transfer constant to worktable speed from servomotor axis:

$$G_s(s) = cK_vK_aK_pG_p(s) / [1 + \beta K_vK_aG_p(s)] \quad (3.2.49)$$

Accordingly, if we can set the amplifying ratio term sufficiently large compared with $\beta, K_vK_aG_p(S) \gg \beta(V \text{ s rad}^{-1})$, the system gain transfer function $G_s(S)$ tends to the following constant system gain K_s , in spite of $G_p(S)$:

$$G_s(s) \rightarrow cK_p / \beta = K_s (s^{-1}) \quad (3.2.50)$$

In practice, K_s has a nearly constant value of 20-30 s^{-1} for an SCR (silicon-controlled rectifier) or thyristor amplifier, and 100-1000 s^{-1} for a transistor.

Moreover, the system gain K_s is inversely proportional to the response time T_s of the worktable motion. That is, from eqn (3.2.50) with on $G_s(S) = [(dy/dt)/\varepsilon]$, we get the following Laplace differential equation:

$$sy = K_s \varepsilon = K_s (F_c / s - y) \quad (3.2.51)$$

the solution of which is

$$dy/dt = F_c [1 - \exp(-K_s t)] \quad (3.2.52)$$

This means that the speed of the worktable (dy/dt) follows with a time lag $T_s = 1/K_s$ the pulse rate of the numerical positioning signal F_c .

(c) Gear ratio to maximize the output angular acceleration

The kinetic equation of a servomotor with a rotating load is as follows (refer to Fig. 3.2.13a):

$$\eta(T_M - J_M \alpha_M) = (T_L - J_L \alpha_L) / n \quad (3.2.53)$$

where α_M and α_L are the angular accelerations of servomotor and rotating load respectively, and η is the efficiency of the reduction gearing.

With $\alpha_M = n\alpha_L$,

$$T_M \cdot n - T_L / \eta = (J_M \cdot n^2 - J_L / \eta) \alpha_L \quad (3.2.54)$$

Now putting $\eta = 1$ and $T_L = 0$, we have

$$\alpha_L = (T_M \cdot n) / (J_L + J_M \cdot n^2) \quad (3.2.55)$$

To obtain the maximum angular acceleration of the load at gear ratio n , make the differential of α_L with respect to n zero; then

$$0 = T_M (J_L - J_M n^2) / (J_L + J_M n^2)^2 \quad (3.2.56)$$

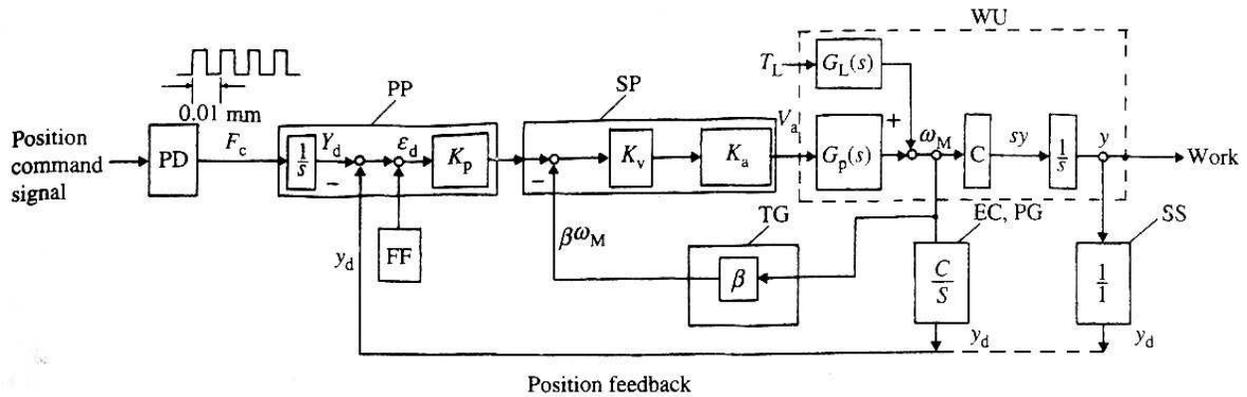


Fig. 3.2.15. Block diagram of NC positioning network using servomotor. Symbols as in Fig. 3.2.14.

that is,

$$J_M = J_L / n^2 \quad (3.2.57)$$

Accordingly, the maximum output angular acceleration is obtained under the condition that the equivalent moment of inertia of the rotating load is equal to the moment of inertia of the armature rotor.

(d) *Example of a machine-tool position control system*

In the systems shown in Figs 3.2.14 and 3.2.15, assume that the feed screw lead $l = 5$ mm and the gear ratio $z_2 / z_1 = n = 1$; then the speed of the worktable is given by

$$(dy / dt) = c\omega_M \quad (3.2.58)$$

where $c = l/2\pi = 8.0 \times 10^{-4}$ m. Therefore the equivalent moment of inertia of the worktable about the armature rotor axis becomes

$$J_{Le} = Mc^2 = 70 \cdot (8.0 \times 10^{-4})^2 = 4.5 \times 10^{-5} \text{ kg m}^2 \quad (3.2.59)$$

as deduced from the angular kinetic energy expressed in the following manner:

$$1/2 \cdot J_{Le} \omega_M^2 = 1/2 \cdot M (c\omega_M)^2 \quad (3.2.59)$$

Therefore, as mentioned previously the moment of inertia of the servomotor armature rotor J_M must be selected to be nearly the same as J_{Le} .

$$J_M \square J_{Le} \quad (3.2.60)$$

Hence the servomotor used for this device should have a moment of inertia of the armature of $J_M = J_{Le} = 4.5 \times 10^{-5} \text{ kgm}^2$. From the servomotor list of Yasukawa Electric Co., the inertia motor UGMMEM-06AA having a rotor moment of inertia $J_M = 5.67 \times 10^{-5} \text{ kgm}^2$ is selected.

Then the starting time-constant of the motor plus machine tools unit is obtained as follows. The relation between the generating torque of the servomotor T_M and the angular speed ω_M is given by eqn. (3.2.40');

$$T_M = N_M / R_i \cdot V_a - N_M^2 / R_i \cdot \omega_M$$

together with the motor data $N_M = 1.06 \times 10^{-1} \text{ V s rad}^{-1} (\text{N m A}^{-1})$, $R_i = 0.84\Omega$, and rated input voltage $V_a = 40.5 \text{ V}$. Hence

$$T_M = 5.10 - 1.64 \times 10^{-2} \cdot \omega_M \quad (3.2.61)$$

On the other hand, the kinetic equation of the servo system with the total moment of inertia around the armature rotor axis J_{ML} is

$$T_M = J_{ML} (d\omega_M / dt) \quad (3.2.62)$$

with $J_{ML} = J_M + J_{Le} = 5.67 \times 10^{-5} + 4.5 \times 10^{-5} = 10.2 \times 10^{-5} \text{ kg m}^2$.

Combining eqns. (3.2.61) and (3.2.62) and solving with the initial condition $\omega_M = 0$ at $t = 0$, we get

$$\omega_M = 3.8 \times 10^2 \cdot [1 - \exp(-t / 7.6 \times 10^{-3})] \quad (3.2.63)$$

Accordingly, the starting time constant T_s of the servomotor plus the machine tool becomes $7.6 \times 10^{-3} \text{ s}$ at the rated condition, and at the final state the angular speed of the motor becomes

$$\omega_{MF} \rightarrow 3.8 \times 10^2 \text{ rad s}^{-1} \quad (3.2.64)$$

The table speed after the starting response time T_s is obtained from the value of ω_{Ms} corresponding to $t = T_s$ in eqn. (3.2.63):

$$(dy / dt)_s = c(\omega_{Ms}) = 8.0 \times 10^{-4} \times 2.4 \times 10^2 = 1.94 \times 10^{-1} \text{ m s}^{-1} \quad (3.2.65)$$

Therefore, in this numerical control system, the minimum time required per unit feed pulse of 0.01 mm must be less than the time constant $T_s = 7.6 \times 10^{-3} \text{ s}$, i.e. the minimum pulse width

$$T_{pm} = 0.01 \text{ mm} / 1.94 \times 10^{-1} \text{ m s}^{-1} = 0.5 \times 10^{-4} \text{ s}$$

or the maximum pulse rate

$$1 / T_{pm} = 2 \times 10^4 \text{ s}^{-1} \quad (3.2.66)$$

In the total numerical control system as shown in Fig. 3.2.15, if we assume the peak pulse voltage to be the rated voltage, 40.5 V, then the system gain K_s of the position control system becomes

$$\begin{aligned} K_s &= (\text{table speed after response time constant/unit pulse feed}) \\ &= 1.94 \times 10^{-1} \text{ m s}^{-1} / 0.01 \text{ mm} = 1.94 \times 10^4 \text{ s}^{-1} \end{aligned} \quad (3.2.67)$$

Therefore for this system, a power transistor type of amplifier becomes necessary.

(e) Example of position control system for a robot hand assembly of R-θ type

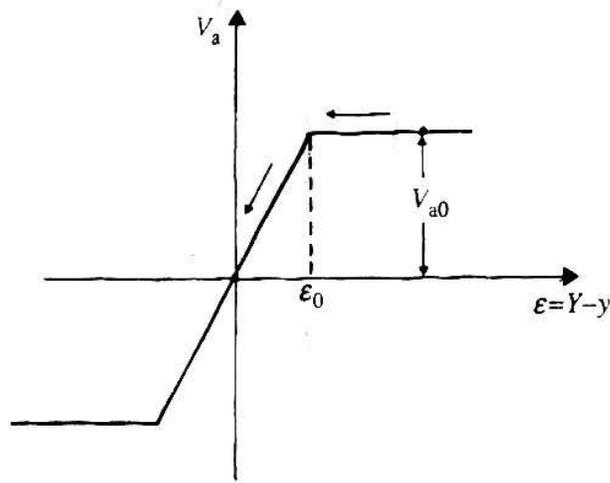
The system is shown schematically in Fig. 3.2.16, where the motion around the 0-axis is the most important. The rotating robot hand unit around the θ-axis consists of servomotor M1, reduction gear of harmonic type RG, and radial motion arm unit RA with assembly grip unit.

The moment of inertia of the rotation unit around the θ-axis is assumed to be

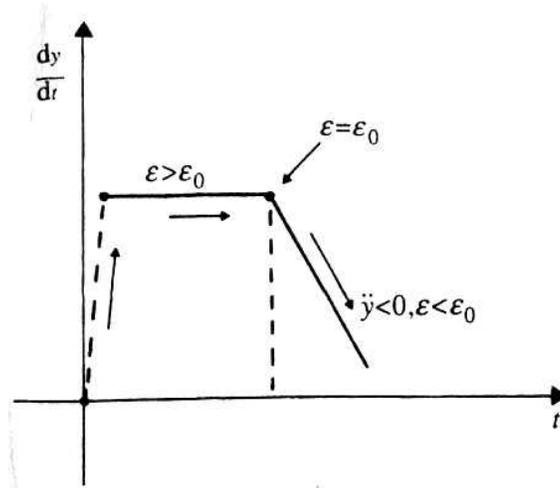
at the same input rated voltage V_a and the same initial condition, the angular speed of the motor armature rotor ω_M becomes

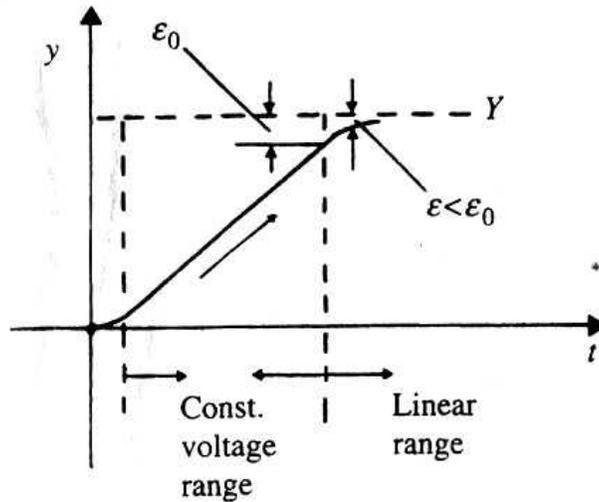
$$\omega_M = 3.8 \times 10^2 \left[1 - \exp\left(-t / 7.6 \times 10^{-3}\right) \right] \quad (3.2.70)$$

And after the starting response time constant $T_s = 7.6 \times 10^{-3}$ s, the grip unit speed S_0 becomes



(a)





(b)

Fig. 3.2.17. Characteristics of servomotor with non-linear driving voltage range, (a), (b) See text

$$s_0 = r \cdot \omega_{Ms} / n = 0.3m \times 3.8 \times 10^2 (1 - 0.368) \text{ rad s}^{-1} = 0.72 \text{ m s}^{-1} \quad (3.2.71)$$

Accordingly, for a continuous-path control system, the positioning resolution δ for a positioning signal (of pulse width $2T_s$) becomes

$$\delta = 0.72 \times 2 \times 7.6 \times 10^{-3} = 1.09 \times 10^{-3} \text{ m}$$

However, for a point-to-point position control system, the position difference at the starting point $\varepsilon = Y - y$ is quite large. Therefore the command driving voltage $V_a = K_a K_v K_p \varepsilon$ falls outside the linear amplification range as shown in Fig. 3.2.17a. As a result, initially the servomotor rotates at a higher constant speed corresponding to V_{a0} , for example twice the rated voltage, and when the difference ε falls below the proportional limit ε_0 , the servomotor is decelerated and reaches the command stopping position, through the ordinary feedback control network, as shown in Fig. 3.2.17b.

(f) Concluding remarks

There are two other types of servomotors in use: the a.c. rotary or linear stepping motor for open-loop control, and the d.c. inverter motor for closed-loop control. Details are not given here, because the operating principle is nearly the same as for the d.c. servomotor dealt with above.

Finally, positioning of nanometre accuracy by a numerical control system using an electric servomotor has recently been achieved with the Fanuc nano-servo-positioner, as mentioned afterwards. Also recently, a low-speed and high-torque servomotor, a so-called direct-drive

motor, has been developed. The direct-drive motor will be the most useful servomotor for positioning of mechanical devices in the near future.

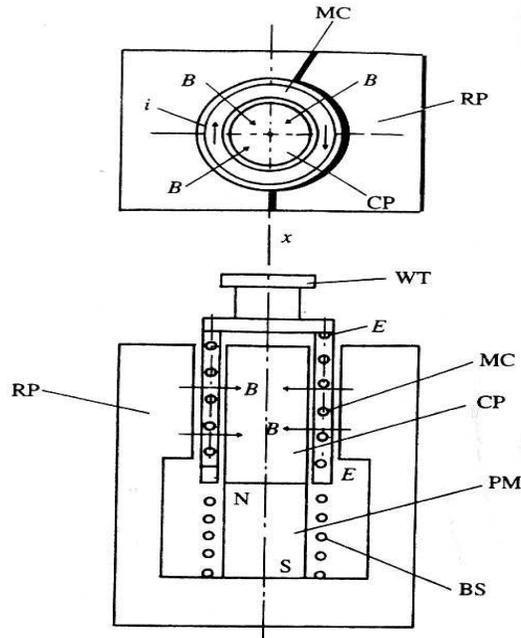


Fig. 3.2.18. Coreless moving-coil-type electrodynamic positioner. MC: moving-coil winding; RP: ring core; CP: iron core; B: mean magnetic flux density (T); i : electric current in moving coil (A); x : displacement of worktable; WT: worktable; E : input terminal voltage for moving coil (V); PM: permanent magnet for excitation; BS: backup coil spring

3.2.6 Electric linear servo-positioner and servo-actuator systems

(a) Voice coil or moving-coil type of linear electrodynamic positioner

As shown in Fig. 3.2.18, the system consists of electric moving-coil winding MC, iron core CP, and ring core RP for magnetic flux path with permanent magnet for excitation PM, and also a backup coil spring BS. The positioning of the moving coil is effected by the electrodynamic principle as in the electric servomotor, but in this system there is always a backup spring for positioning.

The dynamic performance of the system is expressed by the following set of differential equations:

$$M \left(\frac{d^2 x}{dt^2} \right) + kx(t) = Bli(t) \quad (3.2.72)$$

$$L \left(\frac{di}{dt} \right) + Ri(t) = E - Bl \left(\frac{dx}{dt} \right) \quad (3.2.73)$$

where, using the SI unit system,

x = displacement of work table with moving coil (m)

M = mass of working table with moving-coil unit (kg)

k = spring constant of backup spring coil (N m^{-1})

B = mean magnetic flux density in the gap between magnetic flux paths ($\text{T} = \text{Wb m}^{-2} = \text{V s m}^{-2}$)

l = total length of moving coil interlinked with magnetic flux (m)

i = electric current in moving coil (A)

L = internal inductance of moving coil ($\text{H} = \text{Vs A}^{-1}$)

R = internal electrical resistance of moving coil (Ω)

E = input terminal voltage for moving coil (V)

Bl = electromechanical coupling factor of the moving coil ($\text{N A}^{-1} = \text{V s m}^{-1}$).

Equations (3.2.72) and (3.2.73) are transformed as follows by using the Laplace symbol $s = d / dt$:

$$(Ms^2 + k)[x] = Bl[i] \quad (3.2.72)$$

$$(Ls + R)[i] = [E] - Bls[x] \quad (3.2.73)$$

Therefore the displacement of the moving-coil unit $[x]$ is expressed by

$$(Ls + R)(Ms^2 + k) / (Bl) + Bls[x] = [E] \quad (3.2.74)$$

The transfer function of the positioner $G_{ex}(S)$ is given by $G_{ex}(S) = [x] / [E]$, transformed from eqn (3.2.74).

Accordingly, for a step input voltage $[E] = [E]_0$, the steady displacement of the moving-coil unit x_{s0} is obtained as follows, putting terms in s , s^2 , and $s^3 \rightarrow 0$:

$$x_{s0} = [E]_0 / Rk / Bl \quad (3.2.75)$$

That is, the amplification factor K_p of the applied voltage for the command displacement X is

$$K_p = RK / Bl \quad (3.2.75)$$

The starting response time for positioning, T_{sp} , upon a step applied voltage $[E]_0$ is obtained from the Laplace transform eqn. (3.2.74), putting terms in s^2 and $s^3 \rightarrow 0$:

$$\{[(Lk / Bl) + Bl]_s + (Rk / Bl)\}[x] = [E] \quad (3.2.76)$$

then, referring to the previous section, the response time constant $T_{sp}(S)$ is

$$T_{sp} = \{(Lk / Bl) + Bl\} / (Lk / Bl) = (L / R) + (B^2 l^2 / (Rk)) \quad (3.2.77)$$

Therefore it is obvious that T_{sp} , the time lag on positioning, decreases as k increases.

Furthermore, if we take

$$[E] = E_0 \sin \omega t$$

for an acoustic transducer, then the response position or amplitude x_s for a small angular frequency ω becomes

$$x_s = E_0 \sin \omega t / \{Rk / Bl\} \quad (3.2.78)$$

However, when the angular frequency ω is relatively large, the effect of the natural frequency ω_n , of the system should be considered. That is, the Laplace-transformed second-order differential equation deduced from eqn (3.2.74), eliminating the s^3 term, is used as follows:

$$\{(RM / Bl)s^2 + [kL / Bl + Bl]s + (kR / Bl)\}[x] = [E] \quad (3.2.79)$$

Then we can transform the above equation to the differential equation as follows:

$$d^2x / dt^2 + 2\varepsilon dx / dt + \omega_n^2 x = E_0 / (RM / Bl) \cdot \sin \omega t \quad (3.2.79)$$

where $2\varepsilon = (kL / Bl) + Bl / (RM / Bl)s^{-1}$ and $\omega_n^2 = (kR / Bl) / (RM / Bl) = (k / M) \text{ rad}^2 \text{ s}^{-2}$.

Putting $(x = x_s \sin(\omega t - \phi))$ and inserting it into eqn. (3.2.79), we can determine the amplitude x_s and the phase lag ϕ as

$$x_s = E_0 / (RM / Bl) \cdot 1 / \{(\omega_n^2 - \omega^2) + 4\varepsilon^2 \omega^2\}^{1/2}$$

$$\tan \phi = 2\varepsilon \omega / (\omega_n^2 - \omega^2)$$

As an example of a numerical calculation, we take the following specification for a positioner: $E_0 = 1\text{V}$, $B = 1\text{T}$, $l = 0.6 \text{ m}$, $R = 0.2 \Omega$, $k = 5000 \text{ N m}^{-1}$, and $M = 0.1 \text{ kg}$. Then $\omega_n = \sqrt{(k / M)} \approx 200 \text{ rads}^{-1}$, and $|x_s| = E_0 / (RK / Bl) = 0.0006 \text{ m} = 0.6 \text{ mm}$.

(b) Moving-armature type of linear electrodynamic servo-actuator

As shown in Fig. 3.2.19, the system consists of a moving carriage furnished with iron-cored coil windings without a backup spring, and an exciting magnetic field circuit unit.

The motion analysis can be performed by eqns. (3.2.72) and (3.2.73), with a backup spring constant $k = 0$. Accordingly, the starting response time constant T_s for positioning for a step driving voltage E_0 is calculated as follows. Using eqn. (3.2.74) and $k = 0$,

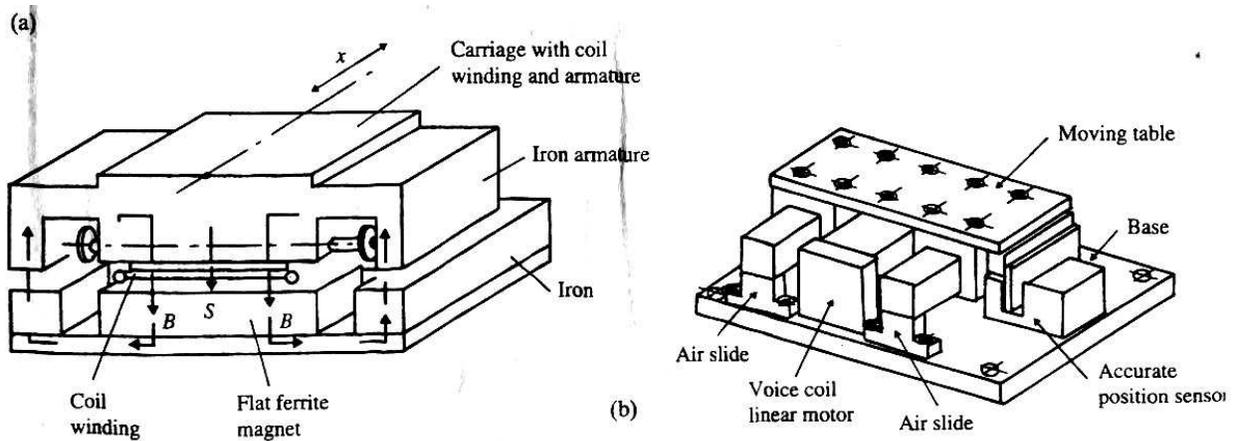


Fig. 3.2.19. Moving-armature-type electromagnetic servo-actuator or servomotor, (a) Linear motor with support roller.(b) Accurate positioner.

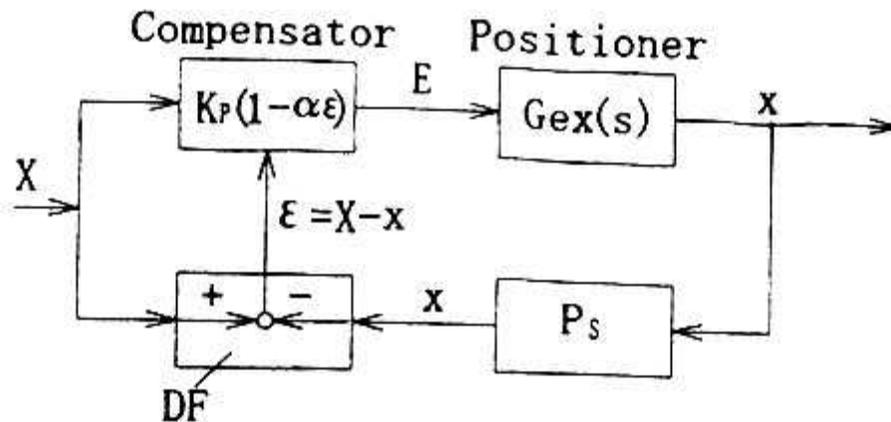


Fig. 3.2.20. Block diagram of complete control system for a positioner. PS position sensor $K_p =$ amplification factor DF: difference finder X : position command signal $\alpha =$ correction factor $x =$ position

$$(Ls + R)(Ms^2 / Bl) + Bls[x] = [E] \quad (3.2.80)$$

then, putting the term in $s^3 \rightarrow 0$,

$$\{(MR / Bl)s + Bl\}[sX] = [E] \quad (3.2.80)$$

Therefore $T_s(S)$ can be obtained as

$$T_s = (MR) / (Bl)^2 \quad (3.2.81)$$

That is, when MR decreases and Bl increases we can perform quicker positioning in response to a step input voltage.

The block diagram for the complete control systems for a positioner is shown in Fig. 3.2.20.

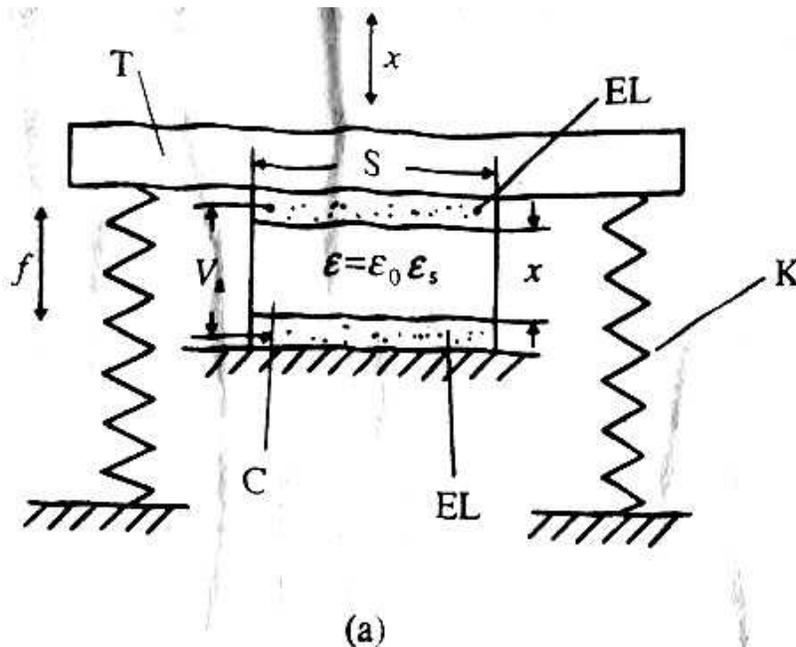
3.2.7 Electrostatic servo-positioner

Electrostatic sources utilize the force of attraction between charged capacitor plates. The force is independent of the sign of the voltage, so a large voltage is necessary for linear operation. Because the forces are relatively weak, a large area is needed to obtain significant output.

The attractive force can be deduced from the storage energy of a parallel flat-plate capacitor. The stored energy E_s (J) of a capacitor at an applied voltage V_a (V), using the SI unit system, is given by

$$E_s = CV_a^2 / 2$$

where the capacitance $C = \epsilon_0 \epsilon_s S / x$ (F) and ϵ_0 is the permittivity (dielectric constant) of vacuum = 8.854×10^{-12} F m⁻¹, ϵ_s the relative permittivity of the material, S the area of the flat plate electrode (m²), and x the distance between the electrodes (m).



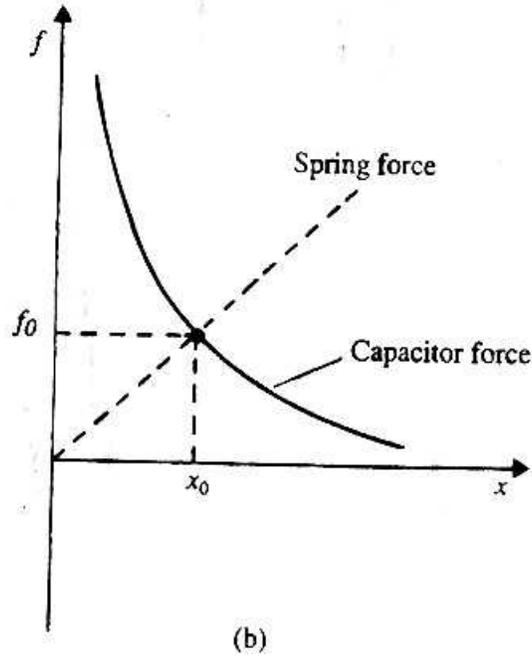


Fig. 3.2.21. Model of electrostatic positioner with backup spring, (a) Construction, (b) Balance of forces (see text). C: capacitor (gas) EL: electrode (applied potential V_a) T: moving table (displacement x) K: backup spring f : attractive force

The attractive force between the flat plates, f (N), is obtained as

$$f = \partial(CV_a^2 / 2) / \partial x = -\epsilon_0 \epsilon_s S V_a^2 / x^2 \quad (3.2.82)$$

If we assume a moving flat-plate electrode capacitor with backup spring of spring constant k (N m^{-1}) as shown in Fig. 3.2.21, then the stable balanced position x_0 of the moving plate at an applied voltage V_a is given by

$$f_0 = kx_0 = (\epsilon_0 \epsilon_s S V_a^2 / x_0^2) \quad (3.2.83)$$

That is,

$$x_0 = (\epsilon_0 \epsilon_s S V_a^2 / k)^{1/3} \quad (3.2.84)$$

Accordingly, $\delta x_0 = (\epsilon_0 \epsilon_s S / k)^{1/3} \cdot (2/3) V_a^{-1} \delta V_a$

Table 3.2.2 Solid-state positioner or actuator characteristics¹.

	Thermal expansion (metal)	magnetostriction (Ni)	piezoelectric ceramics (general electrostrictives, PZT)	Electrostrictive ceramics (limited electrostrictives, PMN)
Hysteresis	small	small	large	small
Response time	s	ns-s	ms	s
Driving power	heat	magnetic field	electric field	electric field
Ageing effect	small	small	large	small
Strain	$10^{-5} - 10^{-3}$	$10^{-5} - 10^{-3}$	$10^{-4} - 10^{-2}$	$10^{-9} - 10^{-3}$

^aFrom Piezoelectric actuators, by K.Uchino, Morikita, 1986

$$\text{so } \delta x_0 / x_0 = (2/3)(\delta V_a / V_a) \quad (3.2.85)$$

For example, if we take an air condenser with $S = 1 \times 10^{-4} \text{ m}^2$, $V_a = 500 \text{ V}$, and $k = 2 \times 10^3 \text{ nm}^{-1}$, then we get $x_0 = 48.0 \mu\text{m}$ and, for $\delta V_0 = 10 \text{ V}$, $\delta x_0 = 0.64 \mu\text{m}$. Therefore, we can perform fine position control with an electrostatic positioner, but it is necessary to apply a very high voltage. Recently an electrostatic servo-motor with specially arranged electrodes and synchronized exciting voltage has been developed, but it is not yet in practical use.

3.2.8 Piezoelectric and electrostrictive servo-positioners and servo-actuators

There are several solid-state positioners and actuators, as shown in Table 3.2.2. We discuss first those of piezoelectric and electrostrictive types.

Certain crystals produce an electric charge on their surface when placed in an electric field. Important piezoelectric crystals include quartz, ADP (ammonium dihydrogen phosphate), lithium sulphate, Rochelle salt, and tourmaline. Lithium sulphate and tourmaline are volume expanders, that is, their volume changes when subjected to an electric field in a suitable direction. Such crystals can detect hydrostatic pressure directly. Crystals which are not volume expanders must have one or more surfaces shielded from the pressure field in order to convert

the pressure to a uniaxial strain that can be detected. Tourmaline is relatively insensitive and is used primarily in high- Q ultrasonic transducers.

It has been found recently that certain piezoelectric ceramics consisting of very fine crystals, such as lead zirconate titanate (PZT), barium titanate, and lead metaniobate, become piezoelectric when suitably polarized. They are sometimes called general electro-strictive materials. They exhibit relatively high electromechanical coupling, are capable of producing very large forces, and are used extensively as sources and receivers for underwater sound. PZT and barium titanate have only a small volume sensitivity, hence they must have one or more surfaces shielded in order to detect sound efficiently.

Piezoelectric ceramics have extraordinarily high relative permittivity and hence high capacitance, and they are thus capable of driving long cables without preamplifiers.

In addition limited electrostrictive ceramic materials such as PMN (lead magnesium niobate, typically doped with $\sim 10\%$ lead titanate) have recently become important. This kind of limited electrostrictive material exhibits a strain which is a quadratic function of the applied electric field. However, unlike PZT, these limited electrostrictive materials are not reversible. That is, they change shape in an electric field but cannot generate an electric field when a strain is imposed by loading. Therefore they cannot be used as receivers. PZT has an inherently large hysteresis because of the domain nature of the polarization, but PMN does not.

Furthermore, it has been discovered that certain polymers, notably poly(vinylidene fluoride), are piezo-electric when stretched. Such piezoelectric polymers find use in directional microphones and ultrasonic hydrophones.

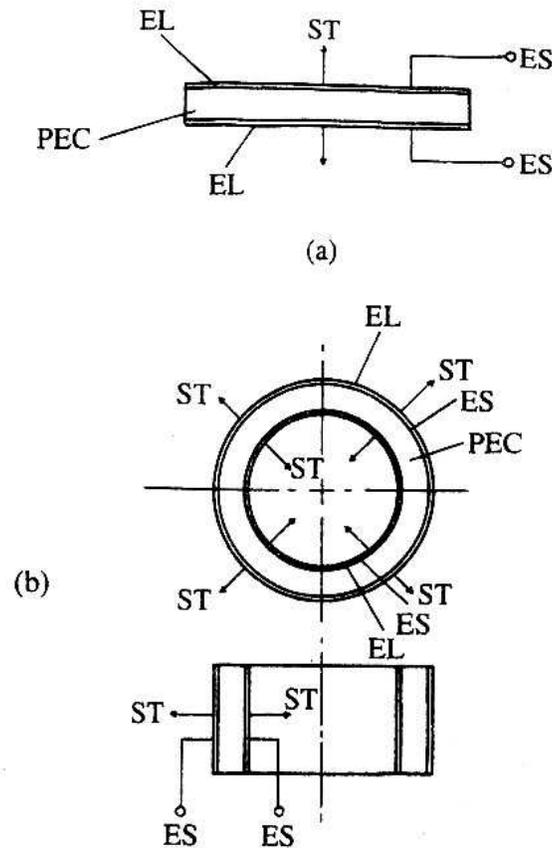


Fig. 3.2.22. Basic shapes of piezoelectric ceramic (electrostrictive) elements, (a) Circular plate, (b) Cylinder. ES: electric source ST: strain ($\delta l / l$) EL: electrode PEC: piezoelectric ceramic

In this section, however, piezoelectric ceramic servo- positioner units are the main topic. They are the most widely used in practice, because piezoelectric ceramics can be formed to suit any required object and can also undergo large deformation with a short response time. Data on piezoelectric materials are shown in Table 3.2.3.

The basic shapes of piezoelectric ceramic elements — circular plate and cylinder — are shown in Fig. 3.2.22. Figure 3.2.23 is a graph of the strain- electric field characteristic, showing that the strain is always positive, regardless of the direction of the electric field, and that there is considerable hysteresis. Accordingly, when a deformation of positive or negative sign about a certain point becomes necessary, biased polarization must be effected. Of course, the hysteresis cannot be removed, but there are several means of overcoming it such as introducing a suitable capacitor into the circuit.

In addition, there are three types of piezoelectric ceramic element, as shown in Fig. 3.2.24 and Table 3.2.4. The most important is the piled stack with laminated thin film elements; the single plate is mainly used in acoustic transducers of Langevin resonance type, and the bimorph bending plate is used in mechanical switching devices.

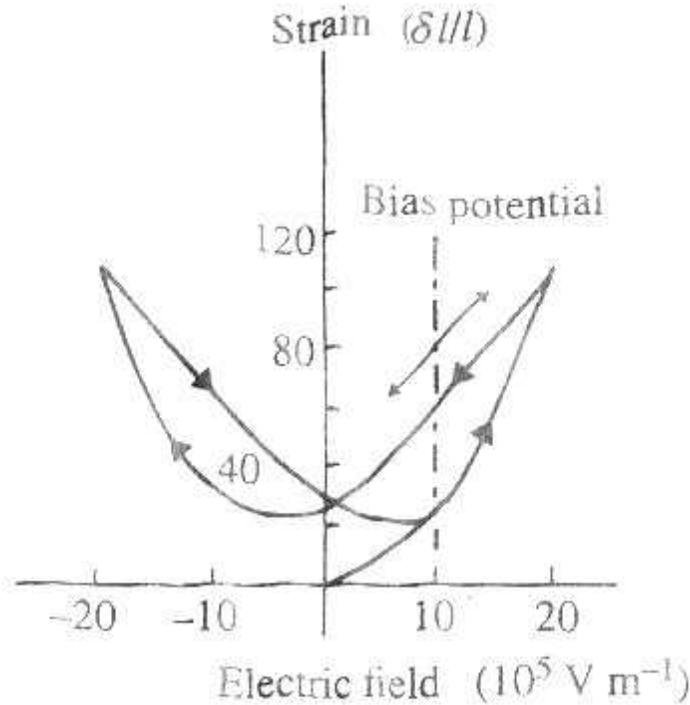


Fig. 3.2.23. Characteristic of piezoelectric ceramics.

The transfer function of an electrostrictive actuator is obtained as follows, referring to the piled stack form as shown in Fig. 3.2.25. The kinetic equation of the actuator is

$$M \left(\frac{d^2 u}{dt^2} \right) + p \left(\frac{du}{dt} \right) + (Ac/L)u = AcdE(N) \quad (3.2.86)$$

and the Laplace transform is

$$\left[Ms^2 + ps(Ac/L) \right] [u] = Acd[E] \quad (3.2.86)$$

where

$u(t)$ = displacement of the centre of equivalent mass (m)

M = equivalent mass of the unit (kg)

p = coefficient of viscous resistance ($N \text{ s m}^{-1}$)

A = sectional area of the unit (m^2)

c = stiffness of the actuator unit ($N \text{ m}^{-2}$)

L = total length of the actuator (m)

d = piezoelectric constant or modulus (strain/ electric field) ($m V^{-1}$, or $C N^{-1}$)

$E(t)$ = electric field ($V m^{-1}$).

Accordingly the transfer function $G(s)$ on $u(s)$ for $E(s)$ is

$$G(s) = [u] / [E] = (Acd) / (Ms^2 + ps + Ac / L) (m^2 V^{-1})$$

Therefore, for an applied unit step voltage $[E(s) = Vm^{-1}]$, the displacement $[u(s)]$ becomes

$$[u(s)] = [Acd / M\omega] \cdot \Omega / \{s + (p / 2M)^2 + \omega^2\}$$

Table 3.2.3 Characteristics of piezoelectric and electrostrictive materials

	Piezo electric ceramics		Piezo electric crystals				
	General electrostrictives	Limited electrostrictives					Unit (dimension)
Component		BaTiO ₃	PZT: 55% PbZrO ₃ 45% PbTiO ₃	PMN: 10% PbTiO ₃	ADP	Quartz SiO ₂ (X-cut)	
Compliance (Stiffness)	s 1/S	20	11			13.6	$[m^2 N^{-1}] \times 10^{-12}$
Relative permittivity	ϵ_s	1150	450		15.5	2.5	$[\epsilon / \epsilon_0] = [1]$
Electromechanical coupling factor	k_{31} k_{33}	0.18 0.48	0.20 0.50		0.28	0.09	[1] [1]
Piezoelectric constant (modulus)	d_{31} d_{33}	60 140-320	39 1000		- 300	$d_{11}=2.31$ $d_{14}=0.727$	$[mV^{-1} = cN^{-1}] \times 10$
Mechanical quality factor	Q	400	200			$10^3 - 10^6$	[1]
Curie point	J_c	120	290				[°C]

Q = holding energy stored per cycle divided by loss per cycle;

$$k^2 = \frac{(\text{energy stored mechanically})}{(\text{total energy stored electrically})} = d^2 / \epsilon_0 \epsilon_s S;$$

ϵ_0 = dielectric constant (electric permittivity) of vacuum = $10^7 / 4C^2 = 8.854 \times 10^{-12}$

k is not the efficiency of energy transfer, because the difference between mechanical and electrical energy stores in the electrical circuit component as the electrical energy; ADP = ammonium dihydrogen phosphate.

$$\text{Where } \omega^2 = (Ac / M) - (p / M)^2 / 4(s^{-2}) \quad (3.2.87)$$

Therefore we get the displacement formula as

$$u(t) = (Acd / (M\omega)) \cdot \exp[-(p / 2M) \cdot t] \cdot \sin \omega t \quad (3.2.88)$$

However, when the term in $s^2 \rightarrow 0$,

$$[u(s)] = (Acd) / (ps + Ac / L) \quad (3.2.89)$$

with the response time constant $T_s = pL / Ac(s)$, and when the terms in s^2 and $s^2 \rightarrow 0$, then

$$u(t) = (d / L)(m^2 v^{-1}) \times 1(Vm^{-1}), (m)$$

A numerical calculation for a PZT element quoted in K. Uchino's *Piezoelectric actuators* (Morikita) is as follows. With $\epsilon_0 = 8.854 \times 10^{-12} \text{ F m}^{-1}$, $\epsilon_s = 3400$, $d = 590 \times 10^{-12} \text{ m V}^{-1}$, and $c(\text{stiffness}) = 5 \times 10^{10} \text{ N m}^{-2}$, the electric field E generated at an applied stress $X = 3 \times 10^7 \text{ N m}^{-2}$ is obtained as follows:

$$P = dX = 590 \times 10^{-12} \times 3 \times 10^7 = 1.77 \times 10^{-2} \text{ Cm}^{-2}$$

$$E = P / \epsilon_0 \epsilon_s = 6 \times 10^5 \text{ Vm}^{-1}$$

Conversely, the strain x at an applied electric field $E = 10 \times 10^5 \text{ Vm}^{-1}$ is

$$x = dE = 590 \times 10^{-12} \times 10 \times 10^5 = 5.9 \times 10^{-4}$$

and in the clamped condition with strain x , then

$$X = xc = 5.9 \times 10^{-4} \times 5.0 \times 10^{10} = 3 \times 10^7 \text{ N M}^{-2}$$

The reason why a weaker electric field of $6 \times 10^5 \text{ V m}^{-1}$ occurs under the same stress condition than the directly applied electric field is that the electromechanical coupling factor k is 100%, given by

$$k^2 = d^2 c / \epsilon_0 \epsilon_s$$

Table 3.2.4 Calculated characteristics of piezoelectric ceramics or electrostrictive elements (prepolarized) as shown in Fig. 3.2.24^a

	Displacement t (/cm)	Generating force (N)	Transfer efficiency (%)	Resonance ! frequency (kHz)	Response time
Single plate	0.36	0.42	8.4	85	-
Piled stack	1.6-10	16-3	38-70	102	10 μ s
Bimorph plate	4.3-300	0.08-1	4.9-10	0.7	1 ms
Langevin transducer	20	-	Q = 200-400	20-40	-

^aNEPEC - 10, length 20 mm, width 5 mm, thickness 0.5 mm, $V_a = 100$ V

$= (590 \times 10^{-12})^2 \times 5 \times 10^{10} / (8.854 \times 10^{-12} \times 3400) = 0.58$, where k^2 is the ratio of mechanical energy stored in the element to input electrical energy.

Practical examples of piezoelectric ceramic servo- positioners are shown in Fig. 3.2.26. The $X—Y—\theta$ stage for ultra-precision positioning, Fig. 3.2.26a, is used to obtain a better production yield in the wafer processing and inspection of semiconductor production, where a positioning accuracy of $0.1 \mu\text{m}$ and a positioning resolution of 30 nm in the range $30 \times 30 \mu\text{m}^2$ is necessary because the design rule or smallest width of an LSI patterned circuit is $0.3 \mu\text{m}$., etc.

The ultrafine controller for mass flow in Fig. 3.2.26b consists of a very fine sensing unit for mass flow and a very fine valve controller consisting of a piezoelectric ceramics servo-positioner which is able to control flow rates of up to 20 ml min^{-1} .

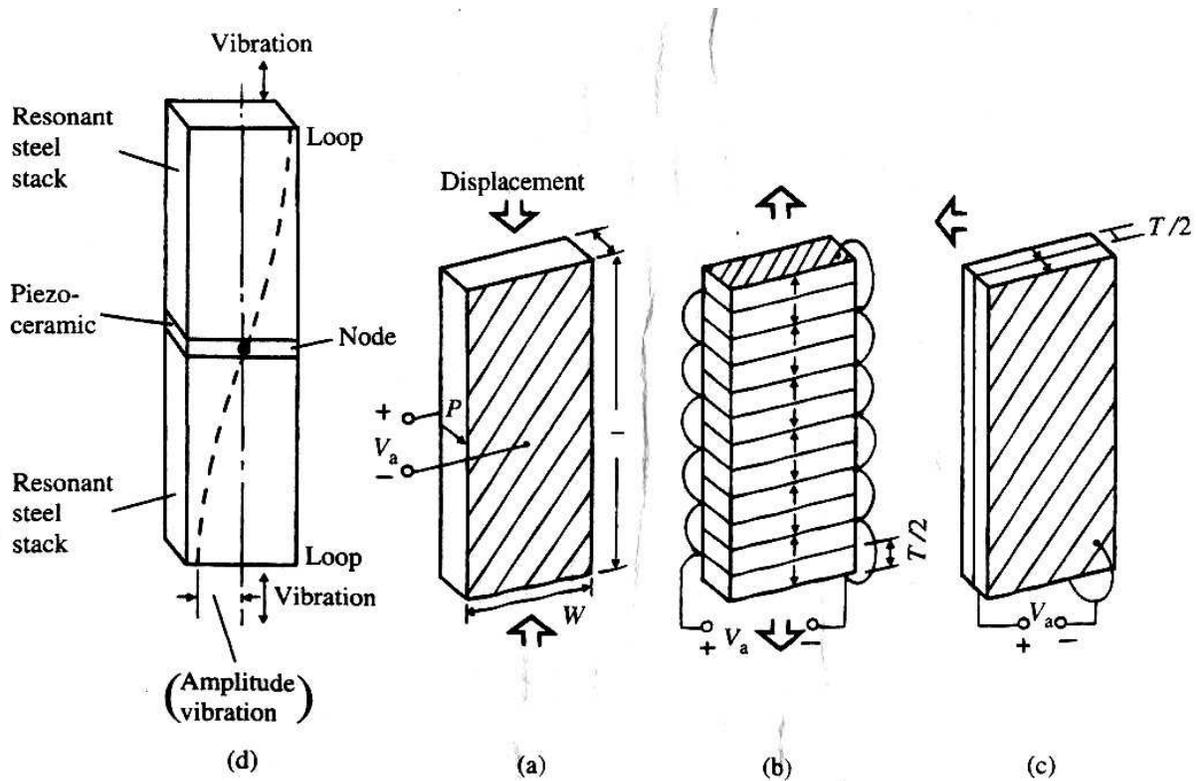


Fig. 3.2.24. Different types of piezoelectric (electrostrictive) element, (a) Single plate, (b) Piled stack, (c) Bimorph (bending) plate, (d) Langevin transducer.

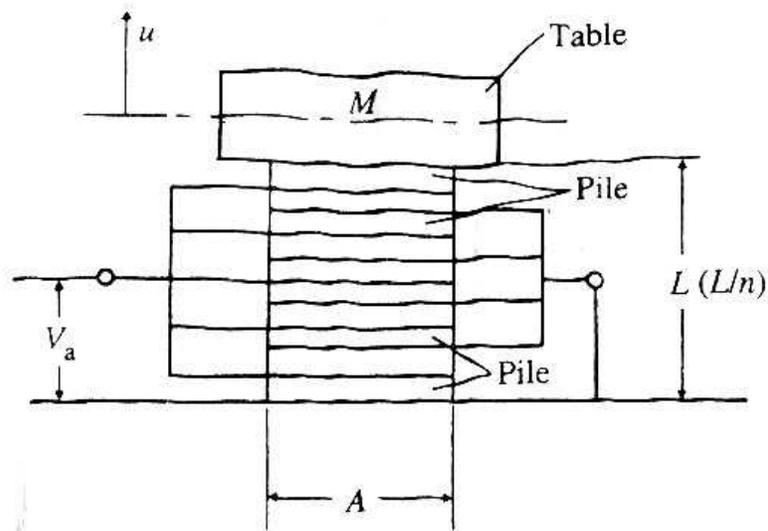
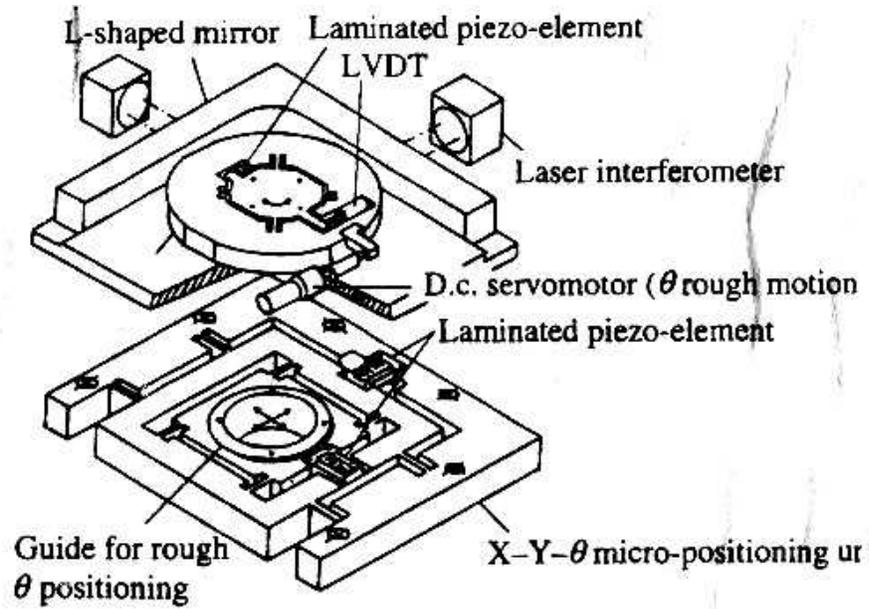
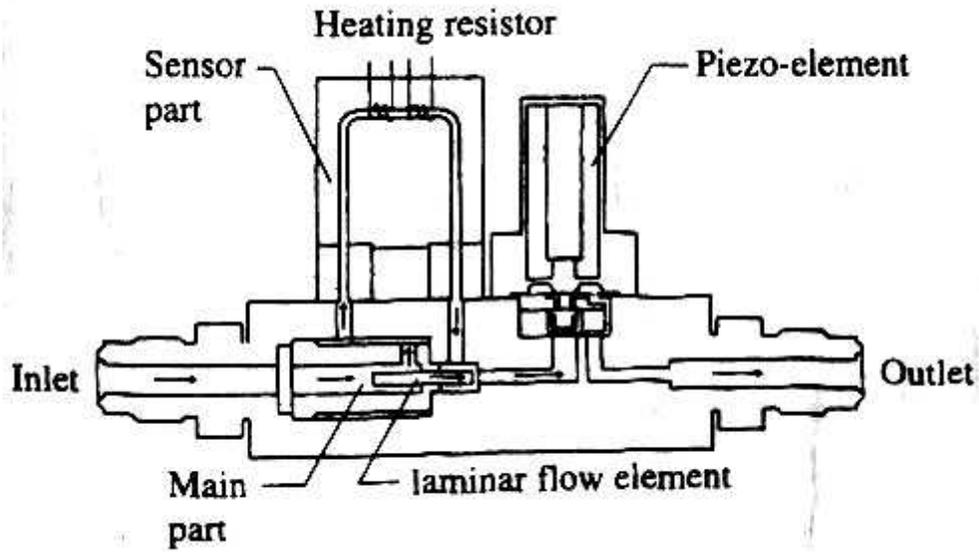


Fig. 3.2.25. Model of piled-stack piezoelectric ceramic actuator, u : displacement of table (m) M : equivalent mass (kg) V_a : applied excitation voltage per pile (V) A : cross-sectional area of stack (m^2) L : total stack length (m) n : number of piles



(a)



(b)

Fig. 3.2.26. Ultra-precision piezoelectric servo-positioner (from New technology for piezoelectric ceramics, Ohm-sha, Japan), (a) X-Y- θ stage for micro-positioning. (b) Mass flow controller.

