Taras Shevchenko National University of Kyiv

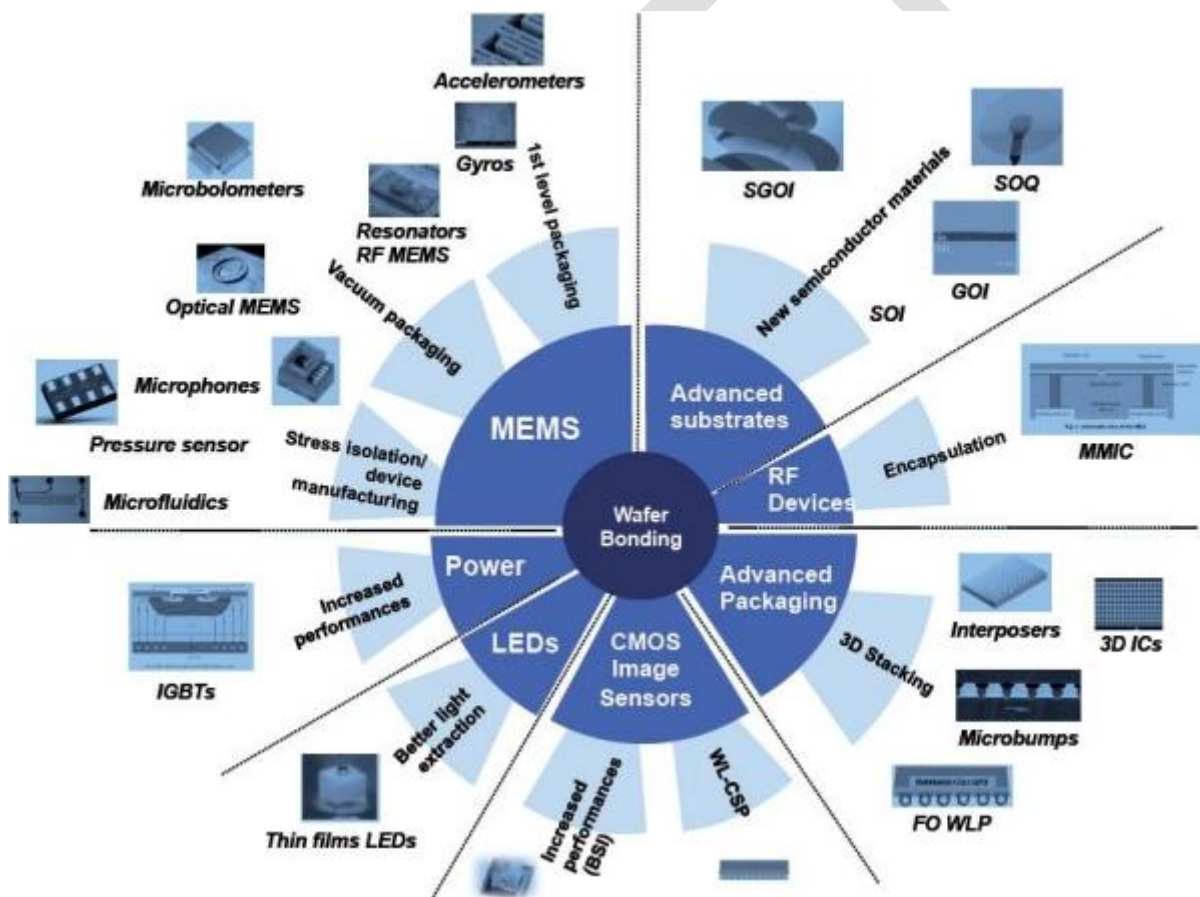**Скришевський Валерій, Іванов Іван**

Application of Microsystems devices

Tutorial

№530785-TEMPUS-1-2012-1-PL-TEMPUS-JPCR

Навчальний посібник " **Application of Microsystems devices** " створено для допомоги вищим навчальним закладам України впровадити нову магістерську навчальну програму "Проектування мікросистем".

Посібник " **Application of Microsystems devices** " створено при підтримці Європейського Союзу за Спільним Європейським Проектом "Curricula Development for New Specialization: Master of Engineering in Microsystems Design" (MastMST), ідентифікаційний номер 530785-TEMPUS-1-2012-1-PL-TEMPUS-JPCR.

Textbook "Application of Microsystems devices" developed to help higher education institutions in Ukraine to introduce new master's educational program "Designing microsystems".

Textbook "Application of Microsystems devices" was created with the support of the European Union within the Joint European Project "Curricula Development for New Specialization: Master of Engineering in Microsystems Design" (MastMST), identification number 530785-TEMPUS-1-2012-1-PL-TEMPUS-JPCR.

The authors express their deep gratitude to the aforementioned universities for full support of the project.

Taras Shevchenko National University of Kyiv

Ivan Ivanov, Valeriy Skryshevsky

Application of Microsystems devices

Tutorial

PREFACE

The work was done under the project Curriculum Development for New Specialization: Master of Engineering in Microsystems Design / MastMST, identification number 530785-TEMPUS-1-2012-1-PL-TEMPUS-JPCR. EC funding programme: Interregional program of the European Neighbourhood and Partnership Instrument.

Target group: students, alumni, lectures and administration of universities, heads of industrial enterprises, the Ministry of Education and Science of Ukraine.

The main objective of the project is to create conditions in Ukrainian technical and classical universities for through 3-level training in the specialty "Design and Engineering microsystems" according to the regional labor market needs.

In the tutorial using of MEMS is considred in everyday life, industry, transport, airspace industry, medicine, in the field of environmental control, energy generation and harvesting. Readers will familiarize oneself with MEMS structures, principles of work, MEMS parameters, MEMS application engineering scenarios.

The proposed tutorial will be useful for students, post-graduate student and researchers who specialize in design and modeling of microelectromechanical systems, applied physics and micro- and nano-electronics.

2.1 Introduction

MEMS (Micro-electromechanical systems) implemented in automotive electronics began in the early 1980s with pressure sensors for engine management systems. In the early 1990s, this

technology was introduced to safety systems with the implementation of accelerometers (Fleming, 2001, 2008). Over the decades, MEMS technology has found many applications within the motor car, enabling innovative electronic systems and the reduction of production and assembly costs of safety systems.

Passenger safety is a high priority for the vehicle manufacturer, often making it a selling feature for the vehicle. Many advances have been made in a number of technologies to more effectively implement protection devices to ensure the safety of occupants in a crash event. The World Heath Organisation reported in 2007 that Western Europe, Canada, Australia and Japan have some of the lowest death rates per annum caused by vehicle crash events. These are between 4.8 per 100 000 persons for the Netherlands (UK 5.4) to 10 per 100 000 persons for Iceland. The USA has a death rate of 13.8 per 100 000 persons compared with Eastern Europe, for example with Bulgaria at 13.2 per 100 000 and Poland of 14.7 per 100 000 persons.

Some of the worst death rates are in North Africa, with around 40 per 100 000 persons per annum. Causes of such crash events are made up of a mixture of road and vehicle conditions, and driver behavior. Road fatalities in the UK started to decrease in the early 1970s, from 7500 deaths per annum, with the introduction of speed limits, annual vehicle testing (the MOT) and other measures, to less than 6000 death per annum in the early 1980s.

Further reductions have occurred during the introduction of the compulsory use of seat belts and the development of vehicle safety systems to less than 2500 deaths per annum in recent years. The UK Department of Transport reports that 1850 persons were killed on the roads in 2010. This reduction is mirrored in other countries with the US Department of Transportation reporting that 32 788 persons were killed in 2010, which is a death rate of 10.6 per 100 000 persons. The reduction in deaths cannot simply be explained in terms of one or two improvements in road conditions or vehicle performance, but consists of a complex mixture of many factors.

An important factor is the development of vehicle safety systems and along with this are the developments in MEMS sensors. Frontal crash, side impact and rollover are the main events that a safety system has been developed to deploy protection devices. MEMS sensing devices are used in such systems to detect the condition of the vehicle in terms of undesirable accelerations and angular rate changes, as well as the detection of the position of occupants. Research and development effort continues worldwide on creating novel safety systems by monitoring passenger position and behavior, vehicle behavior and environmental conditions.

Passenger protection systems

Electronically controlled passenger protection systems are commonplace in most vehicles to some degree. These consist of various airbags, seat belt pretensioners and rollover protection devices, all controlled by an electronic control unit(s) (ECU). These systems are typically integrated into one centrally controlled system made up of the following component parts:

Sensors: – for indicating an event, which may require action.

The ECU:

– for signal processing to identify the critical event that requires action;

– for supplying the appropriate signals to deploy the protection device;

– for system diagnostics to ensure robust behavior.

The protection device:

– for delivering the appropriate protection.

Figure 2.1 illustrates this system consisting of a sensing part, a control unit and an actuation part. Three types of protection devices are shown in this example, for airbag deployment, pretensioner firing and rollover protection. Also three types of sensing devices are shown, each of which is the focus of this chapter concerning crash sensing, vehicle rollover and occupant detection.
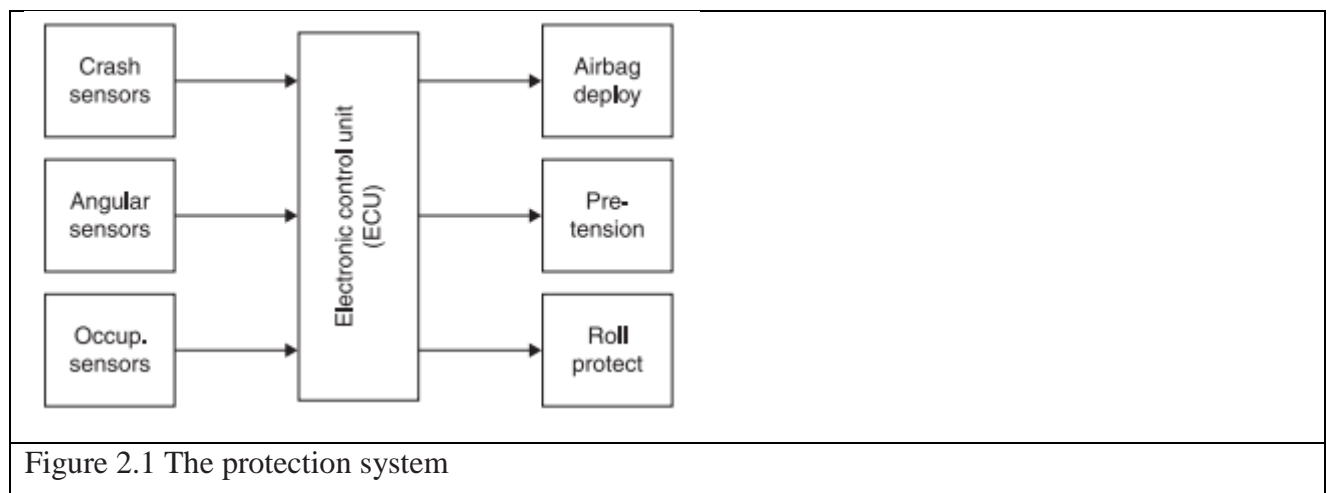
Figure 2.1 The protection system

The crash event

A vehicle in a crash event is usually undergoing a rapid change in velocity in some direction caused by contact with an external obstacle. In a frontal collision of the moving vehicle, there is a rapid deceleration of the vehicle, whereas an unrestrained occupant continues to move forward until abruptly stopped by contact with a rigid part of the interior of the vehicle (such as a steering wheel or windscreen). Any protection system employed looks to ensure smooth deceleration of the occupant to minimize injury.

Pretensioned seat belts and airbags are designed specifically to achieve this. Pretentioners on seat belts are fired by igniting a pyrotechnic charge in the frontal crash event to maximize the restrain of the occupant with the use of the seat belt. This tightens the seat belt to prevent the occupant from jerking forward, which typically takes between 5 and 10 ms (Jurgen, 1999).

Depending on the specification of the vehicle, there are numerous arrangements of airbags and curtain systems. The frontal airbag is the most standard, which is placed in the steering wheel

or dashboard to restrict frontal motion of the occupant. Pyroelectric igniters activate gas generators to inflate the airbag within around 30 ms (Jurgen, 1999).

Rollover protection such as roll- bars (Wanden and Kinnanen, 2001) and headrest-bars (Hehl and Remm, 1999) are usually spring- loaded and released on command. They maintain a volume of space in the occupant's head area during a vehicular roll to minimize injury. Convertible are the typical vehicles with these devices fitted. However, non- convertible vehicles are also considered to require rollover protection devices (Bozzini et al., 2010).

Depending on the crash event (see next section on frontal, oblique, offset, pole, under- ride and side crash events), the firing of the different protection systems is crucial to allow the displacement of the occupant to be within certain limits to ensure time to activate the relevant devices.

Classification of crash events

A passenger vehicle has 6 degrees of freedom, 3 translational movements with usually only the longitudinal translation as desirable, and 3 rotational movements, which are usually all undesirable. Figure 2.2 shows these translational movements.

Figure 2.2. Vehicular translations
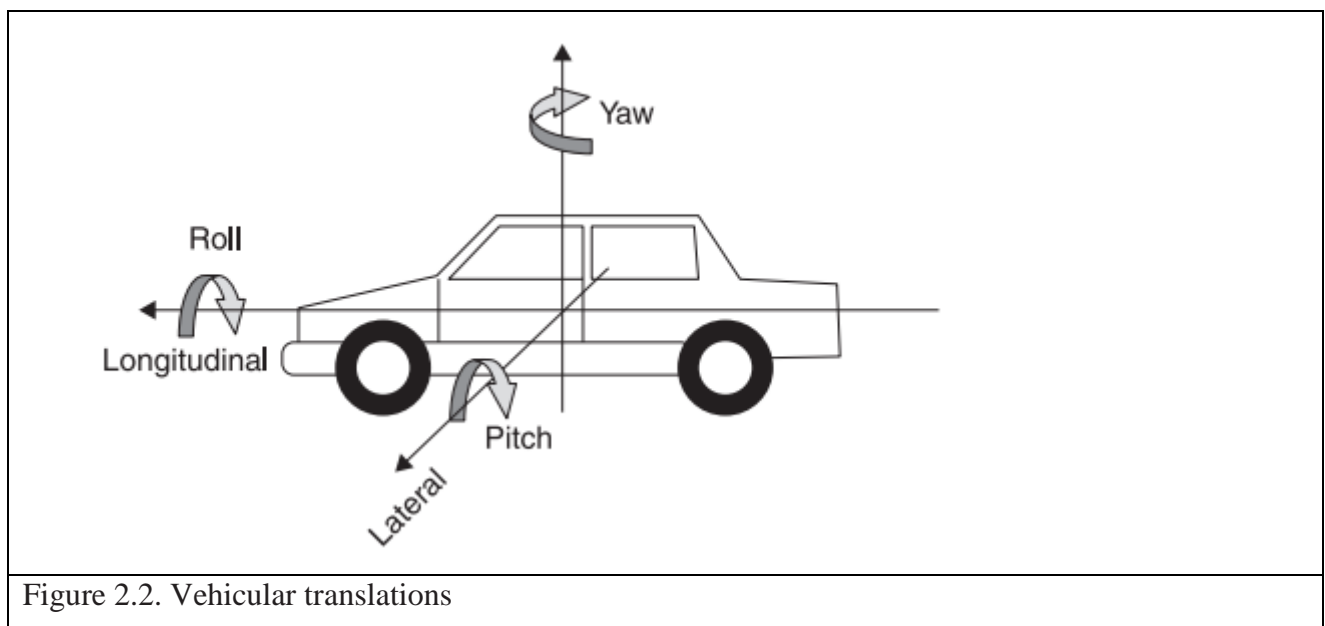
Forward longitudinal crashes have been classified into different types (Fig. 2.3) (Iyoda, 2001). The frontal crash is where the impact has equal force along the front of the vehicle, whereas the oblique crash exhibits a gradient of force level from one side to the other. The oblique crash may cause a yaw rotation. Pole and offset crashes concentrate an impact force at particular areas at the front of the vehicle. The under- ride crash is where forces are concentrated at the upper parts of the vehicle's front, whereas the interference crash affects the underside of the vehicle
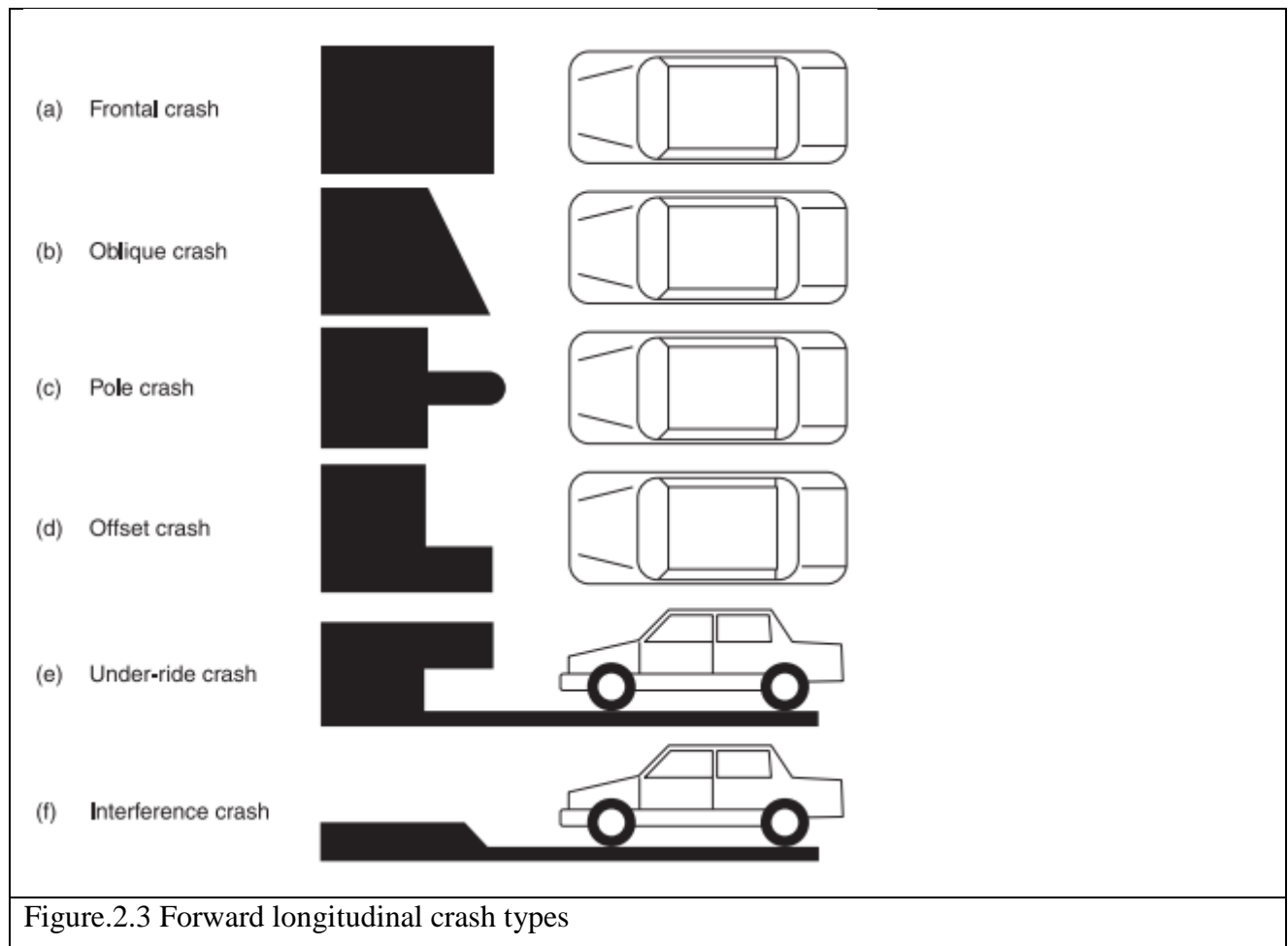
Figure.2.3 Forward longitudinal crash types

Classification of crash events

Considered by most engineers as cutting-edge or fringe technology, MEMS sensors have been embraced by the automotive industry in its quest to improve performance, reduce cost, and enhance the reliability of the family sedan. In fact, hundreds of millions of MEMS sensors have been used in automobiles over the past decade.

Many of these sensors (e.g., MEMS pressure sensors) simply replace older technologies with cheaper, more reliable devices. In contrast, MEMS inertial sensors have enabled many desirable features that are increasingly common in cars today. In this article, I'll examine automotive inertial MEMS sensor applications, describe how they work, and discuss what it is about MEMS sensors that enable or greatly improve the application. Many of the applications will be well known to you because they've become ubiquitous in automobiles. Some applications (often the most interesting) are just appearing in high-end models but are destined to become standard equipment.

Crash Sensing for Airbag Control

Crash sensing for air bag control represents the largest automotive use of inertial MEMS sensors. In this application, an accelerometer continuously measures the acceleration of the car. When this parameter goes beyond a predetermined threshold, a microcontroller computes the integral of the acceleration (i.e., the area under the curve) to determine if a large net change in

velocity has occurred. If it has, the air bag is fired. The decision to fire front air bags has to be made in dozens of millisecond?; the decision to fire side air bags must be made even more quickly because the car door is closer to the occupant than the steering wheel or dashboard.

About 15 or 20 years ago, when air bags first appeared in automobiles, air bag control module makers relied on g switches (an inertial switch made up of a contact, ball, and spring housed in a cylindrical enclosure) distributed throughout the automobile. These switches don't give a lot of information about the nature of the acceleration that is being sensed. They simply provide an on/off signal telling you that the acceleration is above or below a threshold. As a result, a simple center console air bag control module requires several switches (generally three to seven) to decide if the acceleration is the result of road roughness or a crash. Worse yet—because of the reliability and long life required of their contacts—g switches are costly. Wiring them to several locations throughout the car increased their cost and reduced their reliability.

The introduction of MEMS accelerometers to air bag control modules virtually eliminated the use of g switches as the primary acceleration sensor in air bag modules. Because the MEMS accelerometer reads a continuous (analog) measurement, you can replace the g switches with one MEMS device in the center console. The resulting increase in reliability (e.g., Analog Devices' highly integrated accelerometers achieve single-digit ppm defect rates) and reduction in price of the air bag system helped bring about its near universal inclusion in cars. Better still, MEMS accelerometers can perform robust self-testing, allowing the air bag module processor to determine if the sensor's data are reliable or if the air bag module must be serviced.

MEMS accelerometers commonly control side air bags. Because the fire decision must be made quickly, there is no time to wait for the propagation of the sensor's signal through the car's chassis, so the satellite must be placed close to the air bag it conýrols. In addition, because there is virtually no crush zone between the impact and the accelerometer, the measurement range must be above the center console accelerometers. As a result, many vehicles outfitted with side air bags may add two to four more MEMS accelerometers for this function.

Front-looking crash sensors placed just behind the front bumper are being added to some models to help determine the severity of the frontal crash. The acceleration signature of the front-looking sensor is compared with that of the center console accelerometer, allowing the air bag module controller to modulate the inflation rate of the air bag to match the deceleration rate of the car. Here, too, high g range and compact size are important factors in this application.

Vehicle Dynamic Control

Vehicle dynamic control (VDC) systems help the driver regain control of the automobile when it starts to skid. If the VDC works properly, the driver may not even be aware that the system intervened.

A VDC system consists of a gyroscope, a low-g accelerometer, and wheel-speed sensors at each wheel (the wheel-speed sensors may also be used by the ABS). Wheel speed is measured, and the predicted yaw (or turn) rate of the car is compared with that measýred by the gyroscope. A low-g accelerometer is also used to determine if the car is sliding laterally. If the measured yaw rate differs from the computed yaw rate, or if lateral sliding is detected, single-wheel braking or torque reduction can be used to make the car get back in line.

Before the advent of MEMS gyroscopes and accelerometers, VDC for ordinary passenger cars was impractical. Conventional gyros and accelerometers would add thousands of dollars to the cost of the car. Indeed, conventional gyros, built with spinning masses and strain gauges, probably wouldn't be rugged enough to meet the >10-year operational requirement of the automotive market. Even the MEMS gyros are barely up to the task.

A typical MEMS gyro uses a quartz tuning fork. The vibration of the tuning fork, along with applied angular rotation (yaw rate of the car), creates Coriolis acceleration on the tuning fork. An accelerometer or strain gauge attached to the tuning fork measures the minute Coriolis force. Signal output is proportional to tuning-fork size. To generate a strong enough output signal, the tuning fork must vibrate forcefully. You can best accomplish this with a high Q structure. Manufacturers often place the tuning fork in a vacuum to minimize mechanical damping by air around the tuning fork. High Q structures can be fairly fragile.

Because the gyro must be rigidly connected to the car to accurately measure yaw rate, the gyro often experiences shock and vibration. This mechanical noise can introduce ý signal to the Coriolis pick-off accelerometer that is several orders of magnitude higher than the tuning fork?generated Coriolis signal. Separating the signal from the noise isn't easy. Often, the shock or vibration saturates the circuitry and makes the gyro output unreliable for a short time (this explains why your VDC warning light may occasionally come on for no apparent reason).

Conventional MEMS gyros are usually bulky (100 cm^3 or more is not uncommon). This is partly the result of the addition of mechanical antivibration mounts, which are incorporated to minimize sensitivity to external vibration.

New MEMS devices avoid these shortcomings, though. For example, Analog Devices' iMEMS gyro (which is in development) is 7 by 7 by 3 mm (0.15 cm^3). Rather than quartz, it uses a resonating polysilicon beam structure, which creates the velocity element that produces the Coriolis force when angular rate is presented to it. At the outer edges of the polysilicon beam, orthogonal to the resonating motion, a capacitive accelerometer measures the Coriolis force. The gyro has two sets of beams in antiphase that are placed next to each other, and their outputs are read differentially, attenuating external vibration sensitivity.

Rollover Detection

Few vehicles have rollover detection systems, but automakers are rapidly adopting this feature. This is particularly true for vans, pickup trucks, and sport utility vehicles, which are more likely to roll over because of their higher center of gravity. These systems read the roll angle and roll rate of the vehicle to determine if it is tipping over. If it is, the system fires the side curtain air bags to protect the occupants.

Rollover detection systems use a gyroscope to read the roll rate. The roll rate is integrated to determine the roll angle of the vehicle, but roll rate data alone are not enough to predict if a vehicle is (or will be) rolling over. An accelerometer reading vertical acceleration (Z axis) is also required because large roll angles can be encountered in banked curves with no possibility of rollover.

Many rollover detection systems use a second accelerometer to measure lateral acceleration (Y axis). If a car is sliding sideways, it's less likely to roll over if unobstructed. But if it hits a curb or another object, the chances of a rollover increase significantly. The side crash detection accelerometer generally can't perform this task because the magnitude of acceleration when sliding sideways is close to the noise floor of the typical >100 g range used for side impact detection. A low g range, dual-axis accelerometer is best suited to reading the Y and Z axes' acceleration.

Gyros used for rollover sensing don't require the same resolution as those used in VDC systems because the roll rates are almost half an order of magnitude greater, but they must have excellent rejection of external shock and vibration. It's not unusual for a car to roll over immediately after striking another vehicle or stationary object. A gyro whose output is unreliable for a short time after a shock event is next to useless.

Antitheft Systems

One of the more popular ways to steal a car is to simply tow it away. In response to this threat, many automakers (particularly European manufacturers) are including antitheft systems incorporating tilt detection systems. Pan European specifications require accepted tilt detection systems to be capable of measuring a change in inclination of 3° over a 3 min. period (i.e., a rate of 0.016° of tilt per second).

Large electrolytic fluid tilt sensors have the required sensitivity for this application, but they fall short in many other areas. Their tilt range is limited, so it's possible to park a car at an inclination that is greater than the sensor can tolerate. In addition, fluid tilt sensors don't perform well in the widely varying automotive temperature environment. On the other hand, low g MEMS sensors are ideal in this application.

Electronic Parking Brake Systems

When you activate your parking brake, you normally pull it (or push, if pedal actuated) fairly robustly without concern for how much braking force is required. As a result, conventional parking brakes must be substantially overbuilt.

Pushing a button activates electronic parking brake systems. The system measures the inclination of the vehicle, determines how much braking force is required, and applies it. This relieves the manufacturer of the burden of having to grossly overbuild the braking system and allows such features as the automatic parking brake (once the system determines that the car is in park and at rest, it applies the parking brake).

Once again, a low-g accelerometer is used. The performance requirements are similar to those of the antitheft system. In some cases, a single sensor can provide inclination data for both systems.

Vehicle Navigation Systems

Vehicle navigation systems are rapidly becoming a standard feature in American luxury automobiles. In Japan, more than half the cars sold in 2001 were equipped with navigation systems.

A global positioning system (GPS) is a fundamental part of a navigation system, but GPS information alone is insufficient for navigation. The GPS can tell you where you are (position and altitude), but not what direction you are facing. Magnetometers (electronic compasses) are not reliable because they're confused by large ferrous metal objects close by (e.g., a truck full of scrap metal in the next lane).

Navigation systems rely on compass and GPS information when the system is first started. The direction of travel is matched up with map data to give the system more certainty regarding direction. Once initial direction is established, gyroscope information is used to determine when and how much the car has turned, until directional data can be verified by map matching.

In urban settings, it's not unusual for the GPS signal to be obscured by tall buildings or tunnels for short periods. At these times, the navigation system relies on the gyroscope for heading information and a low-g accelerometer for position informatiom. The acceleration signal is integrated twice to derive position (this technique is called dead reckoning).

The Sensor Cluster

Today, many inertial sensors are used in full-featured cars (see Figures 2.4 and 2.5).

■ Crash Detection System	▲ Satellite Sensor
■ Vehicle Dynamic Control System	● Airbag
■ Navigation/Driver Information System	▲ Seatbelt Pretensioner
■ Body/Chassis Control System	
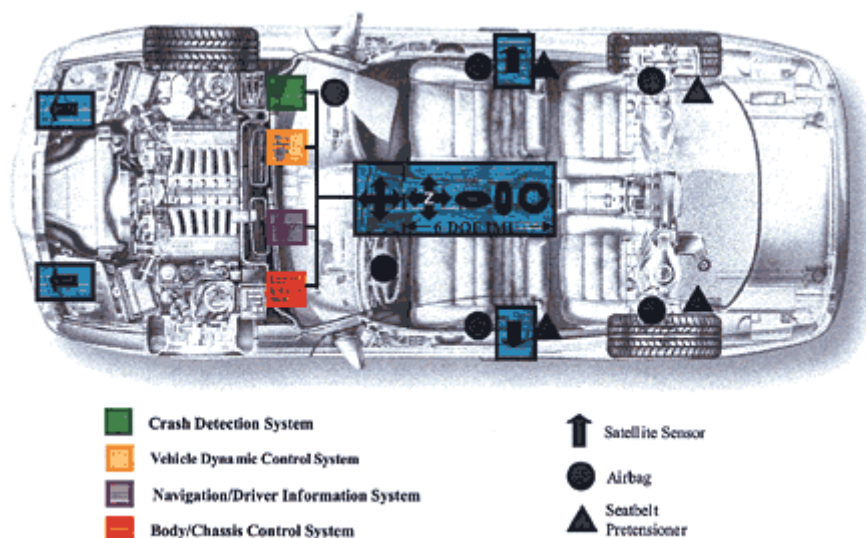
Figure 2.4. By using an architecture that clusters inertial MEMS sensors in the center of the automobile, fewer devices are needed to provide the information required by the various subsystems (e.g., safety and navigation). This approach is becoming the goal of many automotive OEMs.
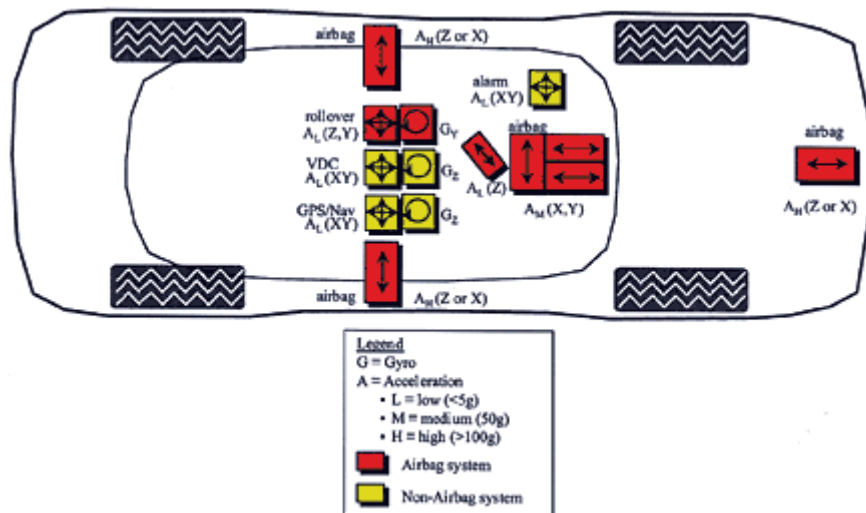
Figure 2.5 The number of inertial MEMS sensors used in cars is substantial and growing. Shown here are the inertial sensors used in full-featured cars today. In some cases, there are as many as 15 axes of inertial sensors (accelerometer and gyro).

In some cases, there are as many as 15 axes of inertial sensors (accelerometer and gyro) per vehicle. But why don't manufacturers use each sensor for multiple functions? The main reason is that to date no one has had the expertise or interest to integrate all the functions in a single system. Manufacturers often deem inertial sensor signals of safety systems off limits to external functions for fear they'll lose the crash sensor signal because another subsystem (e.g., the navigation system) takes the bus down.

Nevertheless, many automotive OEMs are adopting the concept of using a cluster of inertial sensors to send information to whatever system needs it (see Figure 2.4). In this configuration, a six-degree-of-freedom inertial measurement unit (IMU) is located in the center of the automobile. All the inertial accessory systems (e.g., antitheft, VDC, electronic parking brake, and navigation systems) use the IMU signals, and the unit can also pass information to the air bag control unit. Separate stand-alone accelerometers are still placed at locations around the car for crash sensing. This is necessitated by the proximity-to-crash-zone demand of some of the applications.

A side benefit of the inertial sensor cluster concept is that new features can be developed at little additional cost because the inertial information is available for free. All the designer has to do is add some intelligence.

The rollover protection system requires that the sensor detects a less aggressive event than compared to frontal or side crash events. This has been estimated at between 3 and 10 g in terms of acceleration (Jurgen, 1999). Strictly speaking, the angular rate sensor detects angular velocity of between 1.5°/sec (noise level) to 300°/sec.

Technical details of typical devices The Analog Devices ADXRS612 is an angular rate sensor, which uses the principle of oscillating structures and the Coriolis effect. Two seismic masses form two 'dither frames', which are caused to oscillate asymmetrically at 14 kHz in a linear fashion. An electrostatic means is used to drive the oscillation, which requires between 18 and 20 V to achieve the required amplitude. A charge pump circuit is included if only lower voltages are available to supply the device. Afinger structure is fitted to each of the outer parts of each dither frame to enable the sensing of movement orthogonal to the oscillations. This detects the Coriolis movement. The fingers are positioned between fixed fingers to take a capacitive measurement similar to that employed in the ADXL50 accelerator. The signal is demodulated to give an output voltage of 7 mV per angular rate °/sec (Analog Devices, 2007).

Figure 2.6 shows the Bosch SMG061 surface mount angular rate sensor. The cut- open package shows two chips, one being the MEMS sensor and the second being the conditioning electronics and interface. Again, the principle of oscillating structures and the Coriolis effect is utilized to sense the angular rate. A single seismic mass is caused to oscillate at 15 kHz by electrostatic means in a rotational fashion. Any movement orthogonal to the plane of the oscillation (due to angular rate movement) causes a rocking motion, which is out of plane of the oscillating seismic mass. This rocking motion is detected with an electrode under the seismic mass by capacitive means. The signal generated is demodulated to give an output voltage of 7 mV per angular rate °/sec.
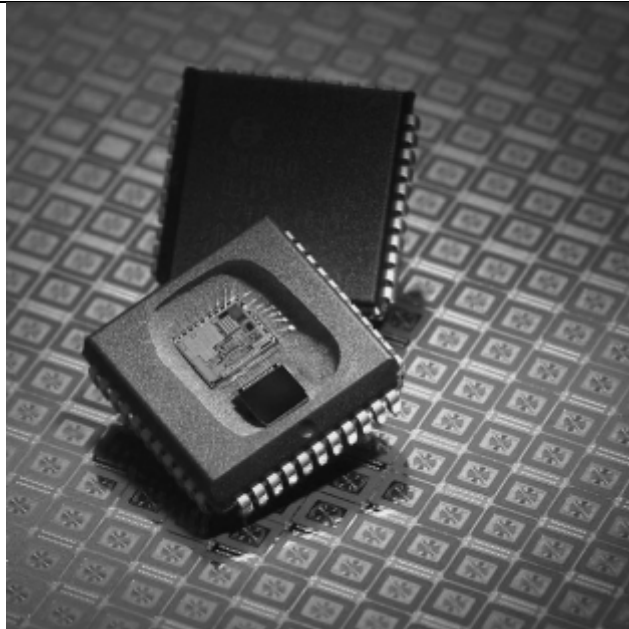
Figure 2.6. Example of sensor

Comments on manufacturing processes

Classen et al. (2007) give a comprehensive description of the developments in the fabrications processes used by Bosch for the angular rate MEMS sensor. These build on the bulk and surface micromachining techniques employed in the accelerometer devices. The silicon wafer is deposited with various combinations of oxide layers and polysilicon layers with an aluminium conductive layer. The oscillating masses are constructed from these layers by the anisotropic deep reactive ion etching method, which has come to be known as the Bosch Process. The formed MEMS structural components are freed by removing the oxide layers using hydrogen fluoride vapor phase etching.

Integration as a system

Rollover protection technology is often integrated with the airbag crash system. Rollover events are not the most frequent type of crash; however, they do incur the most serious injuries. The most common rollover events are due to a soil trip, fall over and curb trip (Viano and Parenteau, 2004).

This system is employed on non- convertible vehicles as well as convertibles for the benefits it brings. Once a rollover event is detected, several protection devices can be deployed. Inflatable side curtains are the most common in nonconvertible vehicles, for protecting against head and upper body injuries.

The rollover protection system requires that the sensor detects a less aggressive event than compared to frontal or side crash events. This has been estimated at between 3 and 10 g in terms of acceleration (Jurgen, 1999). Strictly speaking, the angular rate sensor detects angular velocity of between 1.5°/sec (noise level) to 300°/sec.

Technical details of typical devices The Analog Devices ADXRS612 is an angular rate sensor, which uses the principle of oscillating structures and the Coriolis effect. Two seismic masses form two 'dither frames', which are caused to oscillate asymmetrically at 14 kHz in a linear fashion. An electrostatic means is used to drive the oscillation, which requires between 18 and 20 V to achieve the required amplitude. A charge pump circuit is included if only lower voltages are available to supply the device. Afinger structure is fitted to each of the outer parts of each dither frame to enable the sensing of movement orthogonal to the oscillations. This detects the Coriolis movement. The fingers are positioned between fixed fingers to take a capacitive measurement similar to that employed in the ADXL50 accelerator. The signal is demodulated to give an output voltage of 7 mV per angular rate °/sec (Analog Devices, 2007).

Strain gauges for occupant sensing systems

Background to occupant sensing systems

Occupant classification and position detection have been significant research areas in intelligent safety systems in the automotive field (Hannan et al., 2006). The first occupant detection systems were concerned with identifying whether a seat was occupied.

Occupant sensing has become of critical importance where airbag and inflatable curtains are concerned, as many child and small adults have been injured and killed due to their deployment. An airbag can inflate at over 200 mph. In the USA since 1990, nearly 300 children and small adults have been killed as the result of airbag deployment (source: NHTSA, The National Highway Traffic Safety Administration).

Figure 2.7 gives a typical example of an occupant detection and classification system, which is fitted into a seat. This is Delphi's 'Passive Occupant Detection System – B', which was launched in the early 2000s. The great advantage of such a system is that it does not require action by the driver to suppress the airbag system.

The occupancy of the seat is classified by the system and judges whether the airbag system should be suppressed due to the seat being empty, the occupant being an adult or infant/child. The seat is fitted with a fluid-filled bladder connected to a pressure sensor. The ECU processes data from the seat sensor, and if certain criteria are met, the airbag deployment will be allowed when necessary. The advantages of MEMS pressure sensors in these types of systems are the relatively low costs and the small physical size of the units. MEMS give the options for multiple sensor use for enhanced occupant classification.

Figure 2.7 Passive Occupant Detection System
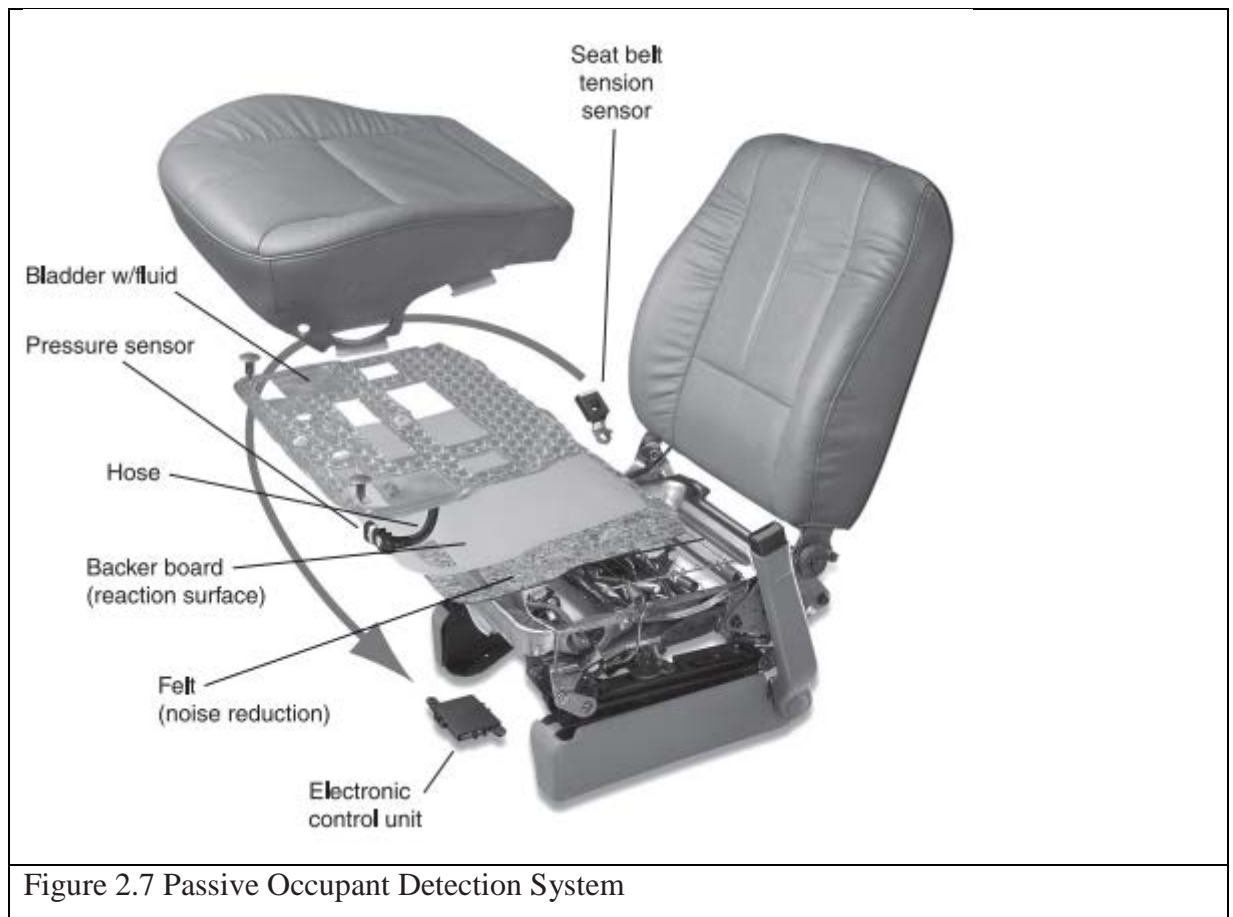
Principles of operation for the MEMS sensor
Force deflector types

Force deflector types of the electronic pressure sensors use a deflection, for example in a diaphragm, to measure the applied force on the deflector. This is illustrated in Fig. 1.13 , where an external pressure supplies a force per the area of the diaphragm to cause the deflection.

Figure 2.8 Force deflection types.
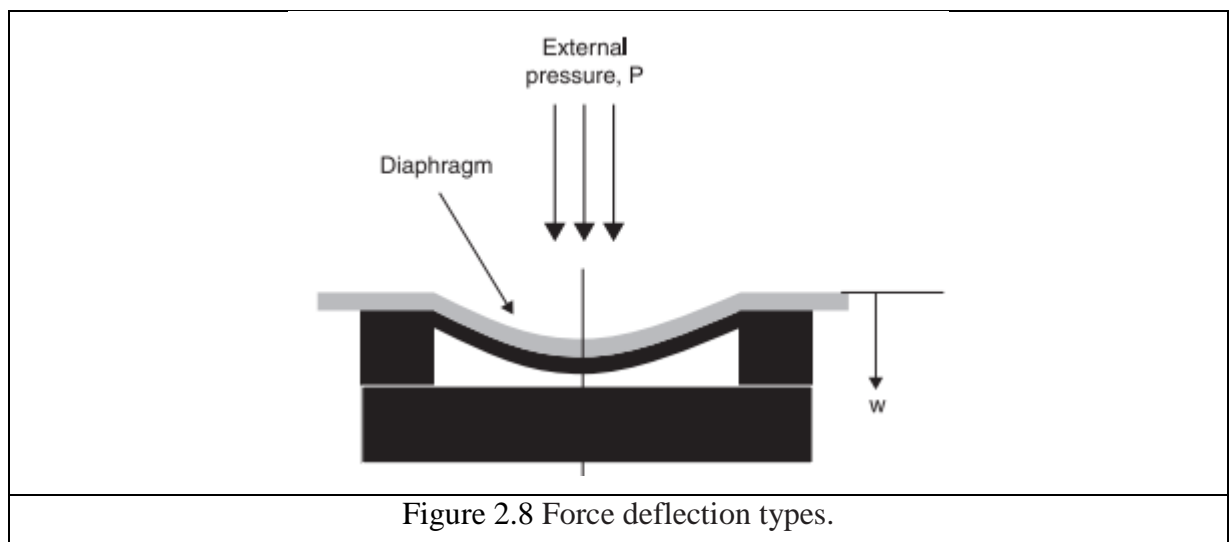
Piezoresistive techniques are one of the most frequently used methods to sense a deflection in a diaphragm structure (Sing et al., 2002; Zaiazmin, 2006). Piezoresistive components are fitted to the diaphragm and electrically connected in a bridge arrangement. Capacitive techniques are also often employed (Hezarijaribi et al., 2008).

Resonant structure types

Resonant structure types use the changes in the resonant frequency of the structure caused by the applied force on the structure to determine that force. This is illustrated in Fig. 1.14 , where an external pressure applies a force across the area of the structure. The structure can be forced to resonate by various means. Welham et al. (1996) report an electrostatic driven comb arrangement with a seismic mass, M. A bias voltage, V_p , is placed on the comb and is oscillated at a resonant angular frequency, ω_0, with an alternating voltage of magnitude, V_d. If a sense voltage, V_s is applied to a sensing comb linked to the driving comb, a current can be detected to measure the movement in the comb

Figure 2.9 Resonant structure type
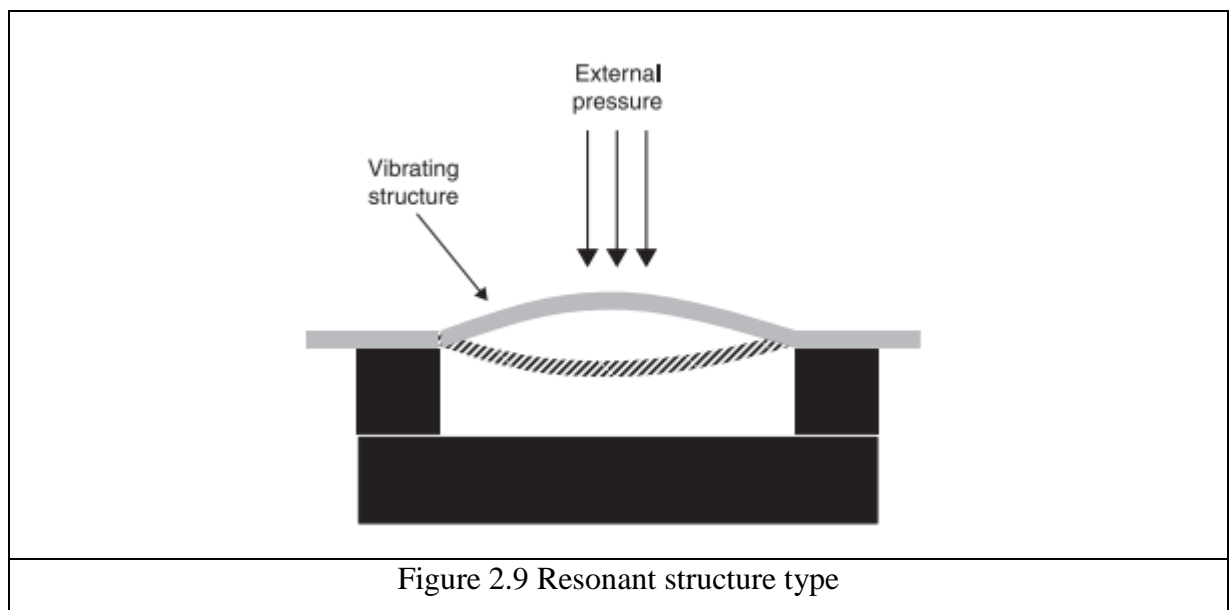
Implementation of the MEMS sensor

The USA have developed airbag standards to ensure airbag systems do 'no harm' to passengers and thus occupant sensors have been developed specifically to enable compliance with such standards (Fleming, 2008).

Technical details of typical devices

The Melexis MLX80807/8 is a device that measures relative or absolute pressure. Piezoresistive material is deposited onto the edges of a diaphragm, which deflects due to pressure differences between the two sides. Four piezoresistive resistors are used at the edge of the diaphragm in a Wheatstone bridge arrangement to modify an electrical signal. Included is a facility to compensate for offset draft due to temperature variations (Melexis, 2011). To ensure that the Wheatstone bridge delivers a change in resistance due to deflection of the diaphragm, typically two of the piezoresistors elements are deposited with positive piezoresistive characteristics and two are deposited with negative piezoresistive characteristics (Adam et al., 2008).

Comments on manufacturing processes

The Melexis MLX80807/8 integrates both the MEMS sensing element and signal processing on a single silicon die. The sensing element consists of a square silicon diaphragm formed by backside etching. Bosch's MEMS pressure sensor fabrication uses surface micromachining of the sensing diaphragm (Adam et al., 2008). A monocrystalline starter layer is first formed and a porous silicon layer is created by using a wet chemical etching process. An epitaxial monocrystalline silicon layer is then applied to the whole wafer. This epitaxial process and subsequent high temperature steps cause silicon atoms from the porous silicon layer to migrate, leaving a void and creating the cavity for the diaphragm.

Integration as a system

Four types of occupant sensor, or occupant behavior sensor, have been the focus of attention by many (Fleming, 2008). The occupant weight sensor measures the passenger's weight to identify whether they are a child or a small adult. If the system identifies that there is no passenger or child seated, then the airbag system is deactivated. If the system identifies a small adult, then the airbag system is programmed to deploy a 'softer' airbag with, for example, less explosive gas injection or other means. A fluidfilled bladder with a pressure sensor is one measuring device that has been patented by Waidner and Fortune (2007). Oestreicher et al. (2006) have patented an idea, which involves strain- gauge sensors placed strategically within the seat.

The seatbelt tension sensor is designed to identify a child seat fitted on the passenger seat. The issue is that the occupant weight sensor can return an adult weight when a child seat is fitted. Therefore, a means needs to be employed to make this differentiation. When a child seat is fitted, the seatbelt tension sensor detects a larger force (than compared to an adult), due to the child seat being tightly belted into place. A typical type of sensor employed uses a magnet and a Hall effect probe to identify displacement of a spring mechanism (Stanley and Takehara, 2007).

The seatbelt buckle sensor identifies whether a seat belt is buckled. If it is not buckled, and the seat is occupied, a higher rate of airbag inflation is used to try to restrain the occupant in a crash event. The type of sensor typically used is a magnet and Hall effect probe arrangement to identify the latched buckle (Almaraz and Martinez, 2006).

The seat position sensor can be used to identify the position of the occupant relative to the airbags. This is particularly relevant for the driver's seat, because if the seat is a long way forward, this may indicate a small adult. Also the driver would then be close to the steering wheel and close to the steering wheel airbag. All this needs to be considered when deploying the protection measures in a crash event. The typical sensor employed is a magnet and Hall effect probe arrangement to identify position (Ventura and Tokarz, 2008).

Vehicle stability control (VSC)

Vehicle stability control (VSC) is a generic term referring to various electronic systems used in modern cars, with the aim of improving the road handling of the vehicle by means of an electronic aid. It can include a number of different systems, such as electronic stability control (ESC), active suspension or active cornering. Considered to be one of the greatest advances in road safety since the invention of the seatbelt, the ESC is an additional improvement to the anti- lock braking system (ABS) and traction control system (TCS), which were introduced in the 1980s. Its basic function is to stabilize the vehicle when it starts to skid, by applying differential braking force to individual wheels and in some cases also by reducing the torque transmitted to the wheel. It keeps the car on course even under extremely difficult conditions, such as icy or wet roads. Additional sensors must be added to the ABS system in order to implement ESC functionality, including a steering wheel angle sensor, a yaw rate sensor and a low g acceleration sensor to measure the dynamic response of the vehicle. Only the last two sensors use MEMS technology and will be part of the discussion in this chapter. This type of system is becoming popular, particularly now that many governments have introduced vehicle safety legislation programs whose focus has shifted from passive safety systems such as airbags to active safety systems, helping to save thousands of lives annually. It is part of the standard equipment of many passenger cars and its penetration rate is greater than 70% in many developed countries. Active damping is also a generic term used to describe semi- or full-active suspension systems that continuously adjust a vehicle's damping levels according to road and driving conditions. It controls the vertical movement of the wheels via an onboard system, with the aim of virtually eliminating variation in the vehicle's roll and pitch, which occurs in many driving situations such as cornering, accelerating and braking. Active damping enables car manufacturers to achieve a higher degree of ride quality and car handling, by keeping the tires in contact with the road, allowing much higher levels of grip and control. A vehicle's chassis movement and/or suspension displacement are detected by a range of sensors located throughout the vehicle. Contactless Hall effect sensors are often used for this purpose, but MEMS inertial sensors are also

used, particularly for detecting the motion of the vehicle body. The data from these sensors are fed into a dedicated control unit, which will act on the suspension, adjusting damping characteristics as necessary. This chapter will present in detail a typical fully- active suspension architecture using MEMS low- g sensors, together with their requirements. Market trends and future developments will also be covered.

Description of an electronic stability control (ESC) system

 The ESC system runs continuously in the background during normal driving conditions, and monitors the behavior of the driver and the vehicle by comparing the driver's intended direction to the vehicle's actual course (Bauer, 2004). Functioning on any road surfaces, it becomes operational only in the case of loss of steering control during critical driving situations, such as an emergency steering maneuver to avoid a crash, understeer or oversteer after misjudgment of a turn, or hydroplaning. By assessing the direction of the skid, preventing the vehicle's heading from changing too quickly (spinning out) or not quickly enough (plowing out), the ESC applies the brakes to individual wheels in an asymmetric fashion. It creates torque around the vehicle's vertical axis, thus counteracting the skid and bringing the vehicle back in line with the driver's intended direction. It may additionally reduce the engine power or operate the transmission to slow the vehicle down.
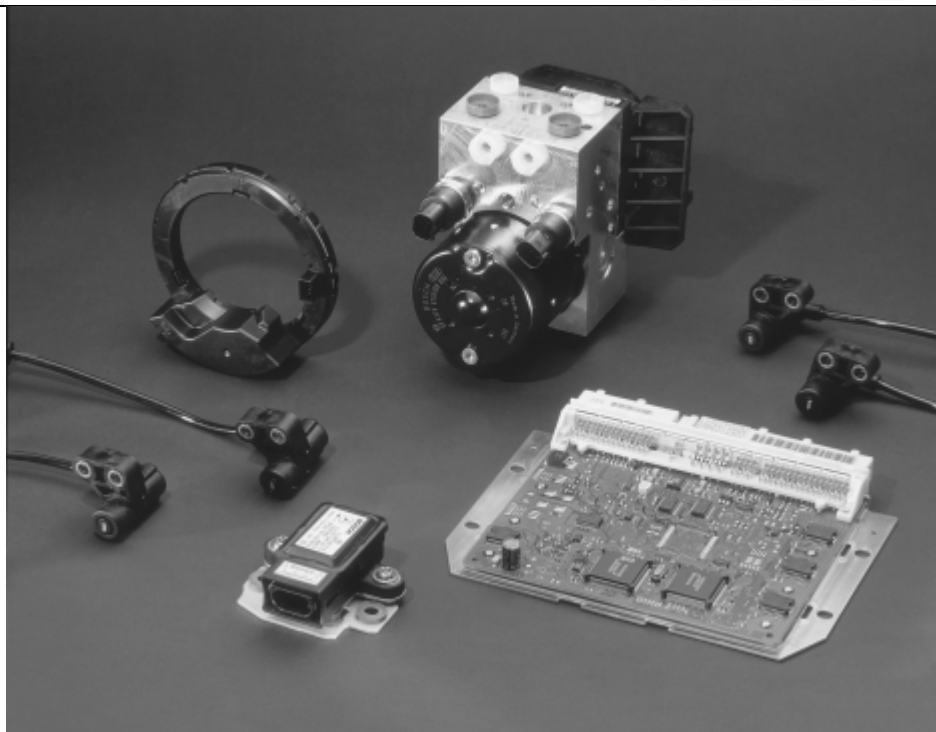
Figure 2.10. Sensotronic brake control system of Mercedes-Benz 211 series
E-Class sedan

The ESC reacts to and corrects skidding much faster and more effectively than a typical human driver, often before the driver is even aware of any imminent loss of control. Indeed, unlike

professional racing drivers who are able to maintain control in many spinout or plowout conditions by using countersteering (momentarily turning away from the intended direction), an average driver would be unlikely to do so in a panic situation. However, an ESC system is not a performance enhancement nor a replacement for safe driving practices, but rather a safety technology to assist the driver in recovering from dangerous situations. It will not prevent cases of vehicles leaving the road due to lack of concentration or drowsiness on the part of the driver.

An ESC system uses a number of sensors to verify what the driver wants (input), while other sensors indicate the actual state of the vehicle (response). Speed and steering angle measurements are used to determine the driver's intended heading. The vehicle response is measured in terms of lateral acceleration and yaw rate by onboard silicon- based sensors, which are manufactured using micromachining technologies. Most automotive ESC systems now use these MEMS sensors, which over the last decade have been able to drastically improve the ESC system in terms of size, cost and reliability, with the MEMS yaw rate replacing the conventional high-precision mechanical sensors.

Both the accelerometer and yaw rate sensors continuously send their data to the brain of the ESC system, a controller that stores the application algorithm, including the set of equations used to model the dynamics of the vehicle (Fig. 2.11).