3.2.9 Magnetostrictive servo-positioner and acoustic transducer systems

Some ferromagnetic materials become strained when subjected to a magnetic field. The relation between the static strain and magnetic field is shown in Fig. 3.2.27. As is easily shown, the effect is quadratic with respect to the field, so a biased field or d.c. bias current is required for linear operation. Important magnetostrictive materials and alloys include nickel and permendur (Co-Fe 50 : 50, but fragile). At one time, magnetostrictive transducers for acoustic devices were used extensively, but they have now been replaced by piezoelectric ceramic transducers.

Magnetostrictive transducers are rugged and reliable but inefficient and configurationally awkward. Recently it has been discovered that certain rare earth-iron alloys such as terbium-dysprosium-iron possess extremely large magnetostriction (as much as 100 times that of Ni). They have relatively low eddy current losses but require a large bias field, are fragile, and have yet to find significant applications. Metallic glasses have also recently been considered for magnetostrictive transducers.

In practice, the static strain due to magnetostriction is too small for direct application to machine tool positioning. Therefore an ultrasonic resonant vibrator as shown in Fig. 3.2.28 is used for positioning, in which the maximum amplitude of the vibrator end can be used as a stopper for solid tools. The positioning accuracy is inadequate, but the response is quite large.

Numerical data on the characteristics of several magnetostrictive materials are shown in Table 3.2.5.

3.2.10 Ultrasonic servomotor or positioner

Ultrasonic servomotors (Fig. 3.2.29) are realized as stationary-wave and travelling-wave systems.

The stationary-wave motor system is directly operated by ultrasonic motion of a piezoelectric oscillator or head of a Langevin-type resonator. A moving or rotating element with lining plate for friction is in contact with a driving element consisting of an elastic body and a piezoelectric ceramic plate for ultrasonic vibration. Figure 3.2.30a shows the principle of stationary-wave generation.

The travelling-wave motion is originated as shown in Fig. 3.2.29b and c, using two layers of piezoelectric plates. The moving element is driven by the travelling wave through the lining plate on the elastic body.

Block diagrams for control of an ultrasonic motor or positioner are shown in Fig. 3.2.31. Of course the positioning accuracy is not very high, because the friction transmission mechanisms are interlinked.

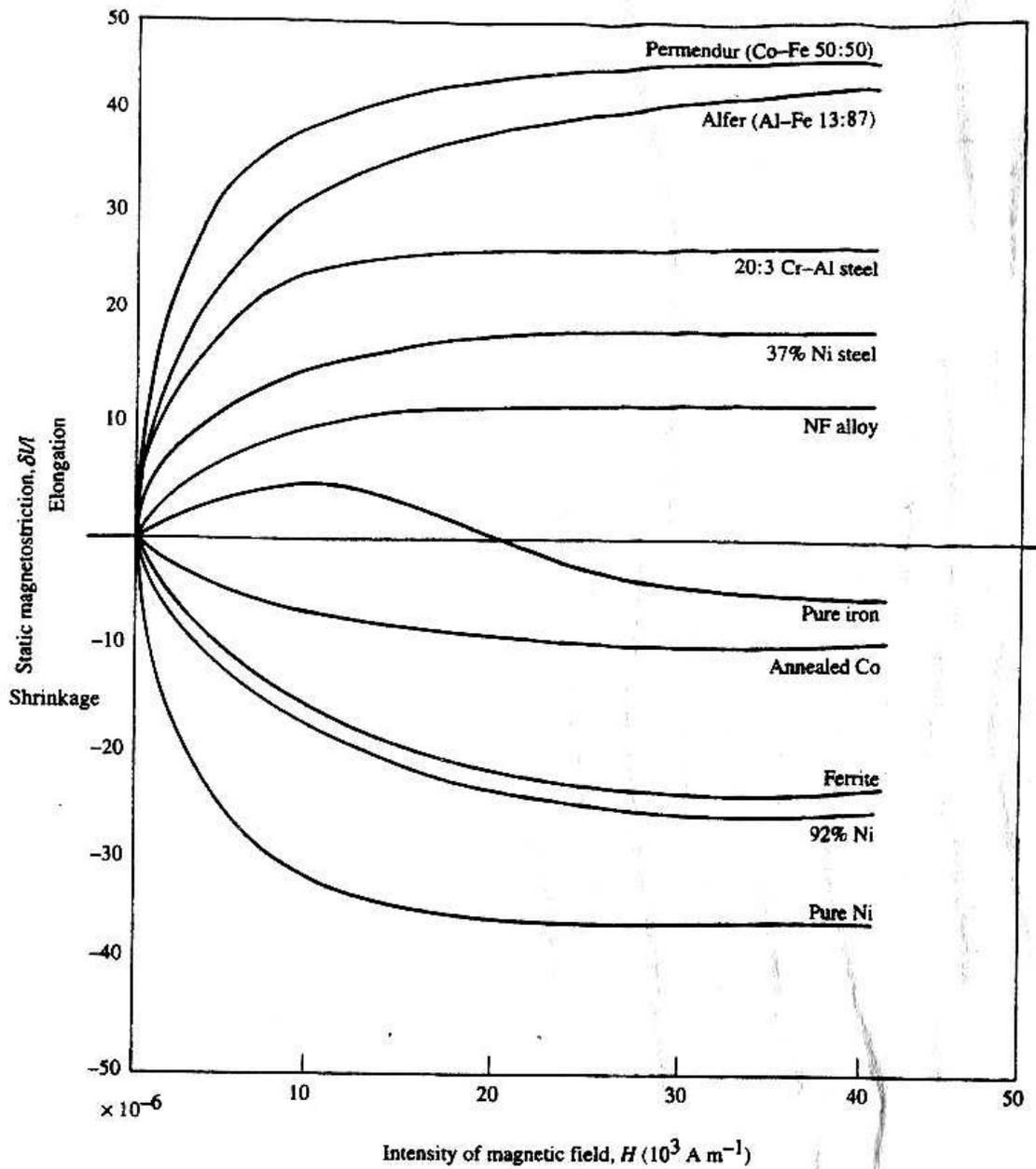


Fig. 3.2.27. Magnetostriction characteristics of metals and alloys.

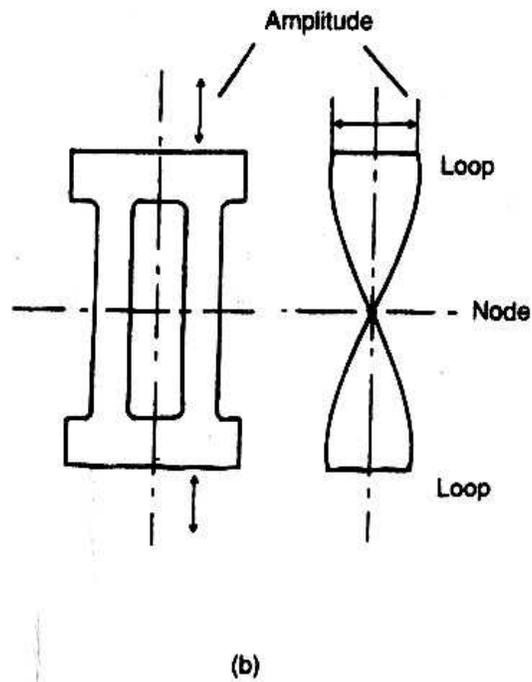
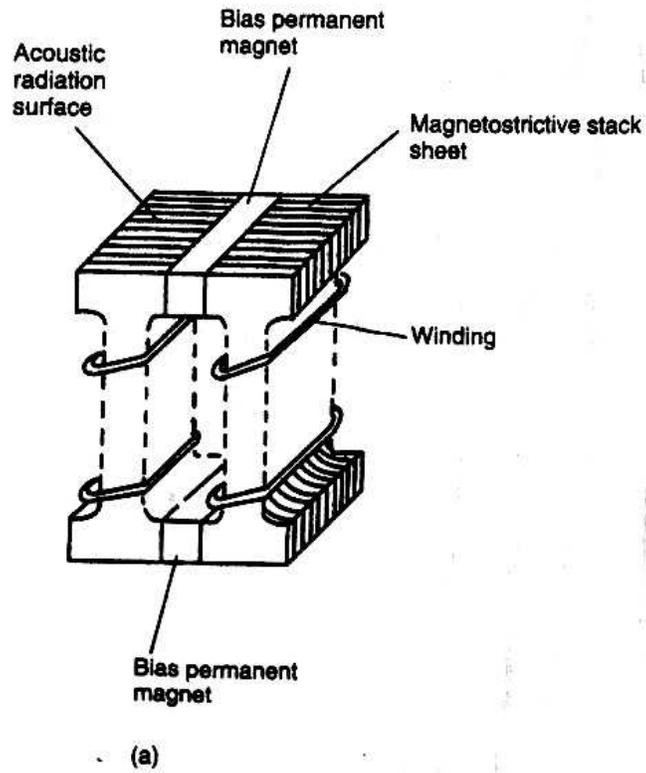


Fig. 3.2.28. Magnetostrictive transducer for ultrasonic vibration, (a) Vibrator, (b) Resonance mode.

Table 3.2.5 Characteristics of several magneto-strictive materials

	Nickel	Alfer	Ferrite
	> 98% Ni	Fe 87% Al 13%	(Ni, Co) Fe ₂ O ₃
Relative permeability, μ_s^0	40	190	20
Resistivity, p (Ω m)	70	910	> 4 x 10 ⁹
Electromechanical coupling factor, k (%)	20-30	20	22
Saturated static magnetostriction, $\delta l / l$	-4 x 10 ⁻⁶	+ 35 x 10 ⁻⁶	-30 x 10 ⁻⁶
Best bias magnetic field, H_b (10 ² A m ⁻¹)	8-12	5-8	8~12
Density (10 ³ kg m ⁻³)	8.9	6.7	5.0
Acoustic speed, v_s (m s ⁻¹)	4800	4700	5700
Elastic constant (stiffness) (GPa)	200	140	0.4 ^b

a magnetic permeability $\mu = \mu_0 \mu_s$ (H m⁻¹), where μ_0 is the permeability of (vacuum) = $4\pi \times 10^{-7}$ H m⁻¹

^b Fracture stress

3.2.11 Other types

(a) A thermal expansion actuator is available for small-range positioning. The actuator is composed of an expanding solid body and a heating and cooling unit, but as is easily seen, the response speed is very low and unstable. However, the force for positioning a tool can be quite large in very fine dimensional motion, so the system is used for position control of grinding wheel heads.

(b) The shape-memory alloy actuator, consisting of Ti - Ni alloy, is used for position control where a large output force is needed, using heating systems.

(c) An anisotropic polymer actuator with a piezoelectric ceramic stack has been developed recently. Special polymers undergo large volume deformation only through their piezoelectric effect.

(d) A metallic hydrogen compound actuator, e.g. consisting of La - Ni alloy, can be used as a result of its large deformation on heating and cooling, with quick response time.

(e) The mechanochemical characteristics of electrolytic polymers can be used for an actuator, as a result of strain variation caused by pH change.

(0 Photomechanical materials such as azobenzene synthetic polymers change their volume upon absorption of photon energy. They have potential for development of control actuators of high sensitivity.

References

1. The electrical engineering handbook. CRC Press Section 5 Electrical effects and devices.
2. Mechatro actuator, NSK (Nippon Seiko Co). Japanese
3. Mechanism of mechatronics, K. Itao, Corona-sha, Japanese
4. Piezo electric actuator, K. Uchino, Morikita, Japanese

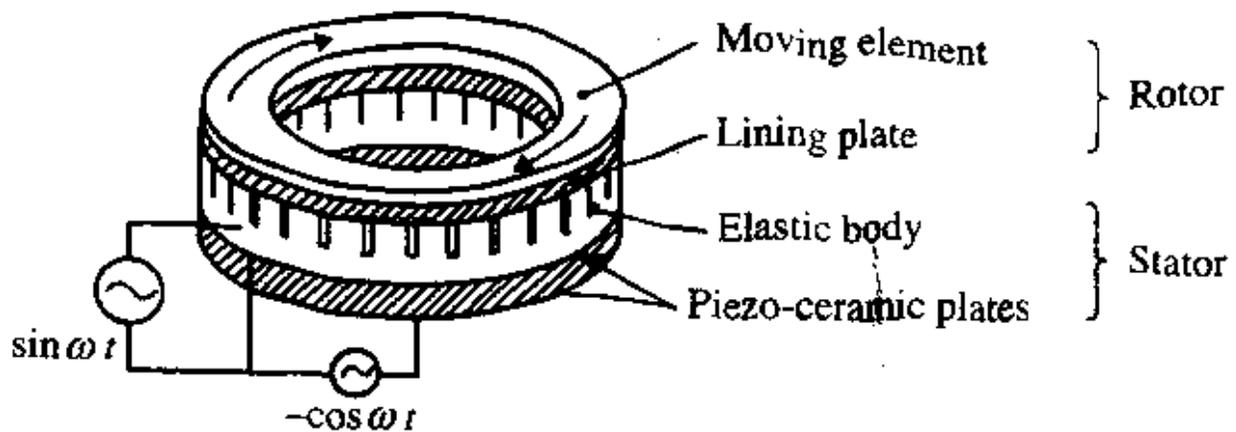
3.3 Computer-aided digital ultra-precision position control

3.3.1 Digital servo systems

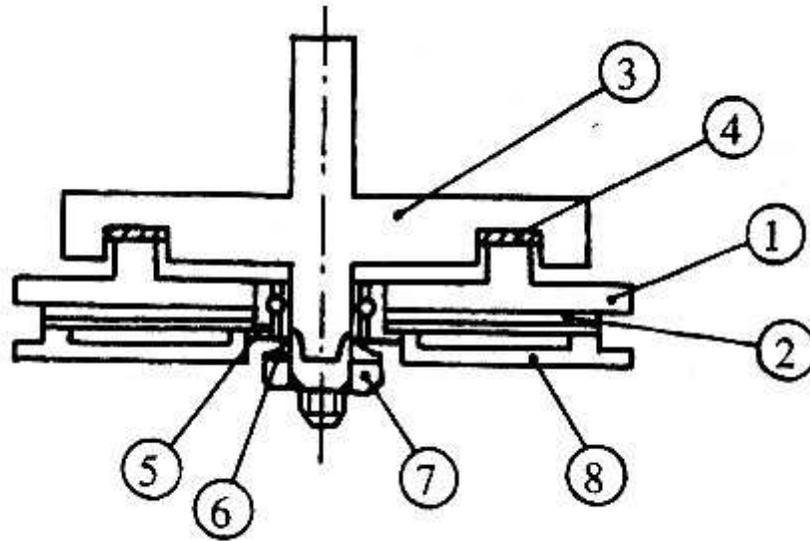
(a) Numerical control systems with nanometre resolution

In this section, the basic requirements for a numerical control system that achieves nanometre resolution are described.

Figure 3.3.1 shows a block diagram for a numerical control system. Processing and machining data are stored in the processing data memory (or NC tape) as G-code data. The G-code data are transmitted to the G-code interpreter, where they are interpreted in real time. The interpreter then generates the motion data and sequence data from the G-code data series. The path (motion) generator processes these motion data and generates position references for the servo system, which, with a motor, drives the machine with force and in turn is affected by reaction forces from the machine. The machine's positions and velocities are detected by motion sensors, which feed these signals back to the servo system. Meanwhile the sequence controller receives sequence data generated by the G-code interpreter to achieve synchronization between the motion references and machine states as detected by the switches and sensors. In the design of a numerical control system with nanometre resolution, several key points must be observed:



(a)



(b)

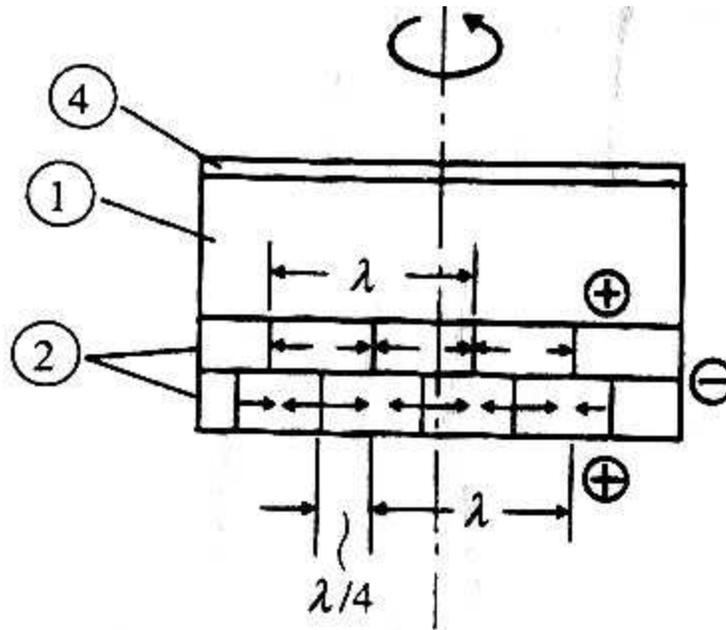


Fig. 3.2.29. Construction of ultrasonic servomotor, (a) General view, (b) Cross-section: 1, elastic body; 2, driving body with two piezoelectric layers; 3, rotating or moving element; 4, lining plate; 5, bearing; 6, plate spring; 7, nut; 8, base.

1. Minimum resolution: 1 nm.
2. Maximum position data: assuming that a 100 m maximum radius can be designated, the maximum number of digital position data is

$$(100m)/(1nm) = 10^{11} \approx 2^{37}$$

requiring > 40 bits.

3. Maximum velocity data: assuming a maximum velocity of 24 m min^{-1} ($= 400 \text{ mm s}^{-1}$), the maximum digital velocity is

$$(400 \text{ mm s}^{-1})/(1 \text{ nm}) = 4 \times 10^8 \text{ Hz} = 400 \text{ MHz}$$

A 400 MHz maximum velocity is a very large value, making it difficult to transmit serially the position pulse signal from the position sensors. If the sampling time for velocity control is assumed to be 1 ms, the incremental displacement in one sampling interval is

$$400 \text{ MHz} \times 1 \text{ ms} = 4 \times 10^5 \approx 2^{19}$$

For a 0.1 ms sampling time, this is

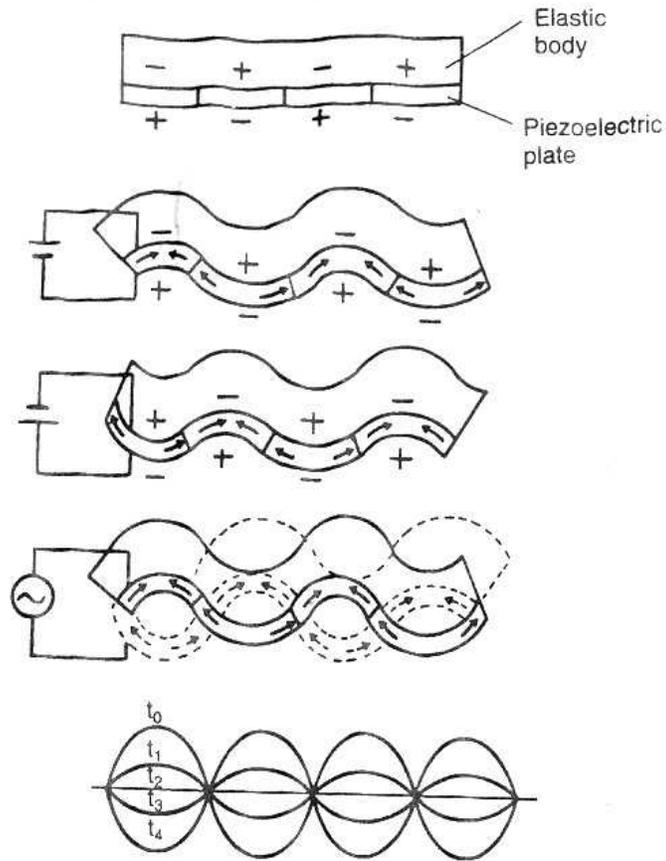
$$400 \text{ MHz} \times 0.1 \text{ ms} = 4 \times 10^4 \approx 2^{16}$$

4. Maximum linear interpolation length for arc: in Fig. 3.3.2, the relation between linear interpolation length L , arc radius R and interpolation error δ is

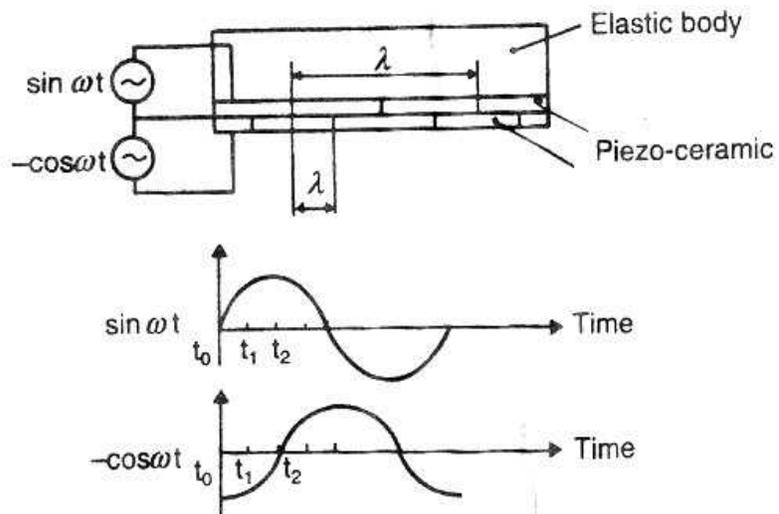
$$R^2 = (R - \delta)^2 + (L/2)^2$$

And

$$\delta \approx (1/2R)(L/2)^2$$



(a)



(b)

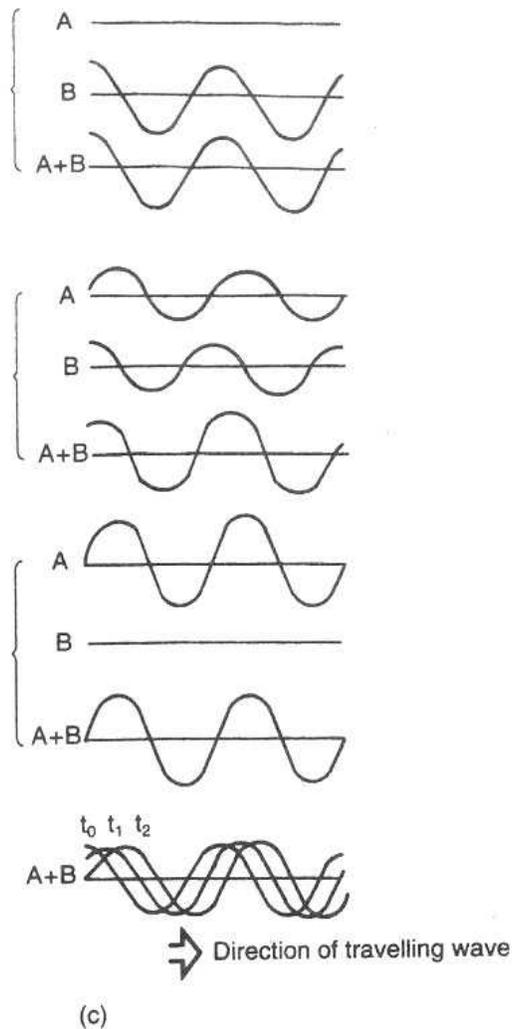


Fig. 3.2.30. Principle of ultrasonic servomotor, (a) Stationary-wave generation, (b) Travelling-wave source, (c) Travelling-wave generation.

Assuming $R = 1 \text{ mm}$ and $\delta < 1 \text{ nm}$, the maximum interpolation length is

$$L < 3\mu\text{m}$$

5. Maximum interpolation rate of G-code blocks: assuming a series of $3 \mu\text{m}$ linear segments and A maximum contouring control velocity of 180 mm min^{-1} ($= 3 \text{ mm s}^{-1}$), the maximum interpolation rate of the G-code blocks is

$$(2\text{mms}^{-1}) / (3\mu\text{m}) \square 1\text{kHz}$$

In other words, one G-code block must be interpreted in <1 ms. This means that NC systems with

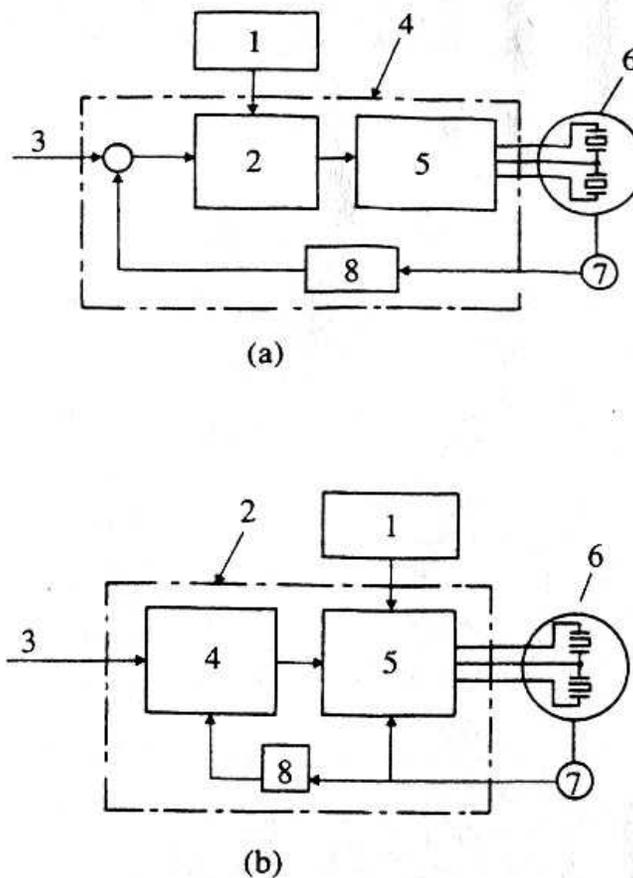


Fig. 3.2.31. Block diagram of ultrasonic servomotor control, (a) For speed control: 1, d.c. source; 2, d.c. voltage control; 3, speed command; 4, speed control circuit; 5, power amplifier; 6, ultrasonic motor; 7, encoder; 8, feedback circuit, (b) For positioning: 1, d.c. source; 2, positioning circuit; 3, position command; 4, positioning unit; 5, speed controller; ultrasonic motor; 7, encoder; 8, feedback circuit; 9, speed signal.

nanometre resolution need a very high data-processing capacity. This is achieved by using ultrahigh-speed microprocessors or parallel data processors.

(b) Fully closed loop control including machine dynamics

There are two servo control methods for a precise NC system: semi-closed and fully closed loop controls: Fig. 3.3.3(a) and (b).

In semi-closed loop control, the motor's position, velocity and current are fed back to the servo-controller, as shown in Fig. 3.3.3(a). The motor drives the machine through a mechanism such as a gear and ball screw. Even if the motor's position can be controlled without error, errors will exist in the machine because of friction and lost motion in the drive mechanism. Semi-closed loop control is therefore not suitable for numerical control systems that need to achieve nanometre resolution. In semi-closed loop control, however, stability in control and good operability are relatively easy to achieve.

In fully closed loop control, on the other hand, the machine's position is directly fed back to the servo-controller, as shown in Fig. 3.3.3(b). Fully closed loop control therefore provides a better control performance than semi-closed loop control. It is important to note, however, that fully closed loop control, which includes machine dynamics such as friction, backlash, and oscillatory characteristics, is prone to control instabilities. Since these machine dynamics are not negligible in nanometre-precision machine control, they must be compensated in various ways. Since backlash cannot be easily compensated, the machine must be designed to avoid backlash by using such methods as the application of a preload. A preload will increase the friction force. The friction force can be compensated by a disturbance observer, which is described later in this section. Oscillatory characteristics can be damped by acceleration feedback. Hence in the design of a fully closed loop control system, the machine dynamics must be precisely measured and taken into account.

(c) A.c. servo system by DSP control

There are two types of servomotors: d.c. and a.c. A d.c. servomotor has a commutator and brushes. Contact between these elements generates a friction force which varies with the motor's rotational speed. An a.c. servomotor, on the other hand, has no contacts and no resulting friction. With similar control performances, therefore, an a.c. servomotor is more suitable to drive a nanometre-resolution machine than is a d.c. servomotor. A brushless d.c. servomotor can be considered as a type of synchronous a.c. servomotor.

By using a DSP (digital signal processor), the servo control performance of an a.c. servomotor can be digitally adjusted to the same level as that of a d.c. servomotor. The DSP is a high-speed microprocessor specially designed for digital signal processing, such as data processing by control algorithms.

Figure 3.3.4 shows a block diagram of an a.c. servo control system. The motor is a synchronous, moving-magnet-type a.c. servomotor. The motor's rotor has several pairs of permanent magnets. The stator has three phase windings, U, V, and W, and the current is controlled so that the current vector is perpendicular to the motor angle, regardless of the rotor's rotation, enabling the motor to generate maximum torque. The rotor's rotational velocity, which is detected by a pulse generator (PG), is used to control the current phase and motor angle. The servo-controller generates the torque reference so that the motor angle will match the motor angle reference. The motor's electrical angles $\sin(n\theta)$ and $\sin[n\theta + (3/2)\pi]$ are calculated from the motor angle θ , where n is the motor's pole number.

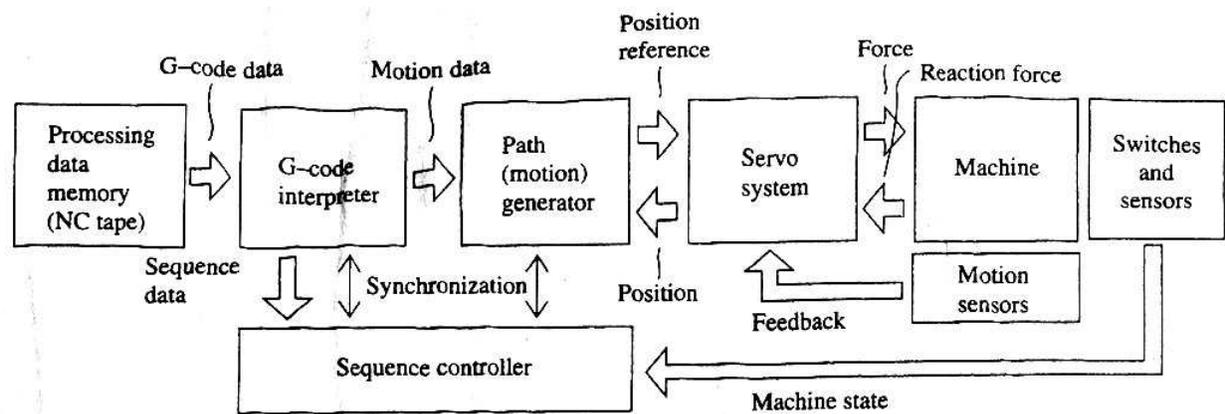


Fig. 3.3.1. Block diagram of numerical control system.

The products of the torque reference and two electrical angles provide the current references for the U- and V-phase windings, which when input to the current amplifiers drive the respective phase currents.

The a.c. servo system has the following features:

- (1) absence of contact, so no need for maintenance
- (2) a higher power rate than a d.c. servomotor, resulting in quicker response
- (3) a higher rotational velocity than a d.c. servomotor.

(d) Servo control with feedforward compensation and disturbance observer

To achieve high-performance servo control with nanometre precision, feedforward compensation and disturbance compensation by an observer are effective methods.

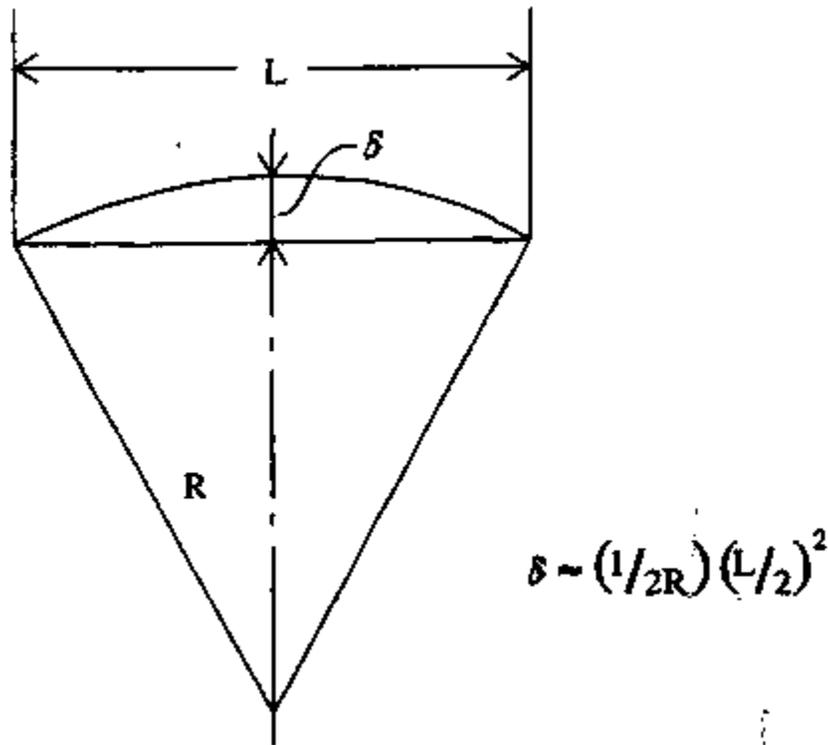


Fig. 3.3.2. Linear interpolation error for arc trajectory.

Figure 3.3.5 shows a servo control system with feedforward compensation and a disturbance observer. In this system, the motor angle is governed by proportional control and the motor's angular velocity by proportional — integral control. The dominant closed-loop dynamics of this control scheme can be approximated by a second-order lag system with response time τ_r and damping ratio ζ . With such control alone, response errors due to response delays and load disturbances are inevitable.

To remove errors due to response delays, therefore, feedforward compensations for acceleration and velocity are added. The gains of the feedforward compensations are set at τ_r^2 for acceleration compensation and $2\zeta\tau_r$ for velocity compensation. These gain settings compensate for response delays in the original servo control, making quick and fine control possible. The response error can be expected to be reduced to less than one-tenth of the original amount by this method.

Next, systematic errors due to disturbance torque are compensated by adding a disturbance observer. Using signals on the motor's angular velocity and motor current, the observer estimates

the load torque. The estimated torque is then subtracted from the motor current reference to achieve load torque compensation. Compensation by the disturbance observer makes it possible to achieve fine control while the system remains insensitive to unavoidable load changes present in precise servo control systems with nanometre resolution.

3.3.2 Software servo systems

Previous servo systems used for industrial robots or precision positioning equipment were inflexible, as shown in Fig. 3.3.6⁽¹⁾ where neither the position loop

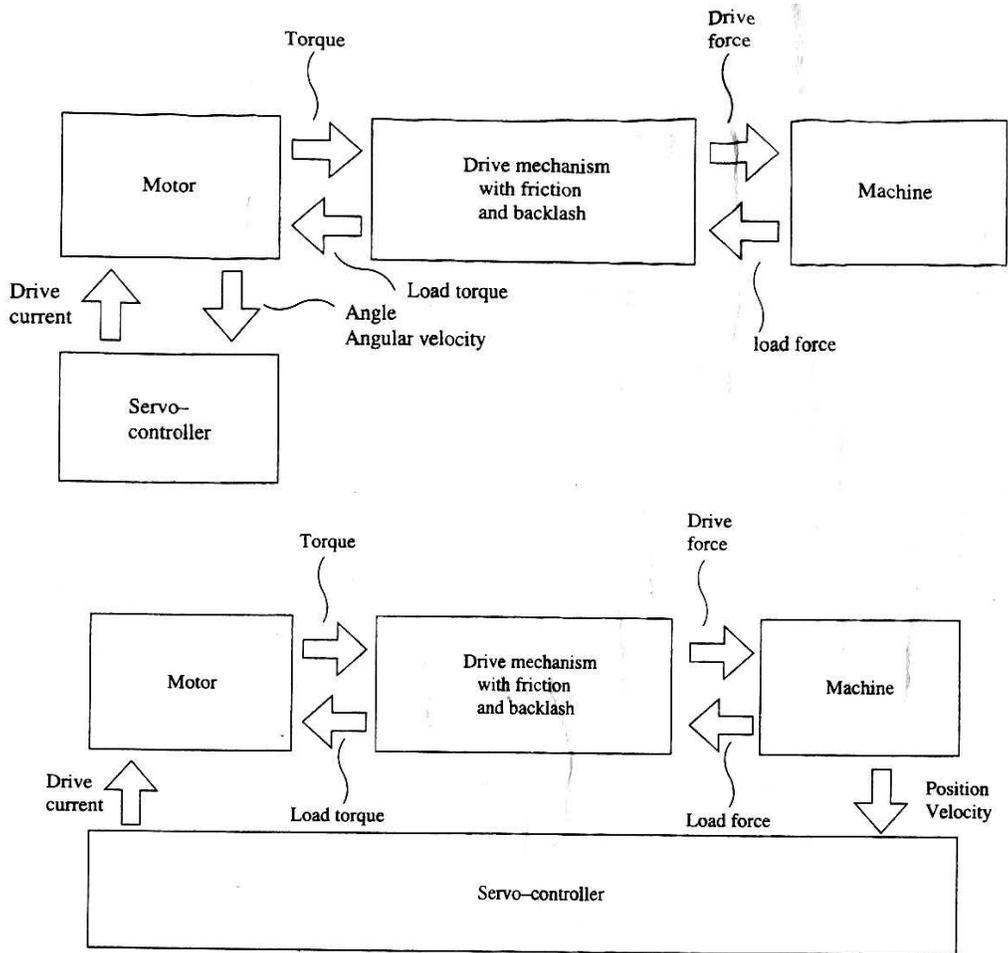


Fig 3.3.3. Loop control: (a) semi-closed; (b) fully closed.

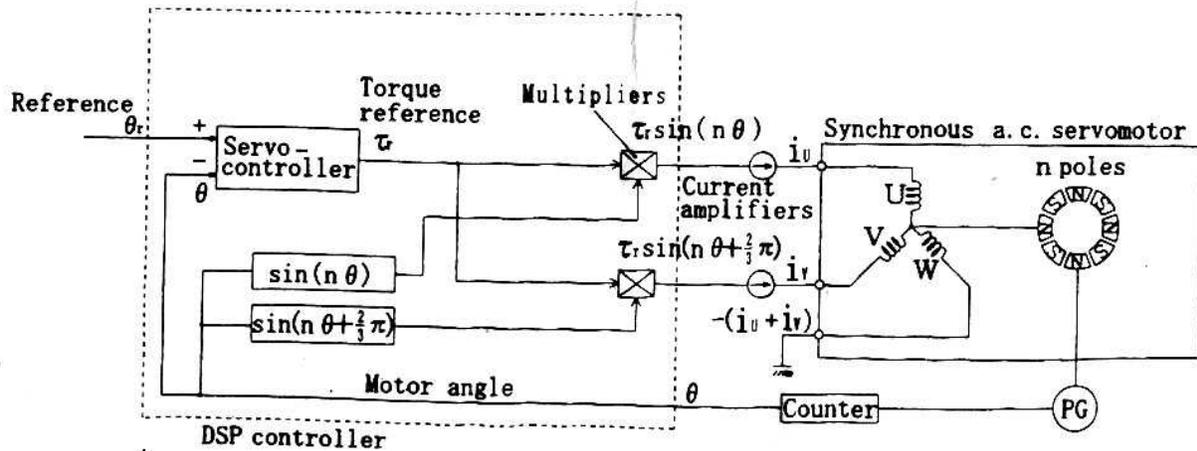


Fig. 3.3.4. A.c. servo system with digital signal processor (DSP) control.

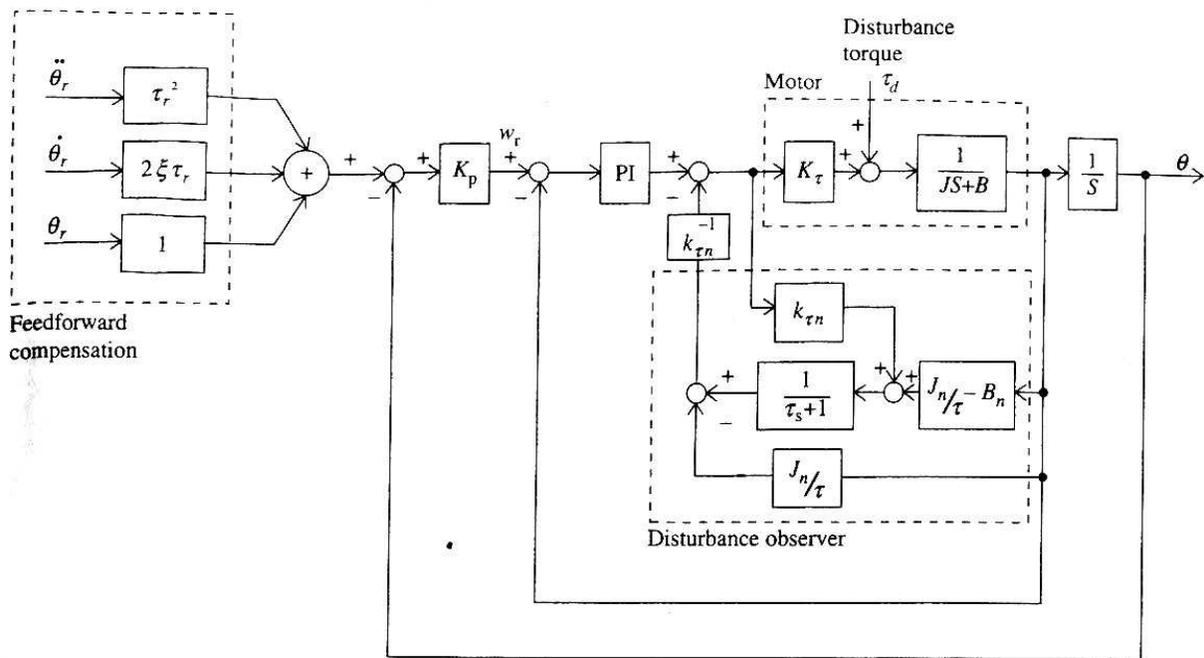


Fig. 3.3.5. Servo control system with feedforward compensation and disturbance observer.

gain or velocity loop gain could be changed after the control system was constructed. This was because the computer was outside the servo system.

However, to attain faster and more precise positioning, it is necessary to be able to change parameter gains and/or control modes, because the inertial moment, say, in an industrial robot changes according to its arm movement. With recent rapid developments in personal computers,

software servo systems such as shown in Fig. 3.3.7⁽²⁾ have become possible. Here, control modes as well as some loop gains can easily be changed according to the operational conditions.

In this section we introduce an example of a rapid positioning mechanism with nanometre accuracy which uses a software servo system developed by the author.

(a) ARV system

In the point-to-point positioning method, a table or stage which must be positioned rapidly is repeatedly subjected to high acceleration and deceleration; the resulting reaction force of the inertia of the table causes residual vibrations in the equipment, increasing the settling time.

we devised a mechanism called an anti-residual- vibration (ARV) system, shown in Fig. 3.3.8, which uses a software servo system to prevent residual vibrations. The table displacement y , measured by a laser instrument, is fed back to the d.c. servomotor. When the table supported on the linear ball guideway is accelerated to the right, as in the figure, the sub-table supported on another linear ball guide begins to move to the left. The sub-table is eventually stopped by the friction force f_h , of the linear ball guideway. The sub table thus absorbs the reaction force of the table inertia and prevents residual vibrations.

(b) Experimental device

The experimental device is schematically depicted in Fig. 3.3.9(a) and its block diagram in Fig. 3.3.9(b). A command from the computer rotates the d.c. servomotor. When the servomotor rotates the screw shaft by an angle θ , the axial force P moves the table by displacement y , which is then measured by the laser instrument (with a resolution of 2.49 nm). The difference Δx between y and the target position x is then registered by the computer and sent to the servomotor. As reported in our previous paper⁽²⁾, the reaction force of the table's inertia causes vibrations in the base centre and also the floating unit, which is positioned between the table and ball nut to absorb the radial runout of the screw shaft.

These vibrations result in a long settling time of 1s

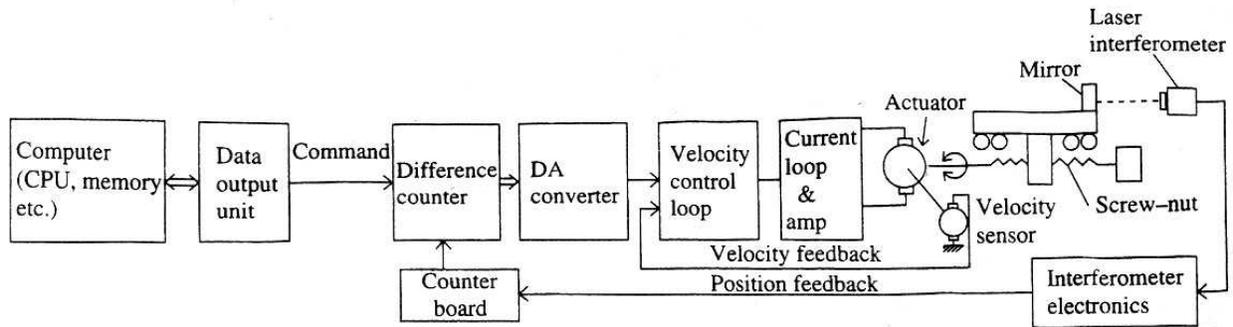


Fig. 3.3.6. Conventional digital software servo system.

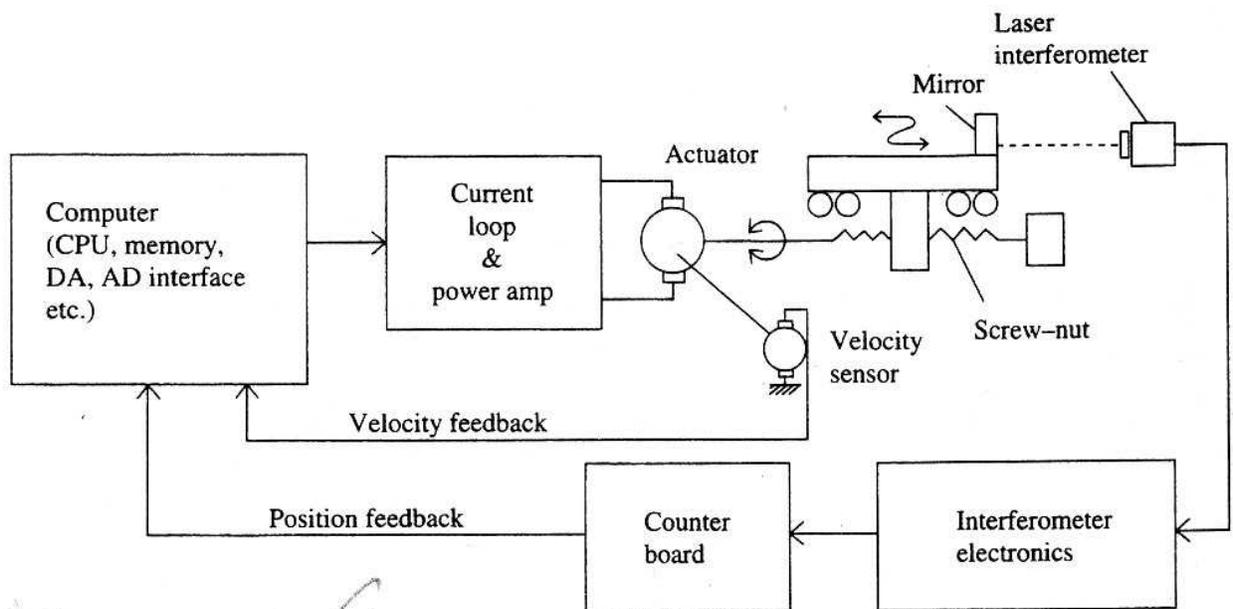


Fig. 3.3.7. Recent digital software servo system.

(i.e. the time required for the table to arrive within 10 nm of the target position). To reduce the settling time and achieve nanometre positioning accuracy, the ARV system mentioned above is used. The software servo system is needed since the ARV creates different dynamic characteristics in the table-guideway-ball- screw apparatus for long and short table travel distances.

(c) Non-linear elastic characteristic

Without the ARV system, when the motor torque is increased linearly, the table displacement y also increases, as shown in Fig. 3.3.10(a) and (b). When the table reaches $x_a = 0.3$ or $0.6 \mu\text{m}$ (at

$t = t_1$ or t_2 respectively), the motor torque is reduced causing the table displacement y to decrease after overshooting by a short distance. Fig. 3.3.10(c) shows the relation between motor torque τ and table displacement y , derived from (a) and (b) and plotted at 1 ms intervals. This non-linear elastic characteristic is caused by elastic interactions between the balls and races in the ball screw and linear ball guide. For $x_a = 0.6 \mu\text{m}$, the balls seem to be rolling between points E (where $y = 0.4 \mu\text{m}$) and F in (b) and (c). Todd and Johnson® have made an in-depth study of this non-linear elastic characteristic in the $0.4 \mu\text{m}$ range, and Futami et al.⁽⁴⁾ have reported that nanometre accuracy was obtained by using this characteristic with position feedback to the linear motor.

(d) Frequency-response test

Figure 3.3.11 shows the results of a frequency-response test using the servo-amplifier voltage V_c as the input and table displacement y as the output: (a)

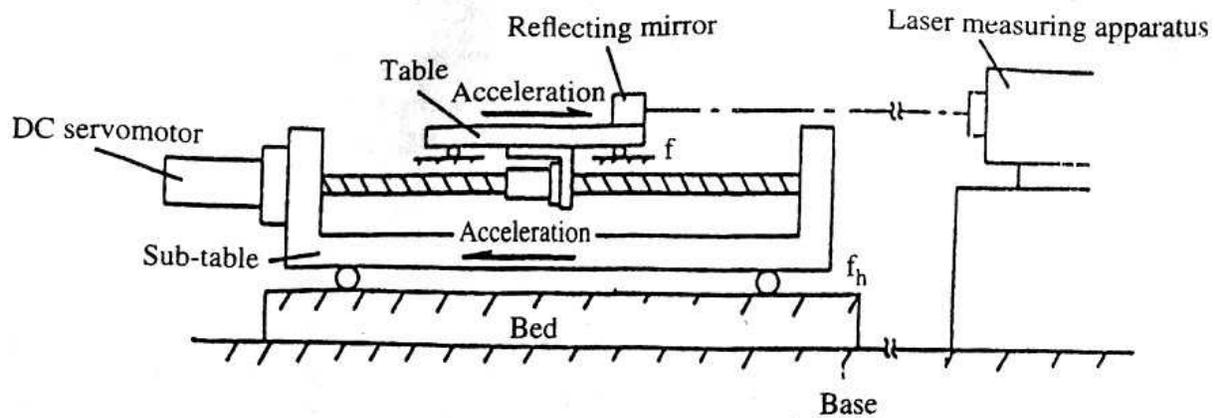


Fig. 3.3.8. Anti-residual vibration (ARV) system.

shows the result when the output amplitude is $> 5 \mu\text{m}$ (out of the non-linear elastic characteristic range), while (b) shows the result when the output amplitude is 50 nm (within the non-linear range). The different characteristics in the two cases explain why the loop gains (proportional gain k_p and integral gain k_i) should be changed when the table displacement moves from the rough positioning for the long travel range into the short travel range. The linear ball guide for the sub-table also has a similar non-linear elastic characteristic.

(e) Positioning experiment

1. Without the ARV system. The following control procedure with a software servo system was used:

- i. For rough positioning, the table is moved by PI control.
- ii. When the table comes within 50 m of the target position, the proportional gain K_p is increased to twice its previous value. This brings the table very close to the target position.
- iii. In the vicinity of the target position where the non-linear elastic characteristic exists, after letting $k_p = 0$ the integral gain k_i is reduced to half its previous value.

The following results were obtained after 10 mm stroke step responses were repeated 20 times: mean value of positioning error E 1 second after table start, $E = 3.7$ nm, and standard deviation of E , $\sigma = 1.8$ nm.

Figure 3.3.12(a) shows the table displacement curve in the vicinity of the target position. We see that the settling time T_s , for the table to come within ± 10 nm of the target position, is 630 ms. In

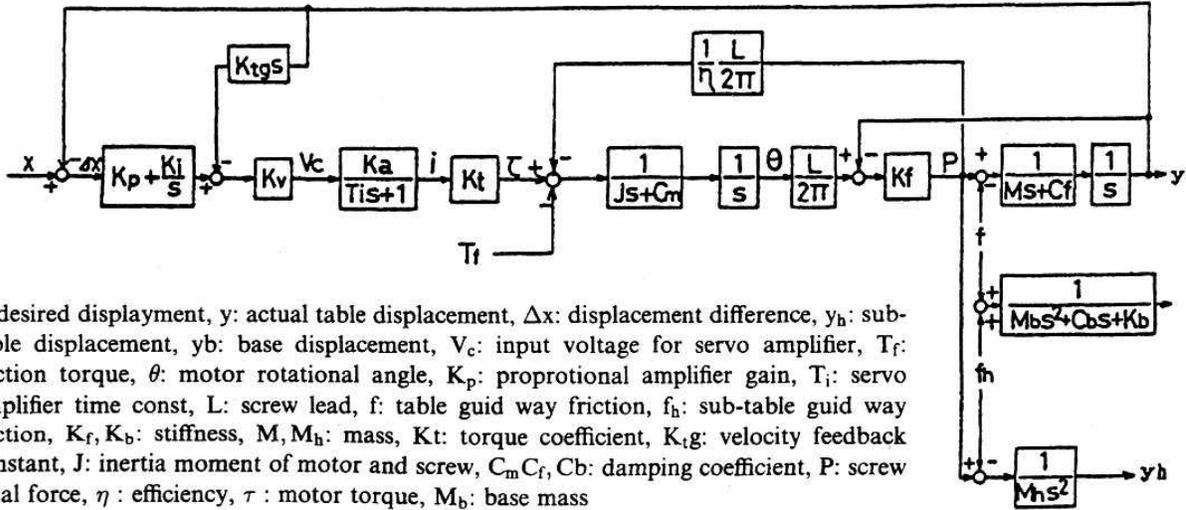
Fig. 3.3.13(a), (i) shows the vertical displacement of the cast-iron base centre, y_b' measured by a capacitance displacement sensor fixed at the floor (see Fig. 3.3.9(a)), and in Fig. 3.3.13(b), (i) shows the results of FFT (fast Fourier transform) analysis, where $f = 17$ Hz represents the vibration of the cast-iron base, $f = 30$ Hz is the structure's resonant frequency, and $f = 50$ Hz is due to the twisting vibration of the base.

2. With the ARV system. The table is moved by the same PI control as before (i.e. without ARV). Figure 3.3.12(b) shows that when approaching the target position, the table displays a vibrational amplitude of over $0.4 \mu\text{m}$ between $t = 100$ ms and $t = 300$ ms, just as in the absence of ARV. This vibration causes the sub-table to move in an unstable manner. Moreover, determination of k_p and k_i is made difficult because of changes in the friction values f and f_h .

To solve this problem, pulse width control (PWC) was used, in which a current pulse (pulse width Δt and height h) is supplied to the servomotor. This is shown in Fig. 3.3.14, which also shows the table displacement y_d . Figure 3.3.15 shows the relation between the pulse width Δt and table displacement y_d after a 15 ms lapse; the dispersion (σ : standard deviation) exists because of the change in friction f .

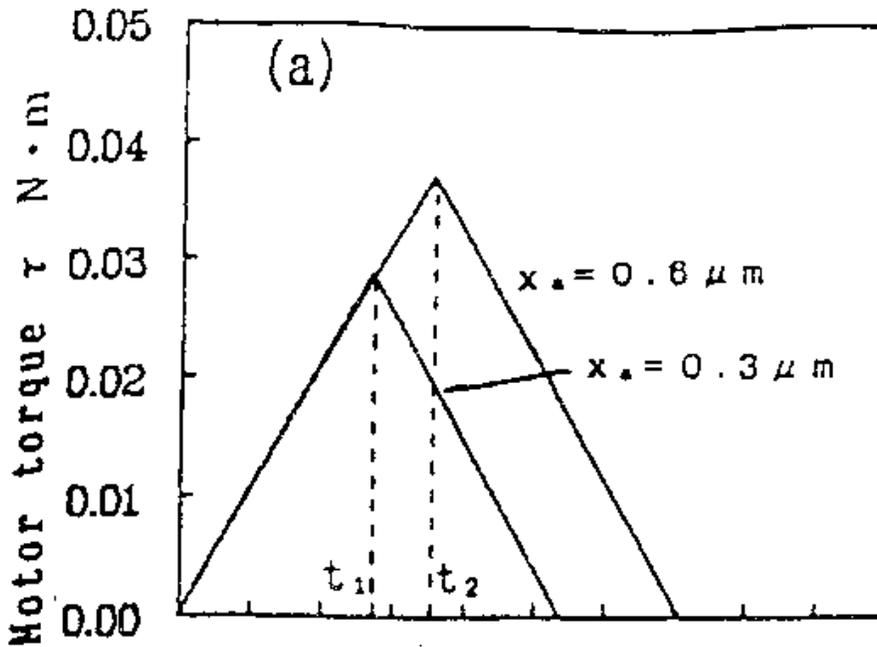
The software servo system developed by the author has the following features (see Fig. 3.3.16):

- i. Rough positioning by PI control ($t = 0$ to t_i). The table is moved in the long travel range.



(b)

Fig. 3.3.9. Experimental rapid-positioning device, (a) Schematic view, (b) Block diagram.



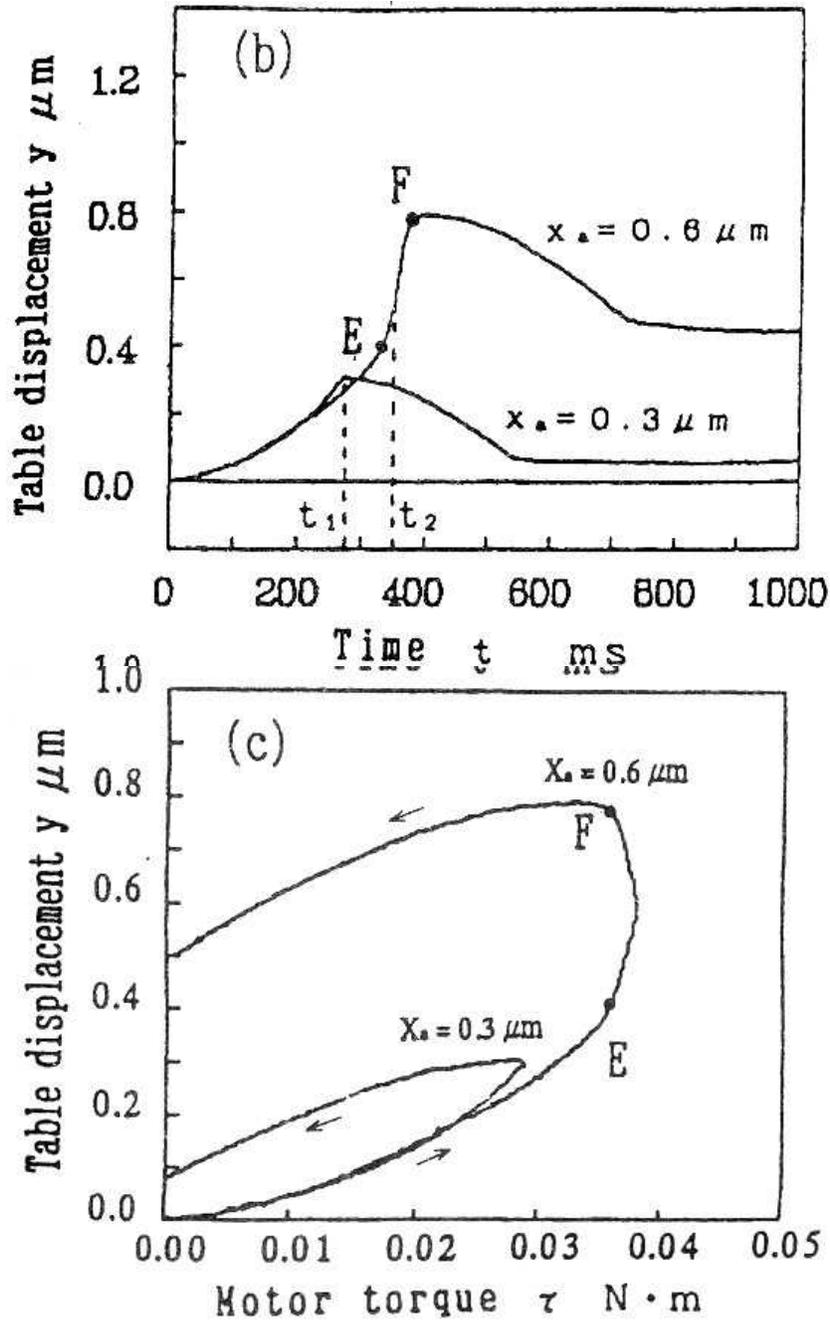


Fig. 3.3.10. Non-linear characteristics of ball screw and table guide without ARV system (from ref. 2).

control ($K_p = 0$), which is particularly effective in the narrow range of the non-linear elastic characteristic.

This software servo system is used for 10 mm stroke step responses, as shown in Fig. 3.3.16: (a) shows the command current i for the servomotor, (b) the table displacement y , and (c) the

magnified displacement in the vicinity of the target position. After the command of $i = 0$, PWC is repeated six times, with $\Delta t = 5, 5, 3, 1,$ and 1 ms. After PWC, integral (I) control is used to stop the table at the target position. In an experiment, the 10 mm stroke step response was repeated 50 times. Figure 3.3.17 shows a histogram of the positioning error E 1 second after table start, showing a mean of $\bar{E} = 1.0$ nm with a standard deviation $\sigma = 3.2$ nm.

(f) Effect of ARV system

Fig. 3.3.12(b) shows the table displacement curve in the vicinity of the target position when the ARV system is used with a table mass of $m = 4.0$ kg. The settling time T_s is 285 ms. Fig. 3.3.13(a)(ii) shows the vertical displacement of the base centre Y_b' , while (b)(ii) shows the FFT analysis results, where we see that the 17 Hz vibration has been completely eliminated.

Using the software servo system we were able to achieve a mean settling time of $T_s = 340$ ms (for 50 trials) and obtain positioning accuracies of nanometre order.

References

1. Yaskawa Electric Co. (1992). Introduction to servo techniques for mechatronics. Nikkan Kogyo Shimbun 97 [in Japanese].
2. Otsuka, J. et al. (1993). Ultraprecision positioning using lead screw drive (2nd report) — nanometer accuracy positioning. Int. Journal of the Japan Society for Precision Engineering, 27, 2.
3. Todd, M.J. and Johnson, K.L. (1987). A model for Coulomb torque hysteresis in ball bearings. International Journal of Mechanical Sciences, 29, 355-65.
4. Futami, S., Furutani, E., and Yoshida, S. (1990). Nanometer positioning and its microdynamics. Nanotechnology, 1, 31.
5. Yang, S. and Tomizuka, M. (1988). Adaptive width control for precise positioning under the influence of stiction and Coulomb friction. Transactions of the ASME, Journal of Dynamic Systems, Measurement and Control, 110, 221.

3.5 Future development of micro-actuators: nano-servo-positioning

3.5.1 Introduction

Microrobots — almost science fiction — do not yet exist but have attracted great interest around the world, as they are considered to be basic technology for the next generation. It is easy to understand the advantage of miniaturization of conventional machines. It makes it possible to use them in confined spaces, and it is expected that they can be mass-produced in a synthesized manner, integrating microprocessors, sensing devices, and micromechanisms, and controlled as autonomous agents. These minute autonomous devices, intelligent mechanical systems combining computer hardware and software, advanced control methods, and microsensors and actuators cooperating synergistically, will force our conventional way of thinking along new paths, and in directions that are not totally clear at present. To realize microrobots, there many unsolved problems remaining, in the areas of micro-actuator and sensing technology, microfabrication and assembly technology, energy supply technology, control and communication technology, materials technology, and so on^(1,2).

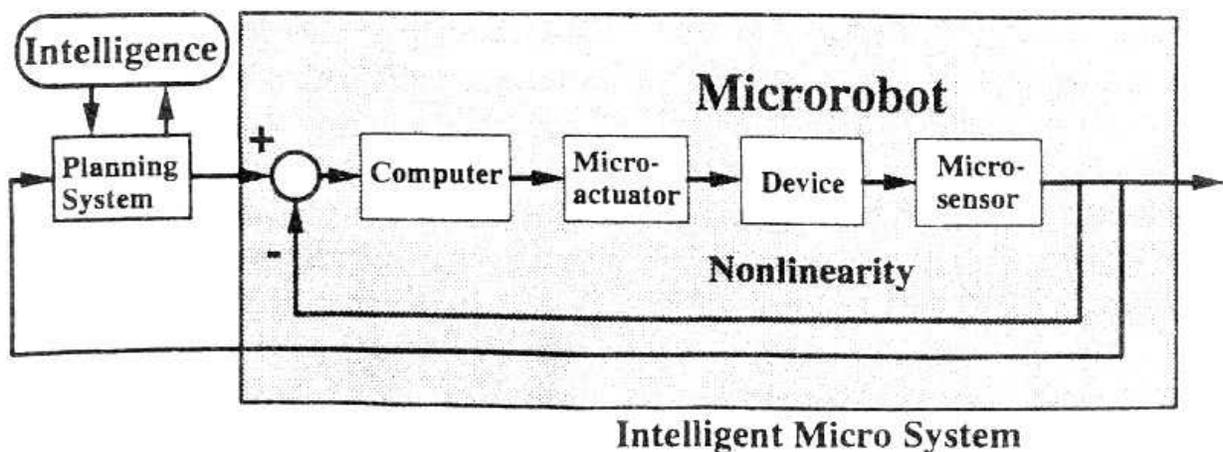


Fig. 3.5.1. Concept of microrobot control system.

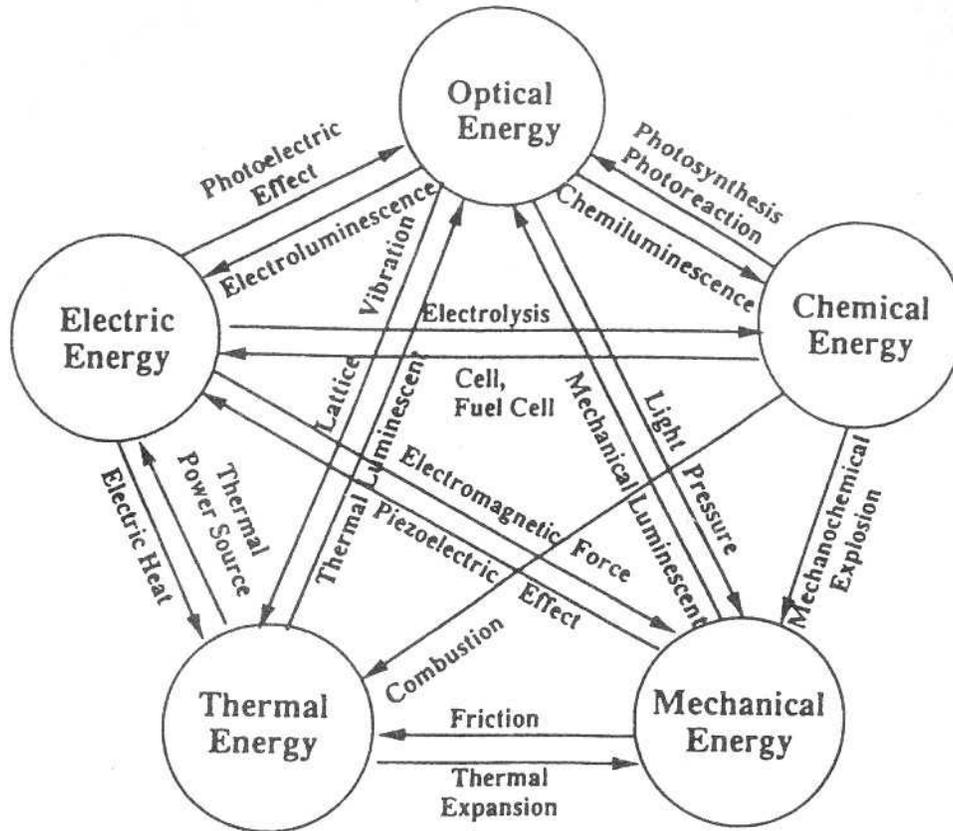


Fig. 3.5.2. Energy transformations.

Microrobots must incorporate several functions, so their parts will be smaller than 1 mm. These parts should be integrated as much as possible to reduce the assembly time. Some may be of sub-micrometre size, and nano-level handling control technology will be required. Moreover, protein engineering is far from a state of development for building and processing the nano-parts, so nano-servo positioning with nano-assembling technology will become essential. From this point of view, recent technology of the STM (scanning tunnelling microscope) and the AFM (atomic force microscope) is promising. To realize such nano-level servo operation, not only actuator control but also sensing, materials and mechanism technologies and their synthesis are important. In these cases, the size of the actuator is not so much important as the configuration and position of the actuators to attain the desired operability. Although operability is not yet adequate, and can be greatly improved, position sensing and control systems are functioning well.

On the other hand, the actuators used for microrobots must have a different nature. This depends on the purpose, but basically they must be small enough to be installed inside a microrobot without sacrificing the power: weight ratio. Moreover, for mobile microrobots, we should consider the energy supply method for the microactuator. A microrobot must have the actuator,

sensor, and processor integrated inside a small body, so the function will be simplified and limited. The concept for the control system of a microrobot is shown in Fig. 3.5.1. Because of the space limitation, the planning system and intelligence (human interface) will be placed outside the robot, and it will have a minimum control system to perform instinctive behaviour or a task commanded by a human.

3.5.2 Micro-actuators and their applications

(a) Classification of micro-actuators

Many different types of micro-actuators have been proposed. The way in which the mechanical energy is obtained is a point of discussion. Energy transformation relationships are shown in Fig. 3.5.2⁽³⁾. For driving an actuator, many different methods are available, such as an electrostatic motor converting electrical energy to mechanical energy. Several conventional micro-actuators have been proposed, such as the electrostatic actuator, electromagnetic actuator, piezoelectric element, GMA (giant magnetostrictive alloy), optical actuator, SMA (shape memory alloy), polymer actuator, and pneumatic actuator. These actuators have their merits and demerits, and several application examples have been proposed⁽¹⁾. Typical examples are given below.

(b) Electrostatic actuator and its applications

When the voltage is applied between the two electrodes, the electrostatic force is generated. From the scaling law, as the size of an object decreases, the mass decreases in proportion to the size cubed. Influence of miniaturization to the electrostatic force between the electrodes is not serious. An electrostatic actuator is suitable for miniaturization and easy to miniaturize. An electromagnetic actuator, which is often compared with the electrostatic actuator, requires a long cable and enough space to produce a magnetic field, and has a large resistance with a large Joule loss, so it is not suitable for miniaturization.

Several kinds of electrostatic actuator have been developed. At first, UC Berkeley succeeded in rotating the side-drive electrostatic micromotor, and then an MIT group improved the speed to more than $10\,000\text{ rev min}^{-1}$ and its lifetime to more than one week⁽⁴⁾. The principal reason for the improvement was said to be solution of the friction problem. For the micro-actuator, friction is the big problem. Means to overcome this problem are considered to be utilization of (1) elastic

deformation, (2) rotary motion⁽⁴⁾, (3) floating devices, and so on. As examples of (1), parallel⁽⁵⁾, quad^(6,7), and comb types of electrostatic actuators have been developed. Examples of (2) are the wobble motor (harmonic electrostatic motor) and the cylinder or cone-shaped rotary type. As examples of (3), the use of air pressure⁽⁸⁾, magnetic force, and superconductive force (Meissner effect)⁽⁹⁾ have been proposed.

(c) *Piezoelectric actuator and its applications*

The piezoelectric actuator has high resolution (order of nm) and good response (order of kHz), and generates a large force. It is suitable for nano-servo positioning, and it has frequently been used as a micro-actuator with great success^(7,10,11). For example, it is quite common to use PZT ($\text{Pb}(\text{Zr,Ti})\text{O}_3$) for the precise positioner of a micromanipulator and in an STM or AFM. Its characteristics may be summarized as follows: (1) precise positioning is possible without any clearances or backlash; (2) response speed is high and effective for the force operations.

Utilizing these properties, a precise positioner with repetitive control of rapid deformation of PZT has been proposed⁽¹¹⁾. A micromanipulator with multiple degrees of freedom has been proposed using stacked

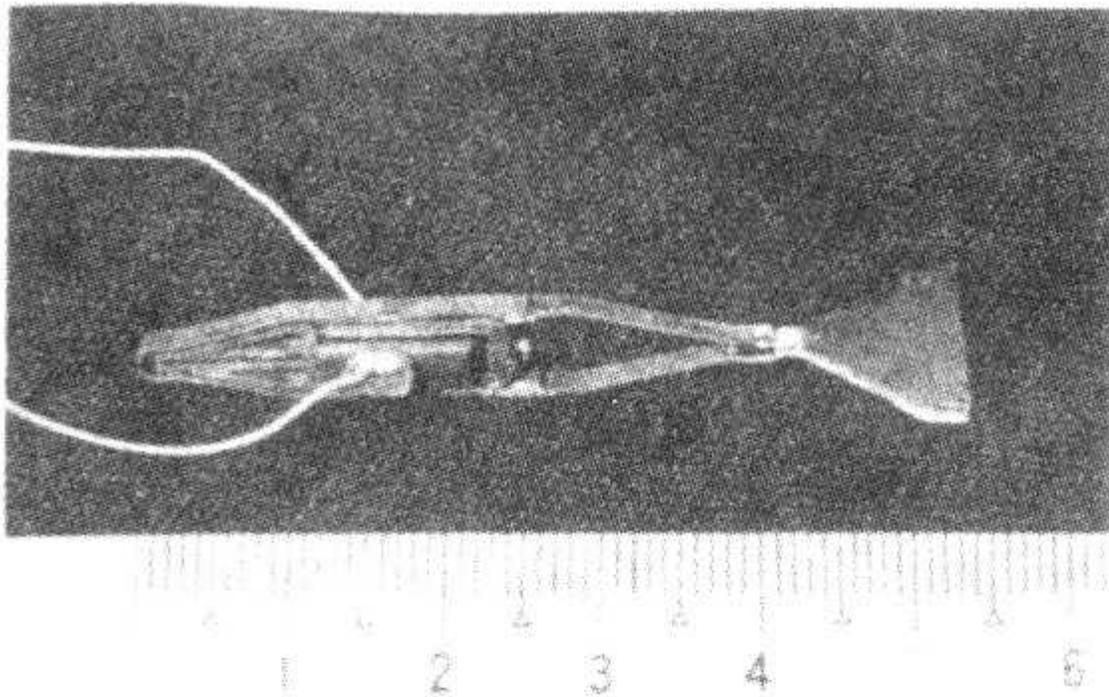


Fig. 3.5.3. Micro-fish.

piezoelectric elements, and position and force experiments have been conducted to demonstrate its effectiveness⁽⁶⁾. A stacked piezoelectric element is not suitable for the miniaturization of the total system. Hysteresis and creep phenomena should be treated properly in the control system. The strain of the piezoelectric element is small (0.1%), and the extension mechanisms have been utilized in some cases.

Figure 3.5.3 shows a micro-fish which can swim in a fluid⁽¹²⁾. This robot uses a stacked piezoelectric element (PZT) with the extension mechanism. For a microrobot in a fluid, the viscosity force dominates over the inertial force on miniaturization. An actuator having a large output and fast response is therefore suitable. This robot uses the rapid deformation of the PZT to generate vibration of the double fins symmetrically (150-750 Hz) to produce a progressive wave as the propulsive force. Because the stacked type has little displacement, this robot augments the displacement of the PZT by up to 326 times (theoretically) by the hinge extension mechanism due to the discharge process. The displacement of the fins is expanded around the resonance frequency and sufficient propulsive force is generated. The swimming speed in water is 3.7 cm s^{-1} .

As other examples, the bimorph type of PZT has been developed for application to a mobile robot, and thin films of PZT and PT have been combined with micromechanical structures and an MOS integrated circuit for robotic applications⁽¹³⁾.

(d) Giant magnetostrictive alloy actuator and its applications

Magnetic material which generates strain when the magnetic field is applied is called magnetostrictive material. Studies on magnetic materials were activated

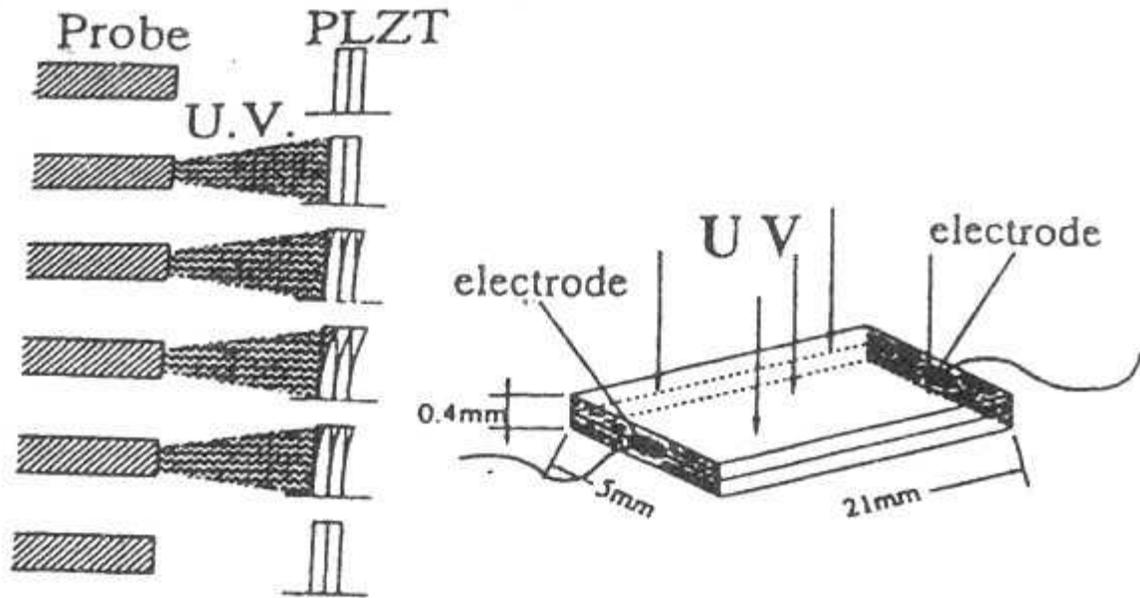


Fig. 3.5.4. UV irradiation experiment with PLZT.

in the USA from 1960, and a material (Tb—Dy—Fe alloy) with a high magnetostrictiveness was developed. Subsequently, with the development of crystal growth technology, a high magnetostrictive property in a comparatively small magnetic field has been obtained. Giant magnetostrictive alloy (GMA) extends in the line of the magnetic field direction and generates strain. GMA has a large force output and large displacement compared with the piezoelectric element (a factor of about two) and the mass per unit stress is small, which can be a great advantage in an actuator⁽¹⁴⁾. To drive the GMA, a magnetic circuit must be provided to control the external magnetic field. However, the element itself can be used as a cableless actuator, and it has been used as a driving actuator for a mobile microrobot⁽¹³⁾.

(e) Optical piezoelectric actuator and its applications

An optical piezoelectric element^(15,16) which shows photostrictive phenomena has received attention as a new actuator for situations where a cable and noise are a nuisance. As applications, a mobile robot and a relay switched on/off by irradiation with a light beam have been developed. The optical response characteristic of the optical piezoelectric element under UV irradiation is characterized by the strain, which arises from a combination of the following three phenomena: (1) the photostrictive effect of generation of the photostrictive voltage, and the strain caused by the piezoelectric effect; (2) the pyroelectric effect of generation of pyroelectric current by the

temperature difference, and the strain caused by the piezoelectric effect; (3) thermal deformation caused by the applied heat flux.

Recently, the bimorph type of PLZT has been developed, increasing the displacement and improving the response time of the strain under UV irradiation

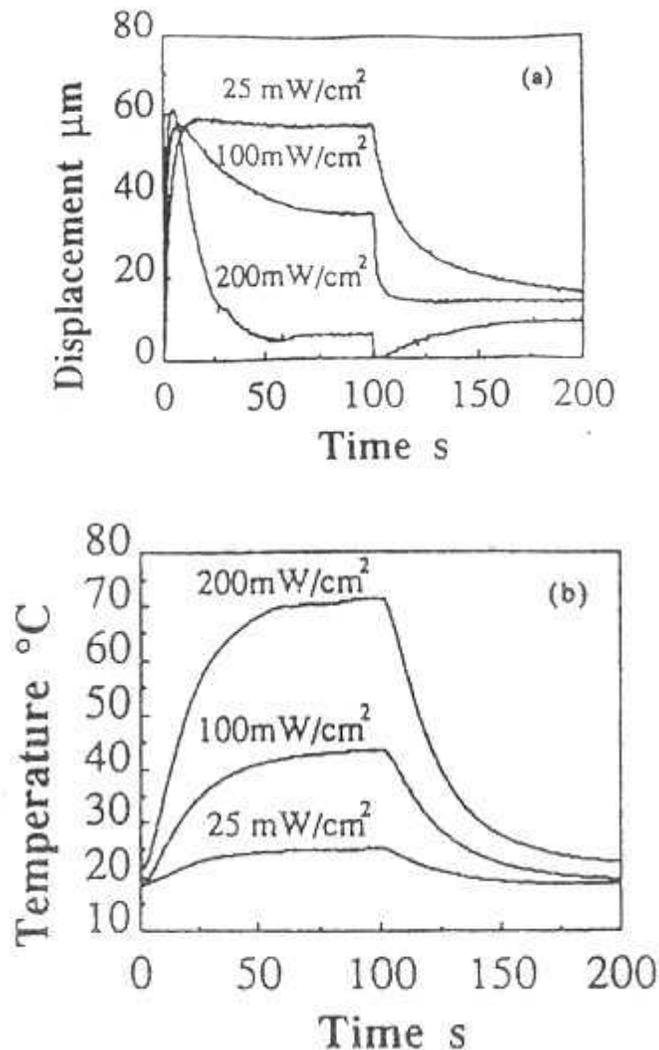


Fig. 3.5.5. Response to UV irradiation, (a) Displacement of tip. (b) Surface temperature of test specimen.

by up to $\sim 20 \text{ s}^{(16)}$. Figure 3.5.4 shows an experimental device for UV (365 nm wavelength peak with narrow spectral band width) irradiation of bimorph PLZT (La: $\text{PbZrO}_3:\text{PbTiO}_3$ 3 : 52 : 48). Radiation was applied for 100 s and then switched off for 100 s. The displacement at the tip of the PLZT and the temperature of the irradiated surface are shown in Fig. 3.5.5. As an example of application this actuator, we have developed an optical micro-gripper and an optical mobile robot

with non-contact energy transmission. At present, the response time of PLZT is slow and this should be improved in future for practical use. We are considering an integrated optical servo control system with energy and information transmission. PLZT has different optical responses in terms of the three different effects stated above. Hence, multifunctional use of PLZT is expected to be developed, not only as an actuator but also as an information transmitter.

(f) Other actuators

Other actuators such as a shape memory alloy (SMA) actuator and a polymer actuator have been developed, and are promising as micro-actuators.

SMA has a shape memory effect depending on the temperature. On miniaturization, heat capacity decreases and heat radiation from the surface relatively increases, so an improvement in the response speed can be expected. SMA has many application examples such as medical operating tools⁽¹⁷⁻¹⁹⁾.

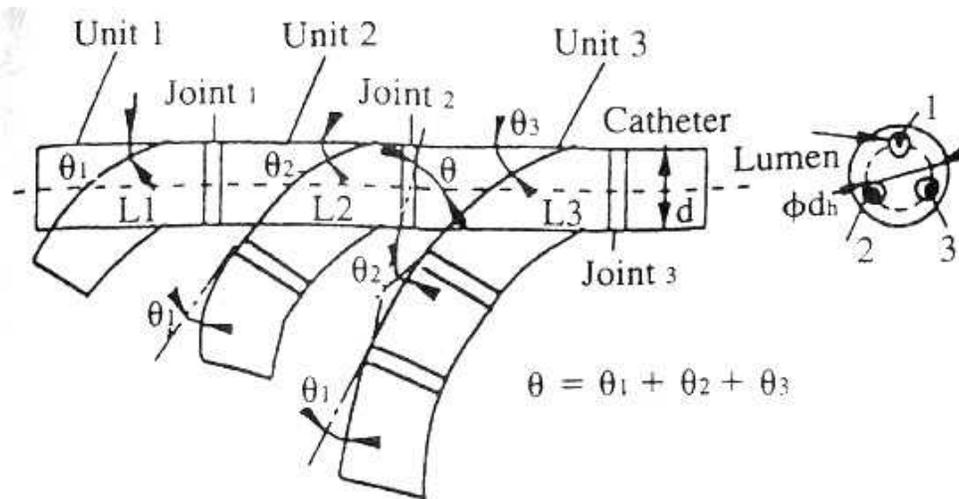


Fig. 3.5.6. Active catheter with multiple degrees of freedom for minimally invasive intravascular neurosurgery.

Figure 3.5.6 shows an active catheter with multiple degrees of freedom using SMA wires⁽¹⁹⁾. This catheter is made of serially connected units which contain SMA wires in the lumina, and its multiple degrees of freedom are obtained by series-parallel structure. This catheter has been developed as a medical tool for the improvement of the operability of minimally invasive surgery. The SMA actuator is promising for integration with micro-mechanisms. Thin-film technology for the SMA is under development. The cooling rate is improved by miniaturization, and a bimetal type which utilizes the difference in thermal expansion has been proposed.

The polymer actuator has the characteristics of strength against impact, force, and moment, ease of processing and light weight. As examples of application, a micro-probe using a piezoelectric polymer actuator, a micro-gripper actuated by a pH-driven film device, and a chemical valve using a polymer actuator contracted by electricity have been proposed. For the medical drug delivery, a micro-pump using a thermo-responsive polymer gel and a highly water-absorbent polymer gel have been proposed⁽²⁰⁾.

3.5.3 Energy supply method

(a) Classification of energy supply methods

One of the final objectives of microrobotics is to realize an antlike mobile robot which is small and intelligent to perform given tasks. Most present microrobots are supplied with energy by the cable. But when a robot becomes small, the cable disturbs its motion with great friction. Hence the method of energy supply to the micro-actuator becomes important. Methods are classified as internal supply (internal energy sources) and external supply (external supply of the energy to the system but without the need for a cable).

(b) Internal supply method

In this case, the energy source is contained inside the moving body. Electrical energy is frequently used as internal energy, and for this purpose, a battery and a condenser have been developed. The battery type is good in terms of output and durability, but it is difficult to miniaturize. Recently a micro lithium battery of micrometre thickness and current density 60 mA cm^{-2} , and which is rechargeable at 3.6-1.5 V has been developed by thin-film technology⁽²¹⁾. As for the condenser type, an autonomous mobile robot $\sim 1 \text{ cm}^3$ in volume (named Monsieur) was developed in 1992 by Epson, based on conventional watch production technology. It uses a high-capacity condenser of 6.8 mm diameter, 2.1 mm thickness, and 0.33 F capacity as an energy source. The electrical capacity of the condenser is small compared with that of the secondary battery, but this microrobot uses two stepping motors with current control by pulse width modulation, and can move for $\sim 5 \text{ min}$ after $\sim 3 \text{ min}$ charging.

(c) External supply method

In this case, energy is supplied to the moving body from outside.

The following methods can be considered:

- (1) optical
- (2) electromagnetic

- (3) ultrasonic
- (4) other.

Method (1) can be classified into (i) optical pressure by irradiating laser beam, (ii) optical energy to strain conversion using UV irradiation and photostrictive phenomena, and (iii) optical energy to heat conversion.

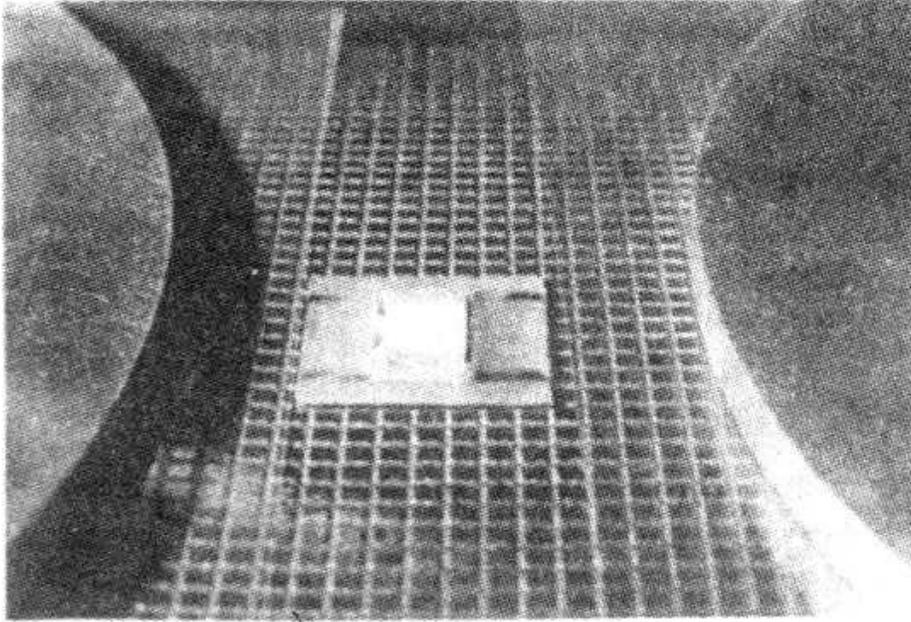


Fig. 3.5.7. Optical mobile robot with non-contact energy transmission on an air table.

As an example of (i), remote operation of the micro-object by a single laser beam as a tweezer has been proposed. As an example of (ii), an optical piezoelectric actuator such as PLZT has been developed, as already mentioned. As an example of (iii), a low-boiling liquid has been used with an optical heat conversion material. Utilization of the pyro-electric effect has also been proposed to supply external energy. Figure 3.5.7 shows a mobile plate on an air table, using pyroelectric current generated by the temperature difference arising from the heat applied by UV irradiation⁽²²⁾. Generally, PLZT can be replaced by another pyroelectric element which can generate current from a temperature change. The field of the mobile plate is made of square electrodes arranged in a square grid. The electrodes are 1.5 x 1.5 mm wide and are placed at intervals of 0.5 mm. The field has many holes (diameter 0.18 mm) placed at intervals of 1 mm, and air is blown through to float the plate. The bottom face of the plate has several electrodes 1 x 1 mm wide placed at intervals of 4 mm. Each electrode is connected to the PLZT. By UV

irradiation, a thrust is generated between the bottom face of the plate and the field, and this can be used as the driving force for the plate. By the use of the air table, friction is considerably reduced and even the weak electrostatic force is enough to move it rapidly. In an experiment, it moved over the field at a speed of 5 cm s^{-1} . Position control of it can be attained by controlling the irradiation selectively.

As an example of method (2), microwaves, which have been used for non-contact energy transmission to aircraft and a solar energy generation satellite, have been considered, but there is few reports on research on their use for micro-actuators. GMA already mentioned can be considered to be an example of this method.

As an example of method (3), radiation pressure from ultra-sonic waves can be used for the non-contact operation and driving of an objects. As another example may be considered the transmission of force through an external medium to a pipeline maintenance pig. In a further example, selective energy transmission to an elastic object on a vibrating plate has been proposed.

3.5.4 Conclusion

For nano-servo positioning, synthesis of the actuator, sensing, control, materials, and mechanical technology is important. This section has focused on and introduced micro-actuators and their application, together with the energy supply method, which can be the crucial feature for practical use of mobile

microrobots. To realize a microrobot, actuator selection and the energy supply method must be discussed together, depending on the task and purpose. Moreover, the physics dominating a macro-object is not always the same for the micro/nano-object. The nano-world in particular is completely different⁽²³⁾. We can observe it by electron microscope or STM/AFM, where quantum mechanical analysis is required because of the interaction between the molecules and the electromagnetic wave effect. In the near future, consideration should be given to the method of control of the actuator depending on the size of the system.

References

1. Fukuda, T. and Arai, F. (1992). Microrobotics - approach to the realization. In Micro System Technologies 92, pp. 15-24. VDE Verlag.
2. Fukuda, T. and Arai, F. (1993). Microrobotics - on the highway to nanotechnology. IEEE Industrial Electronics Society Newsletter, Dec., pp. 4-5.

3. Yamada, A. et al. (1983). Optical energy transformation, Society Publication Center (in Japanese), p. 7.
4. Mehregany, M. et al. (1990). Operation of microfabricated harmonic and ordinary side-drive motors. In Proceedings, IEEE Micro Electro Mechanical Systems.
 1. pp. 1-8.
5. Fujita, H. et al. (1988). An integrated micro servosystem. In IEEE International Workshop on Intelligent Robots & Systems, pp. 15-20.
6. Fukuda, T. and Tanaka, T. (1990). Micro electrostatic actuator with three degrees of freedom. In Proceedings, IEEE Micro Electro Mechanical Systems, pp. 153-8.
7. Fukuda, T. and Arai, F. (1992). New actuators for high-precision micro systems. H.S. Tzou and T. Fukuda (eds.), In Precision sensors, actuators and systems, (eds H.S. Tzou and T. Fukuda), pp. 1-37. Kluwer.
8. Pister, K.S.J. et al. (1990). A planar air levitated electrostatic actuator system. In Proceedings, IEEE Micro Electro Mechanical Systems, pp. 67-71.
9. Kim, Y.K. et al. (1990). Fabrication and testing of a micro superconductive actuator using Meissner effect. In Proceedings, IEEE Micro Electro Mechanical Systems, pp. 61-4.
10. Hatamura Y., and Morishita, H. (1990). Direct coupling system between nanometer world and human world. In Proceedings, IEEE Micro Electro Mechanical Systems, pp. 203-8.
11. Higuchi, T., Yamagata, Y., Furutani, K., and Kudoh, K. (1990). Precise positioning mechanism utilizing rapid deformations of piezoelectric elements. In Proceedings, IEEE Micro Electro Mechanical Systems, pp. 222-6.
12. Fukuda, T., Kawamoto, A., Arai, F. and Matsuura, H. (1994). Mechanism and swimming experiment of micro mobile robot in water. In Proceedings, IEEE Micro Electro Mechanical Systems, (to be published).
13. Polla, D.L. (1992). Micromachining of piezoelectric microsensors and microactuators for robotics applications. In Precision sensors, actuators and systems, (eds H.S. Tzou and Fukuda, T.), pp. 139-4. Kluwer.
14. Fukuda, T., Hosokai, H., Ohyama, H., Hashimoto. H., and Arai, F. (1991). Giant magnetostrictive alloy (GMA) applications to micro mobile robot as a micro actuator

- without power supply cables. In Proceedings, IEEE Micro Electro Mechanical Systems, pp. 210-5.
15. Uchino, K. and Aizawa, M. (1985). Photostrictive actuator using PLZT ceramics. Japanese Journal of Applied Physics, 24, (Suppl. 42-3), 139-42.
 16. Fukuda, T., Hatton, S., Arai, F., et al. (1992). Optical servo system using bimorph optical piezoelectric actuator. In Proceedings, Third International Symposium on Micro Machine and Human Science (MHS92), pp. 45-50.
 17. Ikuta, K. (1988). The application of micro/miniature mechatronics to medical robots. In Proceedings, IEEE/ IROS, pp. 9-14.
 18. Dario, P., Valleggi, R., Pardini, M., and Sabatini, A. (1991). A miniature device for medical) intracavitary intervention. In Proceedings, IEEE Micro Electro Mechanical Systems, pp. 171-5.
 19. Fukuda, T., Guo, S. et al. (1993). Active catheter system with multi degrees of freedom. In Proceedings, Fourth International Symposium on Micro Machine and Human Science (MHS 93), pp. 155-62.
 20. Hattori, S., Fukuda, T. et al. (1992). Structure and mechanism of two types of micro-pump using polymer gel. In Proceedings, IEEE Micro Electro Mechanical Systems, pp. 110-15.
 21. Bates, J.B., et al. (1993). Rechargeable solid state lithium microbatteries. In Proceedings, IEEE Micro Electro Mechanical Systems, pp. 82-6.
 22. Ishihara, H. and Fukuda, T. (1993). Micro optical robotic system (MORS). In Proceedings, Fourth International Symposium on Micro Machine and Human Science (MHS 93), pp. 105-10.
 23. Arai, F., Ando, D. et al. (1995). Micro manipulation based on microphysics. Proceedings of the International Conference on Intelligent Robots and Systems, Vol. 2, pp. 236-41

Module-IV

Applications of Nanotechnology: Nano-grating system, Nano lithography, Photolithography, Electron beam lithography, Machining of soft metal mirrors with diamond turning, Mirror grinding of ceramics, Ultra precision block gauges, balls for rolling bearings, Fabrication CCD's, VCR head assemblies, Optical fibres.

Applications of Nanotechnology

4.1 Nano-grating Systems

4.1.1 Mechanically ruled gratings

The diffraction grating is a typical product for which nanotechnology plays a significant role in its fabrication and evaluation. It is an optical element with extremely fine parallel grooves on a flat or concave optical surface. The groove shape on a mechanically ruled grating is usually triangular, as shown in Fig. 4.1.1, and the spacing of the grooves ranges from submicrometre to a few tens of micrometres, depending on the wavelength range in which the grating is to be used. The reason why nanotechnology is required in the fabrication of diffraction gratings lies not only in the fineness of each groove facet to be shaped but also in the accuracy of the groove positions^(1,2). The errors of the groove positions on typical mechanically ruled gratings are as shown in Fig. 4.1.2 and can be classified into three categories: accumulative, periodic, and random. Accumulative errors of groove positions cause the spectral resolving power to be reduced; the tolerance for the accumulative error to obtain the theoretical resolving power is $\sim 0.1 \mu\text{m}$ in the case of a grating with a groove density of 1200 mm^{-1} used in the first order. Periodic errors of groove positions cause ghost or false spectral lines. The tolerance for the periodic error to reduce the amplitude of the ghost to practically negligible order is $\sim 0.01 \mu\text{m}$ in the case of gratings for visible and/or ultraviolet use. Random errors of groove positions make the incident light scatter and cause background noise in instruments. The tolerance for the random error is also in the region of $0.01 \mu\text{m}$. The groove area (the area of the cross-section of each groove facet) ranges from $10^{-3} \mu\text{m}^2$ for soft-X-ray gratings to a few tens of μm^2 for infrared gratings.

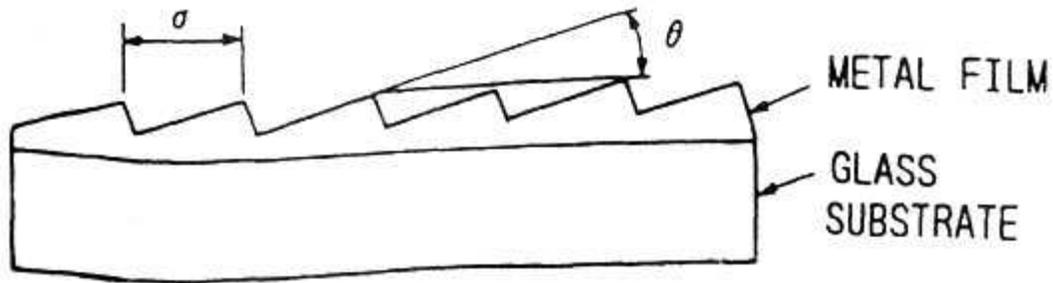


Fig. 4.1.1. Groove shape of diffraction grating.

The accuracy of the groove angle and the smoothness of every groove facet are extremely important, especially in the case of soft-X-ray and ultraviolet gratings, since the imperfection of groove shape has to be small compared with the wavelength at which the grating is to be used.

The ruling engine is a dedicated machine tool for the fabrication of diffraction gratings. It is named after H.A. Rowland who constructed the first machine at the Johns Hopkins University in the 1880s⁽³⁾. Ruling engines were purely mechanical ultra-precise machine tools until the middle of the twentieth century. The key element of such ruling engines was the lead screw and its bearings, and great manual skill was required for the lapping and alignment of the screw⁽⁴⁾. However, the Rowland ghost, caused by periodic errors in the lead screw for the groove spacing, was unavoidable and the ghost was thought to be an intrinsic defect of diffraction gratings.

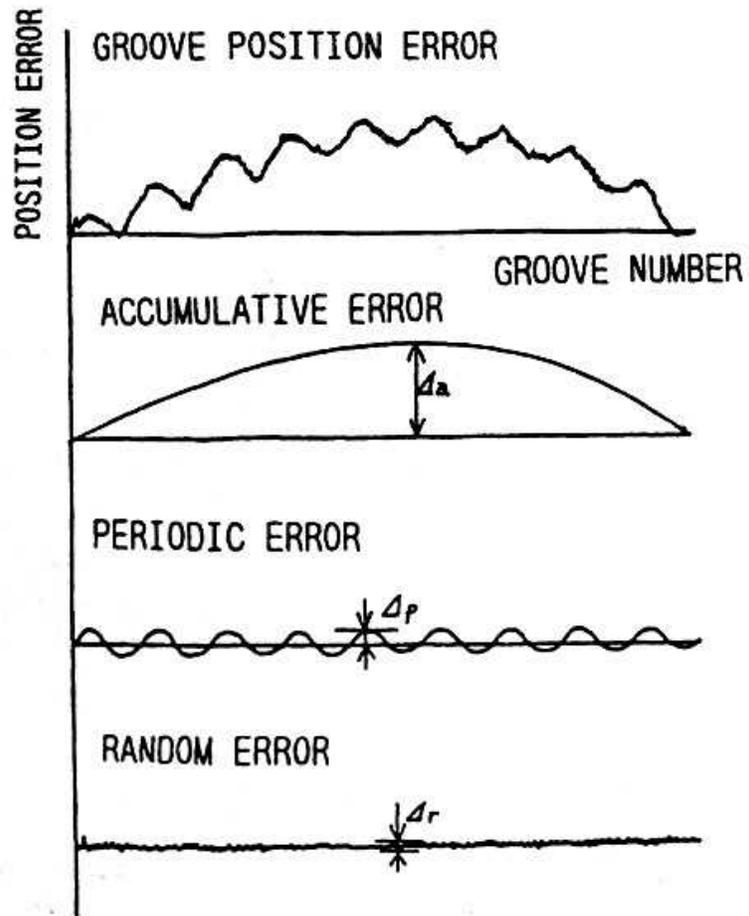


Fig. 4.1.2. Groove position errors of diffraction grating.