Figure 2.11. Main ESC components: 1. Hydraulic bloc (valve actuation) with Electronic Control Unit (ECU); 2. Wheel speed sensor; 3. Steering wheel angle sensor; 4. Lateral and yaw rate sensors module; 5. Communication with engine management ECU.
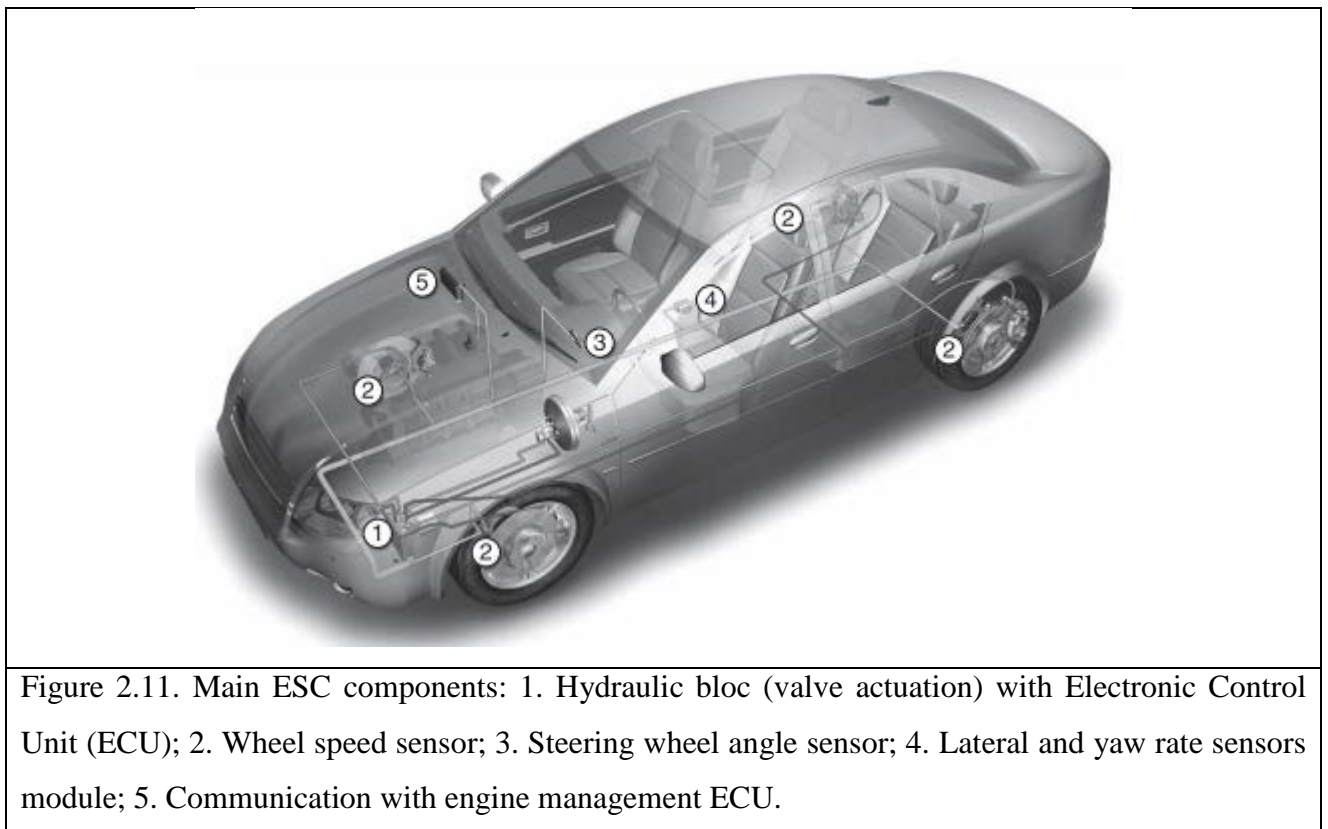
The control algorithm compares driver inputs to vehicle response and decides, when necessary, to apply the brakes driving the valves of the hydraulic modulator and/or to reduce throttle by a calculated amount. If the vehicle is responding correctly to the steering input, the yaw rate will be in balance with the speed and lateral acceleration. Since it is connected to the vehicle network (e.g. controller area network–CAN), the ESC controller is also able to receive data from and issue commands to other controllers such as an all- wheel drive system or an active suspension system, thereby allowing further improvements to vehicle stability and controllability. Most ESC systems have the option of disabling the function, which can be useful in various off- road driving conditions or when using a smaller- sized spare tire, which would interfere with the sensors. Some vehicle manufacturers offer an additional mode in which a driver can utilize the limits of car adhesion with less electronic intervention. However, the ESC system returns to normal operation when the ignition is re- started.   The most important sensors of an ESC system are:

- Steering wheel angle sensor: determines where the driver wants to steer (intended heading). This kind of sensor is often based on Hall effect sensors.
- Yaw rate sensor: measures the rotation rate (or angular speed) of the car along the vertical axis. The data from the yaw sensor is compared with the data from the steering wheel angle sensor to determine regulating action. The concept of 'yaw rate' can be illustrated by imagining a car following a large circle painted on a parking lot, viewed from above. If the car begins pointing north and drives half way around the circle, its new heading is south. Its yaw angle has changed by 180° degrees. If it takes 10 seconds to travel half way around the circle, the yaw rate is 180°/10 sec or 18 °/sec. If the speed stays the same, the car is constantly rotating at a rate of 18 °/sec around a vertical axis that can be imagined to be piercing its roof. If the speed is doubled, the yaw rate increases to 36 °/sec.
- Lateral acceleration sensor: measures the lateral acceleration of the vehicle. Today this relies only on a MEMS-based single or dual- axis capacitive accelerometer, able to measure low- g signal up to 1.5 to 3 g (full scale). The use of a dual- axis low- g sensor also allows the integration of new functionalities, such as hill start assist and electric parking brake (EPB). through an accurate measurement of the tilt angle of the vehicle while on a slope.
- Wheel speed sensor: measures the wheel speed. This kind of sensor is mostly based on contactless magnetic sensors.

3. MEMS microphones in portable devices, (mobile phones, head sets and laptops)

3.1 Introduction

MEMS microphones have been around for more than twenty years, but have been slow to achieve high volume commercial success. One possible reason for this is the availability of very low cost electret condenser microphones that meet the performance requirements of most consumer applications. Recently, however, interest in MEMS microphones has increased due to several factors. These include surface mount capability, integration of signal processing capability and low susceptibility to acceleration effects.   Electret condenser microphones cannot withstand high temperatures and therefore cannot go through a standard high volume surface mount manufacturing flow. The need to add in a custom assembly flow for the microphone can add cost to the total system. A MEMS microphone that can be assembled using standard low cost, high volume surface mount techniques would provide a system cost savings.

Electret microphones are a complex mechanical assembly and although they have been miniaturized tremendously over the years, it is widely believed that further miniaturization will be difficult. Conversion of the mechanical motion into an electrical signal and/or amplification of the signal in an electret can be achieved by including a junction gate field-effect transistor (JFET) or other device in the package. However, inclusion of complex circuit functionality such as that required to convert the analog signal to a digital signal or further sound processing would be difficult due to package space constraints. A MEMS microphone would not necessarily have this limitation. Lastly, electret microphones can be sensitive to acceleration which can often be present

in consumer as well as industrial applications. MEMS microphones can often be designed to be less sensitive to acceleration and thus, more suitable to a wider range of applications.

Fields of MEMS microphones application:

- Consumer electronics
- Laptop/Notebook
- Cellphones/Smartphones
- Earphone handsets
- Medical handsets
- Automotive
- Internet of Things (IoT)

Fig. 6.1. Application of MEMS microphones
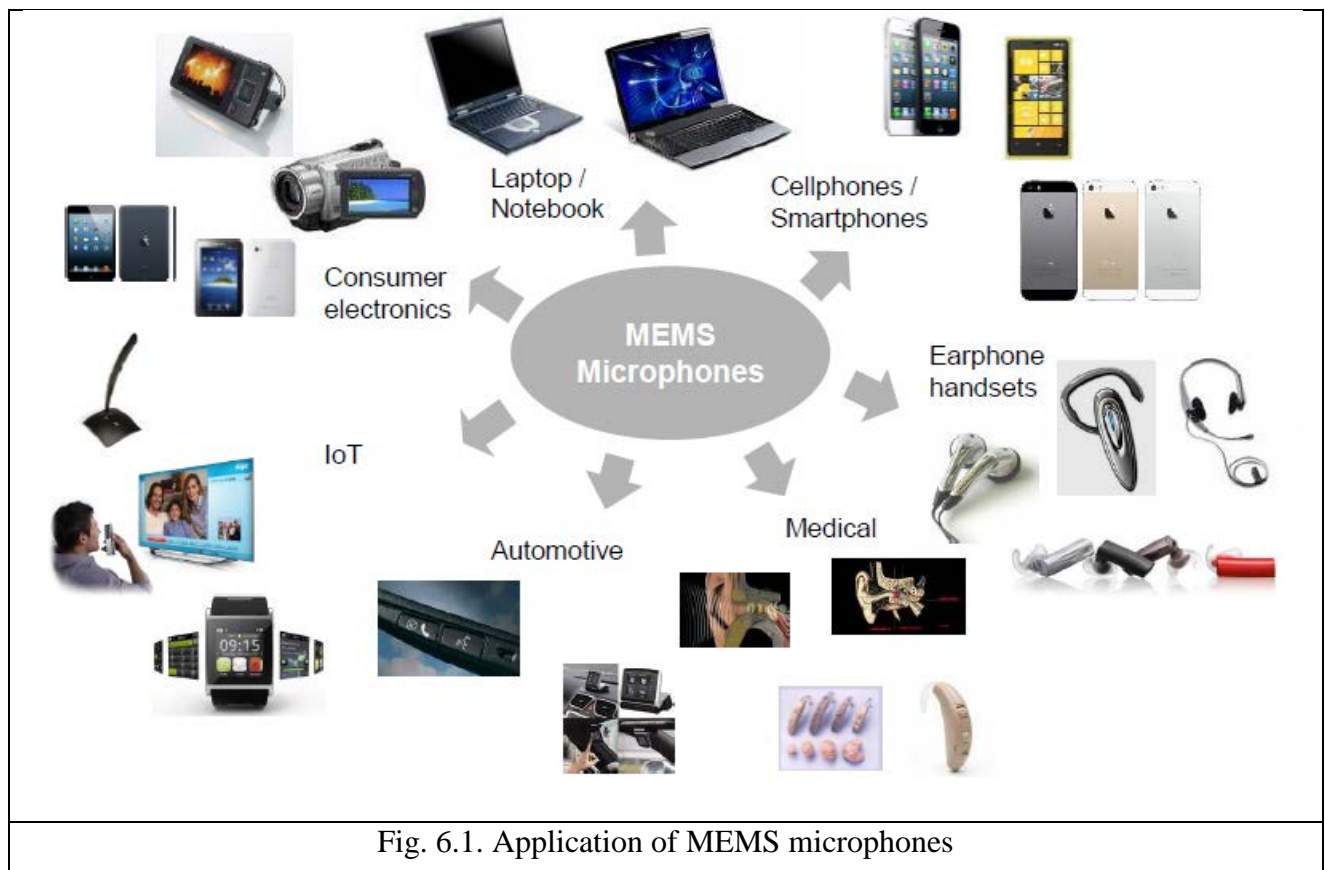
6.1.1 Main world MEMS manufacturers.

MEMS microphone is a very fast growing market with more than 29% growth in units in 2013-2014,driven by smart phone growth. The mobile phone application dominates the market however tablet market seems very promising with fastest growing.

Increasing fields of MEMS microphone applications leads to a corresponding increase in the number of microphones produced annually (Fig. 6.2). At the same time, the development of MEMS

manufacturing technology leads to a reduction of size and cost. Because of the price decrease, the market in value is growing more slowly than in volume.

Total market ($M)	2013	2014	2015	2016	2017	2018	2019
	785,0	929,0	1050,2	1173,7	1317,6	1482,8	1651,4
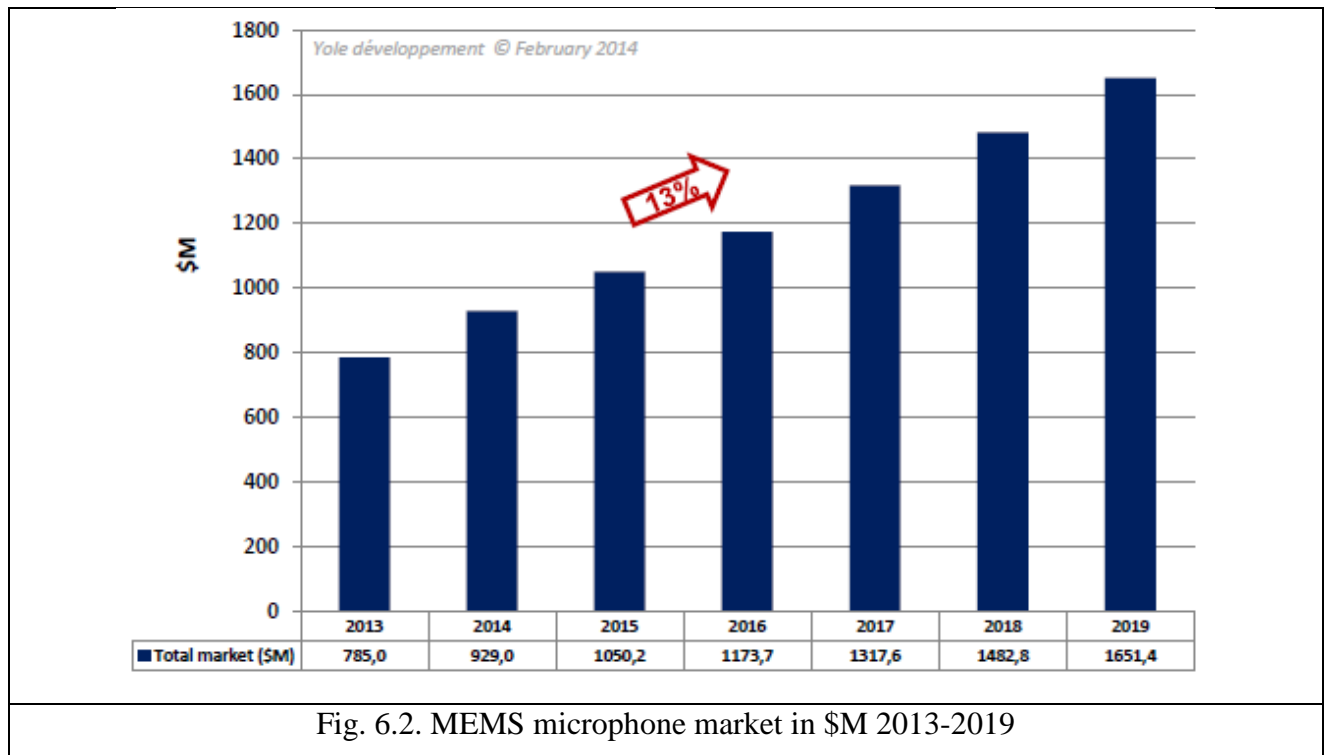
Fig. 6.2. MEMS microphone market in $M 2013-2019

The main MEMS microphone manufactures:

- **Knowles** is the dominant player with 61% market shares in 2013 and 58% in 2014. A decrease is again expected in the next years with new challengers in the market:

- **InvenSense(AnalogDevices):InvenSense** acquired Analog Devices (ADI) MEMS Microphone business and released HD products recently.

- **AAC Technologies** is a fast Chinese growing company that has cross licensing agreement with Knowles and acquired MEMSTECH.

- **Infineon Technologies** is the supplier of MEMS dies for most Knowles competitors and has lots of collaborations with Asian MEMS Microphone players.

- **STMicroelectronics** has collaboration with CEA-Leti on R&D projects and also with OMRON for MEMS Microphones.

- **OMRON** is looking at Infineon business model.

The global market share of these 6 players account for more than 80% of market share in 2014. These players are all developing innovative technical and manufacturing solutions, and, in parallel of course, the right patents to protect their inventions.

6.1.2 Demands for MEMS microphone.

A variety of trends increases the demands for high-quality voice and audio solutions for mobile devices, including:

- users requiring more freedom in how and where they communicate;

- users expecting high-quality voice and audio from their mobile devices;
- voice becoming a preferred interface for mobile device applications;
- users increasingly relying on their mobile devices for far-field interaction, where the mobile device is held remotely from the user, such as in speakerphone mode or video conferencing;
- users' perception of the HD video experience as negatively impacted by poor quality audio;
- OEMs continuing to expand functionality in mobile devices; and
- MNOs deploying wideband communications networks.

These trends, in turn, introduce challenges to delivering high-quality voice and audio in mobile devices, including:

- providing high quality even when used in noisy environments;
- working with the significant limitations on acoustics and signal processing imposed by the size, cost and power constraints of mobile devices;
- implementing voice and audio signal processing techniques that are scalable and adaptable to dynamic sound environments in a way traditional technology has not been able to provide.

6.2 General MEMS microphone construction.

6.2.1 Basic MEMS microphone construction.

MEMS microphones have been fabricated using a variety of methods, each having advantages and disadvantages. Many microphone designs use a polysilicon diaphragm which moves in response to pressure variations. The main common types of MEMS microphones are capacitance microphones and piezoelectric microphones.

When sound wave pushes the capacitance microphone diafragm, it vibrates and the capacitance of the condenser (or capacitor) changes. This is because the capacitance is proportional to the potential difference and inversely proportional to the separation between the plates. Any change in the separation changes the capacitance. The capacitance is also dependent upon the medium but as the medium here remains the same, so this parameter can be ignored. The values of the resistance and the capacitance are chosen such that the change in voltage is immediately reflected in the voltage across the resistance in series. Any change in sound leads to change of the capacitance and leads to voltage change. The voltage is fed to an amplifier to amplify the level of the signal.

Piezoelectric microphone uses the phenomenon of piezoelectricity — the ability of some materials to produce a voltage when subjected to pressure — to convert vibrations into an electrical signal.

6.2.1.1 Capacitor Microphone

The microphone presented here utilizes a polysilicon diaphragm in combination with a perforated backplate, forming two plates of a variable capacitor which are used to read out pressure variations. Often the diaphragm is solid and supported completely around its periphery. However, this design uses a spring supported, thin polysilicon diaphragm in order to increase sensitivity. To avoid an additional deposition of material for a back plate, this design uses the device layer of an silicon on insulator (SOI) wafer to form the backplate. Perforations are formed in the backplate which are designed to provide a minimum acoustic resistance while maximizing capacitive area. A cavity is etched in the handle layer of the SOI wafer to form a backside cavity for the microphone.

MEMS microphone basically is an acoustic transducer is shown in Fig. 6.1 and Fig. 6.2.:

- Transduction principle is the coupled capacity change between a fixed plate (back-plate) and a movable plate (membrane) Fig. 6.1.
- The capacitive change is caused by the sound, passing through the acoustic holes, that moves the membrane modulating the air gap comprised between the two conductive plates.
- The back-chamber is the acoustic resonator.
- The ventilation hole allows the air compressed in the back chamber to flow out and consequently allowing the membrane to move back.

Fig. 6.1. MEMS as acoustic transducer
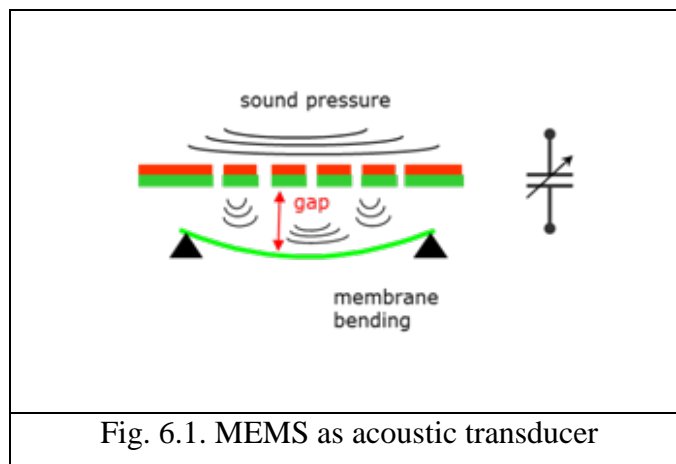
Cross-section of capacitive MEMS microphone construction is shown on Fig. 6.2.

Fig. 6.2. Cross-section of capacitive MEMS microphone construction
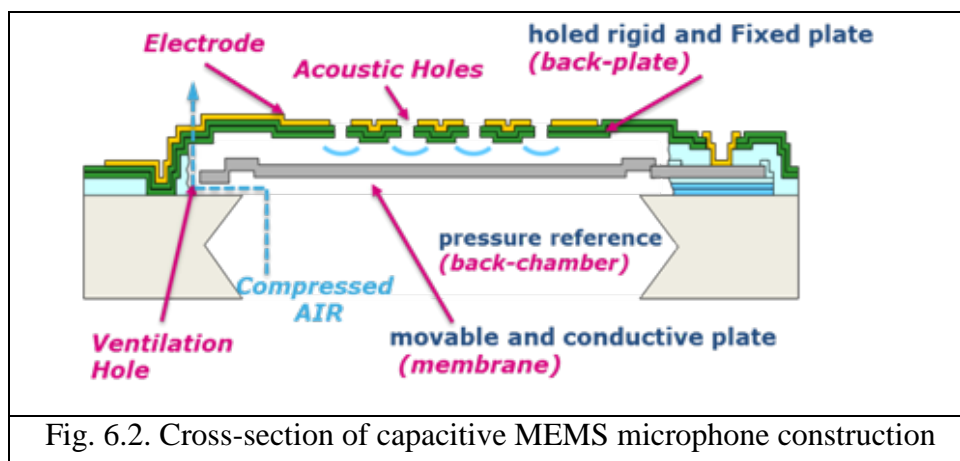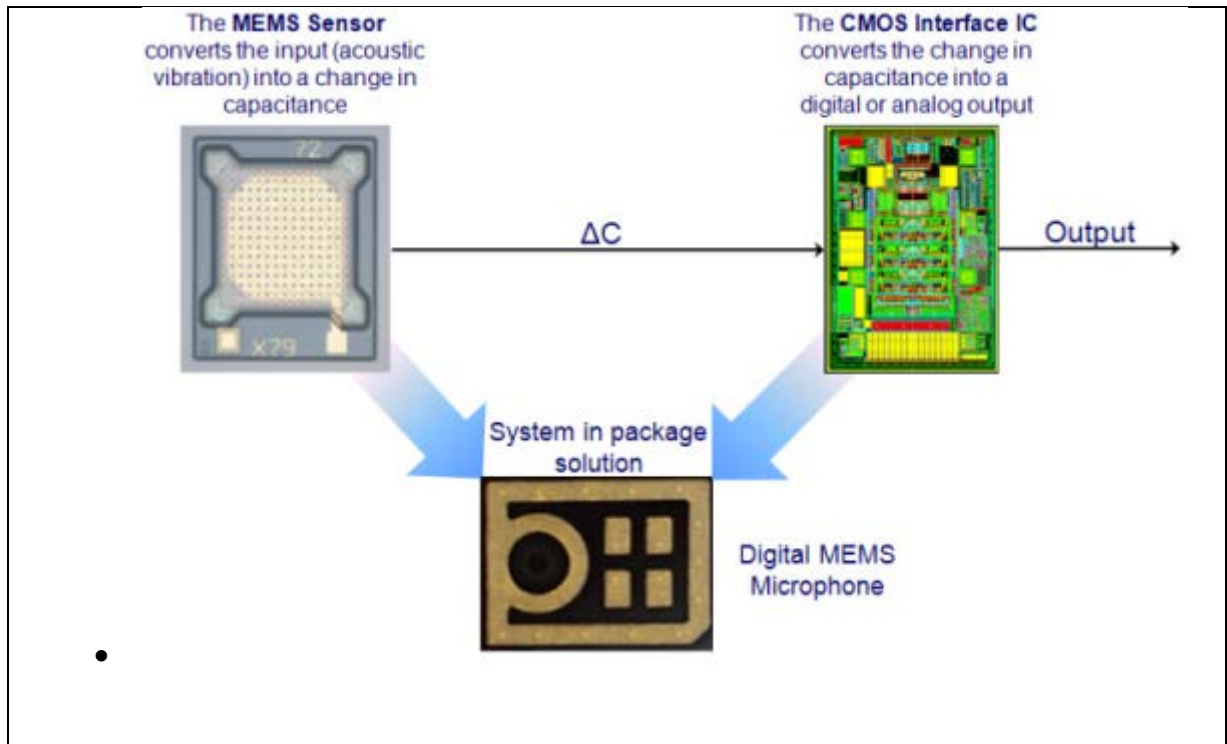
Fig. 6.3. MEMS microphone package components.

The acoustic sensing element is made using a special silicon micromachining process designed to produce audio sensors. The IC interface is manufactured using a CMOS process that allows designing a dedicated circuit able to provide a digital signal externally in PDM format. Manufactured-ready assembled MEMS microphone module consists from properly MEMS microphone, IC interface board and module housing Fig. 6.3.
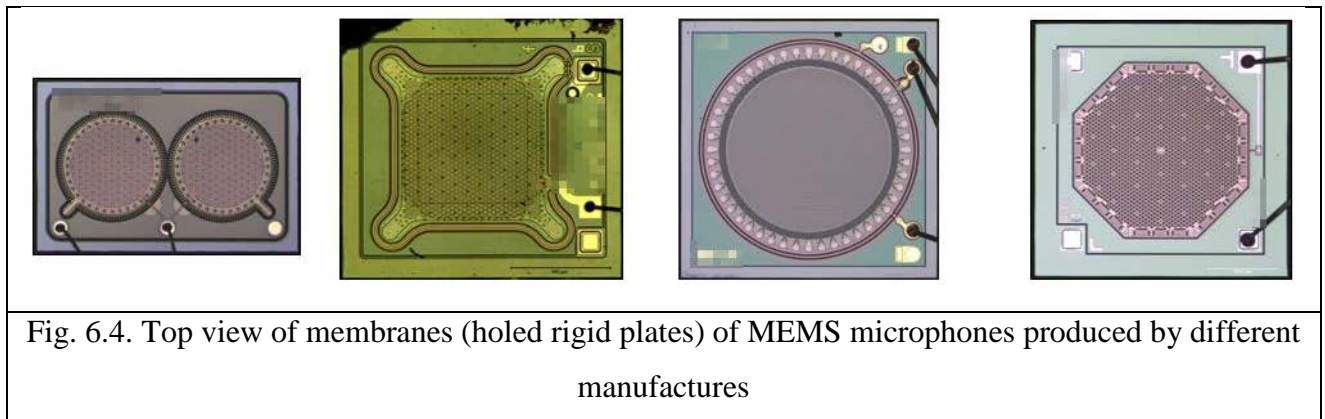
Fig. 6.4. Top view of membranes (holed rigid plates) of MEMS microphones produced by different manufactures

The form of microphone top rigid plate (Fig. 6.4), form of acoustic holes, their size, density locations on the rigid plate (Fig. 6.5) are selected by each manufacturer MEMS microphones so as to ensure the optimum parameters of elements. A scanning electron micrograph of the microphone diaphragm and one of its springs is shown on Fig. 6.6. Frequency response curves for the audio range 100..10000 kHz and ultrasound range 10..80 Hz are shown on Fig. 6.7c and Fig 6.7d. respectively.
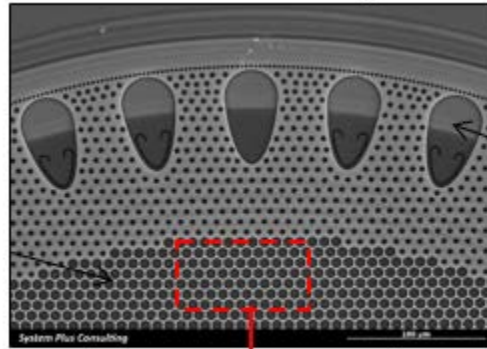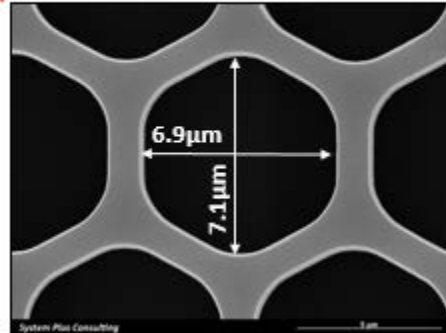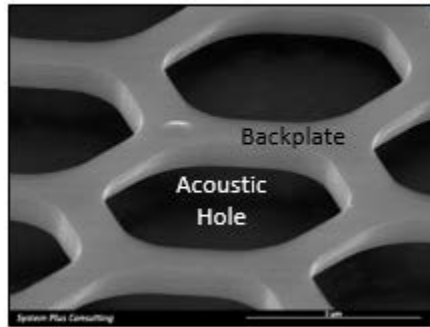
Fig.6.5. SEM view of top holed rigid plate of MEMS microphones.
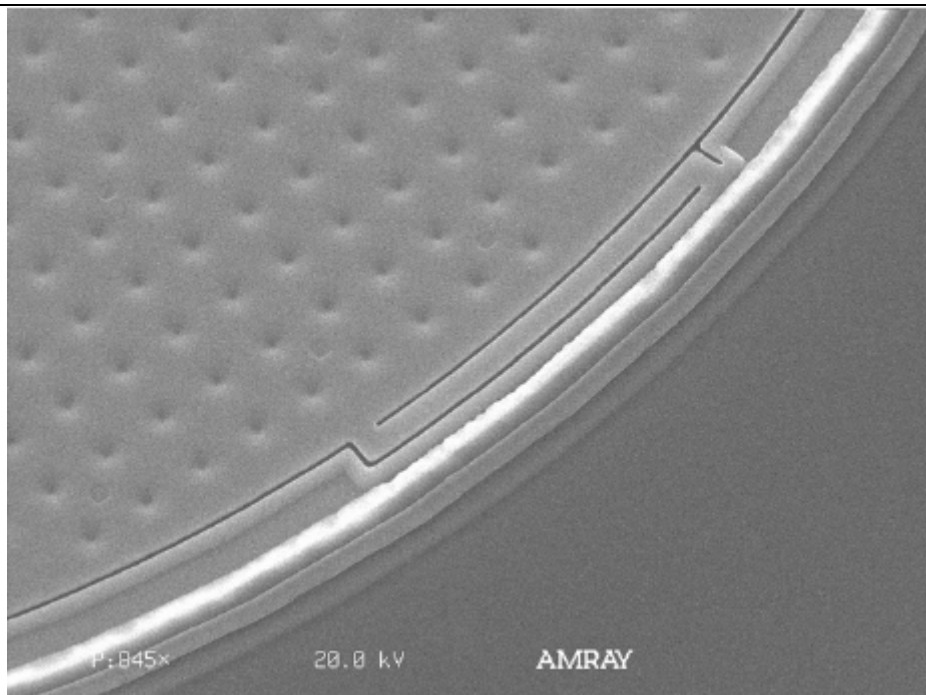
Fig. 6.6 SEM of microphone diaphragm and spring.
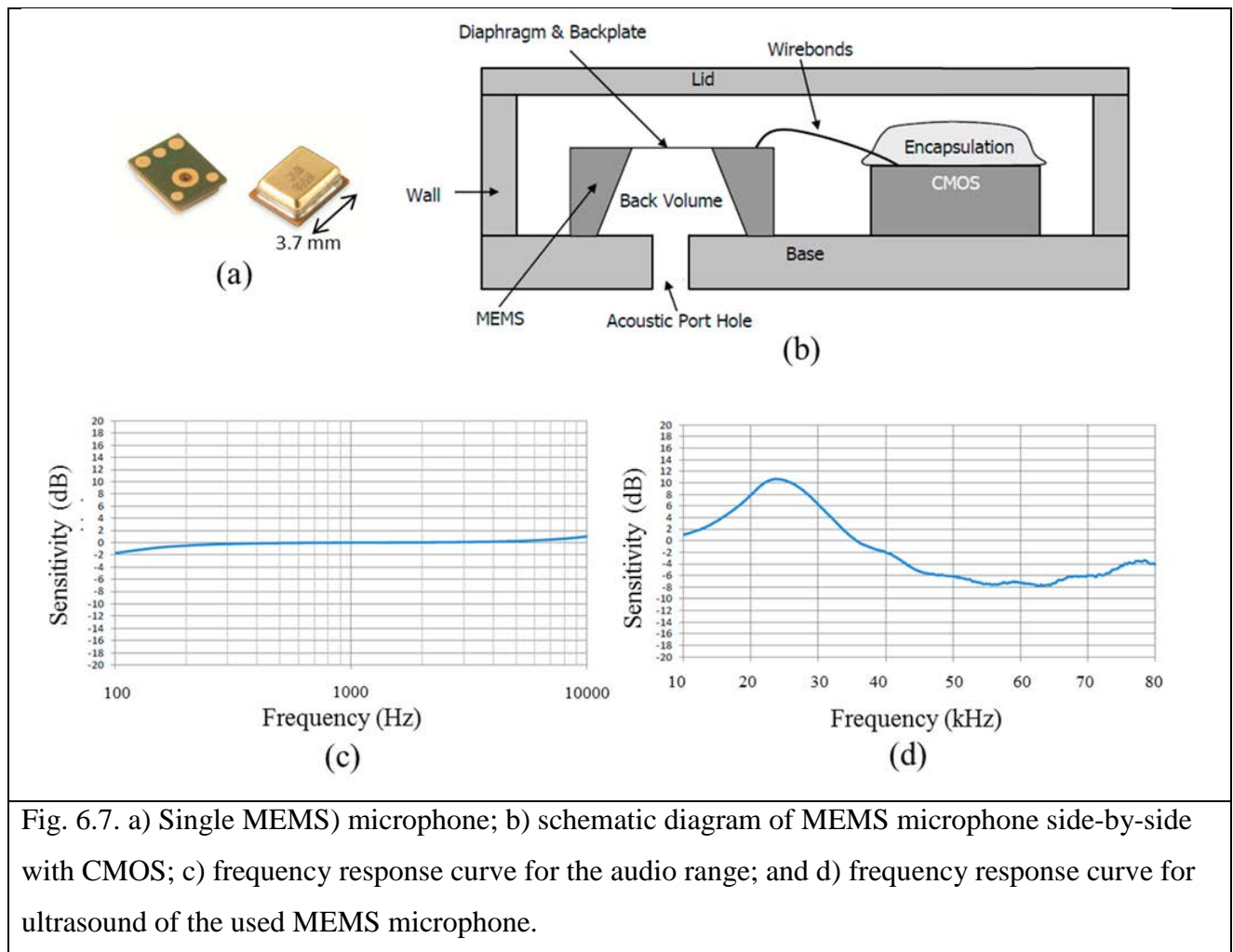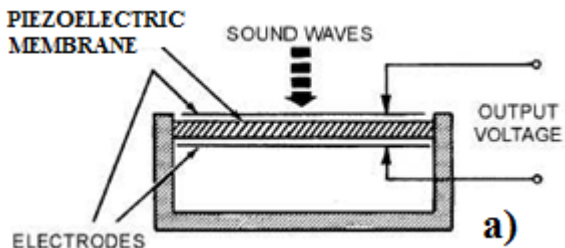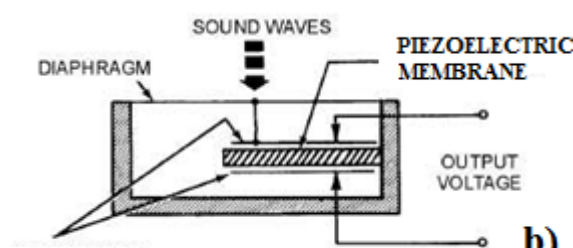
Fig. 6.7. a) Single MEMS) microphone; b) schematic diagram of MEMS microphone side-by-side with CMOS; c) frequency response curve for the audio range; and d) frequency response curve for ultrasound of the used MEMS microphone.

6.2.1.2 Piezoelectric microphone

Piezoelectric microphones rely on piezoelectric materials for the fundamentally linear conversion of mechanical energy to electrical energy. Conventional piezoelectric materials are crystals and ceramics that, when distorted by sound pressure produce voltages proportional to their dimensional strain. Such materials are used in microphones in two fundamentally different ways: A diaphragm responds to sound pressure by moving and the motion is transmitted to a piezoelectric material embedded in the device, or a piezoelectric material itself is used as the diaphragm of a microphone.

The advantage of a piezoelectric microphone is that, like the electret microphone, it does not require the attachment of an external power supply. Figure 6.8 shows a realization of a piezoelectric microphone in which sounds affect piezoelectric membrane directly. Figure 6.8b shows a realization of a piezoelectric microphone in which diaphragm is connected to a pin that drives a piezoelectric membrane internal to the microphone. Construction with diaphragm can provides sound direction selectivity.

.

| Fig. 6.8a. Directly actuated type of piezoelectric microphone. | Fig. 6.8b. Diaphragm type of piezoelectric microphone. |

According to MEMS microphone market analysis, MEMS microphones with better than 64dB Signal-to-noise ratio (SNR) will grow at a compound annual rate of 40 percent until 2017. In addition, up to 12 MEMS microphones are being used in modern automobiles, six in modern smartphones as well as smaller numbers in cameras, camcorders, hearing aids, Bluetooth headsets, wearables and all types of IoT portending a two billion unit market by 2017.

High-definition (HD) microphones will enable applications that were previously impossible for traditional MEMS microphones whose signal-to-noise ratio is limited by the need for a perforated backplate behind the diaphragm. Piezoelectric transducers, on the other hand, have an inherent and enduring advantage over traditional capacitive diaphragms, since piezoelectric tranducers do not require a perforated backplate, which introduces 8db of noise, according to Vesper.

High-definition piezoelectric MEMS microphones can be used for remote-less TVs that answer to voice commands even when they are blaring football scores. Household appliances that answer to voice commands even in the midst of chaotic family background noises, wearables that take dictation, ultrasonic MEMS stethoscopes and security systems whose cameras only need to be turned on when they hear a suspicious noise and all the Internet of Things (IoT) devices on the horizon. Many of these solutions are also being explored with gesture control which often requires video, but audio solutions, on the other hand, require much less power.

Smartphone makers will almost certainly snap-up MEMS microphones to improve not only voice commands, but dictation, noise cancellation and flawless audio-controlled voice-zooming which focuses on a single source of audio while shooting videos by virtue of multiple mic beam steering.

Firm "Vesper" has built a patent portfolio around their proprietary piezoelectric technology, which works quite differently from a traditional microphone. Instead of a traditional diaphragm, which requires a noise-producing perforated backplate, "Vesper" piezoelectric transducer is square and divided into four triangles, in the shape of an "X" drawn between opposite corners, with just enough space between them that they never touch. The piezoelectric material (aluminum nitride)

expands and contracts in response to the sound coming in its port producing an ultra low noise signal. The piezoelectric material is also relatively unaffected humidity and moisture, which causes reduced sensitivity in traditional capacitive MEMS microphones.

Fig. 6.9. a) High-definition low-noise piezoelectric MEMS microphone developed by "Vesper" comparatively to capacitive microphone (b).
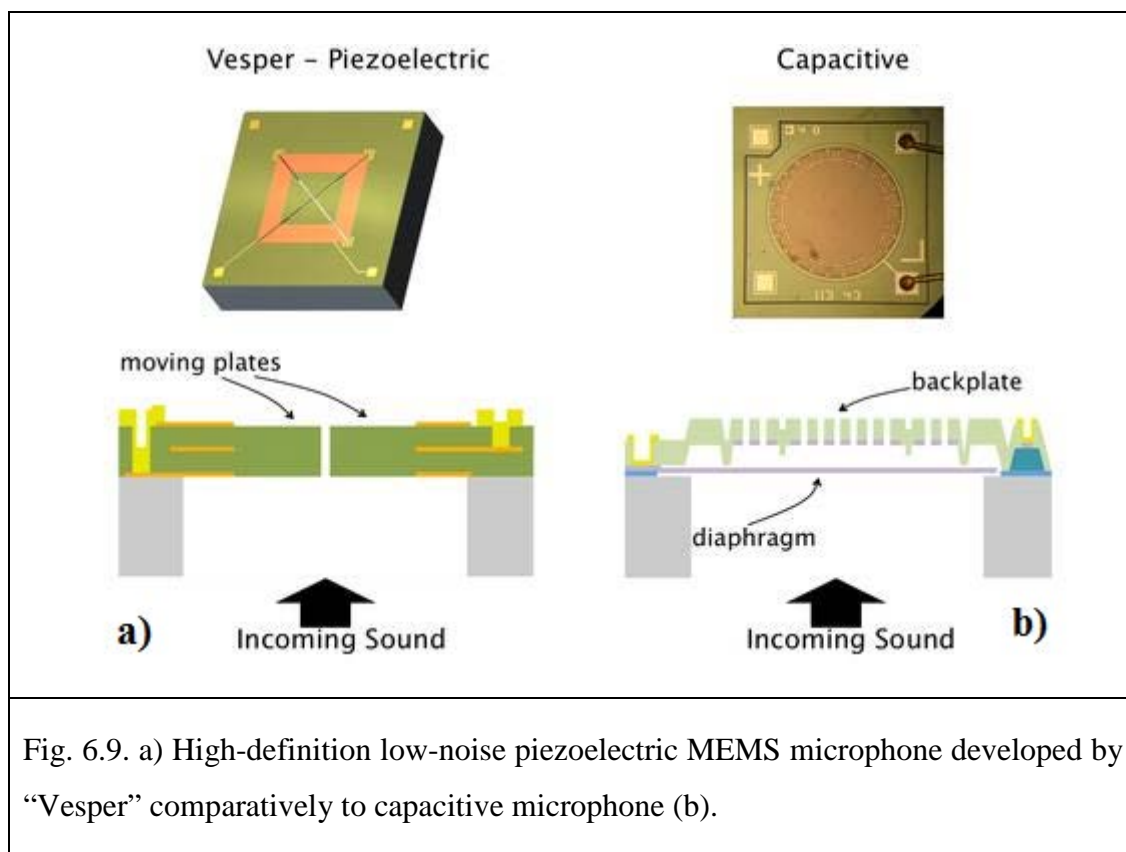
6.2.2 MEMS microphone encapsulation.

MEMS microphone can be encapsulated or side-by-side MEMS diaphragms and ASIC electronics or as monolithic MEMS microphone (Fig. 6.10). In case of separate location of ASIC electronics, area of MEMS microphone can be increased.

Fig. 6.10. Types of encapsulated MEMS microphone: a) microphone and ASIC are separated; b) microphone and ASIC are monolithic.
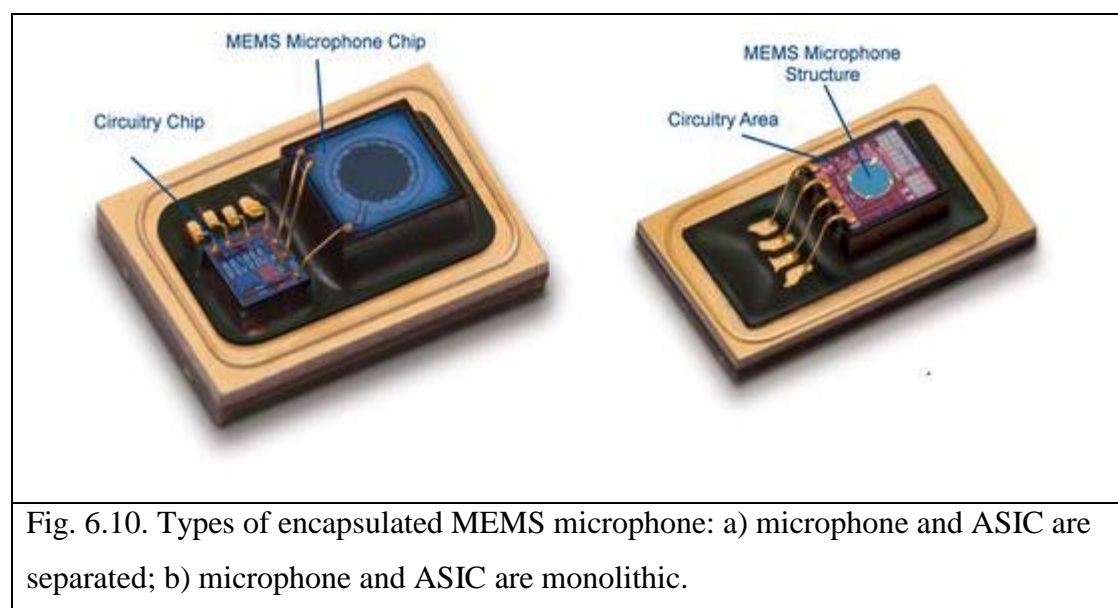
Analysis of MEMS microphone producing shows that the packege size of encapsulated MEMS microphones is decreased every year due to technology development. MEMS microphone package size is decreasing as for analog MEMS microphones as far as for digital MEMS microphones (Fig. 6.11, Fig. 6.12).

Fig. 6.11. MEMS microphone package size history and roadmap

Fig. 6.12. Trend of reducing size of MEMS microphone
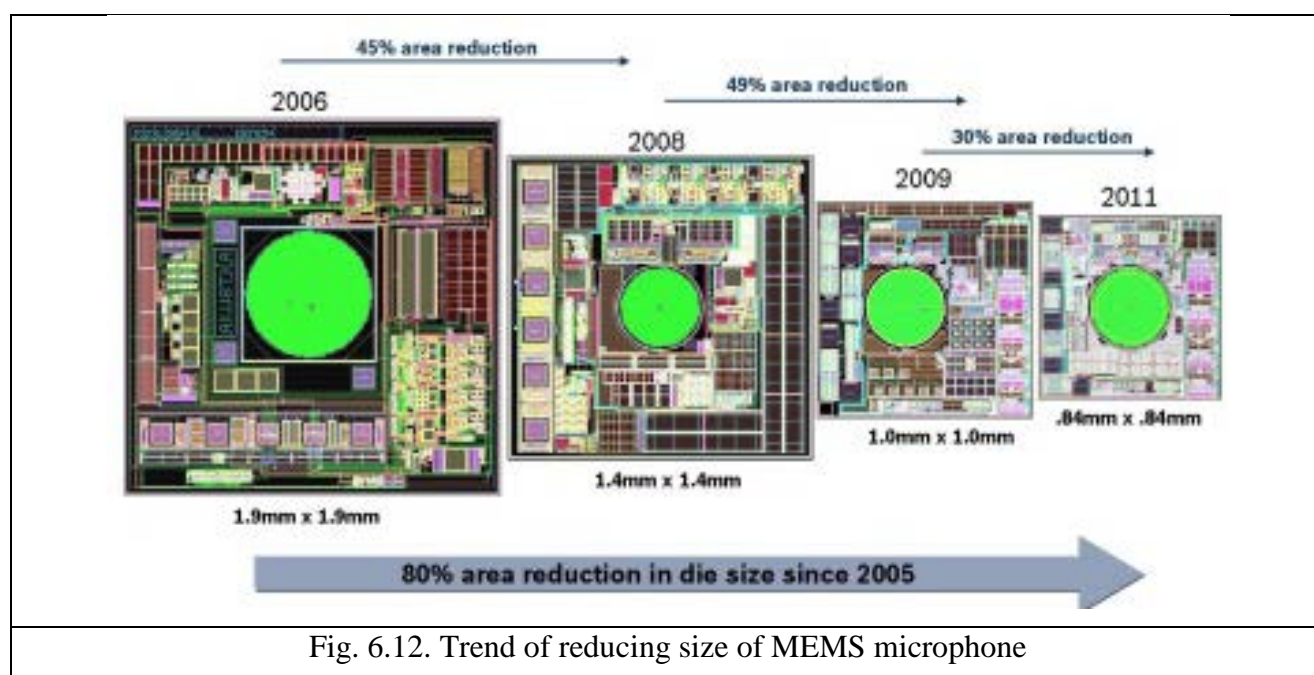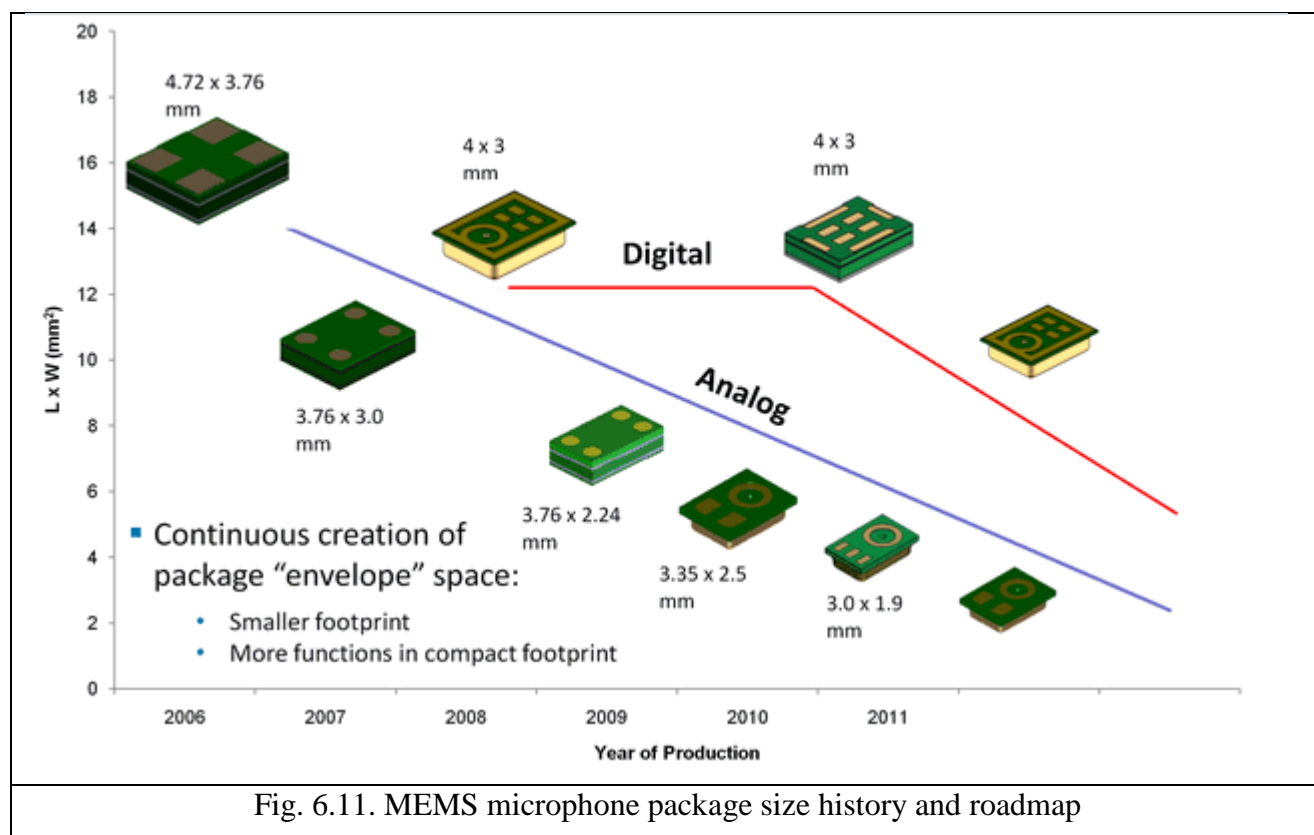
MEMS microphones need to have a hole in their package to allow sound to reach the acoustic sensor. The sound inlet can be located either in the lid (top port) or on the bottom next to the solder pads (bottom port). Bottom port microphones also require a hole in the circuit board they

are mounted on to allow sound to reach the sound inlet. The choice of whether to use a top port or bottom port microphone is usually determined by factors such as the location of the microphone in the product and manufacturing considerations, to mention a couple. Performance can also be a major factor in microphone port selection since top port microphones have traditionally had poorer performance than equivalent bottom port microphones. However, the introduction of high performance top-port mics such as ST's MP34DT01 means this is no longer necessarily true.

The membrane of the acoustic sensor divides the interior of a MEMS microphone into two sections. The area between the sound inlet and the sensor membrane is generally referred to as the front chamber, and the section on the other side of the membrane is known as the back chamber (Fig. 6.13.). The sensor in bottom port microphones is usually placed directly over the sound inlet which provides several benefits.
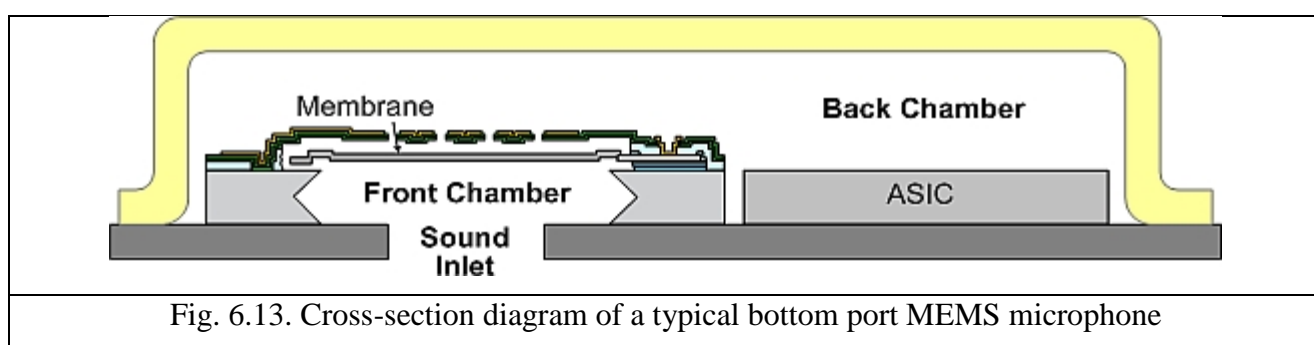
Fig. 6.13. Cross-section diagram of a typical bottom port MEMS microphone

The sensitivity of most MEMS microphones increases at higher frequencies. This increase in sensitivity is caused by the interaction between the air in the sound inlet and the air in the front chamber of the microphone. This interaction creates a Helmholtz resonance, which is the same phenomenon that creates sound when blowing into a bottle. As with bottles, smaller air volumes create higher resonant frequencies and larger air volumes create lower resonant frequencies. The microphone sensor is mounted directly over the sound inlet in most bottom port microphones, which results in a relatively small front chamber and a high center frequency for the Helmholtz resonance. Because the Helmholtz resonance is normally located in the upper part of the audio band, increasing the resonant frequency leads to a flatter frequency response.

Placing the sensor directly over the sound inlet also creates a relatively large back chamber (Fig. 6.14). A larger volume of air in the back chamber makes it easier for the membrane to move in response to sound waves, which improves the sensitivity of the microphone and leads to higher SNR. A large back chamber also improves the microphone's low frequency response. The construction of top port microphones has traditionally been very similar to bottom port microphones, with the sensor and the interface IC mounted on the substrate with a hollow lid enclosing them. Traditionally, the only real difference between top port and bottom port microphones is that the sound inlet is located in the microphone lid instead of in the substrate. For these microphones, moving the sound inlet to the lid turns what was previously the front chamber into the back chamber and vice versa.
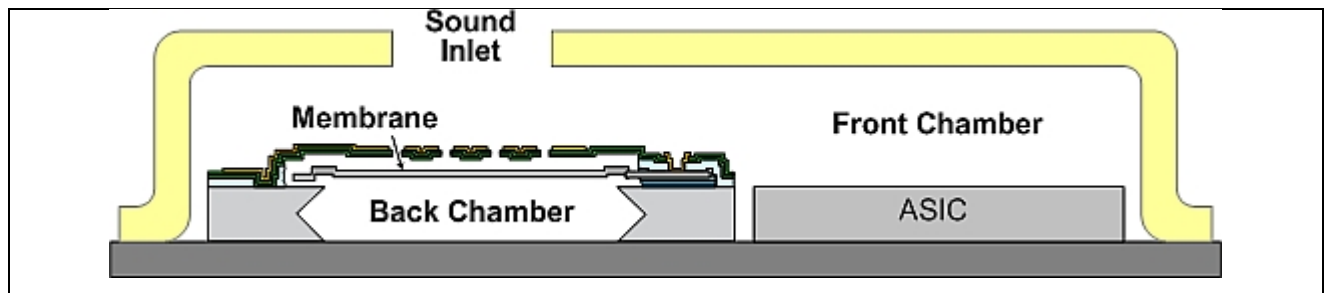
Fig. 6.14. Cross-section diagram of a traditional top port MEMS microphone.

The smaller air volume in the back chamber of traditional top port MEMS microphones makes it more difficult for the membrane to move, which hurts the sensitivity of the sensor and leads to a lower SNR. In addition, the larger air volume in the front chamber between the sound inlet and the membrane lowers the resonant frequency, hurting the microphone's high frequency response. This combination of lower SNR and poorer frequency response at both high and low frequencies is the reason that most top port microphones have poorer performance than an equivalent bottom port microphone.
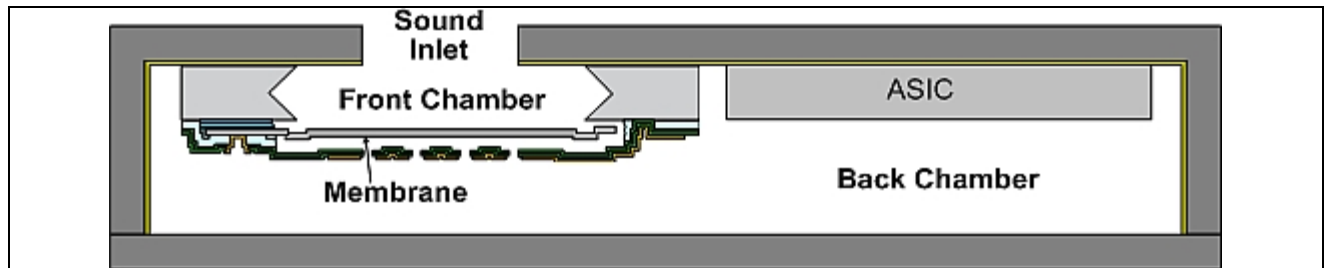
Figure 6.15 - Cross-section diagram of ST's MP34DT01 top port MEMS microphone

An exception to this rule is STMicroelectronics' MP34DT01 top-port digital MEMS microphone. ST's proprietary packaging technology makes it possible to mount the MEMS sensor and the interface IC on the bottom side of the lid of the MP34DT01, directly beneath the sound inlet (Fig. 6.15 and 6.16). This results in a small front chamber and a large back chamber and allows the MP34DT01 to achieve the same level of performance as the bottom port version of this mic, the MP34DB01.
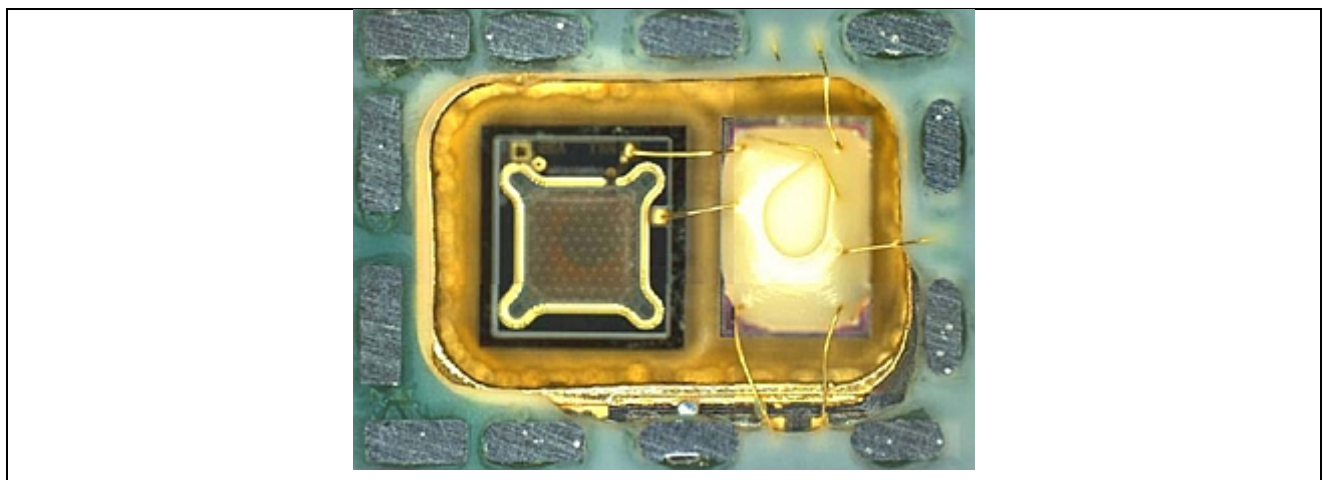
Fig. 6.16. Photo of MP34DT01 top port MEMS microphone

6.2.3 MEMS microphone integration with integrated circle

There are mainly two types of MEMS microphones:

- Analog − convert sound into corresponding voltage output

- Digital − gives a digital output typically pulse density modulation [PDM].

The ASIC inside a MEMS microphone uses a charge pump to place a fixed charge on the microphone membrane.   The ASIC then measures the voltage variations caused when the capacitance between the membrane and the fixed backplate changes due to the motion of the membrane in response to sound waves. Analog MEMS microphones produce an output voltage that is proportional to the instantaneous air pressure level. Analog mics usually only have 3 pins: the output, the power supply voltage (VDD), and ground. Although the interface for analog MEMS microphones is conceptually simple, the analog signal requires careful design of the printed circuit board (PCB) and cables to avoid picking up noise between the microphone output and the input of the IC receiving the signal. In most applications, a low noise audio analog-to-digital converter (ADC) is also needed to convert the output of analog microphones into digital format for processing and/or transmission.
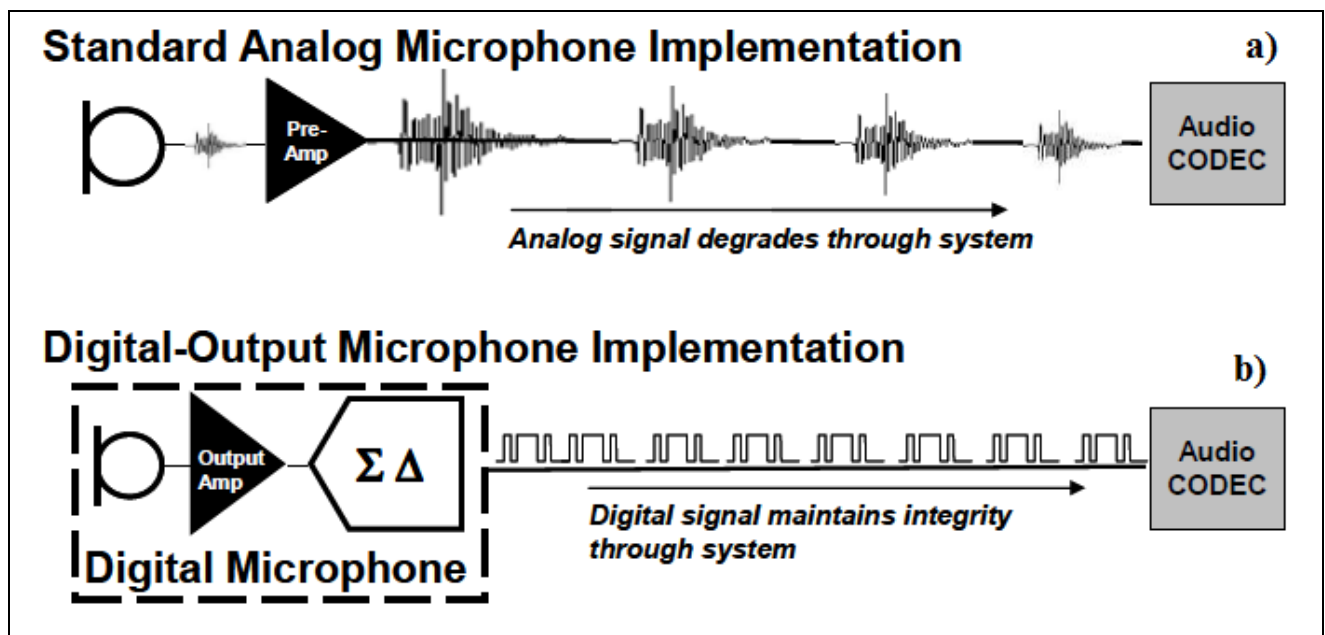
Fig. 6.17. Comparison of analog- vs. digital-output microphone signal through end device

As their name implies, digital MEMS microphones have digital outputs that switch between low and high logic levels. Most digital microphones use pulse density modulation (PDM), which produces a highly oversampled single-bit data stream. The density of the pulses on the output of a microphone using pulse density modulation is proportional to the instantaneous air pressure level. Pulse density modulation is similar to the pulse width modulation (PWM) used in class D amplifiers. The difference is that pulse width modulation uses a constant time between pulses and

encodes the signal in the pulse width, while pulse density modulation uses a constant pulse width and encodes the signal in the time between pulses.

In addition to the output, ground, and VDD pins found on analog mics, most digital mics also have inputs for a clock and a L/R control. The clock input is used to control the delta-sigma modulator that converts the analog signal from the sensor into a digital PDM signal. Typical clock frequencies for digital microphones range from about 1 MHz to 3.5 MHz. The microphone's output is driven to the proper level on the selected clock edge and then goes into a high impedance state for the other half of the clock cycle. This allows two digital mic outputs to share a single data line. The L/R input determines which clock edge the data is valid on.

The digital microphone outputs are relatively immune to noise, but signal integrity can still be a concern due to distortion created by parasitic capacitance, resistance, and inductance between the microphone output and the SoC. Impedance mismatches can also create reflections that can distort the signals in applications with longer distances between the digital mic and the SoC.

The output of a MEMS microphone does not come directly from the MEMS transducer element. The transducer is essentially a variable capacitor with extremely high output impedance in the gigaohm range. Inside the microphone package, the transducer's signal is sent to a preamplifier, whose first function is an impedance converter to bring the output impedance down to something more usable when the microphone is connected in an audio signal chain. The microphone's output circuitry is also implemented in this preamp. For an analog MEMS microphone, this circuit, whose block diagram is shown in Fig. 6.18, is basically an amplifier with a specific output impedance.
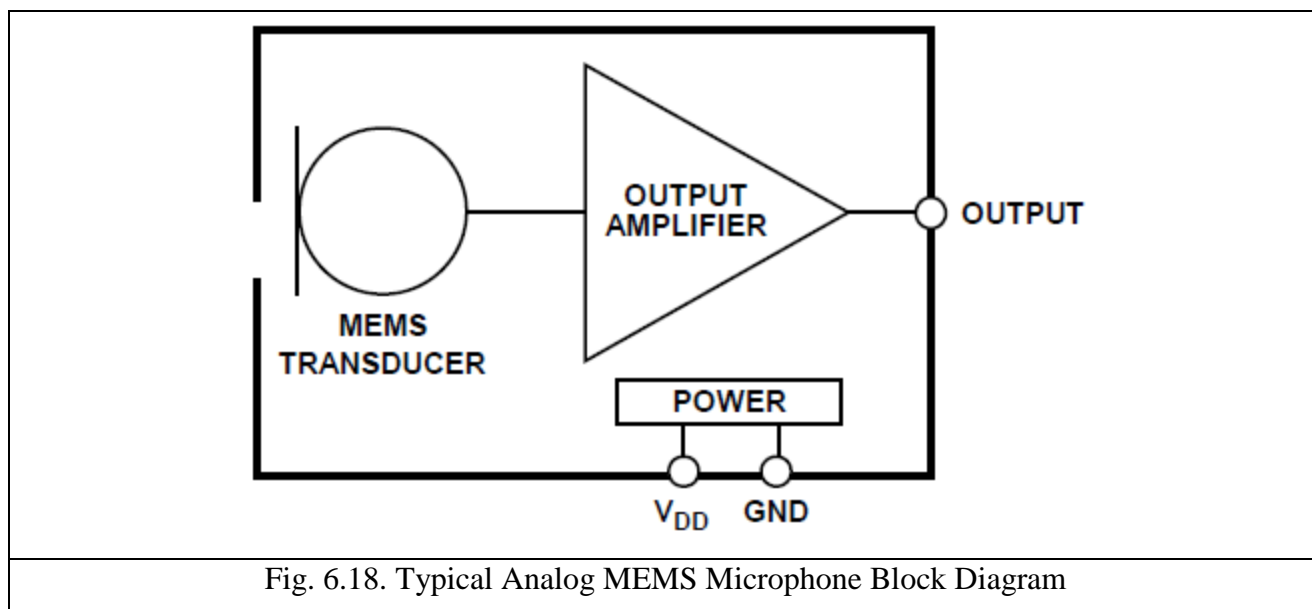
Fig. 6.18. Typical Analog MEMS Microphone Block Diagram

In a digital MEMS microphone, that amplifier is integrated with an analog-to-digital converter (ADC) to provide a digital output in either a pulse density modulated (PDM) or I^2S format. Figure 6.19 shows a block diagram of a PDM-output MEMS microphone.

Fig. 6.19. Typical PDM MEMS Microphone Block Diagram
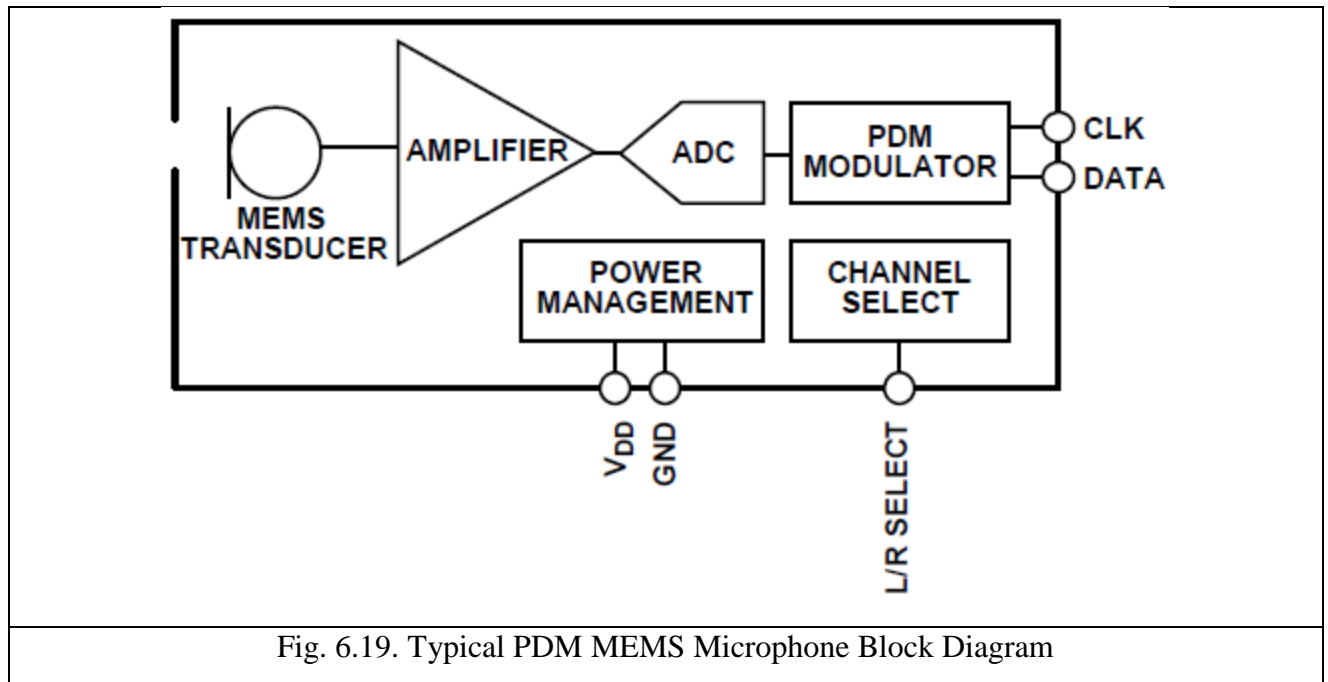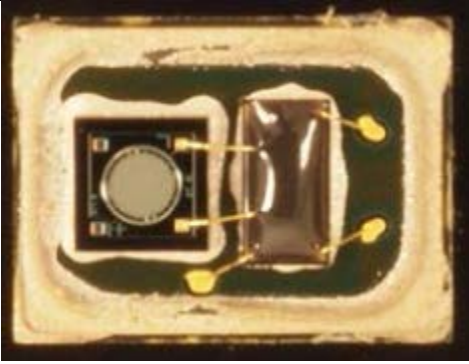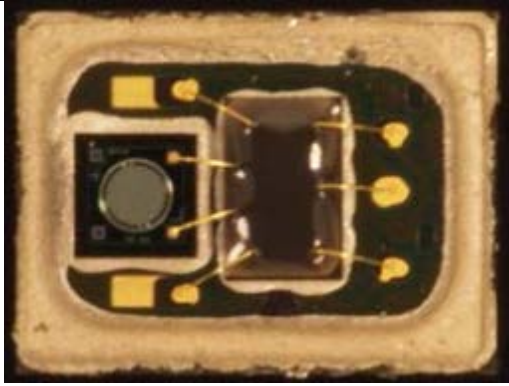
Figure 6.20 shows a typical I²S -output digital microphone. The I²S microphone contains all of the digital circuitry that a PDM microphone has, as well as a decimation filter and serial port.

Fig. 6.20. Typical I²S MEMS Microphone Block Diagram
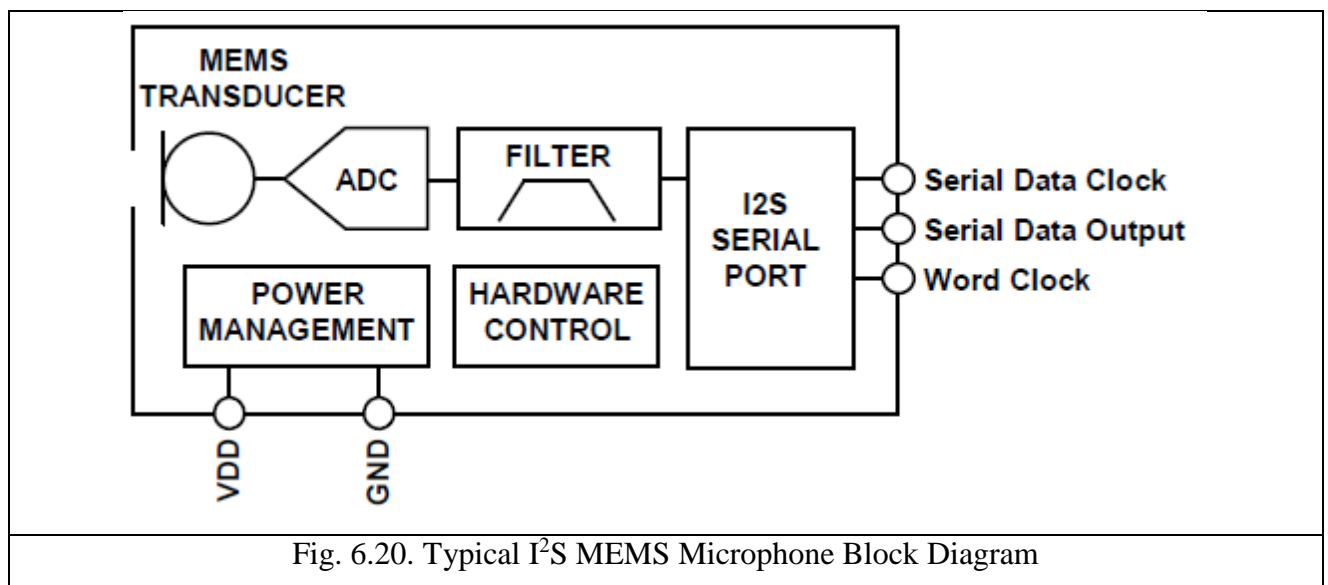
A MEMS microphone package is unique among semiconductor devices, in that there is a hole in the package for the acoustic energy to reach the transducer element. Inside this package, the MEMS microphone transducer and the analog or digital ASIC are bonded together and mounted on a common laminate.

A lid is then bonded over the laminate to enclose the transducer and ASIC. This laminate is basically a small PCB that is used to route the signals from the ICs to the pins on the outside of the microphone package. Figures 6.21a and 6.21b show the inside of analog and digital MEMS microphones, respectively. In these pictures one can see the transducer on the left and ASIC (under the epoxy) on the right side, both mounted on the laminate. The digital microphone has additional bond wires to connect the electrical signals from the ASIC to the laminate.

Fig. 6.21a. Transducer and ASIC of an Analog MEMS Microphone	Fig. 6.21b. Transducer and ASIC of a Digital MEMS Microphone

6.2.3.1 Analog microphones

An analog MEMS microphone's output impedance is typically a few hundred ohms. This is higher than the low output impedance that an op amp typically has, so one need to be aware of the impedance of the stage of the signal chain immediately following the microphone. A low impedance stage following the microphone will attenuate the signal level. For example, some codecs have a programmable gain amplifier (PGA) before the ADC. At high gain settings, the PGA's input impedance may be only a couple of kilohms. A PGA with a 2 kΩ input impedance following a MEMS microphone with a 200 Ω output impedance will attenuate the signal level by almost 10%. Scheme of connection of Analog microphone connection to noninverting Op-Amp circuit isshown on Fig. 6.22.
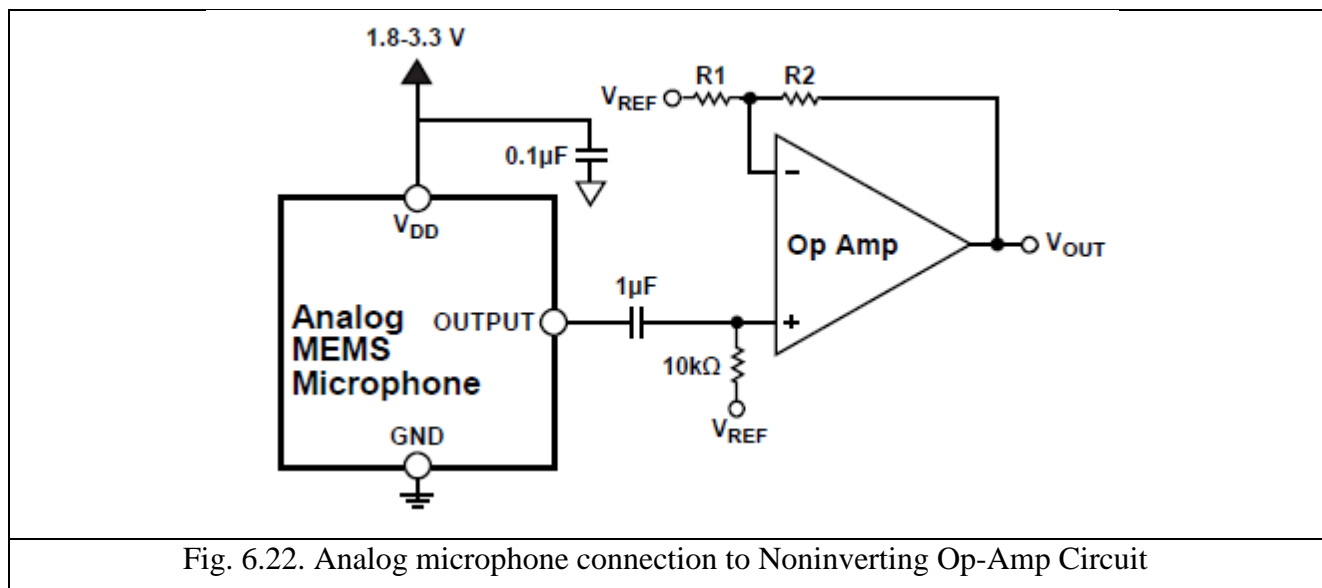
Fig. 6.22. Analog microphone connection to Noninverting Op-Amp Circuit

6.2.3.3 Digital microphones

Digital microphones move the analog-to-digital conversion function from the codec into the microphone, enabling an all-digital audio capture path from the microphone to the processor. Digital MEMS microphones are often used in applications where analog audio signals may be susceptible to interference. For example, in a tablet computer, the microphone placement may not

be near to the ADC, so the signals between these two points may be run across or near Wi-Fi, Bluetooth, or cellular antennae. By making these connections digital, they are less prone to picking up this RF interference and producing noise or distortion in the audio signals. This improvement in pickup of unwanted system noise provides greater flexibility in microphone placement in the design.

Digital microphones are also useful in systems that would otherwise only need an analog audio interface to connect to an analog microphone. In a system that only needs audio capture and not playback, like a surveillance camera, a digital output microphone eliminates the need for a separate codec or audio converter and the microphone can be connected directly to a digital processor.

6.2.3.3 MEMS microphone form factor

Generally, analog MEMS microphones are available in smaller packages than digital microphones. This is because an analog microphone package needs fewer pins (typically three, vs. five or more for a digital microphone) and the analog preamp has less circuitry than a digital preamp. This makes the analog preamp smaller than a digital preamp manufactured in the same fab geometry. Consequently, in the most space-constrained designs, such as in many small mobile devices, analog microphones are preferred in part because of their small size. An analog microphone can be in a package with dimensions 2.5 mm × 3.35 mm × 0.88 mm or smaller, while PDM microphones often come in a 3 mm × 4 mm × 1 mm package, an increase of 62% in package volume.
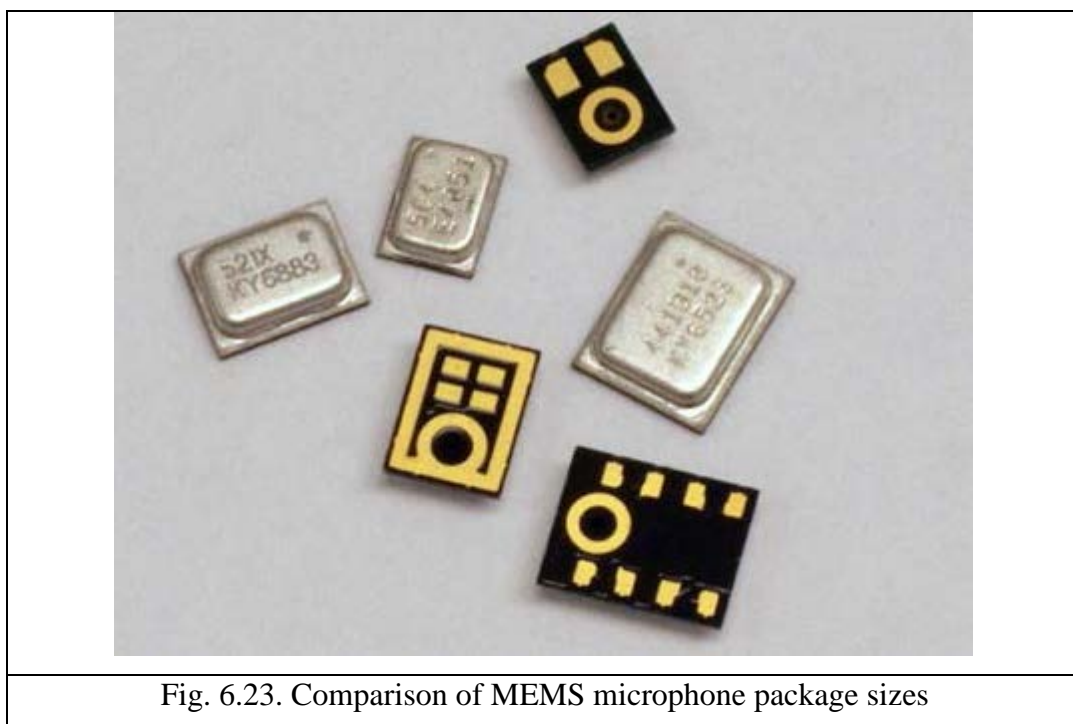
Fig. 6.23. Comparison of MEMS microphone package sizes

Figure 6.23 shows a comparison of three bottom port microphone packages. The smallest is the ADMP504, an analog microphone in the 2.5 mm × 3.35 mm × 0.88 mm, the middle-sized

microphone is the ADMP521, a PDM microphone in the 3 mm × 4 mm × 1 mm package, and the microphone in the largest package is the ADMP441, an I2S microphone in a 3.76 mm × 4.72 mm × 1.0 mm package. This last microphone is in this larger package to support its nine pins. Despite its larger size, a microphone like this is comparable in functionality to an analog microphone and an ADC together, so the savings in PCB area if a converter is otherwise not needed outweighs the slightly larger microphone footprint.

6.3 Main parameters of MEMS microphones

Sensitivity:

The sensitivity is the electrical signal at the microphone output to a given acoustic pressure as input. The reference of acoustic pressure is 1Pa or even 94dBSPL @ 1kHz.

- Sensitivity is typically measured:
  - for Analog microphones in mV/Pa or even dBV = 20 * Log (mV/Pa / 1V/Pa)
  - for Digital microphones in %FS or even dBFS = 20 * Log (%FS / 1FS)

Directionality:

- The directionality indicates the variation of the sensitivity response with respect to direction of arrival of the sound.
- OMNI-Directional means that there is no sensitivity change at every sound source position in the space.
- The directionality can be indicated in a Cartesian axis as sensitivity drift vs. angle or in a polar diagram showing the sensitivity pattern response in the space.

Signal to Noise Ratio [SNR]:

- The signal-to-noise ratio specifies the ratio between a given reference signal to the amount of residual noise at the microphone output.
- The reference signal is the standard signal at the microphone output when the sound pressure is 1Pa @ 1kHz. In other words the microphone sensitivity.
- The noise signal (residual noise) is the microphone electrical output at the silence. This quantity includes both the noise of the MEMS element and the application-specific integrated circuit (ASIC).
- Typically the noise level is measured in an anechoic environment and weighting-A the acquisition. The A-weighted filter corresponds to the human ear frequency response.
- The SNR of current MEMS microphones ranges from about 56 dB to about 66 dB.

Dynamic Range and AOP:

The dynamic range is the difference between the minimum and maximum detectable sound by the microphone without distortion:

- The maximum signal that the microphone can "listen" without distortion is also called acoustic overload point (AOP). For both analog and digital ST microphones the AOP is 120dBSPL as sound pressure.
- The minimum signal that a microphone can "listen" depends on its SNR. In other words, the minimum signal is equivalent to the residual noise in terms of dBSPL.

Frequency response:

The frequency response of a microphone in terms of magnitude:

- Indicates the sensitivity variation across the audio band. Or even, this parameter describes the deviation of the output signal from the reference 0dB.
- Typically the reference for this measurement is the exactly the sensitivity of the microphone à 0dB = 94dBSPL @ 1kHz.
- The typical frequency response of a microphone shows a roll-off at low frequency due to ventilation hole and an rise up at high frequency due to Helmholtz effect.

The frequency response of a microphone in terms of phase:

- Indicates the phase distortion introduced by the microphone. In other words the delay between the sound wave moving the microphone membrane and the electrical signal at the microphone output
- This parameter includes both the distortion due to the membrane and the ASIC

Noise floor

The noise floor of a microphone is the amount of noise on its output in a perfectly quiet environment. Both the sensor and the interface ASIC contribute noise to the output of a microphone. The noise contributed by the sensor is created by the random Brownian motion of air molecules while the noise from the ASIC is created by the preamplifier and, in the case of digital microphones, the delta-sigma modulator. The noise floor is measured across the full audio band and A-weighting filters are used to provide a better measure of the noise level as perceived by human ears.

The noise floor is not always specified in microphone datasheets but it can be calculated by subtracting the microphone's SNR from its sensitivity, providing a result in dBV or dB FS. The noise floor can be expressed as an equivalent input noise in dB SPL by subtracting the SNR from the sound pressure level that the sensitivity is measured at (usually 94 dB SPL).

Distortion (THD)

Distortion is a measure of how accurately a microphone can capture sound. Distortion is usually specified at about 94 dB – 100 dB SPL in order to provide a good indication of a microphone's audio quality at normal sound levels.

Acoustic Overload Point (AOP)

Distortion normally does not increase much as the sound pressure level increases until the sound pressure level starts to approach the acoustic overload point of the microphone. When this happens the distortion starts rising rapidly. The acoustic overload point of a microphone is normally defined as the sound pressure level where the distortion reaches 10%.

Frequency response

The frequency response of a MEMS microphone refers to the change in its sensitivity at various frequencies. The frequency response of a microphone is usually set to 0 dB at 1 kHz to normalize the results. The sensitivity of most MEMS microphones falls off below 100 Hz and starts rising between about 4 kHz – 6 kHz due to the Helmholtz resonance. This is the reason that many MEMS mics only specify their frequency response between 100 Hz and 10 kHz. However, high performance MEMS microphones are available with a relatively flat frequency response over the full audio band from 20 Hz to 20 kHz.

Directional acoustic patterns using MEMS Microphone

An omnidirectional microphone response is generally considered to be a perfect sphere in three dimensions. The smallest diameter microphone gives the best Omni-directional characteristics at high frequencies. This is the reason that makes the MEMS microphone the best Omni-directional microphone.

But MEMS Microphone can also be used in array to modify the response according to desired acoustic patterns Fig. 6.24.
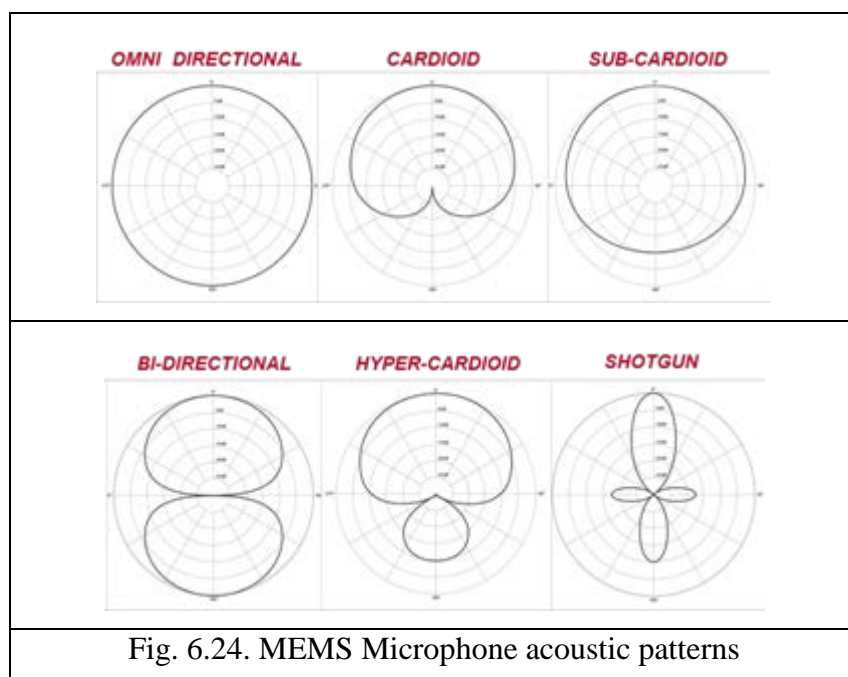
Fig. 6.24. MEMS Microphone acoustic patterns

Power supply rejection (PSR)

The power supply rejection of a microphone is a measure of its ability to prevent noise on the microphone's power supply input from appearing on its output. PSR is usually specified with a 217

Hz square wave to simulate the TDMA noise generated by GSM cellular radios and/or a swept sine wave across the audio band.

6.4 MEMS microphones application in portable devices

6.4.1 Smartphones

Multimicrophone trend in smartphones. A field of auditory research called CASA (computational auditory scene analysis) aims to mimic the intelligibility of the human ear to separate sound sources by using conventional digital signal processing principles. The classic example is the 'cocktail party effect': the human ear is able to hone in on a particular conversation and separate the desired conversation from others. From the perspective of speech intelligibility, these other conversations are considered 'babble' noise, an actual standard type of 'distractor' used to test phones equipped with voice processing.

In a single microphone system, monaural cues such as frequency pitch and onset time can be used to separate sound sources. But humans have two ears for a reason: the additional binaural information provides the brain with the ability to distinguish subtle differences in both time of arrival – inter-aural time difference (ITD) – and the level of audio signals arriving at each ear – inter-aural level difference (ILD). The principle of binaural processing in CASA relies on the interpretation of ITD and ILD cues to separate sound sources Fig. 6.26.

Fig. 6.25. Time and level differences between sounds arriving at both ears provide binaural cues
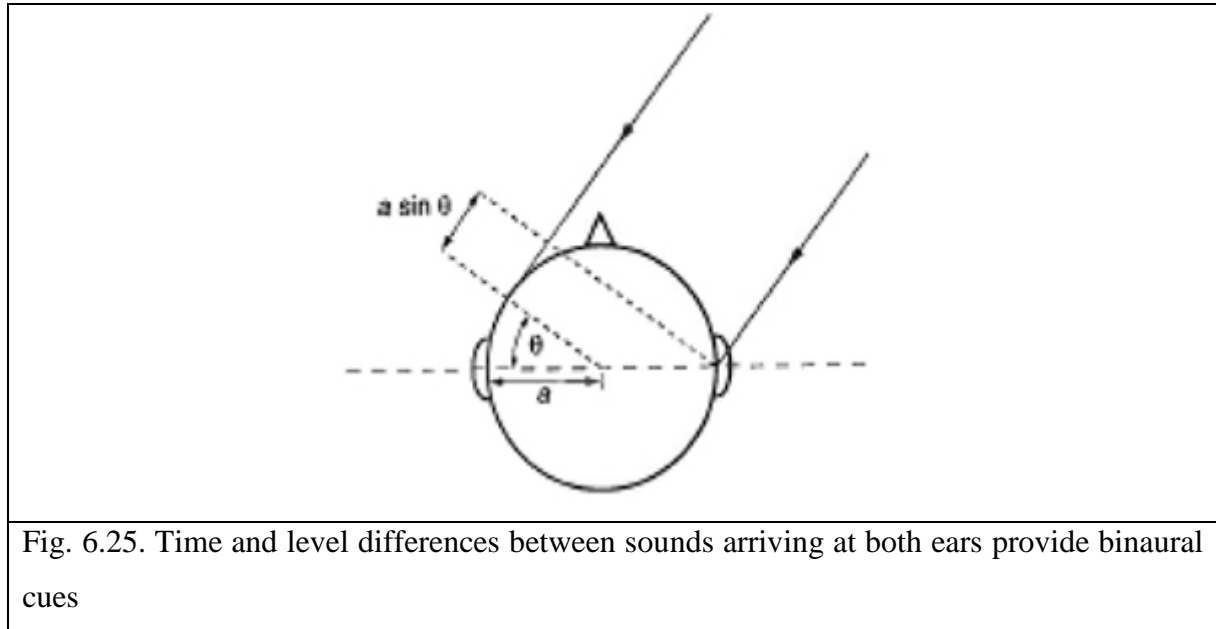
For human listeners, ITD and ILD provide cues in complementary frequency ranges, at least for the ideal scenario of a point sound source in a free-field such as an outdoors location or in an anechoic chamber (i.e., an enclosed space insulated from noise and echoes).

ILDs are most pronounced at frequencies above approximately 1.5 kHz because it is at these frequencies that a person's head is large compared to the wavelength of the incoming sound, thus producing a reflection.

On the other hand, ITDs exist at all frequencies. However, they can only be decoded unambiguously if the ITD is less than half the period of the wavelength at that frequency. For the spacing between ears on a human head, this leads to a maximum frequency of about 1.5 kHz, at which point we can use ITD. Note that since the 2 'ears' of a phone are closer together, we can actually increase this frequency range. This explains why it can be advantageous to have closely spaced microphones – we will come back to this later.

ILD, ITD, and other acoustic attributes such as pitch, harmonics, onset time, and time spacing of sound components are used to characterize, separate, and eventually spatially locate multiple sound sources. Following this sound separation, a classification or grouping of sound sources is made to distinguish between desired (talker) and undesired (distractor) sound sources. Finally, a 'voice isolation' stage identifies the sound of interest, eliminating noise and other audio to deliver clear voice.

Real-world implementation challenges. To arrive at a commercially viable product, developers have to implement these theoretical principles reliably in a real-world environment. For example, users can hold a phone in various positions, either in a 'close-talk' configuration with the phone held to the ear, or a 'far talk' speakerphone type of scenario with the phone held vertically in either portrait or landscape orientation in one's hand, resting horizontally on a table, or even resting in a cup holder inside a car.
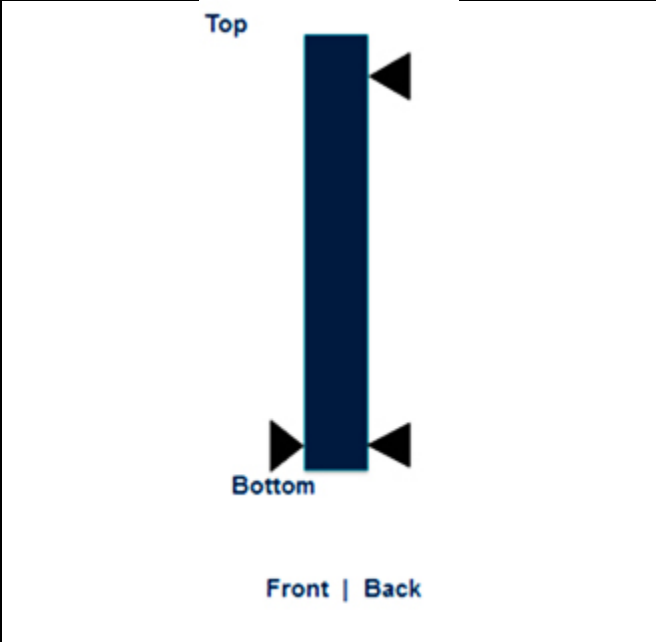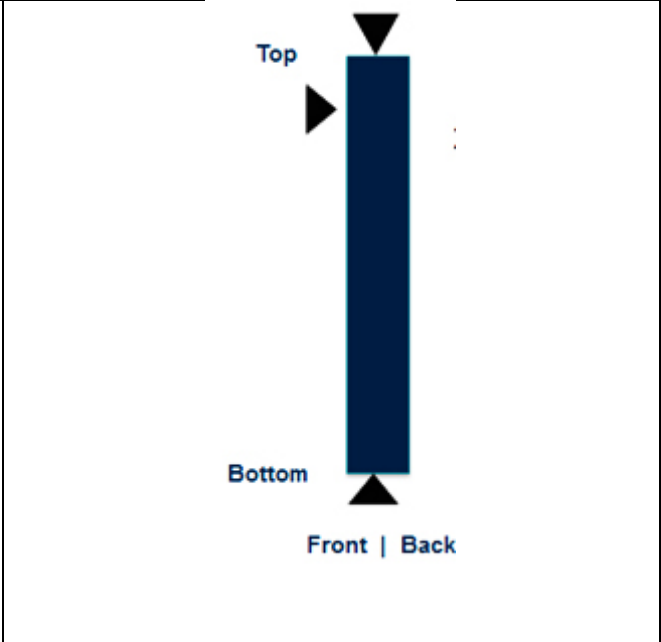
There is always a trade-off between the amount of noise suppression and the resulting voice quality. Therefore, it becomes challenging to have the phone produce high quality output in all scenarios when applying noise suppression. Furthermore, there are several other factors that affect the quality – and complexity – of an implementation, including room acoustics like reverberations that cause multi-path time-of-arrivals at different amplitudes, poor microphone sealing causing microphone crosstalk through the mechanical enclosure, and local pickup of the sound from the far-end into the near-end microphone through the speaker when in speakerphone mode (and thus driving the need for embedded acoustic echo cancellation).

For some binaural processing algorithms that rely on specific cues, mics may need to be spaced close together (CM: close mic) while for others it is acceptable to have a larger distance between mics (SM: spread mic). The correlation of sound between these microphones is important as well. For example, a particular algorithm may rely on one mic having a direct sound path while another mic receives a signal that undergoes a 'shadowing' effect in the path between the sound source and microphone.

Practical industrial design of today's smartphones places additional constraints on microphone placement. For instance, for a close-talk use case, it would be logical to position microphones at the bottom-front and bottom-back locations of the phone. This creates a CM pair as well as direct and shadow paths to the user's mouth in the so-called 'end-fire' configuration known

from phased array theory (**Fig. 26a**). However, bottom-back mic placement is restricted in many devices by the presence of other components (e.g., WiFi antenna, battery, etc.) or by other mechanical restrictions such as an OEM's reluctance to drill a hole in the back casing.

Instead of using such a back-bottom mic placement, many phones use a secondary mic at the top-edge of the unit (**Fig. 26b**).

Fig. 6.26a. Ideal microphone placement	Fig. 6.26b. *Realistic microphone placement*

When combined with the bottom mic, this creates an SM pair instead of a CM pair. It also requires any voice algorithms to be adapted. Furthermore, it becomes more challenging to obtain good separation between voice and distractor sounds as the 'end-fire' phased array approach cannot be used. Thus, there is a trade-off between industrial design, cost, and performance. It is the ability to maintain good performance under such non-ideal conditions that distinguishes one voice processing implementation from the other.

To still take advantage of the benefits of having a CM pair, high-end smartphones can place a third microphone at either the top-front or top-back location of the device. This third mic then forms a CM pair with the top-edge microphone. The voice processor can now select the optimal 2-mic pairing, depending on factors such as:

- Specific customer use case: close-talk, far-talk
- Phone orientation: horizontal/vertical, portrait/landscape
- Applications: narrowband voice call, wideband VOIP, multi-mic multimedia recording, etc.
- Phone placement: concealment of performance loss in case one of the mics is occluded or blocked

Today's next-generation voice processors enable enhanced performance by utilizing input from multiple pairings. For example, Audience's voice processor not only selects an appropriate 2-mic pair depending on these conditions but can also apply simultaneous 3-mic processing to further improve performance.

In the PC space, new 'convertible' designs with screens that can be rotated or flipped introduce a new set of factors that must be addressed. Specifically, users are no longer in a fixed position with respect to the PC microphones, as they are with a traditional notebook. In fact, the microphone(s) may be occluded when the screen is flipped over or the user holds the 'detachable' PC like a tablet.

With these types of form factors, legacy PC-based voice processing technology like beamforming runs into performance limitations. In addition, noise suppression will need to manage PC-related noise such as keyboard tapping and fan noise, as well as cancel loud music from nearby PC speakers. As outlined above, similar problems have already been resolved for mobile phones, and so PC-based voice processing is a natural evolution for this technology into applications like VOIP calls, multimedia recording, and speech recognition.
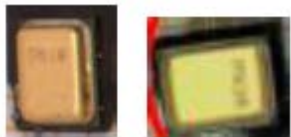
The problem of improving the capture sound and noise reduction can be solved by using multiple microphones. A typical smartphone has only one microphone, to record your voice and to record audio for videos. But nowadays sound technology is advancing.

Later dual microphones started appearing in smartphones (e.g. Nokia smartphones and iPhones) which had 1 microphone specifically for voice, and a second specifically for video audio recording as the type of microphone and type of sound recording required were different.

By 2005/6 technology had advanced to the point where stereo microphones were introduced to help record better audio in video (e.g. the Nokia N93 which came out in 2005/6 has stereo microphones for video recording).

Nowadays we have trend to have 3 microphones in a device. The iPhone 5 has 3 MEMS microphones. The 1st MEMS microphone is used at the bottom of the phone to capture voice. The 2nd MEMS microphone is located on the back and used to delete ambient noise and improve sound quality. The 3rd MEMS microphone (from ADI) is located beside the front camera in order to improve Siri's (voice control) performance. Using of MEMS microphones in different iPhones is shown on Fig. 6.27. Disassembled iPhone 5S is shown on Fig. 6.28. One can see position of front and rear MEMS microphones inside device. MEMS microphones inside iPhone 6 are shown on Fig. 6.29.

iPhone 4S	iPhone 5	iPhone 5s
2 mics from AAC and Knowles	3 mics from AAC, Knowles and ADI	3 mics from AAC and Knowles
a)	b)	c)

Fig. 6.27. MEMS microphones in smartphones: a) iPhone 4S; b) iPhone 5; c) iPhone 5s.

Front facing top microphone:
→ Presumably for FaceTime and speakerphone capabilities.

First Company A
MEMS Microphone

Rear facing top microphone:
→ Partners up with the rear-facing camera for recording video, and is also used for canceling out background noise when on calls.

Second Company A
MEMS Microphone
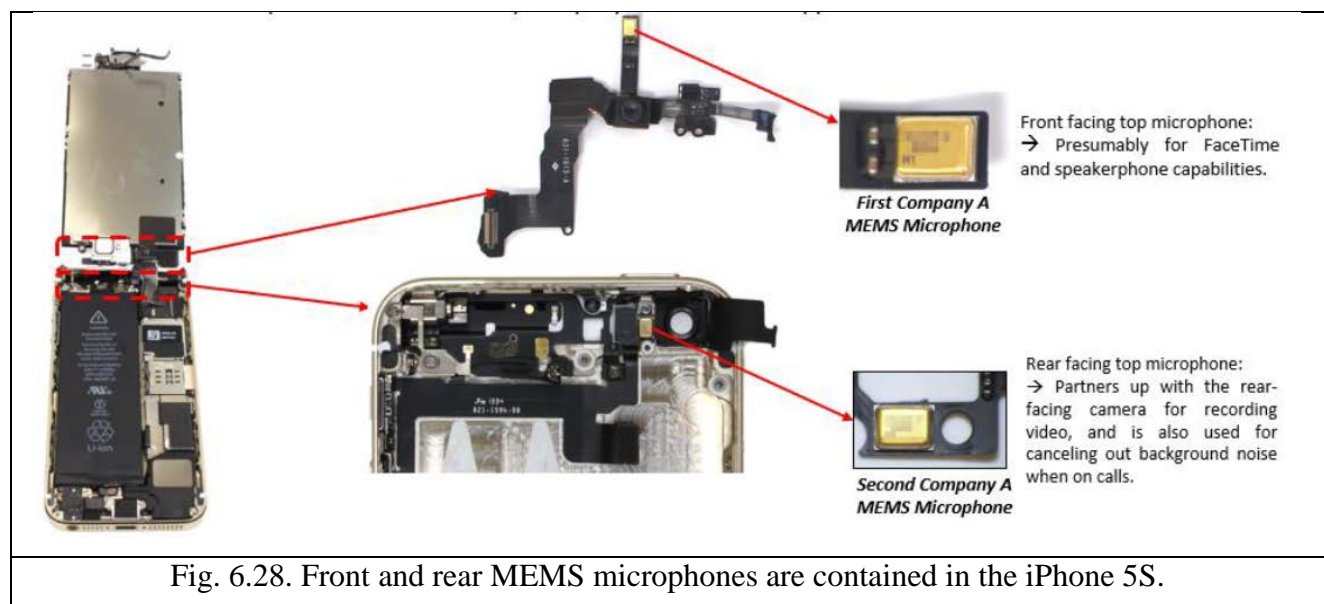
Fig. 6.28. Front and rear MEMS microphones are contained in the iPhone 5S.

Fig. 6.29. MEMS microphone inside iPhone 6.
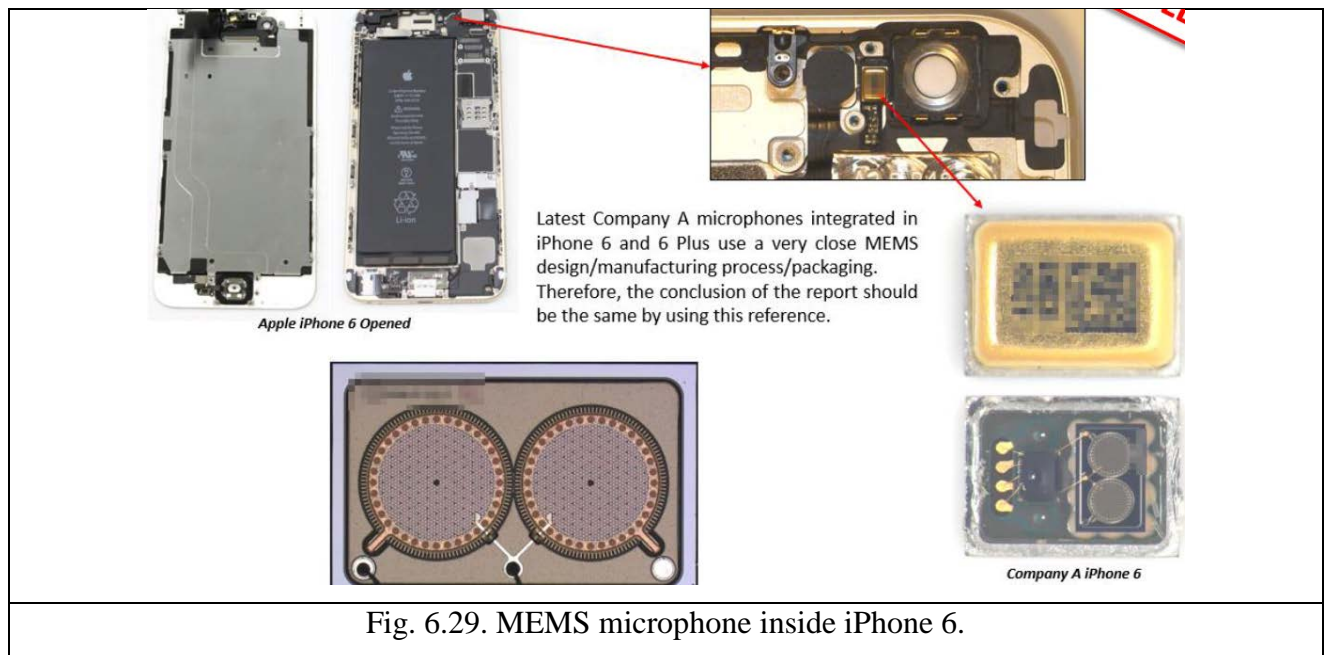
6.4.2 Hearing aids.

HEARING loss is one of the most common human impairments. It is estimated that by year 2015 more than 700 million people will suffer mild deafness. Most can be helped by hearing aid devices depending on the severity of their hearing loss. Contemporary hearing aid device realization requires a dual channel transmitter front end (TFE) for digital hearing aid (DHA) applications that use novel microelectromechanical-systems (MEMS) audio transducers and ultra-low power-scalable analog-to-digital converters (ADCs), which enable a very-low form factor, energy-efficient implementation for next-generation DHA.

The contribution of the design is the implementation of the dual channel MEMS microphones and power-scalable ADC system. The first generation of hearing aids usually consisted of analog variable gain amplifiers, electret microphones and speakers that compensated for hearing loss. They dissipated a considerable amount of power and had flat frequency characteristics that made these devices uncomfortable for most patients since hearing loss usually varies across different frequencies. The next generation of devices adopted analog filter banks in which band-pass filters were used in parallel to amplify the acoustic signal to a specific level in each different frequency band. This design, however, resulted in bulky devices that still required high power consumption. A major breakthrough was achieved through the development of DHAs that exploited the power of digital signal processors (DSPs) that allowed full programmability and customization to a patient's hearing characteristic.

A typical DHA system, shown in Fig. 6.30, consists of a Transmitter Front-End (TFE), DSP, and a Receiver Front-End (RFE). The TFE consists of the microphone, a variable gain amplifier (VGA) and an ADC. The RFE receives the processed digital signal from the DSP and converts it to the analog domain. At the backend, a speaker delivers the acoustic sound to excite the patient's eardrums. One of the major issues with existing DHAs is the rapid degradation of

performance in noisy environments in which the TFE becomes saturated due to the ambient acoustic content and background noise.
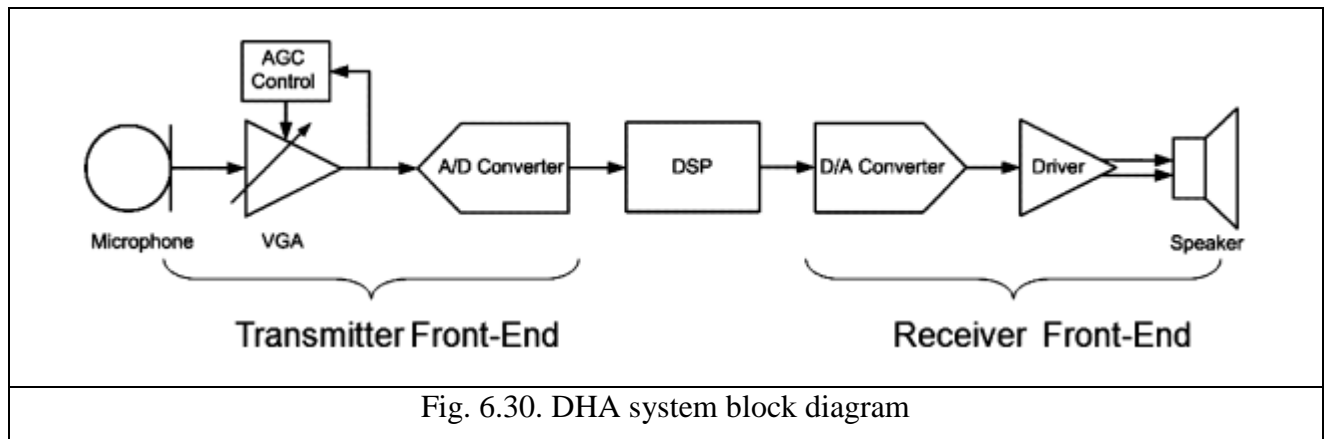
Fig. 6.30. DHA system block diagram

Background noise interferes with the desired conversation thereby impairing intelligibility. While the use of a very high dynamic range TFE can help relieving this problem, it comes at the expense of high power consumption and complexity. Recently developed DHAs also employ microphone arrays combined with adaptive array processing that improve audio quality and perception in real-life environments through noise cancellation mechanisms. Directional DHAs exploit the use of multiple microphone arrays (MMAs) to provide the patient with information on the spatial position of the desired acoustic source, while attenuating the ambient noise at the same time. MMAs apply adaptive beam forming techniques to estimate the signal direction and cancel ambient noise.

There are five basic types of hearing aids, and which ones patients choose depends on their level of hearing impairment, their physiology and health, and their lifestyle. Form factors of hearing aids are depicted on Fig. 6.31 and Fig. 6.32.

Behind-the-ear (BTE): These are the more traditional devices that consist of a durable plastic housing that rests behind the outer ear connected via a thin plastic tube that snakes into the ear canal. The tube carries sound from a receiver in the housing to the tympanic membrane in the ear, so there is no sound directionality. Its relatively large size lets it provide more amplification and larger batteries, so it is well-suited for those with severe hearing loss. They can also contain more processing power, so audiologists, the medical professionals who fit patients with hearing aids, have more options in tailoring the sound. BTE units can have amplification or gains of 35 to 82 dB.

In-the ear (ITE): These devices contain the microphone and speaker, as well as the battery and processor, and fit into the ear. Its size accommodates two microphones for directionality, as well as larger batteries, and makes it is easier to insert and remove for older persons whose hands aren't as nimble as they once were. ITE hearing aids provide 40 to 70 dB of gain.

In-the-canal (ITC): Smaller than ITE hearing aids, this fits directly into the canal. Being smaller makes is less noticeable, but it is only good for mild to moderate hearing loss. Users do, however,

get better sound localization because the microphone takes advantage of the outer ear's sound-reflecting properties. ITC devices can amplify sound by 35 to 50 dB. **Completely-in-the-canal (CIC):** The smallest of all hearing aids, CIC versions are almost invisible to others. There are also more compatible with talking on the phone and they, like ITC models, use the outer ear to reduce wind noise and help in sound localization. Its size limits battery capacity and how much amplification it can provide, but CIC devices do not need much amplification because they place the receiver so close to the tympanic membrane. They are uncomfortable for some patients due to their small ear canals and how deeply the HA must be inserted. It can also take several visits to the audiologist to get the device to remain comfortably in the ear. **Bone anchored hearing aids (BAHA):** Some patients suffering hearing loss have an occluded or non-existing ear canal, or their ear canal secretes too much wax. For these people, traditional hearing aids aren't an option. To get around their limitations, engineers have developed a hearing aid that picks up and amplifies sound, then sends it to the patient's hearing centers in the brain using bone conduction though a titanium anchor inserted in the skull just behind the ear.

Fig. 6.31.Styles of hearing aids Behind the ear (BTE), in the ear (ITE), in the canal (ITC), and completely in the canal (CIC).	Fig. 6.32. Ponto BAHA hearing aids.

Trends of developing of hearing aids are: miniaturization, battery work time increasing, smart algorithms of sound processing, wireless control with help of smartphone, sound source determination, synchronization of information between ears. Examples of hearing aids are presented on Fig. 6.32.
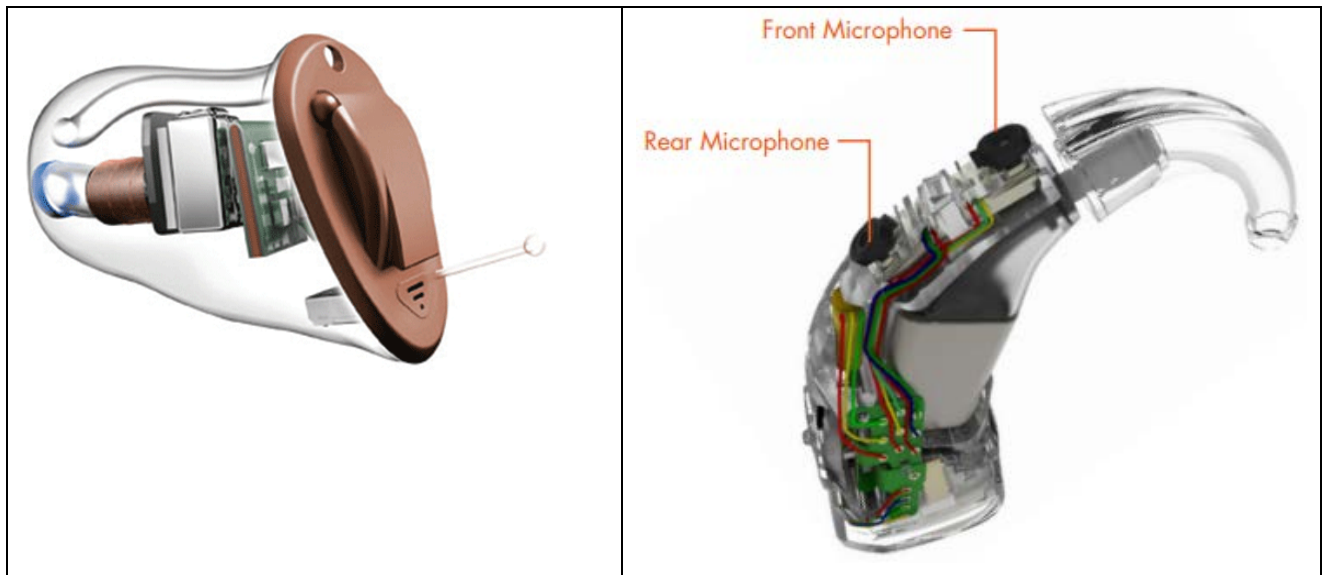
Fig. 6.32. a) hearing aid from Sivantos − device to employ wireless technology for exchanging data with another Insio in the user's other ear to synchronize settings and exchange audio signals. It can also be adjusted by the user through an iPhone or Android device; b) hearing aid with front and rear MEMS microphones.

Conventional electret microphones found in most of today's hearing aids use a polyester diaphragm that is held a few microns away from a metal back plate that holds an electrical charge. When sound hits the suspended diaphragm, it moves toward and away from the metal plate in sympathy with the energy of the sound waves. Because the two surfaces are electrically charged, the movement induces a fluctuating voltage. This change in voltage can be detected, amplified and processed by the hearing aid as the electrical equivalent of the original sound. The diaphragm and surround are weak links in the electret microphone. Selecting a material for the surround requires a number of compromises including, but not limited to, strength, thickness, flexibility, permeability to water vapor, and compatibility with the adhesives required to connect the components of the microphone together. Even the well-engineered components are susceptible to moisture, temperature changes and environmental contamination. Over time, these factors cause unpredictable changes in the sensitivity of the electret microphone, a phenomenon known as "drift."

Assuming there is no change to the directional inlets in a directional system over time, conventional single-diaphragm directional systems are negligibly affected by drift. However, drift spells dire and negative consequences for directional systems with multiple microphones, which are the very systems that fully adaptive directional arrays depend upon. Out of the factory, the microphones in a multi-microphone directional hearing aid are matched in sensitivity. Over time, drift causes a mismatch in sensitivity across the microphones, which then compromises the directional response of the system. As a result, hearing aid patients may report decreased performance in noise environments over time. In addition, because directional systems contain more microphone noise relative to omnidirectional arrays, the patient may report hearing additional

microphone noise if the system is set to a "directional" response. As a result, routine electroacoustic evaluation of microphone performance is recommended; although, without appropriate equipment, it can be a challenge to document directional performance.

Automatic-switching, fully adaptive, dual MEMS microphone directional system microphone technology that minimizes microphone drift was proposed Fig. 6.32b. MEMS microphones are produced from a silicon crystal that is more resistant to changes in humidity and temperature cycling than standard electret microphones. Regardless of the environmental conditions, MEMS microphones produce a stable and consistent response that reduces drift and subsequent microphone mismatch over time. The result is a multi-channel, fully adaptive, directional microphone array designed to provide predictable performance over time.