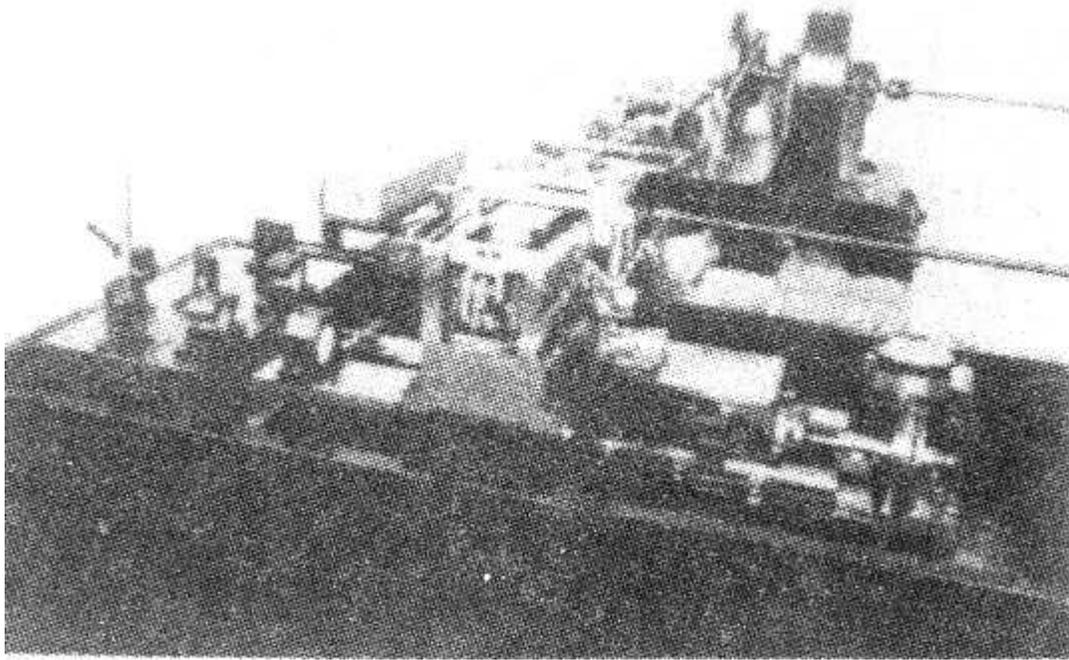


Fig. 4.1.3. Laser interferometrically controlled ruling engine.

In the 1950s, an interferometrically controlled ruling engine was developed at the Massachusetts Institute of Technology⁽⁵⁾. This was designed to measure and control the position of every groove, with the mercury green spectral wavelength as a standard, and the positioning accuracy of every groove was improved to nanometre order so that ghosts in the spectral images were eliminated completely. This interferometric control system was introduced into ruling engines both in the USA and in other countries, especially using a frequency-stabilized He-Ne laser as a light source⁽⁶⁻⁸⁾.

A laser interferometrically controlled ruling engine constructed at the Central Research Laboratory of Hitachi Ltd is shown in Fig. 4.1.3 and its control system in Fig. 4.1.4. This is a so-called shaper-type machine tool; while a diamond tool reciprocates to rule the grooves, the blank carriage is translated continuously by a lead screw for the groove spacing. A laser interferometer monitors the translation of the blank carriage. A reference signal, which is an ideal fringe signal when there is no mechanical error in the translation of the blank carriage, is generated independently and the mechanical translation error is detected as a phase difference between the reference and the fringe signals. This error signal causes a servomotor to rotate to compensate

the mechanical translation error for the groove spacing. The sensitivity of positioning error depends on the wavelength of the light source and the detectable phase difference between the reference and fringe signals. To improve the sensitivity, a multi-reflection laser interferometer as shown in Fig. 4.1.5 was used for the ruling engine. Here, one fringe signal interval corresponds to one-eighth of the laser wavelength ($\sim 0.08 \mu\text{m}$) in the carriage translation. As the

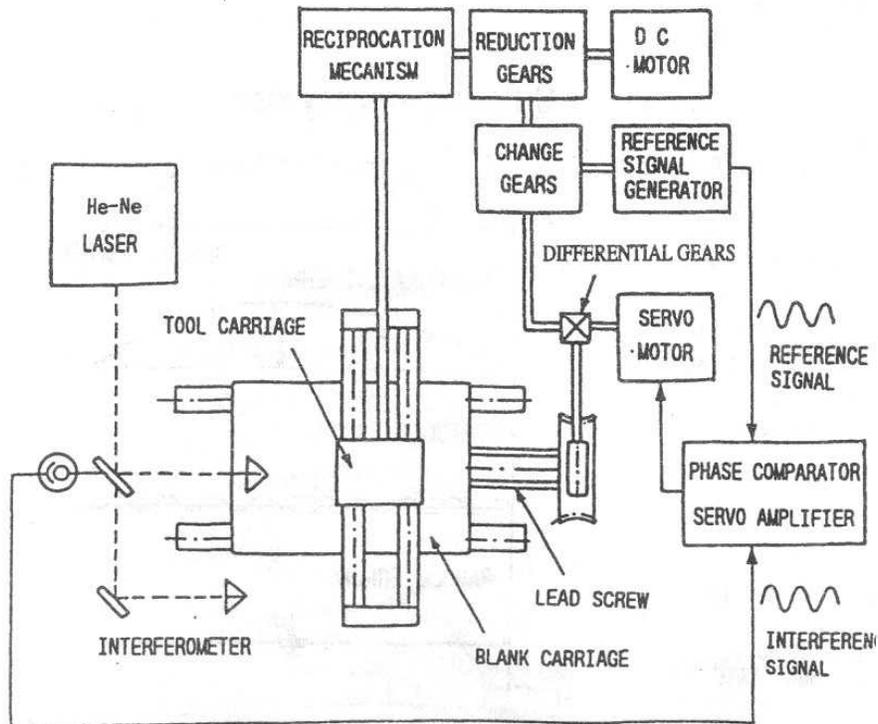


Fig. 4.1.4. Control system of ruling engine.

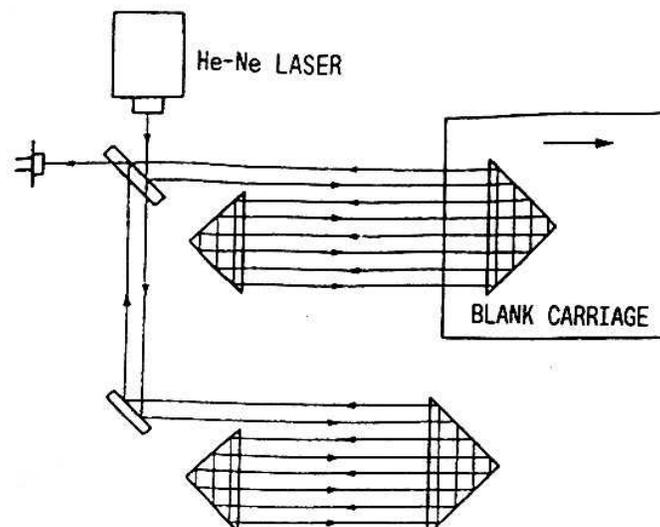


Fig. 4.1.5. Multi-reflection laser interferometer.

minimum detectable phase difference is approximately one-fortieth of the fringe period, the interferometer system is able to detect a position error of approximately 2 nm in the grooves⁽⁹⁾. A Twyman - Green interferogram of the diffracted wavefront of a 600 mm^{-1} grating with purely mechanical ruling .and with laser interferometrically controlled ruling is shown in Fig. 4.1.6. The periodic error of the grooves with a lead screw pitch of 2 mm is eliminated by the laser interferometric control.

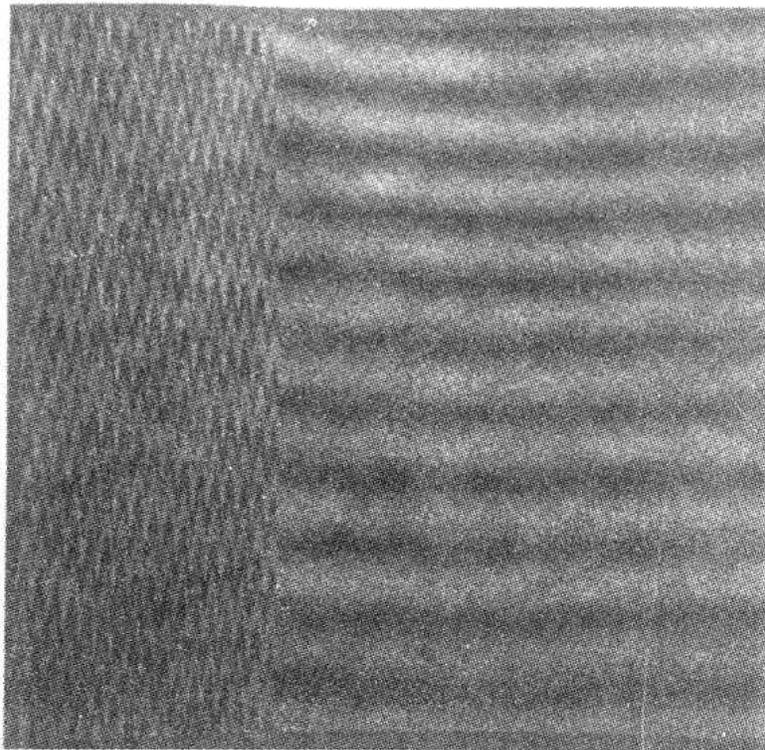


Fig. 4.1.6. Interferogram of mechanically ruled diffraction grating (600 mm^{-1} 1st-order): contrast between purely mechanical and laser interferometric control.

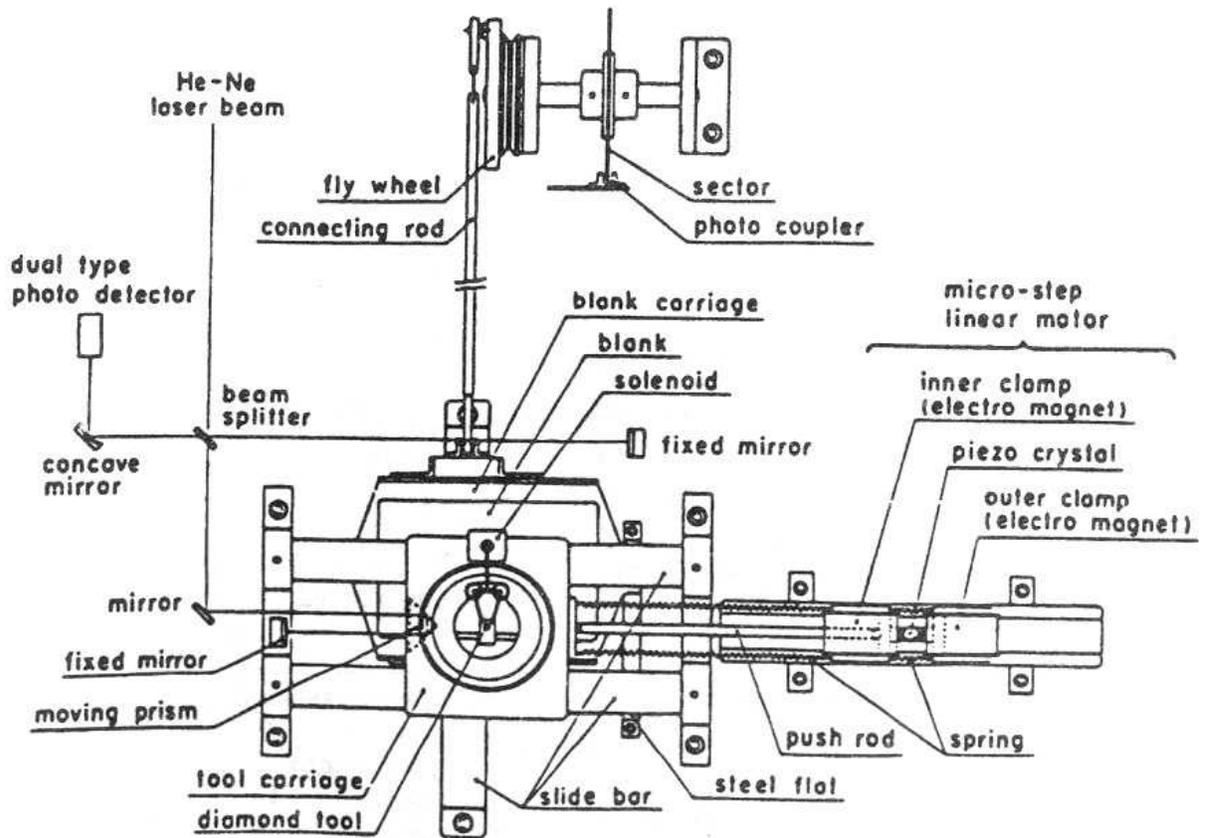


Fig. 4.1.7. Control system of piezoelectrically controlled ruling engine.

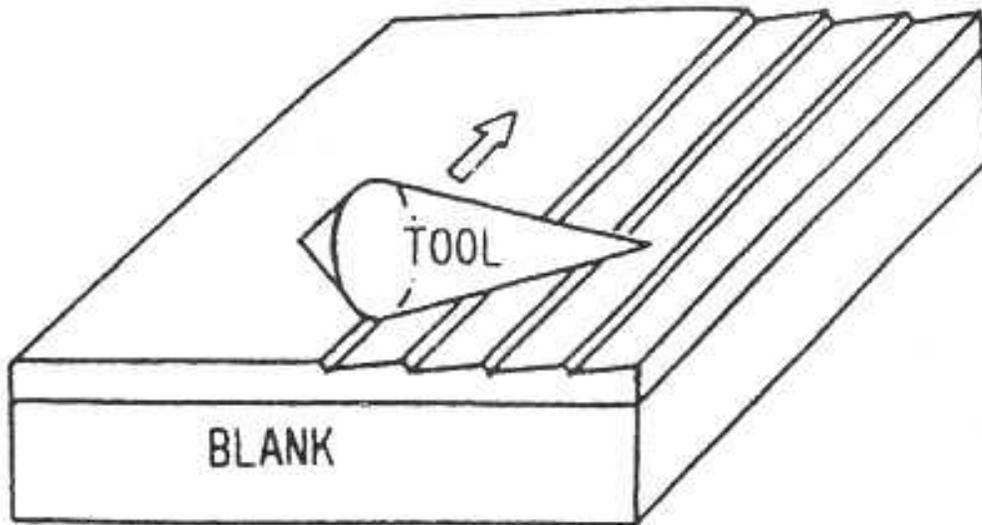


Fig. 4.1.8. Burnishing process for grating grooves.

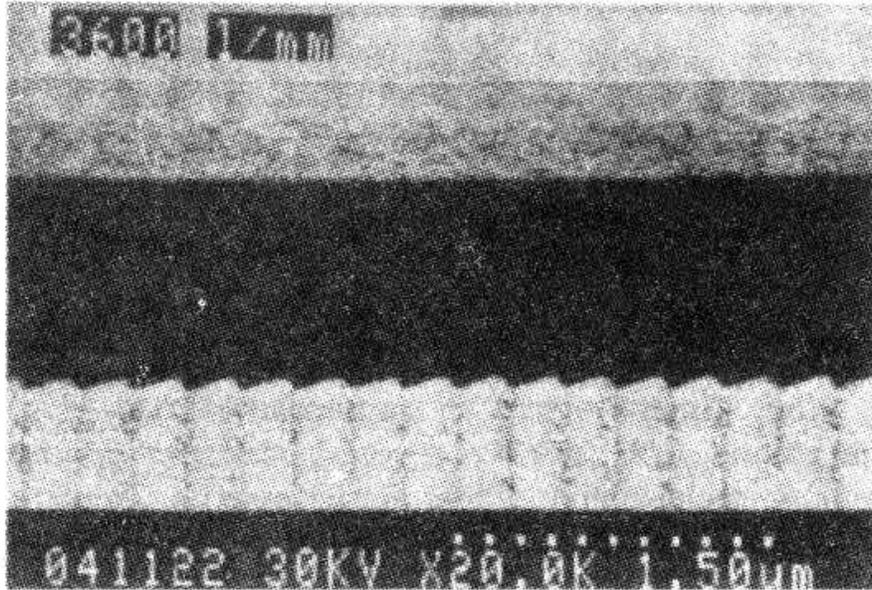


Fig. 4.1.9. Cross-section of mechanically ruled grating grooves (3600 mm^{-1}).

With the improvement of laser interferometric measurement and electromechanical actuators, hydraulic or piezoelectric drives have been introduced for translation of the groove spacing in recent ruling engines with digital control systems^(10,11). A piezo-electrically driven planer-type ruling engine was constructed at the Tohoku University; its control system is shown in Fig. 4.1.7⁽¹²⁾. A micro-step linear motor with two magnetic clamps and a ceramic piezo-crystal cemented between the clamps is used as the actuator for tool translation. For groove spacing, the piezo-crystal expands until the laser interferometer counts the fringes for the exact position of the groove, and the magnetic clamp holds the tool carriage stationary while the blank carriage reciprocates for ruling a groove.

The groove facets of diffraction gratings are formed by a burnishing process (plastic deformation of metal), using a shaped diamond tool as shown in Fig. 4.1.8. An aluminum film vacuum-evaporated onto a polished glass substrate is usually used as a grating blank. After appropriate alignment and loading of the diamond tool, triangular groove facets as shown in Fig. 4.1.9 are obtained⁽¹³⁾. The groove depth ranges from a few tens of nanometres to a few tens of micrometres, depending on the wavelength range in which the grating is to be used.

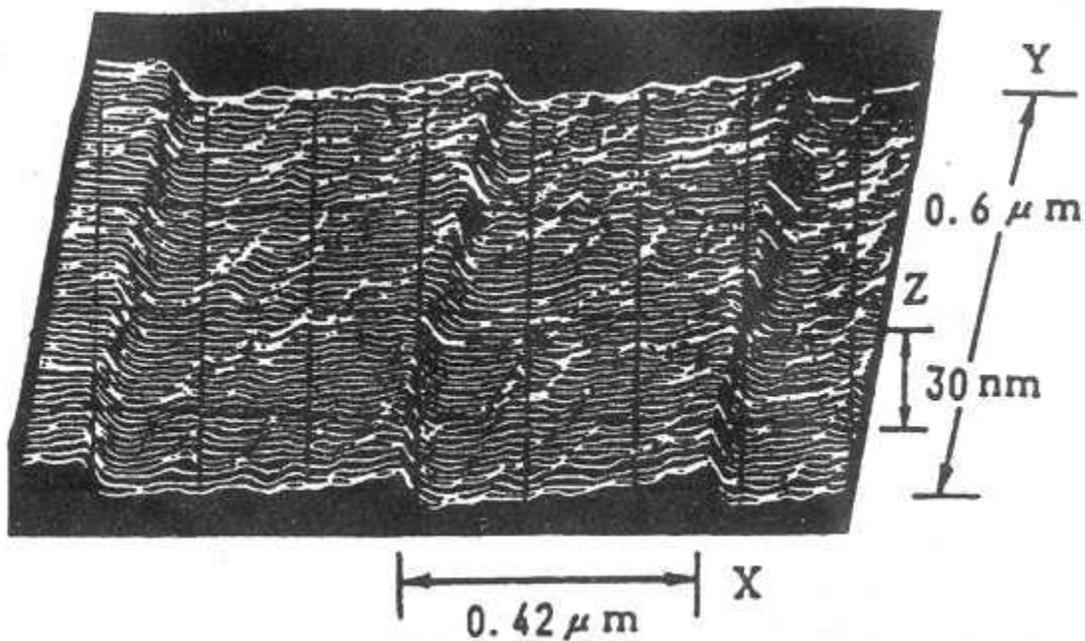


Fig. 4.1.10. Scanning tunneling micrograph of grating grooves (2400 mm^{-1} , 1.7°).

To evaluate the groove shape, especially for the extremely fine and shallow grooves of soft-X-ray gratings, it is necessary to use either a scanning tunnelling microscope or an atomic force microscope to measure the cross-sectional profile. A scanning tunnelling micrograph of a soft-X-ray grating with a groove density of 2400 mm^{-1} and 1.7° groove angle is shown in Fig. 4.1.10⁽¹⁴⁾. The highest groove density obtained so far by mechanical ruling is $10\,000 \text{ mm}^{-1}$ ⁽¹⁵⁾. Here the grooves are ruled directly on a polished glass surface.

References

1. Harrison, G.R. (1949). The production of diffraction gratings, I. Development of the ruling art. *Journal of the Optical Society of America*, 39, 413-26.
2. Hutley, M.C. (1982). *Diffraction gratings*. Academic Press, London.
3. Strong, J. (1960). The Johns Hopkins University and diffraction gratings. *Journal of the Optical Society of America*, 50, 1148-52.
4. Strong, J. (1951). New Johns Hopkins ruling engine. *Journal of the Optical Society of America*, 41, 3-15.

5. Harrison, G.R. and Stroke, G.W. (1955). Interferometric control of grating ruling engine with continuous carriage advance. *Journal of the Optical Society of America*, 45, 112-21.
6. Harrison, G.R. (1973). The diffraction grating -an opinionated appraisal. *Applied Optics*, 12, 2039-48.
7. Jarell, R.F. and Stroke, G.W. (1964). Some new advances in grating ruling, replication and testing. *Applied Optics*, 3, 1251-62.
8. Loewen, E.G. (1970). Diffraction gratings for
 1. spectroscopy. *Journal of Physics E: Scientific Instruments*, 3, 953-61.
 2. Instruments, 3, 953-61.
9. Harada, T. and Kita, T. (1980). Mechanically ruled aberration-corrected concave gratings. *Applied Optics*, 19, 3987-93.
10. Horsfield, W.R. (1965). Ruling engine with hydraulic drive. *Applied Optics*, 4, 189-93.
11. Bartlett, R. and Wildy, P.C. (1975). Diffraction grating ruling engine with piezoelectric drive. *Applied Optics*, 14,1-3.
12. Takashima, K. and Nawata, S. (1978). Diffraction grating ruling engine with piezoelectric drive. *Japanese Journal of Applied Physics*, 17, 1445-6.
13. Harada, T., Taira, H., Kita, T., and Itou, M. (1987). Groove profile measurement of diffraction grating using electron microscope. *SPIE Proceedings*, 815, 118-23.
14. Oshio, T., Sakai, Y., and Ehara, S. (1987). Observation of grating surface by the use of scanning tunnelling microscope. *SPIE Proceedings*, 815, 124-6.
15. Takashima, K. (1988). Ultra-precision machining of diffraction gratings. *Optical and Electro-optical Engineering Contact*, 26, 218-23 (in Japanese).

4.1.2 Holographic gratings

A novel optical process for the fabrication of diffraction gratings was developed in the 1960s by adoption of laser holography⁽¹⁾. The holographic grating is a diffraction grating which utilizes the interference fringes of laser beams as grooves. The manufacture of holographic gratings is much faster than mechanical ruling and has the advantage of ghost-free groove positioning.

The fabrication process for a holographic grating is shown in Fig. 4.1.2. A grating blank with a photoresist material coated on to a polished glass substrate is prepared. Two collimated laser beams separated by a beamsplitter are applied so as to form parallel and equally spaced

interference fringes on the blank surface. Usually the visible radiation of an Ar laser (457.9 nm) or an He-Cd laser (441.6 nm) is used. The interference fringes are exposed and sinusoidal grooves are formed after development. Finally, a metal film is vacuum-evaporated on to the photoresist grooves for the reflective surface.

The spacing a of a holographic grating is given by eq. (4.1.2.1) as a function of laser wavelength λ_0 and angles of incidence of the laser beams γ and δ :

$$\sigma = \frac{\lambda_0}{\sin \gamma - \sin \delta} \quad (4.1.2.1)$$

A groove density as high as 4000 mm^{-1} can be obtained using visible laser radiation. The exposure time for holographic grating grooves is usually a few minutes, whereas the mechanical ruling of a 4000 mm^{-1} grating 100 mm wide takes more than 20 days. The positioning accuracy of the grooves depends on the quality of the wavefronts for the fringe exposure,

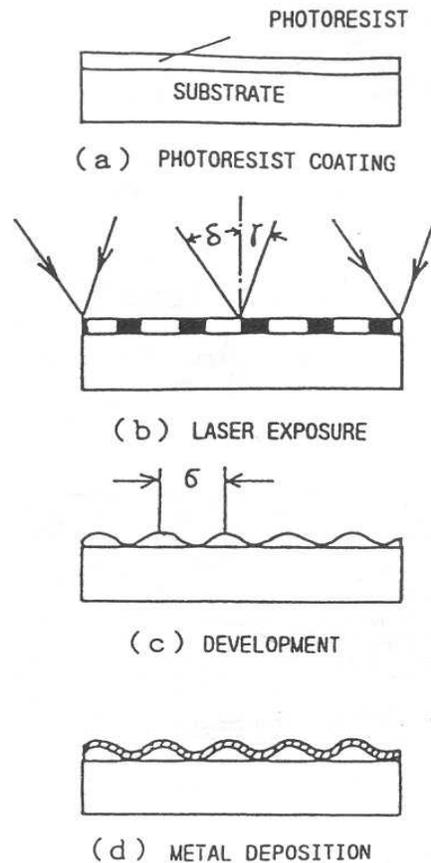


Fig. 4.1.2.1 Fabrication process for holographic grating.

so a high resolving power and a ghost-free spectral image quality can be expected from holographic gratings.

One disadvantage of holographic gratings is that the groove shape is basically sinusoidal. With a mechanically ruled grating, the diffraction energy can be concentrated into a certain wavelength range by forming triangular grooves with an appropriate groove angle. This is called the blaze effect, and the groove angle is called the blaze angle. Several processes for converting the sinusoidal shape of holographic grating grooves to a triangular shape have been developed so far⁽²⁻⁴⁾. The most practical is the use of ion etching technology as developed at RIKEN, the Institute of Physical and Chemical Research⁽⁵⁾; the process is shown in Fig. 4.1.2.2. First, a holographic grating with photoresist grooves on a glass substrate is prepared. Then the grating is

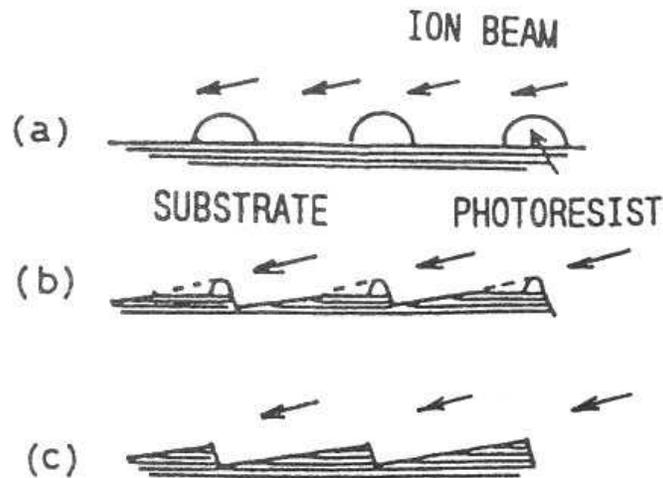
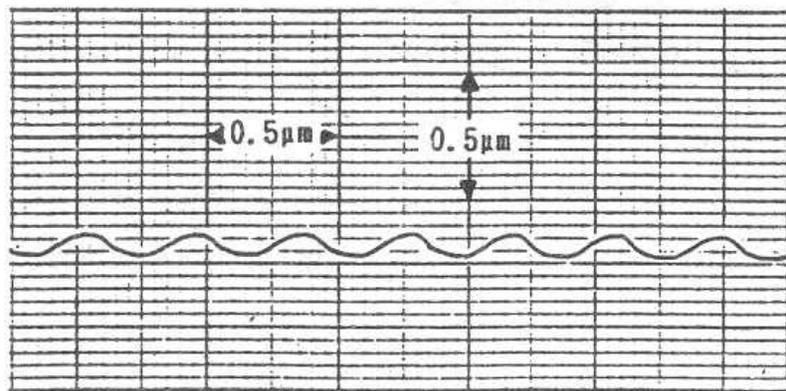
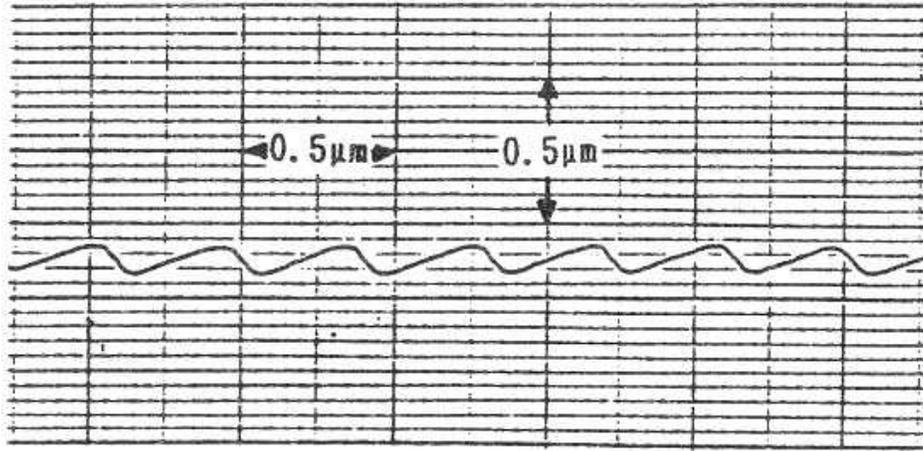


Fig. 4.1.2.2. Fabrication process for ion-etched blazed holographic grating.



(a) Photoresist grating



(b) Ion-etched grating

Fig. 4.1.2.3. Cross-sections of holographic gratings (2400 nm^{-1}). Courtesy of K. Sano, Shimadzu Corp.

mounted in a vacuum chamber and the surface is etched obliquely with accelerated ion beams. By taking advantage of the difference in etching rate between the photoresist and the substrate, triangular grooves are formed. The factors determining the groove angle to be shaped are the tilt angle of the substrate and the etching rates of the photoresist and substrate materials. Figure 4.1.2.3 shows cross-sections of holographic gratings before and after ion etching. Blazed holographic gratings are commercially available for use in visible, ultraviolet and soft-X-ray spectroscopic instruments.

References

1. Labeyrie, A. and Flamand, J. (1969). Spectroscopic performance of holographically made diffraction gratings. *Optics Communications*, 1, 5-8.
2. Sheridan, N.K. (1968). Production of blazed holograms. *Applied Physics Letters*, 12, 316-18.
3. Nagata, H. and Kishi, M. (1975). Production of blazed holographic gratings by a simple optical system. *Japanese Journal of Applied Physics*, 14, (Suppl. 14-1), 181-6.
4. Tsang, W.T. and Wang, S. (1975). Preferentially etched diffraction gratings in silicon. *Journal of Applied Physics*, 46, 2163-6.
5. Aoyagi, Y. and Namba, S. (1976). Blazed ion-etched holographic gratings. *Optica Acta*, 23, 701-7.

4.1.3 Diffraction gratings with variable spacing

The concave grating is a diffraction grating with grating grooves on a spherical concave substrate. Since Rowland invented and supplied concave gratings for spectroscopy in the 1880s, the grooves of concave gratings have been straight, parallel, and equally spaced. The concave grating possesses the optical property of both a plane diffraction grating and a concave mirror, so that focused spectral images can be obtained without using any auxiliary optical elements such as concave mirrors or lenses. However, the concave grating cannot avoid various types of aberration such as astigmatism, coma and spherical aberrations; poor image-focusing is therefore intrinsic to the concave grating⁽¹⁾.

In the 1970s, two approaches for reducing the aberrations of concave gratings were introduced, one by holographic and the other by mechanical means. The aberration-corrected holographic concave grating is fabricated by exposing interference fringes of two spherical wavefronts, instead of the plane wavefronts used in fabricating conventional concave gratings^(2,3). A schematic diagram for the fabrication of aberration- corrected concave gratings is shown in Fig. 4.1.3.1. Here the position of the n th groove $P(u, w, l)$ counted from the centre $O(0, 0, 0)$ satisfies eq. (4.1.3.1), where λ_0 is the laser wavelength for the fringe exposure and C and D are the light source positions:

$$n\lambda_0 = (\langle CP \rangle - \langle DP \rangle) - (\langle CO \rangle - \langle DO \rangle) \quad (4.1.3.1)$$

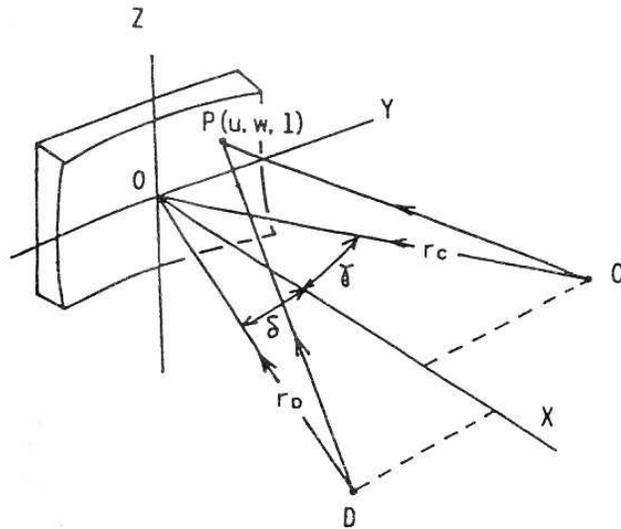


Fig. 4.1.3.1. Optical arrangement for fabrication of aberration-corrected holographic concave grating.

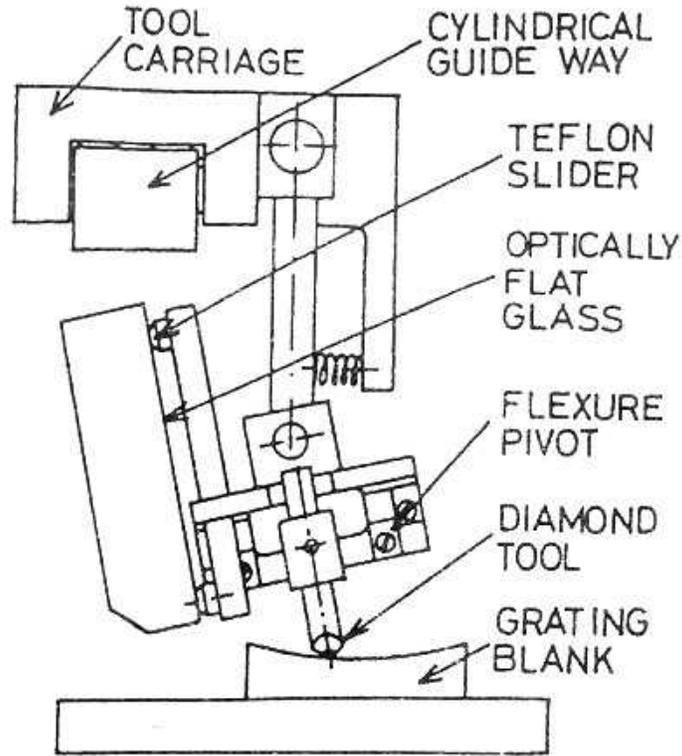


Fig. 4.1.3.2. Mechanism for ruling curved grooves.

The track of P along the spherical surface of the blank substrate is no longer straight, and the increment of groove position ($\partial w / \partial n$) varies with the value of n , so the grooves are curved and variably spaced.

In the design of the aberration-corrected focal condition for a holographic concave grating, the values of four variables — four the distances of the points C and D from O and the incidence angles of laser beams from C and D to O — are chosen so as to minimize each aberration term such as astigmatism, coma or spherical aberration.

The mechanical process for fabricating aberration-corrected concave gratings is to rule curved grooves by reciprocation of the tool in a tilted plane as shown in Fig. 4.1.3.2 and numerical control of blank translation as shown in Fig. 4.1.2.3^(4,5). The relation between the groove number n and the blank translation w_θ is given by eq. (4.1.3.2), where σ_0 is the groove spacing at the centre and R is the radius of curvature of the spherical substrate:

$$n = \frac{1}{\sigma_0} \left(w_\theta + \frac{b_2}{R} w_\theta^2 + \frac{b_3}{R^2} w_\theta^3 + \frac{b_4}{R^3} w_\theta^4 + \dots \right) \quad (4.1.3.2)$$

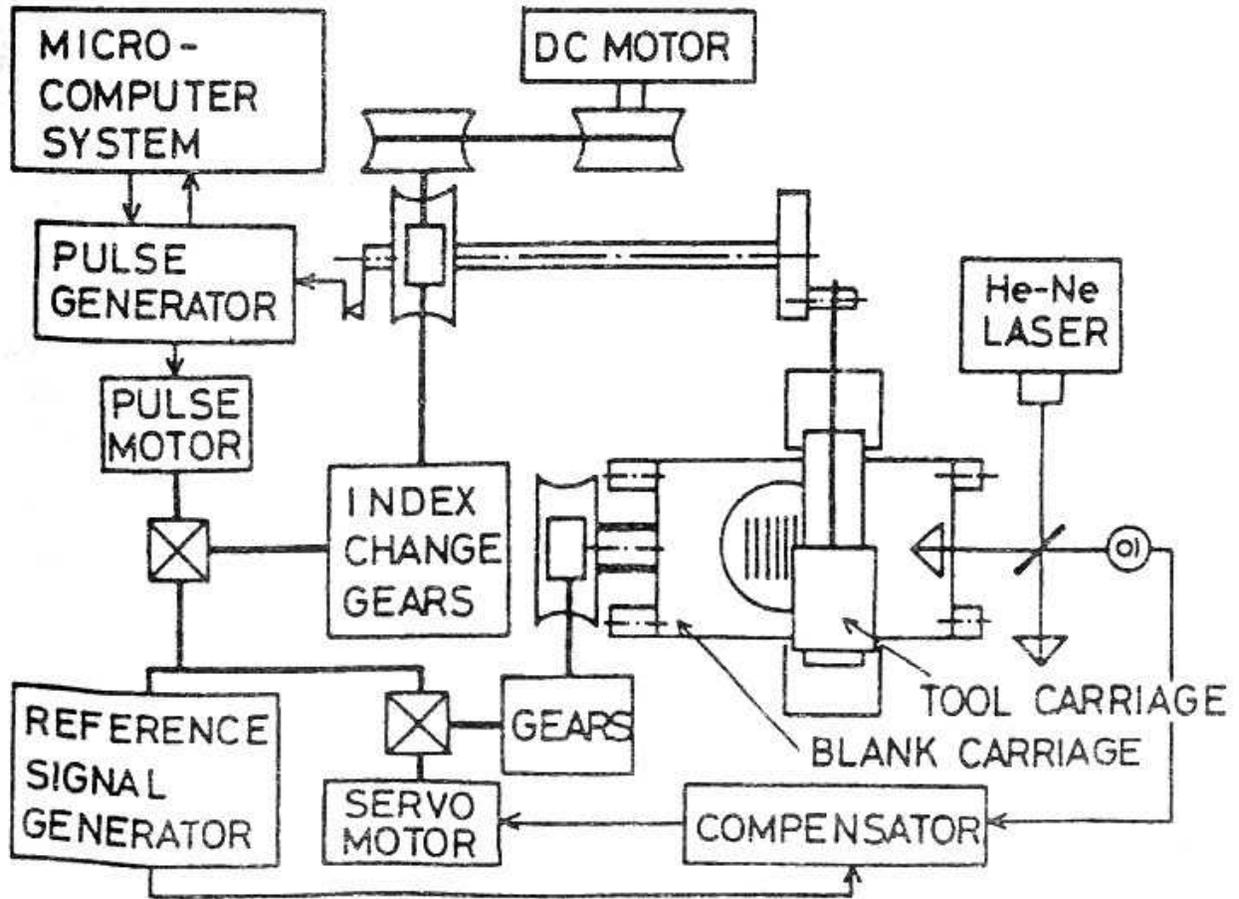


Fig. 4.1.3.3. Numerical control system for variably space grooves.

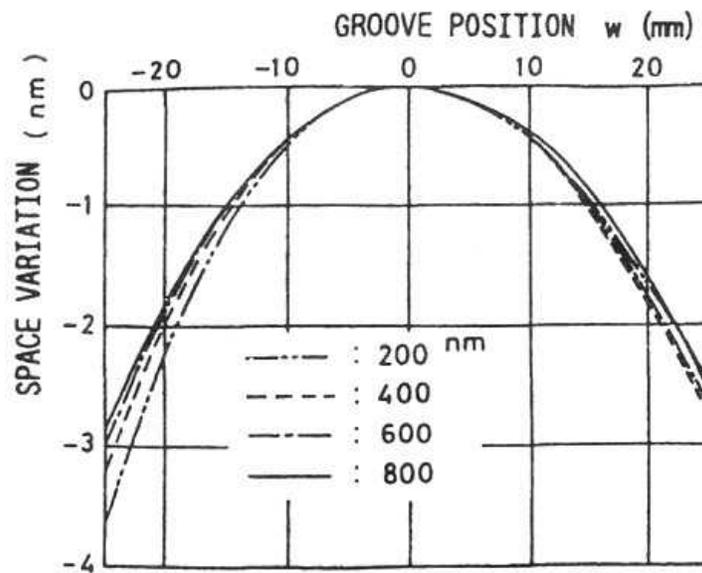


Fig. 4.1.3.4. Space variation of grating grooves for aberration-corrected Seya-Namioka monochromator (600 mm^{-1} , $R = 500 \text{ mm}$).

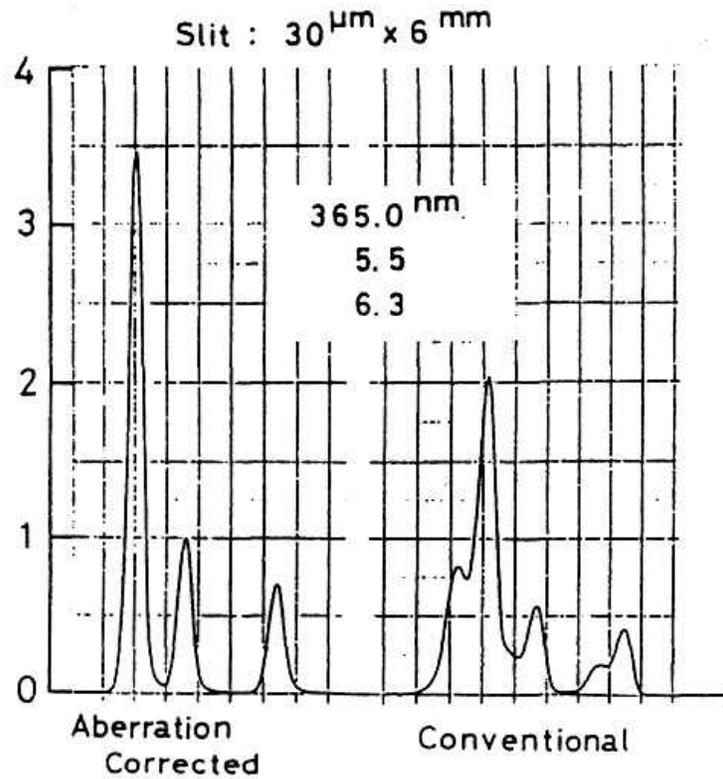


Fig. 4.1.3.5. Comparison of spectral images from Seya - Namioka monochromator (600 mm^{-1} , $R = 500 \text{ mm}$) for the Hg lines at 365.0, 365.5 and 366.3 nm.

The ruling parameters for space variation b_2 , b_3 , b_4, \dots and the tool tilt angle θ for groove curvature are chosen so as to minimize each aberration term for the concave grating. Here, the ruling parameter b_2 and θ are related mainly to focal position and amount of astigmatism, b_3 , to coma, and b_4 to spherical aberration.

An aberration-corrected concave grating to enhance the spectral image-focusing property of the Seya - Namioka monochromator has been designed and fabricated⁽⁶⁾. This monochromator is an instrument for obtaining monochromatic light by simple rotation of a concave grating while both incidence and exit slits are fixed at an angular position of 70° from the grating centre. However, the monochromator cannot avoid large coma and spherical aberration on the spectral image when a conventional concave grating with equally spaced grooves is used. The most appropriate spacing for reducing such aberrations for a concave grating with a radius of curvature of 500 mm and a groove density of 600 mm^{-1} at the centre is shown in Fig. 4.3.4. A concave grating with the ruling parameters to reduce coma and spherical aberration at 400 nm was ruled mechanically.

The spectral images obtained with the aberration-corrected concave grating and a conventional one are compared using the mercury spectrum; the result is shown in Fig. 4.3.5. The effectiveness of a space variation as much as 3 nm is clearly seen.

The nanotechnology for obtaining enhanced image-focusing by diffraction gratings using curved and/or variably spaced grooves has had fruitful results in the field of advanced science and engineering such as astrophysics^(7,8), plasma diagnostics^(9,10), application of synchrotron radiation^(11,12), and optical communication⁽¹³⁾.

References

1. Namioka, T. (1959). Theory of the concave grating. *Journal of the Optical Society of America*, 49, 446-60.
2. Flamand, J., Labeyrie, A., and Pieuchard, G. (1969). Diffraction gratings. US Patent 3,628,849.
3. Noda, H., Namioka, T. and Seya, M. (1974). Geometric theory of the grating. *Journal of the Optical Society of America*, 64, 1031-6.
4. Harada, T., Moriyama, S., and Kita, T. (1974). Mechanically ruled stigmatic concave gratings. *Japanese Journal of Applied Physics*, 14, (Suppl. 14-1), 175—9.
5. Noda, H., Namioka, T., and Seya, M. (1974). Design of holographic concave gratings for Seya — Namioka monochromators. *Journal of the Optical Society of America*, 64, 1043-8.
6. Kita, T. and Harada, T. (1980). Mechanically ruled aberration-corrected concave grating for high resolution Seya — Namioka monochromator. *Journal of the Spectroscopy Society of Japan*, 29, 256-62.
7. Hettrick, M., Bowyer, S., Malina, R.F., Martin, C., and Mrowka, S. (1985). Extreme Ultraviolet Explorer. *Applied Optics*, 24, 1737-56.
8. Harada, T., Kita, T., Bowyer, S., and Hurwitz, M. (1991). Design of spherical varied line-space gratings for a high resolution EUV spectrometer. *SPIE Proceedings*, 1545, 2-7.
9. Kita, T., Harada, T. Nakano, N., and Kuroda, H. (1983). Mechanically ruled aberration-corrected concave gratings for a flat-field grazing incidence spectrograph. *Applied Optics*, 22. 512-13.

10. Nakano, N., Kuroda, H., Kita, T., and Harada, T. (1984). Development of a flat-field grazing incidence XUV spectrometer and its application in picosecond XUV spectroscopy. *Applied Optics*, 23, 2386-92.
11. Harada, T., Kita, T., Itou, M., and Taira, H. (1986). Mechanically ruled diffraction gratings for synchrotron radiation. *Nuclear Instruments and Methods in Physical Research*, A246, 272-7.
12. Koike, M., Harada, Y., and Noda, H. (1987). New blazed holographic grating fabricated by using an aspherical recording with an ion-etching method. *SPIE Proceedings*, 815, 96-101.
13. Okai, M. and Harada, T. (1991). Novel method to fabricate corrugation for distributed feedback lasers using a grating photomask. *SPIE Proceedings*, 1545, 218-23.

4.2 Nano-lithography

4.2. Photolithography

4.2.1 Introduction

Since photolithography was first applied to semiconductor circuit fabrication, the performance of semi-conductor circuits has been extended up to ULSI. Photolithography is the most important and key technology in the semiconductor fabrication system.

Photolithographic technology has been improved in accordance with the demands of higher circuit integration, and now lines of width several hundreds of nanometres have been fabricated by photolithography. The light source for photolithography mainly determines the resolution achievable. The wavelength of the light source has been successively shortened, from the g-line (436 nm) of the high-pressure mercury arc lamp to its i-line (365 nm), the 248 nm radiation of the KrF excimer laser, and now the 193 nm radiation of the ArF excimer laser.

The optical wafer stepper, referred to simply as the stepper, is used on almost all production lines for mass production of ULSI as the photolithographic device. The stepper involves many of the most advanced component technologies, including nanotechnology.

4.2.1.2 Optical configuration of the stepper

The optical system of the stepper is shown in Fig.4.2.1.1. The key component is the projection lens for imaging the mask pattern on to the wafer with some reduction ratio. Theoretically, the resolution of the lenses is defined by the formula

$$R = K \cdot \lambda / NA \quad (4.2.1.1)$$

where R is the resolution (nm), AT is a process factor, a constant defined by the process of pattern duplication and equal to 0.8 under production conditions and 0.65 in R & D, λ is the wavelength of the light source (nm), and NA is the numerical aperture ($= 1 / \cos \theta$, where θ is the angle formed by the optical axis and the outermost light beam to the image). Hence the resolution of the optical system is higher, the shorter the wavelength or the larger the NA. The development of lens design and fabrication technology has improved the lens system so that shorter wavelengths and a larger NA can be used. The performances of typical lens systems supplied from 1981 to 1993 are shown in Table 4.2.1.1.

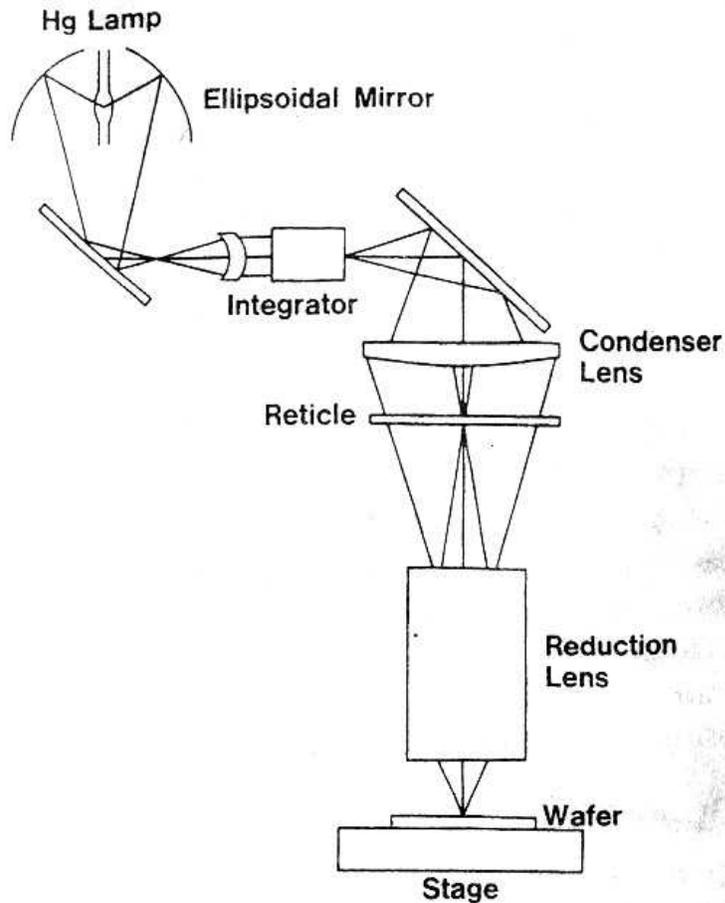


Fig. 4.2.1.2. Optical system of water stepper.

A lens system is made up of about 30 component lenses, each of which has a maximum diameter of 250 mm and a maximum mass of 10 kg. The total mass of the whole lens system is up to 500 kg.

A stepper which uses an excimer laser as the light source for exposure is called an excimer stepper. The excimer laser is used to obtain higher resolution, having a shorter wavelength in the far ultraviolet region than the g-line or i-line in the ultraviolet region. The basic configuration of the excimer stepper is almost the same as the ordinary stepper except for the light source.