Studies conducted by Starkey Hearing Technologies and Knowles Electronics investigated long-term microphone performance after accelerated aging testing. These tests subjected hearing aids with MEMS and electret microphones to high humidity, salt fog and temperature cycling. Figure 2 shows the results of accelerated aging tests, which revealed that after extended exposure to high heat and humidity, the magnitude and variation of MEMS microphone drift was greatly reduced relative to the electret microphones. Nominal drift in MEMS microphones was consistent in direction and phase, minimizing mismatch between microphones. In other words, although drift occurred within dual-MEMS microphones, it was consistent across the two microphones such that mismatch was minimized. These test results suggest that repeated calibration of microphone sensitivity is unnecessary and that directional performance is consistent over the life of the hearing instrument with MEMS microphones.
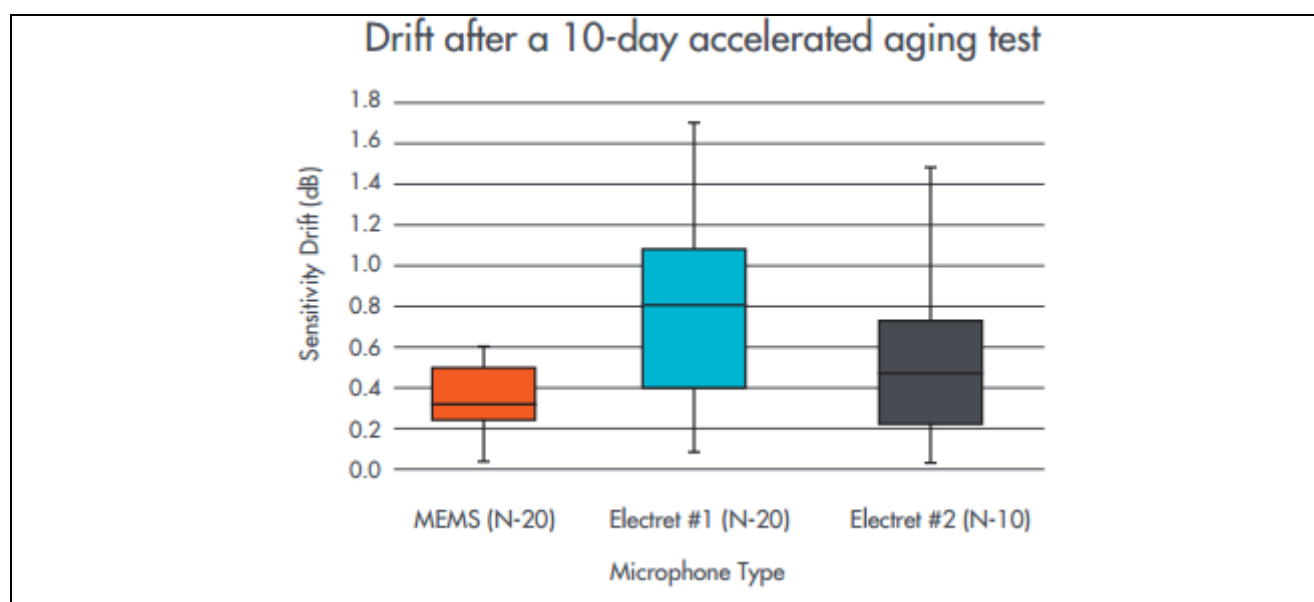
Fig. 6.33. Sensitivity drift is plotted as box-and-whisker plots after a 10-day accelerated aging test for MEMS microphones and two commonly used electret hearing aid microphones. For each box, the center horizontal line represents the median drift, and the top and bottom horizontal lines represent the interquartile range. Each whisker denotes the minimum and maximum drift per box.

6.4.3 MEMS micropones in laptops.

Dd

Speaker Component

Microphone Component

Speaker Component

Microphone Component

Microphone Module

Speaker Component

Microphone Component

MEMS Component
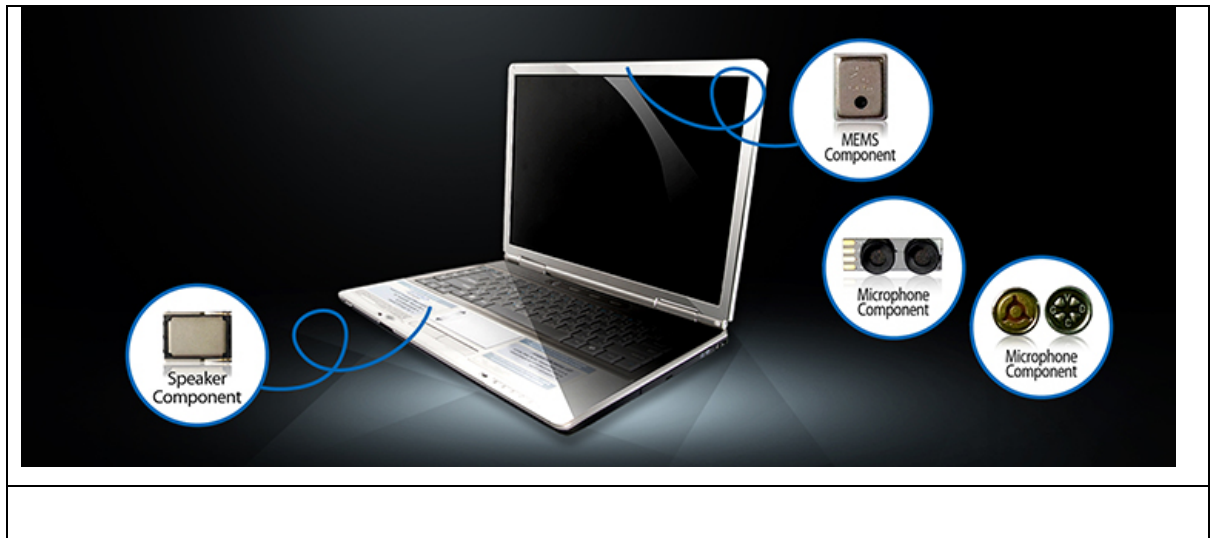
REFERENCE LIST

- A mems condenser microphone for consumer applications Doi: 10.1109/MEMSYS.2006.1627742

- A MEMS-Based Power-Scalable Hearing Aid Analog Front End

- Advances Multi-Channel Adaptive Directional Systems

4. MEMS based displays

4.1 Introduction

Ink and paper are arguably the de facto standard for information display. Developed over 5,000 years ago, today's inks and dyes provide lifelike color imagery. Display technologies, on the other hand, are relatively new. The CRT was developed less than 100 years ago and the increasingly popular flat-panel display less than 40 years ago. For roughly a decade now, engineers have been working to create a display technology capable of providing a paper-like reading experience, not only with regards to superior viewability, but also with respect to cost, power and ease of manufacture. Display technologies such as backlit LCDs, reflective LCDs, electroluminescent (EL) displays, organic light-emitting diodes (OLED) and electrophoretic displays (EPD) were all steps in this direction. mirasol displays, based on industry-proven MEMS technology, promise to take the quest for paper-like displays to a new level. A wide variety of display technologies are aiming to

capture the key characteristics of ink and paper. In this section we will compare them, with particular emphasis on energy consumption and readability.

4.1.1 Emissive/Transmissive Displays

Displays are classified as one of three types: emissive/transmissive, reflective or transflective. A transmissive LCD consists of two transmissive substrates between which the liquid-crystal material resides. By placing a backlight underneath one of the substrates and by applying a voltage to the liquid-crystal material the light reaching the observer can be modulated so as to make the display pixel appear bright or dark. A display can also directly emit light, as in the case of an OLED display, whose active display material emits light. In the case of an LCD, a constant source of power is required to both modulate the liquid-crystal material and to power the backlight. An LCD requires constant refreshing—at least sixty times per second—in order to prevent the liquid-crystal material from transitioning to a different modulation state, resulting in image degradation or flicker. Such is also the case with OLED and EPD—constant power must be provided to the light-emitting materials in order to prevent screen flicker.

4.1.2 Reflective Displays (Continuous refresh type)

In a reflective display, one of the substrates found in a transmissive display is replaced with a reflective substrate. Reflective displays usually employ liquid-crystal material on top of the reflective substrate so as to modulate the ambient light reflecting off the reflective substrate. Since there is no backlight in reflective displays, they consume substantially less power than emissive displays. However, since the material providing modulation is liquid-crystal, the majority of these types of displays must constantly be refreshed or the displayed image will be lost. So far, most portable devices employing reflective displays are the continuous refresh type.

4.1.3 Reflective Displays (Bistable type)

A bistable display is capable of maintaining one of two states (on or off) without any external influence such as an electric field. A bistable reflective display employing liquid-crystal material for light modulation is in many ways identical to the continuous-refresh reflective display. The key difference is the type of liquid-crystal material that is used. Through proper choice of chemistry, manufacturing and drive schemes, the liquid-crystal material can be locked into one of two states. Once the material has been locked into a certain configuration, it is not necessary for the display to be refreshed. In fact, power can be completely removed from the system and the display will maintain the last image shown. EPD and mirasol displays are also bistable. EPDs typically consist of charged microcapsules containing dye suspended between two substrates. The microcapsule, generally a sphere, is black on one half and white on the other. Depending on the electric field

applied between the two substrates, the microcapsule will flip orientation to position either the black or the white half toward the observer. Depending on the capsule orientation, the ambient light will either be reflected toward the observer or be absorbed. In a mirasol display, a flexible thin-film mirror is fabricated on a transparent substrate, leaving an air gap of a few hundred nanometers between the thin film and the substrate such that when ambient light enters this cavity and reflects off the thin-film mirror, it interferes with itself, producing a resonant color determined by the height of the cavity. A mirasol display produces iridescent color, similar to what you would observe in a butterfly's wings. Depending on the electric field applied between the substrate and the thin film, the film can be positioned in one of two states. Because mirasol displays are bistable, they don't require a refresh until the image is changed. As a result, they consume very little power, providing extended battery life for the user.

4.1.4 Transflective Displays

Transflective displays are a hybrid of emissive and reflective display technologies. Transflective displays were engineered to overcome the shortcomings of emissive displays, namely the backlight's high power consumption, and the shortcomings of reflective displays, such as poor image quality at low ambient light levels. Transflective displays employ a partially transmissive mirror as the secondary substrate, as well as a traditional backlight. In low light situations, the device operates as a transmissive display, employing the backlight. In high ambient light conditions, the backlight turns off and the display functions as a reflective display. A transflective display is a compromise and its image quality is generally subpar. In sunlight they are not as bright as purely reflective displays, while indoors they are not as bright as emissive displays. Regardless, they offer a compromise for applications where a wide variety of lighting conditions are seen and transflective displays are widely used in the portable device market.

Overview of IMOD Technology in mirasol Displays

MEMS-based display technologies have been under development for over a decade, but have only recently started to gain traction. Display systems based on arrays of movable mirrors are now widely available in the consumer marketplace. Deformable mirrors and mechanical shutters are also making use of MEMS-based displays. Their digital nature and fast response make them ideal for display applications. However, their role has been limited to applications with fixed-angle light sources rather than portable direct-view displays, as they are not effective when removed from a fixed-angle light source. Developed to address these shortcomings, mirasol displays are based on the principle of interference, which is used to determine the color of the reflected light. IMOD pixels are capable of switching speeds on the order of 10 microseconds. Additionally, mirasol

displays fabricated to use IMOD technology have shown reflectivities of greater than 60 percent, contrast ratios greater than 15:1 and drive voltages of as low as 5 volts. Though simple in structure, IMOD elements provide the functions of modulation, color selection and memory while eliminating active matrices, color filters and polarizers. The result is a high-performance display capable of active-matrix type functionality at passive-matrix cost. Qualcomm's mirasol displays are a strong contender in the display industry, with the potential to offer many of the benefits of ink and paper

Interferometric modulator display (IMOD)

Interferometric modulator display (IMOD, trademarked mirasol) is a technology used in electronic visual displays that can create various colors via interference of reflected light. The color is selected with an electrically switched light modulator comprising a microscopic cavity that is switched on and off using driver integrated circuits similar to those used to address liquid crystal displays (LCD). An IMOD based reflective flat panel display includes hundreds of thousands of individual IMOD elements each a microelectromechanical systems (MEMS)-based device. Display devices play a vital role in the sharing of information, and are used in our daily lives in a range of applications. CRT (cathode ray tube) displays ruled the display sector for more than 70 years. However, the push towards ever higher quality and more compact displays has resulted in a variety of new technologies being developed. Nanotechnology research has produced a number of advanced technologies for flat displays, enhancing their performance and quality whilst also considering environmental aspects such as resource conservation and energy use. These advancements have also led to novel features such as flexibility and foldability.

A number of competing technologies have evolved and each of these have their advantages and disadvantages. Interferometric modulator displays (IMOD), developed primarily by Qualcomm under the trademark Mirasol, is a display technology capable of creating a range of colors by simply controlling how light interferes with itself. Manipulation of the light is done by a MEMS-based device (micro-electromechanical system). A tiny electronic light modulator, which has a microscopic cavity that is switched with the help of driver-integrated circuits similar to those used for addressing liquid crystal displays (LCDs), creates interference patterns which produce different apparent wavelengths of light. An IMOD-based reflective flat panel display comprises a large number of single IMOD elements, each of which is a microelectromechanical systems (MEMS)-based device.

4.2 Principlles of work

The basic elements of an IMODbased display are microscopic devices that act essentially as mirrors that can be switched on or off individually. Each of these elements reflects only one exact wavelength of light (such as a specific hue of red, green or blue) when turned on, and absorb light (appear black) when off. Elements are organised into a rectangular array in order to produce a display screen An array of elements that all reflect the same color when turned on produces a monochromatic display, for example black and red (in this example using IMOD elements that reflect red light when "on"). As each element reflects only a certain amount of light, grouping several elements of the same color together as subpixels allows different brightness levels for a pixel based on how many elements are reflective at a particular time.

Multiple color displays are created by using subpixels, each designed to reflect a specific different color. Multiple elements of each color are generally used to both give more combinations of displayable color (by mixing the reflected colors) and to balance the overall brightness of the pixel.

Because elements only use power in order to switch between on and off states (no power is needed to reflect or absorb light hitting the display once the element is either reflecting or absorbing), IMOD-based displays potentially use much less power than displays that generate light and/or need constant power to keep pixels in a particular state. Being a reflective display, they require an external light source (such as daylight or a lamp) to be readable, just like paper or other electronic paper technologies.

4.2.1 IMOD Display Technology:

Working prototypes of low-powered interferometric modulation screens were first unveiled at CES 2010. It is believed that Qualcomm's Mirasol IMOD displays will succeed the monochromatic "E-Ink" screens that presently dominate the e-reader industry. The IMOD display uses just 1 mW whereas an ebook size TFT LCD screen needs up to 10 W of power. This is mainly because the IMOD does not need a backlight for viewing. The IMOD can be easily viewed in direct sunlight as it uses the surrounding light instead of a backlight for illumination. IMOD displays can handle videos at a refresh rate of up to 15 fps. Though this may not be ideal for high-quality video, the switching speed is sufficient to make the display video capable, with no motion-blur effects.
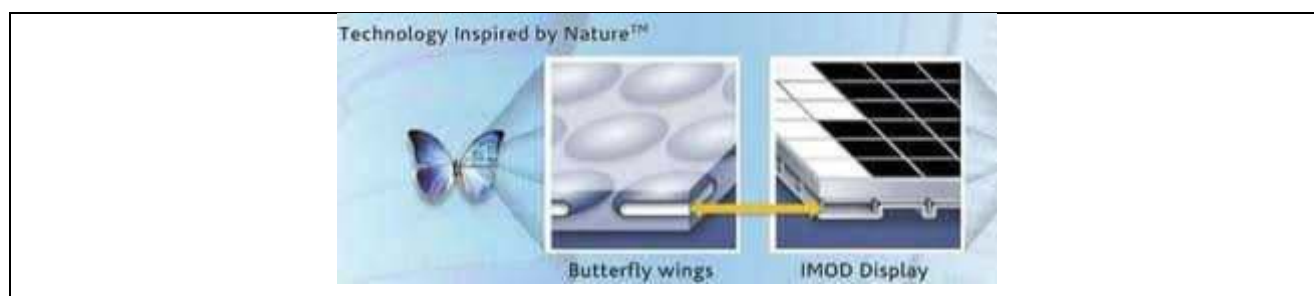
Figure 4.1 IMOD matrix

4.2.2 IMOD Displays Work

The Mirasol display from Qualcomm is the first in the industry to use interferometric modulation (IMOD). The IMOD element, which is the core building block of Mirasol display technology includes two stable states. When there is no voltage, separation of the plates occurs and the light striking the substrate is reflected. On application of a small voltage, electrostatic attraction pulls the plates together and light is absorbed, causing the element to turn black. This bistable nature is key to mirasol display's excellent energy efficiency. Fig 3.1 Basically, a pixel in a Mirasol display is an optically resonant cavity.
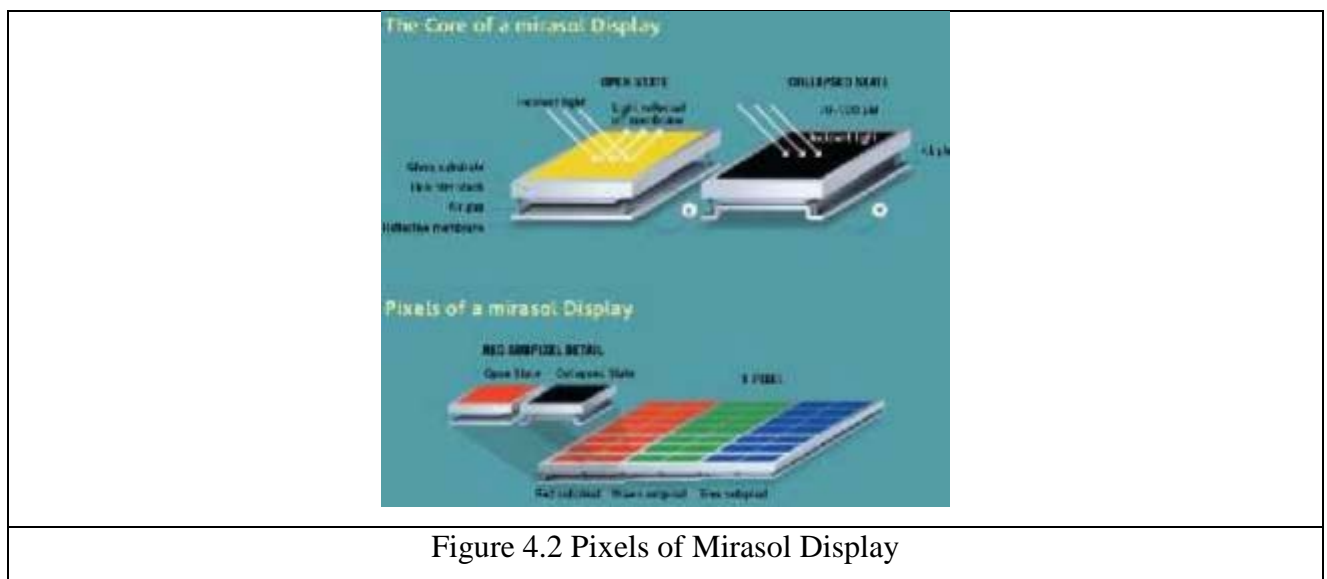
Figure 4.2 Pixels of Mirasol Display

The device includes a deformable reflective membrane, which is self-supporting and a thin-film stack, each of which behave as a single mirror of an optically resonant cavity with both residing on a transparent substrate. Striking of ambient light on the structure causes it to be reflected both off the top of the thin-film stack and also off the reflective membrane.
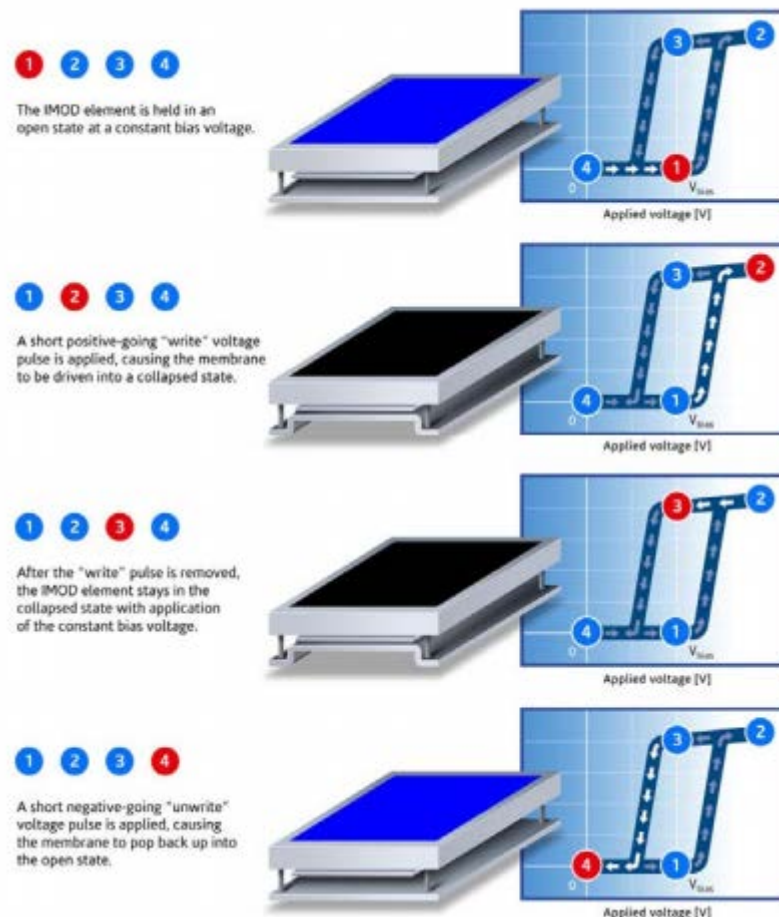
Figure 4.3 Hysteresis Effect in an IMOD Pixel

Based on the optical cavity height, light of specific wavelengths that bounce off the membrane will be slightly out-of-phase with the light that reflects off the thin-film structure. Some wavelengths constructively interfere and some destructively based on the phase difference. A color is perceived by the human eye as certain wavelengths will be magnified with respect to others. The mirasol display image is capable of switching between black and the selected color by changing the state of the membrane. This is achieved by the application of voltage to an electrically conducting thin film stack protected by an insulating layer. The application of voltage that results in electrostatic forces causes collapse of the membrane. The change in the optical cavity causes constructive interference at ultraviolet wavelengths, which cannot be seen by the human eye. Hence the image seen on the screen looks black. The IMOD elements that reflect in the RGB wavelengths are spatially ordered and assembled to form a full-color display

Pixel Information:

A pixel in an IMOD-based display consists of one or more subpixels that are individual microscopic interferometric cavities similar in operation to Fabry–Pérot interferometers (etalons), and the scales in butterfly wings. While a simple etalon consists of two half-silvered mirrors, an IMOD comprises a reflective membrane which can move in relation to a semitransparent thin film

stack. With an air gap defined within this cavity, the IMOD behaves like an optically resonant structure whose reflected color is determined by the size of the airgap. Application of a voltage to the IMOD creates electrostatic forces which bring the membrane into contact with the thin film stack. When this happens the behavior of the IMOD changes to that of an induced absorber. The consequence is that almost all incident light is absorbed and no colors are reflected. It is this binary operation that is the basis for the IMOD's application in reflective flat panel displays. Since the display utilizes light from ambient sources, the display's brightness increases in high **ambient** environments (i.e. sunlight). In contrast, a back-lit LCD display suffers from incident light

For a practical RGB color model (RGB) display, a single RGB pixel is built from several subpixels, because the brightness of a monochromatic pixel is not adjusted.

A monochromatic array of subpixels represents different brightness levels for each color, and for each pixel, there are three such arrays: red, green and blue.

Figure 4.4 RGB Mirasole pixel
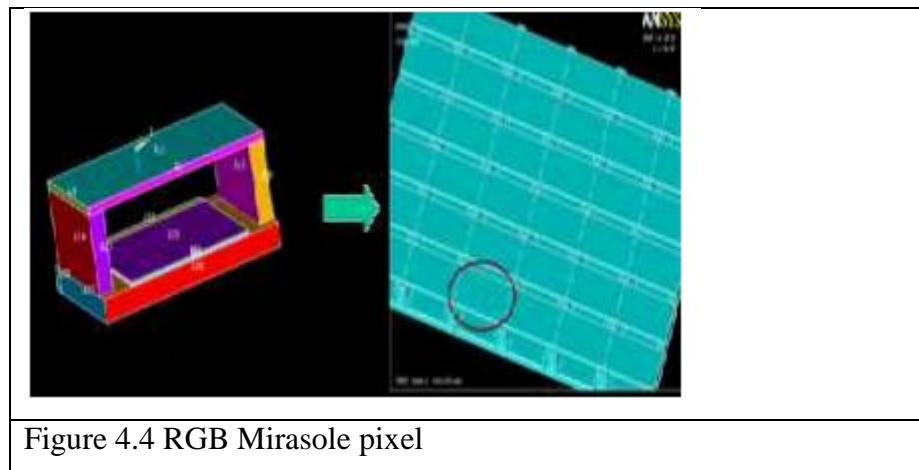
Competing Display Technologies:

The use of OLEDs in displays has been restricted by the cost and scalability of production for large screens, although OLED remains a significant competitor in the smartphone display market. Carbon nanotube-based displays have also been investigated. Since the chemical nature of carbon nanotubes (CNTs) is quite different to other carbon forms, the toxicity of CNTs must be studied carefully. CNT synthesis is challenging to control, making it quite expensive. Furthermore, CNT processing for displays need uniform dispersion and CNT mats result in material wastage. Reports suggests that Qualcomm will be investing around $120 million to obtain a 5% stake in struggling consumer electronics and liquid crystal display manufacturer Sharp.
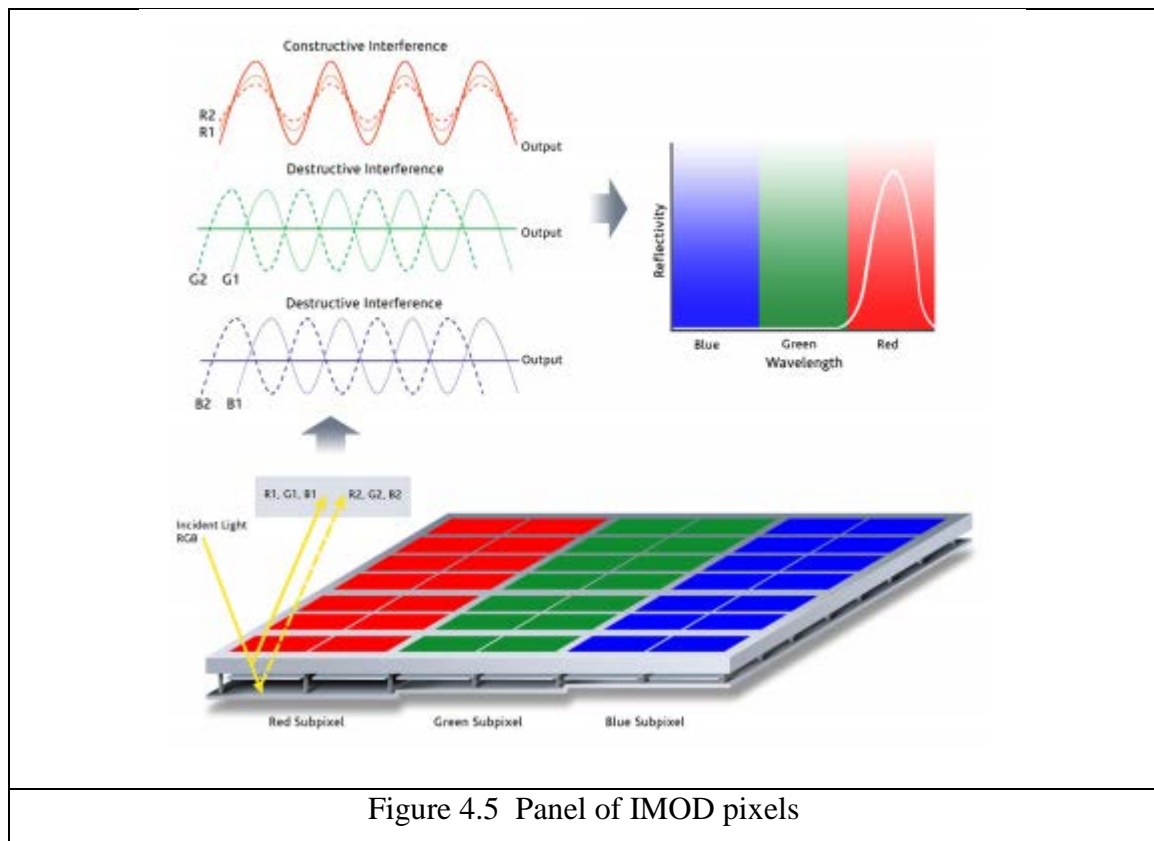
Figure 4.5 Panel of IMOD pixels

The proposed equity investment will be utilized for the expansion of a contract between Qualcomm subsidiary Pixtronix and Sharp for the development and commercialization of MEMS-based displays that can be manufactured with Sharp LCD production lines. PerfectLight is another MEMS-based display from Pixtronix that is based on a digital micro shutter that modulates light from an RGB LED backlight. It has a high switching speed that makes it useful for high quality video, and according to Pixtronix, the display was capable of viewing angles over 170°, 24-bit color depth and more than 3,000:1 contrast ratio at one quarter of the power consumption of LCDs of similar size and resolution.

Micromechanical display uses interferometric modulation Iridigm Display Corp. (San Francisco, CA) has developed a new way of using micromechanics that may eventually provide an alternative to liquid crystal displays (LCDs). The company says that it has already demonstrated reflectivities of more than 80 percent using Digital Paper technology, and that contrast ratios of 12:1, viewing angles of ±60 deg., and drive voltages <5 V have been achieved. In addition, the new displays can be made using fabrication processes that have already been developed for the LCD industry. According to Iridigm, major companies have already expressed interest in using the new technology, including Palm, Motorola, and Handspring. The basic component of the new display is the Interferometric modulator, or IMod.[1] Previous micromechanical devices have used interference,[2] but as the result of diffraction from surface relieve structures. This is not unlike the way embossed holograms (such as those on credit cards) work. To combine the desired diffractive orders and eliminate unwanted colors, these systems require a relatively sophisticated optical system. Plus, the angle of view is narrow. For these reasons, development of this technology has

been geared around the projection display market where the cost of the optical system and the geometrical restriction can be accommodated

Figure 4.6 Construction of pixel
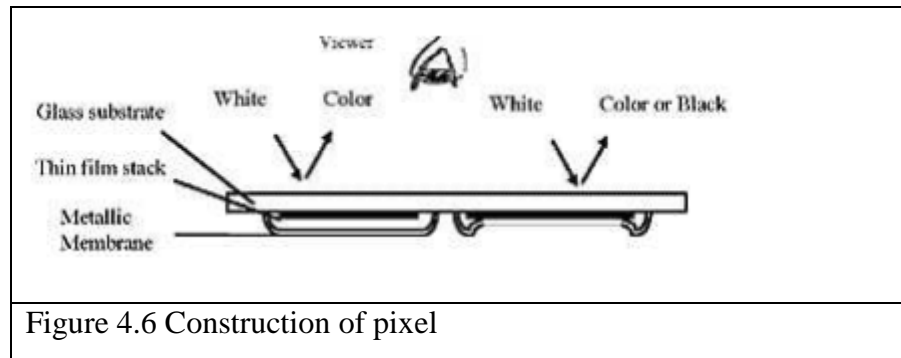
The color reflected by the Interferometric modulator (IMod) device is set by the optical distance between the thinfilm stack and the metallic membrane. The thickness of a protective insulating layer between the two determines the color when the two mirrors are in contact (right). To set the color reflected when the two surfaces are separated (left), the designer must calculate the additional distance required and set the height of the membrane accordingly.

The IMods, on the other hand, use interference in a way that has much more in common with reflection holography (Figure 1). The device consists of a conducting metallic membrane, a thin-film stack protected by an insulating layer, and a transparent substrate (in this case, glass). The stack and metal membrane act as two mirrors in an optical resonator. The optical distance between them determines which wavelengths constructively interfere with each other -- and so are reflected by the device -- and which destructively interfere and are absorbed.

Where no current is flowing, the optical distance between the two mirrors is determined by the thickness and refractive index of the insulator, and the height of the membrane. When the device is switched, however, the current causes the membrane to be electrostatically attracted to the insulator, stack, and substrate. When it makes contact, the optical distance drops to that through the insulator, and the color changes. By using this method, some of the problems inherent to diffractive systems are immediately solved. In particular, there is almost no "rainbow effect," where the color changes with viewing angle.

Development Of IMOD:

The IMOD technology was invented by Mark W. Miles, a MEMS researcher and founder of Etalon, Inc., and (co-founder) of Iridigm Display Corporation. Qualcomm took over the development of this technology after its acquisition of Iridigm in 2004, and subsequently formed Qualcomm MEMS Technologies (QMT). Qualcomm has allowed commercialization of the technology under the trademark name "mirasol", and this energy-efficient, biomimetic technology sees application and use in portable electronics such as e-book readers and mobile phones The restoration force of the membrane is linear whereas the electrostatic force opposing it is nonlinear. As a result, if the voltage is kept at the correct level (here, somewhere between 3 and 5 V) the

device only needs to be switched when its state is actually required to change; it will then maintain that state until it is switched again. This bistable operation means that the device has inherently low power requirements. Iridigm has produced a number of different test displays using their first-generation design, including a Palm-Pilot-sized screen. In these systems, tens or hundreds of 40- X 30-µm IMods were used within each pixel.Though most of the demonstrations so far have been monochromatic, the company says that, using conventional RGB additive color mixing, full-color displays are a straightforward extension of the existing technology
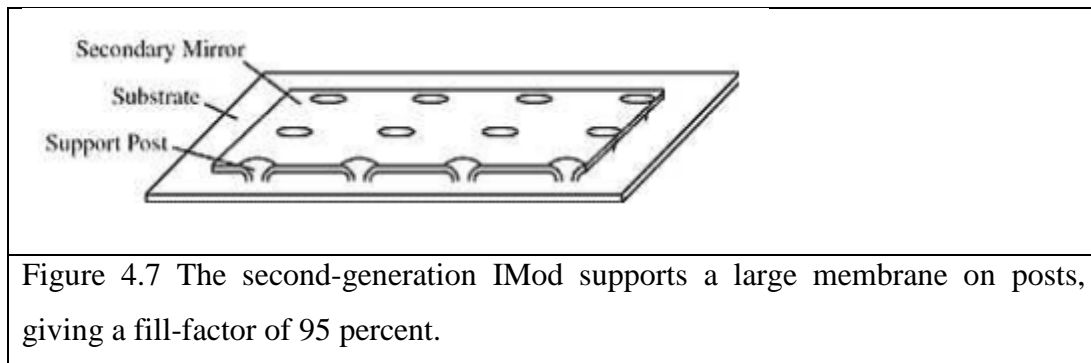
Figure 4.7 The second-generation IMod supports a large membrane on posts, giving a fill-factor of 95 percent.

Among the improvements that Iridigm has been working on is its second-generation design (Figure 4.7), which has already been fabricated and tested. Here, a large membrane is raised up using a number of support posts, rather than having lots of small, separate devices. The new design not only supports a fill factor of up to 95 percent, but it has also has good long-term behavior. In tests, it has been shown to be reliable for a number of switch cycles equivalent to using a Palm Pilot four hours a day for 18 years.

.Uses of IMOD: Future IMOD panels manufacturers include, Qualcomm in conjunction with Foxlink, having established a joint-venture with Sollink in 2009 with a future facility dedicated to manufacturing IMOD panels. Production for this began in Jan 2011, with the fabricated panels intended for devices such as ereaders

IMOD displays are now available in the commercial marketplace   □ QMT
IMOD technology, are found in the Acoustic Research ARWH1 Stereo Bluetooth headset device.   □
the Showcare Monitoring system (Korea), theHisense C108 mp3 applications from Freestyle Audio
and Skullcandy.   □ In the
Comp have announced phones with mirasol displays, and LG claims to be developing "one or
more" handsets using mirasol technology.   □ These
other) "bichromic" displays.

Advantages of IMOD Displays (Mirasol)

Bistability: This nature inspired MEMS-based innovation is bistable. This allows nearzero power usage in situations where the display image is unchanged. This bistability derived from the inherent hysteresis of the material causes considerable power savings, especially compared to displays that continually refresh, such as LCDs. [3, 4]

Speed

Since visible light wavelengths operate on the nanometer scale (i.e., 380nm to 780nm), the deformable IMOD membrane only has to move a short distance—a few hundred nanometers— in order to switch between two colors. This switching happens extremely fast, on the order of tens of microseconds. This switching speed directly translates to a video rate-capable display with no motion-blur effects. Traditional STN- or cholesteric-based passive matrix displays have switching speeds as slow as tens or hundreds of milliseconds. An IMOD element's switching time is 1000 times faster than traditional displays. In addition, switching speed of mirasol displays is maintained across a wide temperature range, unlike organic liquid-crystal-based displays, whose switching speeds decrease as temperatures go into low environmental ranges.

Readability

Humans view the world by sensing the light reflecting from various surfaces. As a result, a reflected image from a newspaper is more appealing and easier to view for the human eye, compared to a backlit image. Based on human perception, there are two critical factors which determine readability: luminance and contrast. Luminance is the amount of light that reaches the human eye. In the case of a reflective display, it is the amount of ambient light that is reflected from the display, rather than absorbed. The key metric is the reflectivity of the display's white state, which is measured by comparing it to the reflectivity of a standard white source. A white sheet of paper measures between 70 and 90 percent reflectivity, and a newspaper measures on the order of 60 percent reflectivity. Contrast is the ratio of the display's white state reflectivity to its dark state. This metric dictates whether or not the human eye will be able to perceive transitions between the dark and light areas on the display, which translates to spatial detail. If the contrast is too low, the display will appear washed out and the user will have difficulty perceiving image details. A high contrast ratio makes the image look sharper and improves readability.

For reference, a newspaper has a contrast ratio of approximately 4:1. Comparing the readability of reflective displays to that of emissive displays, it is clear that emissive displays work well at low ambient light levels. The problem with these displays, however, is when ambient light levels increase from room lighting to levels found outdoors on a sunny day, making it difficult for the user to discern spatial detail as shown in Figure 5. This is illustrated by the fact that a user must typically shield their portable-device screen when they are outdoors in bright sunlight. Two factors account for this: first, the increase in light that is reflected from the device pixel in the black state and second, the ambient light exceeding the light levels being emitted from the display. Both of these factors reduce the display's contrast. In the case of reflective displays, the black state suffers from

the same problem as emissive displays—the black-level luminance increases as ambient light levels increase. However, the display's white state offers superior viewability. As ambient light levels increase, so does the mirasol display's white-state reflectivity. As a result, a mirasol display offers a superior contrast ratio in brightly lit environments. In darker environments, supplemental illumination is provided by a low-power frontlight.

Figure 4.8 Brightness/Contrast Versus Ambient Light Levels
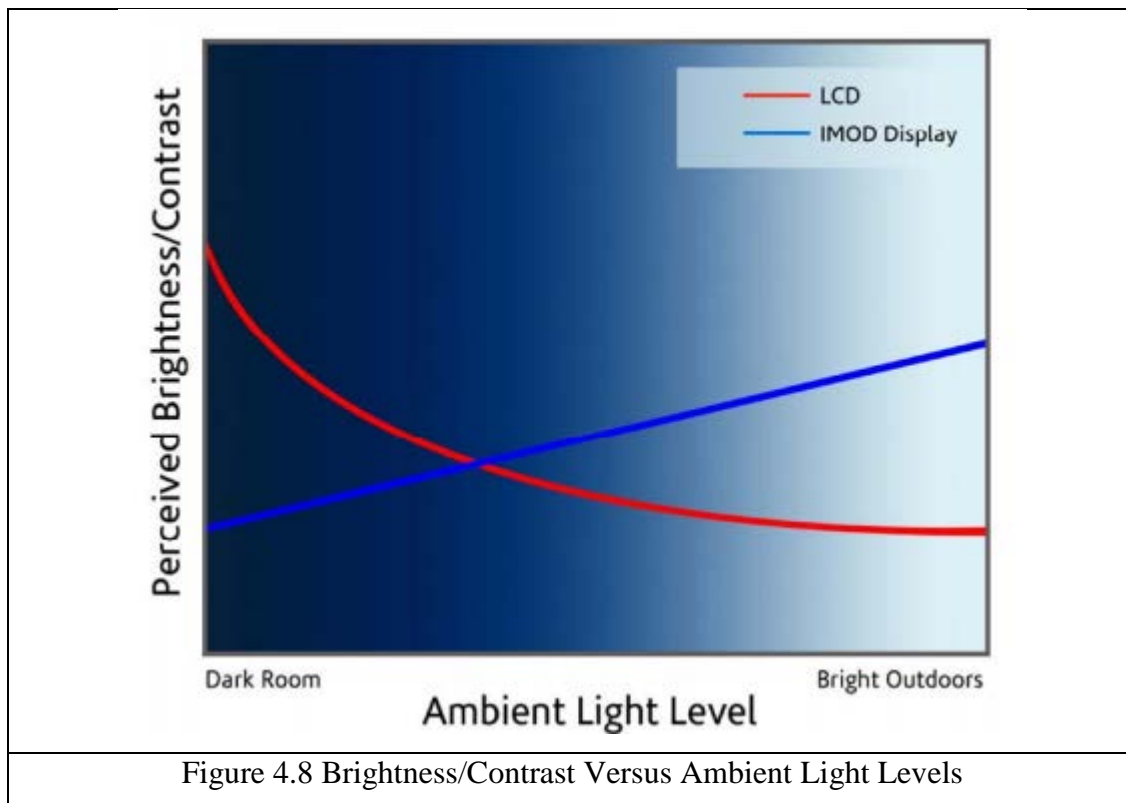
An additional benefit of mirasol displays is their wide viewing angle. Unlike an LCD display, which exhibits grayscale inversion when viewed at angles varying in elevation from normal (looking directly at the display, head-on), the mirasol display shows a non-grayscale-inverted image. Images shown on mirasol displays are also impervious to rotations around the normal, once again unlike LCD-based displays. In this sense, the IMOD element provides the benefit of an emissive—a wide symmetrical viewing angle. Qualcomm's mirasol displays offer reflectivity on the order of 50 percent and contrast ratios greater than 8:1. By comparison, the Wall Street Journal newspaper offers a reflectivity of 60 percent and a contrast ratio of around 4:1.

Ease of Manufacture

The MEMS elements that constitute a mirasol display have been designed for ease of manufacture. Qualcomm's mirasol displays are produced using a process known as surface micro- machining, which is derived from the wafer scale roots of MEMS fabrication. The name refers to the idea of

building all of the structure and components of the MEMS device on the surface of the underlying substrate. In the case of mirasol displays, these comprise an array of deposited metal and metal-oxide films which are lithographically patterned to produce a microscopic planar structure. The result is a monolithic electro-optic display which requires fewer process steps to build than the TFT array in a LCD. The overarching manufacturing benefit of mirasol display production is that the process was engineered to utilize infrastructure already in place in FPD fabs. All of the materials used for mirasol display fabrication currently exist within the FPD palette and, in most cases, substitute materials may be utilized. The end result is a flexible and robust process that enables conversion, with minimal modification, of many FPD fabs into mirasol display foundries, minimizing the time needed to bring IMOD technology to market.

Robustness

The biggest problem with LCD lifetime is the use of organic materials. Both the liquid-crystal material and the alignment material are organic and, as a result, break down over time when exposed to high temperature and light (both artificial and sunlight.) By relying on inorganic materials, mirasol displays are capable of performing over an extended temperature range and at the same time are impervious to high-intensity visible and UV radiation. Additionally, even when exposed to extreme temperatures, a mirasol display's response times are unaffected, and the impact on drive voltages and image content is minimal. Mechanically, the IMOD element is extremely robust. mirasol displays have demonstrated reliability over 12 billion cycles.

Industry Compatibility

Qualcomm's mirasol display are designed to conform to industry standards, enabling mirasol modules to be "plug-and-play" compatible with standard mobile systems. The mirasol module offers standard industry interfaces, and standard power supplies. This use of industry standards ensures that there is minimal risk to adopting this powerful new technology ensures that there is minimal risk to adopting this powerful new technology will allow for vastly improved performance and highly differentiated products to end users.

5. Inkjet printers (piezoelectrics or thermal bubble ejection to deposit ink on paper).

5.1 Introduction

Microdroplet generators are becoming an important research area in MEMS, not only because of their historically valuable marketing device, the ink-jet printhead, but also because of many other emerging applications in precise- or micro-amount fluidic control. Recently, with the emerging applications in biomedicine, fuel injection, chemistry, pharmaceuticals, electronic fabrication, microoptical devices, IC cooling, and solid free form, much research has focused on microdroplet generators. microdroplet generators are defined as devices that generate microsized droplets in a controllable manner; that is, droplet size and number can be controlled and counted accurately.

Microdroplet generators usually employ mechanical actuation to generate high pressure that overcomes liquid surface tension and viscous force and permits droplet ejection. Depending on the droplet size, the applied pressure is usually higher than several atmospheres. The operation principles, structure/process designs, and materials often play key roles in the performance of droplet generators. In addition to the well-known ink-jet printing, applications for microdroplet generators cover a wide spectrum in various fields such as direct writing, fuel injection, solid free form, solar

cell fabrication, LEPD fabrication, packaging, microoptical components, particle sorting, microdosage, plasma spraying, drug screening/delivery/dosage, micropropulsion, integrated circuit cooling, and chemical deposition. Many of these applications may become key technologies for integrated microsystems in the near future.

Fig. Types of inkjet technology

Mechanisms of drop formation
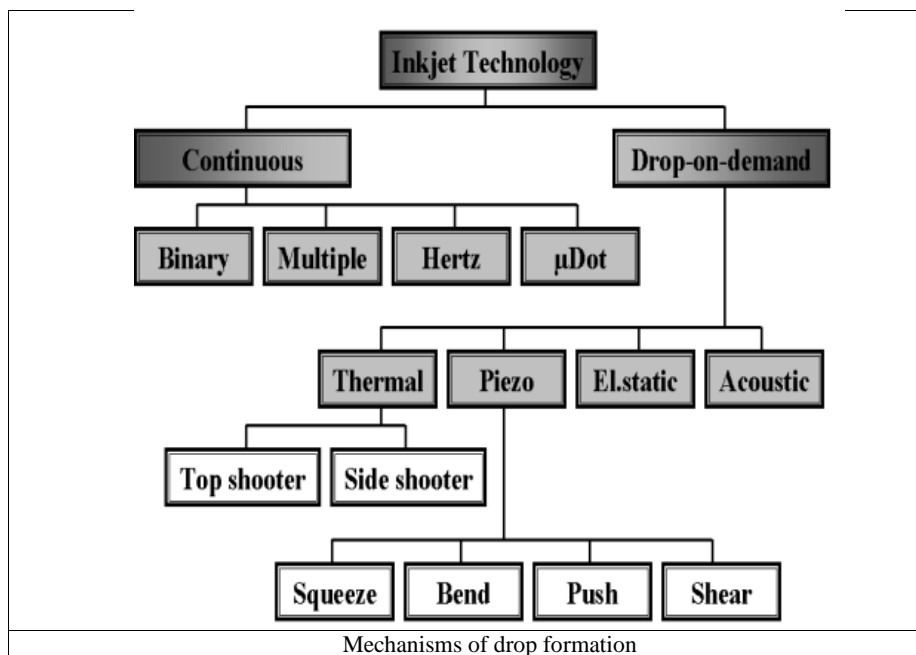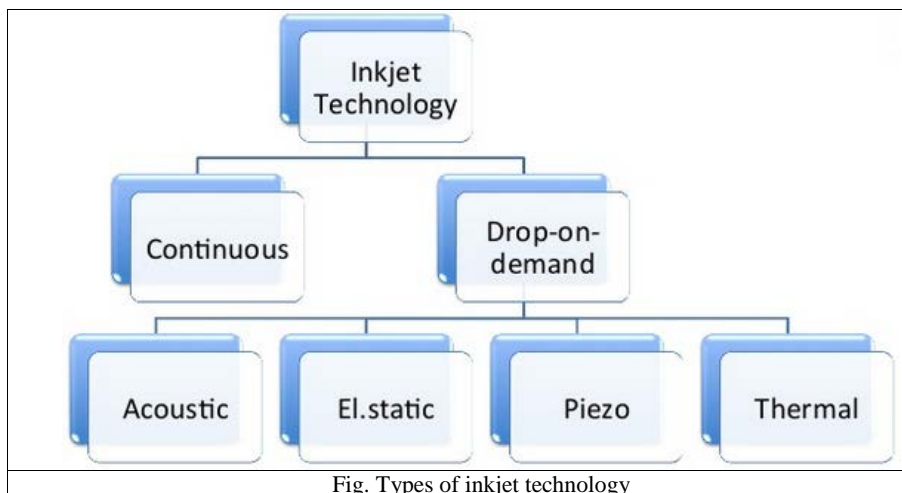
9.5 Operation Principles

9.5.1 Pneumatic Actuation

Spray nozzles are presently one of the most commonly used devices for generating droplets, as in airbrushes or sprayers. Two types of spray nozzles are shown in Figure 9.1 and Figure 9.2. Figure 9.1 shows an airbrush that lowers pressure at the outer area of the capillary tube by blowing air across the tube's end; this forces the liquid to move out of the tube and form droplets.
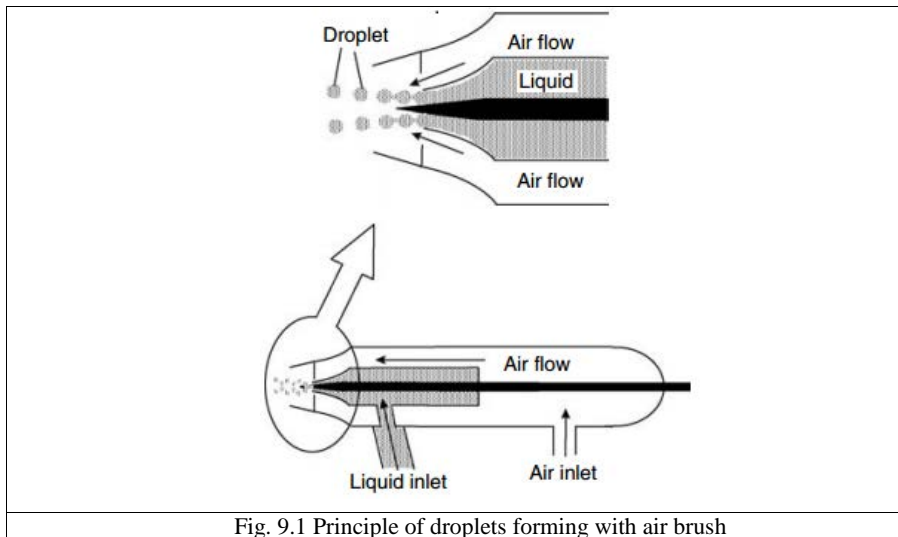
Fig. 9.1 Principle of droplets forming with air brush

The second device, a sprayer shown in Fig. 9.2, employs high pressure to push liquid through a small nozzle to form droplets. Typical sizes of the droplets generated by spray nozzles range from around tens to hundreds of μm in diameter. These devices can be fabricated in micro sizes by micromachining technology. However, controlling individual nozzles in an array format is difficult.
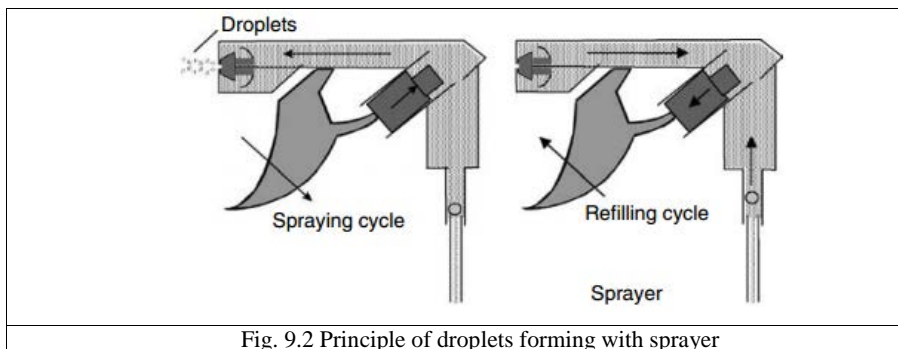
Fig. 9.2 Principle of droplets forming with sprayer

9.5.2 Piezoelectric Actuation

Droplet ejection can be based on piezoelectric actuation. Two types of droplet ejection devices are based on piezoelectric technology. One is the continuous ink-jet[]. Figure 9.3 schematically shows the operational principle of this type of ink-jet. Conductive ink is forced out of the nozzles by

pressure. The jet would break up continuously into droplets with random sizes and spacing, but uniform size and spacing of the droplets is controlled by applying an ultrasonic wave at a fixed frequency to the ink through a piezoelectric transducer. The continuously generated droplets pass through a charge plate, and only the desired droplets are charged by the electric field and deflected to print out, while the undesired droplets are collected by a gutter and recycled. One piezoelectric transducer can support multiple nozzles, so the nozzle spacing can be as small as desired for high-resolution arrays. The complexity of this device's droplet charging and collecting system is the major obstacle to its practical application.
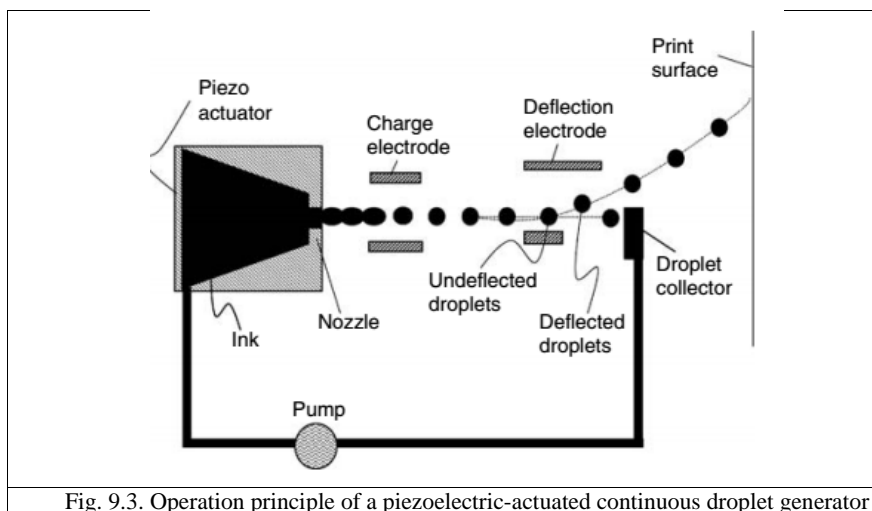
Fig. 9.3. Operation principle of a piezoelectric-actuated continuous droplet generator

The other device, the droplet-on-demand ink-jet, utilizes a piezoelectric tube or disc for droplet ejection only when printing a spot is desired []. Figure 9.4 shows a typical drop-on-demand drop generator. The operational principle is based on the generation of an acoustic wave in a fluid-filled chamber by a piezoelectric transducer through the application of a voltage pulse. The acoustic wave interacts with the free meniscus surface at the nozzle to eject a single drop. The major advantage of the drop-on-demand method is that it does not require a complex system for droplet deflection. Its main drawback is that the size of piezoelectric transducer tube or disc, in the order of sub mm to several mm, is too large for high-resolution applications. It was reported that the typical frequency for a stable operation of piezoelectric ink-jet would be tens of kHz.

Fig. 9.4. Operation principle of a piezoelectric-actuated droplet-on-demand droplet generato
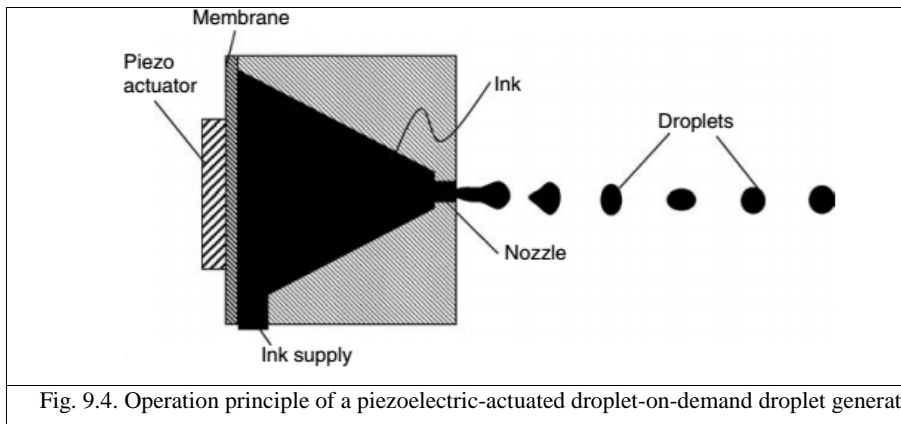
Principles of work of piezoelectric droplet generators that based on deformation mechanisms as squeese, bend, push, shear is shown on Figure XXX.

Figure Constriction of piezoelectric MEMS droplet generators

D
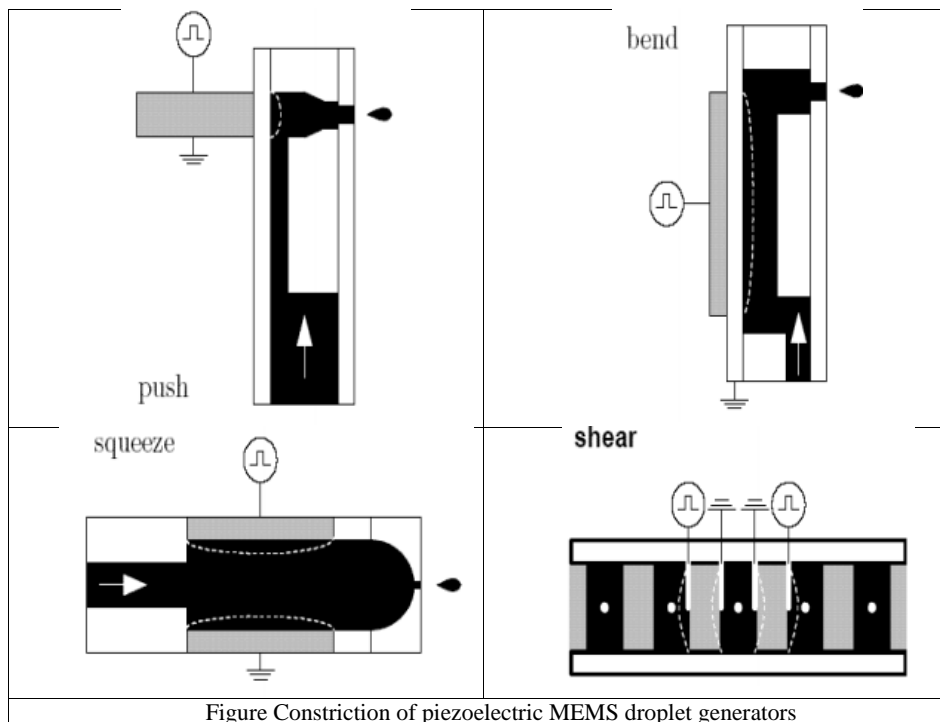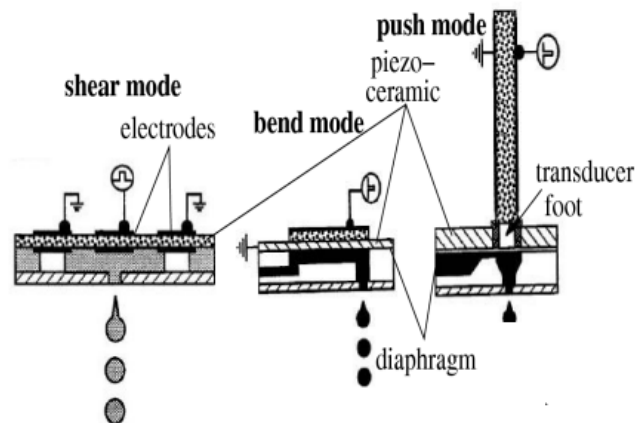
Figure Constriction of piezoelectric MEMS inkjet heads operated in shear, bend and push mode

Dd
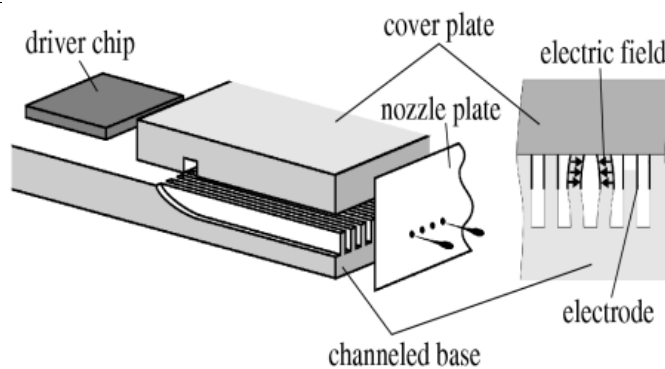
Figure. Main parts of piezoelectric MEMS inkjet head that operate in shear mode.

9.5.3   Thermal-Bubble Actuation

Thermal bubble jet technology has been studied by HP in the United States and by CANON in Japan since the early 1980s. Figure 9.5 shows the cross-section of a thermal bubble jet device []. Liquid in the chamber is heated by applying a pulse current to the heater under the chamber. The temperature of the liquid covering the surface of the heater rises to around the liquid critical point in microseconds, and then a bubble grows on the surface of the heater, which serves as a pump.

Commented [RbD3]: Allen, R.R., Meyer, J.D., and Knight, W.R. (1985) "Thermodynamics and Hydrodynamics of Thermal Ink Jets," *Hewlett-Packard J.*, May, pp. 21–27.
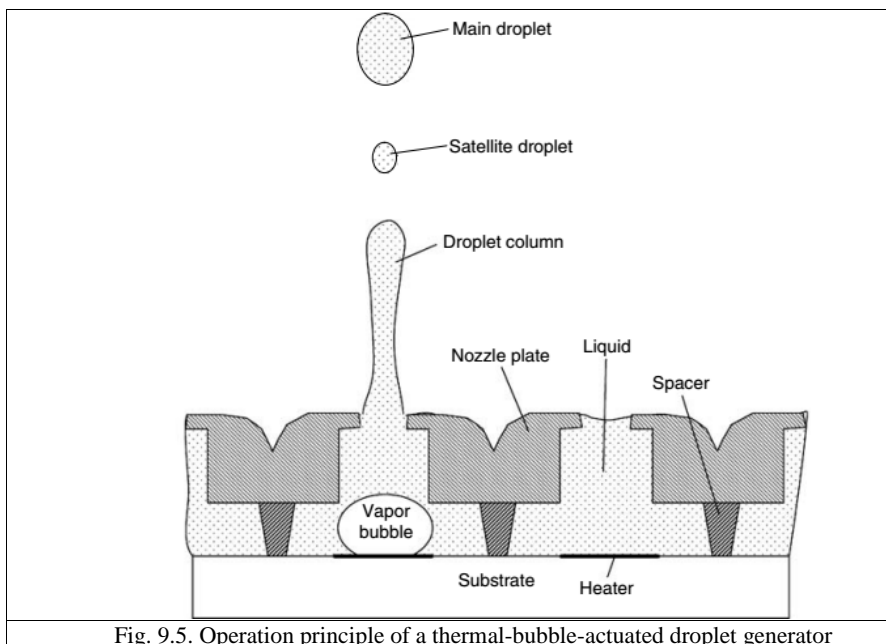
Fig. 9.5. Operation principle of a thermal-bubble-actuated droplet generator

The bubble pump pushes liquid out of the nozzle to form a droplet. After the droplet is ejected, the heating pulse is turned off and the bubble starts to collapse. Liquid refills the chamber when surface tension on the free surface of the meniscus returns it to the original position. A second pulse generates another droplet. The energy consumption for ejecting each droplet is around 0.04 mJ for an HP ThinkJet printhead. Because bubbles can deform freely, the chamber size of the thermal bubble jet is smaller than that of other actuation means, which is important for highresolution applications. The resolution reported in the literature ranges from 150 to 600 dpi and 1016 dpi. The typical operational frequency for contemporary thermal bubble jets is around several to tens of kHz.

9.5.4 Thermal-Buckling Actuation

A buckling diaphragm was porosed for droplet generation. Figure 9.6 schematically shows the technique's basic operational principle. A composite circular membrane consisting of a silicon dioxide and nickel layer is fixed on the border and remains separated from the substrate by a small gap. A heater is placed at the center of the composite membrane and electrically isolated from it.

Pulsed current is sent to the heater, and then the membrane is heated for several µs. When the thermally induced stress is greater than the critical stress, the diaphragm buckles abruptly and ejects a droplet out of the nozzle. The power required to generate a droplet at a speed of 10 m/s is around 0.1 mJ using a 300 µm diameter diaphragm. The power consumption and the size of the buckling membrane ink-jet device are much larger than those of the thermal bubble jet. The reported

frequency response of a membrane buckling jet ranges from 1.8 to 5 kHz depending on the desired droplet velocity.
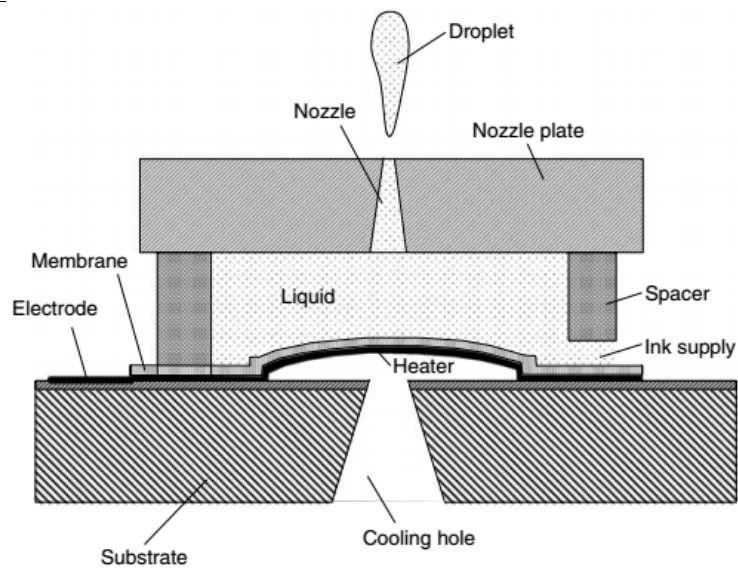
Fig. 9.6. Operation principle of a thermal-buckling-actuated droplet generator

9.5.5 Acoustic-Wave Actuation

A lensless liquid ejector using constructive interference from acoustic waves to generate droplets [] is shown on Figure 9.7. A PZT thin-film actuator with the help of an on-chip Fresnel lens was employed to generate and focus acoustic waves on the air–liquid interface for droplet formation. The actuation results from the excitation of a piezoelectric film under a burst of RF signal. The device does not need a nozzle to define droplets, reducing the clogging problems that trouble most droplet generators employing nozzles. Droplet size also can be controlled by using acoustic waves with specific frequencies. However, due to the acoustic waves' vigorous agitation of the liquid, it is difficult to maintain a quiet interface for reliable and repeatable droplet generation. As a result, a "nozzle area" is still needed to maintain a stable interface. The applied RF frequency ranges from 100 to 400 MHz, and the burst period is 100 µs. The power consumption for one droplet is around 1 mJ, which is high compared to other techniques. The droplet size ranges from 20 to 100 µm depending on the RF frequency. The reported size of the device is 1×1 mm^2, which is much larger than the droplet generators mentioned previously.
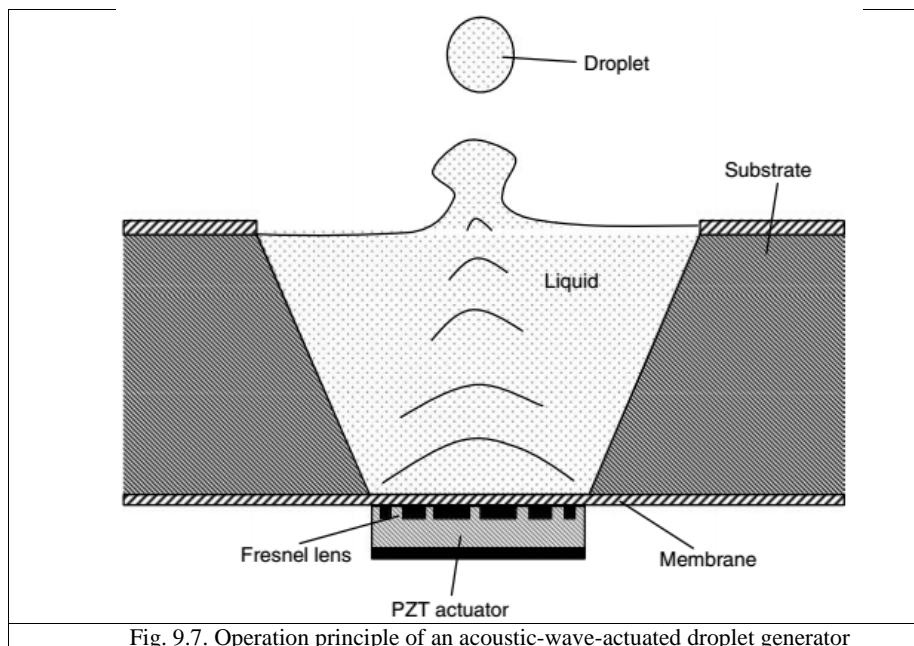
Fig. 9.7. Operation principle of an acoustic-wave-actuated droplet generator

9.5.6 Electrostatic Actuation

The electrostatically driven ink-jet printhead was first introduced by the Seiko Epson Corporation [] for commercial printing. As shown in Figure 9.8, the actuation is initiated by applying a DC voltage between an electrode plate and a pressure plate; this deflects the pressure plate for ink filling. When the voltage stops, the pressure plate reflects back and pushes the droplet out of the nozzle. This device has been developed for use in electric calculators due to its low power consumption, less than

0.525 mW/ nozzle. The driving voltage for a SEAJet is 26.5 V, and the driving frequency can be up to 18 kHz with uniform ink ejection. A 128-nozzle chip with 360 dpi pitch resolution also has been demonstrated. This device was claimed to offer high printing quality (for bar code), high speed printing, low power consumption, and long lifetime (more than 4 billion ejections) under heavy-duty usage as well as low acoustic noise. However, the fabrication comprises a complex bonding process among three different micromachined pieces. Further, making the pressure plate requires a very precise etching process to control the accuracy and uniformity of its thickness. Due to the deformation limitation of solid materials and alignment accuracy in the bonding process, the nozzle pitch may not be easily reduced any further for higher resolution applications.
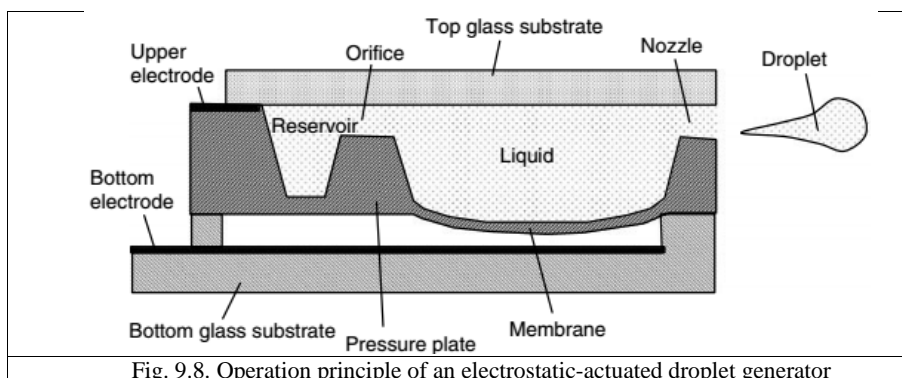
Fig. 9.8. Operation principle of an electrostatic-actuated droplet generator

9.5.7 Inertial Actuatio

Inertial droplet actuators apply a high acceleration to the nozzle chip to cause droplet ejection. This type of apparatus is shown in Figure 9.9. The print module consists of large reservoirs on the top plate, which is connected to nozzles on the bottom plate. The print module is mounted on a long cantilever beam with a piezo-bimorph-actuator for acceleration generation. Generating 1 nl droplets from 100 μm diameter nozzles requires 500 μs. Twenty-four liquid droplets of different solution types were demonstrated to be ejected simultaneously from the nozzles in a 500 μm pitch.
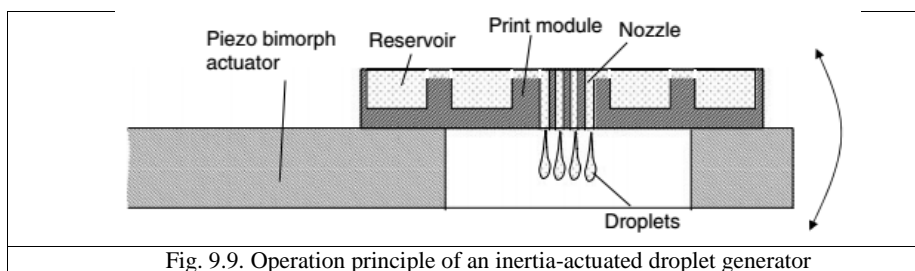
Fig. 9.9. Operation principle of an inertia-actuated droplet generator

The smallest droplet claimed to be generated is 100 pl from 50 μm diameter nozzles. This principle provides a gentle ejection process for bioreagent applications. However, the ejection of smaller droplets may encounter strong surface tension and flow-drag forces in micro scale that are much

larger than the droplet inertial. That the droplets cannot be selectively and individually ejected from the desired nozzles also limits this technique's applications.

9.6 Aplications

9.6.1   Ink-jet Printing

Graphics and text printing. Ink-jet printing, which involves arranging small droplets on a printing medium to form texts, figures, or images, is the most well-known microdroplet application. The smaller and cleaner the droplets are, the sharper the printing is. However, smaller droplets cover a smaller printing area and thus require more printing time. Therefore in printing, high-speed microdroplet generation with stable and clean micro sized droplets is desired for fast and high quality printing. The printing media can be paper, textile, skin, cans or other surfaces that can adsorb or absorb printing solutions.

Principle of work of Epson Micro Piezo print head is shown on Figure XXX

Diagram 1: How an Epson Micro Piezo™ print head works

Ink   Print nozzle   Meniscus

Piezoelectric Element   Pull

Ink is drawn from main ink chamber

Stage 1: An electrical pulse is applied to the piezoelectric element behind each nozzle. This causes the element to bend, creating a negative pressure that draws ink from the main ink chamber into the firing chamber.

Ink droplet is propelled out at great speed when piezoelectric element is bent the other way

Piezoelectric element returns to original shape after cycle is complete

Push

Stage 2: The electrical charge is altered, causing the piezoelectric element to rapidly push in the opposite direction, propelling the ink droplet out.

Figure. Principle of work of Epson Micro Piezo print head

For comparision priciple of work of thermal inkjet print head is depicted on Figure XXX and advantages and disadvantages of ecach methos are shown in Table 9.1

Diagram 2: How a thermal inkjet print head works

Heating resistor

Ink drawn in from main chamber

Stage 1: A resistor rapidly heats the ink in the firing chamber until it boils and starts to form an air bubble

Stage 2: In a micro second, the air bubble grows until it explodes the ink out through the nozzle and onto the medium

Stage 3: After propelling the ink out, the bubble cools, contracts, and collapses, and the resulting vacuum draws more ink in from the main ink chamber to repeat the cycle

Figure. Principle of work of thermal inkjet print head

Table 9.1

	Piezo	**Thermal**
Advantages	• Precise and Variable Droplet Sizes • Can use wide range of inks due to low temperature (solvent, UV, pigment, dye) • Droplet sizes down to 1.5 picoliter • Run longer due to cooler temperature	• Less Expensive Print Heads • More Print Heads Per Printer
Disadvantages	• More Expensive Heads • Fewer Print Heads Per Printer	• Only 2 different droplet sizes • Larger droplet size • Limited ink options due to extreme heat • Need to be replaced more often due to high heat

Microscopic ink droplets and precise ejection are essential to high image quality. With conventional ejection systems, however, the finer the ink droplets, the more vulnerable they are to airflow disturbances from print head movement and changes in ink viscosity due to temperature fluctuations, resulting in inconsistent ejection volumes and ink dot placement. Canon's FINE print head construction is shown on Figure XXX. Canon's FINE print head ensures that all the ink under the heater is pushed out in a single ejection to ensure efficient performance. Ink droplet speed has been increased to more than 1.5 times that of conventional systems, reducing their susceptibility to airflow disturbances to enable increased placement accuracy.

Figure FINE's Ink-Ejection Process
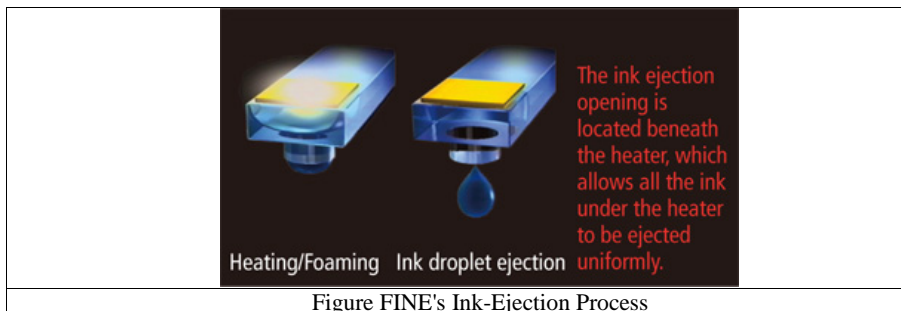
Achieving smaller droplet sizes and faster print speeds requires high-precision fabrication technologies to enable the placement of a greater number of nozzles over a wider area. Canon's FINE print heads are made by integrating the heaters and nozzles into a single unit on a wafer using a process that makes the most of the company's expertise in semiconductor production, as well as its original material technologies and innovative processing technologies. The ability to fabricate nozzles with high precision over large areas without using bonding processes makes it possible to arrange 6,000 or more nozzles within an area measuring a mere 20 mm x 16 mm, roughly the size of a thumbprint (Figure XXX)

Figure Print Head with Enlarged View of Nozzle Configuration
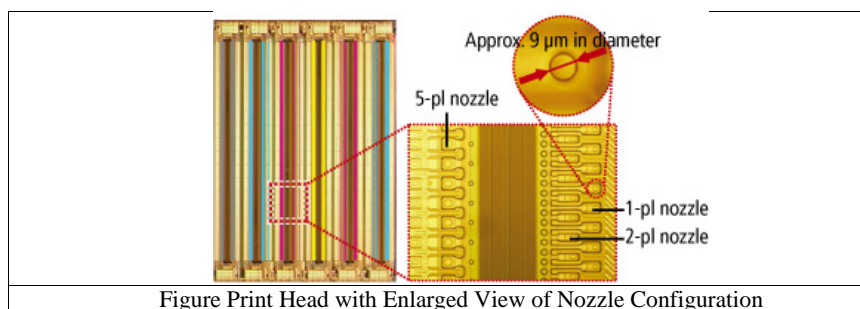
Main parts of an inkjet printer.

The main parts of an inkjet printer is shown on Figure XXX
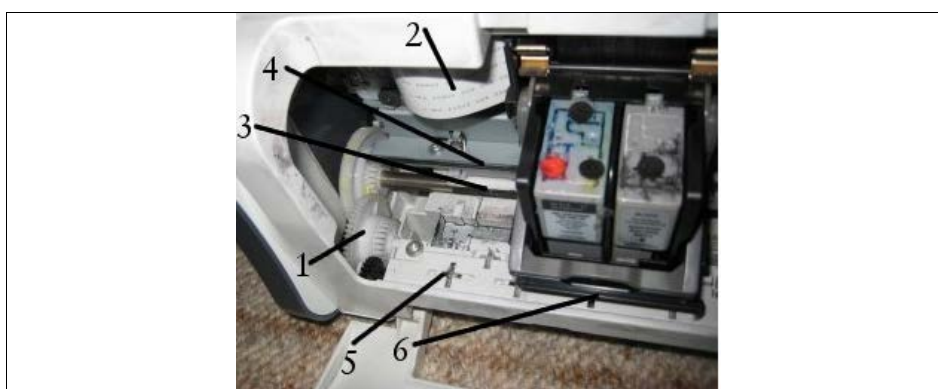
Figure. Construction of inkjet printer

1 Gears driven by an electric stepper motor turn rollers that advance the paper through the printer.

2 A flexible ribbon cable carries printing instructions from the electronic circuit inside the printer to the moving cartridge. (Inkjet printers contain circuits that translate the instructions from your computer into precise movements of the printhead. The electronic bits and bobs aren't actually visible on this photo. There's usually a single, large circuit board somewhere in a printer with all the components mounted on it, including the control switches, LED display lights, and various connections to the printer mechanism and power supply.)

3 Plastic and rubber rollers pinch the paper tightly so it can be moved through the printer with absolute precision.

4 A sturdy metal rail guides the printer head as it moves back and forth.

5 Spiked wheels at the front of the printer help to grip the paper securely and move it precisely.

6 The print cartridge prints from left to right then reverses the print information and prints backwards from right to left. This is known as bidirectional printing and allows pages to be printed much faster.

An important role in the work of inkjet printer plays a print head, which through microscopic nozzles is sprayed onto the surface of the carrier tiny ink droplets. Most old inkjet printers had 12 nozzles. The latest cutting-edge printers have the several tens of thousands of of nozzles, with the diameter of the nozzle is much smaller than the diameter of the needle, or even human hair (Fig. XXX)
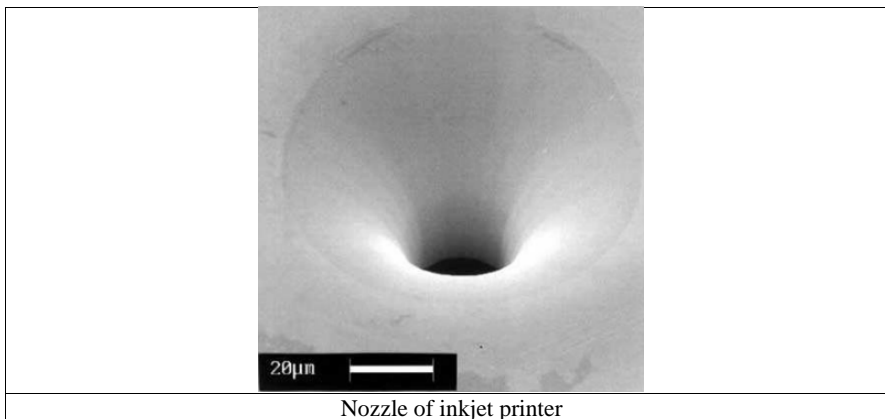
Nozzle of inkjet printer

.

Manufacturers of inkjet heads independently determine the number of nozzles in the print head, their size and location. For example, Lexmark company uses chess principle arrangement of small and

large nozzles, the Canon company prefers to equip their printers plurality of nozzles with a small diameter (Figure XXX).
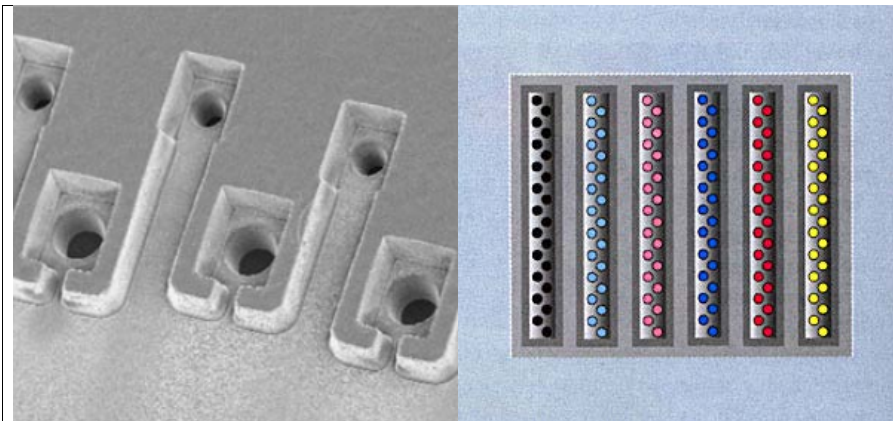
Figure Nozzles of printer head of Lexmark	Figure Nozzles of printer head of Canon

Modern inkjet printers form images from millions of microscopic droplets, and therefore they are very good quality. The picture of droplets under the microscope on the surface of the paper is shown on Figure XXX

.
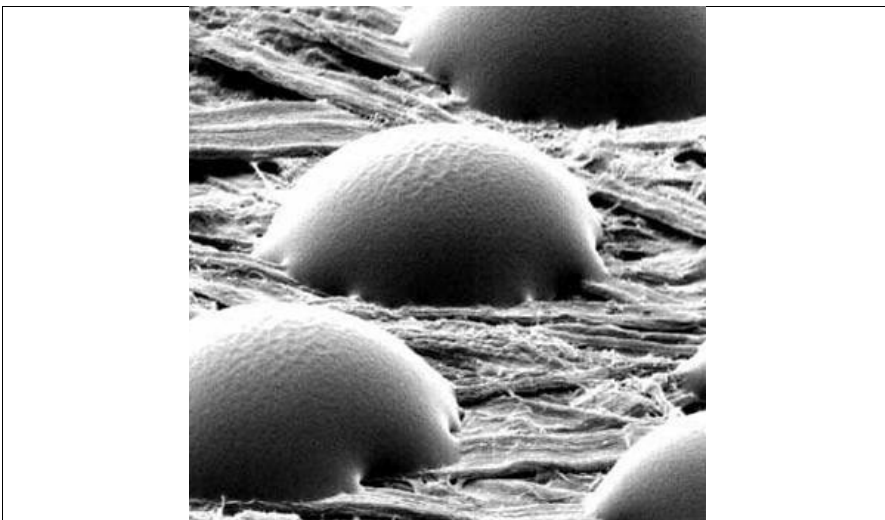
Figure. SEM image of ink droplets on surface of paper.

Electronic circles printing. Printed electronics will enable the development of ultra-low-cost RFID circuits for use as electronic barcodes, since it eliminates the need for lithography, vacuum processing, and allows the use of low-cost web manufacturing []. Recently, there have been several

demonstrations of printed transistors with mobilities approaching or exceeding 1 cm^2/V·s; however, all such devices have been fabricated using silicon substrates with thermally grown oxides or using vacuum sublimated materials. In order to achieve ultra low cost, performance must be maintained without silicon substrates or vacuum processing. Development of an inkjet-printed precursor with mobilities as high as 0.2 cm^2/V·s was reported This performance is more than adequate for a 135 kHz application. However, this study was also done on silicon substrates. The process flow for such a device is shown in Figure XXX while the typical I_D-V_{DS} curve is shown in Figure XXX.
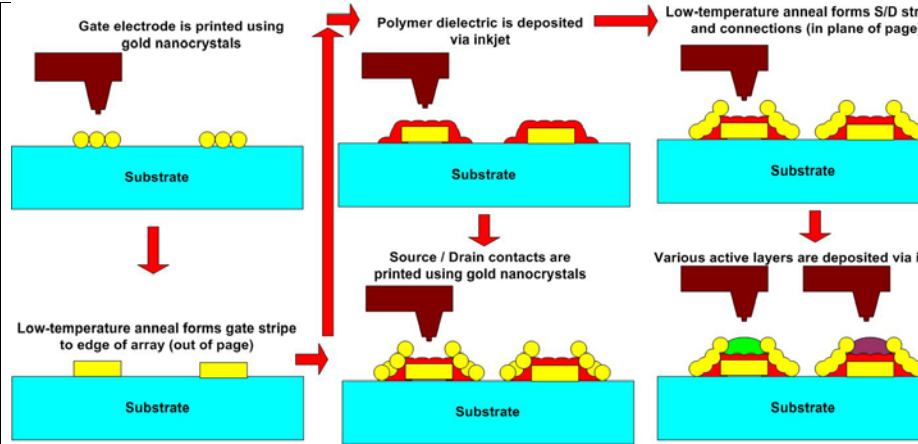
Fig. Process flow for an all-inkjet-printed transistor
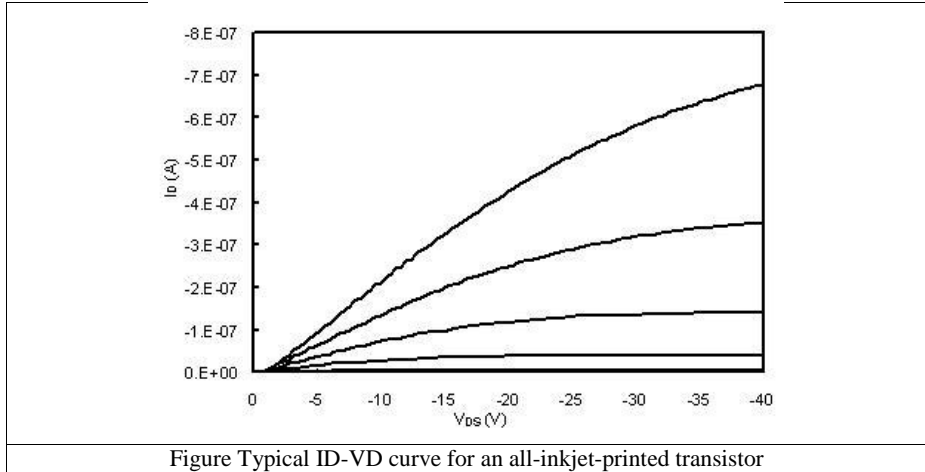
Figure Typical ID-VD curve for an all-inkjet-printed transistor

General principle of printing electronic circles using MEMS based inkjet printers is shows on Figure

Scheme of non contact direct printing onto substrate for electronic applications
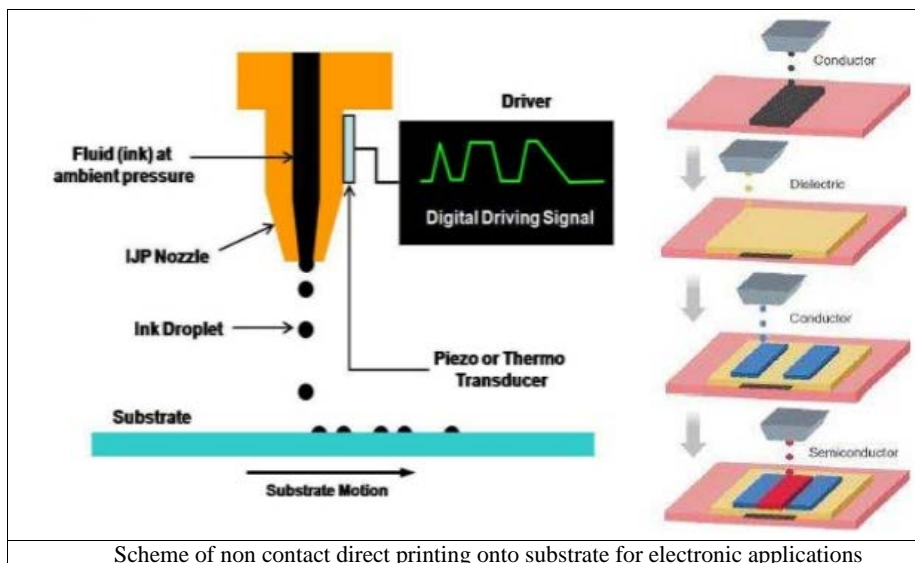
9.6.2 Biomedical and Chemical Sample Handling

The application of microdroplet generators in biomedical sample handling is an emerging field that has drawn much attention in the past few years. Many research efforts have focused on droplet volume control, droplet size miniaturization, compatibility issues, the variety of samples, and high-throughput parallel methods. Piezo- and pneumatictype droplet injectors was developed for mass spectrometry []. Figure XXX schematically shows the design of the injectors, which generate submicron to micron sized bioreagent droplets for sample separation and analysis in a mass spectrometer, as shown in Figure 10.29. Luginbuhl et al. (1999) employed silicon bulk micromachining to fabricate silicon nozzle plates and Pyrex glass actuation plates, while Wang et al. (1999) employed a combination of surface and bulk micromachining to fabricate the droplet generator. These injectors are part of the lab-on-the-chip system for incorporating microchips with macroinstruments. Microdroplet generators were also used by Koide et al. (2000), Nilsson et al. (2000), Goldmann et al. (2000), and Szita et al. (2000) for the accurate dispensing of biological solutions. Piezo- and thermal-type injectors were used in those investigations for protein, peptide, enzyme, or DNA dispensing. With an operation principle similar to ink-jet printing, the devices provided for precisely dispensing and depositing a single biological droplet onto a desired medium, and they also could dispense droplet arrays. The arrayed bioreagents can be bioprocessed further for high-throughput analysis. Continuous jet-type droplet generators were reported by Asano et al. (1995) to effectively focus and sort particles by using the electrostatic force. The experimental setup is shown in Figure 10.30. A syringe pump pressurizes the sample fluid to pass through a nitrogen sheath flow for focusing, and then the sample is ejected from a piezoelectric transducer disturbed

Commented [RbD7]: Luginbuhl et al. (1999), Miliotis et al. (2000), and Wang et al. (1999)

nozzle to form droplets. Droplets containing the desired particles were charged at the breakup point and deflected into collectors. The reported separation probability for 5, 10, and 15 µm particles can be as high as 99%. However, the inner jet diameter limits the particle size for separation. Other than solid particle separation, this method potentially can be applied to cell sorting for biomedical applications. In addition to biomedical reagent handling, microdroplet generators were widely used in chemical handling. For example, Shah et al. (1999) used an ink-jet system to print catalyst patterns for electroless metal deposition. This system used a commercial ink-jet printhead to ejecte a Pt solution as a seed layer for Cu electroless plating. The lines produced by this method were reported to be 100 µm wide and 0.2–2 µm high.

9.6.3 Fuel Injection and Mixing Control

9.6.4 Direct Writing and Packaging

9.6.5 Optical Component Fabrication and Integration

9.6.6 Solid Free Forming

9.6.7 Manufacturing Process

9.6.8 IC Cooling

12. Micro-opto-electromechanical systems (MOEMS)

12.1 Introduction

Optics and photonics have benefited greatly from miniaturization through Micro-Electro-Mechanical Systems (MEMS). Optical MEMS is a focus for extensive research into the fundamentals of optical interactions because it enables the application of silicon fabrication technology to micro-optical systems and because it provides a unique set of tools for scientific measurements across a wide range of disciplines, making it a driver for the commercial development of optical systems.

The precision mechanics of MEMS, microfabrication techniques, and optical functionality all make possible a wide variety of movable and tunable mirrors, lenses, filters, and other optical structures. In these systems, electrostatic, magnetic, thermal, and pneumatic actuators provide mechanical precision and control. The large number of electromagnetic modes that can be accommodated by beam-steering micromirrors and diffractive optical MEMS, combined with the precision of these types of elements, is utilized in fiber-optical switches and filters, including dispersion compensators. The potential to integrate optics with electronics and mechanics is a great advantage in biomedical instrumentation, where the integration of miniaturized optical detection systems with microfluidics enables smaller, faster, morefunctional, and cheaper systems. The precise dimensions and alignment of MEMS devices, combined with the mechanical stability that comes with miniaturization, make optical MEMS sensors well suited to a variety of challenging measurements. Micro-optical systems also benefit from the addition of nanostructures to the MEMS toolbox. Photonic crystals and microcavities, which represent the ultimate in miniaturized optical components, enable further scaling of optical MEMS.

What is difference between 'micro-optics' and 'macro-optics'? Even though the optical physics does not change as optical structures become smaller, three essential qualities which distinguish the two domains are proposed:

- micro-optics relies on a wide variety of phenomena, as diverse as optical diffraction, surface tension or micromechanical movement, which can play a much more important role in defining the performance of an optical component than in macroscopic optics;

- micro-optics may realize optical functions employing materials and states of matter, such as liquids, which are not ordinarily used for macroscopic optics;
- micro-optical components are manufactured using the technologies of microfabrication, which are fundamentally different from those traditionally used for macroscopic optical devices.

12.2   MOEMS applications

12.2.1 MEMS tunable optical elements

12.1.1.1 MEMS and Micro-Optics

The early work on MEMS-actuated optics, combined with the long traditions and capabilities of classical optics, has led to the development of a very rich discipline of tunable micro-optics. By using microfabrication techniques to manufacture microscale lenses, mirrors, filters, and other types of optical structures, the field of micro-optics has made available a broad spectrum of ultra-miniaturized optical devices. The combination of these with MEMS has, however, expanded the functionality and applicability of these components even further, primarily by making them tunable. A few examples of these MEMS tunable optical elements are discussed in this section.

Tuning Micro-Optics: A "tunable" optical component or system is one in which the optical characteristics may be controllably changed; the parameters to be tuned may include focal length, magnification, beam direction, spectral composition, or a host of other factors. In macroscopic optics, this tuning typically requires mechanical motion of one or more elements with respect to others, but in micro-optics entirely new means for optical tuning become possible. These results from the fact that in micro-optics, materials such as liquids, polymers, soft elastomers or flexible semiconductors, to name a few, are employed, using structures and technologies that have no analogs in macro-optical systems. These novel materials and technologies imply that optical elements may be tuned by changing the characteristics of the components themselves. Combining MEMS technology with microoptics means that microlenses may change their curvature, micromirrors accurately shift their reflection angle, or microfilters change their transmission wavelength, all by altering the optical properties of components on the microscale.

The role of MEMS: Two aspects of MEMS play an important role in the realization of tunable micro-optics. The first is MEMS fabrication technology itself, since microoptical structures are manufactured using microtechnology, not the traditional production means for macroscopic optics. The second role of MEMS in tunable optics is the generation of means to actuate tunable components. Whereas the field of optical MEMS initially relied strongly on electrostatic actuation for mechanical movement or positioning of microoptical components, the spectrum of actuation techniques has greatly expanded. Pneumatic, fluidic, thermal, magnetic and even chemical mechanisms for optical tuning have been employed, thus greatly expanding the repertoire of conceivable tunable optical systems.

12.1.1.2 Tunable Micromirrors

Actuated micromirrors are perhaps the oldest micro-optical devices fabricated using MEMS techniques, and since they were used from the start as a means to variably deflect an optical beam, they may be considered the "original" tunable micro-optical component. Movable micromirrors with high positioning and attitude accuracy have been made using myriad designs, and served as early demonstrators for the successful combination of micromechanics and micro-optics. As a result, micromirrors have been applied to a wide range of applications in areas as diverse as optical telecommunications and consumer electronics, thus representing an optical MEMS success story. Difference by the drive type of the MEMS based mirror is shown in Table 12.1

Table 12.1 Properties of MEMS based micromirrors.

Drive system	Merits	Demerits
Electromagnetic	· Low driving voltage (5V or less) · Large driving power →Supportive of large mirrors and dissonance drive	· Large number of parts (magnets, yokes, etc.) · Characteristics degradation due to the use beyond the storage temperature (decrease of flux density due to high temperature) · Difficult to control the mirror arch (Film stress of the drive coil affects the mirror arch.)
Electrostatic	· Small numbers of parts · Consistent with LSI processes · Easy to control the mirror arch	· High driving voltage (100V or less) · Small driving power →Not supportive of large mirrors and dissonance drive · The comb electrode limits the amplitude. →Difficult to control the amplitude →Not supportive of large mirrors and large amplitude
Piezoelectric (thin film type)	· Small numbers of parts · Easy to control the mirror arch · Highly efficient forming technology of piezoelectric thin film has been established.	· High driving voltage (20V or less) · Small driving power →Not supportive of a large mirror and dissonance drive (a large drive unit is needed to support them.)

Electrostatic Actuation: As MEMS technology has advanced, the spectrum of mirror concepts has likewise broadened and these have become firmly established in the MEMS device repertoire. Whereas the optics of the mirror is relatively simple, one area on which considerable research effort has been expended is actuation for achieving angular and translational movement. Many original concepts used electrostatic actuation, which was shown to be suitable for precise positioning and thus accurate beam steering with a resolution of 0.2° in a scan range of 20°. When combined with micromechanical hinge structures for the mirrors, the typically planar electrostatic actuators were also employed to generate out-of-plane mirror movement for more flexible three-dimensional motion.
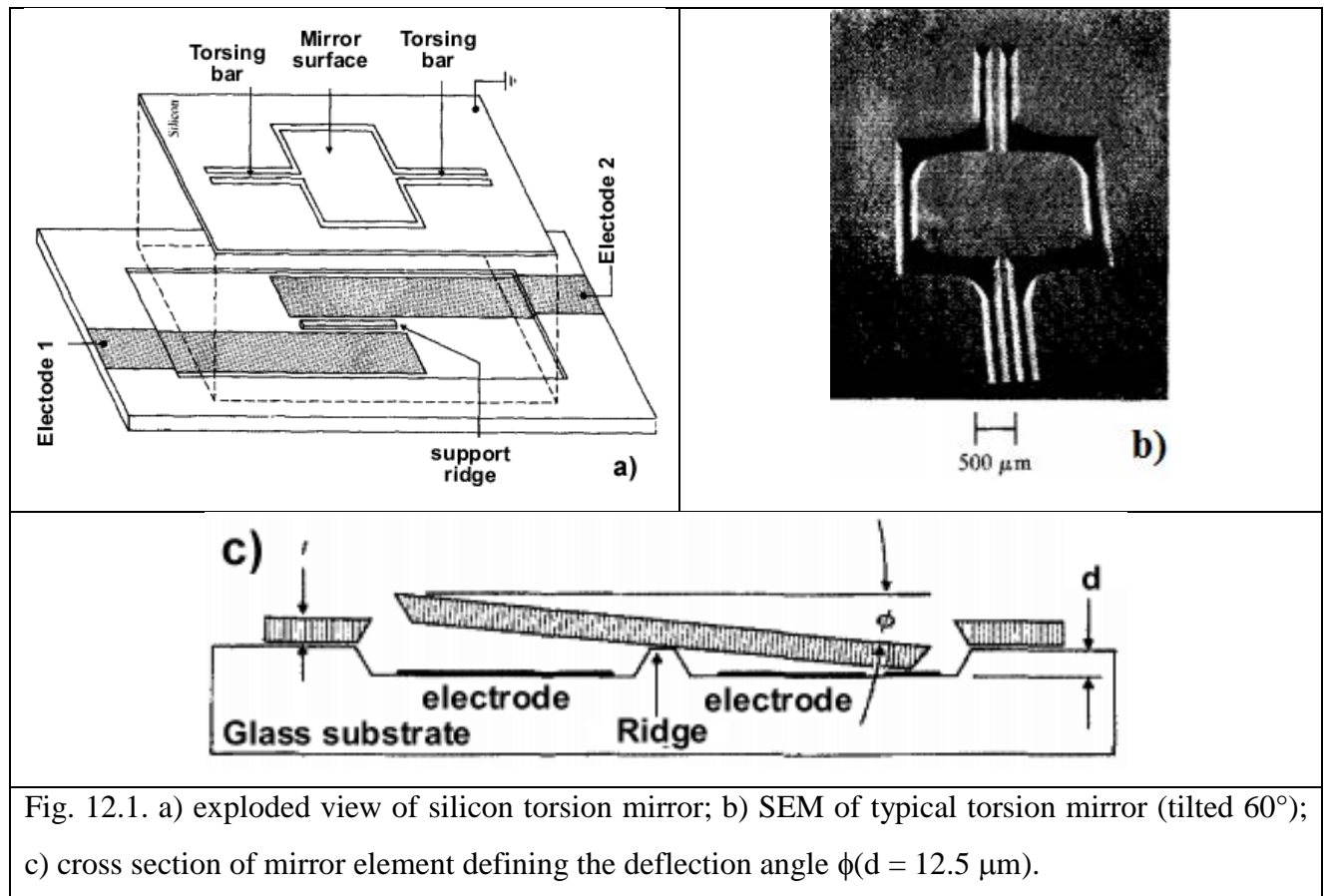
Fig. 12.1. a) exploded view of silicon torsion mirror; b) SEM of typical torsion mirror (tilted 60°); c) cross section of mirror element defining the deflection angle ϕ(d = 12.5 μm).

Optical deflector for display or galvanometer applications, silicon torsion mirrors is shown on Fig. 12.1a. A glass substrate was etched to form a well into which the torsion plate could deflect. The differential electrostatic attraction between the mirror and the electrodes causes the mirror to switch about the hinged diagonal (Fig. 12.1c). A support ridge is used to prevent it from being pulled down in parallel to the base of the well, and a pair of metal deflection electrodes were fabricated in the well. For a mirror size of 2x2 mm, deflection angle of ±1 o was achieved.

Continuous operation of the torsional mirror for months, followed by dislocation-sensitive silicon etching showed no damage. This highlights one of the important advantages of using single-crystal materials and of scaling down to these dimensions: fatigue mechanisms are not the same as those found in macroscopic materials due to the absence of grain boundaries.

Fig.12.2. a) 2-dimensional silicon scanner based on micromirror. b) SEM of 2-dimensional silicon actuated micromirror
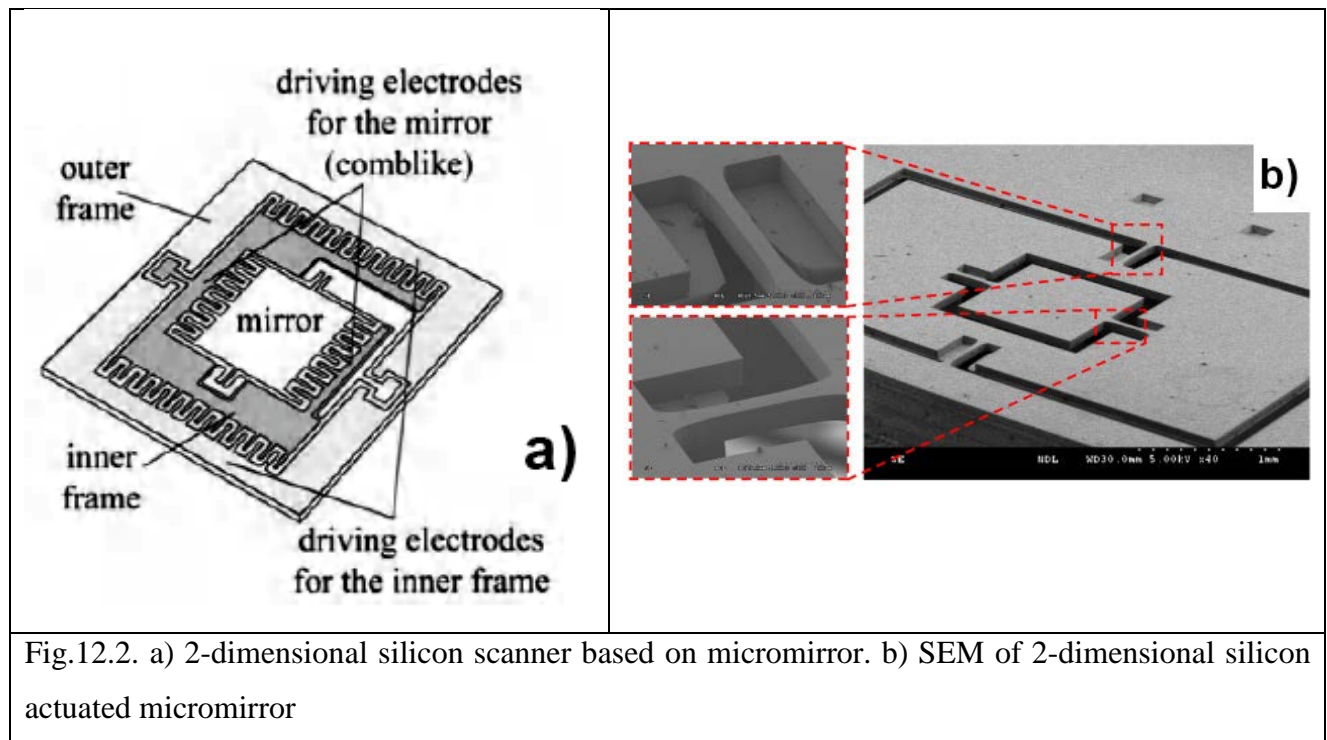
The Technical University of Chemnitz, Germany, demonstrated the fabrication of single-crystal silicon torsional light deflectors as large as 3 x 3 mm using combinations of dry and wet etching. They also demonstrated the use of square silicon plates for 20 light beam scanning Fig. 1.2.2. These plates were designed with four underlying deflection electrodes, and were supported by compliant silicon members to allow semi-independent actuation in two orthogonal directions.

Fig. 12.3. Design of MEMS micromirrors: a) cross-section of continuous DM; b) cross-section of segmented DM. Photograph of a fabricated DM with 140 active actuators supporting a continuous DM (c).
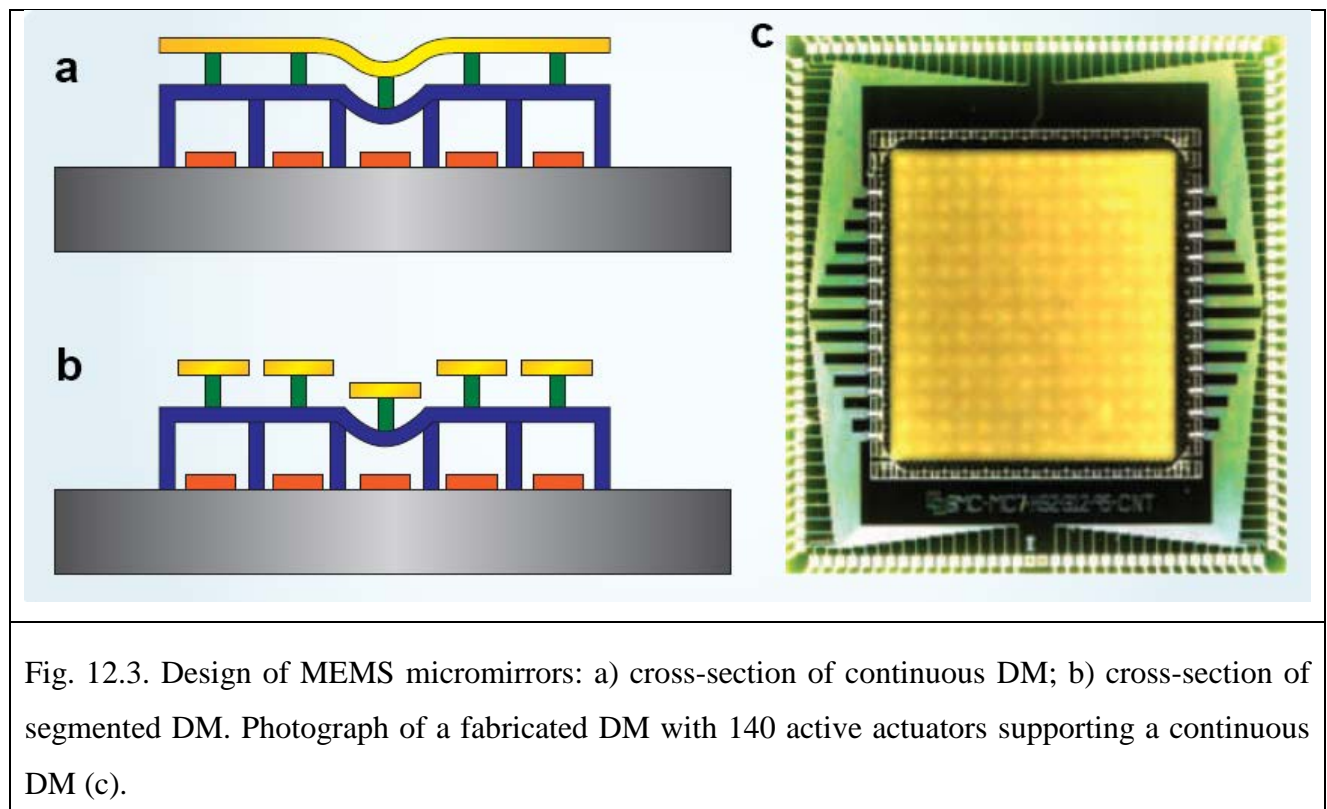
Design of MEMS deformable mirrors (DM) is shown on Fig. 12.3. A metal-coated thin-film mirror (gold) is attached by silicon posts (green) to an array of locally anchored compliant electrostatic actuator membranes (blue). Actuator deflection is controlled by applying independent voltages to an array of rigid silicon actuator electrodes (red) on the wafer substrate (black). Electrostatic attraction between the energized actuator electrodes and the electrically grounded actuator membrane deflects the actuators precisely and repeatably, thereby shaping the mirror. All structural elements are silicon.

Example of DM made of metal-coated glass plates bonded to an underlying array of actuators is shown on Fig. 12.3. Boston University group had developed macro- and microscale DMs in both segmented (Fig. 12.3a) and continuous-mirror configurations (Fig. 12.3b). Photograph of a fabricated DM with 140 active actuators supporting a continuous DM is shown on Fig. 12.3c.

MEMS DM devices bring a number of inherent benefits in their design and manufacturing processes. In terms of design, MEMS DM devices are scalable; increasing the size and spatial resolution of a DM is achieved by adding duplicate lithographic mask features. MEMS DMs are also mechanically stiff and lightweight, allowing their bandwidths to be controlled over tens of kilohertz. The non-contact electrostatic actuation mechanism is repeatable to subnanometre precision, consumes almost no power, exhibits no hysteresis and is unaffected by trillions of operation cycles. In terms of manufacturing, the MEMS DM production approach does not call for exotic materials or manufacturing tolerances: it can exploit a MEMS foundry and begins with an optically smooth, flat and inexpensive substrate. Devices are batch-produced twenty wafers at a time, so although development costs are high, commercial production and replication costs are low. In addition, hundreds of devices can be produced on each wafer, allowing broad parameter variation in a single batch-production cycle, which accelerates research and prototyping. Perhaps the most important benefit, however, is cost. MEMS DM systems cost almost an order of magnitude less than the technology they replace. This economic advantage has catalysed rapid growth in the field of AO, in both astronomy and other imaging applications in which aberration compensation is critical, such as improved *in vivo* retinal imaging systems and the subsurface imaging of biological tissue.

Astronomical applications. Last years adaptive optic (AO) has become indispensible as a means of compensating for aberrations introduced by atmospheric turbulence in large ground-based telescopes. The resulting gains in resolution are especially important in narrow-field imaging with large telescopes, and have led to exciting recent advances in the observation of exoplanets (planets outside the Solar System), the characterization of planetary rings and atmospheres, and studies of galactic structure. AO will become increasingly important for the coming generation of extremely large telescopes, as its benefits increase nonlinearly with telescope aperture. 'Extreme' AO — the direct detection and characterization of exoplanets through coronagraphic imaging or spectroscopy

combined with high-performance AO — is a key driver for the development of advanced MEMS DMs. Boston Micromachines Corporation (BMC) recently produced a 4,092-actuator, 26 mm aperture, 4 μm stroke MEMS DM for a new planet-imaging instrument known as the Gemini Planet Imager (Fig. 12.4).
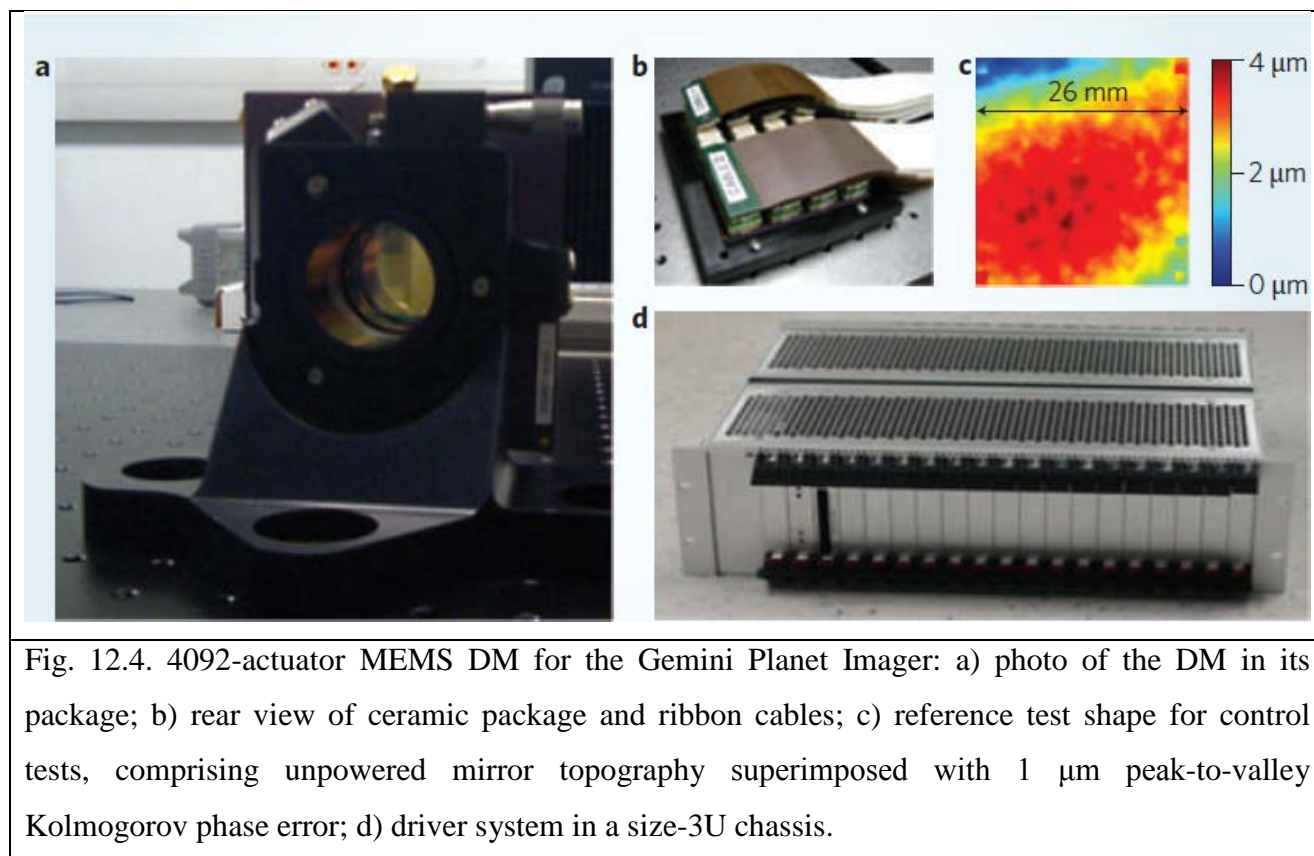
Fig. 12.4. 4092-actuator MEMS DM for the Gemini Planet Imager: a) photo of the DM in its package; b) rear view of ceramic package and ribbon cables; c) reference test shape for control tests, comprising unpowered mirror topography superimposed with 1 μm peak-to-valley Kolmogorov phase error; d) driver system in a size-3U chassis.

Retinal imaging. *In-vivo* retinal imaging using AO-enhanced ophthalmoscopes was a key driver for the development of MEMS DMs. The motivation was straightforward: uncompensated aberrations of the eye generally prevent ophthalmoscopic instruments from resolving cellular structures of the retina. The diffraction limit for the optics in the eye would allow cellular resolution, however, if such aberrations could be corrected. A pioneering effort by David Williams' research group at the University of Rochester demonstrated closed-loop AO in 1997, yielding substantial improvements in image quality and the first clearly resolved images of cone photoreceptors. A decade of subsequent research by many groups yielded numerous additional advances, including the integration of AO into confocal scanning laser ophthalmoscopy and optical coherence tomography instruments, and the development of MEMS DMs designed specifically for use in retinal imaging AO. The capacity of confocal AO systems to resolve cellular and microvasculature structures has led to a rapid increase in the number of AO publications and development grants over the past decade.