A KrF excimer laser (248 nm) is now in use for test production of ICs, and an ArF excimer laser (193 nm) is under development. It is necessary to involve features such as a narrow waveband of the radiation generated high stability of light power, and long operating life when using an excimer laser for photolithography. A kind of optical monochromator is inserted in the cavity of the laser to achieve a narrow bandwidth. Some characteristics of a typical excimer laser

for photolithography are listed in Table 4.2.1.1. In the application of the excimer laser, optical spatial pattern noise, called speckle noise, is generated

Table 4.2.1.1 Performance of typical stepper lens systems

Type	Delivery date	Resolving power (gm)	NA	Spectral line	Demagnification	Exposure area (mm)
NSR-1505G	1981	1.2	0.30	g	1/5	15
-1505G2A	1984	1.0	0.35	g	1/5	15
-1505g4D	1987	0.75	0.45	g	1/5	15
-1505EX	1988	0.75	—	KrF	1/5	15
-1505G6E	1988	0.65	0.54	g	1/5	15
-1505i6A	1989	0.65	0.45	i	1/5	15
-2005G8C	1990	0.55	0.60	g	1/5	20
-2005EX8A	1992	0.4	0.50	KrF	1/5	20
-2005i9C	1993	0.35 ^a	0.57	i	1/5	22

a With SHRINC1 option (see subsection 4.2.1.6)

Table 4.2.1.2 Characteristics of excimer laser for the excimer stepper

Output power	4-6 W
Repetition frequency	400-600/Hz
HzWavelength band width	1.2-2.5 pm
Wavelength stability	±0.3-0.5pm
Gas life	8x10 ⁴ -5x10 ⁷ pulses
Laser window maintenance	2-5 x10 ⁸ pulses
Power life	1x10 ⁹ pulses
Laser tube maintenance	6x10 ⁸ -2 x10 ⁹ pulses
Monochromator maintenance	3-4 x 10 ⁹ pulses

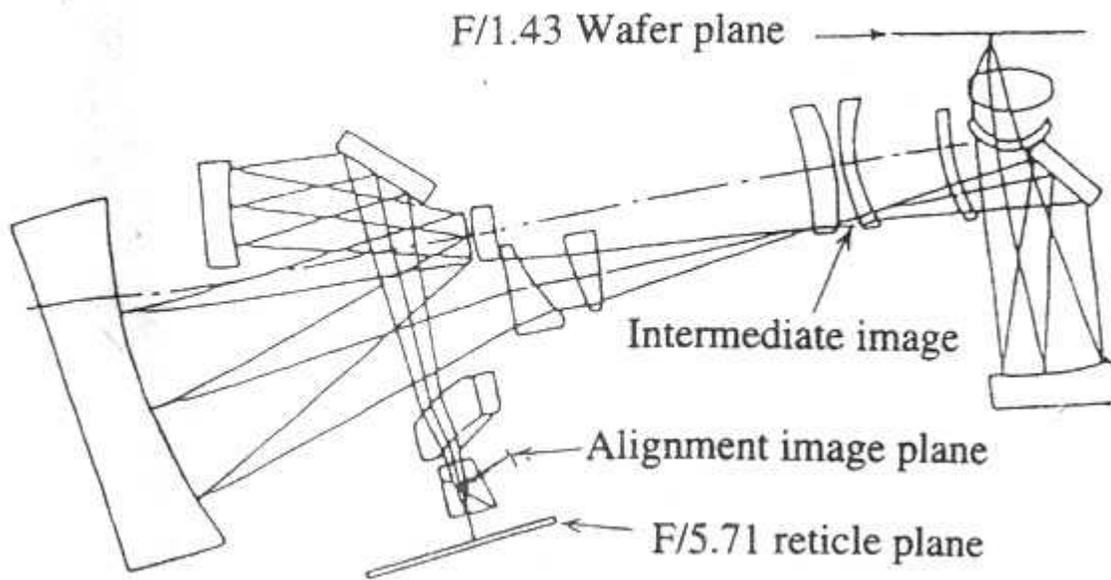


Fig. 4.2.1.2. Optical system of mirror projection optics (Micrascan, Perkin-Elmer Co.).

by small dust particles in the light path, owing to the high coherence of the laser beam⁽¹⁾. A laser beam direction distributor should be inserted in the light path in the illumination system to reduce coherence. In some systems, the distributor is a swinging mirror which is driven synchronously with the excimer laser pulses. Multiple exposure with different directions of illumination hides the speckle noise pattern.

Shorter wavelengths allow less material to be used in excimer laser projection lenses, owing to the low transparency of the materials at the excimer laser wavelength. Typical optical glasses—green soda-lime glass and white crown glass (BK7) - cannot be used in the ultraviolet region. Quartz (SiO_2) and fluorspar (CaF_2) are possible candidates.

A combined system using reflecting mirrors has been developed to reduce the number of lenses. An example is shown in Fig. 4.2.1.2. A problem is the small exposure area with this system, so scanning methods have to be used, as in subsection 4.2.1.5.

4.2.3 Alignment system

Integrated circuits are fabricated by applying some 10 to 15 different pattern masks for the multilayered structure. The alignment between a previously exposed pattern on a wafer and the

succeeding pattern on a mask that will be exposed on the wafer is a critical factor determining the minimum pattern width.

The stepper has a number of alignment systems, each with a certain attainable accuracy. Wafer pre-alignment is achieved by means of two rolling pins fitting the facet of the wafer within $\pm 3 \mu\text{m}$

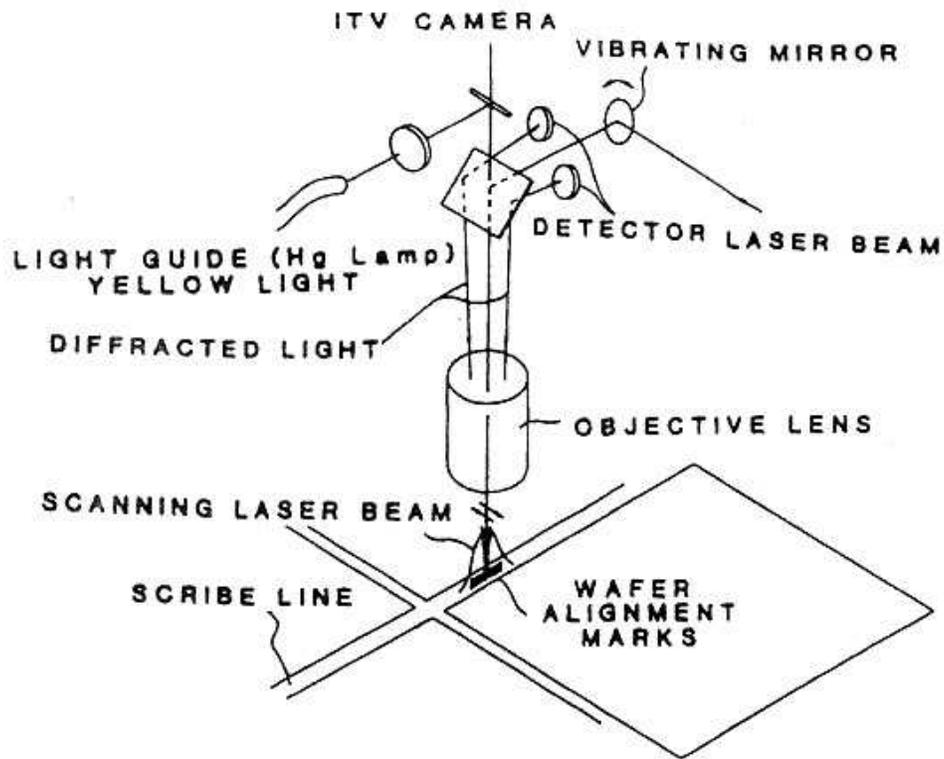


Fig. 4.2.1.3. Principle of a wafer alignment system.

mechanically. The wafer alignment system aligns the wafer itself to the reference line of the stepper stage motion; this is also called global alignment. For more accurate alignment, chip alignment is applied. Each chip on the wafer is aligned with the optical axis of the stepper lenses or mask alignment mark, depending on the alignment method.

Essentially, alignment is achieved by detection of the centre of the alignment pattern and adjustment of the position of either wafer or mask so as just to overlap each other. The detection of the centre of the alignment pattern is analogous to the technique used to detect the centre of the line on the standard scale of a photoelectric microscope. The same technique is used in the alignment system of IC steppers, although some improvement has been achieved.

An example of the wafer alignment system is shown in Fig. 4.2.1.3. A beam from an He-Ne laser illuminates the pattern on the wafer through an alignment microscope. A vibrating mirror in the

light path as shown modulates the illuminating position on the wafer sinusoidally. The signal reflected from the wafer varies, depending on the reflectivity of each point on the pattern and the unpatterned surface, and the waveform of the signal changes according to the relative positions of the pattern and the centre of deflection of the vibrating light beam.

The signal from a photo detector is amplified and fed to a phase-sensitive detector (PSD). The discriminated signal varies depending on the displacement of the alignment pattern from the axis. The centre of the pattern can be estimated from the zero-voltage point of the signal. The position-sensing repeatability of this method is within 10 nm. The alignment patterns are formed as a grating as shown in Fig. 4.2.1.3 to obtain a better signal-to-noise ratio.

An example of the chip alignment system is shown in Fig. 4.2.1.4. The laser beam illuminates the chip alignment pattern through the projection lens. The diffracted light from the wafer returns along the same optical path and is split by the beamsplitter to the detector. The spatial filter stops the zero-order diffracted light to eliminate effects due to the strong beam containing error factors through surface roughness or pattern irregularity.

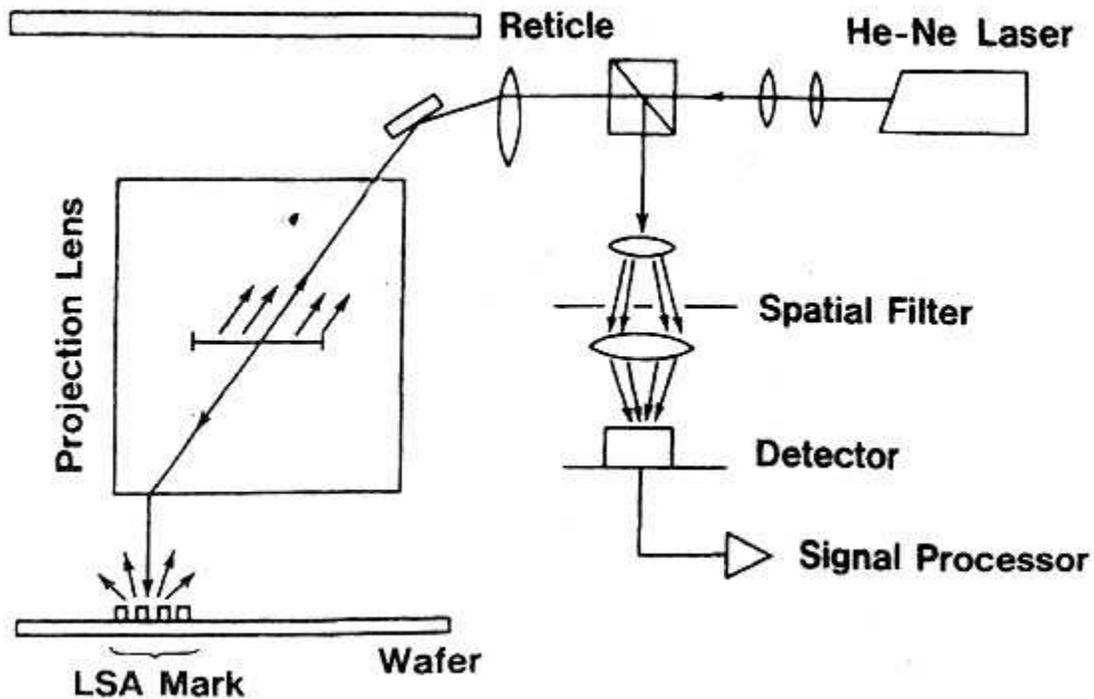


Fig. 4.2.1.4 Principle of a chip alignment system.

The mark is scanned by the stage motion. The signal from the detector shown in Fig. 4.2.1.5 is interpolated by the fringe signal of the stage interferometer, allowing the accurate centre of the pattern to be detected. The total configuration of the alignment system is shown in Fig. 5.2.6.

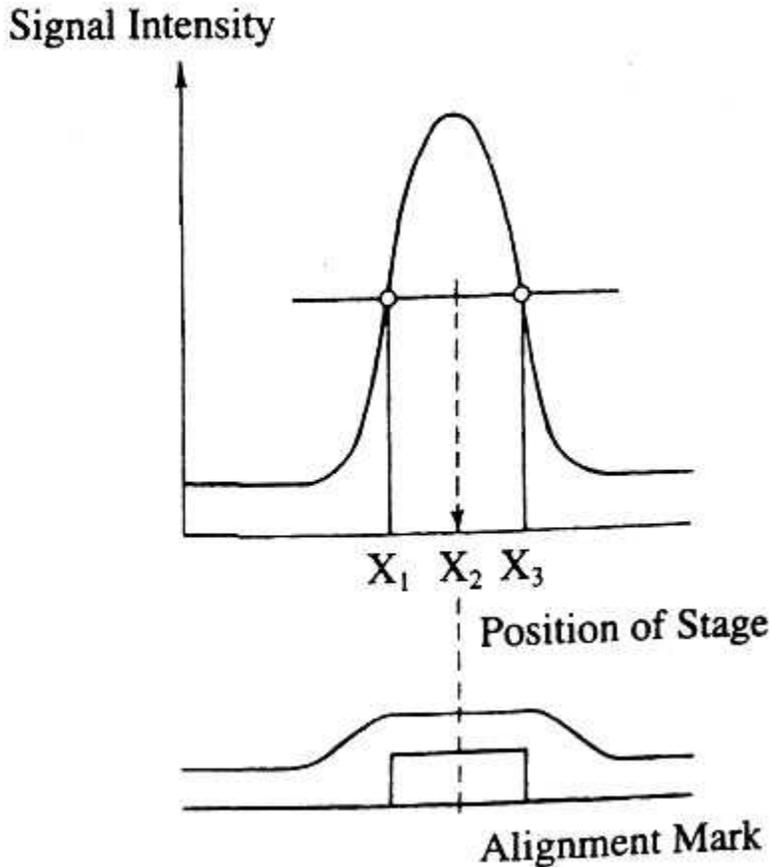
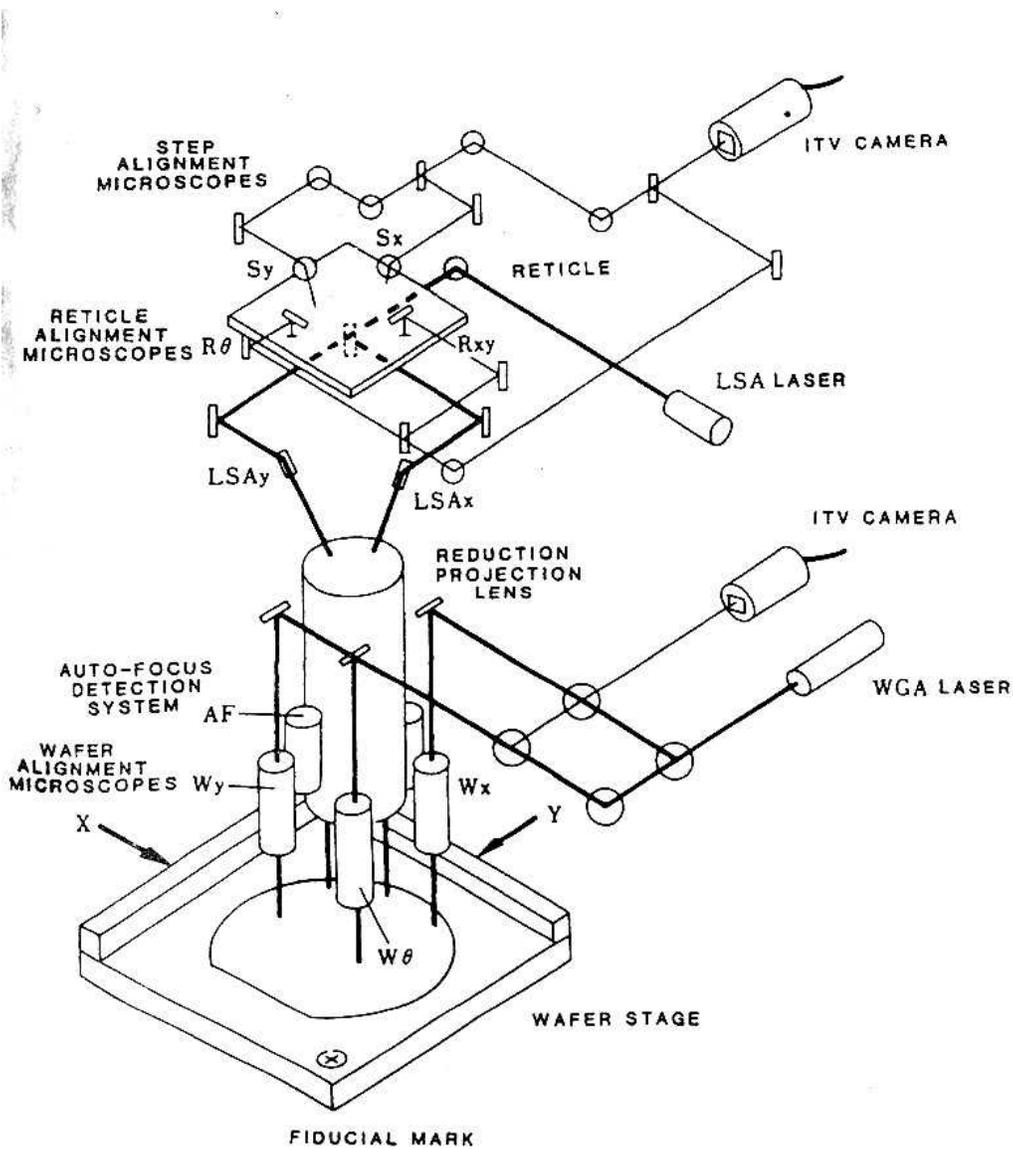


Fig. 4.2.1.5. Chip alignment signal processing.

As already mentioned, the alignment accuracy determines the minimum pattern width of ICs. Improvement of the alignment system is a crucial factor in developing higher-resolution steppers.

A new alignment sensor has been devised using a diffraction and frequency-heterodyning method⁽²⁾. The configuration of the laser interferometric alignment (LIA) system is shown in Fig. 4.2.1.7, and the optical signal processing principle in Fig. 4.2.1.8. Two different laser beams are modulated by an AOM (acoustic optical modulator) with frequencies f_1 and f_2 . The two beams are reflected through the projection lens below the reticle and imaged on to the wafer. The alignment pattern in grating form on the wafer is thus illuminated by two laser beams from two

different directions as shown in Fig. 4.2.1.8. The beams diffracted by the grating, +1 order of f_1 and -1 order of f_2 interfere and frequency-heterodyning takes place.



From phase modulation theory, the phase of the heterodyne frequency $f_h (f_h = f_1 - f_2)$ varies according to the displacement of the relative positions of the alignment pattern and the laser beam. On the other hand, the reference signal f_r is processed by the reference signal generator as shown in Fig. 4.2.1.7. The phase difference between f_r and f_h corresponds to the displacement of

the alignment mark from the fiducial position. The stage is moved to reduce the phase difference to zero and thus achieve alignment.

This heterodyne method allows greater disturbance to be tolerated from surface roughness of the mark, low step dimension of the mark, and high reflectivity of materials such as aluminium. The alignment accuracy has been found to be from 43 to 85 nm for various alignment marks. The results of an alignment test are shown in Fig. 4.2.1.9 for two types of alignment mark.

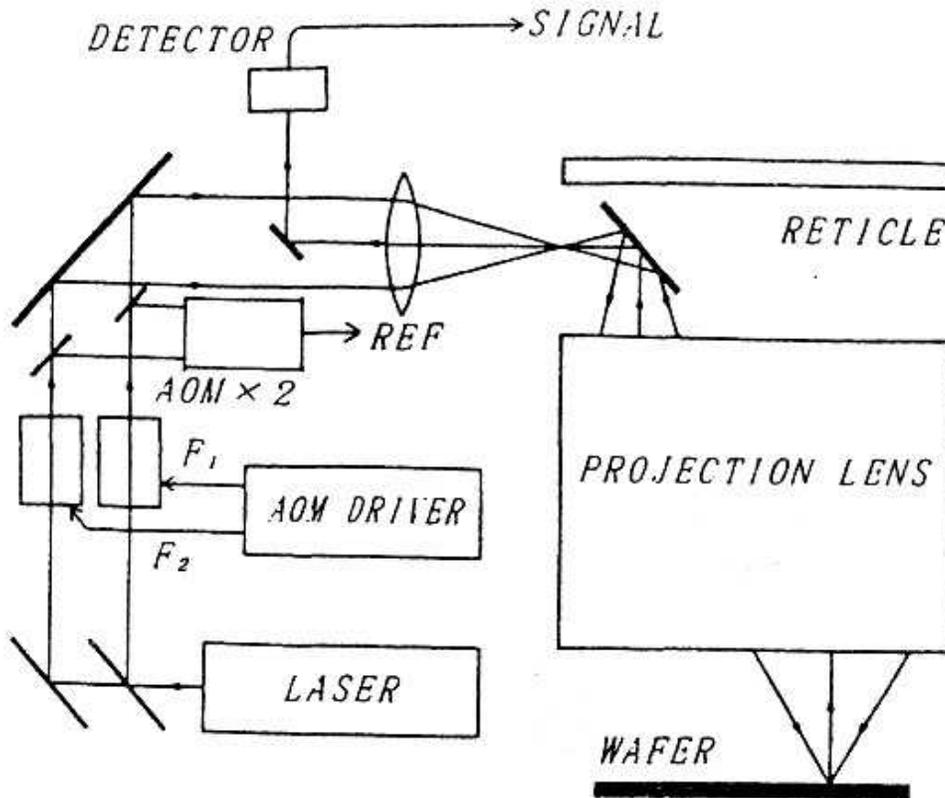


Fig. 4.2.1.7. Interferometric alignment system.

4.2.4 Autofocus, auto-levelling system

The depth of focus at which a specific resolution can be obtained is defined by

$$DOF = k \cdot 1 / \lambda \cdot (NA)^2 \quad (5.2.2)$$

Since high-NA lenses are used for high-resolution imaging, focusing is very critical. Line-width variation with out-of-focus displacement is shown in Fig. 4.2.10. Autofocusing systems are used in almost steppers. The

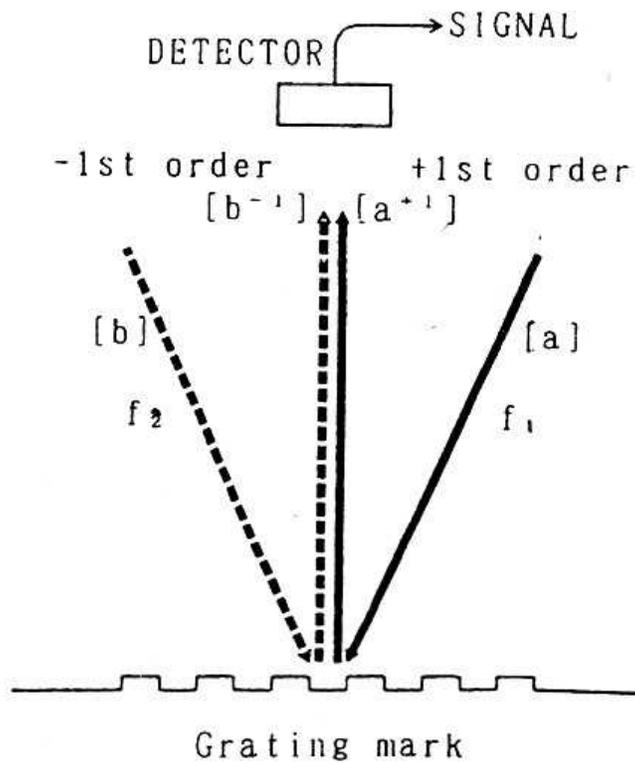


Fig. 4.2.8. Principle of LIA optical signal processing.

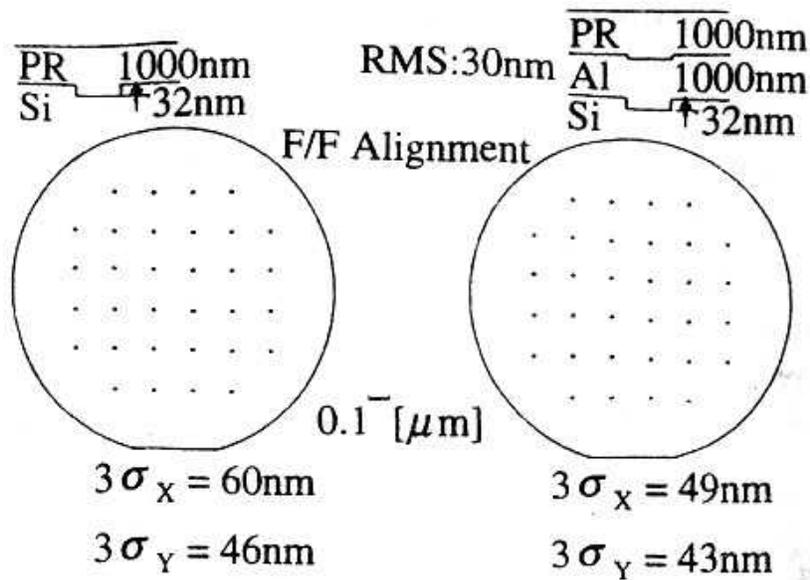


Fig. 4.2.9. Results of LIA tests.

principle of the stepper autofocus system is shown in Fig. 4.2.11. This system is an application of the photoelectric microscope used for pattern alignment. A photo detector senses the horizontal

shift of the projected line of the laser beam on the wafer depending on the vertical position of the wafer. The error signal from the photoelectric sensor is fed back to a control amplifier to drive the motor positioning the wafer stage on its vertical axis. This error signal decreases to zero as the vertical position approaches and finally reaches the focus point. The focus point can be controlled within 0.1 μm .

Hitherto the level of the wafer surface has been regulated only once, at the start, before exposure. However, a large wafer with a diameter of up to 200 mm cannot be allowed only one initial adjustment of level, since the surface is not at the same height throughout. Hence a level control system for adjustment of each chip has been designed for a recently developed stepper which has a high resolving power and can handle large wafers. The principle of the system is that of the auto-collimator. The system may be imagined from the autofocusing system shown in Fig. 4.2.11 with the right-hand section of the autofocusing system replaced by the projection side of the auto-collimator and the left-hand section by its receiving side. The wafer corresponds to the mirror of the auto-collimator.

4.2.5 Mechanical stage for wafer stepping and alignment

The stage carrying the wafer needs to have both speed and high positioning accuracy to attain high resolution and high throughput. Positioning accuracy is determined by two major technical components: the accuracy of construction of the mechanical stage, and that of position control.

The wafer supported by the stage is very light, but the required accuracy of wafer positioning means that the stage must be very rigid and heavy: its mass is up to 50 kg. The moving table is driven by an electric motor through a precision lead screw and nut system. The guide system is composed of a V-flat guide and needle bearing or a bar and roller guide and double flat needle bearing. In an early state of the development, the stage was composed of coarse and fine tables driven by two motors, one for each table. However, the need for high throughput led to a change to a one-motor system to reduce positioning time. The accuracy and performance of the stage system are shown in Table 4.2.3.

Position control is achieved as follows; a position command is given by a computer and the stage position is sensed by a laser interferometer with feedback to the controller, which drives the motor until the error between command and stage position is reduced to zero. After positioning is completed,

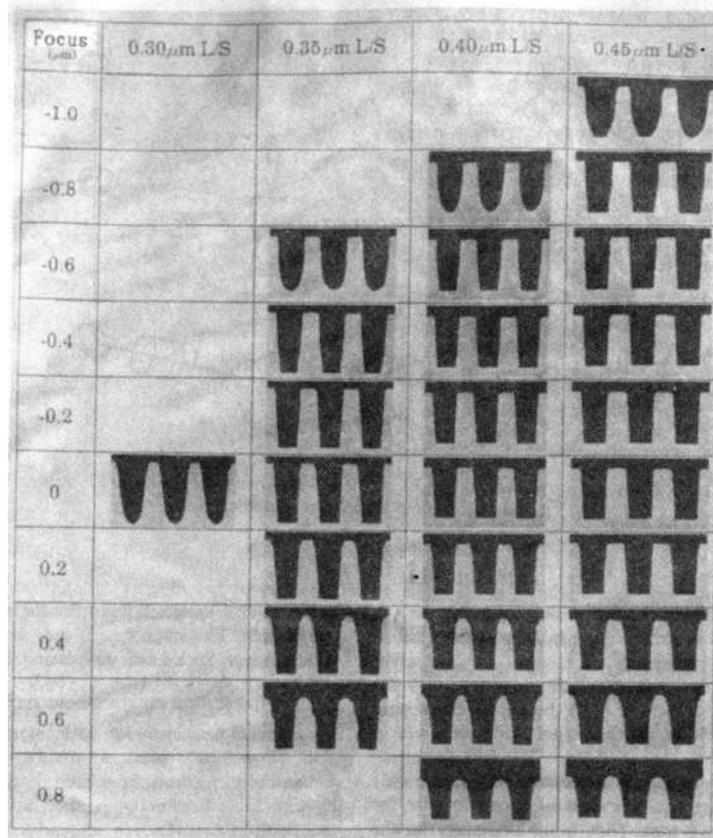


Fig. 4.2.10. Line patterns as a function of focusing conditions (i-line, 365 nm, NA 0.57, photo-resist PFi-28 1.06 μm , normal incidence illumination) for some line and space (L/S) values.

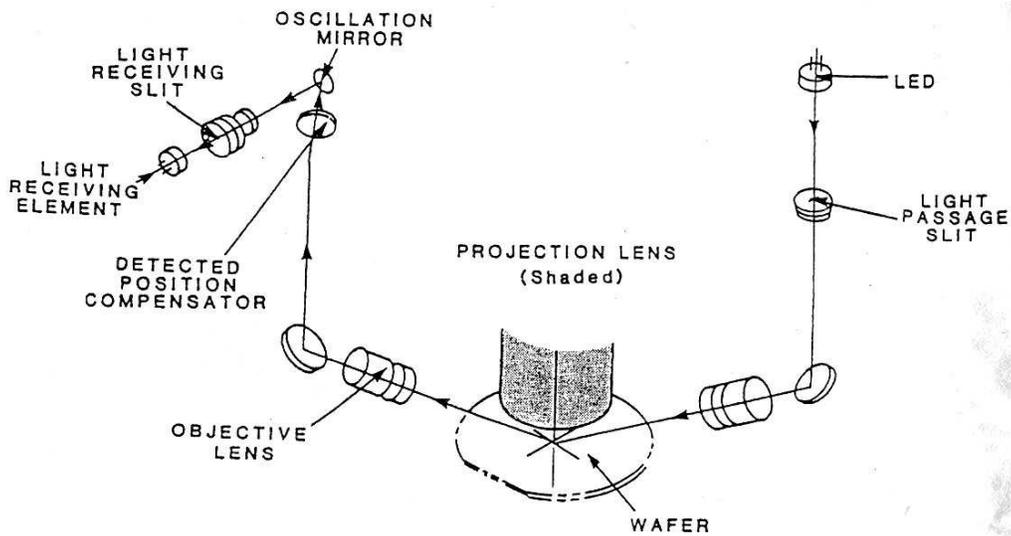


Fig. 5.2.11. Configuration of an autofocus system.

focusing, alignment, and exposure are achieved for one chip. This process is repeated step by step, hence its name 'step-and-repeat'.

With a higher-resolution stepper, the step-and-repeat system cannot be used, because the exposure area is not able to cover a whole chip area with a resolution better than 0.25 μm . The lens or mirror projection system will be able to achieve imaging only in the slit zone covering part of a chip of size for example 22 x 4 mm. Then a slit-like image on the mask is projected on to the wafer, and the full image of the mask is exposed after scanning of the mask and wafer synchronously over a chip. For this purpose, a

Table 4.2.3 Performance and accuracy of stepper stage (single-motor drive)

Yaw	$\pm 0.5 \text{ arcsec} / 200 \text{ mm}$
Pitch	$\pm 1.0 \text{ arcsec} / 200 \text{ mm}$
roll	$\pm 1.0 \text{ arcsec} / 200 \text{ mm}$
Orthogonality	0.2 arcsec (compensated)
Maximum speed	100 mm/s^{-1}
Positioning accuracy	0.07 μm (3σ)
Positioning time	0.4 s/20 mm

step-and-scan system is used for the stage system of the stepper. The stage system may be replaced by a different type, such as an air-bearing stage.

4.2.6 Resolution enhancement technology

To obtain higher resolution with a normal light source and projection lenses, a number of resolution enhancement methods have been developed.

A modified illumination method has been developed and used in a production stepper(3). When a narrow line and space pattern is illuminated, transmitted light is diffracted by the slit composed of lines as shown in Fig.5.2.12(a) according to Fresnel diffraction theory. The diffraction angle θ is given by

$$\sin \theta = \lambda / p \quad (5.2.3)$$

where λ is the wavelength of the illuminating light and p is the pitch of the lines and spaces.

If p is larger than the mask-side aperture of the projection lens, zero-order light and diffracted ± 1 - order light pass through the lens and focus the image of the mask on the wafer. But if p is

smaller than the critical value for the aperture, the ± 1 -order light beam is stopped by the aperture. The focused image is impaired because the ± 1 -order light beam does not contribute to the image. However, if the illuminating angle is made oblique as shown in Fig. 4.2.12(b), the

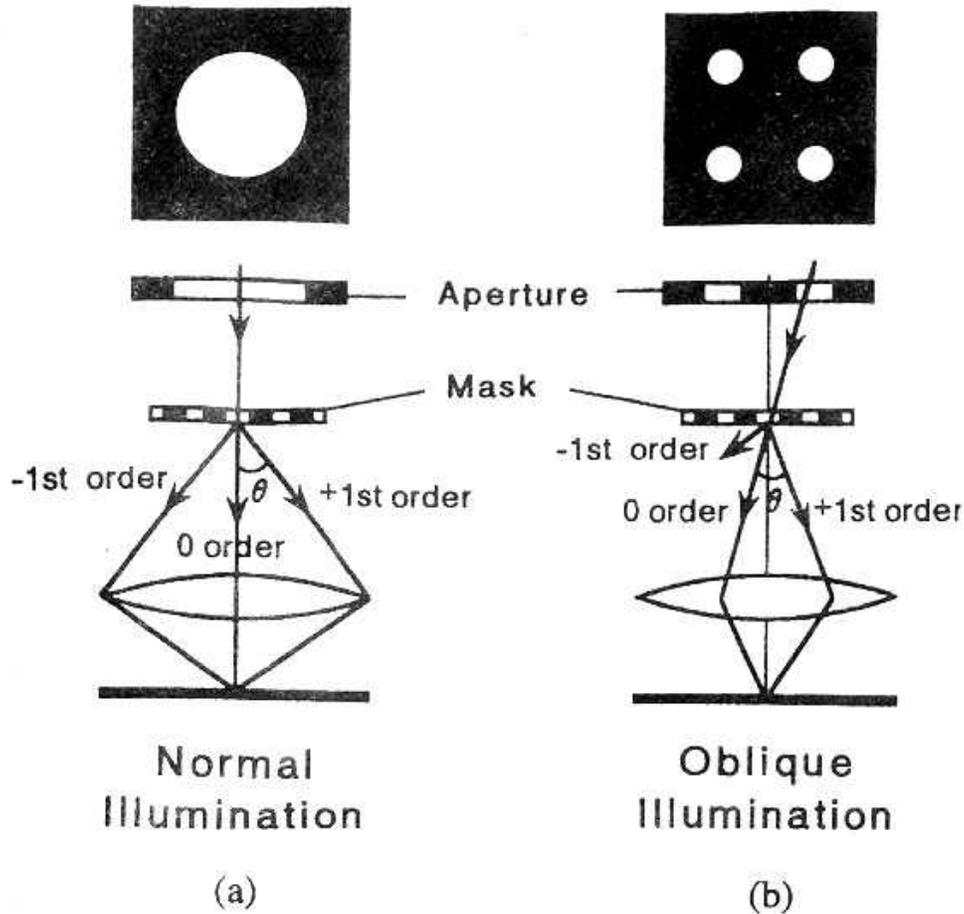


Fig. 4.2.12. Principle of the oblique illumination system.

zero-order light and either the +1-order or the -1- order light can pass through the aperture. The image is then better than that composed by only the zero-order light. Oblique illumination is achieved by means of the aperture plate shown in Fig. 4.2.12(b). This method has been applied in a Nikon commercial stepper,

under the acronym SHRINC (super-high resolution by illumination control).

Another resolution enhancement method is the phase-shifting mask method(4). A phase-shift mask can be formed with a line and space pattern and a phase-shifter covering the spaces in the pattern as shown in Fig. 4.2.13(a). The phase of the light passing through the phase-shifter is

changed by 180° . As a result, the contrast of the light intensity on the wafer is better than that with no phase shifter, as shown in Fig. 4.2.13(b).

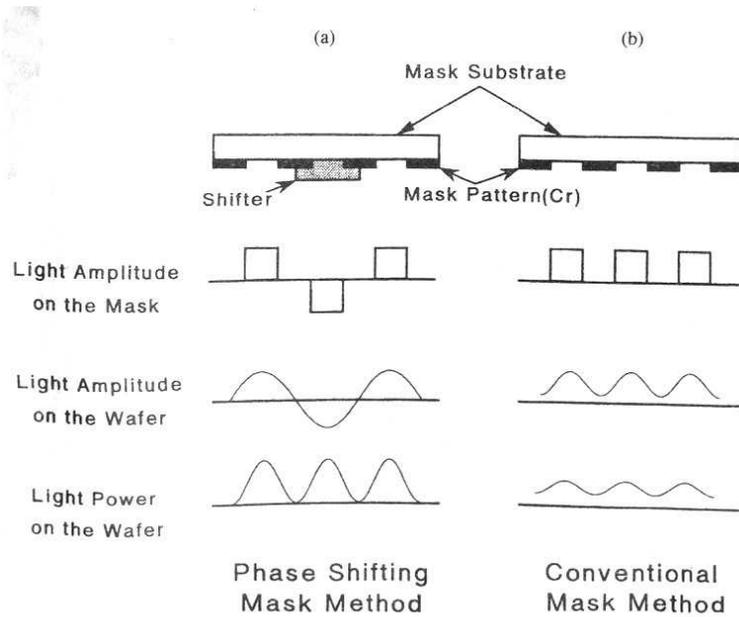


Fig. 4.2.13. Principle of the phase-shifting mask method.

References

1. Bennewitz, P.V. et al. (1986). Excimer laser based lithography. SPIE Proceedings, 633, pp. 6-16.
2. Ota, K. et al. (1991). New alignment sensor for wafer stepper. SPIE Proceedings, 1463, p. 307.
3. Shiraishi, N. et al. (1992). New imaging technique for 64M-DRAM. SPIE Proceedings, 1674, pp. 741-52.
4. Levenson, M.D. (1982). Improving resolution in photo-lithography with a phase-shifting mask. IEEE Transactions on Electron Devices, ED-29, pp. 1828-36.

4.3 Electron beam lithography

4.3.1 Introduction

Electron beam (EB) lithography has been an essential mask fabrication technology for ULSI devices since Bell Laboratories developed a high-throughput and reliable EB system, EBES1⁽¹⁾. Mask fabrication technology is now very important for ULSI miniaturization, because a phase-shifting mask improves the resolution of optical lithography.

It has been forecast that progress in the miniaturization of ULSI devices achieved by the revolution in optical lithography will result in a 1 Gbit DRAM (dynamic random access memory) with a size of 0.15 μm by about the year 2000. However, since optical lithography for a size of 0.25 μm is still under development, optical lithography is not always used in the development of 256 Mbit to 1 Gbit DRAM devices. Although the EB system throughput is very low, it is sufficient for use in R & D on devices because of its high-resolution capability.

The problem is that it is impossible for optical lithography to fabricate a pattern smaller than 0.15–0.1 μm , owing to its resolution limitation. EB lithography, in addition to X-ray lithography, is a promising technique for such smaller patterns. The key features to be developed for a production-stage EB system are throughput and writing accuracy.

4.3.2 EB lithography for masks

Figure 4.3.1 shows a lithography scheme. Nowadays the main lithography technique is for an optical projection printing machine to duplicate a pattern from a mask of severalfold magnification (reticle) on to a Silicon (Si) wafer. The reticle or mask is fabricated by the EB writing process. So far, a Gaussian beam system has been used for reticle-making in most mask shops. Figure 4.3.2 shows an example of a writing method for a reticle EB system, EBM-130/40, produced by Toshiba Machine Co⁽²⁾. The stage moves continuously in one direction, while the beam is scanned in a direction perpendicular to the movement. LSI data are divided into addresses represented by a 1 or a 0. The electron beam is on or off according to whether the address is 1 or 0 respectively. The data transfer rate is 40 MHz. The Gaussian beam spot size has to be reduced for ULSI miniaturization, as shown in Table 5.3.1. The throughput is $> 1 \text{ h}^{-1}$ for a 4 Mbit DRAM-class reticle when writing with a 0.5 μm spot size. However, the throughput decreases to 0.05 h^{-1} for a 64 Mbit DRAM-class reticle, since the spot size has to be reduced to 0.1 μm . Furthermore, the reticle writing speed is reduced by proximity effect correction such as a GHOST exposure method⁽³⁾, by an increase in reticle size from 125 to 150 mm and also by

phase- shifting mask writing, which requires about double exposure. A high-speed reticle writing system is therefore strongly required. MEBES (ETEC Co.) has been developed as a high-speed reticle writing system which adopts a Zr-O-W thermal field-emission electron gun⁽⁴⁾. Figure 4.3.3 shows the electron optical column of MEBES. The spot size is 0.1 μm , the current density 400 A cm^{-2} at an acceleration voltage

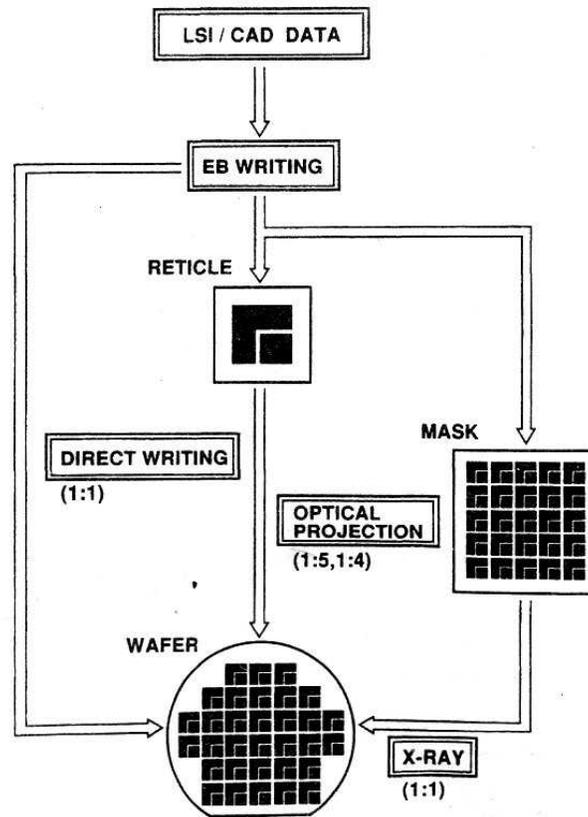


Fig. 4.3.1. Lithography scheme

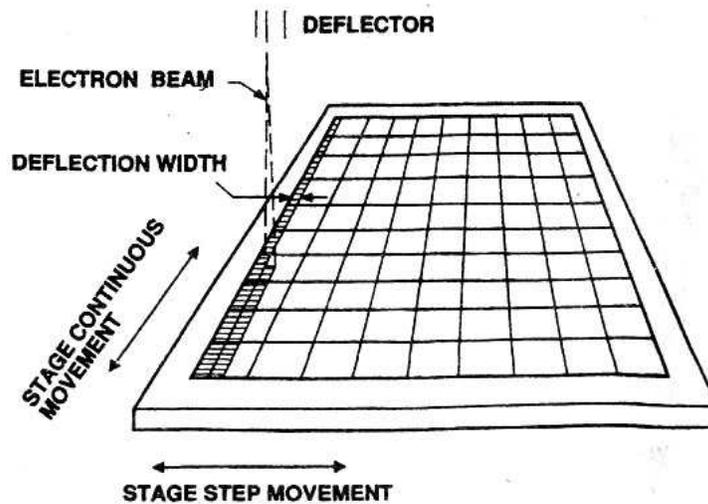


Fig. 4.3.2. Writing method for reticle writing system EBM 130/40.

Table 4.3.1 Throughput of the EBM-130/40

DRAM capacity	Reticle pattern size (x5) (μm)	Beam diameter (μm)	Throughput (h^{-1})
4M	3.5	0.5	1.3
16M	2.5	0.25	0.3
64M	1.5	0.1	0.05
256M	1.25	0.05	0.0125
1G	0.75	0.05	0.0125

of 10 kV, and the data transfer rate 160 MHz. The writing speed of MEBES is four times that of EBM- 130/40, resulting in a reticle writing time of ~ 5 h for a 64 Mbit DRAM class.

The EBES4 developed by Bell Laboratories is similar in concept, using a thermal field-emission electron gun⁽⁵⁾. The spot size is $0.125 \mu\text{m}$, the current density 1600 A cm^{-2} at an acceleration voltage of 20 kV, and the data transfer rate 230 MHz.

Another method of achieving a high throughput system is to adopt the variably shaped beam (VSB) concept. The writing method for the EX-8 (Toshiba Co.) is shown in Fig. 4.3.4, which adopts VSB, a continuously moving stage, and vector scanning (beam flies from pattern to pattern⁽⁶⁾). The EX-8 has the ability to write a 1 Gbit DRAM-class reticle pattern, because the address size, corresponding to the beam size in a Gaussian beam system, is 0.01 μm . The EX-8 can generate triangular beams and rectangular beams from 0.1 to 2.56 μm by using a keyhole-type second shaping aperture. LSI patterns are composed of slanting-angle patterns in addition to x - y patterns. A conventional VSB-type EB system approximates a slanting angle by small rectangles. On the other hand, the EX-8 writes a slanting-angle pattern with a combination of triangular and rectangular beams, which results in a higher throughput than with conventional VSB-type EB systems. The throughput is $\sim 2 \text{ h}^{-1}$ for a 64 Mbit DRAM-class pattern. Moreover, the EX-8 is equipped with an alignment function using marks on a glass plate, which enables it to write a Levenson-type phase-shifting mask⁽⁷⁾. It is also equipped with a fully automatic glass plate loading system, as shown in Fig. 4.3.5. A robot takes a glass plate from a cassette magazine and puts it into the I/O chamber. After the I/O chamber has been evacuated, the glass plate moves to the loading chamber, from where a shuttle takes it to the writing chamber.

Writing accuracy, as well as throughput, is very important for a reticle writing system. The required reticle accuracy is indicated in Table 4.3.2. Figure 4.3.6 shows pattern errors appearing in a pattern written by a VSB and continuous stage-moving system. The patterns dimensional accuracy is limited by the resist process in addition to problems inherent in EB lithography, for example the proximity effect and resist heating. Stitching accuracies are composed of shaped beam stitching, sub-field stitching caused by sub-deflection, and stripe stitching caused mainly by the main deflection and stage movement. The improvement of stripe stitching accuracy is the most difficult among stitching accuracies. A stripe stitching error arises from residual distortion of the main deflector after distortion correction, stage attitude control error, mechanical vibration between the electron optical column and the substrate, electrical noise, and beam drift. Moreover, stripe stitching overlaps beam stitching and sub-field stitching. The only way to reduce stripe stitching error is to discover the origin of the error factors and to eliminate them one by one. Long-range dimensional accuracy is related to the difference in writing position from the ideal position. The accuracy has to be better than 5×10^{-7} , since the required long-range dimensional accuracy is $\sim 0.05 \mu\text{m}$ per 100 mm.

Furthermore, long-term stability of accuracy is required. It needs great effort to maintain an accuracy of 5×10^{-7} for a long time⁽⁸⁾. Many monitoring systems for acceleration power supply, stage temperature, beam size, beam drift, and so on have been developed. Generally, an EB system is very complicated, which poses a problem to be solved in the future.

Data conversion from LSI/CAD data to EB system data is essential for large-volume LSI data. VSB systems are used in parallel to conventional Gaussian beam systems in most mask shops. Data conversion from conventional Gaussian beam system data to the VSB system data is therefore necessary. An example of a data conversion system, in which the EX-8 VSB-type system and the EBM-130/40 Gaussian beam-type system are used, is shown in Fig. 4.3.7. LSI/CAD data are converted to EBM-format data for the EBM- 130/40. It takes about two days for 16 Mbit DRAM- class pattern data to be converted from LSI/CAD' to EBM format, because data compaction using a hierarchical structure of the LSI data cannot be applied effectively to EBM data. EBM data are converted to VSB data (EX-8 data) in a very short time. The reticle-making speed is not so high, because it is limited by conversion from LSI/CAD data to EBM data in spite of the high writing speed of the EX- 8 and high-speed data conversion from EBM data to VSB data. A high-speed data conversion system from LSI/CAD data to VSB data has therefore been developed using a hierarchical structure of the LSI pattern data and a parallel computer processing method⁽⁹⁾. The data conversion time is ~ 30 min for a 64 Mbit DRAM-class pattern.

References

1. Herriot, D.R., Collier, R.J., Alles, D.S., and Stafford, J.W. (1975). IEEE Transactions on Electron Devices, 3. ED-22, 385.
2. Takigawa, T., Shimazaki, K., and Kusui, N. (1986). 4. SPIE Proceedings, 632, 175.
3. Owen, G. and Rissman, P. (1983). Journal of Applied Physics, 54, 3573.
4. Gesley, M. (1991). Journal of Vacuum Science and Technology, B9, 1877.
5. Alles, D.S., Biddick, C.J., Bruning, J.H., Clemens, J.T., Collier, R.G., Gere, E.A., et al. (1987). Journal of Vacuum Science and Technology, B5, 47.
6. Takigawa, T., Ogawa, Y., Yoshikawa, R., Koyama, K., Tamamushi, S., Ikenaga, O., et al. (1987). Journal of Vacuum Science and Technology, B8, 1877.

7. Levenson, M.D., Viswanathan, N.S., and Simpson, R.A. (1982). IEE Transactions on Electron Devices, ED-29, 1828.
8. Anze, H., Tamamushi, S., Nishimura, E., Ogawa, Y., and Takigawa, T. (1992). In Digest of papers, MicroProcess 92, 5th International MicroProcess Conference, p. 132.
9. Magoshi, S., Koyama, K., Ikenaga, O., Watanabe, S., Saito, T., Ooki, S., and Sakamoto, S. (1992). In Digest of papers, MicroProcess '92, 5th International MicroProcess Conference, p. 128.
10. Pfeiffer, H.C. and Langner, G.O. (1978). In Extended abstracts, 8th International symposium on Electron and Ion Beam Science and Technology, p. 893. Electrochemical Society, Princeton, NJ.
11. Yoshikawa, R., Wada, H., Goto, M., Kusakabe, H., Ikenaga, O., Tamamushi, S., et al. (1987). Journal of Vacuum Science and Technology, B5, 70.
12. Takigawa, T., Wada, H., Ogawa, Y., Yoshikawa, R., Mori, I., and Abe, T. (1991). Journal of Vacuum Science and Technology, B9, 2981.
13. Sohda, Y., Nakayama, Y., Saito, N., Itoh, H., and Todokoro, H. (1991). Journal of Vacuum Science and Technology, B9, 2940.
14. Berger, S.D., Gibson, J.m., Camarda, R.M., Farrow, R.C., Huggins, H.A., Kraus, J.S., and Liddle, J.A. (1991). Journal of Vacuum Science and Technology, B9, 2996.
15. Abe, T., Yamasaki, S., Yoshikawa, R., and Takigawa, T. (1991). Japanese Journal of Applied Physics, B3, L528.