BMC MEMS DMs in particular have been a critical and enabling component for aberration compensation and AO in vision science instruments. BMC MultiDMs were specifically designed to compensate for aberrations of the eye, with a 4.4 mm aperture, 6 μm stroke and 140 actuators. In a

recent initiative supported by the National Science Foundation through the Center for Adaptive Optics and a grant from the National Institutes of Health Bioengineering Research Partnership, a team of scientists produced more than a half-a-dozen variations of retinal imaging instruments that are AO compensated using BMC MEMS DMs. These include both scanning laser ophthalmoscopes and optical coherence tomography systems, and have resulted in the sharpest images ever obtained of microscale features in the living retina. In collaboration with Steven Burns at the University of Indiana, USA, BMC recently integrated a 6-μm-stroke DM into an AO laser scanning ophthalmoscope (AOSLO). The AOSLO's lateral resolution of ~2 μm enables *in-vivo* high-resolution imaging of previously indiscernible retinal structures. This instrument used to evaluate the clinical usefulness of enhancements in resolution, contrast and brightness when monitoring the progression of disease, as well as the effectiveness of treatments that affect cellular structures in the retina.

The adaptive optics scanning laser ophthalmoscope (AO-SLO) is developed to enhance the value of the retinal examination, a procedure essential for the diagnosis of eye diseases. Retinopathy and other disorders of the eye can develop as a result of not only hereditary transmission and aging, but also diseases such as diabetes or other causes. Because eye disorder can cause reduced visual acuity or loss of sight, early disease detection is essential for healthy eyes, just as it is for other parts of the body. Among the many organs of the body, the fundus of the eye is the only organ with a vasculature clearly viewable by direct observation. Retinal examinations reveal not only vascular conditions associated with eye disorders, but possibly also how hypertension, arterial sclerosis, and other diseases affect the eye.

The early detection of abnormalities in the retina requires an ophthalmoscope that can identify photoreceptors. Modern AO-SLO enables the viewing of minutely detailed dynamic images with a resolution of around 5 μm (microns). This has made possible the detailed observation of individual photoreceptors as well as the white blood cells flowing in the blood vessels of the retina. A AO-SLO technology can be divided into three basic parts (Fig. 12.5): an ophthalmoscope system configured with a laser, MEMS mirror and lenses; an AO control system with a wavefront sensor and wavefront correction device; and an image analysis system for analyzing captured images of the retina.

Several large hurdles must be cleared before high-definition images of the photoreceptors can be reliably captured. One of these is ocular aberrations. Aberration refers to the blurring or distortion of an image while the image is being formed through a lens system. Because the eye, as a lens, has aberrations, when observing the eye from the outside using an ophthalmoscope, distortions invariably occur in the resulting images. Aberrations of the human eye vary from person to person and also according to the condition of the eye during examination. Adaptive optics (AO) is an advanced technology that enables precise correction for aberrations. AO control system can

measure aberrations using a sensor and controls a wavefront correction device to make possible high-speed correction, and integrated the system into the scanning laser ophthalmoscope (SLO).
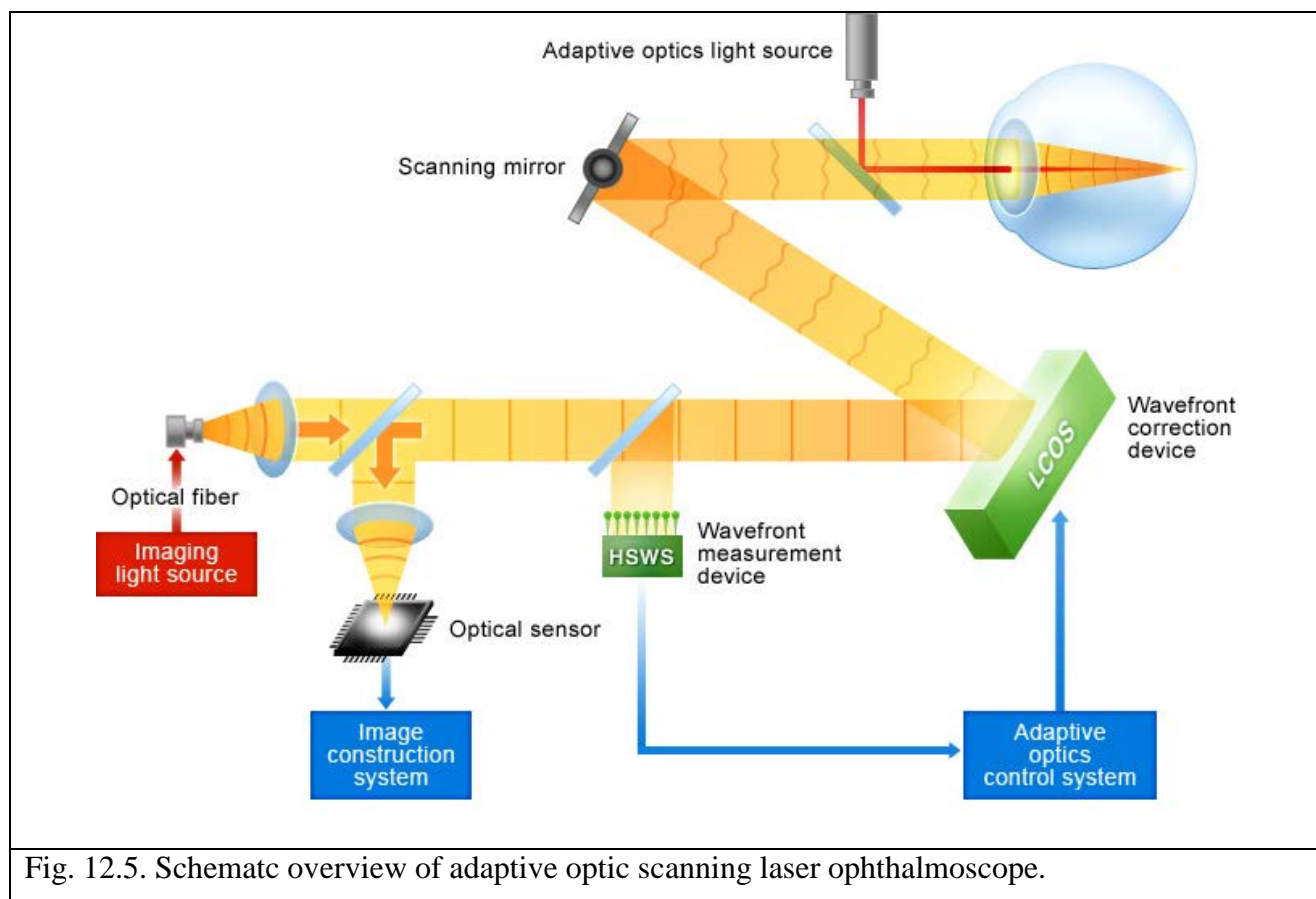
Fig. 12.5. Schematc overview of adaptive optic scanning laser ophthalmoscope.

Example of MEMS-based adaptive optics scanning laser ophthalmoscope image of the individual cellular layers of the retina is shown on Fig. 12.6.
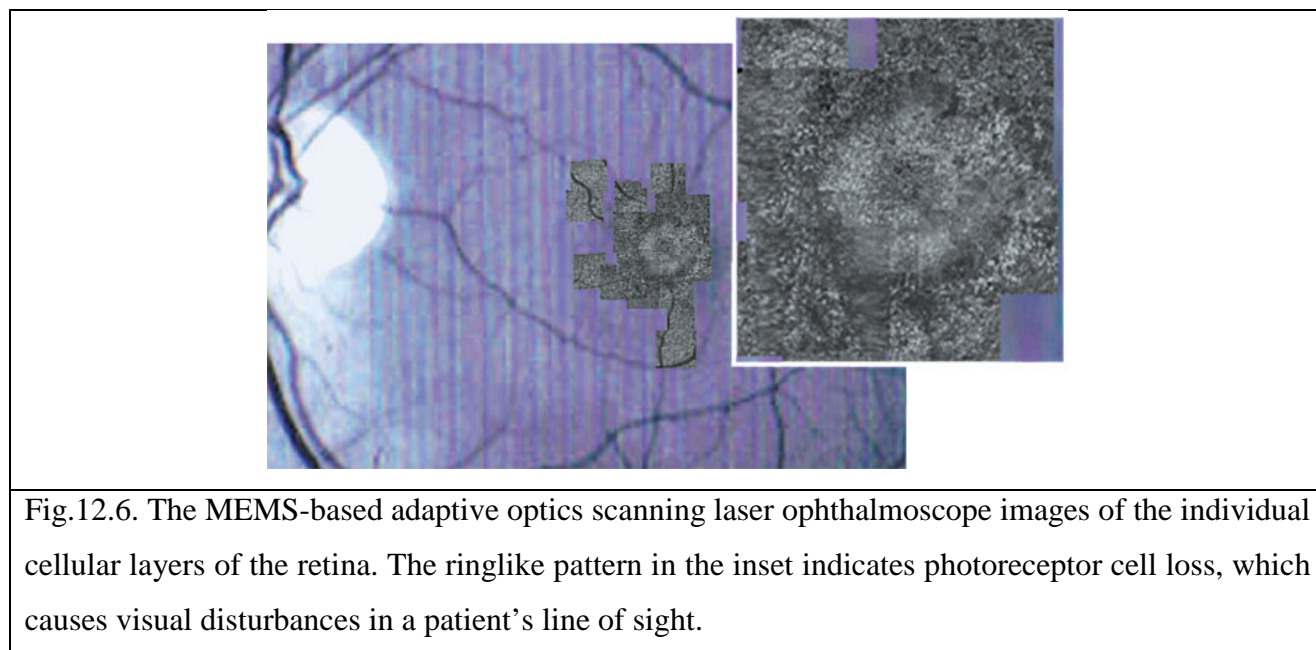
Fig.12.6. The MEMS-based adaptive optics scanning laser ophthalmoscope images of the individual cellular layers of the retina. The ringlike pattern in the inset indicates photoreceptor cell loss, which causes visual disturbances in a patient's line of sight.

Biological microscopy application. MEMS DMs are compact and inexpensive, allowing them to be easily integrated into many experimental microscopy platforms that might benefit from AO-enhanced performance. Two prominent examples are multiphoton microscopy for the subsurface imaging of biological tissue, and adaptive scanning microscopy for high-resolution, wide-field

imaging (Fig. 12.7). In multiphoton microscopy for the subsurface imaging of biological tissue, image quality is affected by both scattering in the biological medium and by an index-of-refraction mismatch between the tissue and the water/oil on which the microscope objective is optimized.

Fig. 12.7. Schematic (left) and photograph (right) of the ASOM — the first commercially available AO microscope — produced by Thorlabs and built using the BMC MultiDM system.
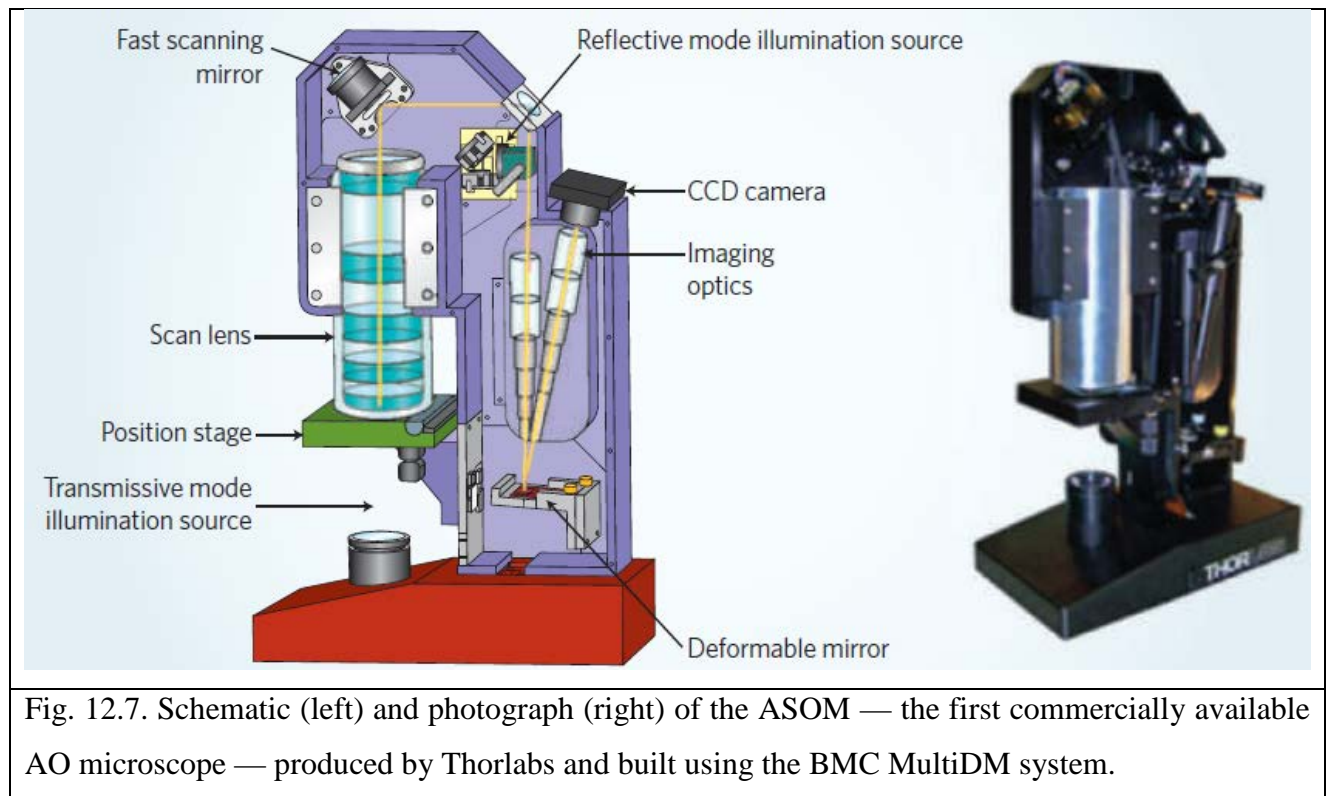
AO can be used to overcome the deleterious effects (primarily spherical aberration) resulting from this index mismatch. However, multiple scatting makes it challenging to measure such aberrations. As a result, this application relies on 'sensorless' AO techniques, in which an optimization scheme iteratively reshapes the DM and then measures the resulting image quality. MEMS DMs have proved to be well-suited to this type of control because they are fast (frame rates of >10 kHz) and can be shaped precisely and predictably in an open loop.

Exploiting this shaping characteristic makes it possible to improve the performance of iterative AO control techniques by intelligently selecting perturbation shapes (spherical aberration, for example). AO example applications use MEMS DMs to correct for intrinsic aberrations in the optical beam path of an imaging instrument (atmospheric turbulence, misshapen cornea/lens and heterogeneous biological tissue). It is also possible to use AO to broaden the design space for optical instruments themselves, even in the absence of sample aberrations. A prominent example is the adaptive scanning optical microscope (ASOM) developed by the Center for Automation Technology at the Rensselaer Polytechnic Institute in the USA. This instrument uses AO to replace stage scanning with beam scanning in a high-resolution microscope that images over a wide field-of-view. Imaging such a wide field at high resolution (in pathology screening, for example) normally requires a narrow-field, high-resolution objective to be positioned above the translating stage, after which the sample itself is moved to construct a composite mosaic image. Wider field

objectives are possible and have been used in lithographic applications, but making such objectives with sufficient quality to achieve diffraction-limited image quality across the field is prohibitively expensive for most applications. In the ASOM, a wide field objective is manufactured at low cost. Its aberrations at different scanning angles are measured in a calibration process and then compensated for using a MEMS DM inserted in the microscope beam path. The ASOM offers an expanded field-of-view, rapid scanning speeds, low light imaging capabilities and no specimen movement.

Magnetic actuation. As MEMS technologies have developed, numerous other principles have become feasible for micromirror actuation. Magnetic actuation using silicon and magnetic polymers, as shown in the example of Fig. 12.8, has led to angular deflections at resonance of ±10° for fields of only 0.07mT. Alternatively, purely metallic (nickel) micromirror-based scanners, which do not require expensive silicon on insulator (SOI) substrates for fabrication, have also been developed to provide excellent performance, yielding deflections of ±7° for fields of 0.23mT. Magnetically actuated mirrors do not require the high voltages of most electrostatic concepts, and they only require low magnetic fields that can be generated using integrated or ultra-miniaturized coils.
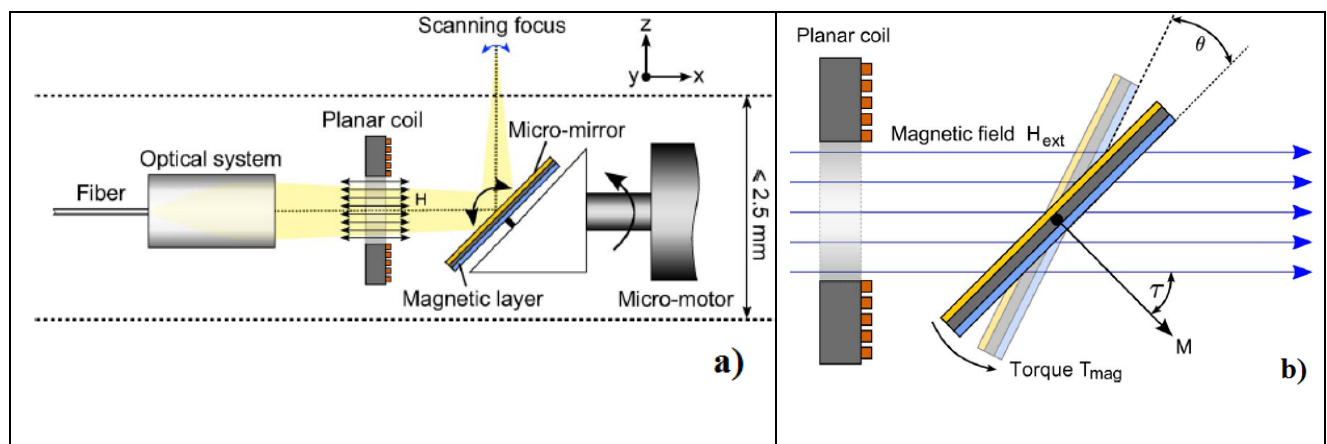
Fig. 12.8. a) Schematic sketch of a tubular scanning endoscopic probe. The magnetic micromirror is rotated by an integrated micromotor about the x-axis while it delivers a fast 1-D scan, yielding in combination a 2-D tubular scan; b) Actuation of the mirror by applying an external magnetic field at an angle τ to the surface (and, thus, direction of magnetization of the mirror M). The resulting torque leads to a rotation of the mirror membrane.

One area in which magneticactuated MEMS micromirrors are potentially of great interest is imaging endoscopy: Mirrors are used for scanning an optical field either for diagnostics, using, for example, optical coherence tomography (OCT); for treatment, employing, for example, laser ablation; or for imaging. For use in clinically applicable endoscopes, the micromirrors need to be compact (less than 5 mm diagonally) and actuated without prohibitive voltages or temperatures. One advanced such implementation is shown in Fig. 12.8a. A micromirror is mounted at 45∘ on the

shaft of a micromotor, allowing a fully circumferential scan. Actuation of the mirror by applying an external magnetic field can be performed at an angle τ to the surface (Fig. 12.8b).

The resulting torque leads to a rotation of the mirror membrane. By concomitantly scanning the mirror in one dimension, a 2-D 360° arc may be scanned; if the arrangement is combined with an OCT setup, for which the A-scan is in the radial direction, a 3-D tubular ring may be scanned. This functionality is presently realized in clinical practice either by rotating the entire endoscope using external motorized stages, an approach which often results in signal artifacts induced by vibration and bending of the optical fibers, or by the use of a rotating prism. These approaches provide only 1-D circular lines, making a synchronized lateral movement necessary for 2-D imaging. A micromirror used, as shown in Fig. 12.8a, can only be realistically actuated magnetically, since its rotation does not allow hard-wired electrical contacts

Due to their efficiency, the magnetic micromirrors may be actuated using miniaturized coils, also integrated into the endoscope, requiring only low drive currents and tolerating large mirror/coil spacings. Using a combination of SOI and polymer technologies, these micromirrors may be mass fabricated using standard MEMS processes (Fig. 12.9).

Fig. 12.9. Hard magnetic micromirror using a polymeric magnet attached to the backside of the tiltable mirror.
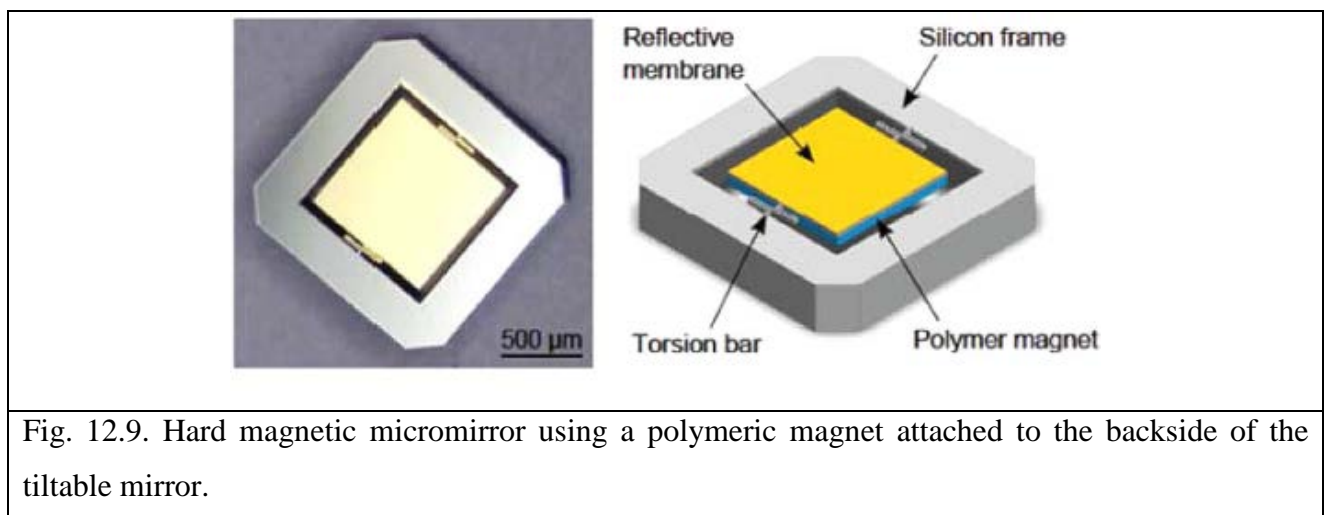
Electromagnetic drive MEMS mirror can be used to reliaze optical scanner. Fig. 12.10 describes the product configuration. This scanner is configured by aligning permanent magnets around the monocrystal silicon substrate that forms the moving plate (with the mirror and coil on its surface), beams, and supporting section. When the electric current flows through the coil formed on the periphery of the moving plate, running torque (Lorentz force) is generated by the interaction between the permanent magnets and the magnetic field. As a result, the moving plate can tilt to the position balanced with the righting moment of the beams. Since the Lorentz force is proportional to the current, by changing the current value, the moving plate can tilt freely. That is, the optical scanning angle (hereinafter "deflection angle") can be changed freely within the scope of rating.

Fig. 12.10. Electromagnetic drive MEMS optical scanner
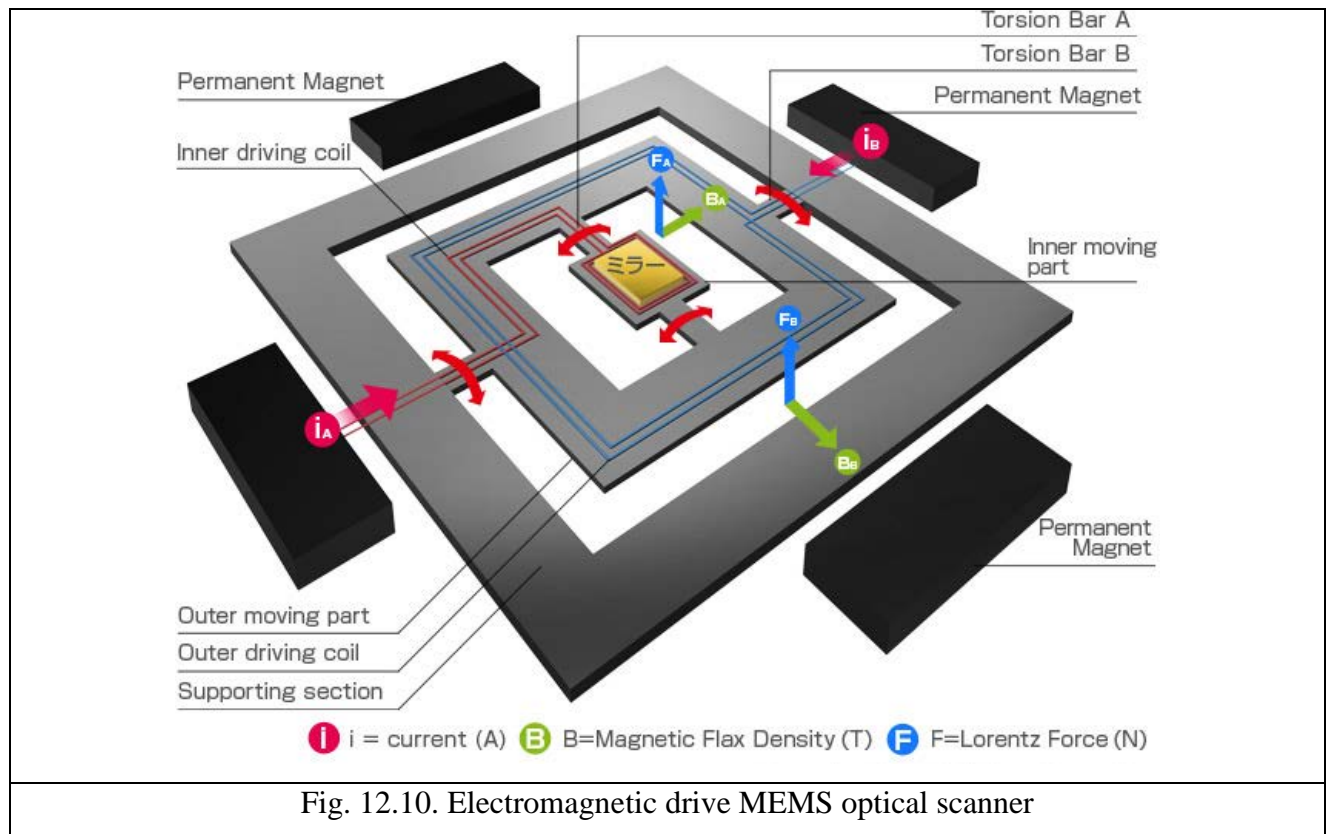
Thermal Actuation. Thermal actuators have also been used in a variety of micromirror configurations, including vertically-actuated mirrors for adaptive optic. Using bimorph flexures, which are actuated thermally and rely on residual stress incorporated into the supporting beams during fabrication, several micrometers of vertical mirror motion have been demonstrated in two-dimensional arrays with over 95% fill factor, the latter parameter of considerable interest in wavefront control applications.

The thermally actuated structure are based on bimorph layer on silicon-on-insulator wafers with the rextangular geometry with three arms for each contacting pad (Fig. 12.11a) The micromirror structure is composed by two layers, one layer is gold (0.5 pm thick) and the other one is thermal oxide (2 ptm thick). Thermal expansion coefficient of gold is larger than the thermal expansion coefficient of oxide, therefore the structure will bend down due to the Joule effect in the layers which are the actuator arm. The maximum stress apears in the metal (gold) over the refelective plate. This is due to different thermal expansion coefficients of the layers that form the reflective plate of the mirror. Gold has the highest thermal expansion coefficient therefore is the layer that is highly stressed. In Fig. 12.11b. it is shown that the minimum displacement on the vertical direction (Z), at the edge of the device is 13 μm for 100 mV applied. SEM image of thermal micromirror is depicted

SEM image with micromirror than can be bended due thermal actuator is shown on Fig. 12.11c. Frame thermal actuator and mirror actuator were used to obtain larger mirror displacement.
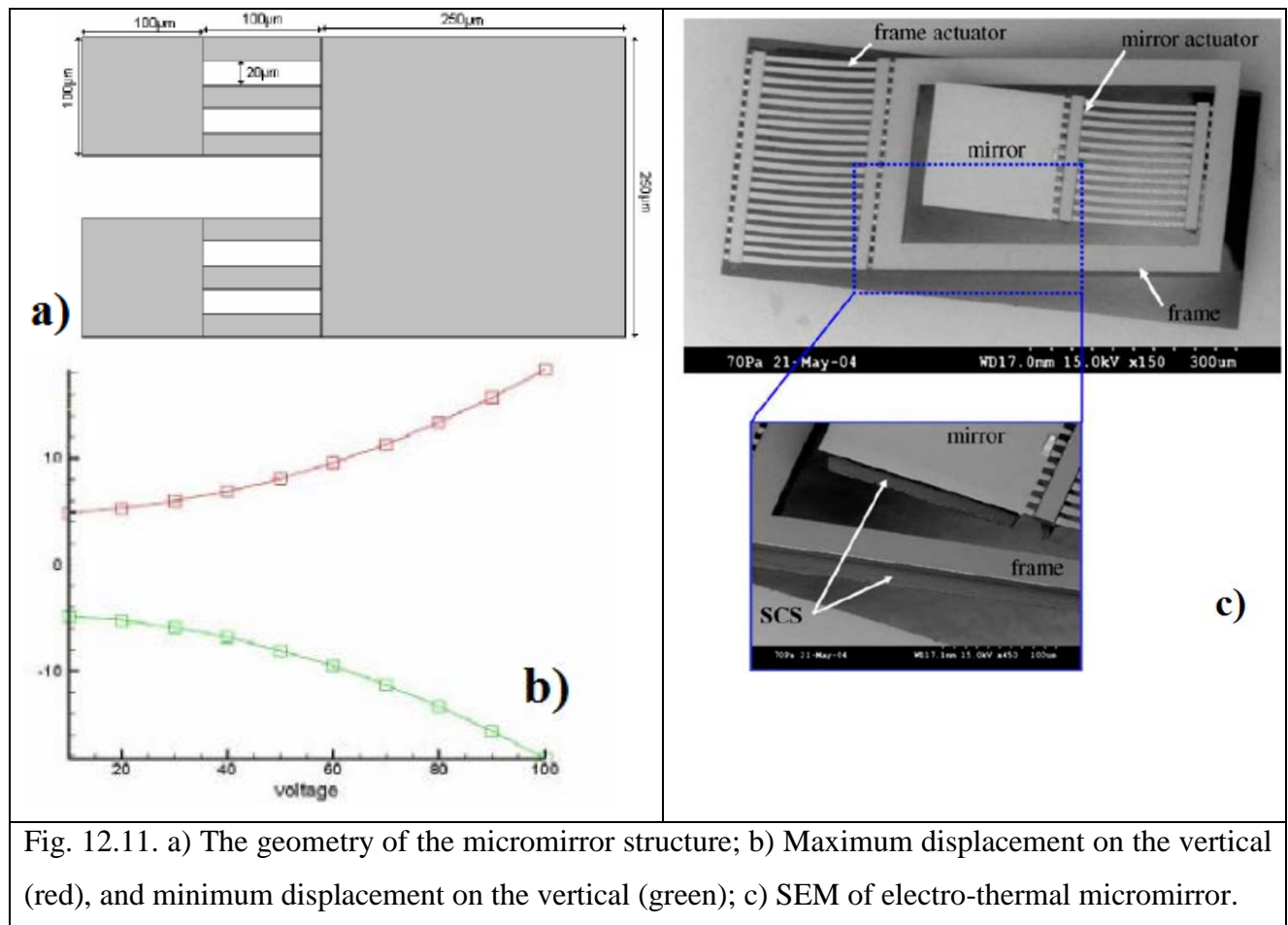
Fig. 12.11. a) The geometry of the micromirror structure; b) Maximum displacement on the vertical (red), and minimum displacement on the vertical (green); c) SEM of electro-thermal micromirror.

Pneumatic Actuation: Pneumatically actuated optical components, including mirrors and lenses can be used to assemble micropotical systems. Using gas-filled microfluidic chambers attached to cavities over which flexible polymer membranes are stretched, increase in gas pressure results in a distension of the membrane, typically in the vertical direction. Using silicon micromachining to generate micromirrors attached to these membranes, the pressure-based distension results in vertical or angular mirror motion. Mirror tilt angles of up to 75 μm and purely vertical motion of 80 μm have been demonstrated using on-chip integrated heaters for thermo-pneumatically actuated devices.

Sketch on Fig. 12.12a shows the pneumatically actuated mirror in deflected position at an arbitrary angle. The membrane-based, pressure-actuated, micro-mirrors employ a 60 μm thick silicon micro-mirror which is fixed to a 50 μm thick polydimethylsiloxane (PDMS) membrane mounted on a structured silicon wafer. The mirror surface itself is made of a gold or aluminium evaporated layer, which is located on the polished silicon wafer, below the PDMS membrane, implying that the light which is reflected at the metallized mirror surface travels through the PDMS membrane twice. The PDMS membrane and a glass substrate bonded to the backside of the structured silicon chip form a cavity, connected to the external driving pressure by a micro-fluidic channel inside the silicon chip. This cavity forms the pressure chamber. The mirror chip is silicon. 3D sketch of a deflected micro-mirror is shown on Fig. 12.12b. The silicon chip (top) is bonded to

the Pyrex substrate (bottom) and the mirror is attached to the inside of the PDMS membrane. The pneumatic access channels are seen on the left.
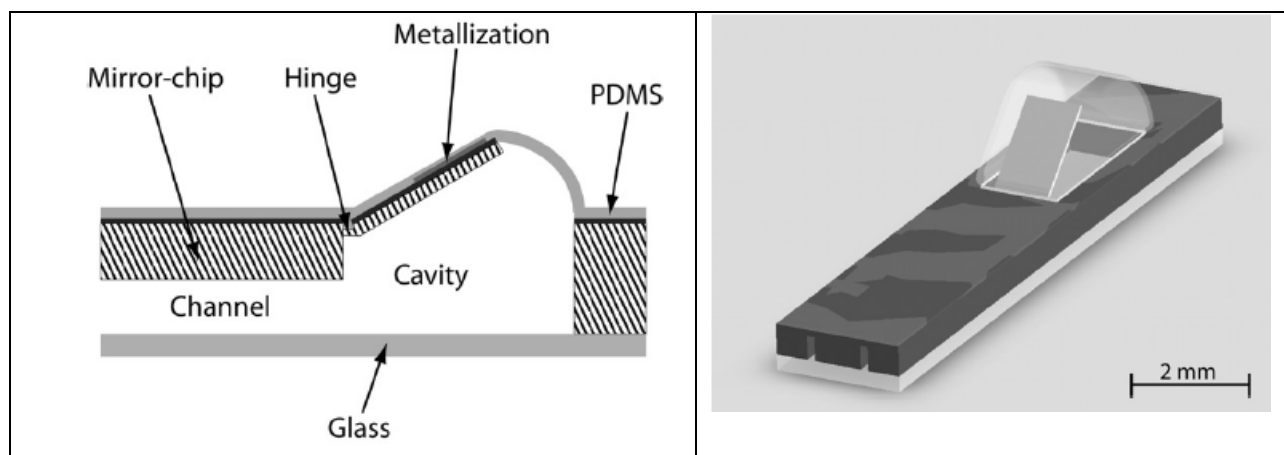
Fig. 12.12 Pneumatically actuated mirror device: a) cross-sectional view; b) 3D sketch of a deflected micro-mirror.

Figure 12.13a depicts a pneumatically actuated mirror device, glued onto a glass substrate, provided with a fluidic cavity. A plastic tube provides the pneumatic pressure. The distended PDMS actuator and the attached elliptical mirror at its maximum extension are clearly seen. Depenndence of mirror angle on pneumatic pressure is shown on Fig. 12.13b.
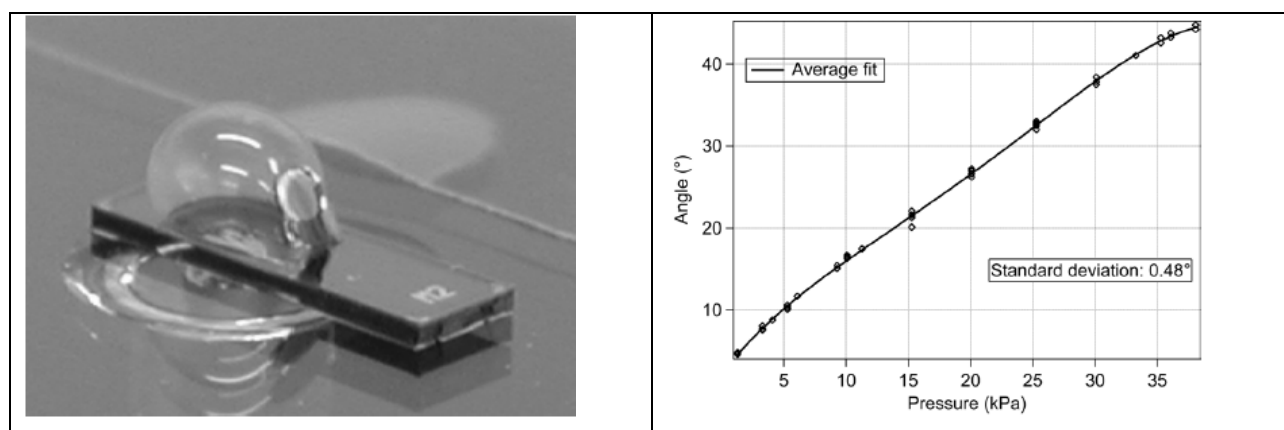
Fig. 12.13. a) pneumatically actuated mirror extended to an angle of 60°; angle versus pressure measurement of mirror.

Piezoelectric actuation. A Piezoelectric (PZT) actuator can be integrated with MEMS mirror to provide 2-D scanning opportunities as shown on Fig. 12.14. The mirror plate is driven by a S-shaped PZT actuator, which is capable of 6 degrees of freedom of movement. The actuator main composition consists of a top electrode layer (Pt/Ti), a piezoelectric thin film (PZT) and a bottom electrode layer (Pt/Ti). The top and bottom electrodes are each connected to their individual bond pads. A proof mass is left beneath the mirror to maintain the rigidity and flatness of the reflecting surface during motion. Bending mode occurs in both static and dynamic actuations. When an AC or DC bias is applied to the piezoelectric actuator, the actuator bends and causes the mirror to undergo

translational and rotational movement along the y-axis. Torsional mode is induced during dynamic actuation, when ac resonant frequency corresponding to rotational motion along x-axis is applied to the device.
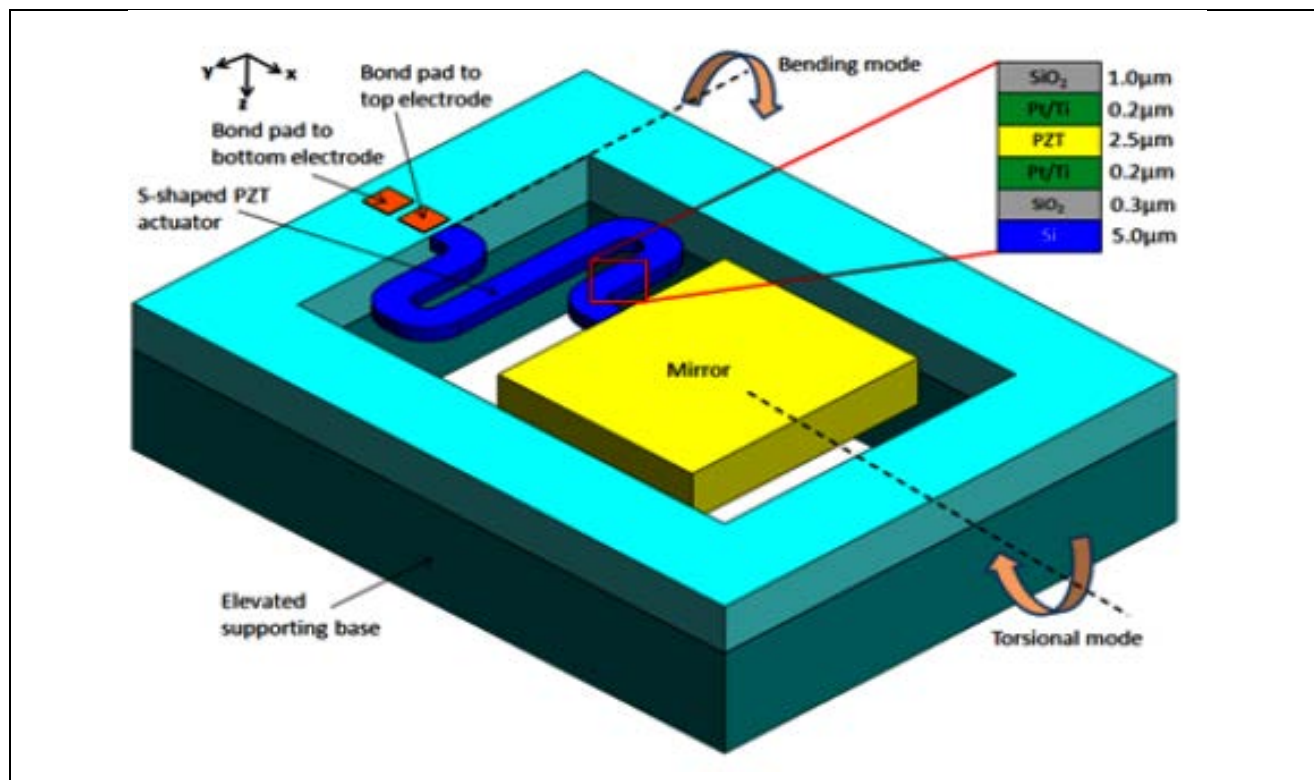
Fig. 12.14. Schematic drawing of 2-D scanning mirror actuated by S-shaped PZT actuator. Bending and torsional mode occur when the device is excited at their resonant frequencies respectively.

12.1.1.3 Tunable Fluidic Optics

Fluidics has become increasingly important for advanced micro-optics. The combination of MEMS fabrication techniques, controlled-liquid configurations and optics has led to the concept of optofluidics, a broad spectrum of technologies, in which fluids and photons play a dominant role in microoptofluidic systems. Optofluidics has attracted considerable attention due to its utility in realizing new types of displays but has also seen interesting applications in tunable lateral (in-plane) lensing structures, tunable photonic crystals, and fluidic lasers. Optofluidics for tunable micro-optical components, in particular lenses and apertures is briefly considered in this section.

Liquid Lenses. Liquid droplets on a surface make excellent spherical lenses. It has been shown that the curvature, and thus focal length, of these lenses can be controlled using electrowetting, an electrically-induced change in the liquid contact angle, and this concept has been extensively used to realize tunable microlenses. Using micromachining to define placement structures, individual tunable lenses with a stable lateral position and a tuning range between 2.3mm and infinity with less than 50V driving voltage can be fabricated. Alternative approaches using the dielectrophoretic effect, which requires the application of non-uniform electric fields, have also been employed, albeit with a smaller tuning range.

Electrowetting-on-Dielectrics.

The curvature of the spherical droplet surface (whether of a liquid in air or of a liquid embedded in a second liquid) is given by the contact angle θ_V, which is in turn defined by the surface energies of the boundary between liquid, substrate, and ambient. Contact angle θ_V can be varied by applying a bias between the fluid and its surroundings, an effect known as electrowetting. This effect forms the basis for an important branch of tunable fluidic optics. A variation of this phenomenon, electrowetting-on-dielectrics (EWOD), allowes controlled variation of liquid contact angles using reasonable (i.e., below 100V) voltages. EWOD uses the capacitive arrangement shown in Figure 12.15, in which the liquid droplet is deposited on an insulating dielectric over a conducting substrate. When a bias is applied between the droplet and the substrate, the contact angle θ_V then varies with applied voltage V. A liquid droplet is separated from a conducting substrate by an insulating dielectric with thickness t. A change in θ_V leads to a change in droplet curvature and thus tuning of the focal length EWOD has extensively been used for droplet manipulation in microfluidic systems and in displays.
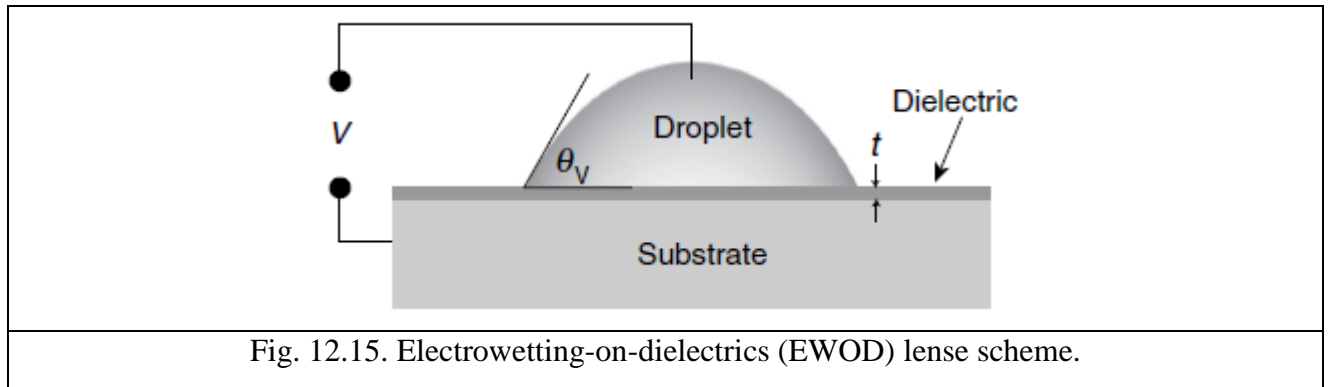
Fig. 12.15. Electrowetting-on-dielectrics (EWOD) lense scheme.

When used to tune the curvature, and hence the focal length, of a liquid lens, the contact angle change, due to EWOD given in Equation 1.1, results in a variation of focal length f of the form:

$$f(V) = \frac{D}{2(n_L - n_A)\sin\theta_V} \tag{12.1}$$

for aperture diameter D and refractive indices of the lens liquid and ambient, nL and nA, respectively.

For practical tunable liquid lens structures, EWOD typically uses a fluid ambient, such that the lens droplet is surrounded by a liquid electrolyte, through which the droplet is then electrically contacted. The lens and ambient fluids are chosen to have densities as closely matched as possible, thereby reducing or eliminating the effects of variable orientation, movement, or vibration Important for practical applications of this type of liquid tunable microlens is fabricating the entire structure in a laterally stabilized, sealed package. Using a glass cylinder for mechanical stability, as

seen in Fig. 12.16., one of the earliest fully packaged liquid lenses then employed ITO electrodes coated with parylene, a family of poly(p-xylylene) polymers, as an insulator and a hydrophobic surface coating on the inside surfaces. The two electrodes are on the bottom and side surfaces of the cylinder, where the variable contact angle is generated on the side surfaces.

Fig. 12.16. A liquid lens system packaged in a glass cylinder: a) convex lense; b) concave lense.
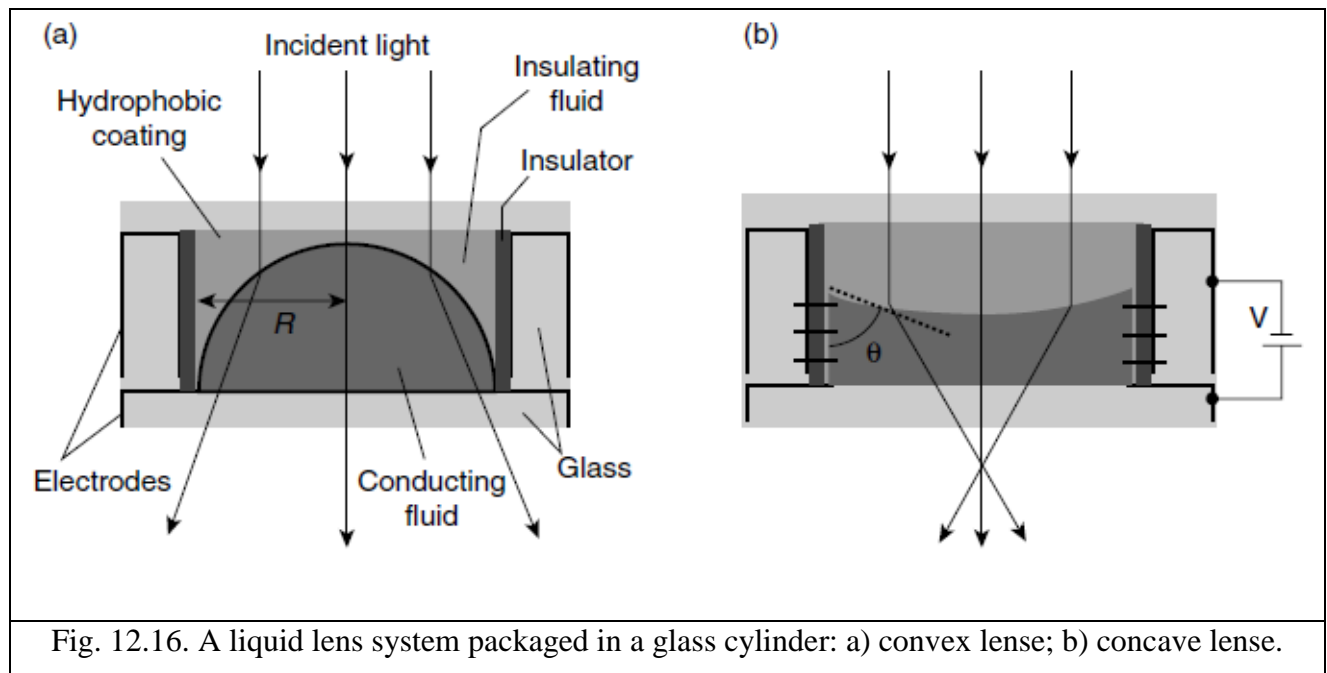
Ultimately integrated with a charge-coupled device (CCD) image sensor, the 5.5mm tall water/oil lens system demonstrated a focus variation between 50 and 200mm. Concepts for using this lens in an optical zoom system and correcting chromatic aberrations have suggested means for optimizing its optical performance.

Fig. 12.17. A microsystems-based liquid lens system fabricated using silicon and glass technologies. (a) Schematic cross section showing the etched silicon structure for lens positioning bonded to a pyrex substrate, the glass cover and the electrodes. (b) Photograph of the completed 8×8×8 mm3 system.
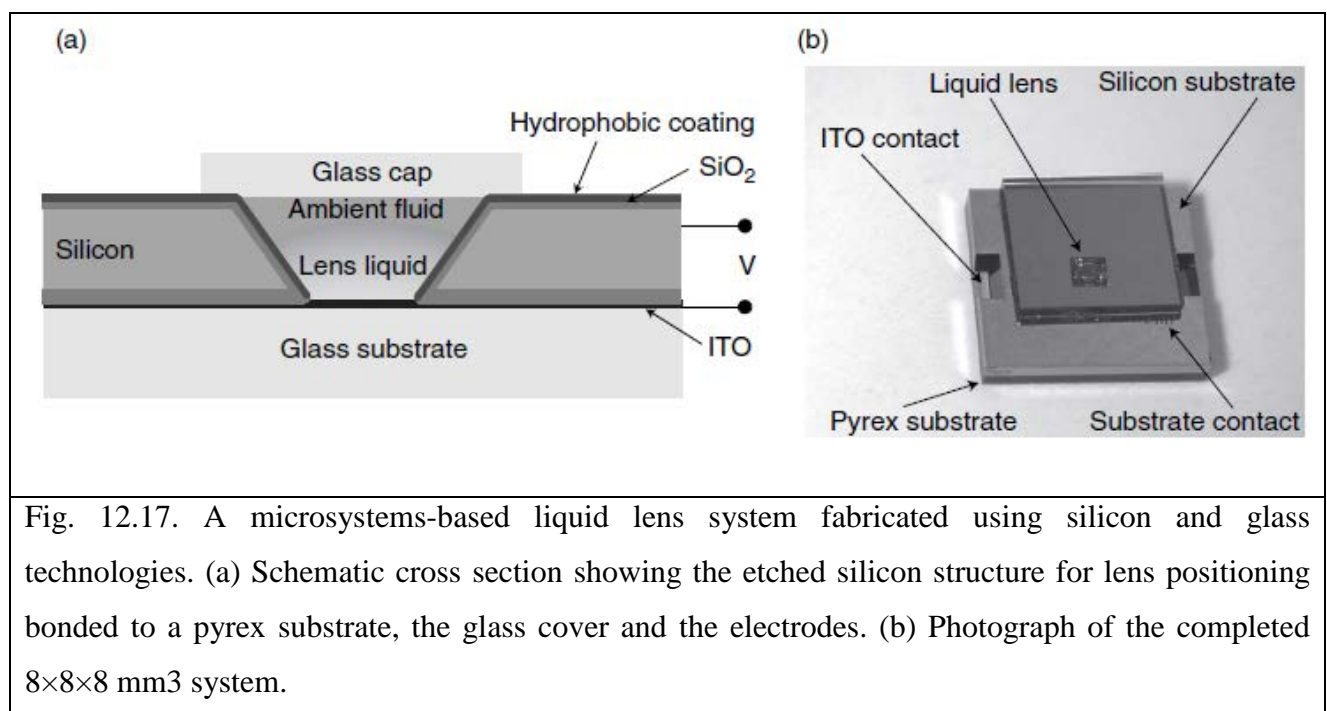
An alternative approach for liquid lens stabilization using microsystems fabrication techniques is shown in Fig 12.18. As seen in the schematic sketch of Fig 12.18a, the structure is

based on an etched V-groove structure in a silicon substrate, on which ITO electrodes are deposited. After filling with an aqueous inorganic salt solution, with a density of 2.1 g/cm^3 and a refractive index of 1.51, for the lens and a density-matched perfluorocarbon for the ambient fluid, with refractive index of 1.293, the entire structure was sealed using a pyrex substrate and a glass cover plate.

Means have been proposed to separate the electrowetting actuation from the lens itself, to essentially "pump" the lens liquid and thereby change the curvature of a spherical liquid surface pinned into position. This concept currently suffers from the fact that the lenses are formed by a fluid/air interface, and are thus sensitive to vibration, movement, and orientation.

Fig. 12.18 Liquid microlens array. a) finite imagery of a pattern in panel; b) generated image; c) scheme of microlens array tuned by electrowetting actuation of the fluid in the reservoir below.
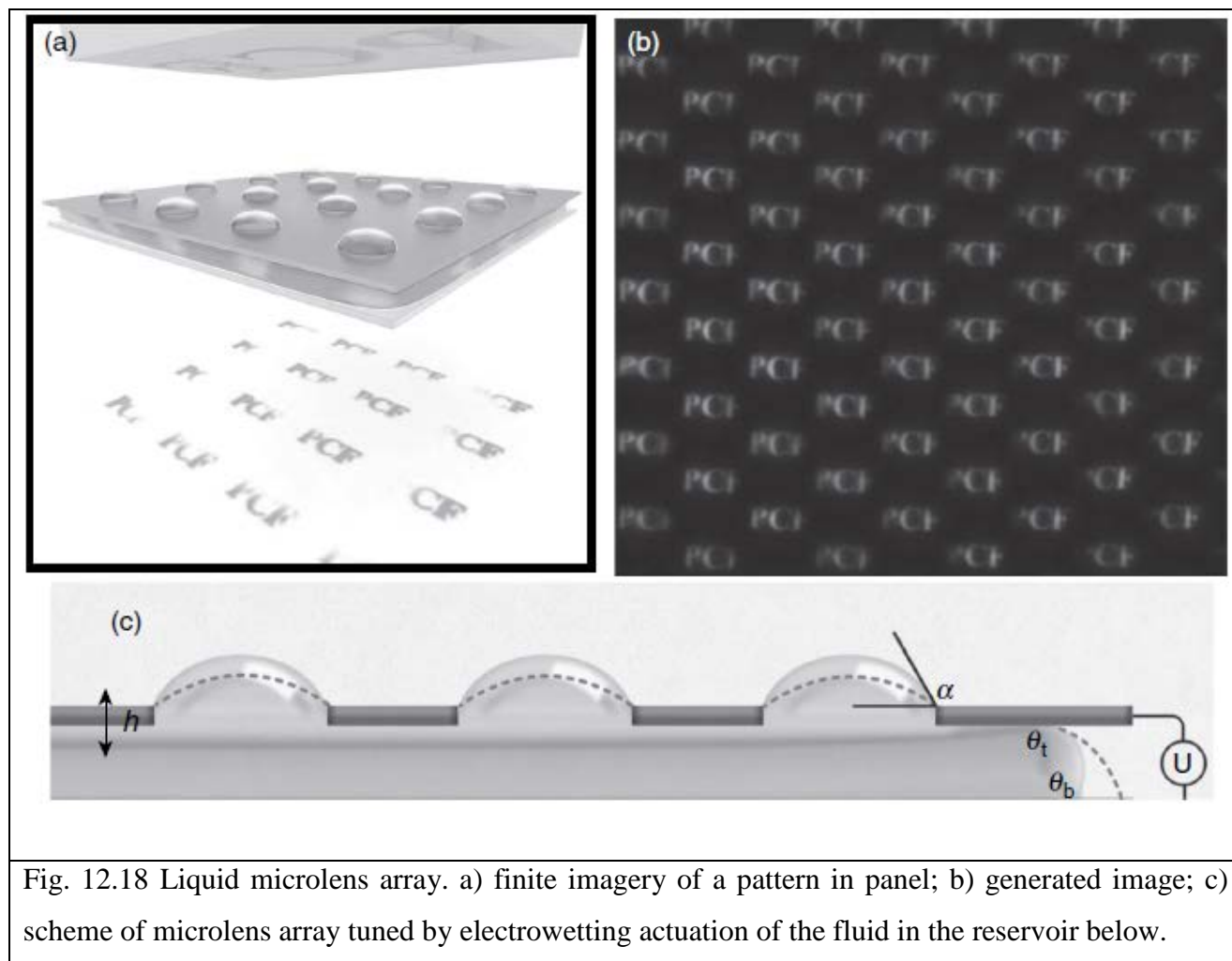
This idea has been extended to actuation of liquid lens arrays: Fig. 12.18 shows a two-dimensional array of lenses, all actuated from a single liquid reservoir. Using a water-based fluid, electrowetting actuation changes the pressure in the lower chamber, causing a distension of the lens menisci. The lenses may be tuned in focal length from 2 to 8 mm at frequencies greater than 1kHz, implying potential for use in tunable array imagers. Electrowetting-on-dielectrics allows one bit of further flexibility in lens design, namely the realization of reconfigurable two-dimensional tunable lens arrays. Since EWOD is well established as a mechanism for liquid droplet movement, as has

been demonstrated in a wide variety of microfluidic systems, the technique may also be used to reposition liquid lenses on a planar substrate.

The approach employs a microstructured surface incorporating a rectangular grid structure; this grid provides a mechanism for pinning the edges of a droplet on the surface which may be moved using the forces of electrowetting generated by a buried electrode array. It has been shown that two-dimensional positioning accuracy of 70 μm can be achieved. Using the same electrode configuration, the positioned lenses may also be tuned in focal length, in the range of 0.58 to 1.24 mm.