4.4 Machining of soft metal mirrors with diamond turning

High-precision machining using diamond cutting tools has been used in finishing of soft metals such as copper and aluminium, for which it is difficult to perform such machining by grinding, lapping and polishing. This diamond turning technique is finding wide application in the finishing of parts used in precision machines and optics such as information equipment, laser machining equipment, and space equipment. Ultra-precision machines have been developed for mirror-like surface finishing of precision parts and optics by the use of these diamond tools.

i. Surface roughness evaluation and surface damage observation

Mirror-finishing of soft metals by grinding and lapping is considered difficult because these processes have the shortcoming that scratches occur in the finished surface or undesired flatness values remain at the end of the machined surface. Diamond turning using an ultra-precision machine can eliminate these defects, providing a satisfactory mirror-finished surface⁽¹⁻³⁾.

Two kinds of diamond tool with circular and straight cutting edges have been used in cutting experiments. Surface roughness can be calculated geometrically based on the cutting edge geometry and the tool feed rate per spindle revolution. Geometrically calculated surface roughness R_{\max} values are given by eqns. (1) and (2):

circular arc cutting edge

$$R_{\max} = f^2 / 8r \quad (4.4.1)$$

straight cutting edge

$$R_{\max} = f / [\cot \delta + \cot(\pi - \varepsilon - \delta)] \quad (4.4.2)$$

Where f is the feed rate (mm rev^{-1}), r is the nose radius (mm), ε is the nose angle (rad), and δ is the working- end cutting-edge angle, the angle between finished surface and end cutting edge (rad). In the cutting operation, the surface roughness becomes small when the nose radius r is large and the working-end cutting edge angle δ is small.

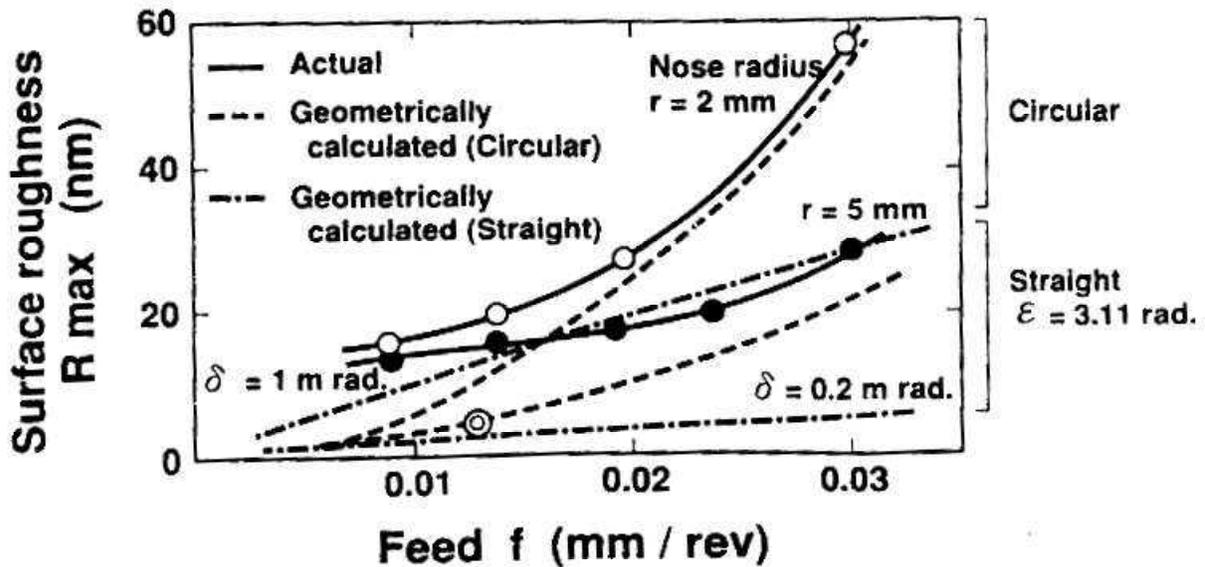


Fig. 4.4.1. Relation between surface roughness and feed.

Figure 4.4.1 shows a comparison between geometrically calculated surface roughness and actual surface roughness obtained by using a diamond circular tool in machining aluminium at different feed rates. The depth of cut was chosen as $2 \mu\text{m}$, and nose radii for the single-point diamond tool were chosen as 2 mm and 5 mm. The ultra-precision machine used in these cutting experiments was designed for practical fly cutting operation. The tool holder can be dynamically balanced, and the workpiece is mounted in a vacuum

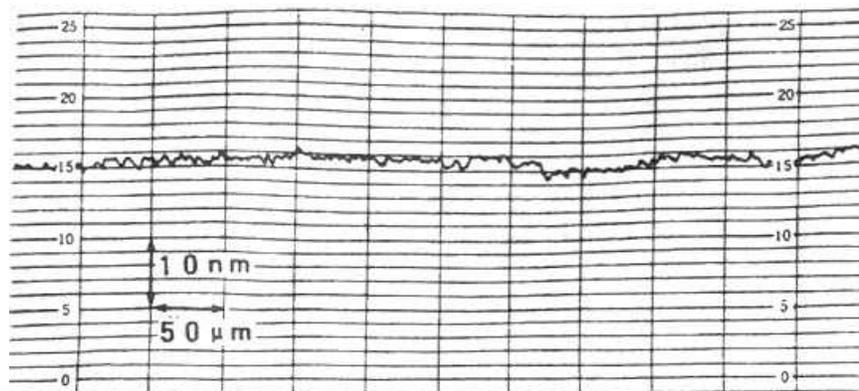


Fig. 4.4.2. Surface roughness on machined surface of OFHC copper (straight diamond tool).

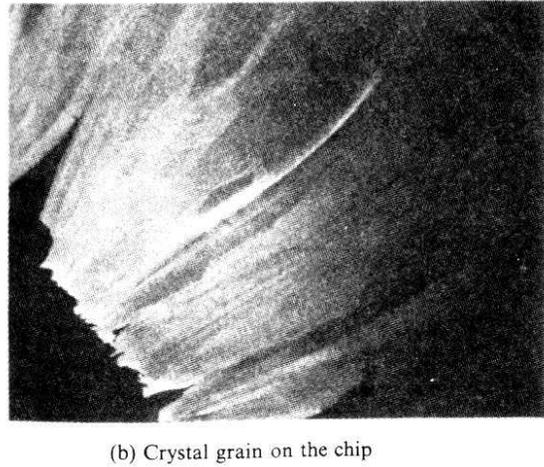
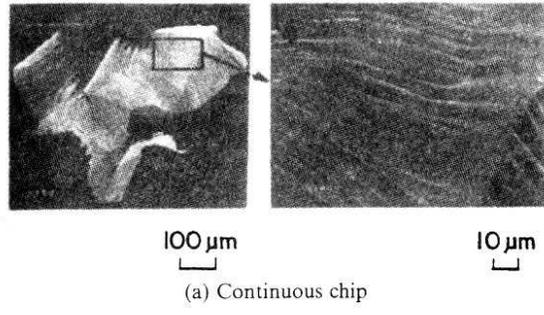


Fig. 4.4.3. SEM photographs of chip formation from OFHC copper machined with a diamond tool, (a) Continuous chip, (b) Crystal grain on the chip.

chuck on an *X-Y* table. The spindle is an externally pressurized air bearing spindle constructed with spherical bearings⁽¹⁻³⁾. The spindle was rotated at 1500 rev min⁻¹ (cutting speed 350 m min⁻¹). Surface

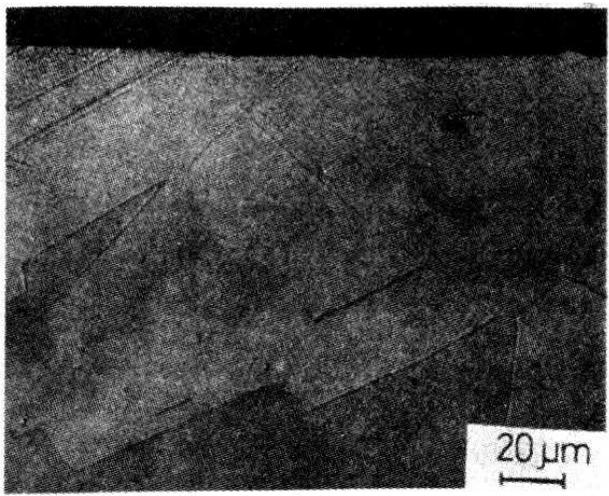
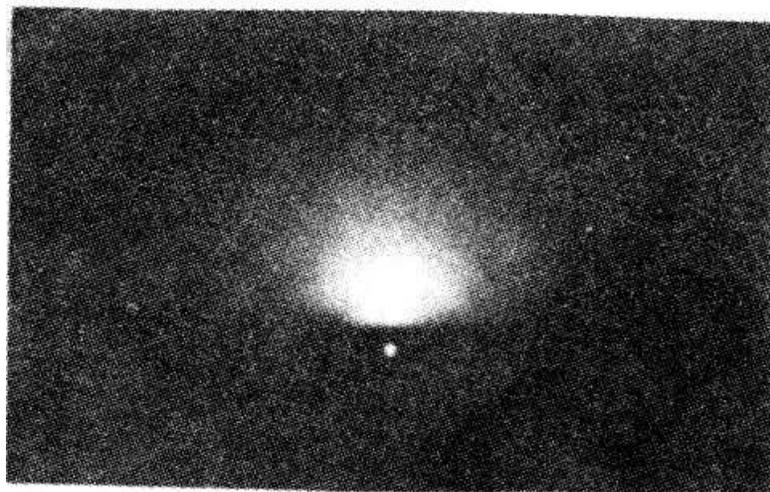


Fig. 4.4.4. Optical photomicrograph of machined surface of OFHC copper.

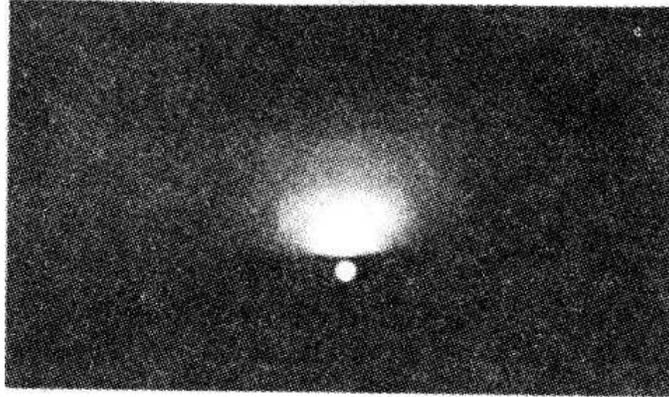
roughness was measured by a thin-film thickness tester (Talystep). The experimental results for 2 mm nose radius practically coincided with the geometrically calculated surface roughness when the feed exceeded $0.019 \text{ mm rev}^{-1}$, i.e. surface roughness R_{max} exceeding 30 nm. However, when the geometrically calculated surface roughness was $< 30 \text{ nm}$, the values did not coincide with the actual surface roughness.

OFHC copper was machined with a finishing grade (surface roughness R_{max} 10-20 nm) almost equivalent to that obtained with aluminium by using a diamond tool with nose radii of 2 mm and 5 mm. The depth of cut and feed were $2 \mu\text{m}$ and $0.013 \text{ mm rev}^{-1}$ respectively.

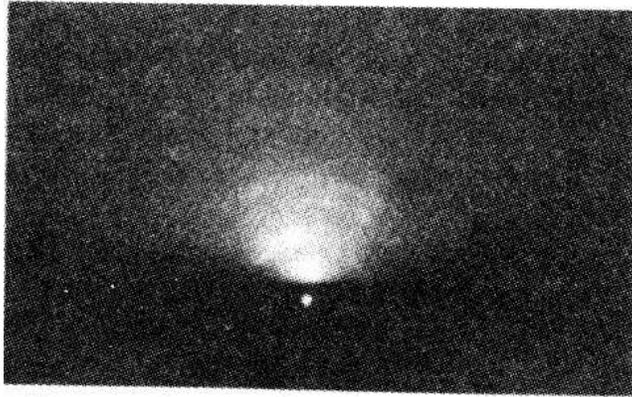
When a straight cutting edge diamond tool is used, the geometrically calculated surface roughness R_{max} value is given by eqn. (6.1.2), so it is strongly influenced by the working-end cutting edge angle δ . Therefore δ should be as small as possible in order to make the surface roughness small. Figure 6.1.2 shows examples of surface roughness measurements for OFHC copper obtained by using a straight-edge diamond tool. Here, ϵ was 3.11 rad and δ was adjusted to 0.2 mrad; the feed was $0.013 \text{ mm rev}^{-1}$, the cutting speed was 350 m min^{-1} , the depth of cut was 0.005 mm, and no cutting oil was used. The surface roughness R_{max} measured with a Talystep was 4 nm, indicated by the double circle in Fig. 4.4.1. The actual surface roughness coincides roughly with the geometrically calculated value. This is so far the best finished surface roughness achieved with a diamond tool. Smooth surfaces were obtained by using a highly accurate machine with an ultra-precision spindle and a sharp diamond tool with δ adjusted to a very small value.



(a) Machined surface



(b) Machined surface from which
 $0.2 \mu\text{m}$ metal was removed
by ion etching



(c) Machined surface from which
 $0.5 \mu\text{m}$ metal was removed
by ion etching

Fig. 4.4.5. Electron diffraction patterns of machined surface, (a) As machined, (b) After removal of $0.2 \mu\text{m}$ of metal, (c) After removal of $0.5 \mu\text{m}$ of metal.

Figure 6.1.3 shows an SEM photograph of chips produced when machining OFHC copper. This photograph suggests that the present mirror-like surface machining gives continuous chips by microscopic cutting instead of burnishing.

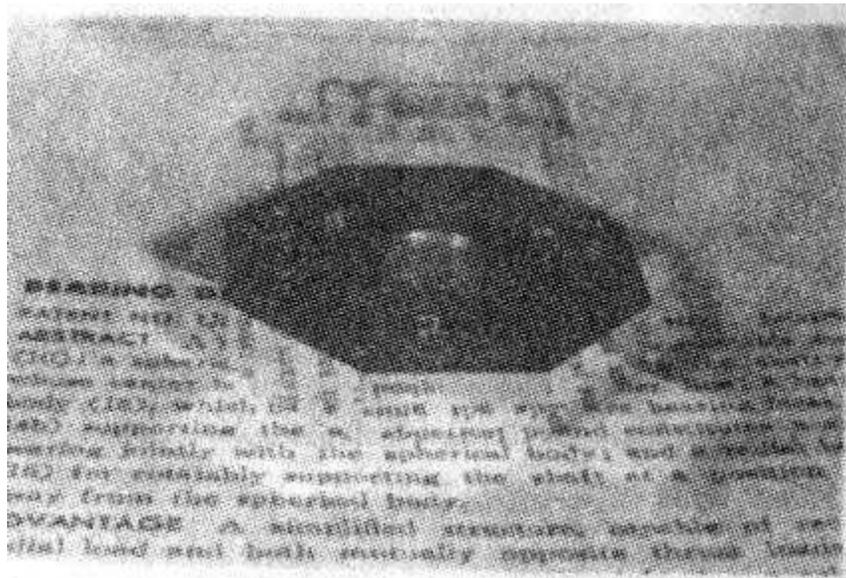
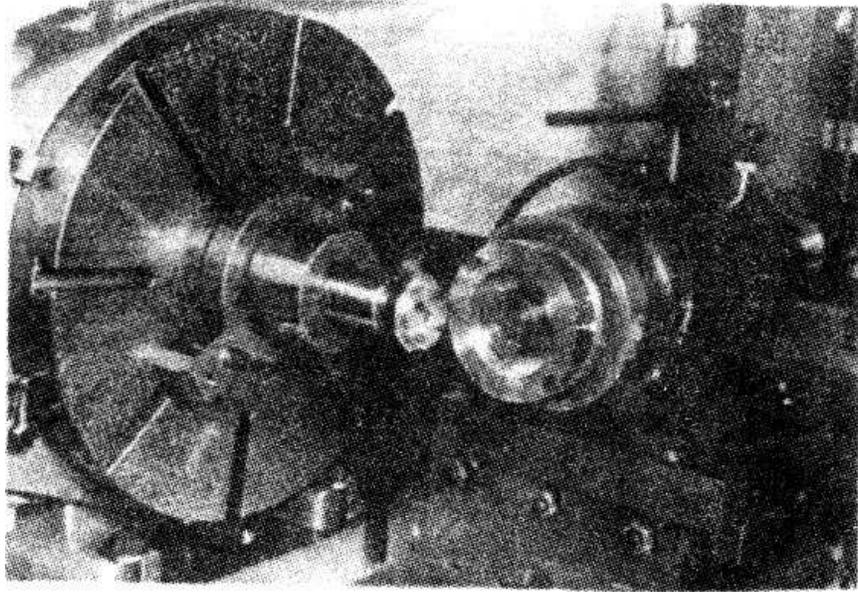


Fig. 4.4.6. Polygonal mirror machining arrangement (a) and machined mirror (b).

Generally a damaged layer remains on the machined surface of a metal. Figure 4.4.4 is an optical photomicrograph which shows a mirror-finished section of OFHC copper machined to 10-20 nm R_{\max} surface roughness under the conditions of 0.013 mm rev^{-1} feed and 5 μm depth of cut. From this photomicrograph it can be presumed that the crystal structure of the metal surface is very little disturbed and that only a slight surface damage layer is produced by diamond cutting. This surface was therefore observed by electron diffraction. Figure 6.1.5(a) is an electron diffraction pattern of a machined surface. In this pattern, the electron diffraction ring is indefinite; the

crystal structure of the surface appears distorted, showing the presence of strain. Figure 4.4.5(b) shows the pattern of the machined surface from which $0.2\ \mu\text{m}$ stock was removed by ion etching. In this figure the electron diffraction ring is clearly seen. Figure 4.4.5(c) shows the machined surface from which $0.5\ \mu\text{m}$ metal stock was removed in a similar manner. In this case the electron diffraction ring is much clearer. From this it can be deduced that the mirror-finished metal surface has a $0.2\ \mu\text{m}$ surface damage layer.

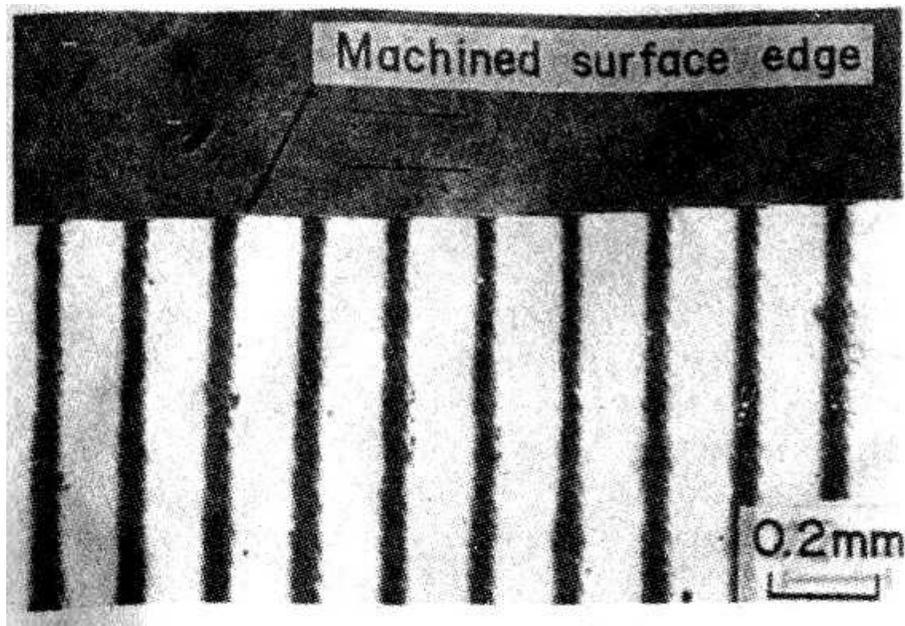


Fig. 4.4.7. Interference pattern of machined surface edge obtained with an interference microscope (wavelength 660 nm).

4.5 Mirror grinding of ceramics

Grinding, as well as cutting, is a typical machining method to meet the requirements of surface accuracy and quality. If the depth of tool action against the work surface is gradually brought closer to the limit in such machining, and it finally becomes possible to achieve removal of nanometre-size material, approaching the depth of a single-atom layer, surface accuracy and quality will be further enhanced. Also, when such machining conditions are applied to hard and brittle materials, a significantly mirror-like surface can be obtained, without the microcracks produced in conventional polishing.

Let us assume that constituent atoms are individually removed as the smallest chips from the work surface. Whether chip formation is effected mechanically, physically or chemically in such a case, the energy consumed should not differ much if the work consists of the same material. Therefore it may be said that there is no difference in the material removal mechanism between the grinding and cutting methods in nanometre-order chip formation. In practical nano-cutting, a single point diamond tool has been used, while an abrasive wheel on which there is an irregularly shaped edge of randomly distributed grains has to be applied in nano-grinding. Although there seems to be a great difference between the two methods, studies on micro- or nano-behaviour in the cutting edge region of single-point tool cutting have provided strong evidence as to the mechanism of mirror grinding.

4.5.1 Nano-grinding requirements

To ensure a mirror-like surface on certain kinds of materials, polishing was the primary machining method at one time. The mirror cutting method for soft and non-ferrous metals, such as aluminium alloy or oxygen-free copper, was developed in response to the requirement for fabricating devices in the electronics and optics fields. The material removal mechanism in nano-cutting to obtain mirror-like surfaces was clarified, and this method has subsequently played an important role in the fabrication of various high-tech devices.

Nano-cutting can be used for mirror finishing not only of such soft metals but also of polymers such as PMMA, amorphous metals such as electroless Ni-plated surfaces, single crystals and polycrystals of KDP, Ge, and Si, or amorphous materials including optical glass. Nano-cutting methods have made such remarkable progress, that the surface roughness R_{\max} of hard and brittle materials has been reduced to several nanometres. The realization of such nano-cutting has been supported by the use of ultraprecise and highly rigid cutting equipment, sharpened single point

diamond tools, and a variety of peripheral technologies. This implies some points of similarity in realizing nano-grinding.

The surface roughness of worked faces is conceptually formed as the profile resulting from the production of chips, and a sharpened cutting edge is necessary for obtaining mirror surfaces. Diamond single crystal as a tool material is very hard, the cutting face and flank are polished smoothly, and the cutting edge formed by their intersection can be extremely sharp.

In the indentation hardness test on a specular glass surface, when an indenter with minimal tip radius is gradually forced into the surface, the behaviour of a minute area on the glass comprises elastic deformation at first, then plastic deformation, and finally micro cracking, as is well known⁽¹⁾. Accordingly, when a hard and brittle material is to be cut to a mirror-like surface as in cutting aluminium alloy, the desirable working condition is plastic behaviour over a minute area, for instance 1/10-1/100 of the actual depth of cut for metals, to avoid crack generation⁽²⁾.

Figure 4.5.1 shows a model of the turning or fly cutting of a hard and brittle material⁽³⁾. In the area where the actual depth of cut becomes considerable, the removal of material in the brittle mode is unavoidably accompanied by some cracking. In the limited area of very small cutting depth, however, the plastic behaviour in the ductile mode occurs. It is important in nano-cutting or the nano-grinding to avoid cracks completely, i.e. to adopt a working condition that eliminates the brittle mode.

When a tool is used with a negative rake angle, e.g. -10° to -30° , for a hard and brittle material, notable cracks are found to develop in front of the tool edge. Taking account of these results in nano-cutting, nano-grinding has developed as a more practical technology.

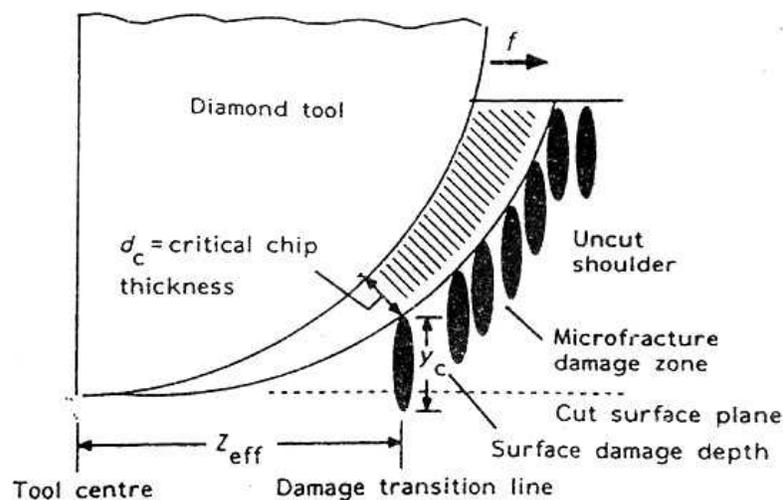


Fig. 4.5.1. Machining geometry for cutting model: cross-section along the cutting direction, with effective cutting depth increasing with distance from centre of tool. Ductile response is obtained when the depth of damage, d_c , initiated at cutting depth y_c does not extend below the plane of the cut surface.

The need for nano-grinding becomes particularly pressing in fabrication of devices from a hard and brittle material. In cutting a hard material, wear and chipping of the artificially sharpened cutting edge are unavoidable problems in practice. When nano-cutting is used for the fabrication of a large area of several optical devices, a fatal problem such as having to change the cutting tool halfway through working can occur. When applying nano-cutting to the fabrication of hard and brittle material devices, the continuous use of single point tool has been found to be difficult. A grinding wheel, on the other hand, has innumerable tips of fixed abrasives and can maintain an equilibrium state of budding, flattening and dislodgement of the grains for a fairly long time. Since it will be possible to obtain a satisfactory mirror-like surface under such a condition, nano-grinding can be expected to be a valuable mirror finishing method for hard and brittle materials.

4.5.2 Nano-grinding technology

(a) Grinding equipment

In the grinding process, as the grinding wheel of a facial tool acts against the work surface, the force applied may reach several hundred times that obtained with ordinary single-point diamond tools. Accordingly, grinding equipment and wheels of high rigidity and high accuracy are necessary⁽²⁾.

The development of improvement of nano-grinding equipment may be regarded as a continuation of that of nano-cutting machines. Accuracy of movement of the machine tool is certainly important, because the shape and track of the cutting edge are transferred to the whole work surface, so that it becomes possible to secure precise flat or spherical surfaces.

In cutting or grinding systems, various elements such as air bearings, hydrostatic pressure bearings, piezoelectric actuators, and friction drives have been investigated by various organizations to achieve higher precision and rigidity of movable units. Concerning the materials of construction of the machines, consideration has been given to thermal expansion and external

or internal vibration; super-invar, low- thermal-expansion glass, and granite have therefore been used. Furthermore, various improvements including a new cooling system, and ultra-precision laser measurement system, and a high speed response control system have been adopted to upgrade grinding technology.

In order to develop the next generation of ultra precision machines, cutting and grinding systems have been thoroughly reviewed, and several new machine forms have been proposed. For instance, the tetraform structure, which is in the form of a tetrahedron and consists of six supported arms, has been reported as well-balanced and subject to minimal distortion under forces and heat generated during machining, taking account of the fact that there can be no completely rigid structural material for machines^(4,5). In the ductile mode of grinding, a quartz crystal has been finished to a mirror-like surface with a surface roughness R_a of 2.76 nm. In grinding instead of conventional lapping of silicon wafers, ultra-precision grinding equipment with a loop stiffness between grinding wheel and work table of 150 N/ μm was developed and a flatness of TTV < 0.6 μm was obtained on 150 mm wafers⁽⁶⁾.

(b) Grinding wheel and grinding characteristics

In general, a grinding wheel is composed of three elements, that is, abrasive grains, bonding materials and pores. In normal metal grinding, grinding wheels consisting of these three elements in a suitable ratio to facilitate chip generation and elimination have been preferred. When the self-dressing phenomenon occurs over a long period of grinding with a vitrified-bond wheel of alumina grains, normal grinding is said to be in progressing smoothly.

In the self-dressing phenomenon, it is presumed, a grain tip becomes flat as the grinding wheel surface wears, part of the grain breaks and is dislodged by the grinding force, and a new tip is produced. This the grinding wheel radius, which is unnecessary in nano-grinding. However, it may not be realistic to expect such self-dressing for hard and tough diamond grains. In addition, the undesirable dislodgement of fine grains will be rather more important than the self dressing phenomenon, when using a specified diamond wheel of fine abrasive grains for the production of micro- or nano-chips.

In the early fabrication of Mn-Zn ferrite magnetic heads, a mirror grinding condition was required to obtain a mirror-like surface and a chipping-free track width. Grinding machines and wheels were selected by considering their accuracy and rigidity, because the track width of the

magnetic disks was of the order of $10^0 - 10^2 \mu\text{m}$, which was the same size as the microcrystals of the polycrystalline ferrite. As one countermeasure, heat-resistant resin-bonded diamond wheels containing antifriction materials such as MoS_2 , WS_2 , and graphite were developed, with fairly good results.

With a resin-bonded wheel it is possible to attain a higher quality of ground surface than with a metal-bonded wheel. However, since the bond material is soft, abrasion of the tool cannot be disregarded. Moreover, deformation of the face of the grinding wheel due to the grinding force impairs precision grinding. For greater rigidity, vitrified-bond diamond wheels have advantages. Since these grinding wheels have many pores, which facilitate elimination of the chips, they are considered suitable for high-precision grinding.

Many diamond wheels are bonded with metal, for example bronze-sintered wheels or nickel-plated wheels. It has attracted special interest recently that ferrous-bonded diamond wheels, first developed for the advanced ceramics, can be applied to the grinding of various materials. These grinding wheels are sintered with cast iron fibres or powders, high-purity iron powders by the carbonyl method, and diamond grains. Compared with conventional bronze-bonded wheels, iron-fibre-bonded wheels have higher rigidity and stronger grain-holding power. Moreover, much stronger grinding wheels with steel powders are obtaining excellent results. For such grinding wheels, a new electrolytic dressing method is proposed to restore the worn wheel surface to the previous standard. An aqueous grinding liquid acts as an electrolytic solution and direct current is supplied from the wheel of the anode to the negative pole, as illustrated in Figure 6.2.2. This can be effected simultaneously with the grinding process, and the whole is called the Elid (electrolytic in-process dressing) grinding method^(7,8). The bond material on Coolant and dressing fluid are the same and form the grinding fluid. The negative electrode is usually copper or graphite. Insulation between machine and wheel can usually be eliminated if the machine frame and positive pole for Elid are earthed. the wheel surface is dissolved or converted into non-conductive films liable to wear without affecting the bonded diamond grains mechanically, so the diamond grains continue to protrude beyond the bond material. A modification of the electrolytic dressing method, the twin-ECD (electrochemical dressing) method with alternating current has been proposed⁽⁹⁾.

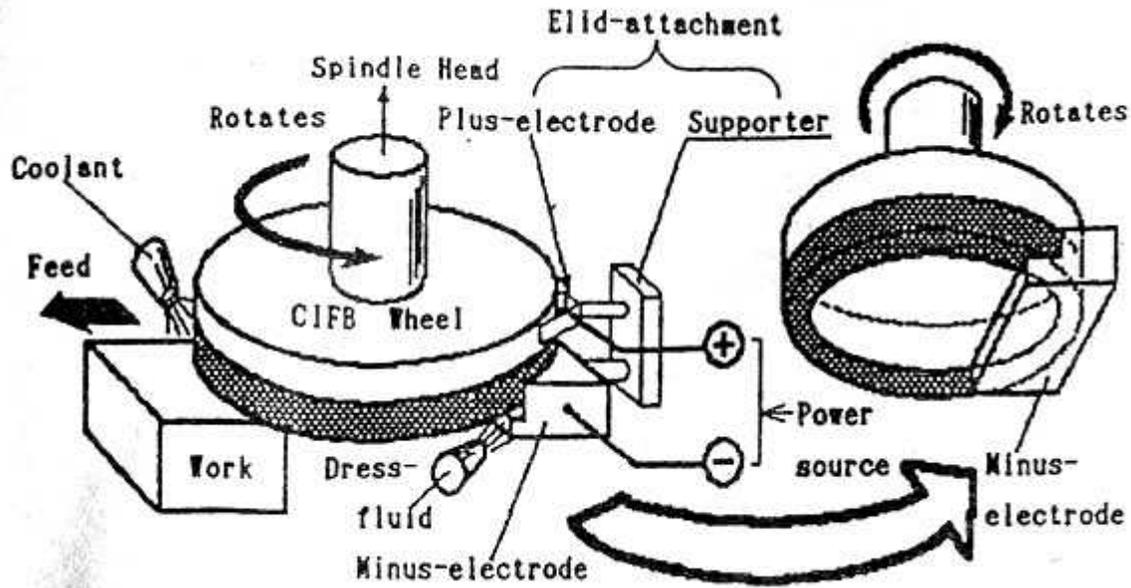


Fig. 4.5.2. Schematic diagram of grinding with electrolytic in-process dressing (Elid).

In general, a metal-bonded wheel with small diamond grains, $\leq 5 \mu\text{m}$ in size is not practical because mechanical dressing is usually applied without sufficient consideration and leads to the dislodgement of the diamond grains. In the Elid grinding method, however, effective dressing is possible even if the diamond grains on the wheel surface are $\leq 1 \mu\text{m}$ in size. During electrolysis, the bond material on the wheel surface does not simply dissolve into the grinding solution but is converted to oxidized or hydroxide films. Therefore, in the application of the Elid method to nano-grinding, metal-bonded wheels consisting of rather smaller grains are used. The work-material-removing action, however, should be interpreted to be done by the non-metal bonded wheel.

The Elid grinding method can be carried out on rotary grinding and reciprocal grinding machines of the vertical or horizontal type and can be applied to nano-grinding of crystalline materials, such as silicon wafers and GaAs wafers, in addition to various kinds of advanced ceramic elements and glass lenses, and enables these materials to be finished to mirror quality with a surface roughness R_{max} of several nanometres.

A comparison between conventional mechanical dressing and the Elid method has been carried out in the grinding of BK7 glass. As shown in Figure 4.5.3,

when using a mechanically dressed grinding wheel, in spite of a rather smaller grinding force, the ground surface is covered with microcracks and has a surface roughness of $R_a \leq 460$ nm. On the other hand, in the Elid grinding method the grinding force is ten times as high, a surface roughness $R_a \leq 11$ nm is obtained, and microcracks are never observed⁽¹⁰⁾. Although the Elid grinding method should satisfy the excellent nano-grinding conditions, it is important to watch out for the tendency of the grinding force to rise and to take appropriate action.

In nano-grinding, some grinding wheels made of cerium oxide powders or silicon oxide powders have been successful in special cases, in addition to diamond grinding wheels. As grinding wheels of fine abrasive powders can give smaller surface roughness, a method of manufacturing excellent grinding wheels by the dispersion of abrasive powders has been devised, utilizing electrophoretic deposition principles. Colloidal silica (10-20 nm) is mixed with sodium alginate solution, and the abrasive powder is collected at the positive electrode by means of electrophoretic deposition, then dried, and pelletized into grind stones. By application of a wheel of such stones to silicon wafer grinding, a mirror-like surface with a surface roughness $R_{max} \leq 10$ nm and without grain tracks is obtained with 2 μm depth of cut and 800 m min^{-1} wheel speed⁽¹¹⁾. The use of poly (vinyl alcohol) as a bond material instead of sodium alginate for fine diamond powders has been proposed.

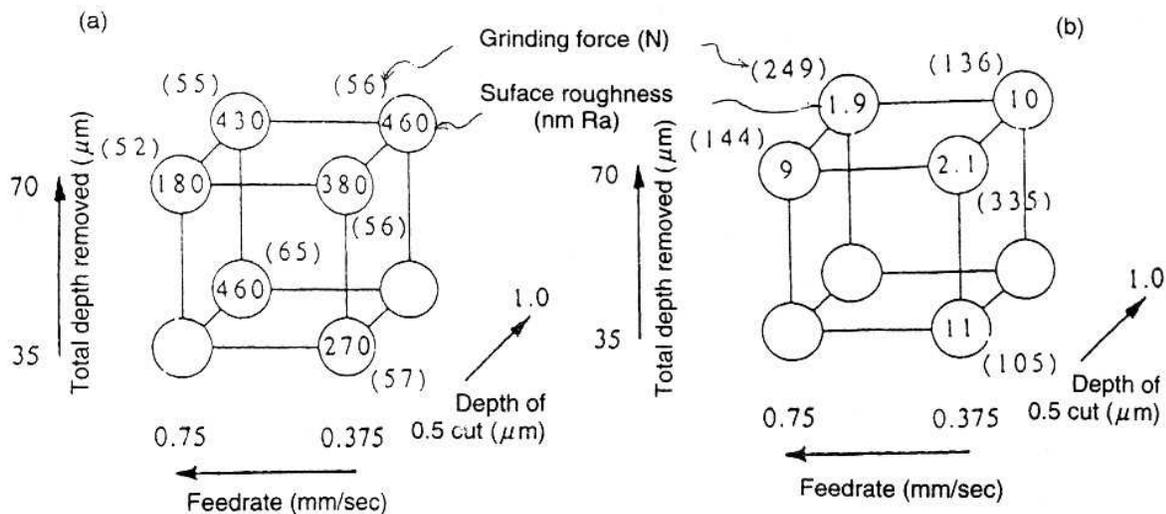


Fig. 4.5.3. Grinding characteristics in (a) mechanical and (b) electrolytically assisted diamond grinding of BK7. Figures in circles are surface finish R_a (nm); figures in parentheses are normal grinding force (N) during final cuts.

References

1. Taniguchi, N. Chapter 1 of this book.
2. Miyashita, M. (1988). Review of ultraprecision grinding technology. In Proceedings of the International Congress for Ultraprecision Technology, Aachen, 41-57.
3. Blackley, W.S. and Scattergood, R.O. (1991). Ductile- regime machining model for diamond turning of brittle materials. ,13, 95-103.
4. Lindsey, K., Smith, S.T., and Robbi, C.J. (1988). Sub-nanometre surface texture and profile measurement with NANOSURF 2. *Annals of the CIRP*, 37, 519-22.
5. Anon. (1991). Design news: Tetraform system revolu-tionizes machine design. *Precision Engineering*, 13, (1), 95-61.
6. Abe, K., Yasunaga, N., Miyashita, M., Yoshioka, J., and Daito, D. (1993). Development of ultra precision grinding equipment for ductile mode surface finishing of brittle materials. In Proceedings of the 7th International Precision Engineering Seminar, 153-64.
7. Suzuki, K., Uematsu, T., Yanase, T., Honma, M., and Asano, A. (1991). Development of a simplified electro-chemical dressing method with twin electrodes. *Annals of the CIRP*, 40, 363-6.
8. Ball, J.M., Murphy, N.A., and Shore, P. (1992). 'Ductile' mode diamond grinding of optical glasses using electrolytic techniques. In XVI International Congress of Glass, Vol. 6, 259-64.
9. Ikeno, J. and Tani, Y. (1990). Nanometer grinding using ultrafine abrasive pellets — Manufacture of pellets applying electrophoretic deposition. *Annals of the CIRP*, 39, 329-32.

4.6 Ultra-precision block gauges

4.6.1 Introduction

Block gauges have several features: they possess a high accuracy as end standards; the dimensions remain stable over long periods; any arbitrary gauge dimension can be built up by wringing together several gauges; various applications are possible by combining them with auxiliary parts; they are easy to use; etc. Currently they are widely used in the machine industry as length standards.

In manufacturing plants, various length-measuring instruments are used to inspect the dimensions of fabricated parts and products. To guarantee the accuracy of such length-measuring instruments, they must periodically be compared with block gauges and calibrated. In turn, the block gauges used for inspection must be checked against a set of higher- grade calibration block gauges, and the highest (00)- grade block gauges must be calibrated by absolute measurement with an optical interferometer. The optical interferometer is a length standard that uses the wavelength of a standard light (see Section 6.4.3 below). The measured results using the standard light are traceable to national as well as international length standards.

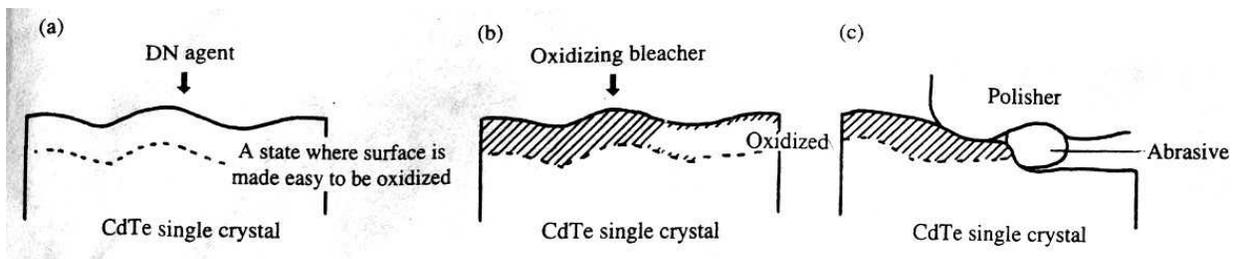


Fig. 4.6.1. Processing mechanism for chemical polishing of CdTe single crystal, (a) Effect of DN agent, (b) Effect of oxidizing bleach, (c) Effect of mechanical work by abrasive.

To meet such specifications and applications, the dimensions, flatness, parallelism and surface roughness of the end surfaces of block gauges must be finished to the highest accuracy. Since their invention, improvements have been made in the material characteristics, heat treatment methods, and methods of measurement. We briefly discuss below the accuracy, methods of measurement and fabrication block gauges.

4.6.2 Accuracy of block gauges

The length of a block gauge is defined in ISO 3650⁽³⁾ as follows:

The length of a block gauge at a particular point of the measuring face is the perpendicular distance between this point and a rigid plane surface of the same material and surface texture upon which the other measuring face has been wrung.

Dimensional measurements of block gauges are thus based on this definition. Since the advantage of block gauges lies in being able to create any arbitrary dimension by wringing together individual blocks, it is important to improve the 'wring' strength and minimize dimensional errors due to the combination.

Flatness, parallelism, and surface roughness are thus related to the accuracy of combined block gauges as well as to the final dimensional accuracy. The dimensional tolerances and permissible variations in parallelism are shown in Table 4.6.1 and Fig. 4.6.2 (Detailed specifications are omitted here.)

4.6.3 Measurement of block gauges

If the block gauge is to be used in industry as a length standard, it must be measured according to the definition of a metre given by national and international standards. In 1889, the metre was defined as the length of a prototype standard; since 1983, it has been established as the distance that light travels in vacuum during $1/299\,792\,458$ of a second⁽⁴⁾. The dimension of a block gauge, on the other hand, is defined in ISO 3650 by two methods: by interferometry and by comparison with a reference block gauge, as shown respectively in Figs 6.4.3 and 6.4.4. The former method is used to measure the dimensions of highly accurate grade-00 gauges and reference block gauges, ^{96}Kr , ^{198}Hg , and ^{114}Cd , recommended by the 17th CGPM (1983) as standard light sources to realize the definition of the metre, are also used as light sources process of for block-gauge interferometry.

4.6.4 Measurement by interferometry

There are two methods of measurement using interferometry: the coincidence method and the counting method. The former was developed early on and is still in wide use today. Although we do not discuss its principles here, Fig. 4.6.5⁽⁵⁾ shows the construction of a block gauge interferometer developed jointly by the National Research Laboratory of Metrology of Japan and Tsugami Corp.

In the coincidence method, the fraction of the interference fringes is measured to determine the block gauge's dimensional error. The measured result is then corrected by incorporating the following factors to obtain the 'true' dimensional error:

- (1) temperature measurement to correct for thermal expansion (thermometer resolution 0.01 K when the coefficient of thermal expansion of the gauge material is $11.5 \pm 1.0 \times 10^{-6} \text{ K}^{-1}$);
- (2) measurement of characteristic indices of ambient air to correct for variations in the index of refraction;
- (3) correction to account for the size of the pinhole or slit of the interferometer collimator;
- (4) correction to account for differences in the base plate material (e.g. glass, fused quartz).

4.6.5 Fabrication

Although the block gauge is very simple in structure, several processing steps are needed to satisfy such requirements as dimensional accuracy and long-term stability⁽⁶⁾. In addition, to the traditional materials

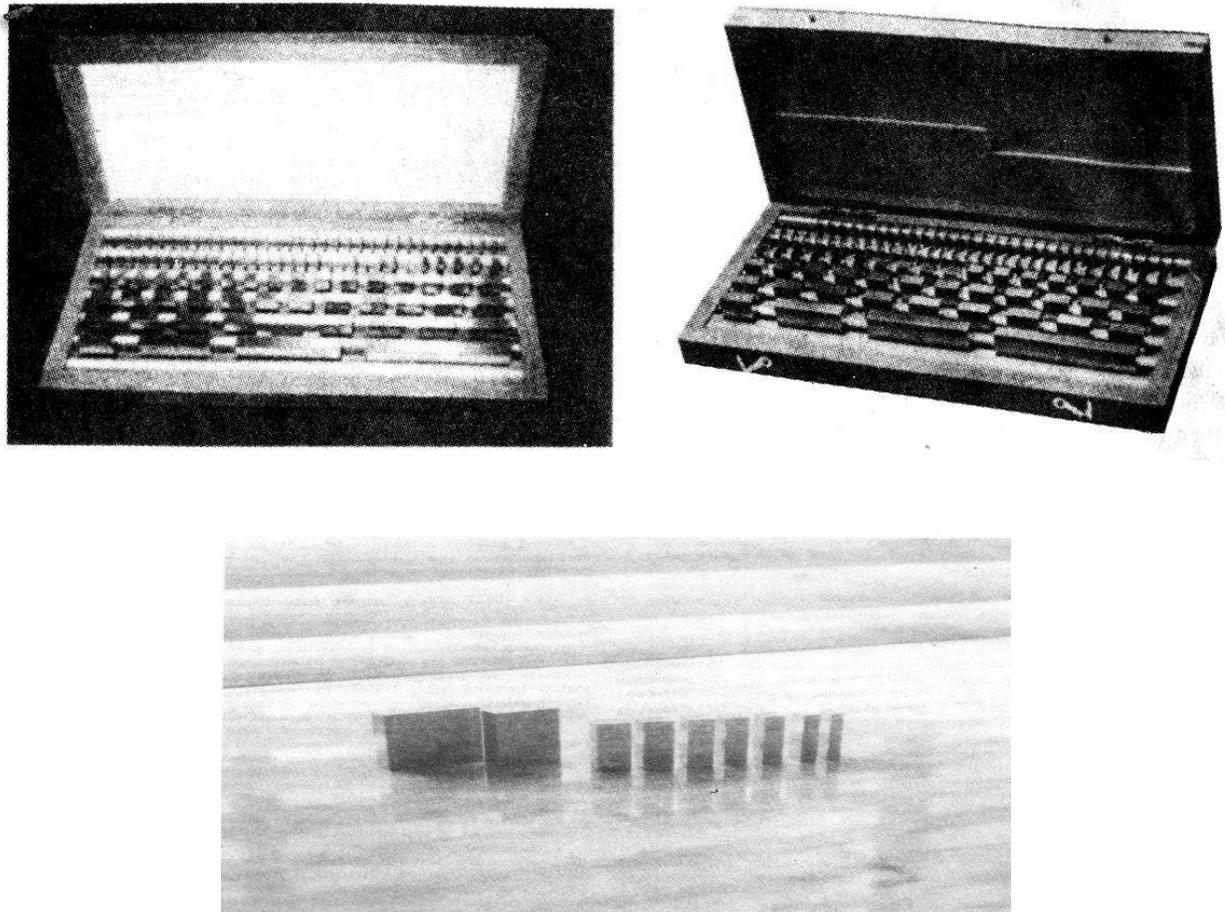


Fig. 4.6.2. Block gauges.

such as high-alloy steel and cemented carbide, new materials such as zirconia ceramics⁽⁷⁾ are being used today. Here we describe a manufacturing process for block gauges made from high-alloy steel (the lapping process will be described in the following section):

- (1) purchase of stock material: high alloy steel DC- 8 (see Table 6.4.2 for its composition);
- (2) milling;

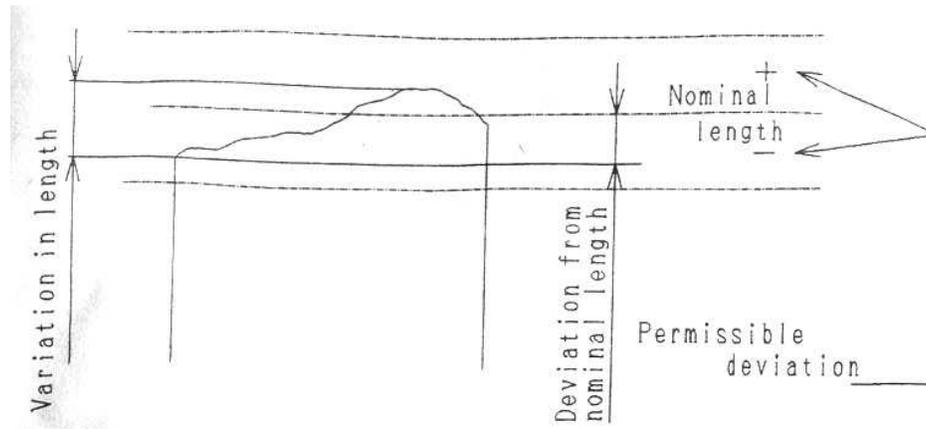


Fig. 4.6.3 Variations in length.

- (3) cutting to standard lengths;
- (4) heat treatment: after quenching, sub-zero treatment (-70 to -80°C) and tempering (100 to 120°C) are repeated alternately several times;
- (5) accelerated ageing: although ageing time depends on the particular heat treatment, it is usually 6-12 months;
- (6) rough grinding: to prevent machining strains, heavy grinding is avoided; the surfaces of opposing faces are alternately ground a little at a time;
- (7) finish grinding; ,
- (8) rough lapping — first stage;
- (9) rough lapping — second stage;
- (10) finish lapping — first stage;
- (11) finish lapping — second stage;
- (12) marking;
- (13) inspection: grading.