Electrowetting actuated tunable lenses with reconfigurable position were also proposed. As is seen on Fig. 12.19, a liquid microlens is accurately positioned on a structured substrate; once in the correct position its focal length may then also be tuned. In the shown structures, a positioning accuracy of 70 μm and a focal length tuning range of 580 μm to 1,240 μm was demonstrated

Fig. 12.19. An electrowetting-actuated repositionable liquid microlens moving in the vertical direction; when in position, its focal length can subsequently be tuned.
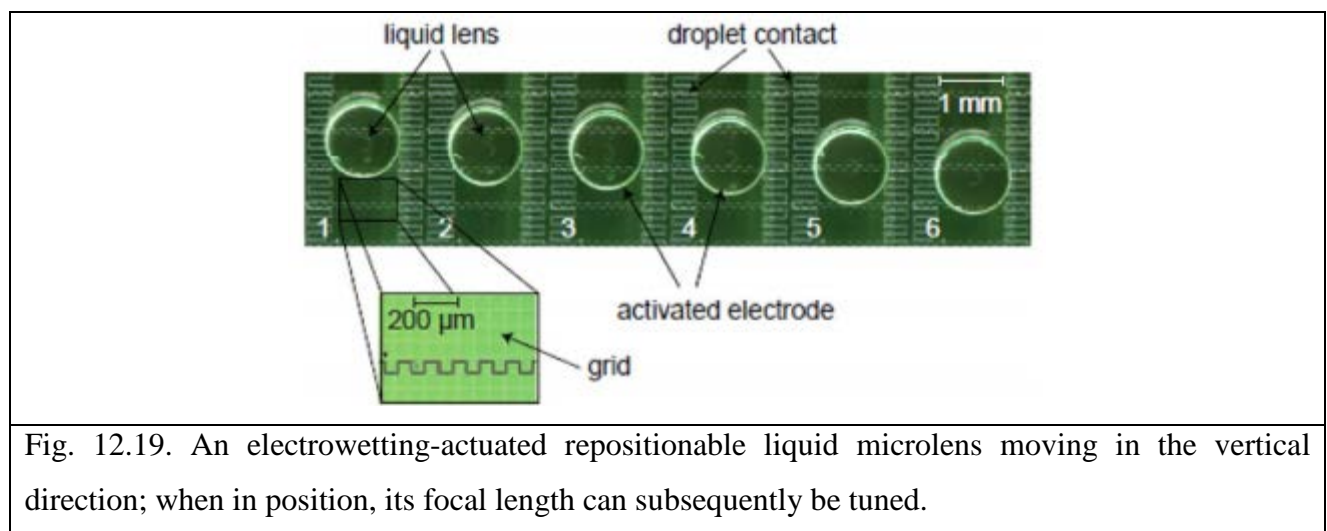
Dielectrophoresis lenses. An alternative to electrowetting for tunable liquid lens actuation is dielectrophoresis (DP). As we saw above, electrowetting is a surface phenomenon in that the contact angle is varied using an applied electric field; it requires a liquid with suitably high conductivity. In contrast, dielectrophoresis is a bulk effect and requires the use of two nonconducting liquids with differing dielectric constants. The application of a nonuniform electric field to this pair of liquids results in a movement of the interface due to induced polarization in the dielectric, so that the dielectrophoretic force F_D is proportional to the gradient of the electric energy and the difference between the dielectric constants of the two liquids (ε_A, ε_B), or for electric field E. When such liquids are considered, they should satisfy the following conditions:

1. They are nonconductive.
2. They are immiscible.
3. Their dielectric constants are largely different.
4. Their refractive indices are quite different.
5. Their gravity densities are well matched.

6. They are highly clear in the visible spectral region.

7. Both liquids should have no chemical reaction by light, thermal, or any other means.

8. They are not harmful or hazardous.

Particularly since no contact between electrode and droplet is required, dielectrophoresis has also been extensively applied to droplet manipulation in microfluidic systems. Typical dielectrophoretic systems use concentric rings of electrodes to generate the required field gradients, and while requiring relatively high voltages, power consumption is low since there is no current flow. Individual dielectrophoretic microlenses have been successfully demonstrated using two immiscible liquids and a suitably patterned electrode array, as seen in Fig. 12.20.
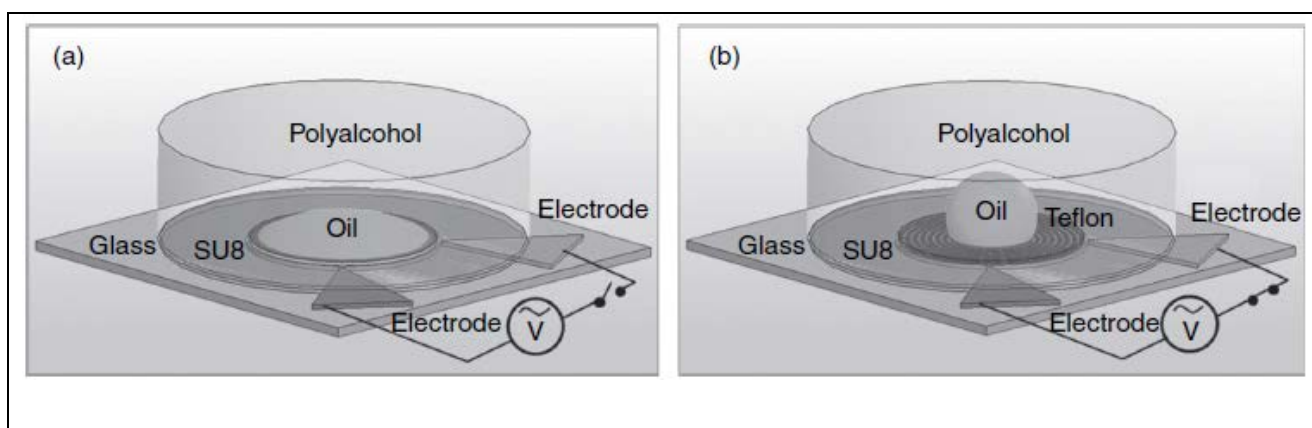
Fig. 12.20. Dielectrophoretic liquid lens, consisting of an oil droplet suspended in alcohol actuated using a structured electrode (droplet size is 500 μm): a) no applied voltage; b) actuated.

For a lens with 3mm aperture, a focal length tuning range of 12 to 34 mm was demonstrated, although applied voltages of up to 200V were required. As with all liquids, the refractive index of the fluids used for dielectrophoretic lenses vary with temperature, such that focal length is temperature dependent; optimization of the liquids employed and limiting the temperature ranges can minimize this effect. Dielectrophoresis has also been used to actuate liquid-crystal-based lenses of interest, since liquid crystals have a high refractive index and are not susceptible to electrolysis, a problem for aqueous materials for high electric fields. It has also been shown that dielectrophoretic lenses can be fabricated on flexible substrates, albeit with relatively slow response times and requiring high voltages.

A number of mechanical actuation mechanisms, with which the pressure on the lens liquid may be varied, have also been employed for tuning liquid lenses. One useful technique has been the use of hydrogels, part of a family of highly hydrophilic polymer networks which can absorb large volumes of water, thereby undergoing significant expansion. As seen in Fig. 12.21, using an immiscible oil/water combination, the pressure increase in a water-filled chamber due to the expansion of a surrounding hydrogel induces a change in the curvature of the water/oil interface and thus the focal length of the resulting convex lens. . The hydrogel in the water-filled chamber

absorbs or expels water as a function of temperature, changing the pressure and thus the meniscus shape of the curved oil/water interface.

Fig. 12.21. An oil/water-based liquid microlens actuated using a hydrogel: a) 3d view of construction; b) cross-section of construction.
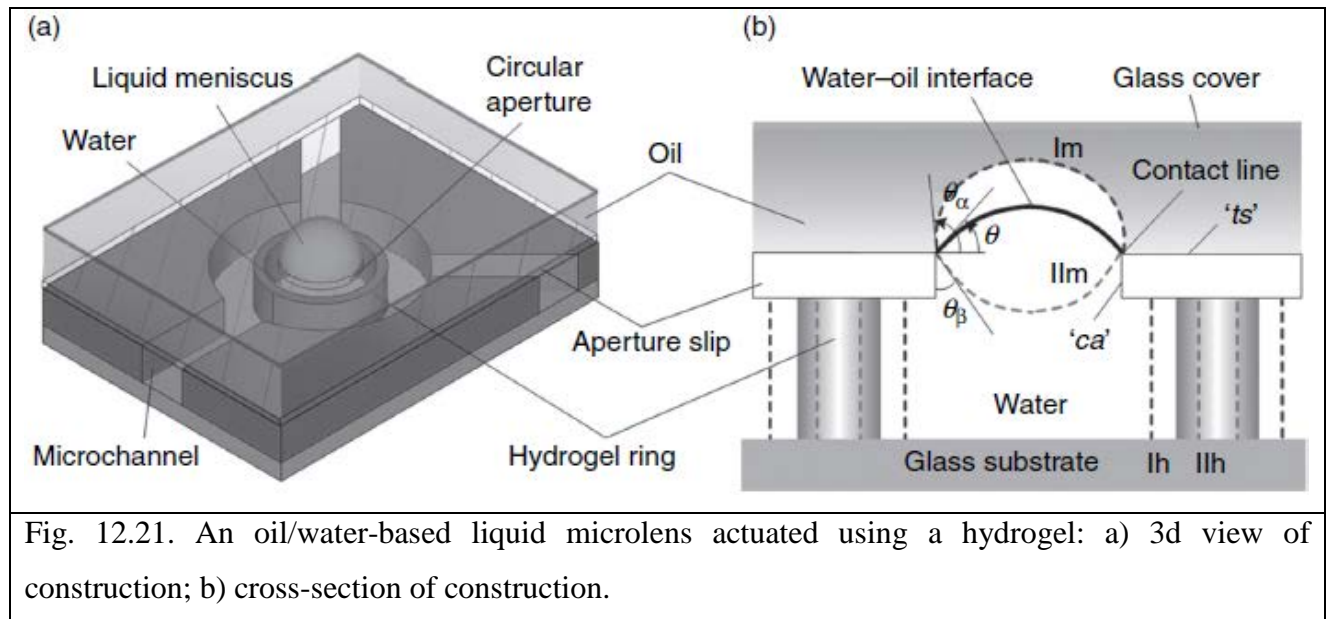
To lower the operating voltage for a large droplet size (due to the thick cell gap), using fringing field to actuate the liquid lens has been considered. Figure 12.22a depicts the side view of the device structure. The ITO electrode on the bottom glass plate is etched with concentric rings (Fig. 12.22b). A dielectric layer is coated over the substrate surface to cover the electrode. When a voltage is applied across the terminals A and B, fringing electric fields are generated across the neighboring electrode rings with a gradient. The gradient of the electric field is mainly determined by the external voltage and the gap between the electrode rings.

Fig. 12.22. Dielectrophoresis lense contacts: a) schematic cross-sectional view of the bottom substrate of dielectrophoresis lenses with electrode and dielectric layer; b) the ITO electrode with concentric pattern; c) schematic cross-sectional view of a deformable liquid crystal droplet (not to scale) in voltage-off state (solid line) and DEP force actuated state (dashed lines).

At V = 0, the droplet is in relaxed state with a minimal curvature. When a voltage is applied to the electrodes, the border of the liquid droplet is subject to an electric field with the highest gradient. Because the dielectric constant of liquid is larger than that of air, the generated DE force pulls the liquid droplet to expand outward along the substrate surface where the electric field is the strongest and occupied by air. This liquid expansion causes the droplet shape to change. As a result, the focal length of the droplet changes accordingly. At a certain voltage, the shape change of the droplet will stop when the generated DE force is equal to the surface tension of the droplet.

Fig. 12.23. Images observed through an dielectrophoretic liquid droplet lens at different voltages.
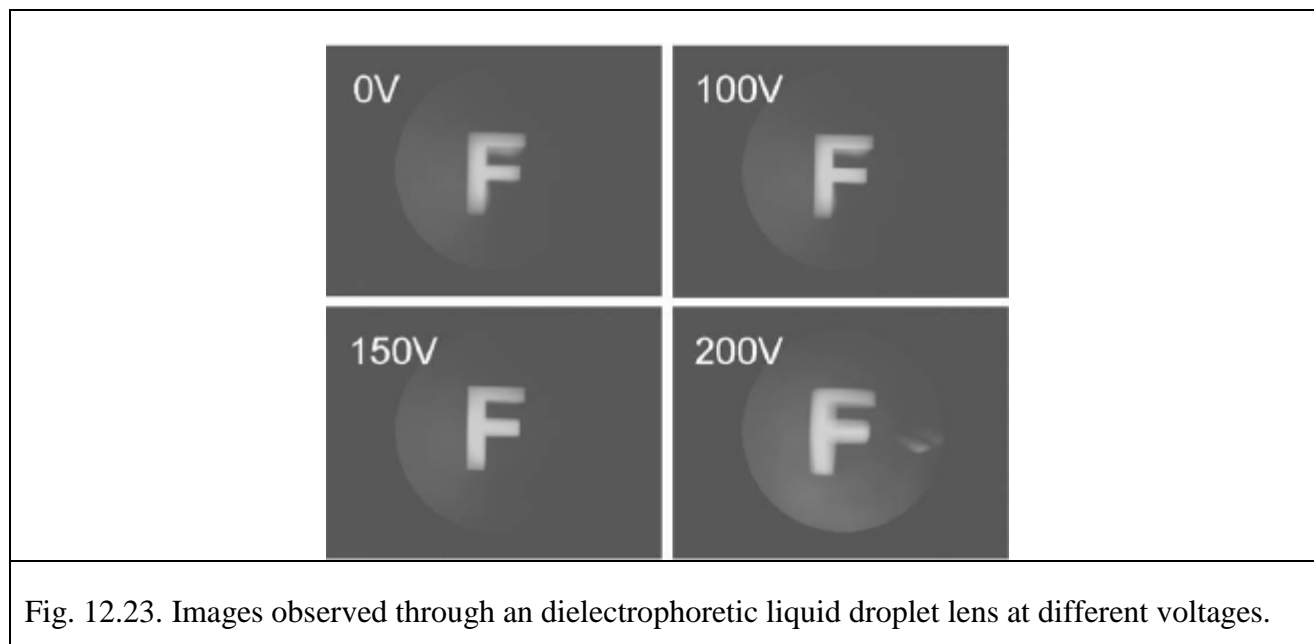
Pneumatic membrane lenses. An alternative to purely liquid lenses, in which surface tension defines the curvature, is the use of membrane-based lenses, actuated pneumatically. Using the same technology as for actuating micromirrors, the distended membrane covering a fluid-filled cavity forms a pressure-tunable refractive surface. Pneumatic tuning allows the generation of convex, planar and concave profiles in the same lens, and yields a correspondingly large tuning range; the lack of required electrical power means that lens structures like this are ideal for use in medical or endoscopic imaging, for which high voltages present challenges. Alternatively, completely integrated pneumatic lenses, with on-chip thermal actuation, allow compact tunable lens systems with no required pneumatic connections. The technologies of MEMS also allow realization of more complex optical arrangements using the membrane lens approach. Combining two independently tunable lenses, a zoom system with a zoom ratio of 2.8 in a compact, 18 mm long optical system, has been demonstrated. Furthermore, the use of custom liquids, with well-defined refractive index and dispersion properties, allows the realization of tunable achromatic doublet lenses, which are continuously corrected for chromatic aberration throughout the tuning range. These structures combine classical micromachining with silicone, rather than silicon, materials.

Mechano-fluidic lens systems, however, usually require external mechanical pumping components, which make the system large and impracticable with respect to integration. MEMS-

based mechanisms may be considered for actuation, and a variety of pneumatic and hydraulic micro-actuators have been proposed; piezoelectrically actuated micro-pumps, thermo-pneumatic micro-pumps and a variety of pneumatically actuated optical devices have been demonstrated.
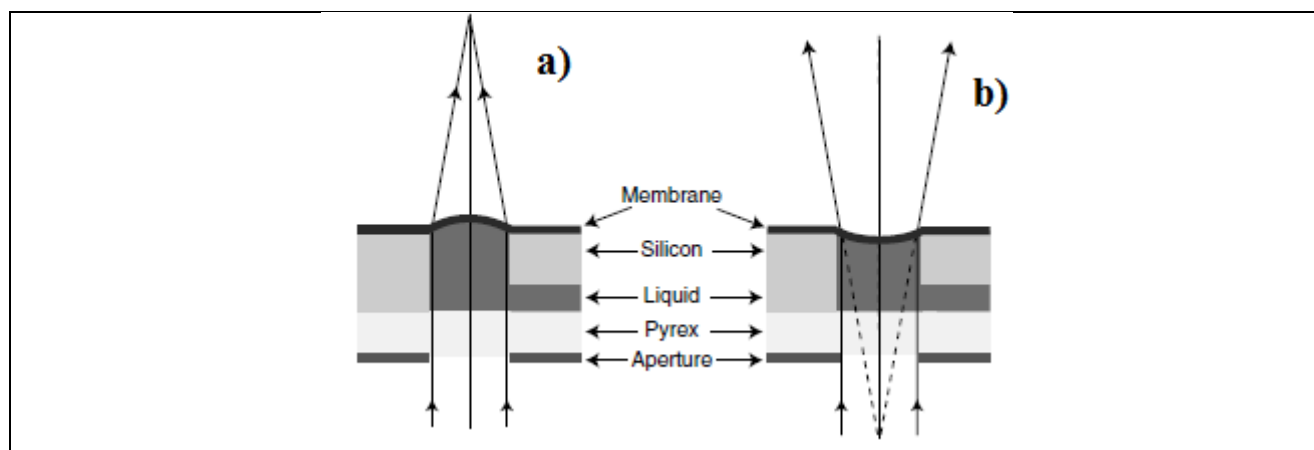
Fig. 12.24. Schematic cross section of a hydraulically tunable membrane microlens. Pressure is applied to a liquid-filled microfluidic channel etched into silicon and sealed using a glass substrate and a polydimethylsiloxane (PDMS) membrane; pressure applied to the liquid causes the membrane to distend, forming a convex (a) or concave (b) lens.

A second family of tunable liquid microlenses is that of hydraulically tunable fluid-filled membrane lenses. In contrast to the liquid lenses, hydraulic lenses employ optical fluids fully enclosed in microfluidic systems. As shown in Figure 1.7, the lens may be realized by capping part of the microfluidic chamber with a distensible, optically transparent membrane; by increasing the pressure on the fluid in the chamber, the membrane expands, generating a convex profile, whereas negative pressures yield concave profiles. The lens functionality is still generated by the body of the fluid, but its profile is defined by the distended membrane. Hydraulic or pneumatic actuation1 has also been shown to be useful for tuning other micro-optical components, most notably micromirrors.
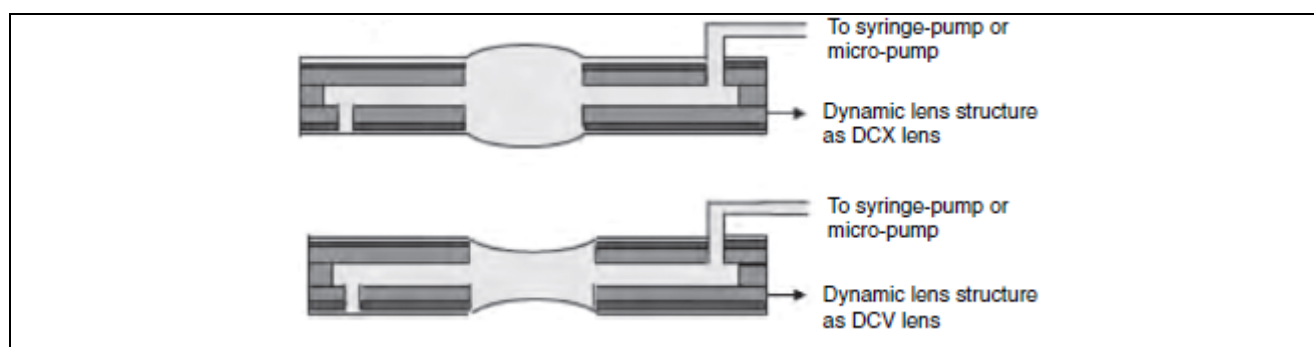
Fig. 12.25. Schematic diagram of a double-sided hydraulically actuated microlens that may be configured as a biconvex (labeled DCX) or biconcave (DCV) lens, depending on the applied pressure.

Using of microfabrication techniques, especially deep dry etching, allows realization of biconvex or biconcave tunable lenses with a focal length tuning range of −76 to +76mm (Fig. 12.25)

Piezoelectrically actuated micropumps have very distinct advantages in reliability and fast response speed. However, the required voltages are generally high between 10 and 500 V. Thermo-pneumatic pumps feature usually a large membrane displacement at low voltages and low power consumption.

Integrated thermo-pneumatic actuation can be combined with a mechano-fluidic tunable microlens to realize a completely integrated pneumatically-tunable microlens system. A membrane-based fluidic lens is integrated with a thermal pump actuator and temperature sensor chip, and the optical performance is extensively characterized. The design, fabrication, and operation of the lens is discussed, along with an analysis of the thermal structure, properties of the liquids employed, design of the control circuitry and the response time of thermo-pneumatic lenses. As shown schematically in the cross-section of Fig. 12.26, the thermo-pneumatically tunable lens is based on separate pump and optical cavities in a microfluidic system. A liquid-filled lens cavity is bounded by a thin, distensible membrane, which forms the refractive lens surface. Inside this cavity are four thermal pumps, air-filled cavities with integrated heater and temperature sensor structures, separated from the lens cavity by a second set of flexible membranes. The lens cavity is filled with an optical liquid and the chambers of the thermal pump are filled with air. Thermal expansion of the air causes compression of the liquid and distension of the lens membrane, shown as a concave surface at the top of the structure.

Fig. 12.26. Schematic design of the thermally tunable lens, with the device dimensions shown.
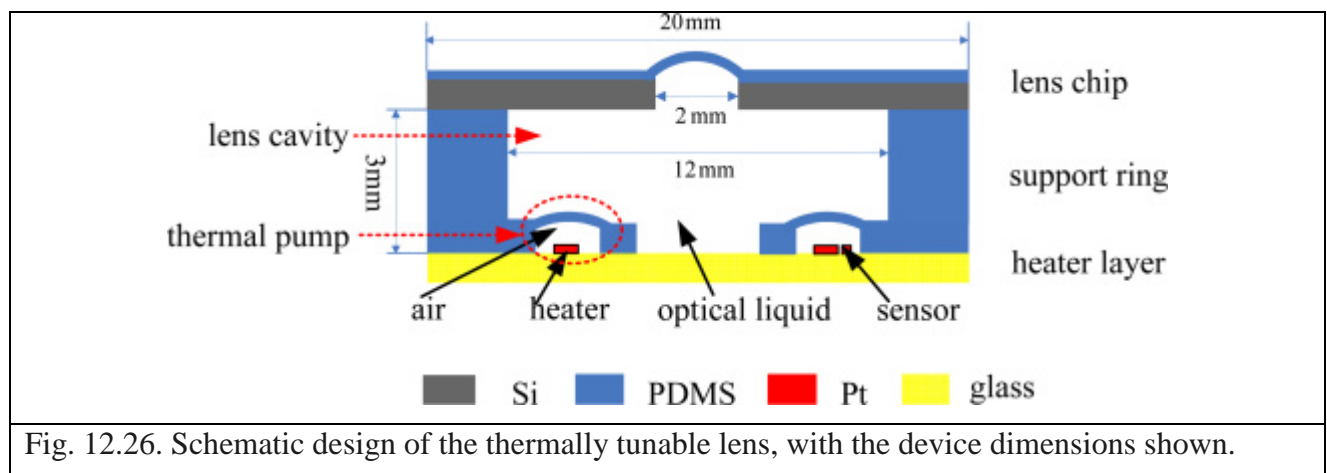
Upon heating, the air in the thermal pump expands, causing the membranes to likewise expand and compress the optical liquid. This compression, in turn, causes a distension of the optical lens membrane, and the resulting change in curvature gives rise to a tuning of the lens focal length.

Fluidic apertures. In addition to lens structures, optofluidics has also proven to be of value in generating novel, high performance tunable apertures and shutters. Rather than using liquids to

define refractive surfaces, these devices employ opaque and transparent liquids, which are switched using electrowetting and other microfluidic techniques.
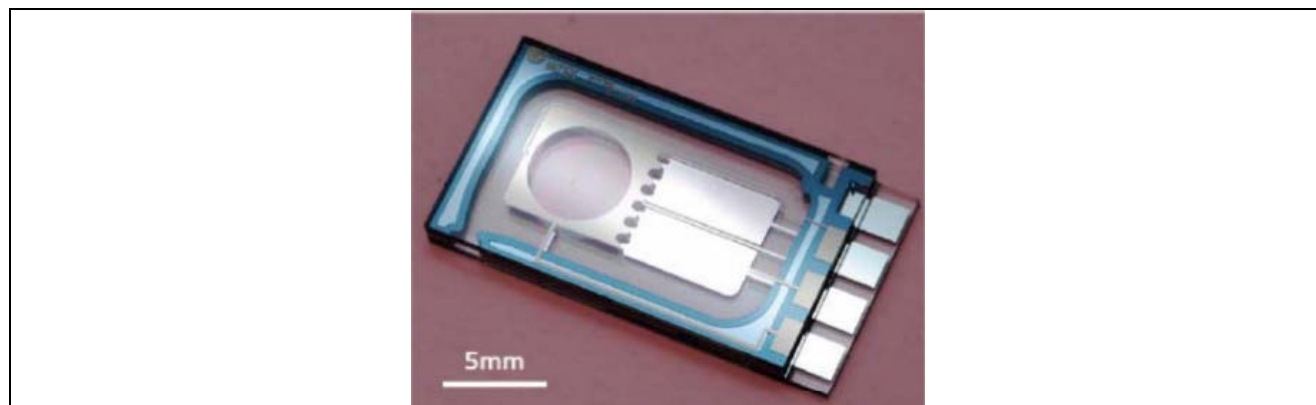
Fig. 12.27. An optofluidic shutter after chip dicing and prior to liquid filling; an opaque liquid is pumped into and out of the circular aperture using electrowetting.

The optofluidic shutter on Fig. 12.27 uses a highly absorbing ink, which is alternately switched into and out of the circular aperture. This completely integrated device, requiring only electrical contacts, demonstrated switchable attenuation of 47 dB with switching times below 100 ms; the only moving parts were the two liquids, one absorbing and the other transparent.

The attenuation concept can be refined further through the realization of a tunable iris, shown in Fig. 12.28, which allows variable attenuation of an optical beam, again using only moving opaque and transparent liquids. An opaque ring of ink forms a circular aperture that is tuned by means of electrowetting on dielectrics.
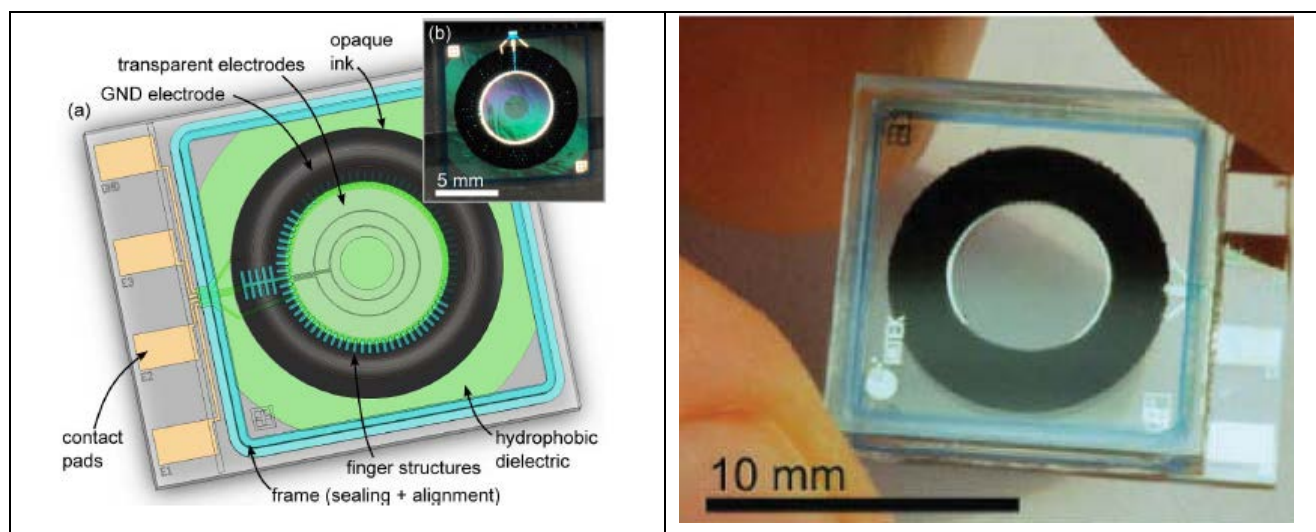
Fig. 12.28. Optofluidic micro-iris: a) illustration of the optofluidic microiris design; b) a fully integrated prototype optofluidic micro-iris in its fully open setting.

Using circular electrode arrays, the aperture could be opened and closed in four discrete steps with response times in the range of several hundred milliseconds. Using index-matched transparent fluids in this fully-integrated design, transmission characteristics in the open parts of the aperture showed highly uniform transmission with very low wavefront distortion.
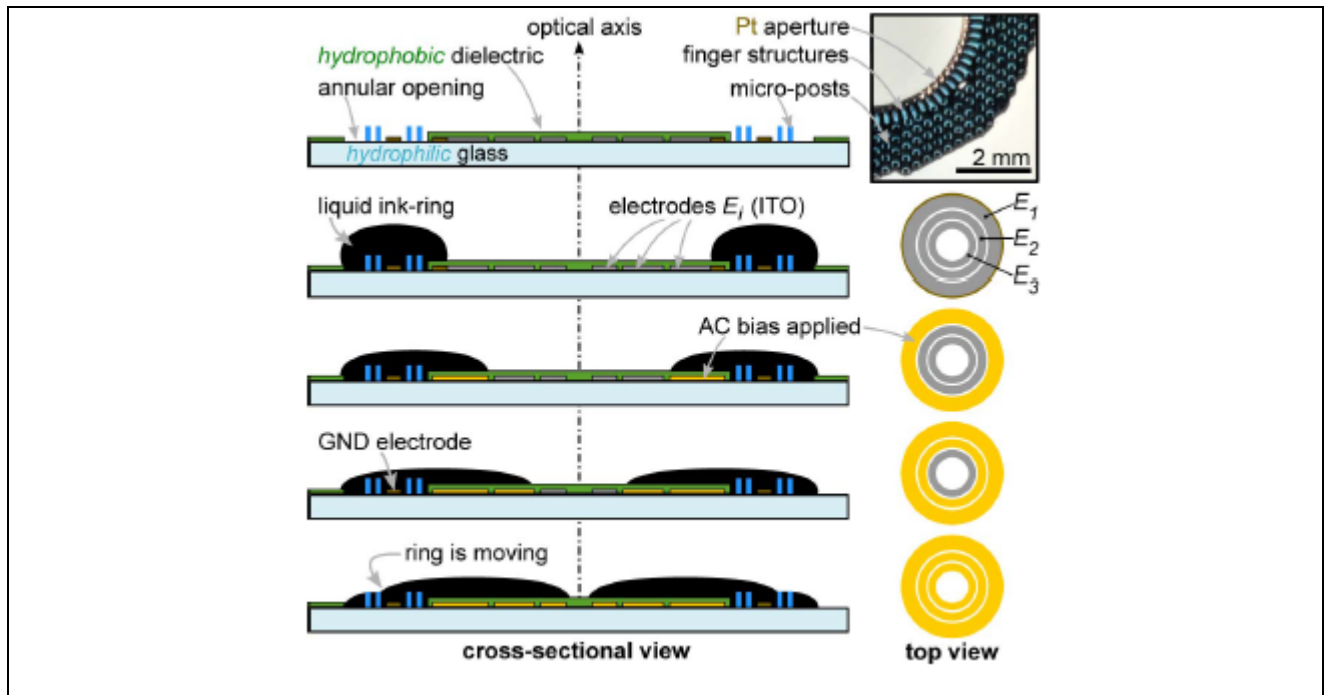
Fig. 12.29. Schematic of the *digital* tuning mechanism: the aperture is closed (opened) stepwise by applying (removing) ac bias to the three concentric actuation electrodes *Ei*, from outermost to innermost (vice versa). The small top-view photograph on the upper right shows microposts, which become visible after removing most of the ink. These structures, made from permanent dry film resist, are slightly hydrophilic and therefore highly beneficial for positioning and dosing of the liquid ring.

As shown by the cross-sectional view of Fig. 12.29, an annular opening is etched into the dielectric. Thereby, a strong wettability contrast is created between the hydrophilic glass area in the opening and the strongly hydrophobic dielectric. This contrast is enhanced further by micropost structures in the annular glass area, also is shown on Fig. 12.29, which increase the effective hydrophilic surface per substrate area. As a result, the aqueous ink will self-assemble into the desired toroidal shape, as soon as a defined volume of ink is dispensed onto the hydrophilic ring structure.

Fig. 12.30. Photographs of the actuation of a microiris chip, demonstrating reversible aperture tuning.

The aperture of the microiris may be tuned by applying/removing ac bias to the three individual actuation electrodes E_i. Such digital aperture tuning is seen clearly in the sequence of Fig. 12.30. The advantages of the digital tuning mechanism are apparent, namely, robust and fully reversible tuning and well-defined circular apertures that remain concentric in all aperture states.

The size of the electrodes was designed such that the aperture area decreases/increases by a factor of two at each of the four tuning steps. However, any other aperture ratio, as required by a specific application or, for example, elliptical apertures, may be defined, simply by changing the size or the shape of the actuation electrodes, respectively.

Shutters types. Variable optical attenuators (VOAs) with a single shutter blade as shown in Fig. 12.31a have been constructed using different MEMS technologies . Generally, the blade motion is in plane, but depending on the configuration the optical beam may travel parallel to or perpendicular to the substrate. When operated dynamically, such devices may act as optical choppers. Electrostatically driven twin-blade devices have recently been demonstrated and a translating aperture has been used in a spectrometer.

Fig. 12.31. Types of shutter: a) single blade; b) two blades.

Fig. 12.32. MEMS optical choppers: a) scheme of work; b) SEM image.
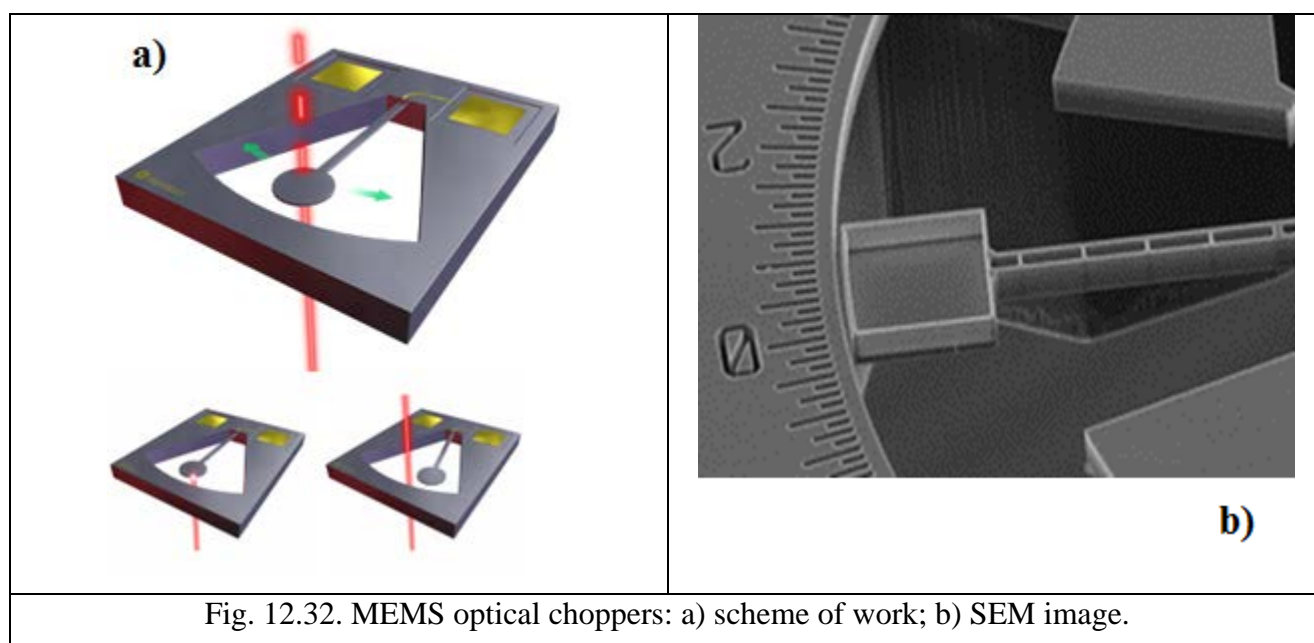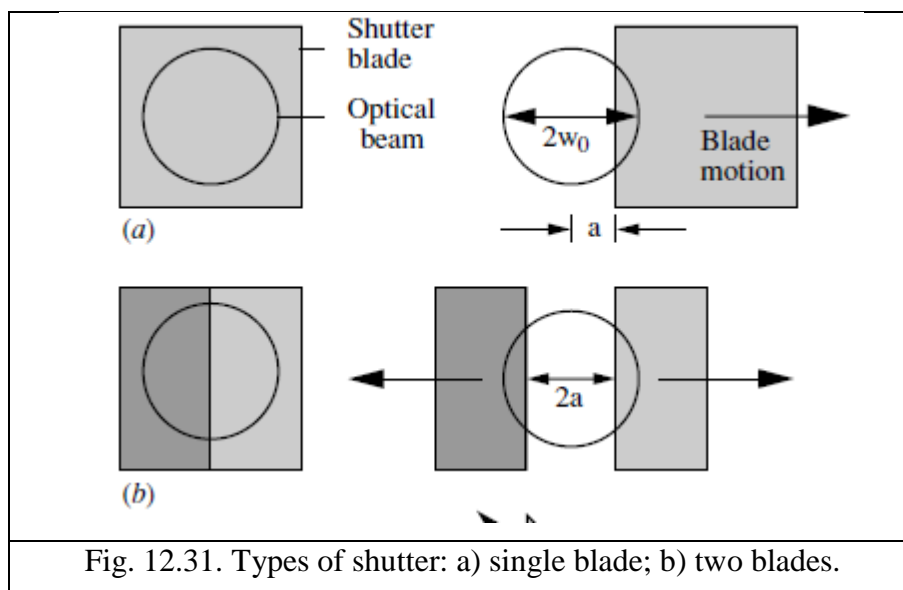
Devices with two shutter blades operating symmetrically, as shown in figure Fig. 12.31b, have the advantage of predictable wavelength-dependent loss (WDL), since the main peak of the far-field diffraction pattern obtained when an apertured beam exits the device must lie on-axis.

Example of MEMS optical chopper is shown on Fig. 12.32. The shutter chip has a clear optical path which can be interrupted by a blocker that moves in the plane of the chip (Fig. 12.32a). SEM image with clearshown blocker is shown on Fig. 12.32b.

A two blade device is merely a special case of a general *N*-blade polygonal shutter, which can be created from a set of symmetric triangular elements. These elements are initially tiled together to close the pupil, with one of their vertices lying on-axis, and are then translated in quasi-tangential directions to open the pupil, with each element sliding against its neighbours. General scheme of control aprerture diameter is shown on Fig. 12.33. If the blades and their motions lie in plane, a complete iris may be formed, by removing the substrate beneath the pupil.
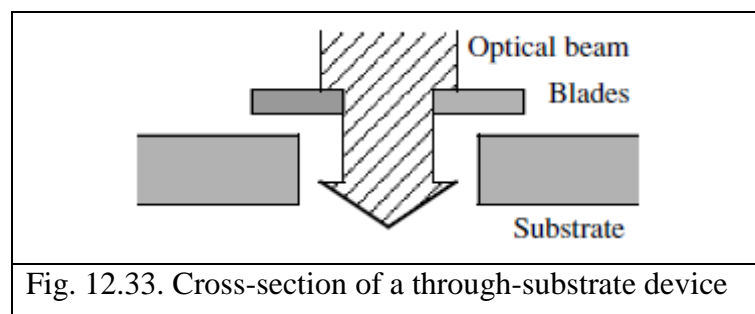
Fig. 12.33. Cross-section of a through-substrate device

Figures 12.34a and 12.34b show four-blade and eight-blade devices based on this principle, respectively. In the former case, a square aperture is created, and in the latter, an octagonal one. These devices may all be constructed in 'normally open' variants. As the number of blades rises, the pupil tends to a circle. Pupils with increased symmetry may be useful in VOAs with low WDL and polarization-dependent loss (PDL). Alternatively, when combined with other variable components, such as deformable mirrors and movable or tunable lenses, they may find applications in miniature imaging systems or adaptive optics. The use of a multi-blade shutter is clearly similar to the conventional photographic iris, and hence refer as an iris VOA. In a macroscopic iris, the blade motions are linked by a sliding rotary drive. Using MEMS technology, which allows complex mechanisms to be duplicated very simply and accurately, it is possible to obtain a similar effect using a number of entirely separate linear actuators operating together. In a two-blade device, the shutter blades will clearly strike each other if driven in the opposite direction to that used to open the pupil. However, in a four-blade device, the blades may be driven in either directionwithout colliding. This characteristic allows some protection against damage caused by under-damped oscillations. SEM images of four-blades iris iris at different drive powers is depicted on Fig. 12.34.
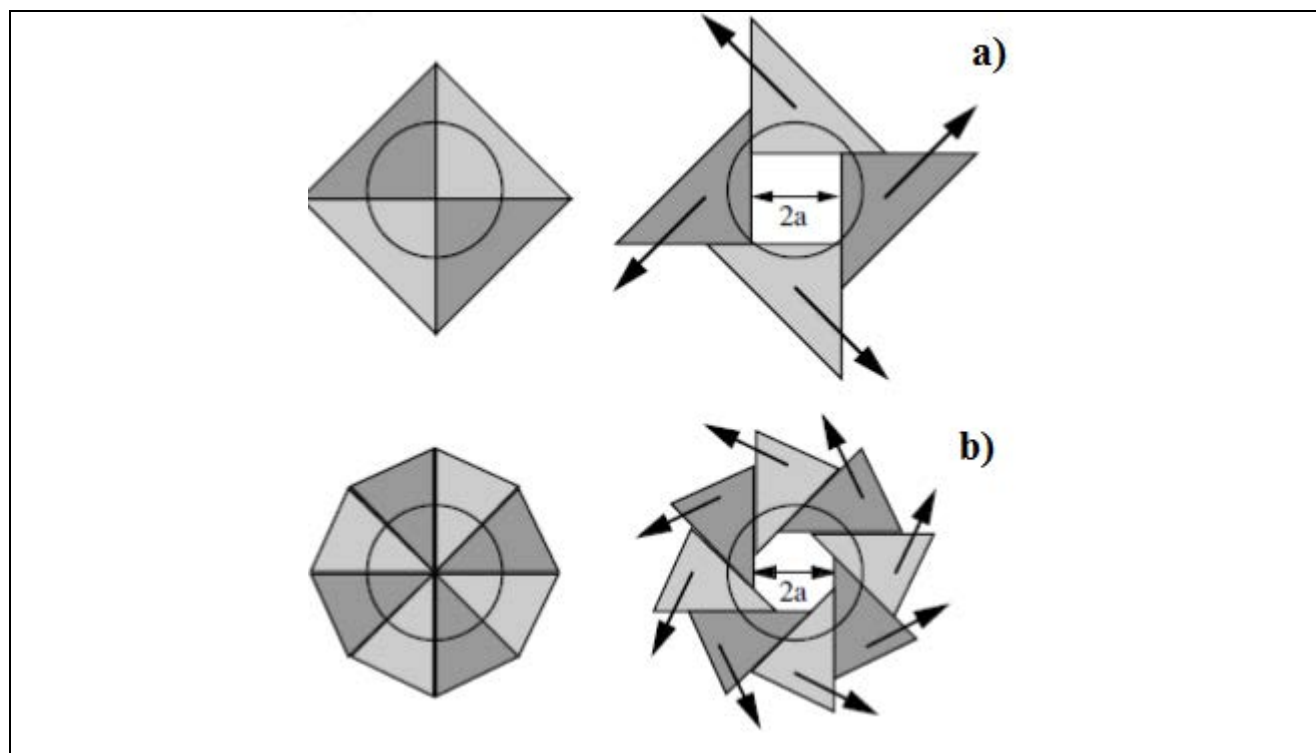
Fig. 12.34. Types of shutter: a) four blade; b) eight blades.

(i)

(ii)

(iii)

(iv)
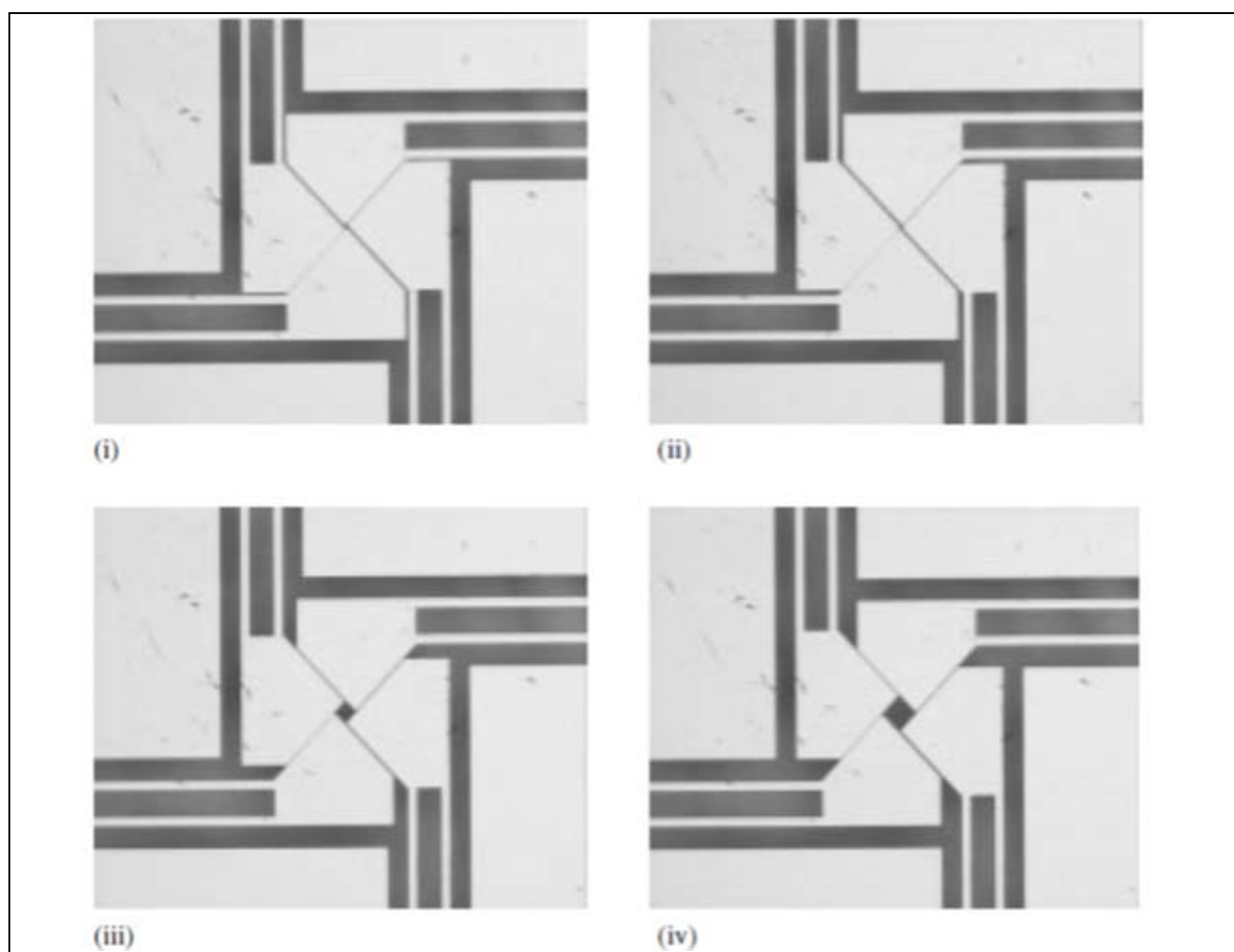
Fig. 12.34. Optical microscope views of iris at drive powers of (i) 0 mW, (ii) 80 mW, (iii) 460 mW, (iv) 1.1 W;

MEMS adjustable iris diaphragm design is proposed, in which eight blades are combined to form a polygon-shaped aperture. Different from the commonly used single layer configuration, these blades are evenly arranged in two separate chips in current design.

Fig. 12.35. Schematic of the tunable iris diaphragm: one chip layer.
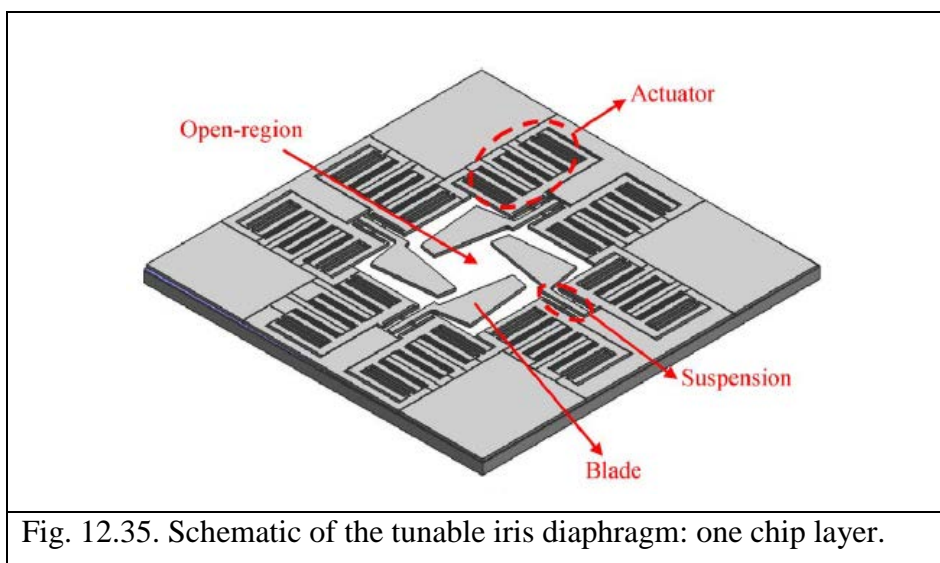
An octagonal aperture can be obtained by overlapping these two chips with a small separation gap along the optical path. Therefore, the friction issue between the moving blades is completely avoided, and the undesired diffraction effect caused by the gap between blades, the dimension of which is mainly limited by the etching process, can also be eliminated. During operation, all the blades are simultaneously actuated to rotate in the clockwise or counterclockwise direction using the electrostatic comb drive actuator combined with a specially designed suspension structure.

Fig. 12.36. Schematic of the proposed tunable iris diaphragm (a) one chip, micro-assembly procedure
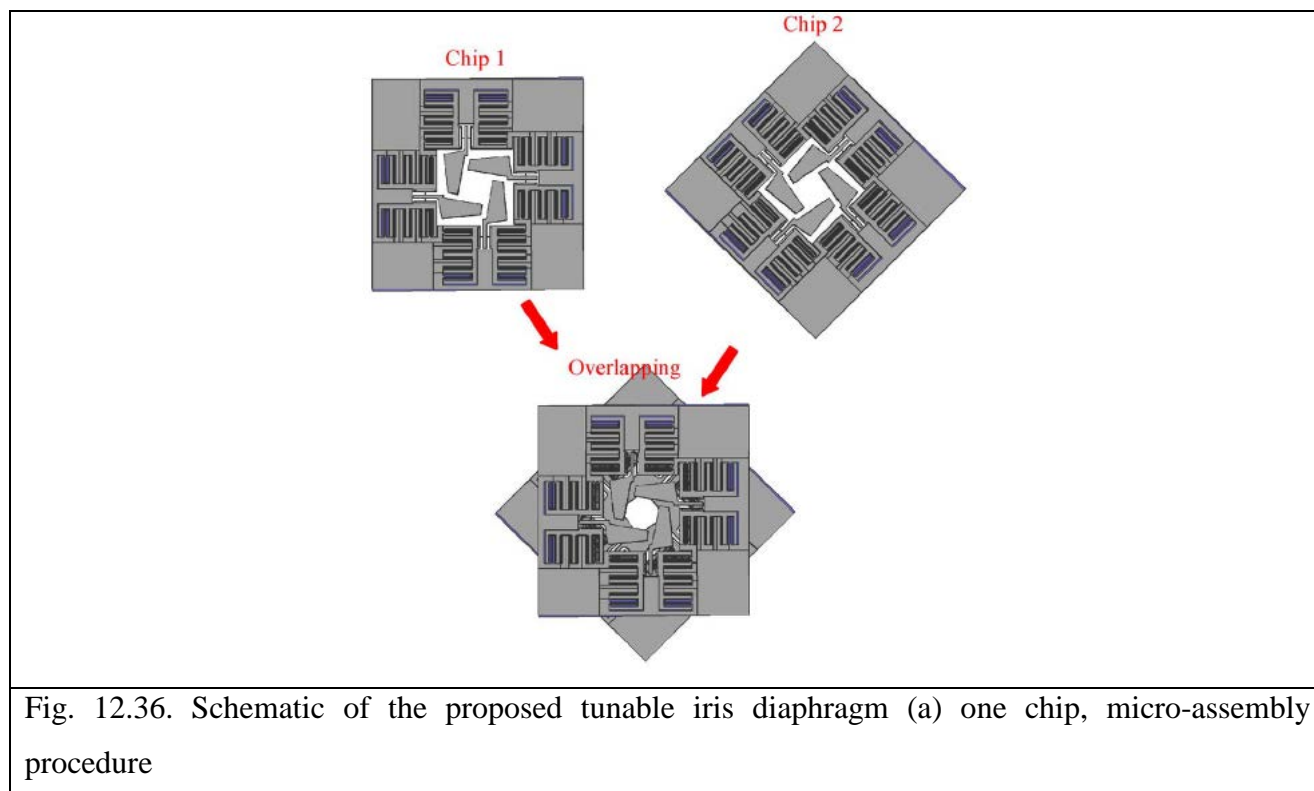
As a result, the aperture size can be selected to open or close, respectively. The proposed iris demonstrates clear aperture, thus avoiding the relatively low light transmission (the maximum transmission is always within 80%) associated with most optofluidic configurations, which is mainly caused by the light reflection at multiple interfaces between different mediums, such as air-glass/polymer and liquid-glass/polymer, located along the light transmission path. At the same time, relatively large tuning range of aperture size (from 0.45 mm to 1.56 mm) has also been successfully achieved with current design (Fig. 12.37).
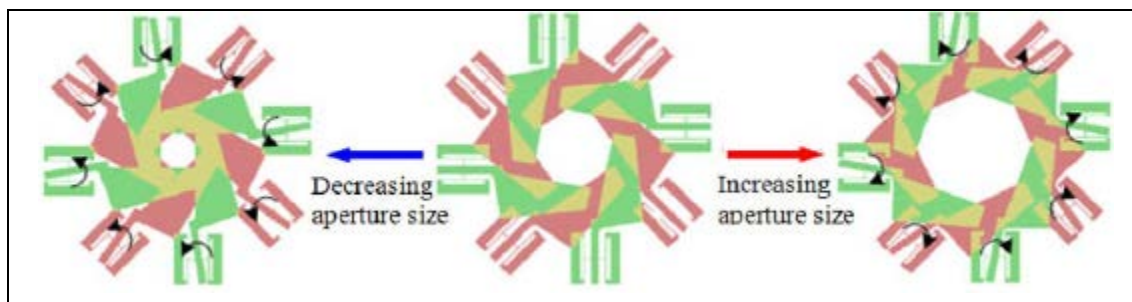
Fig. 12.37. Schematic of the proposed tunable iris diaphragm iris tuning procedure

MOEMS diffraction gratings

The MOEMS devices that work on the principle of diffraction are classified as diffractive optical MEMS. These devices are wavelength-sensitive and differentiate between the spectral components of light incident on them. Such reconfigurable grating structures produce diffraction patterns comparable to the fixed gratings of the past, but with the added flexibility of tuning the output characteristics (angle, intensity of diffraction) during operation. The MEMS tunable diffraction gratings can be used in the development of hand-held digital spectrometers, optical switches, and in display devices. Actuation of grating beams can have two types of motion:

- out-of-plane actuation;
- in-plane actuation.
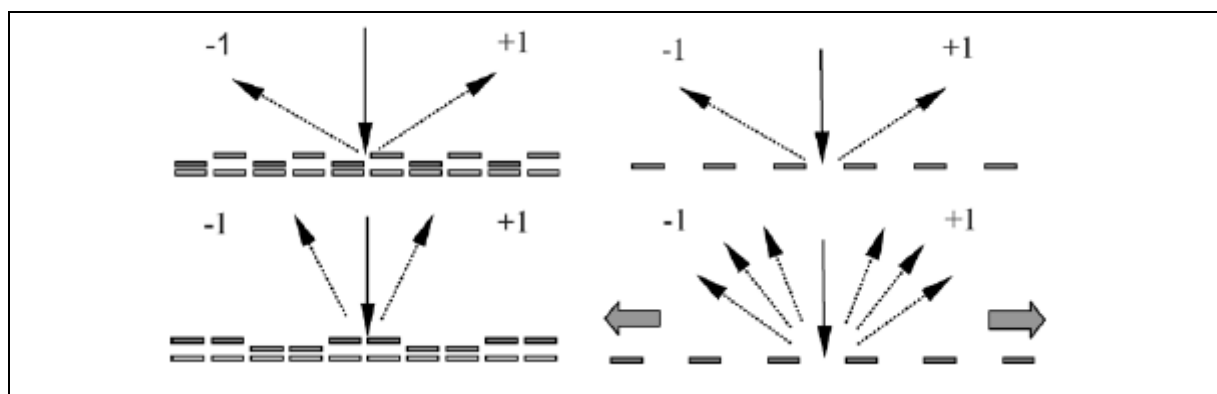
Fig. 12.38. Actuation concept of analog (in-plane actuation) and digital (out-of-plane actuation) tunable gratings. In-plane actuation permits analog control over the diffracted angles.

Analog tuning of the period involves in-plane actuation of the grating beams and digital tuning of the parameters uses out-of-plane actuation of the beams. The analog tuning mechanism has very fine angular resolution, but the deflection range is limited. Different types of MEMS grating actuation were demonstrated:

- MEMS analog tunable gratings with electrostatic actuation.
- MEMS tunable gratings with piezoelectric actuation for tuning the period of the grating.
- MEMS tunable gratings with thermal actuation to obtain a large displacement in the lateral direction.
- MEMS deformable gratings with wide tunability using comb-drives.

One well-known digitally actuated MOEMS device is the grating light valve (GLV). The GLV makes use of an array of electrostatically tunable beams Fig. 12.39a. In its initial state, when all beams are at the same height, a light signal with wavelength, λ, incident on the array is simply reflected in the direction of the zeroth order Fig. 12.39b. By applying a voltage to the device, every alternate beam is vertically deflected by λ/4, causing the majority of the reflected signal to shift from the zeroth order to the first order as depicted in Fig. 12.39c. With this device, a phase shift between the upper grating and lower reflector is always present, therefore, some interference will always be present. A broadband light signal will never be completely reflected into the zeroth order unless the gap between the beams is very small in the initial unactuated state.
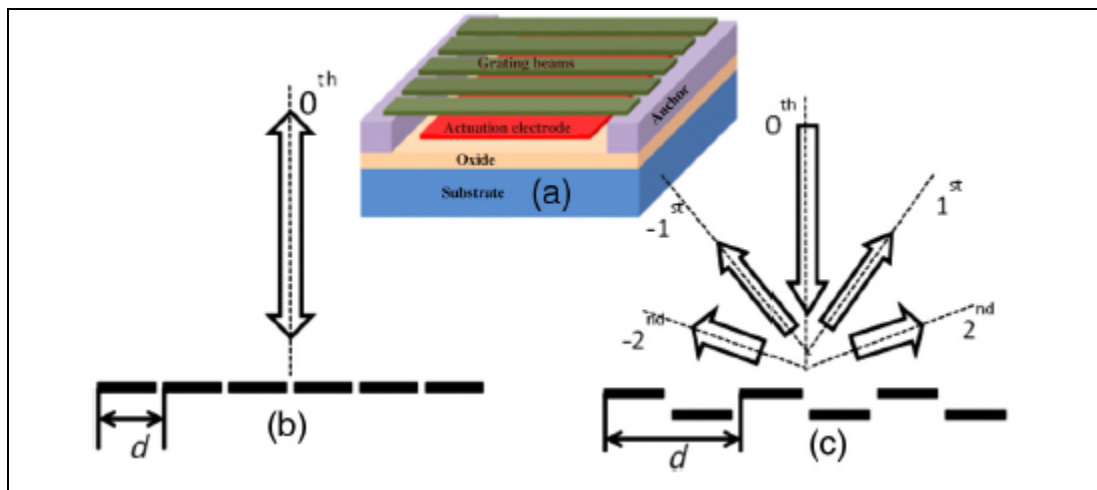
Fig. 12.39. (a) Schematic of grating light valve developed by silicon light machines (b); "ON" position where all mirrors are in a single plane reflecting the light signal into the zeroth order; (c) "OFF" position where 50% of the mirrors are deflected out-of-plane by λ/4 causing most of the light directed into first and other higher orders.