Table 4.6.1 Tolerances and permissible variations

Tolerances and permissible variations (/mi)									
RANGE OF NOMINAL LENGTHS (MM)		GRADE 00		GRADE 0		GRADE 1		GRADE 2	
OVER UP TO INCLUDING	AND	TOLERANCE ON NOMINAL LENGTH AT ANY POINT	PERMISSIBLE VARIATION IN LENGTH	TOLERANCE ON NOMINAL LENGTH AT ANY POINT	PERMISSIBLE VARIATION IN LENGTH	TOLERANCE ON NOMINAL LENGTH AT ANY POINT	PERMISSIBLE VARIATION IN LENGTH	TOLERANCE ON NOMINAL LENGTH AT ANY POINT	PERMISSIBLE VARIATION IN LENGTH
—	10	±0.06	0.05	±0.12	0.10	±0.20	0.16	±0.45	0.30
10	25	±0.07	0.05	±0.14	0.10	±0.30	0.16	±0.60	0.30
25	50	±0.10	0.06	±0.20	0.10	±0.40	0.18	±0.80	0.30
50	75	±0.12	0.06	±0.25	0.12	±0.50	0.18	±1.00	0.35
75	100	±0.14	0.07	±0.30	0.12	±0.60	0.20	±1.20	0.35
100	150	±0.20	0.08	±0.40	0.14	±0.80	0.20	±1.60	0.40
150	200	±0.25	0.09	±0.50	0.16	±1.00	0.25	±2.00	0.40
200	250	±0.30	0.10	±0.60	0.16	±1.20	0.25	±2.40	0.45
250	300	±0.35	0.10	±0.70	0.18	±1.40	0.25	±2.80	0.50
300	400	±0.45	0.12	±0.90	0.20	±1.80	0.30	±3.60	0.50
400	500	±0.50	0.14	±1.10	0.25	±2.20	0.35	±4.40	0.60
500	600	±0.60	0.16	±1.30	0.25	±2.60	0.40	±5.00	0.70
600	700	±0.70	0.18	±1.50	0.30	±3.00	0.45	±6.00	0.70
700	800	±0.80	0.20	±1.70	0.30	±3.40	0.50	±6.50	0.80
800	900	±0.90	0.20	±1.90	0.35	±3.80	0.50	±7.50	0.90
900	1000	±1.00	0.25	±2.00	0.40	±4.20	0.60	±8.00	1.00

4.6.6 Lapping

Lapping is a critical process that determines the dimensional accuracy of the block gauges. This entire process is carried out in constant-temperature rooms ($20 \pm 0.2^\circ\text{C}$), with separate rooms reserved for rough and finish lapping.

In the early days, the accuracy of block gauges was achieved by hand lapping. A set of eight gauges was

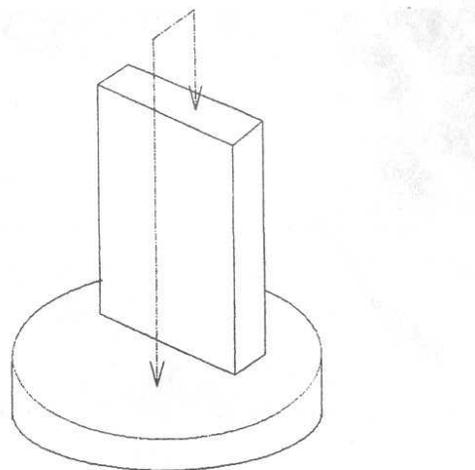


Fig. 4.6.3. Centre length measured by interferometry.

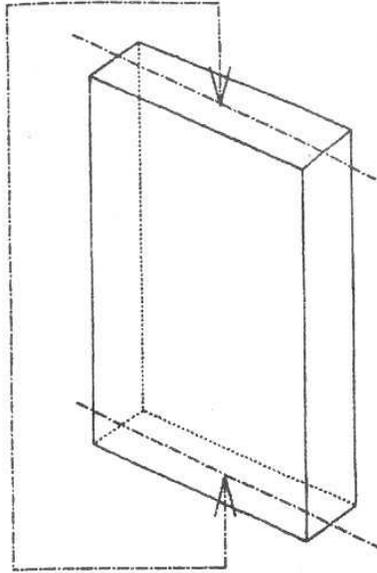


Fig. 4.6.4. Measurement of length by comparison with a block gauge as reference standard.

used, four attached on each side of opposing faces of the block gauges and one side finished at a time, as shown in Fig. 4.6.4⁽⁸⁾. It was a process requiring highly skilled craftsmanship. Today most block gauges are machine-lapped, and the lapping speed, pressure, and time are precisely controlled.

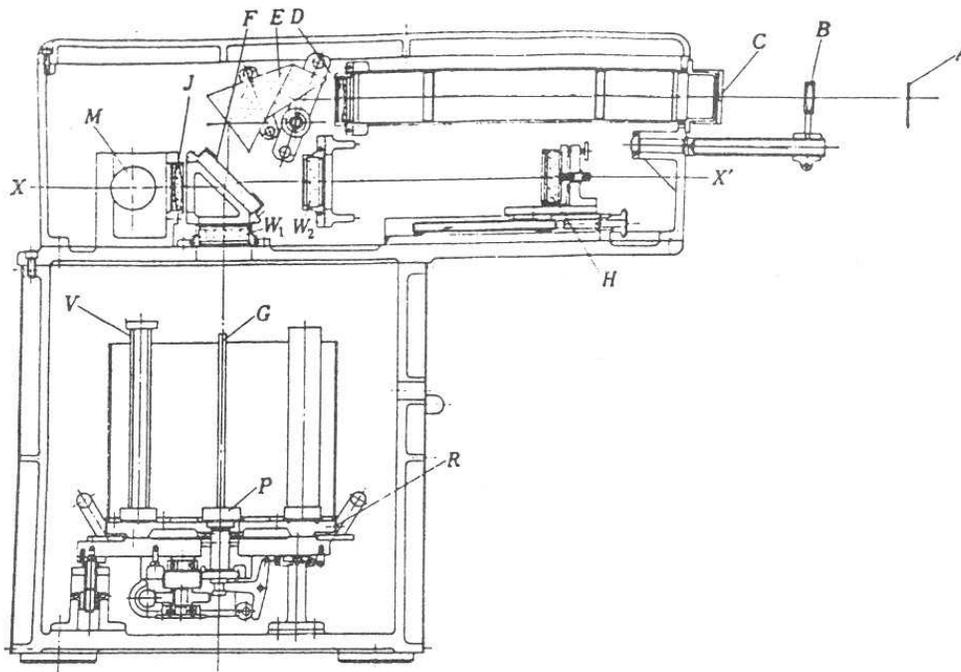
A rotary lapping machine has a set of two circular lap tables of the same diameter; the work is placed between the lap tables and opposing parallel surfaces are lapped simultaneously. The workpieces are laid out and their positions are switched from time to time, as shown in Fig. 4.6.7^(9,10), so that a particular piece will not follow the same path on the lap table. The lapping process is divided into four stages. Each stage uses progressively finer abrasives to remove the machining margin in stages and attain the final dimensions and form. The machining margin and abrasive grain size for each lapping stage are shown in Table 6.4.3. Chromium oxide, silicon carbide, aluminum oxide, and diamond are some materials used as lapping abrasives. According to one report⁽¹¹⁾, after measurements were made on steel block gauges made by six Japanese and ten overseas manufacturers, a surface roughness R_{\max} of 20-40 nm was found to be typical.

4.6.7 Traceability of block gauges

To maintain the level of measurement accuracy using end standards among public inspection agencies and block gauge manufacturers in Japan, since 1960 the National Research Laboratory

of Metrology (NRLM) has conducted round robin tests of 100 mm-length block gauges using optical interferometry every two years.

The relevant department of NRLM first measures three such gauges and sends each to three groups, namely (a) various sections within NRLM, (b) public inspection agencies, and (c) block gauge manufacturers and their users. Each organization then takes turns in measuring the blocks, after which they are returned to NRLM, which takes a second set of measurements. Any deviation from the first set of NRLM measurements is considered to be caused by secular change, and this is used to correct the block gauge dimensions at the time they were measured by the respective organizations. If an organization's measurements deviate by < 30 nm from the corrected value, its measurement capability is considered as satisfactory. If on the other hand the deviation exceeds 30 nm the organization is given a second chance to take another set of measurements. The results of three round robin tests in Japan up to 1986 are shown in Table 4.4.4⁽¹²⁾.



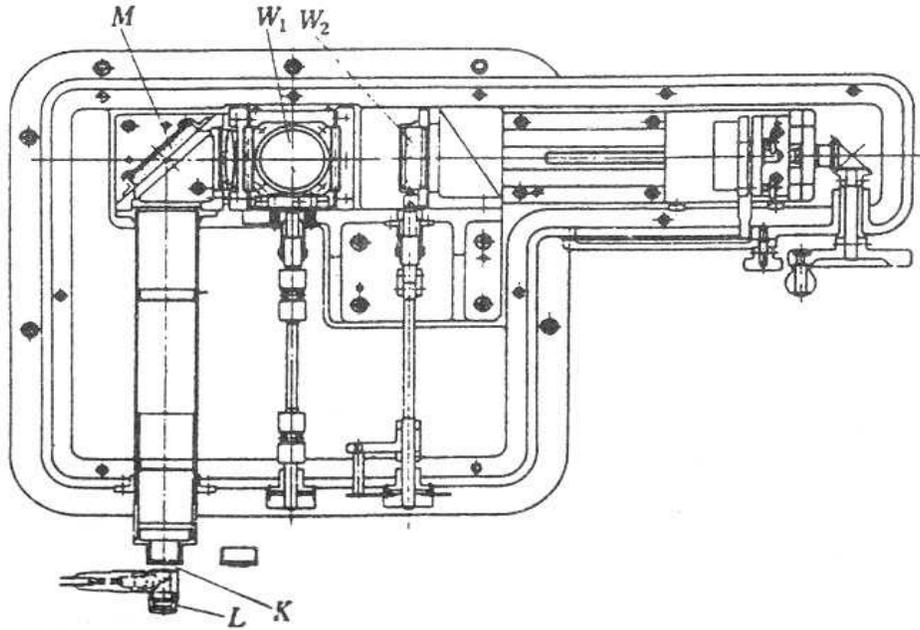
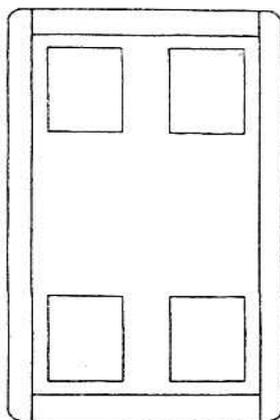


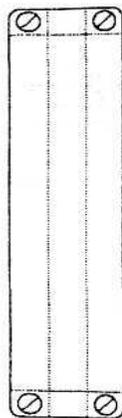
Fig. 4.6.5. Diagram of block gauge interferometer, with section X-X'. A, light source; B, condenser lens; C, entrance slit. D, collimator lens; E, prism; F, beam splitter; G, block gauge; H, reference surface; K, observation slit; M, reflecting mirror; P, base plate.

Table 4.6.2 Chemical composition of DC8 high-alloy steel (Daido Steel Co. Ltd.)

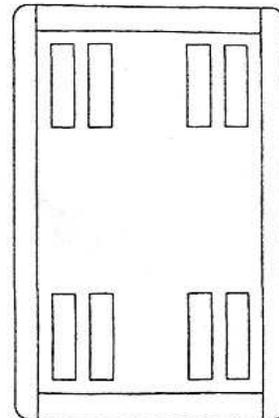
Elements	C	Si	Mn	P	S	Cu	Ni	Cr	W
Spec. (%)	2.00-2.20	≤ 0.40	≤0.50	≤0.025	≤0.02	≤0.25	≤0.10	12.00- 13.00	0.60- 1.00



Back View



Side View



Front View

Fig. 4.6.6. Hand lapping.

References

1. Tsugami, K. (1962). Block gauges. Nikkan Kogyo Shimbun.
2. Kuroda Precision Industries Ltd. (1991). Gauge catalogue no. 3.
3. International Organization for Standardization (1978). ISO 3650. Geneva.
4. Seino, S. (1971). Precision Machines, 50 (7).
5. Sakurai, Y. and Seino, S. (1965) Kikai no Kenkyu, 17, (3).
6. Kuroda, S., Yokoyama, K., and Matsukura, T. (1961). Seimitsu Kikai, 27, (10).
7. Mitsutoyo Co. (n.d.). Gauge block catalogue no. 4177.
8. Rolt, F. H. (1929). Gauges and fine measurements, Vol.I. Macmillan, London.
9. Gierlich, R. (1950). How gage blocks are made. American Mach...
10. Gierlich, R. (1953). Producing gage blocks to "millionths". Machinery, 59, (6).
11. Hoshina, N. and Mori, Y. (1971). Toshiba Review, 26, (6).
12. Seta, K., Matsumoto, K., and Seino, S. (1988). Bulletin of NRLM, 37, (4).

4.7 Ultra-precision balls for rolling bearings

4.7.1 Introduction

Rolling bearings are considered to possess the highest precision among mechanical structural parts. Rolling bearings are produced and used in the largest quantities worldwide, despite the many types of bearings available, including sliding bearings and magnetic bearings. However, this does not necessarily mean that rolling bearings are superior in performance to other types of bearings. For example, the runout accuracy of rolling bearings, when mounted on a spindle, is actually inferior to that of properly designed fluid bearings.

In terms of the accuracy of the component parts, however, rolling bearings are superior to fluid bearings. In a fluid bearing the averaging effect of a fluid film does not allow the geometrical inaccuracies of its component parts to exert a direct influence upon the accuracy of the spindle. In contrast, the accuracy of the component parts of a rolling bearing directly influences the overall accuracy of the bearing. Rolling elements such as balls and rollers, in particular, can greatly affect the overall bearing accuracy, and therefore must be of a very high grade of accuracy.

This section discusses ultra-precision balls for ball bearings and includes a description of the application background that requires such ultra-precision. It also describes the manufacturing processes and the principle followed in the processes used to produce ultra-precision balls.

Table 4.7.1 Machining margin and abrasive grain size in lapping (figures in μm)

Stage	Hand lapping	Machine lapping	Grain diameter
Finish grinding	20	15-20	-
1st stage rough lapping	5	1.1-2.0	4-8
2nd stage rough lapping	3	0.4-0.6	2-4
1st stage finish lapping	1	0-0.2	0-1
2nd stage finish lapping	Nominal dimension	Nominal dimension	0-0.5

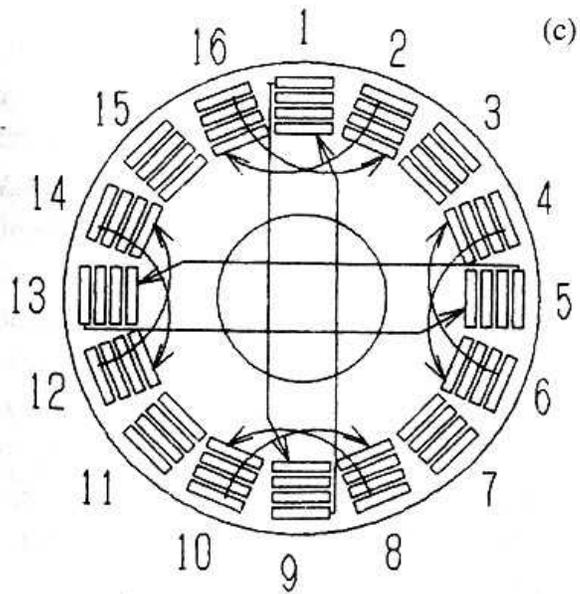
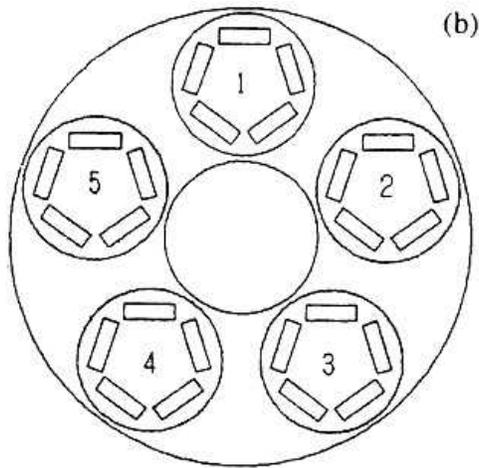
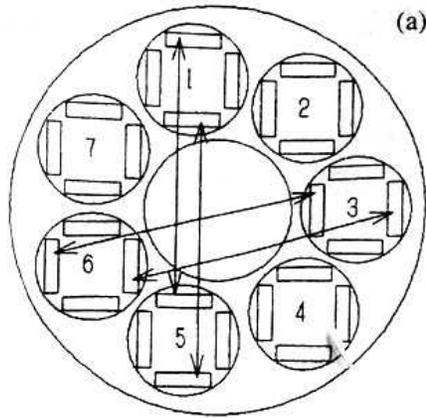


Fig. 4.7.1. Machine lapping, (a) Workpiece layout and position switch (28 pieces), 1st-stage rough lapping, multiple carriers, (b) Workpiece layout (25 pieces), 1st- stage rough lapping, multiple carriers, (c) Workpiece layout and position switch (16 pieces), 2nd-stage rough lapping and after, single carrier.

4.8 CCDs (charge-coupled devices)

The CCD is currently used in two applications. One is the linear sensor in facsimile equipment. The other is the area sensor for camera fields. Home video cameras are common, and in industry, video cameras are used for quantity inspection and as robot eyes, in broadcasting for TV cameras, and in the medical field as very small cameras for stomach diagnosis. In the near future, there may be applications in visual telephones and as electric still cameras. Figure 4.8.1 shows a 1/3-inch (8.47 mm) 320 000 (320k)-pixel interline CCD (PAL format) made by Toshiba; Table 4.8.1 shows its main specifications.

Figure 4.8.2 shows the trend for CCDs estimated by the author, compared with the general trend for DRAMs. In general, the main specifications for CCDs are the diagonal length (i.e. optical format, e.g. 1/3 inch) and the (number of pixels (e.g. 320 000)). Home video cameras have become smaller and lighter year by year, accompanying the decrease in the size of CCDs. Currently among MOS devices, DRAMs are at the forefront of technology. CCDs have design rules similar to those for main-generation DRAMs, but in mass production, CCDs have about a one-year lag. For example, the 4M generation of DRAMs corresponds to the 1/3-inch 270 000-350 000-pixel CCDs. In CCDs, the size has decreased from 1/2-inch (12.7 mm) to 1/3-inch and a 1/4-inch (6.35 mm) size will soon appear. The successful production of 1/5-inch (5.08 mm) CCDs hinges on improving their sensitivity. On the other hand, the number of pixels has been between 270 000 and 400 000. For high-definition TV, CCDs with 2 million pixels will be used.

4.8.1 Fabrication and configuration of CCDs

A CCD consists of three parts: the silicon chip, the colour filter on the chip, and the package. As shown in Fig. 4.8.1, the silicon chip with colour filter is placed inside the package and sealed with a glass lid. The silicon chip, which is the main component of the image sensor, has four circuits: the photodiodes (PD), vertical CCD (VCCD), horizontal CCD (HCCD), and sense amplifier (SA) (see Fig. 4.8.3). The colour filter is directly formed on the chip by dyeing. To enhance the colour resolution, a complementary colour arrangement is used, as shown in Fig. 4.8.4. The cross-section and plan of a pixel are shown in Figs 4.8.5 and 4.8.6 respectively. The unit cell of a pixel is small: $9.6\ \mu\text{m} \times 8.4\ \mu\text{m}$ for 1/2-inch CCDs and $9.6\ \mu\text{m} \times 6.4\ \mu\text{m}$ for 1/3-inch

CCDs. The exposed area of the photodiode, the gate width, and the width of the buried channel are even smaller.

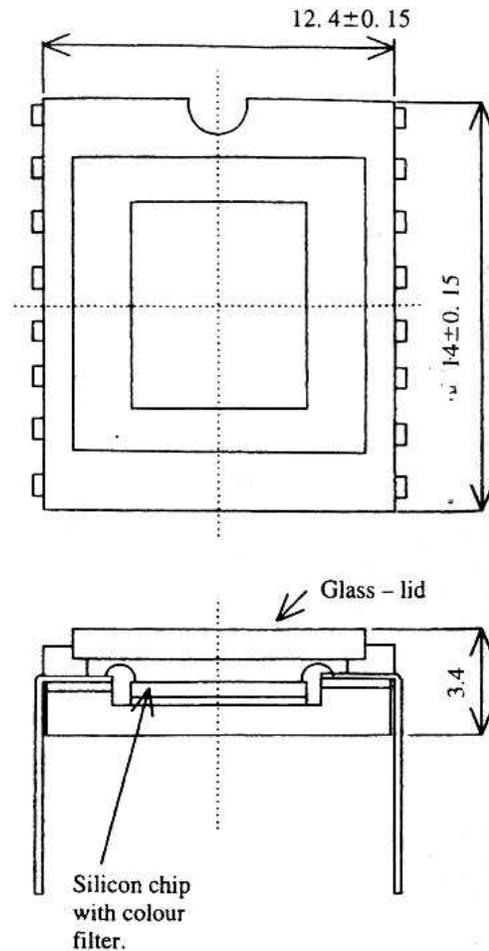


Fig. 4.8.1. External shape and cross section of CCD. ($\frac{1}{3}$ inch 320000 TOSHIBA).

Each photodiode in a pixel stores charges (electrons) in proportion to the light intensity and transfers these charges to the VCCD (see Fig. 4.8.3). The VCCD then transfers these charges to the next stage using a four- phase clock ($\Phi V1 - \Phi V4$). Thus the charges of an entire pixel column are transferred to the HCCD. The HCCD transfers these charges at a very high speed to the sense amplifier using a two-phase clock ($\Phi H1 - \Phi H2$). The sense amplifier amplifies these charges to drive the outside load. All charges in the pixel columns are transferred to the HCCD during one field period of TV scanning⁽¹⁾.

In the VCCD, charges are transferred as follows. Figure 4.8.7 shows the transfer timing for a VCCD.

Table 4.8.1 Main specifications of the Toshiba TCD 525 ID

Optical format	1/3 -inch (8 mm)
Total number of pixels	545 (H) x 587 (V)
Effective number of pixels	514 (H) x 581 (V)
Image size (mm)	4.9 (H) x 3.7 (V)
Unit cell size (/µm)	9.6 (H) x 6.4 (V)
Colour filter	Complementary mosaic type
Packaging	16 pin cerdip
Transfer method	Interline
Sensitivity (mV lx ⁻¹)	50
Saturation voltage (mV)	600
Dark current noise (at 60°C) (mV)	1 (standard), 3 (maximum)
Image lag (mV)	0(standard), 1(maximum)
Smear ratio (%)	0.015 (standard), 0.03 (maximum)
Blooming margin	1000 times standard luminous energy
Electric shutter speed (s)	1/50-1/10 000

First, for the odd fields of TV scanning, charges in the photodiodes of the n th and $(n+1)$ th pixels are transferred to their VCCD and combined. Then, according to the four-phase clock pulses, the combined charges are transferred to the next stage one by one. Next, for even fields, the charges in the photodiodes of the $(n-1)$ th and n th pixels are combined and transferred. The colour information comprises a set of eight pixels as shown in Fig. 4.8.4. For odd fields, the two combinations Gr + Ye and Mg + Ye are made in one row of the VCCD, and the other two combinations Mg + Cy and Gr + Cy in the adjoining row. For even fields, combinations Ye + Mg and Ye + Gr are made in one row, and Cy + Gr and Cy + Mg in the adjoining row. In a TV receiver the RGB information is formed from this information. In the visual spectral region, complementary colours are generally as follows: Ye = R + G; Cy = B + G; Mg = B + R.

4.8.2 Principle of CCDs

(a) *Photodiode*

When a field shift pulse is applied to the second polysilicon gate, free charges in the n⁺ layer of the photodiode are swept out to the VCCD; the n⁺ layer is thus depleted and new charges will be stored in proportion to the light intensity during the next field period of TV scanning. When light is focused on the photodiode, electrons in the valence band absorb light energy and are excited to the conduction band. This creates pairs of electrons and holes. The holes are carried out to the p⁺ layer of the photodiode and channel stopper (which surrounds the photodiode), and electrons are stored in the n⁺ layer of the photodiode. Figure 4.8.8 shows the potential curve of the photodiode.

(b) *VCCD*

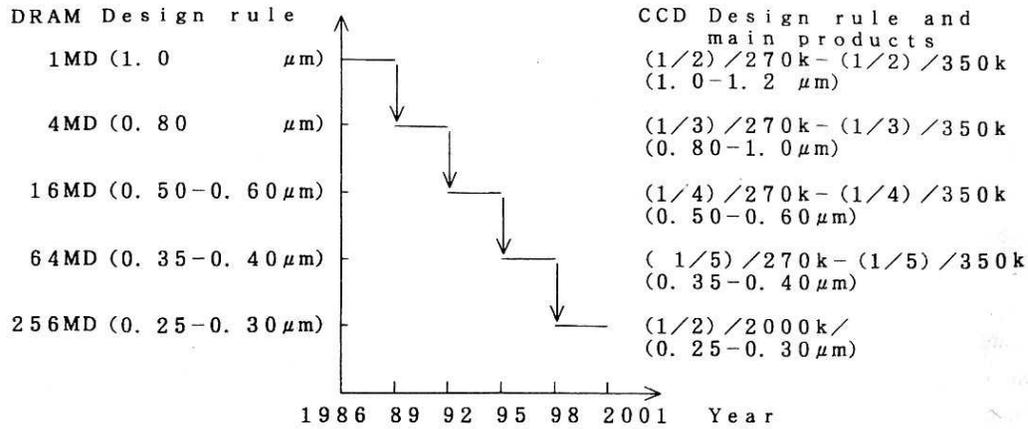
It is first necessary to remove all free charges in the n-type buried channel of the VCCD. This is done by applying a positive pulse to the top of the VCCD gate, which has a high voltage drain (OD), as shown in Fig. 4.8.3. Thus the n-type buried channel layers of the VCCD are depleted and will have positive charges. Figure 4.8.9 shows the potential curve of the VCCD. To transfer charges to the next stage, a negative gate voltage is applied to set the bottom of the potential at ~ 0 V. Conversely, to receive charges, the gate voltage is set equal to 0 V to set the bottom of the potential at a positive voltage.

(c) *HCCD*

The HCCD is different from the VCCD in three respects: (1) the buried channels of a HCCD consist of two types, n and n⁻, as shown in Fig. 4.8.10, and in their depleted state, the potentials of both layers are different; (2) the HCCD is driven by a two-phase clock; (3) the transfer speed is much faster than for the VCCD. To transfer charges to the next stage, phases Φ_{H1} and Φ_{H2} are set at 0 V and +5 V respectively. Charges stored in the n layer of H1 are transferred to the n layer of Φ_{H2} as shown in Fig. 4.8.10.

(d) *Sense amplifier*

The sense amplifier is constructed of two-stage source-followers, a floating capacitor, reset gate, and an output gate, as shown in Fig. 4.8.11. The gate of the first-stage source-follower is connected to the floating capacitor, which is the source of the reset gate. The output gate, located between the floating capacitor



- * Solid lines show the main generation of DRAMs.
- * In mass production of CCDs, the design rule for DRAM for that generation seems to be adopted within one year after mass production start of that DRAM.

Fig. 4.8.2. Trend for CCDs estimated by the author, compared with the general trend for DRAMs.

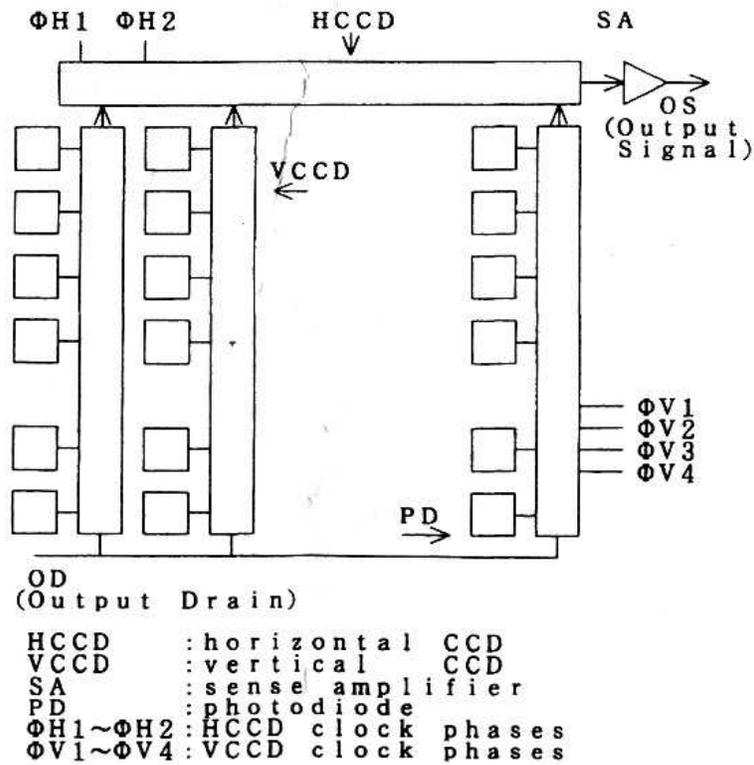


Fig. 4.8.3. Configuration of CCD.

and the end gate of the HCCD, blocks charges from the HCCD. To provide a reference potential for the floating capacitor, a reset pulse is applied to the gate so that the potential of the floating capacitor becomes

equal to the potential of the reset drain. When the end gate of the HCCD is set equal to 0 V, charges flow into the floating capacitor, lowering its potential by the negative charges. The potential difference from the reference is amplified by the two-stage source- followers.

4.8.3 Manufacturing process

Figure 7.4.5 shows the cross-section of a unit cell and Table 4.8.2 the total process flow for CCDs.

(a) Wafer process

The photodiode has four layers of impurities in the longitudinal direction of the silicon: p⁺, n⁺, third P-well, and n-substrate. The n-substrate is usually an epitaxially grown layer on the base n-type substrate for preventing swirls, i.e. image defects that look like arcs. The third P-well is formed by an Mev implanter and its depth is 3-4 μm from the silicon surface. The p⁺ and n⁺ layers are formed by a self-alignment implant process applied to the polysilicon gate. An Mev implanter is used to form the n⁺ layer to stabilize the amount of impurities, leading to a stable saturation voltage of the signal (V_{sat}).

The p⁺ layer has three important functions. The first is to reduce the dark current generated from the surface states to 1/5-1/6 of that without the p⁺ layer.

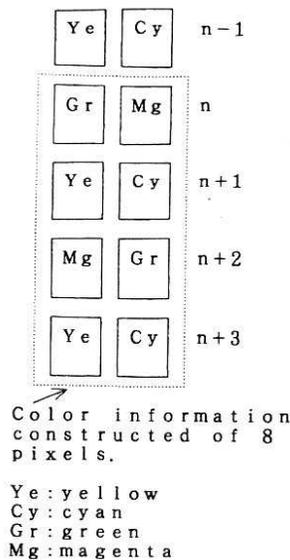


Fig. 4.8.4. Colour arrangement.

The second is to remove the image lag effectively by maintaining the potential of p + at 0 V. The third is to make it possible to dump charges of the photodiode to the overflow drain quickly by maintaining the potential of p + at 0 V in electronic shutter timing, making the variable-speed shutter possible⁽²⁾.

The function of the third P-well is to sweep out extra charges rapidly to the overflow drain. If this is not done, when the photodiode is exposed to strong light, the screen TV image will have blooms due to the extra charges pouring into the VCCD from the overflowing photodiode⁽³⁾.

The VCCD has five layers of impurities: n-type buried channel, first P-well, second P-well, third P-well, and n-substrate. Below the n-type buried channel, the first P-well is formed to avoid smears that appear as bright spots. The oxide film of the gate is usually thermally grown oxide or ONO (oxide- nitride-oxide). The gate thickness is larger than that of DRAMs or gate array (11-15 nm), because a higher voltage is applied and the leakage current must be minimized.

The first and second n+ polysilicon gates overlap each other, as shown in Fig. 4.8.6, to make the charge transfers smooth.

The first metal layer is formed on the first insulator layer to shield the VCCD from light. It is necessary to make the thickness < 300 nm to reduce smears. To

Table 4.8.2 CCD process flow.

Process	Remarks
Silicon wafer process	Forming the silicon image sensor
On-chip colour filter process	Forming colour filter on the chip
First test	Selecting good chips for packaging
Packaging	Assembling the chip with colour filter on a
Second test	Final test before shipment

reduce smearing further, the insulator thickness should be made as thin as possible, as shown in Fig. 7.4.12. The material of the first metal layer is usually Al, W, or some other heavy metal⁽⁴⁾, Even a 5 nm pinhole will cause an image defect. Therefore a second shield metal layer is applied over the first metal layer. The insulators are deposited on the metal to form a smooth plane and protect the metal.

On top of the silicon chip, a nitride film is deposited to block penetration by heavy metals or alkali metals such as Na, prevent the filtration of humidity, and reduce the surface states of silicon.

Table 4.8.3 summarizes the film thicknesses and their tolerances.

(b) Colour filter

In colour CCDs, one colour is assigned from yellow, cyan, magenta, and green on each pixel. On top of this, a microlens is formed. The colour filter is formed directly on the silicon chip. First, the base layer is coated to make the surface smooth. Next, a dyeing layer, usually of casein, is coated and engraved by a photo-engraving process. The residual part is dyed by dipping in a dyeing liquid. Then the interlayer is coated, separate between the upper and lower dyed layers. To reduce the thickness of the colour filter, the interlayer can be omitted by special treatment of the dyed layers, i.e. fixed colour treatment.

After three alternate layers of dyed layer and interlayer are repeated, the top layer is coated for protection and a microlens is formed, the thickness and shape of which are controlled to focus light on the photodiode. Green colour is produced by overlaying yellow and cyan. The total thickness of the colour filter is $\sim 6-9 \mu\text{m}$.

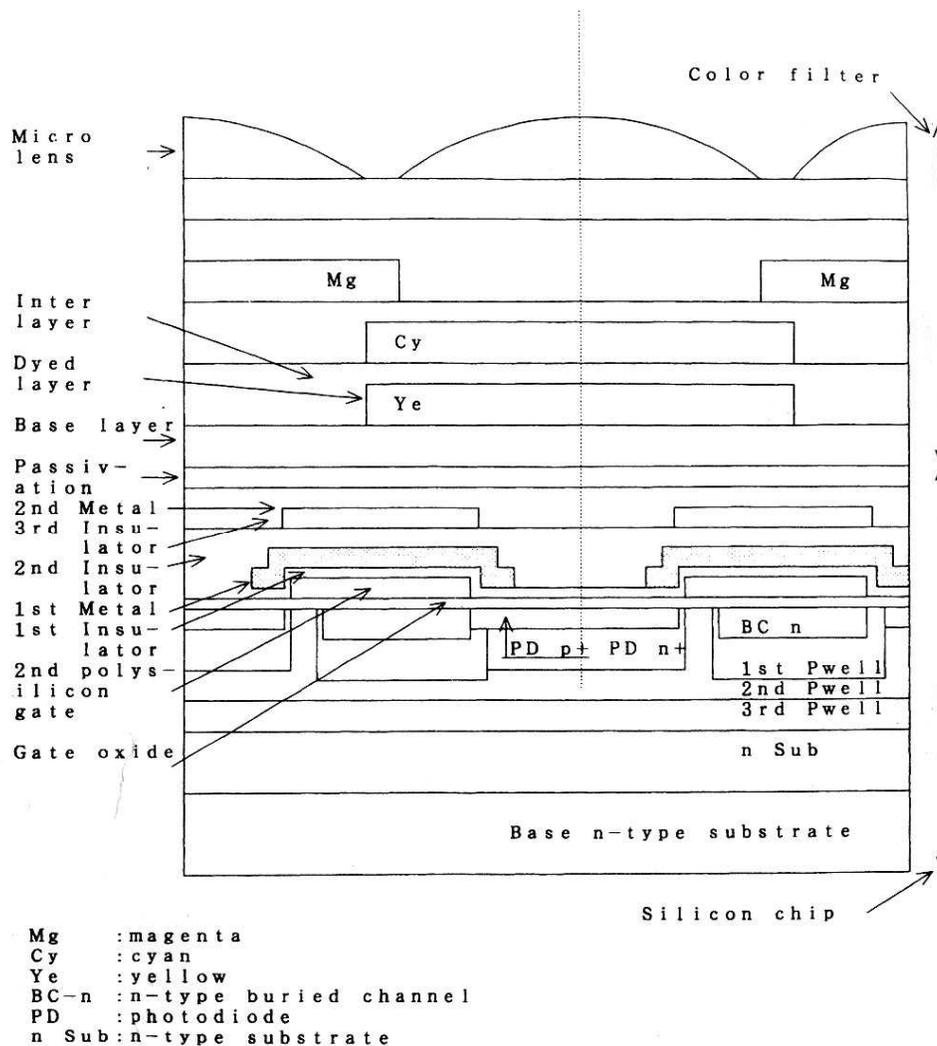


Fig. 4.8.5. Cross section of unit cell.

(c) *Packaging*

After the colour filter is formed, a first test is made to eliminate inferior chips. Only good chips are packaged. Figure 4.8.13 shows the packaging process, which consists of dicing, mounting, wire-bonding, glass-lid setting, and sealing. During the process, if a 1 μm particle is deposited on the photodiode, the chip's performance will suffer. So the packaging process has to be carefully conducted to prevent foreign particles on the chips. In particular, silicon chippings produced by the dicing process are carefully removed.

(d) *Testing*

Two tests are performed for CCDs. The first is conducted at high temperatures to detect defects such

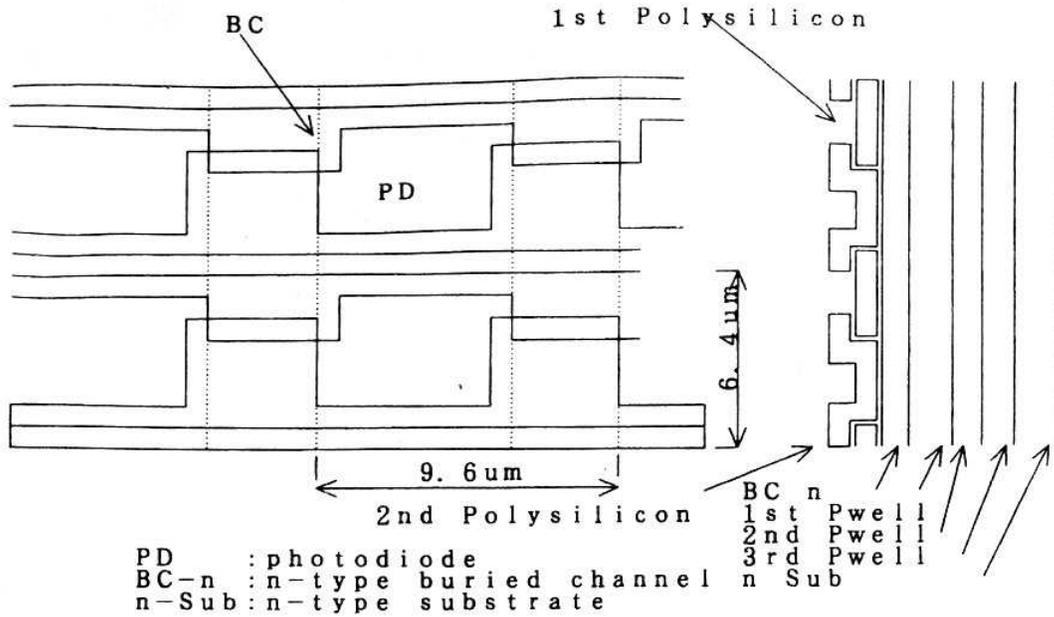


Fig. 4.8.6. Basic construction of VCCD.

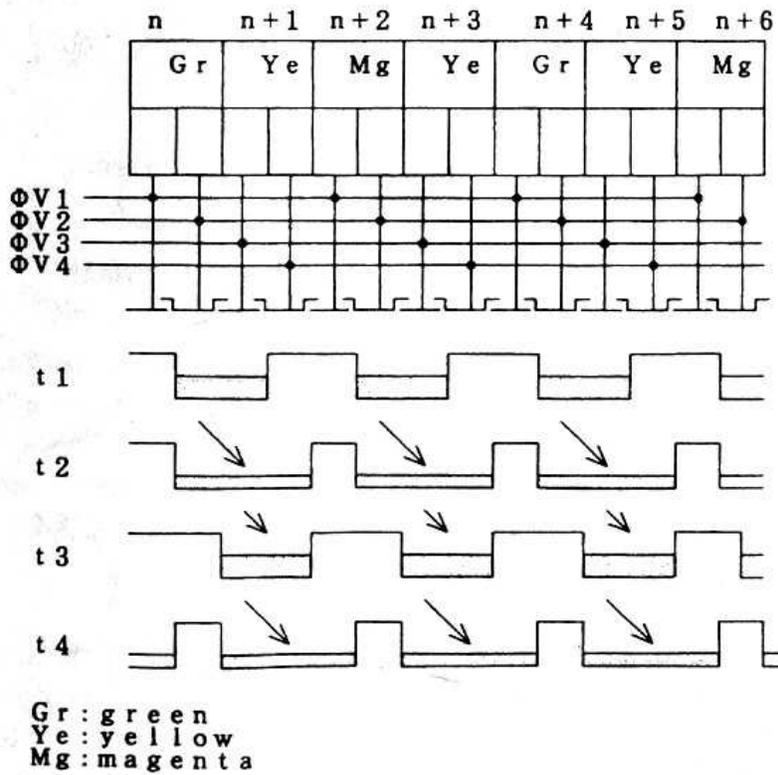


Fig. 4.8.7. Transfer timing of VCCD.

as currents generated by crystal defects in the silicon. The second test is conducted at room temperature to detect defects resulting from packaging and defects such as the presence of fixed patterns in the vertical lines, which are easier to detect at room temperature. Although automated testing to replace visual inspection by humans is important from the point of view of

Table 4.8.3 Film thicknesses generally used in CCDs.

	Film thickness	Thickness variation ($\pm 3\sigma$) (nm)
Gate oxide	60-100	3-5
Gate polysilicon	300-600	30-0
1st insulator	100-300	10-30
1st metal	400-1000	40-00
2nd insulator	350-1000	35-100
2nd metal ^a	500-800	50-80
3rd insulator	350-500	35-50
Passivation	350-500	35-50

^a Not now used.

economy and exactness, it is difficult and time-consuming, whereas the human eye can detect flaws of 100nm as well as very small image irregularities.

4.8.4 Nanotechnology in the mass production of CCDs

There are three applications of nanotechnology in the mass production of CCDs: (1) the technology dealing

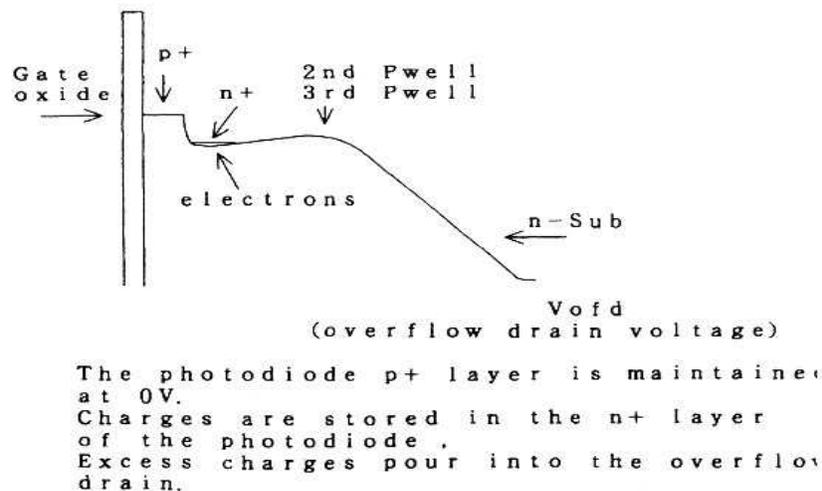


Fig. 4.8.8. Potential profile of PD.

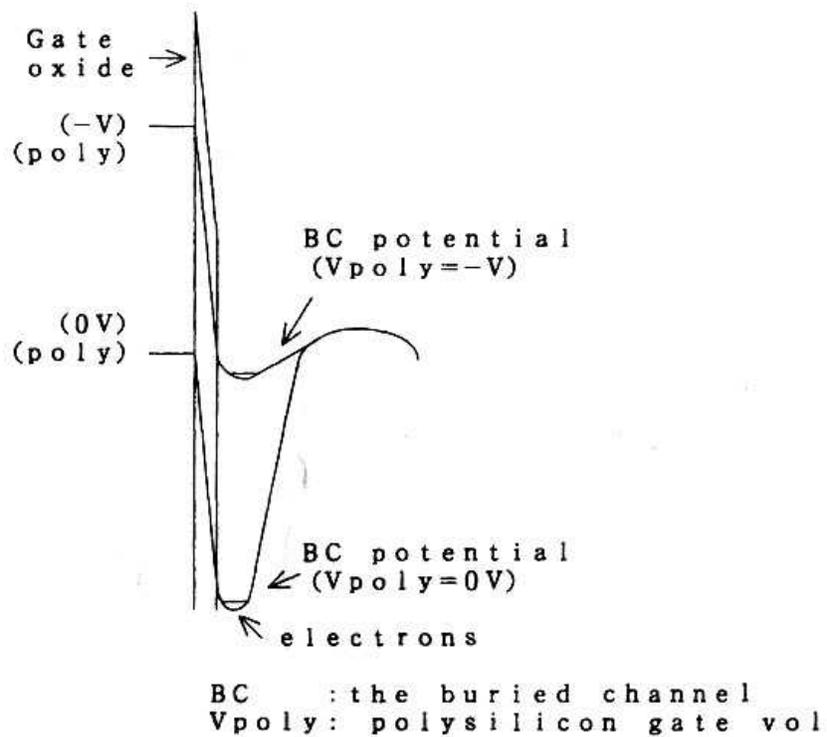


Fig. 4.8.9. Potential profile of VCCD.

with very small signal currents of 1 fA-1 pA; (2) the technology involved in the total process of eliminating image defects such as 100 nm flaws and observable image irregularities; and (3) that needed to achieve accuracy in aligning and working each element to within ± 300 nm.

(a) Technology of very small currents

Figure 4.8.14 shows the relation between the output voltage of the sense amplifier and the number of input electrons, in which the main defects are shown.

The relation between the number of input electrons (combined charges of two photodiodes) to the sense

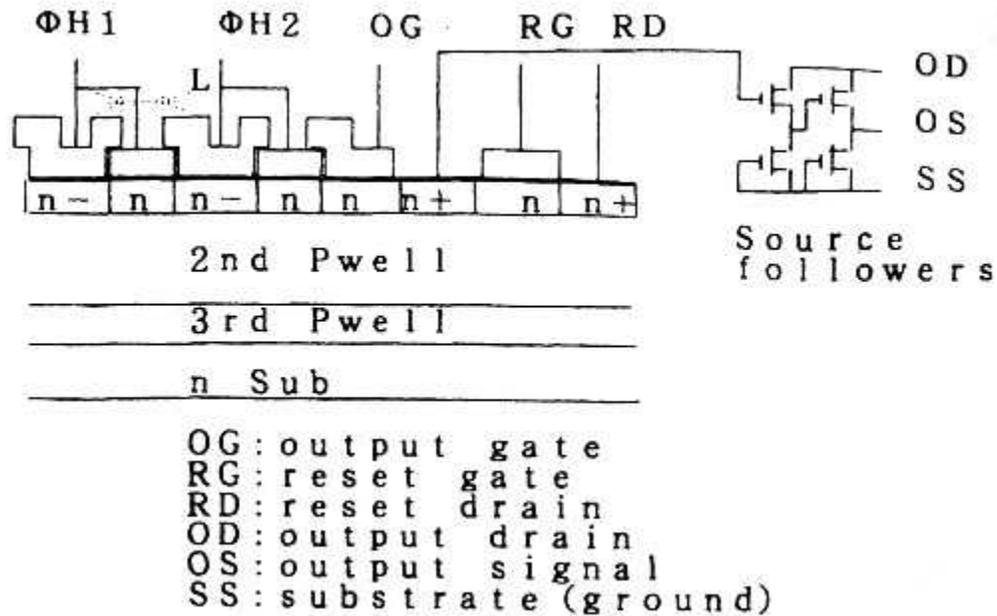


Fig. 4.8.10. Basic structure of HCCD.

amplifier and the sum of currents generated in two photodiodes during one field period of TV scanning is

$$I = nq / t$$

where I is the sum of currents generated in the two photodiodes (A), n is the number of electrons stored in two photodiodes during one field period of TV scanning, t is one field time of TV scanning, and q is the charge of an electron, 1.6×10^{-19} C. If $t = 20$ ms and $n = 125$, then $I = 1$ fA.

The relation between the output voltage and the number of input electrons is

$$V = ns$$

where V is the output voltage (V) and s is the sensitivity of the amplifier ($\mu V e^{-1}$). If $s = 16 \mu V e^{-1}$ and $n = 125$, then $V = 2$ mV.

The main defects in CCDs, their current levels, and preventive measures are discussed below. Figure 7.4.15 illustrates the image defects.

1. *Vertical lines.* Fine vertical lines are generally caused by leakage currents of > 0.1 fA (13 electrons) from the VCCD gate, or a bad transfer from VCCD to HCCD. Preventive measures include the use of a gate oxide with a low leakage structure and field strength concentration or ONO, and making the transfer path of the charges smooth, so that the flow of electrons is not obstructed.

2. *Dark-current noise.* This appears as image noise with low brightness, and is also called fixed pattern noise. It is created by the current generated from the surface states ($\sim 0.75\text{-}1.5\text{ fA}$, or 94-195 electrons at 60°C). Countermeasures include reducing contamination by heavy metals, annealing in an inert gas, terminating dangling bonds in the silicon structure,

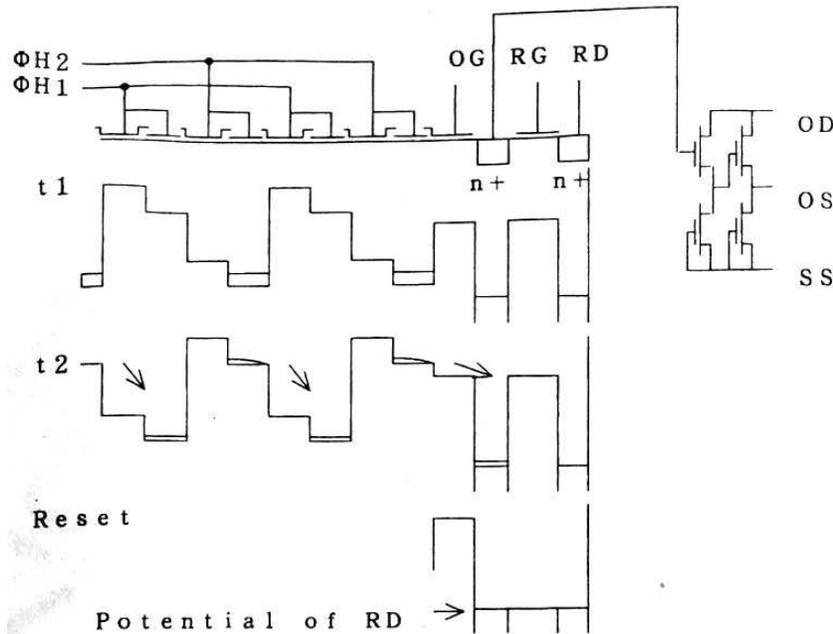


Fig. 4.8.11. Timing of HCCD.

and forming a p+ layer on the n+ layer of the photodiode.

3. *Pixel defects*

(a) *White spots.* These appear to be caused by the currents generated from defects in the crystal silicon by contamination by heavy metals, or by the complex interaction of implant damage and contamination. The current involved is $\sim 1.5\text{-}3\text{ fA}$ (195-375 electrons at 60°C). The main countermeasures are: (1) to use an n-substrate which has a layer with low oxygen concentration, such as an epitaxial growth layer, to create an intrinsic gettering layer in the base n-type substrate; (2) to reduce contamination by heavy metals in all processes; (3) to perform gettering near the end of the process; and (4) to keep the number of impurities in the photodiode n+ layer as small as possible, so that V_{sat} is satisfied.

(b) *Black spots.* These are mainly caused by particles on the photodiode. A projection of the Al hillock in the photodiode or a notch in the metal shield fine can also cause a black spot. The current involved is above $\sim 5\text{ fA}$ (625 electrons). Countermeasures include minimizing foreign particles in all processes from wafer fabrication to packaging.

4. *Deficiency of V_{sat}* - This appears as an irregularity in brightness, and is caused by a deficiency in the number of impurities in the n + photodiode and variations in the concentration in the third P-well. For a V_{sat} of 600 mV, 3.75×10^4 electrons are stored in the two photodiodes. The main countermeasure is to stabilize the implantation process by using an Mev implanter.

If we compare the number of input electrons to the sense amplifier of a CCD with that of a 4M DRAM (which stores $\sim 6 \times 10^5$ electrons in a unit capacitor of ~ 40 fF to build a voltage of 2.5 V), we see that the signal electrons in the CCD are much smaller in number. The number of electrons in CCDs is $\sim 1/6000-1/16$ of that for DRAMs. Hence to reduce the current generated from defects in the silicon to a level of 1.5 fA, it seems at least necessary to reduce contamination by heavy metals to a level of 10^{10} atoms cm^{-2} .

(b) Technology to prevent image irregularities

1. *Horizontal line.* This appears as a fine black line, and is caused by irregularities in the mask pattern. If this occurs, the mask pattern must be checked for irregularities > 100 nm.

2. *Diagonal line of colour filter.* This appears as fine diagonal lines at the corner of the picture, and is caused by an unevenness of ~ 50 nm in the thickness of the surface coating, which is the result of unevenness in the silicon chip pattern. Countermeasures include stricter control of the coating conditions and making the silicon chip pattern even.

3. *Colour flicker.* Between the even and odd fields of TV scanning, the signals carrying colour information

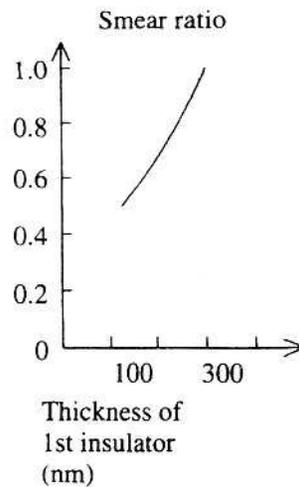


Fig. 4.8.12. Relation between smear ratio and thickness of 1st insulator.

are transferred among different pixel combinations. Therefore if the signals have different magnitude, this appears as flicker in colour. The countermeasure is to reduce variation among pixels in the colour spectrum. 4. Fine flaws. A 100nm flaw can be apparent to the eye. Handling is therefore carefully done in all processes, especially in packaging. These defects are checked by the highly sensitive human eye. In other non-image devices, these flaws create no problems.

(c) *Technology for achieving accuracy in aligning and working elements in the wafer process.*

1. *Accuracy in aligning and working.* Table 4.8.4 shows the author's estimated targets for accuracy of aligning and working elements in CCDs, compared with the general targets for DRAMs. The design rule for CCDs is nearly the same as for DRAMs. However, in DRAMs the minimum working dimension is the gate length of a unit cell, whereas in CCDs it is the etched length of the HCCD gate, indicated by L in Fig. 4.8.10. Moreover, the separation between unit cells is achieved in a different manner. In DRAMs, separation between unit cells is usually effected by insulators such as LOCOS (local oxidation of silicon) and the implanted impurities underneath. In CCDs however, it is effected only by implanted impurities called channel stoppers. This is because in CCDs there are no wirings over the gates, crossing the separations. Signal separation between adjoining rows of VCCDs is affected by alignment errors in the implantation process in these adjoining cells.

As shown in Fig. 4.8.16, signal charges in the

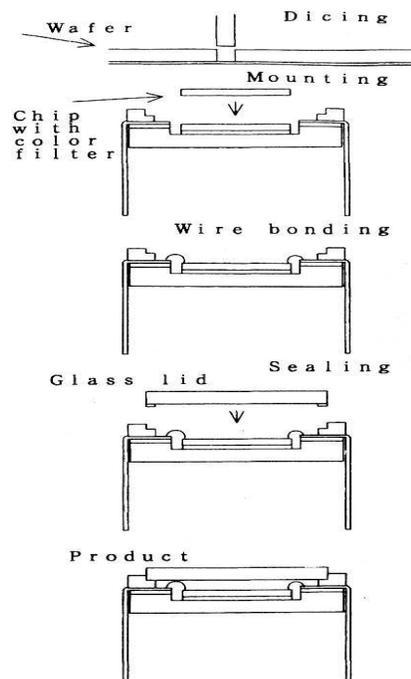


Fig. 7.4.13. Process flow of packaging.

photodiode can pour into three different ports when errors in the alignment and distributed concentration of the implanted impurities exceed certain thresholds. The first is the port to its own buried channel, when the shift pulse is applied to the gate. The second is the port to the overflow drain, when an intense light impinges on the photodiode. The third is the port to the adjoining buried channel, which when taken will cause a defect. This can be caused by an alignment error of > 300 nm in the first P-well.

2. *Angle of ion beam in the implantation process.* The angle of the ion beam when implanting the n^+ and p^+ layers of the photodiode has a large effect on the

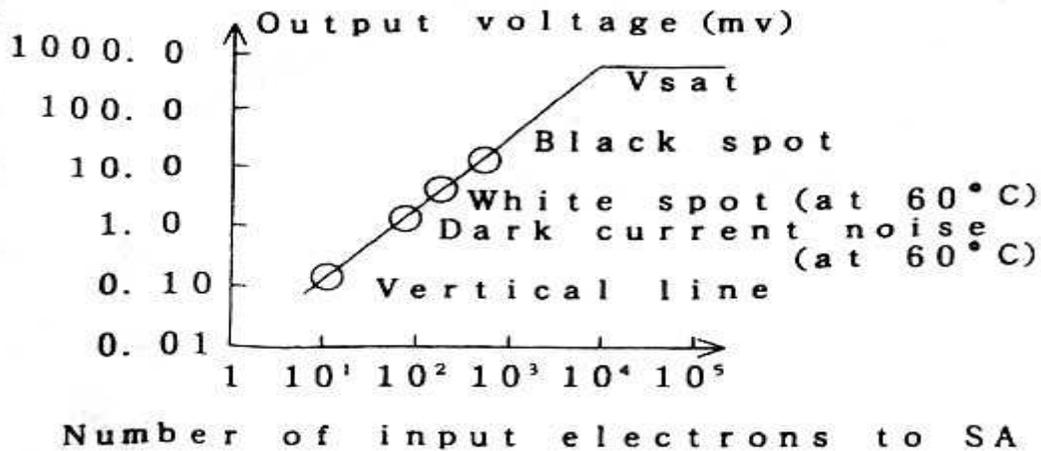


Fig. 4.8.14. Relation between number of input electrons to SA and output voltage, in which main defects are plotted.

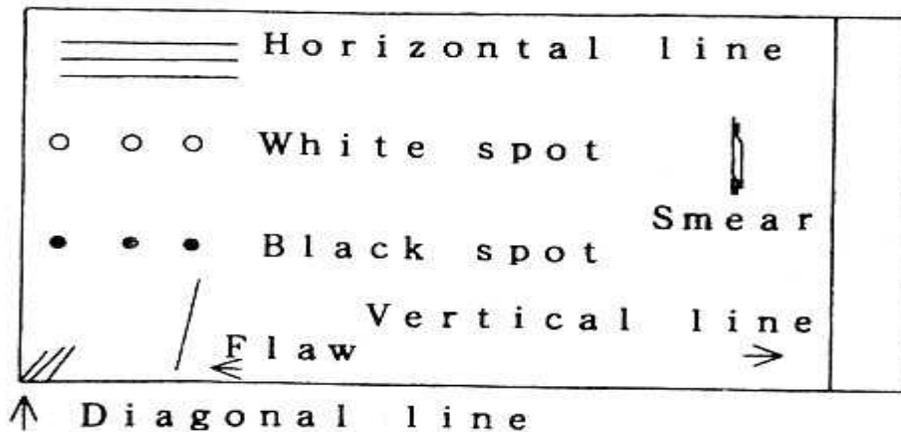


Fig. 4.8.15. Image defects.

magnitude of the field shift pulse voltage (V_{fs}), which is the voltage at which the image lag is zero. Figure 4.8.17 shows the relation between the angle conditions for n+ and p+ layers and V_{fs} . Both are implanted by self-alignment to the gate. Angle A is $+10^\circ$ and angle B is -10° . If the angle for the p+ layer is A and for the n+ layer is B, V_{fs} is small. If the angle for the p+ layer is B and for the n+ layer is A, V_{fs} is large. The difference between A and B creates a 140 nm difference under the silicon surface. When the p+ layer projects from the edge of the gate towards the buried channel, this seems to create a potential barrier and raises V_{fs} to a high level. So the n+ and p+ layers of the

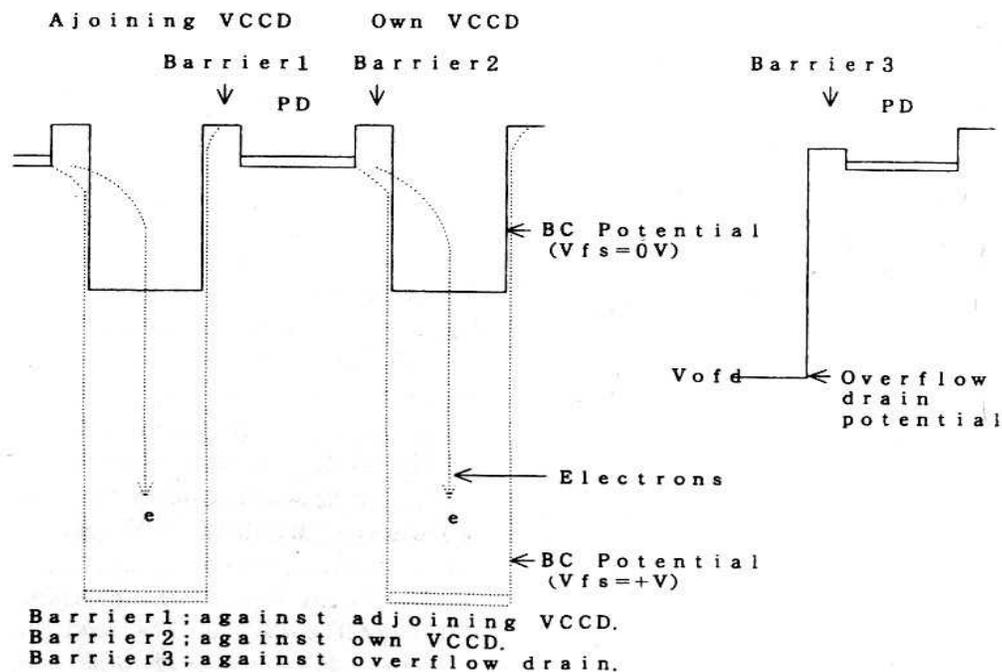


Fig. 4.8.16. Potential barriers of the photodiode against three ports.

photodiode are implanted in a manner so as to keep V_{fs} small.

The author would like to thank M. Kubo, T. Yamada, K. Sato, K. Sekine, and T. Usami of the CCD Image Sensor Engineering Dept, Y. Iizuka and other staff members of the CCD Process Engineering Section, the staff of the CCD Image Sensor Application Dept. and Iwate Toshiba Electronics for- valuable advice, discussions and technical support in the preparation of this text.

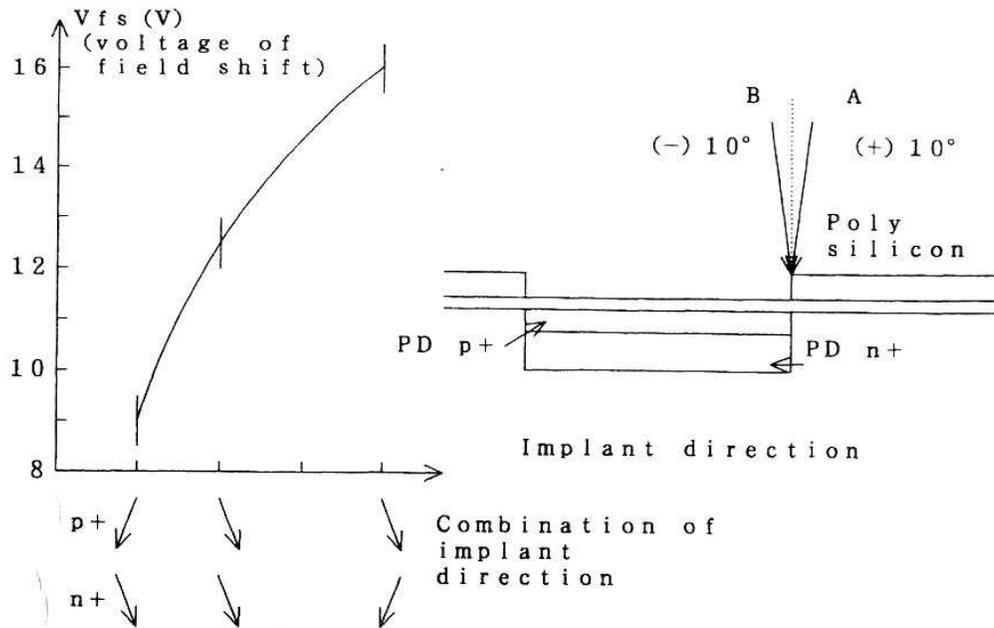


Fig. 7.4.17. Relation between V_{fs} and combination of implant direction.

Table 7.4.4 Estimated targets for design rule and accuracy of aligning and working CCD elements, compared with those generally used for DRAMs.

CCD	1/2/270k	1/3/270k	1/4/270k	1/5/270k	1/5/3
	1/2/350k	1/3/350k	1/4/350k		50k
DRAM	1M	4M	16M	64M	
Design rule (μm)	1.0	0.8	0.5-0.6	0.35-0.40	
Accuracy of aligning ($\pm 3\sigma$) (μm)	0.30	0.20	0.10-0.12	0.060-0.080	
Accuracy of working ($\pm 3\sigma$) (μm)	0.30	0.20	0.10-0.12	0.060-0.080	

References

1. Yamada, T., Yanai, T., and Kaneko, T. (1987). 2/3 Inch 400,000 pixel CCD area image sensor. *Toshiba Review*, (162), 16-19.
2. Kuriyama, T., Kodama, H., Kozono, T., Kitahama, Y., Morita, Y., and Hiroshima, Y. (1991). A 1/3-in 270000 pixel CCD image sensor. *IEEE Transactions on Electron Devices*, 38, 949-53.
3. Hojo, J., et al. (1991). A 1/3-in 510(H) 492(V) CCD image sensor with mirror image function. *IEEE Transactions on Electron Devices*, 38, 954-9.
4. Toyoda, A., Suzuki, Y., Orihara, K., and Hokari, Y. (1991). A novel tungsten light shield structure for high density CCD image sensors. *IEEE Transactions on Electron Devices*, 38, 965-8.

4.9 VCR Head Assemblies

A video cassette recorder (VCR) is a device used for recording and replaying video and audio signals. These signals are recorded on and read out from magnetic tapes using magnetic heads. In a VCR system, magnetic tape runs on a tape guide and a cylinder unit with magnetic heads rotates to read signals from and write signals on to tapes. Key VCR mechanisms are the tape running guideway and cylinder unit. As shown in Fig. 4.1 .la, there is a very narrow gap (0.5 μm) at the top of the magnetic head. In replaying, this narrow gap must follow the track very precisely. The setting accuracy for heads on the cylinder unit is therefore a very important factor in picture quality and compatibility.

On conventional production lines, these magnetic heads are set manually by skilled operators, by observing optically enlarged head images on monitor displays. In this adjustment work, the position of the narrow gap is used for reference. This work is tedious, slow, and unreliable. We have therefore developed an automatic adjustment system for VCR magnetic heads on the cylinder unit.

4.9.1 Adjustment items and specifications

There are three adjustment items and setting errors for the magnetic heads, as shown in Fig. 4.9.1b:

- (1) setting angle error for two symmetrically oppositely positioned heads;
- (2) rotational error for every individual head relative to the setting screw axis;
- (3) distance of protrusion of head tip from cylinder edge.

As shown in Table 7.7.1, adjustment of micrometre order accuracy is required.

4.9.2 System configuration

In this system, the head gap position is detected by a small interferometer with a reference plane as shown in Fig. 4.9.2. The positioning errors for heads are

Table 4.9.1 Adjustment specifications for VCR head assemblies

Setting angle error (μm)	± 3.0
Rotational error (arc min)	± 14.0
Protrusion distance (arc s)	± 25.0

calculated by an image-processing computer using two head images (with and without interferometric fringes).

As shown in Fig. 4.9.3, two head images through two optical systems are transferred to image processing computers and calculated parameters are transferred to the adjustment mechanism.

As shown in Figs 4.9.4 and 4.9.5 the concepts for calculating head positioning errors involve the following. 1

1. The setting angle error is calculated from the mutual positions of the two head gap images relative to cursor positions on the monitor display. These cursor positions are calibrated as 180° opposite each other using the standard master cylinder unit.

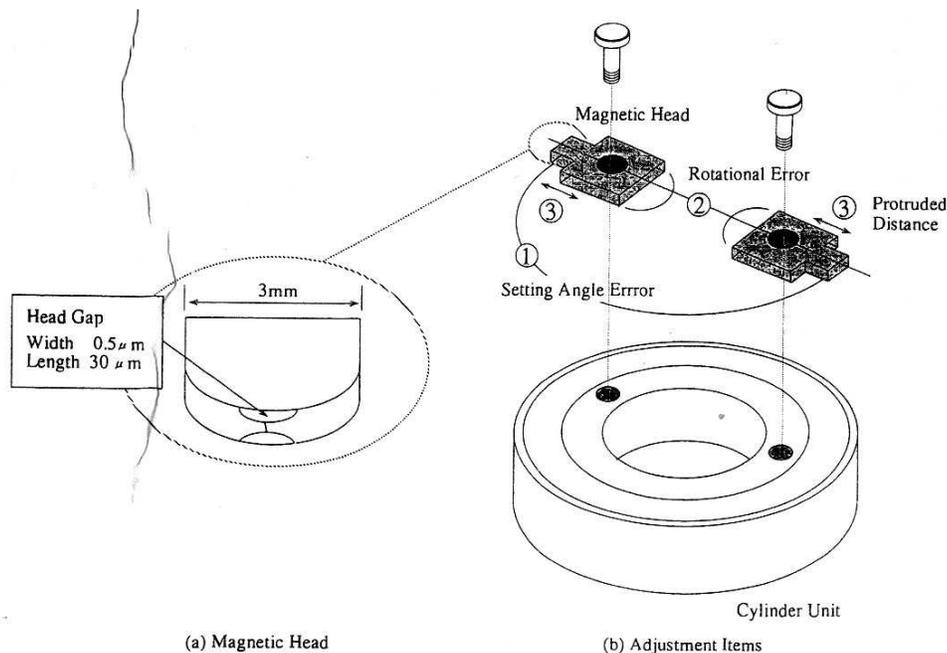


Fig. 4.9.1. VCR head assembly, (a) Magnetic head, (b) Cylinder unit.

2. Rotational error for the screw axis is calculated from the distance between the centre of the interfero-metric fringes and the centre of the gap.

3. The protrusion distance from the cylinder edge is adjusted, where the fringe contrast is higher than a certain threshold level, using interferometric optics. The interferometric optics are also calibrated using the master cylinder unit.

Calculation accuracies for the gap position and fringe centre directly affect the adjustment accuracies. High-speed and highly accurate and reliable algorithms are therefore required, even when a head image is noisy or unclear because of dust particles on the head gap.

4.9.2 Image processing algorithm

The software configuration of this system is as follows:

- (1) calibration for optical magnifying force and cursor positions;
- (2) autofocusing for the optical system;

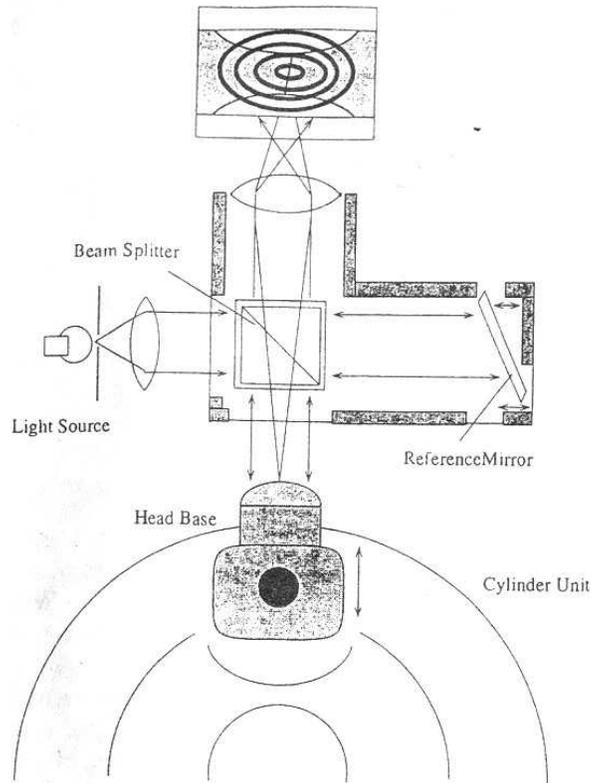


Fig. 4.9.2. Optical system for setting.

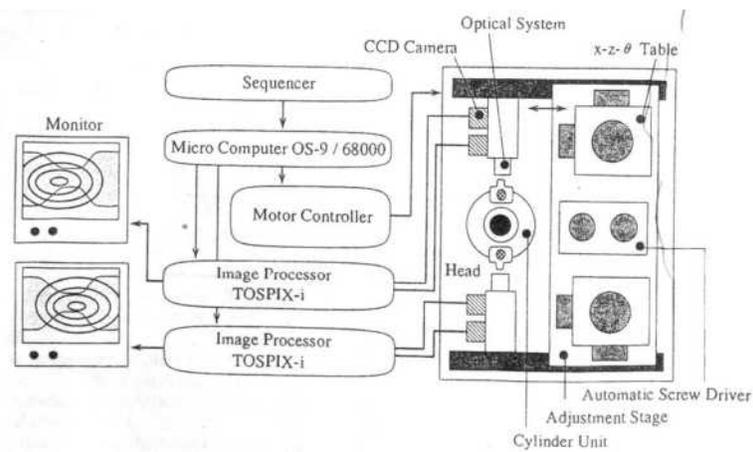


Fig. 4.9.3. Total system configuration.

- (3) gap position detection: rough and fine;

(4) fringe analysis: pairing left and right side fringes and calculating fringe centre and fringe size.

The main algorithm is used for calculating the gap position and fringe analysis. On a production line, over a million heads must be assembled per month, so robustness, flexibility, and adaptivity of the algorithm are very important factors.

First, image data are binarized by threshold level $L1$ and the rough position of the gap centre is determined by x , y projections as shown in Fig. 4.9.6. Then two windows are set on the upper and lower gap positions, and within the window the second binarized level $L2$ is determined by the equation

$$L2 = (\text{Average intensity on a line in the window}) - C, \text{ where } C \text{ is a constant.}$$

Next, as shown in Fig. 4.9.7, using the image binarized by $L2$, the x -position of the head is decided at the deepest point (G_x) in the projection on the x -axis. Again, calculating the T -projection/ in the image binarized by $L1$, the precise y -position of the gap (G_y) is determined by the x -position of the gap (G_x). This algorithm depends on calculating the projection for two different images binarized by different threshold levels $L1$ and $L2$. This method is not affected by noisy head image and dust particles in the gap.

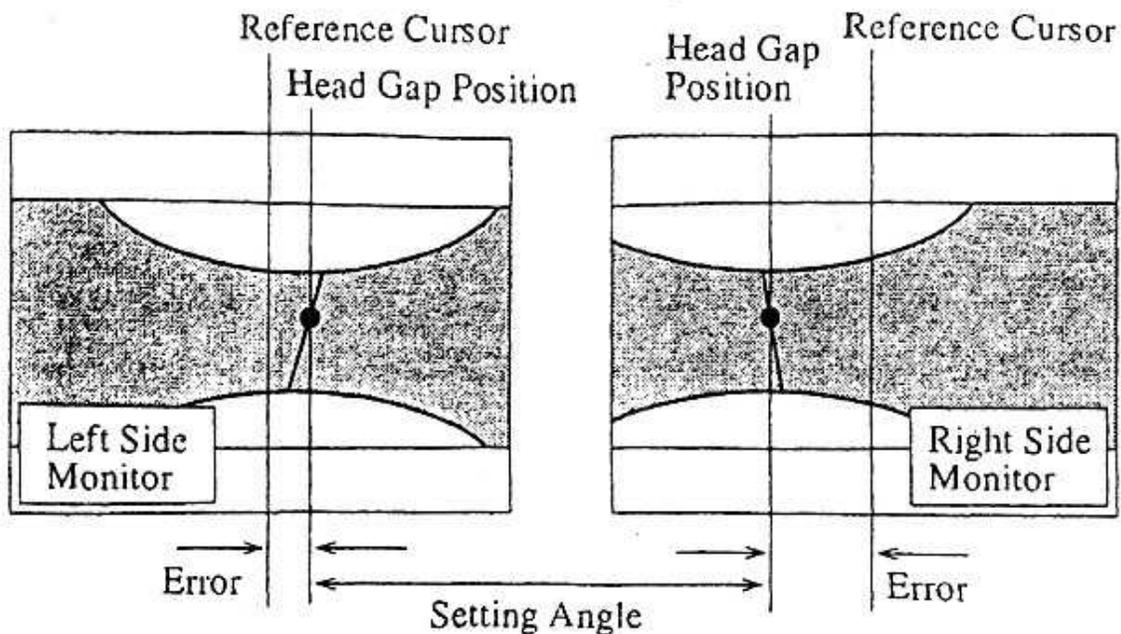


Fig. 4.9.4. Angle error detection.

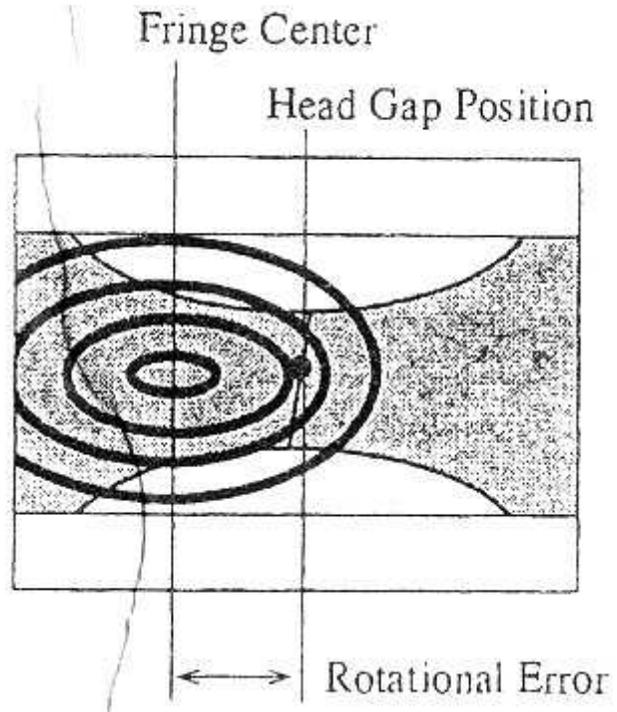


Fig. 4.9.5. Rotational error and protrusion distance detection.

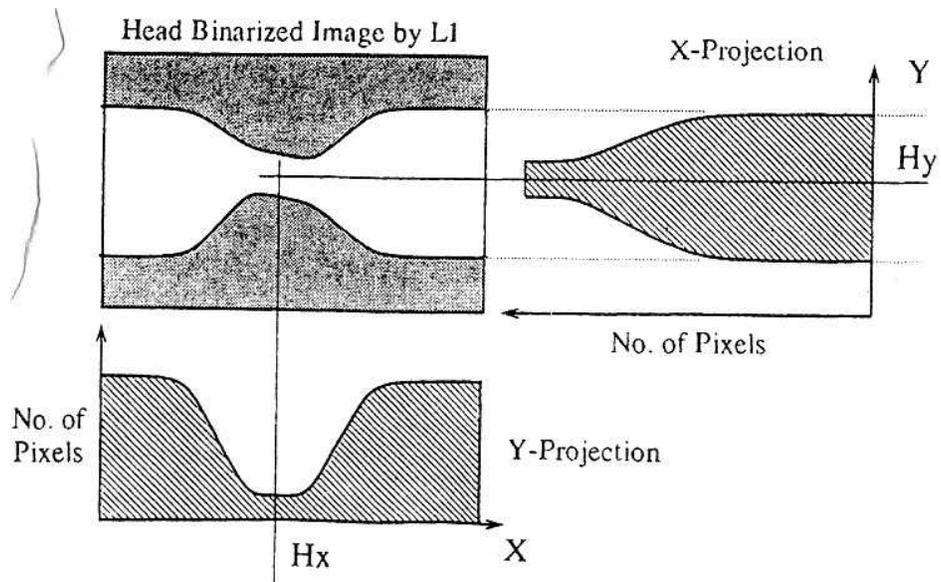


Fig. 4.9.6. Rough positioning of head.

For fringe analysis, only the fringe pattern is extracted by subtracting two images, with and without the fringe pattern. After calculation of the run-length and circumscribed rectangles for the fringe, these fringes are classified and paired as shown in Fig. 4.9.8. After omission of irregular fringes, the precise fringe centre is decided. Also, fringe size is represented by the distance between the highest contrast in the left- and right-hand fringes as shown in Fig. 4.9.9.

4.9.3 Adjustment mechanism and sequence

The adjustment mechanism is one of the key points in this system. Pulse motors are used as actuators for accurate control of rotational angle, speed, and return angle of the screw. To control the screwing torque, mechanical slip clutches are used. Also, to prevent

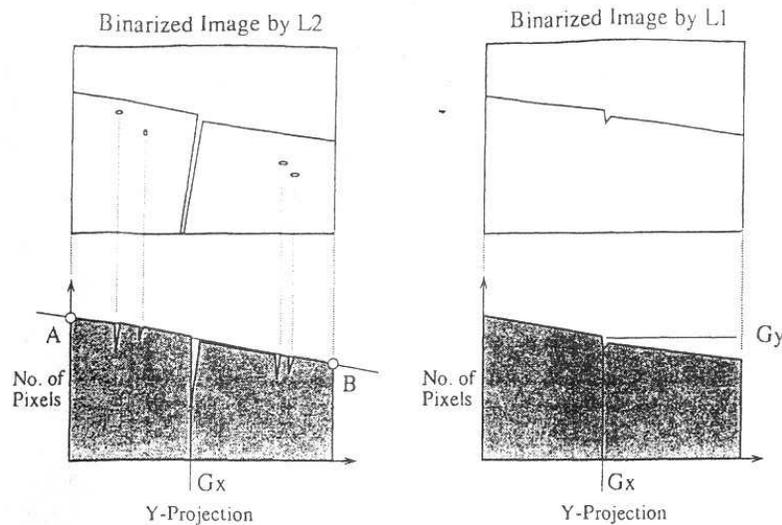


Fig. 4.9.7. Gap position detection in the window.

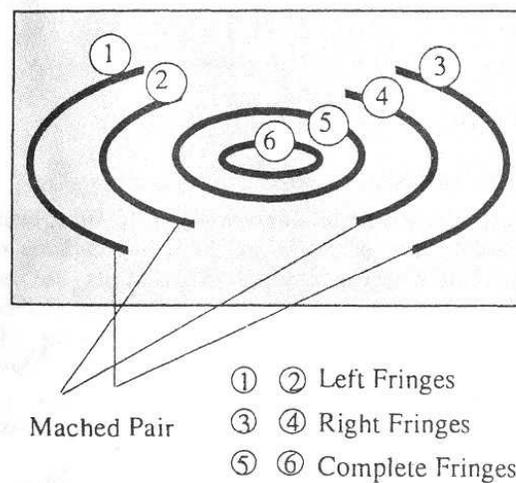


Fig. 4.9.8. Fringe analysis.

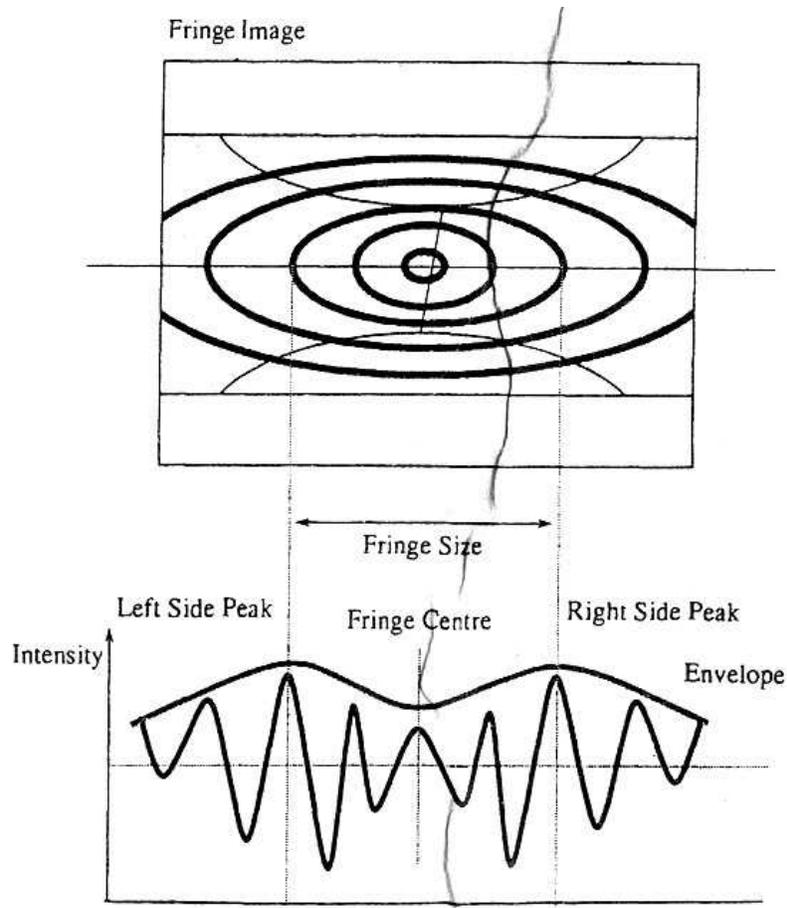


Fig. 4.9.9. Fringe intensity profile.

head movement due to unbalanced force applied by the driver bit, a remote centre compliance mechanism is used.

The adjustment sequence is shown in Fig. 4.9.10. After final checking for fringe pattern, if the result is unsatisfactory the magnetic heads are adjusted and a second check is made. This sequence is repeated up to three times, and if the results are still unsatisfactory the heads are considered defective.

There some statistical tendencies for head movement, such as slipping and elastic deformation of the head base, depending on the head maker or head lots. These parameters are therefore incorporated in the feedback software to compensate for irregular head movement.

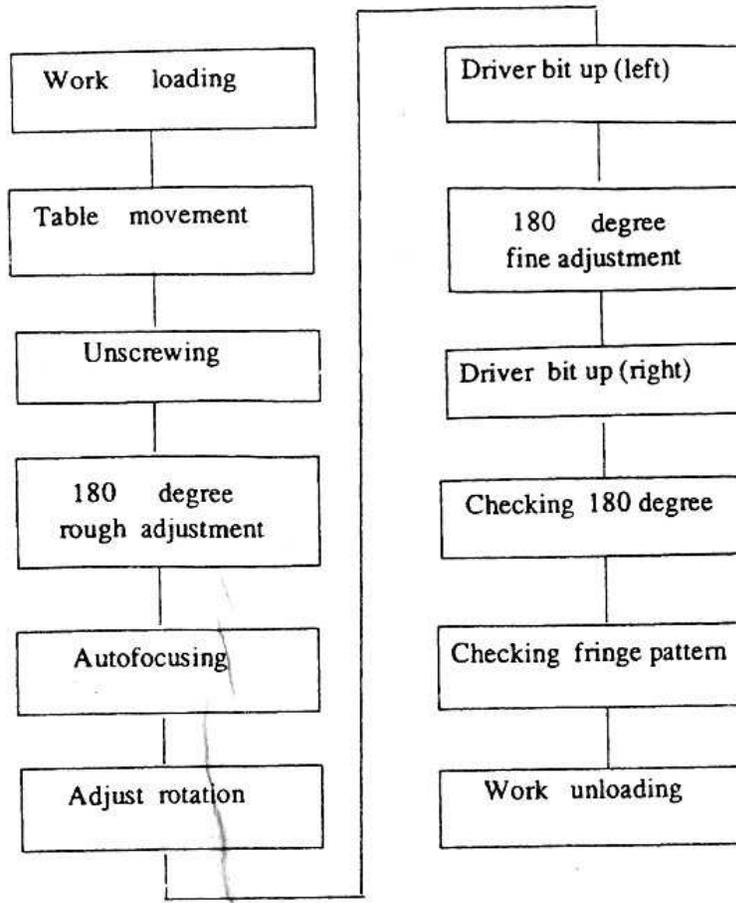


Fig. 4.9.11 Adjustment sequence.

Table 4.9.2 Results of VCR head adjustment

	Protrusion distance (/xm)	Rotational error (arc min)	Setting angle error (arc s)
Specification	± 3.0	± 14.0	± 25.0
Average	0.12	0.68	-0.87
σ	0.81	2.69	5.53

4.9.4 Adjustment results

Adjustment results are shown in Table 7.7.2. The adjustment speed is 30 s per unit. One machine corresponds to three skilled operators in overall productivity. Adjustment accuracies and deviations for the items are also better than achieved by manual adjustment.

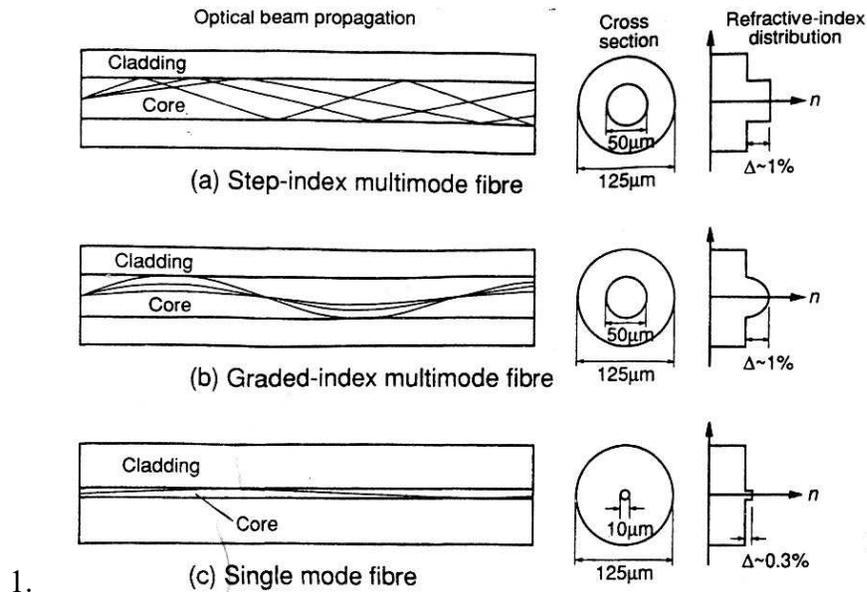
Reference

1. Komatsu, T., Nagashima, S., Tsukada, H.; (1989). An automatic adjustment system for VCR magnetic heads on cylinder units. *Annals of the CIRP*, 38, 9-12.

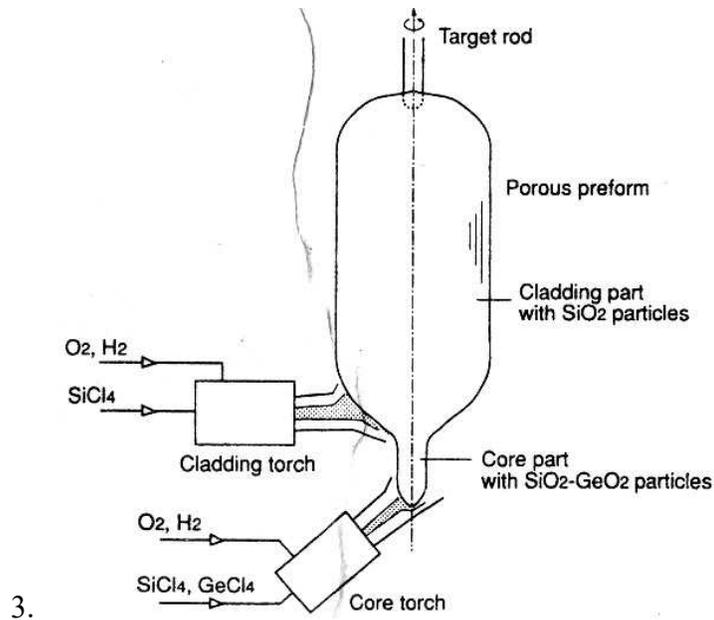
4.10 Optical fibres and related optical components

Optical communication systems are already in commercial use in long-haul transmission lines where the low-loss and high-bandwidth characteristics of silica-based optical fibres can be effectively utilized, and are now ready to penetrate into subscriber systems for realizing 'fibre-to-the-home' (FTTH) networks, connecting individual homes with optical fibres⁽¹⁾. Optical fibres most commonly used in such systems are silica-based single-mode optical fibres which have an outer diameter of 125 μm and a core diameter as small as 10 μm . This section describes optical fibres and related optical components from the viewpoint of microfabrication or nanofabrication technologies.

4.10.1 Optical fibre fabrication



2. Fig. 4.10.1. Optical fibre structures.



3. Fig. 4.10.2. Optical fibre preform fabrication by VAD.

Optical fibres can be classified into (1) step-index multimode fibres, (2) graded-index multimode fibres, and (3) single-mode fibres, as shown in Fig. 7.9.1. The fibre diameter is typically 125 μm . The relative refractive-index difference Δ between core and cladding in Fig. 7.9.1 is typically 0.3-1%, depending on the fibre structure. Step-index and graded-index multimode fibres, which have larger core diameters, several tens of micrometres, find use in short-haul transmission systems. Single-mode fibres with smaller core diameters, $\sim 10 \mu\text{m}$, are the most widely used both in long-haul systems and in subscriber systems, because of their low transmission losses and high transmission bandwidths. In comparison with multimode fibres, much higher precision is required in handling single-mode fibres and related optical components.

The first step in the fabrication of silica-based optical fibres is to compose a fibre ‘preform’ which has a similar cross-sectional geometry to the fibre product but with much larger outer diameter of 20-50 mm and a shorter length of 30-100 cm. Three popular methods of preform fabrication are (1) modified vapour-phase deposition (MOCVD), (2) outside vapour-phase deposition (OVPD), and (3) vapour-phase axial deposition (VAD)⁽²⁾. A VAD set-up is shown in Fig. 7.9.2. Fine glass particles, synthesized by flame hydrolysis of SiCl_4 and GeCl_4 in an oxy-hydrogen torch, are deposited in an axial direction on the end of a rotating fused silica target rod. The porous glass preform consisting of $\text{SiO}_2\text{-GeO}_2$ particle core and SiO_2 particle cladding is

then heated to $\sim 1450^\circ\text{C}$ in an electric furnace for consolidation. The relative refractive-index difference A between core and cladding is adjusted to the desired value by controlling the GeCl_4 flow rate in the core torch during deposition. The fibre preform thus fabricated is heated to higher temperature ($\sim 2000^\circ\text{C}$) in a carbon furnace and drawn into long optical fibres as illustrated in Fig. 7.9.3. During fibre-drawing, the outer diameter of the fibre is precisely controlled to $125 \pm 0.5 \mu\text{m}$ by using a non-contact monitoring apparatus with a scanning He-Ne laser beam. Simultaneously with drawing, the fibre is coated with polymer resin for surface protection and easy handling. An optical fibre, several tens of kilometres long can be drawn from a single preform. It should be stressed that the excellent transmission characteristics of silica-based optical fibres are partly due to the vapour-phase deposition process by which ultra-high-purity glass preforms can be synthesized, and partly to the fibre drawing process by which an extremely smooth boundary (probably sub-nanometre roughness) between core and cladding can be realized.

4.10.2 Fibre-to-fibre interconnection

One of the most crucial problems in optical fibre systems is how to interconnect fibres with high coupling efficiency and low cost⁽³⁾. To attain low coupling losses, the optical axes of the two fibres have to be aligned with submicrometre accuracy. This is in sharp contrast to metal cables, where mutual coupling can be readily attained by rough contact of two metal conductors. Figure 7.9.4 shows the coupling loss characteristics of single-mode optical fibres, calculated from the overlap integral of the optical field distribution of a single-mode optical fibre. It can be seen in Fig. 7.9.4 that to attain a coupling loss $< 0.1 \text{ dB}$, for example, the fibre axis displacement ΔX and angle tilt θ must be $< 0.5 \mu\text{m}$ and 0.5° respectively. Practical fibre-to-fibre interconnection takes the form of splices (i.e. permanent or quasi-permanent interconnection) or connectors (i.e. matable/dematable interconnection), as follows

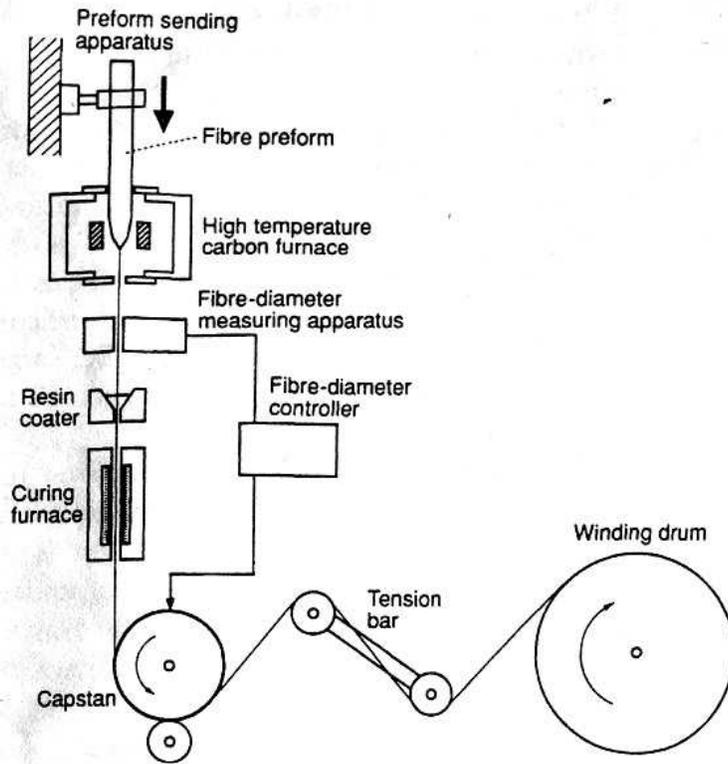


Fig. 4.10.3. Optical fibre drawing process.

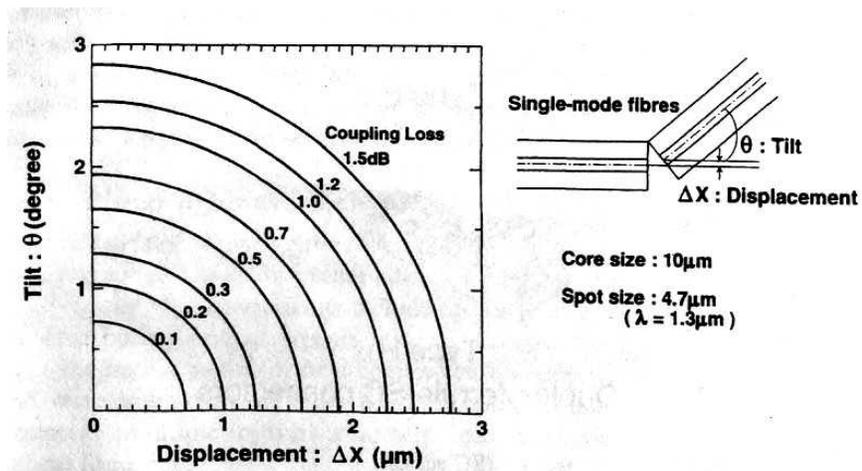


Fig. 4.10.4. Coupling loss characteristics of single-mode fibres.

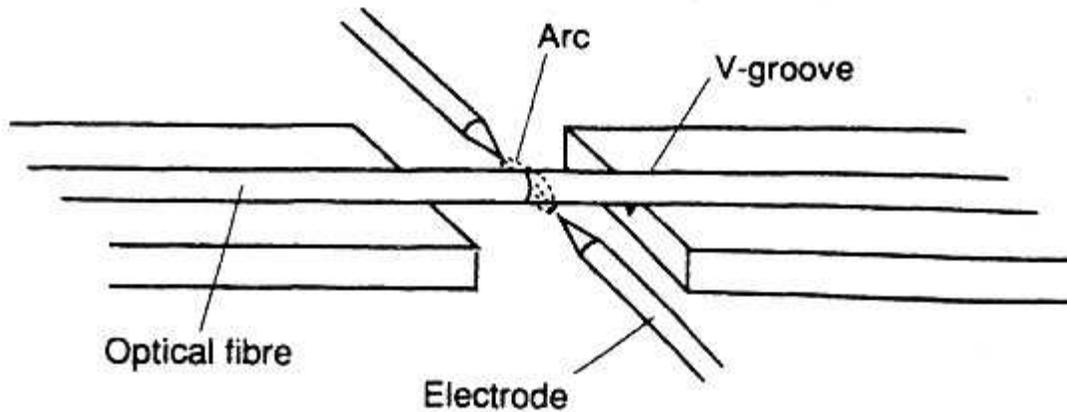


Fig. 4.10.5. Optical fibre splicing by arc fusion.

Arc fusion splices (Fig.4.10.5) have the lowest coupling losses, typically averaging < 0.06 dB for single-mode fibres. Back-reflection from fusion splices can be as low as -60 to -70 dB. Environmental stability is excellent (< 0.02 dB), since there is no potential for movement between the fused fibres. Several types of optical fibre connectors have been developed for use with single-mode fibres. Figure 4.10.6 shows the arrangement and alignment mechanism of the SC connector series, which features (1) low loss (~ 0.1 dB) and high return loss (-40 dB) through the use of zirconia ceramic ferrule and physical contact techniques, and (2) compactness and durability through the use of a simple push-pull mechanism⁽⁴⁾. At present, optical connectors are manually assembled. It is necessary to simplify and automate the assembly processes in order to reduce the assembly cost, which currently accounts for half the total cost of the connector.

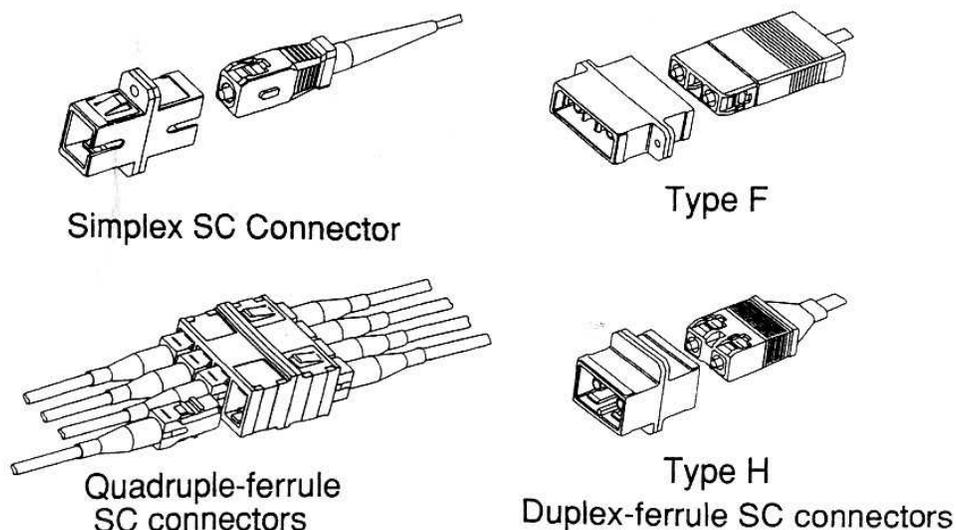


Fig. 4.10.6. Optical fibre connectors (SC series)