From grating theory, the diffraction efficiency in the first order is maximum when the duty cycle of the grating is 50%. Therefore, the grating to be fabricated should have a duty cycle as close

to this value as possible in all the tuned periods. This will ensure high diffraction efficiency for all the tuned periods. Fig. 12.40 shows the cross-section and top views of various states of a grating tuned by this method. In Fig. 12.40a, the structure is unactuated and acts as a mirror since we assume that the dimension "a" is small compared to "b." In Fig. 12. 40b, every alternate beam has been actuated so that the period is now equal to two beam widths. In Fig. 12. 40c every alternate pair of beams is actuated so that the duty cycle is almost equal to the previous case even though the period has been doubled. In digital tuning, a large deflection range can be achieved at the cost of angular resolution.

Ideally, the gap between the beams should be as small as possible. In the ideal case, when the grating is in its unactuated state it should act like a mirror. However, due to the 2-μm gap, it actually behaves like a grating.
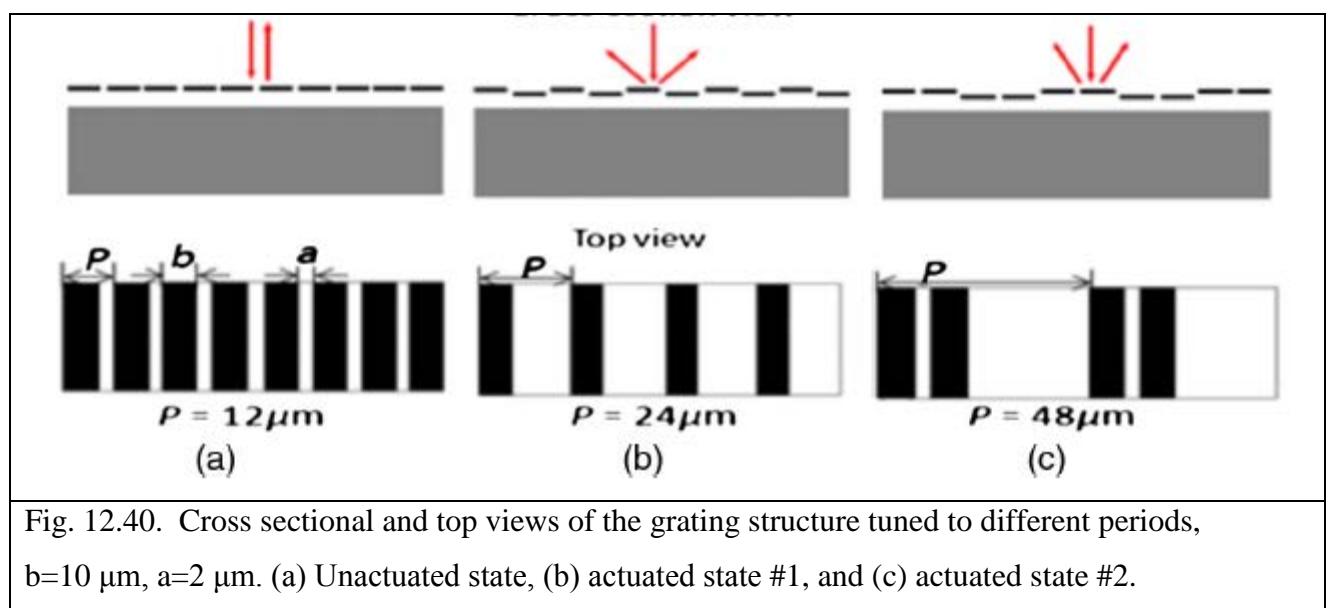
Fig. 12.40. Cross sectional and top views of the grating structure tuned to different periods, b=10 μm, a=2 μm. (a) Unactuated state, (b) actuated state #1, and (c) actuated state #2.

Microscope images of the fabricated devices are shown in Fig. 12.40. The grating structure is shown in Fig. 12.40a. The magnified image (16;800×) of a single beam is shown in Fig. 12.40b. The edges are not straight but possess a wavy nature due to the wet etching process.
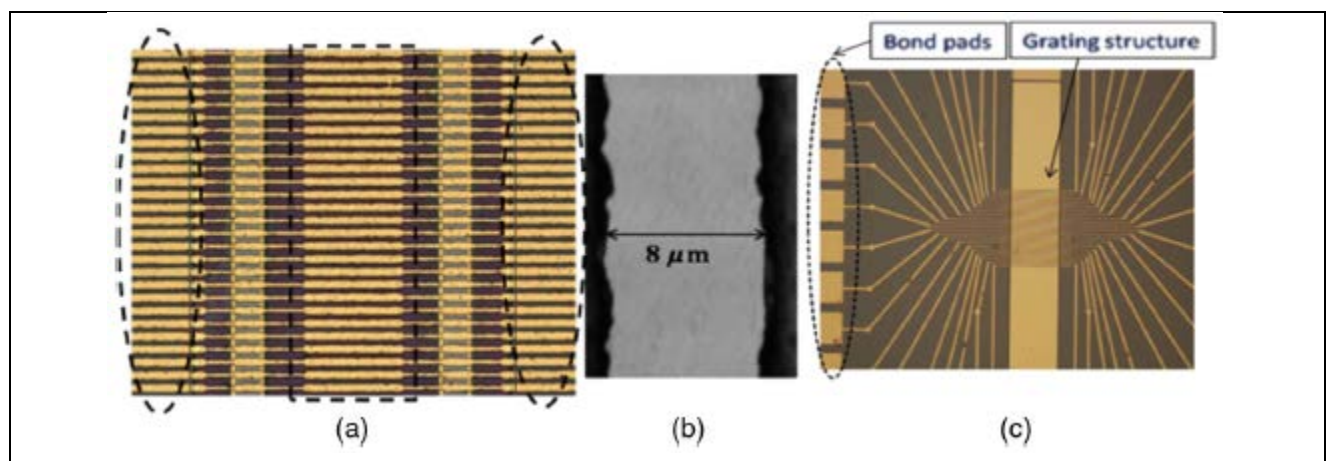
Fig. 12.40. a) fabricated grating structure; b) 16800× magnified image of the beam of grating structure; c) grating with bond pads to have individual access to beams.

The grating structure with bond pads to provide individual access to beams is shown in Fig. 12.40c. After fabrication, the roughness of the gold measured over a region of 100×100 µm2 was 9 nm.

Piezoelectric tunable gratings. Grating structure with integrated thin-film piezoelectric actuators on a deformable membrane can be produced to achieve tunability in the diffracted angle up to 400 µrads, with a resolution of 2 µrads,. Compared with the digital designs, this concept trades a large tuning range for fine angular resolution. Moreover, the use of integrated piezoelectric actuators allows fast dynamic responses in the order of MHz, low-power consumption, low driving voltage (compared to thermal actuation), and better localization of tunability on individual sets of gratings, in contrast with the analog methodologies mentioned above. In this design, the diffraction grating period is tuned by transverse (in-plane and perpendicular to the grating grooves axis) actuation forces applied on the grating structure. This is illustrated in Fig. 12.41., where the grating period is increased continuously, resulting in steering of the diffracted orders. To achieve transverse actuation, the grating grooves was defined on a supporting deformable membrane. The membrane was then mechanically stretched by thin-film piezoelectric actuators, as shown in Fig. 12.41.
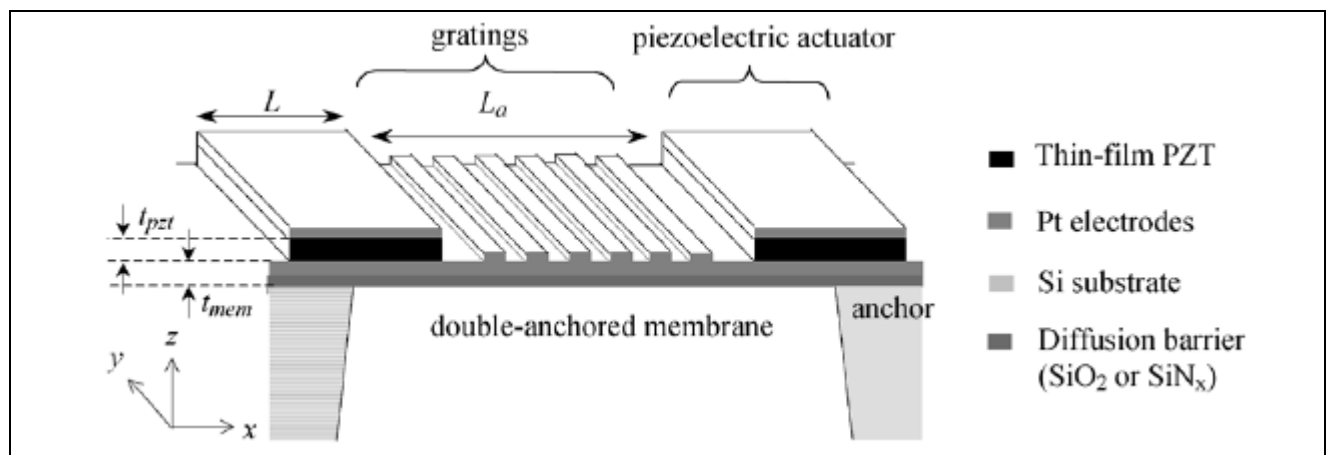
Fig. 12.41. Design schematic of double-anchored deformable membrane, driven via thin-film piezoelectric actuators. The gratings, defined on top of the membrane, are tuned progressively along with the membrane.

The piezoelectric actuators, a lead zirconate titanate (PZT) material located at both ends of the membrane, are capable of producing sufficient force to strain the membrane up to 0.3%, or equivalently, 0.3% diffracted angle change. This is accomplished through application of an electric field across the PZT film to induce strain, due to reorientation of electric polarizations. This correspondingly strains the membrane in the transverse direction and, subsequently, the 4 period binary phase grating imprinted on the membrane. The double-anchored membrane configuration, illustrated in Fig. 12.41 minimizes out-of-plane displacements of the grating grooves from unbalanced film residual stresses and bending moments of the PZT under actuation.
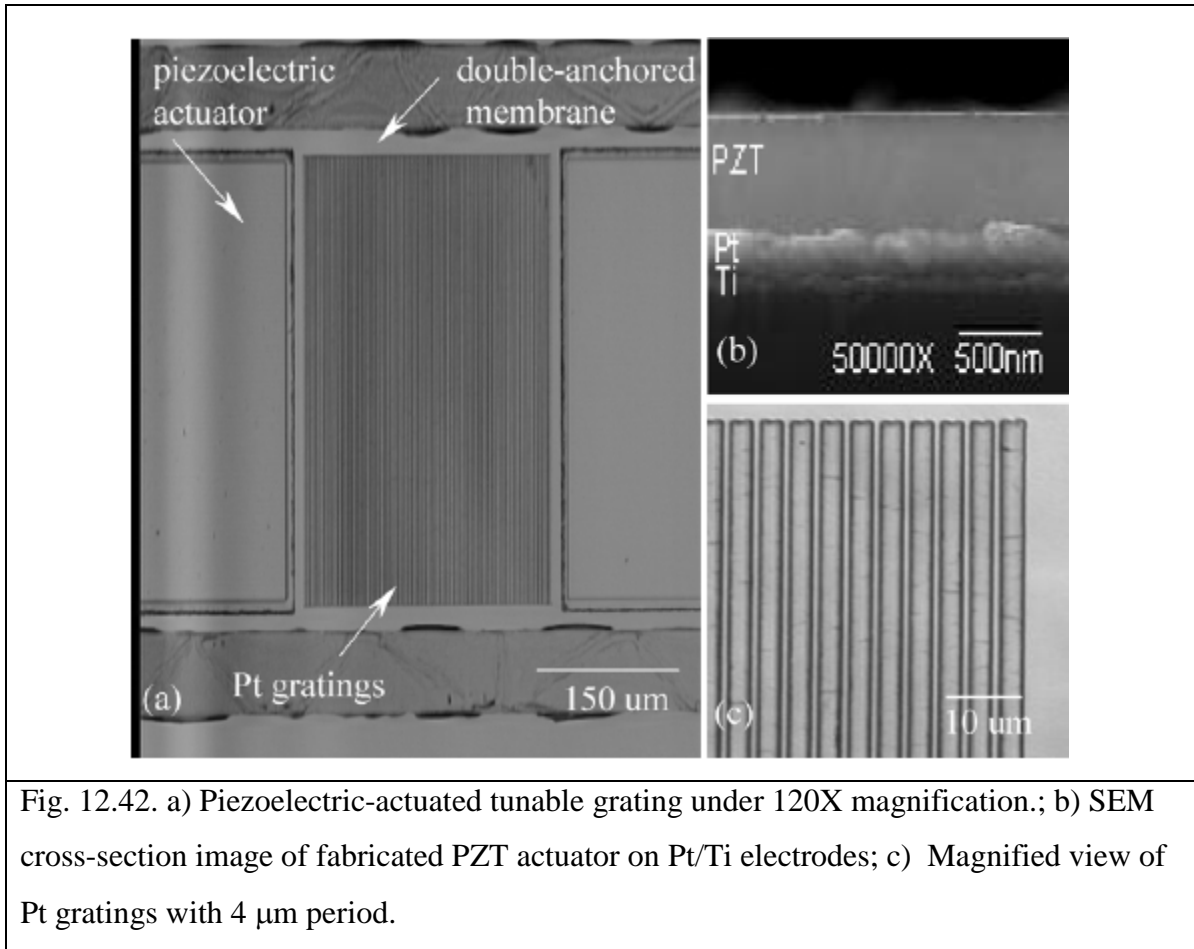
Fig. 12.42. a) Piezoelectric-actuated tunable grating under 120X magnification.; b) SEM cross-section image of fabricated PZT actuator on Pt/Ti electrodes; c) Magnified view of Pt gratings with 4 μm period.

MEMS grating based on thermal actuator to vary the spacing of the grating beams is shown on Fig. Each silicon beam within the grating was connected to a diamond shaped structure made from other silicon beams, as seen in Figure 2.7(a) and 2.7(b). With a current applied through the diamond shape, the beams would thermally expand and alter the spacing between grating 19 beams. Electron beam lithography was used during fabrication to decrease the grating period to less than 1 μm. A grating period shift from 700 nm to 1140 nm was demonstrated. One problem with thermal actuators, however, is slow response time, especially during cooling to return to its initial state. The small silicon beams are also susceptible to burning up if too much actuation current is applied.

12.1.1.4 Tunable Systems.

A further advantage of MEMS-based tunable optics, beyond compact tuning mechanisms, which become viable on the microscale, is that microsystems fabrication processes may also be applied to construct tunable systems of considerable mechanical and optical complexity. Such systems may consist of large-scale two-dimensional arrays of tunable lenses, each independently actuated, where microfabrication techniques assure good uniformity, high fill factors and compact dimensions not achievable with classical optics manufacturing approaches.
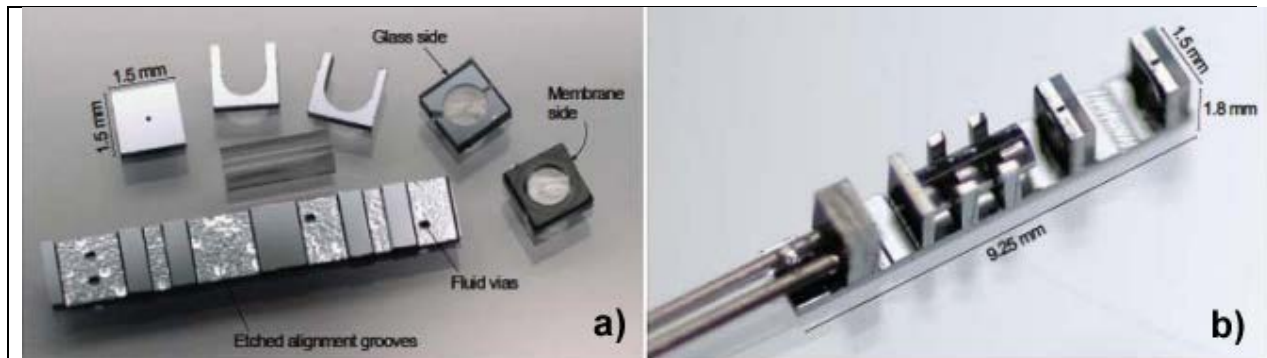
Fig. 12.43. An MEMS-based optofluidic micro-bench assembled from individually microfabricated components (a) into a system with two independently tunable pneumatic microlenses on a silicon optical bench (b).

On the other hand, MEMS-based processes, particularly those underlying the concept of the silicon optical bench, allow the assembly of complex systems including tunable lenses, mirrors and shutter. One such example is shown on Fig. 12.43: the system includes two independently pneumatically tunable lenses mounted on a silicon substrate machined to allow precision optical alignment and containing the necessary microfluidic channels for actuation as well as fixed-focus GRIN lenses. More-advanced systems include scanning micromirrors, rotational micromotors, and tunable fluidic irises, all to allow three-dimensional imaging with a system whose size scale permits integration into a rigid 3mm endoscope.

12.2.2 Biology and biomedicine

Optical MEMS are widely for biomedical applications. For example, MOEMS based miniature scanning confocal microscope that utilizes an off-axis grating as the focusing objective lens and micromachined torsional scanning mirrors to realize real-time confocal imaging with a working distance of 1 mm is shown on Fig. 12.44.
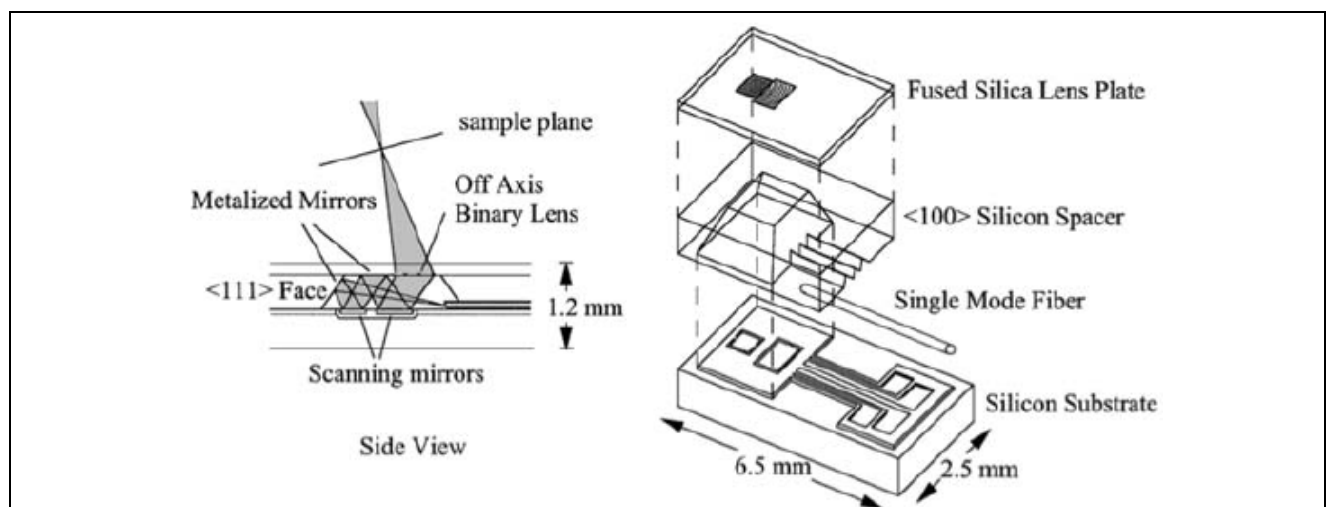
Fig. 12.44. Micromachined confocal scanning microscope, consisting of a single-mode optical fiber for illumination and detection, two tortional mirrors for scanning, and a binary transmission grating

MOEMS are used to create three-dimensional raster scanning module that fit into 1 mm^3 as a key part of a microconfocal imaging array for biological lab-on-a-chip applications. In this module, low voltage and large static displacement vertical actuators are integrated with microlenses on high-aspect-ratio silicon-on-insulator (SOI) structures. The isolated vertical comb drives and the coupled-torsion flexures provide upward and downward piston motion with low driving voltages (Fig. 12.45).
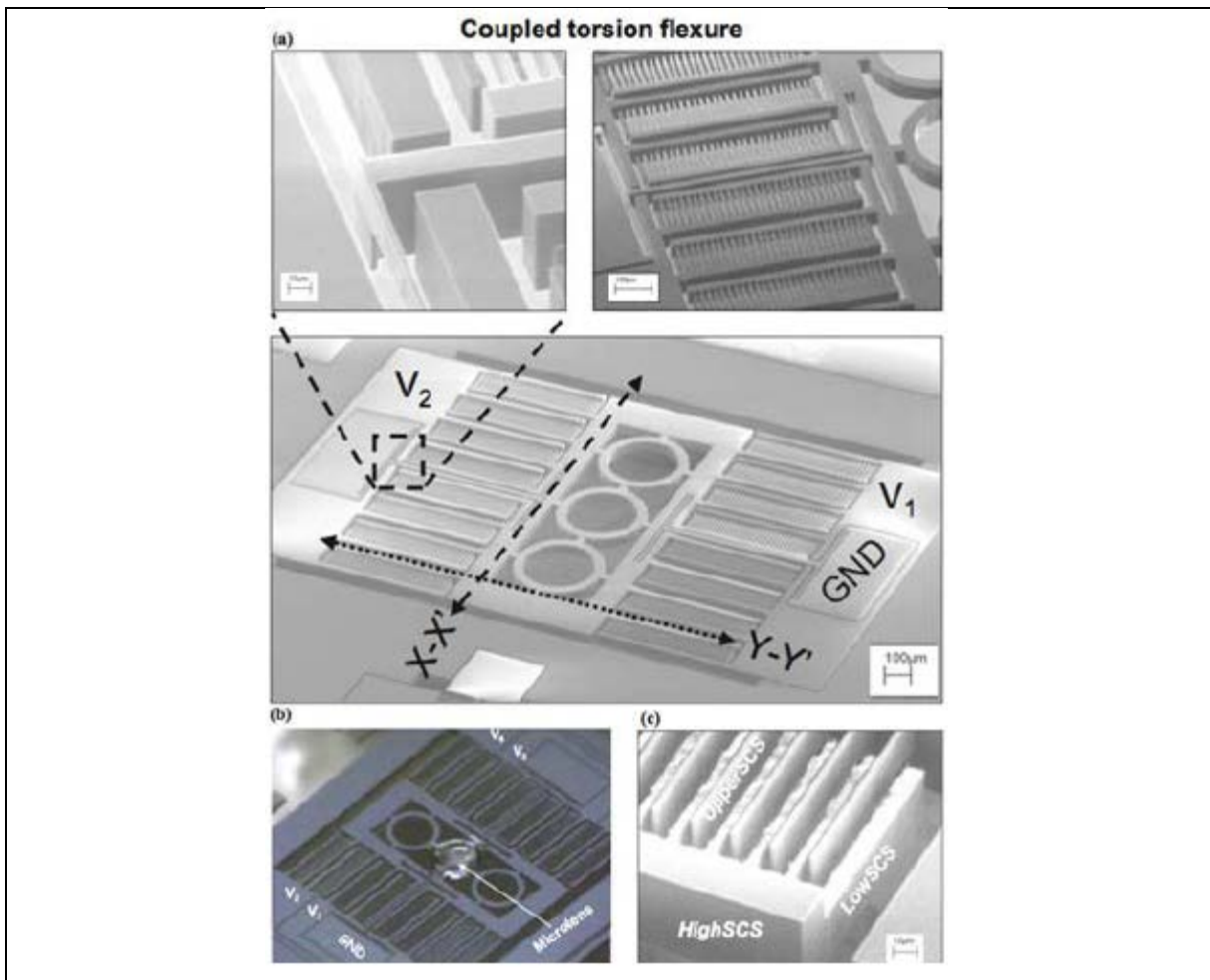
Fig. 12.45. SEM and picture of fabricated vertical microlens scanner: (a) SEM of unidirectional device, (b) Microscopic picture of bi-directional device with polymer lens, (c) Self aligned vertical comb fingers.

Vertical comb-drive-based two-dimensional gimbaled MEMS scanners with large static rotation were demonstrated (Fig. 12.46). Two axis optical MEMS mirrors are useful in biomedical imaging, raster scanning and image projection. This method of backside island isolation provides electrical isolation as well as mechanical coupling of SOI structures without additional dielectric backfill and planarization. It allows a gimbal structure with electrical isolation, enabling two-axis rotation of MEMS scanners for biomedical imaging applications.
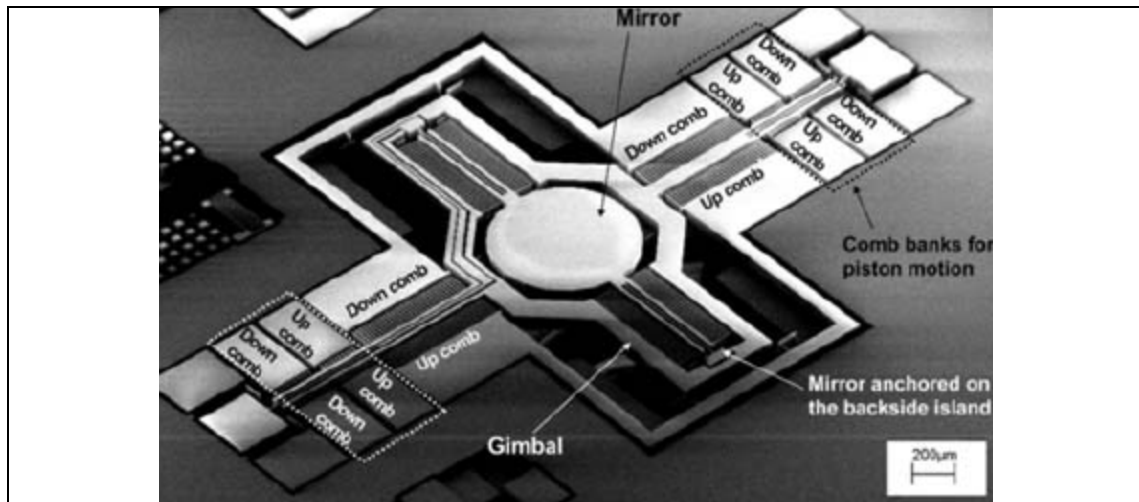
Fig. 12.46. SEM of 2D MEMS scanner for biomedical imaging applications

The first miniature scanning confocal microscope with electrostatically actuated microlenses for focusing and scanning is demonstrated on 12.47. In this microscope, the objective lenses, scanners, pupil, and pinhole of the confocal microscope are integrated in a volume smaller than 2 mm^3. The objective lenses are composed of two vertically cascaded polymer microlenses integrated onto micromachined comb actuators. Raster scanning is implemented by electrostatically actuating each microlens in orthogonal directions. The system achieved reflection confocal imaging with 3 μm transverse resolution over a 100 single molecule μm field of view and a 0.38 mm working distance at λ = 633nm. This device demonstrates the combination of silicon micromachining, polymer micromachining, and vertical stacking of different substrates
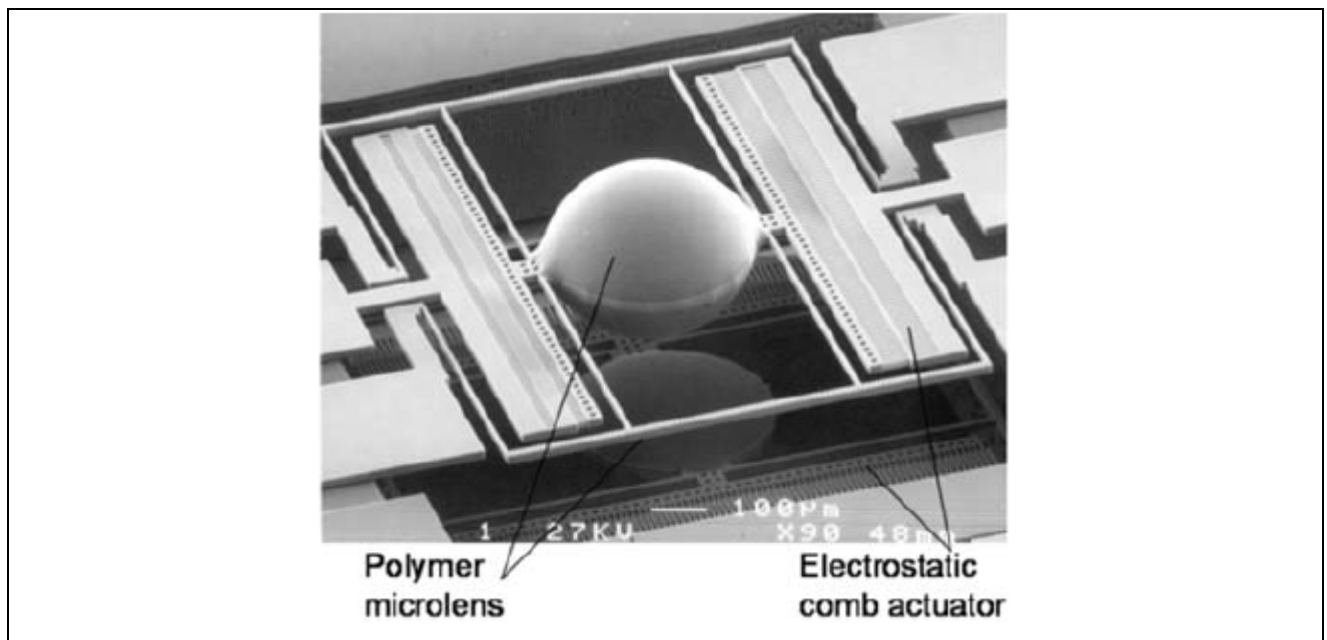
Fig. 12.47. Scanning electron micrograph of MEMS microlens scanners. Two microlens scanners are orthogonally aligned and vertically stacked together.

A two-dimensional MEMS scanner for dual-axis confocal microscopy is shown on (Fig. 12.48). Developers fabricated the scanner by using a double SOI wafer that enabled actuation by

self-aligned vertical electrostatic comb drives. Maximum optical deflections of ±4.8° and ±5.5° are achieved in static mode for the outer and inner axes, respectively. Reflectance images were achieved with a field of view of 344 μm × 417 at 8 frames/s. The transverse resolutions are 3.94 and 6.68 μm for the horizontal and vertical dimensions, respectively. As a result, device demonstrates real-time high-resolution *in-vivo* imaging with two-dimensional MEMS scanner.
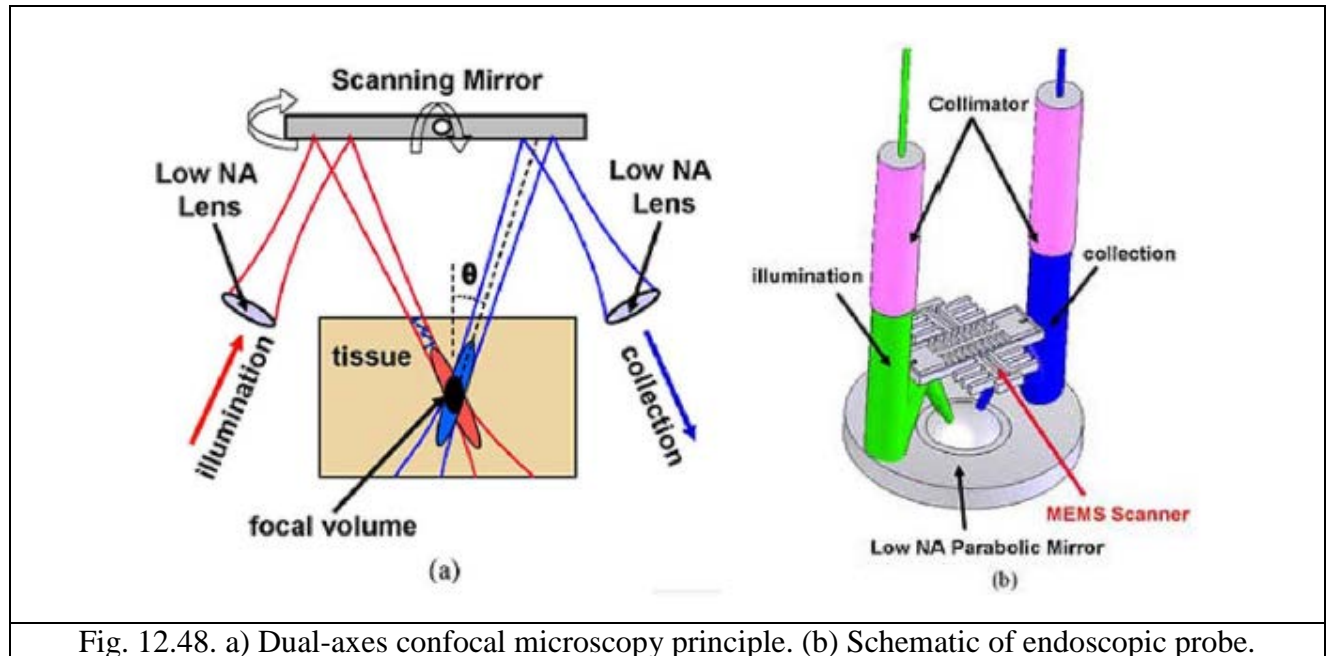
Fig. 12.48. a) Dual-axes confocal microscopy principle. (b) Schematic of endoscopic probe.

Handheld dual-axis confocal MOEMS-based microscope is depicted on Fig.. The fully packaged microscope has a diameter of 10 mm and acquires images at 4 Hz frame rate with a maximum field of view of 400 μm × 260 μm. The transverse and axial resolutions of the handheld probe are 1.7 μm and 5.8 μm, respectively. The capability to perform real time small animal imaging is demonstrated *in vivo* in transgenic mice. These optical MEMS components enable biomedical applications such as optogenetics and noninvasive endoscopic screening for disease, etc.
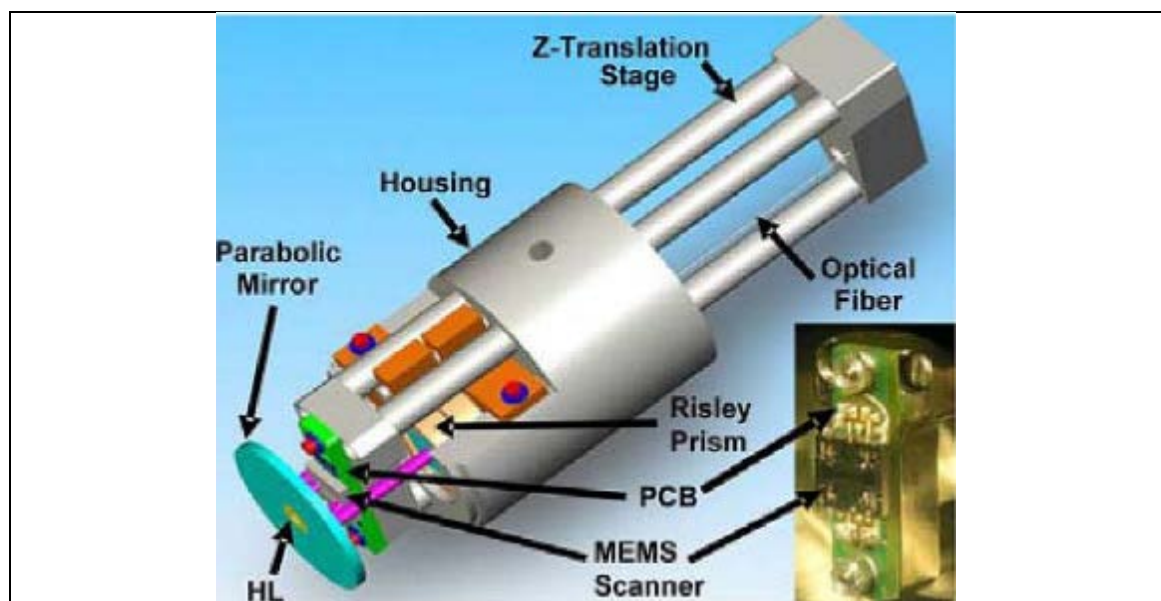
Fig.12.49. Schematic of a 10 mm diameter handheld probe with the packaged MEMS

scanner in the inset.

A fast-scanning two-photon fluorescence imaging based on optical MEMS, which has the acquisition rates of lines up to ~3.5 kHz is shown on Fig. 12.50.
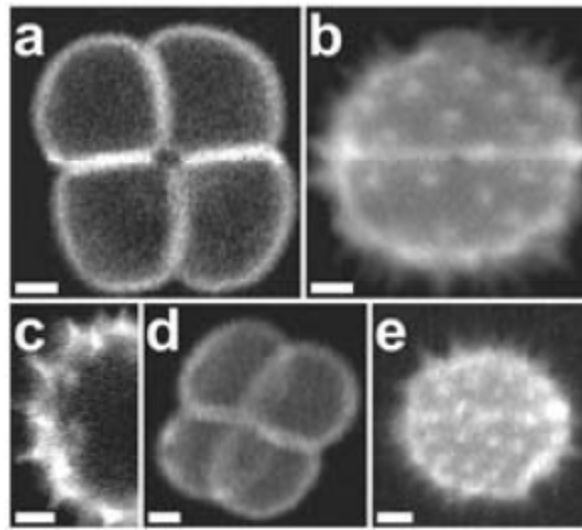
Fig. 12.50. Two-photon fluorescence images of pollen grains acquired using instrumentation based on a MEMS scanner.

A miniature microendoscope, which consists of a combination of the MEMS scanning mirror, microlenses and fiber-optic technology was developed (Fig. 12.51). The microscope is sufficiently small that an adult mouse can carry the 2.9-gram device on its head while allowing the microscope to capture deep brain images of awake and moving animals.
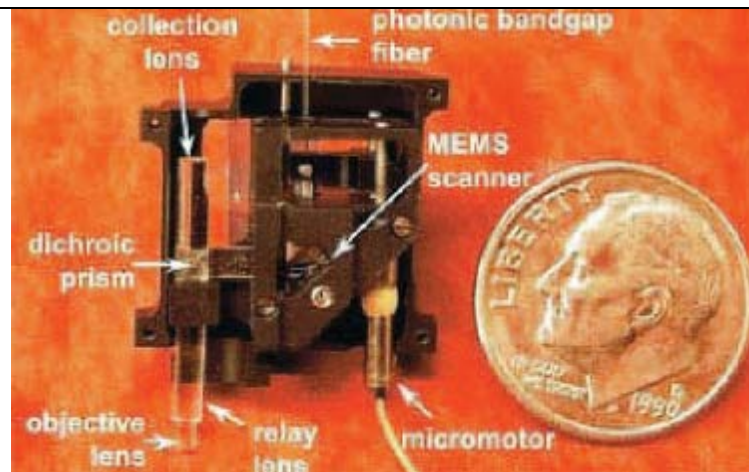
Fig. 12.51. Miniature microendoscope, which consist of a combination of a MEMS scanning mirror, microlenses, and fiber-optic technology.

MOEMS single molecule detection.

The ultimate goal of most bioMEMS systems is to monitor biological processes with sensitivity, selectivity and good temporal resolution. Current advanced Optical MEMS/NEMS enables the miniaturized tools that are used to investigate individual bio molecules to achieve high precision

and throughput. One of the first attempts to investigate single-molecular-level detection was made by Seo *et al* [83]. As in Fig. 21, they utilized integrated SixNy microfluidic devices with microelectrodes to achieve a bondless nano- to microfluidic interface, low transmission loss and good stability under challenging external environmental parameters such as rapidly changing temperature and the presence of corrosive gasses. They demonstrated detection of a single 500 nm size florescent bead and fluorescently-labeled DNA. Autonomous biophotonic MEMS without a conventional microscope was achieved by designing single-layer-based two-dimensional optical scanners. Following this development many similar MEMS designs were realized for the detection and manipulation of single biological entities [84]. Significant progress in single-molecule detection with Optical MEMS has been achieved in the genomic sequencing field. Pacific Biosciences has developed a next-generation DNA sequencer based on a zero-mode wave-guide system, which provides extreme optical confinement and enables singlenucleotide sensitivity while DNA polymerase constructs the sequence as shown in Fig. 22. The principle of this NanoElectroMechanical System (NEMS) is straightforward and simple. All waveguides have a specific cut-off wavelength beyond which propagation cannot be sustained. At wavelengths longer than the threshold, the field decays exponentially along the propagation direction resulting in an evanescent wave [85]– [87]. The practical decay length comprises zeptoliter observation volumes within the guide for ultra-sensitive biomolecular detection.

Fig. 21. Biophotonic MEMS for single molecule detection and manipulation. Left: (a) cross sectional diagram of silicon nitride microfluidic system with integrated micropores; (b) 500 nm fluorescent beads trapped outside of microfluidic channels; (c) 500 nm fluorescent beads transported from outside the microfluidic channels to the inside by optical trapping; Right: Detection of single carboxy-rhodamine 6G molecules excited with a 488 nm Ar laser and detected

by an APD.

Fdfd

Fig. 22. MEMS implementation of a four-color real-time sequencing platform is shown. Zero-mode waveguides provide ultra-compact detection volume allowing high signal-to-noise detection.

The optofluidics field has combined flexible microfluidic systems with tunable optical MEMS/NEMS and advanced optics to create lab-on-a-chip instruments, biomolecular sensors and molecular-imaging tools [88]. Integration of photonic circuits plays a key role for devices that are more functional and compact [89]. In addition to optical sensing, active biological manipulation has been achieved utilizing both optical and non-optical forces, including hydrodynamics and dielectrophoresis (DEP). This dual-purpose use of optics (actuation and detection), makes optofluidics a very attractive approach to biomedical instrumentation that requires biomolecular manipulation and observation. Liu *et al.* demonstrated the use of optofluidics to control photothermal nanoparticles [90]. Their work was based on direct optical-to-hydrodynamic energy conversion using suspended photothermal nanoparticles near the liquid–air interface utilizing submilliwatt power. They were able to drive and guide liquid flow in microfluidic channels to transport biomolecules and living cells at controlled speeds and directions (Fig. 23). This work laid the foundation for optically driven, large-scale optofluidic circuits for biomolecular sensing and manipulation. As shown in Fig. 24, Yang *et al.* demonstrated active DNA manipulation (λ-DNA) by creating stronger optical confinement and higher intensities than those of diffraction-limited systems [91]. Their technique simultaneously utilizes nearfield optical forces to restrain matter inside the waveguide and scattering/adsorption forces for transport. Lien *et al.* demonstrated optofluidic integration with eight solid-core waveguides to observe fluorescent particles [92]. A similar concept with metal-clad, dynamicallyreconfigurable liquid lenses in a microfluidic channel was demonstrated by Tang *et al.* [93]. Figure 25 shows a schematic diagram of the device and how it focuses light from an optical fiber through the liquid lens. The lens is formed by three laminar liquid flows that interact to create the desired refractive-index profile in the region where the lens forms. In addition to biomolecules such as DNA, RNA, and proteins, a larger biological entity – Caenorhabditis elegans (C. elegans) – has been investigated using microfluidics with embedded low-cost and high-resolution microscopy [94]. Figure 26 shows the implementation and application of a high-resolution, lensless and fully on-chip microscope system. Utilizing either gravity-driven

flow or electrokinetic flow for positioning yields a compact, low-cost imaging solution. A variety of optical interactions enable various biological/ medical applications such as Surface Enhanced Raman Spectroscopy (SERS) as in the microfluidic platform in Fig. 27 [95]–[97]. A very promising approach is to use optofluidic resonators for SERS signal amplification within micro- or nanofluidics. Various biological molecules, including Dengue virus and A β protein, have been targeted for maximum signal amplification by SERS. Taking advantage of the amplification ability of SERS and implementing a miniaturized immunoassay, Lee *et al.* [98], presented a programmable and fully automatic gold array-embedded gradient microfluidic device for providing a convenient and reproducible surface-enhanced Raman scattering (SERS)-based immunoassay platform for cancer biomarkers as shown in Fig. 28. They achieved automatic serial dilution by the gradient microfluidic generator and a total analysis time less than 60 minutes. Micro-optofluidics also enables particle synthesis. Sunghoon Kwon and his group at the Seoul National University developed a technique that can synthesize freefloating microstructures in microfluidic channels at the desired time and location within the field of view of the lithography system [99], [100]. Figure 29 shows a detailed procedure of Kwon's maskless particle-generation system. They demonstrated a real-time in-situ polymerization process to dynamically synthesize extruded polymeric microstructures with various two-dimensional shapes. They used highspeed spatial light modulators to dynamically control shapes of polymerized microparticles and to achieve high throughput. The Institute for Optofluidics and Integrated Nanophotonics (IONAS) has developed reconfigurable optofluidic nanophotonic circuits, which are promising in applications such as optical information processing and highly sensitive and localized optical sensing [101]–[103]. They have investigated optofluidic PhC microcavities (Fig. 30) and their properties in the context of (bio) sensing. High quality factors of the order of 6 × 104 have been demonstrated.

Optical MEMS and Microfluidics: Optofluidics

MOEMS sensors.

12.2.3 Nano opto-electromechanical systems

12.2.4

Reference list

- Optical MEMS: From Micromirrors to Complex Systems
- AdAptive imaging mems deformable mirrors thomas Bifano
- A 2-D MEMS scanning mirror based on dynamic mixed mode excitation of a piezoelectric PZT thin film S-shaped actuator
- Integrated Optofluidic Iris Philipp Müller, Roland Feuerstein, and Hans Zappe

- Elastic Tunable Grating Powered by MEMS Actuators for Fluorescence Spectroscopy Applications

7.   Optical switching technology

7.1 Introduction

      7.1.1   Electro-optical switches

      7.1.2   Thermo-optical switches

      7.1.3   Magneto-optical switches

      7.1.4   MEMS-based optical switches

      7.1.5   SOA-based optical switches

      7.1.6   Switching based on optical nonlinear effects

      7.1.7   Liquid crystal optical switches

      7.1.8   Photonic crystal all-optical switches

      7.1.9   Fiber, holographic, quantum optical swithes

7.2 Optical switch architectures

      7.2.1   Two-dimensional optical switches

          7.2.1.1 $N \times N$ crossbar switch

          7.2.1.2 Polygon switch

          7.2.1.3 Re-arrangeable non-blocking switch

          7.2.1.4 L-switching matrix

          7.2.1.5 Multistage 2D switch

      7.2.2   Three-dimensional optical switches

          7.2.2.1 Lucent Microstar

          7.2.2.2 Microlens scanner

7.3 Actuating principles of MEMS-based optical switches

      7.3.1   Electrostatic

      7.3.2   Electromagnetic

      7.3.3   Thermal

7.4 Application

      7.4.1   Optical computers

      7.4.2   Optical Signal Switching and Routing

      7.4.3   Optical Network Protection and Restoration

      7.4.4   Optical Network Monitoring

      7.4.5   Video Distribution

      7.4.6   Instrumentation Resource Sharing

      7.4.7   Military Communications

7.1 Introduction

In order to appreciate the benefits of using MEMS to implement optical switches, one must understand the wavelength division multiplexing (WDM) system. At the transmitter side, optical signals of different frequencies first need to undergo gain adjustment by variable optical attenuators (VOA) so that signals from various sources will have comparable power. An optical multiplexer (MUX) is then used to combine the signals into a single optical fiber. Transporting optical signals of different wavelengths increases the capacity of the system and reduces implementation cost per unit bandwidth. The reverse process occurs at the receiver. An optical de-multiplexer (DEMUX) is utilized and signals of different wavelengths are separated into different optical fibers and routed to their destinations

7.1.1 Electro-optical swithes

1×1 optical switch is usually a 1×1 optical modulator. It can be fabricated in III–Vs materials, Si(Ge) materials, LiNbO3, or polymers. A single-mode rib waveguide with a width of W is composed of an n-type silicon guided-wave layer on a SiO2 layer. An abrupt p+-n junction is formed below the top surface of the ridge waveguide to inject the carriers into the waveguide. If the p+-n junction of the waveguide is forward-biased when guided-mode optical signals are end-coupled into the rib waveguide, a large number of carriers will be injected into the guided-wave layer of the waveguide, and the refractive index in the waveguide will decrease because of the plasma dispersion effect, which can make the guided-mode convert into the radiation mode of the substrate and the cover. This causes a lot of guided-mode energy to be lost and absorbed in the rib waveguide, which will cause the rib waveguide to cut off, resulting in the so-called waveguide 'vanishing.'

Consequently, there is no output light in the waveguide, and thus switching is achieved. For the fabrication, SOI material is used. This is produced by cleaning followed by oxidizing the substrate wafer (the SiO2 is 400–500 nm thick), bonding at a high temperature (in O2, at 1200°C, for 2 h), and thinning (grinding precisely rear face down to 20 µm), and then, polishing to a thickness of 6 µm with r.m.s. roughness <±0.5 µm. The second stage is the preparation of the SOI ridge waveguide, which consists of lithography, oxidizing, photo-etching, and potassium hydroxide liquidphase anisotropic etching at 80°C. Silicon (3 µm) is etched away to form the ridge waveguide, i.e. a ridge height of 3 µm and a width of 6 µm.

The third stage is the fabrication of the optical waveguide intensity modulator, and the processes involved are the making of diffusion masks, photo-etching, diffusing phosphorus to form two n+ collectors either side of the ridge waveguide, oxidizing, photoetching, and diffusing boron to form a p+-n junction on the ridge waveguide. The gap between each n+ region and the ridge waveguide is 12 µm and the area of the p+-n junction is 6×200 µm2. The metal contacts are obtained by sputtering a Ti/Al electrode. To improve the response speed of the switch, electron

irradiation is performed at an electron energy of 14 MeV. After that, the wafer is diced to an overall length of 6 mm, but the real operation length of the switch is <0.5 mm. The chip is end-polished and wire-bonded for characterization. The switching current of the switch is 45 mA. The modulation depth is 96% at an injection current of 45 mA. The insertion loss is 3.65 dB at a wavelength of 1.3 µm. The switching time is about 160 ns.

O   ptical switches are key components for applications in 1.31–1.55 µm optical communications, networks, and microsystems. They can reduce the cost of the network and increase fiber transmission capacity and at the same time, distribute optical signals to different subscribers. Optical switches in Si-based waveguides make use of changes in the refractive index induced by carrier injection and offer advantages of small device size, polarization independence, and capability of integrating with other Si-based optoelectronic devices. By adding Ge into Si, the bandgap of SiGe shifts towards the optical communication wavelength while the refractive index increases, which is good for wave guiding. The p-n junction can also be formed easily during SiGe epitaxy. In this chapter, six switching approaches at the optical communication wavelength are introduced. Optical switches utilizing carrier-induced refractive index change have tremendous potential for the application of optical processing because of their small size, single-mode operation, and polarization independence. According to the types of carrier injection, optical waveguide switches can be divided into vertical injection and lateral injection. For the former, one electrode is located on the top of the switch and the other at the bottom of the substrate, and the substrate must be n+ or p+ type for good ohmic contact. For either n+ type or p+ type substrate, a large amount of free carriers will be present at the interface between the substrate and the waveguide layer. Hence, carrier absorption loss will be very large at the interface when the switch is in operation. However, the injection current and injection current density cannot be reduced by any other means. With lateral injection, the carrier absorption loss at the interface between the substrate and the waveguide layer can be eliminated due to the avoidance of either n+ or p+ substrate. However, the injection current and injection current density still cannot be reduced.

7.1.2 Thermo-optical swithces

Thermo-optical switches are key elements in the construction of all-optical transmission and switching networks (Yao et al., 2000). They are very attractive due to their small size, large scalability and potentiality for integration with waveguide DWDM (de)multiplexers. They play an important role in optical telecommunication applications, such as optical cross-connection (OXC), protection switching and switch arrays for optical add-drop multiplexing (OADM).

Historically, the most common examples of thermo-optical switches are based on polymers and silica. The polymer devices are fabricated and marketed starting from standard polymer materials or particular patented molecules, while the silica-based devices are derived from the well-

established silica-on-silicon technology for passive waveguide components. In both cases the refractive index should be chosen in order to reduce the coupling losses with the optical fiber.

Channel waveguides in both technologies are typically made by first depositing the bottom cladding layer, followed by the deposition of the core layer. Then, reactive ion etching is used to etch the core ridge while a following coating process realizes the upper cladding layer. Typically, in the commercial thermooptical devices based on polymeric or silica technology, silicon wafers are used as substrates because of their compatibility with standard IC process equipment, good surface quality and excellent heat conducting property. This last characteristic is very important for thermo-optical components because it allows the substrate itself to act as a good heat sink. Switching between output channels can be induced by creating a difference in their propagation characteristics or, in other words, in their effective refractive indices. This can be done by driving the resistive stripe heater electrodes deposited on the top of the cladding layer above the waveguides. Polymer and silica planar lightwave circuit designs having milliwatt thermal powers and millisecond switching times have been realized. Thermo-optical switches in these materials are commercially available.

The thermo-optic effect (TOE) is present in all materials. The thermo-optic coefficient (TOC), given by $d\varepsilon/dT$, is the fundamental optical parameter, describing the temperature dependence of the refractive index of materials, where ε is the complex dielectric function of the material at temperature T. Being the TOC of considerable interest in many optics and optoelectronics applications, such as guiding, coupling and modulation of radiation, it is highly desirable to obtain a theoretical prescription, i.e. its absolute magnitude and its sign, over a wide range of frequencies, based on, at most, a small number of known values at a few frequencies. Therefore, in the next section, the physical principles of the TOE are discussed. We review the most important models presented in the literature. The aim is to provide not all details but their potentiality and limitation, pointing out the main hypothesis and conclusions. We also introduce the basic physics of thermodynamic problem involved in thermo-optical switches. Again, a complete description is outside the scope of our work; therefore, considering a simple case, we introduce the fundamental parameters describing the basic tradeoff of thermo-optical switches. Finally, we conclude the section by pointing out the importance of thermal effects in nonlinear optical devices.

Regarding materials, it is well known that in photonic devices a large number of materials have been investigated. Of course, their properties, i.e. mechanical, thermal and optical, influence the performances of devices. Many times, when we design a device, a major difficulty is caused by the lack of experimental data about the properties of materials. Therefore, in section 3.3, a wide review of the most common materials utilized for the fabrication of thermo-optical switches is reported, providing for each material an updated literature reporting the value of the TOC.

A number of thermo-optical switches have been developed. They include interferometric devices, such as directional coupler and Mach–Zehnder, and digital optical switches based on modal evolution in the conventional Y-junction branch and based on total internal reflection (TIR), micro-electro-mechanical system (MEMS) technologies.

7.1.3 SOA-based optical switches

The application of semiconductor optical amplifiers (SOAs) is not limited only to the amplification of optical signals. It is also used as an important element in optical switching (Tanaka et al., 2009), all-optical signal processing (Teimoori et al., 2008), demultiplexing (Tangdiongga et al., 2007), regeneration (Ezra et al., 2009) and wavelength conversion (Banchi et al., 2010). The well-known benefits of SOAs include fast switching times, high extinction ratio (especially in switching applications), sizable operating gain bandwidth and a relatively compact footprint. SOAs, however, introduce amplified spontaneous emission (ASE) noise added to the input optical signal as an inevitable side effect of the amplification.

Electrical control of the SOA gates is achieved by modulation of the SOA bias current. Therefore, operating the SOA in the linear regime is preferred (no saturation or pattering effect). Likewise, low threshold current and fast switching time are important issues in design. These gates are interesting in optical burst switching or optical packet switching networks; the intermediate nodes in the network are transparent and opaque in terms of data and control, respectively; the packet control data (header information) is converted and processed within the electronic control plane. A vast amount of data, however, remains in the optical domain (optical transmission plane). Thus, an optical space switch can establish transparent connections between multiple inputs and outputs. The bottleneck of this approach is the lack of all-optical memory to hold the data in the optical domain, while waiting for the electronic routing task to complete (Srivastava et al., 2009).

In the all-optical switching approach, the SOA is operated in the nonlinear regime where one or more auxiliary light beams are injected into the SOA. Nonlinear effects in an SOA can be classified into two different phenomena: new frequency component generation inside the SOA medium, i.e. four-wave mixing (FWM), and alteration of a property of the input light, i.e. polarization (crosspolarization modulation), phase (cross-phase modulation) or amplitude (crossgain modulation). An element sensitive to the property altered by the SOA provides the gating action. For example, an arrayed waveguide grating (AWG), polarization beam splitter (PBS) or SOA-assisted interferometric architecture (e.g. Mach–Zehnder) can, respectively, act upon variation in wavelength, polarization and phase in the SOA.

Optical cross-connect

The optical system that is described so far has only fixed connections. In other words, signals from one source can only be hard-wired to a certain destination. An optical cross-connect (OXC) is used to redirect optical beams from one input fiber to multiple output fibers. In essence, OXC serves as a routing matrix that allows any input port to connect to any output port. Originally, the switching system is implemented by converting the incoming optical signal into an electrical signal by a photo-detector before being switched electrically to a defined output, and subsequently the electrical signal is converted back into an optical signal. This optical-electrical-optical (OEO) conversion process is expensive to implement and maintain. An increase in the encoding speed of the optical signal will require the electronic switching system to be upgraded. As a result, OEO-based optical switches could easily become the bottleneck of the network. Moreover, the unnecessary conversions will bum more power and add noise to the signal. MEMS technology, on the other hand, allows the processing of optical signals without the OEO conversion. Micromirrors can be built, as can be seen in the next section, to redirect the optical signals into pre-defined optical fiber outputs.

Wavelength selective cross-connect

Combining the WDM and OXC would result in a useful system called wavelength selective cross-connect (WSXC). As shown in Fig. 5.1, a WSXC consists of N OXCs, where N is the number of wavelengths in each optical fiber. At the input side, the signals are split into different wavelengths using optical DEMUXs, the same process as in the receiver side of a WDM system. Those signals of various wavelengths can then be switched to the desired outputs and combined with optical MUXs just like a WDM transmitter. The WSXC allows signals from multiple sources to be routed to the same destination as long as those signals have distinct wavelengths.

Figure 7.1 Wavelength selective cross-connect

Figure 7.2. Optical add/drop multiplexer (OADM): implementation using
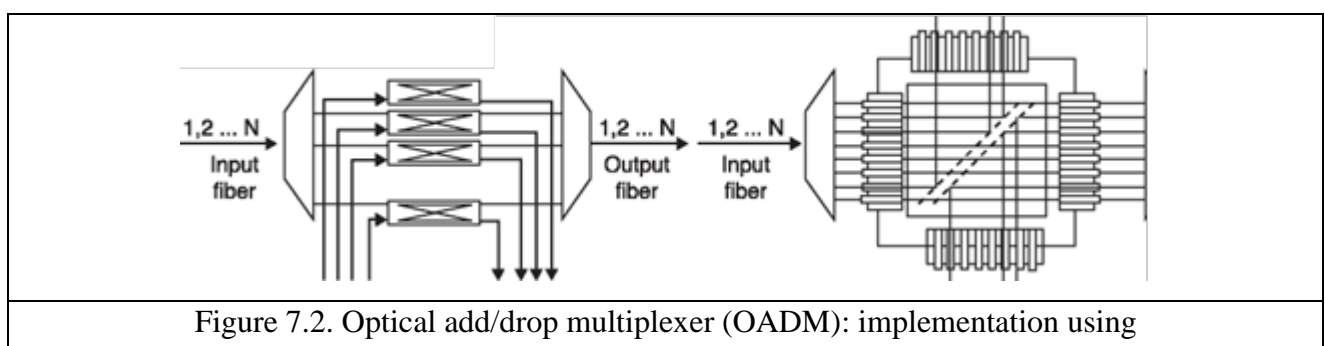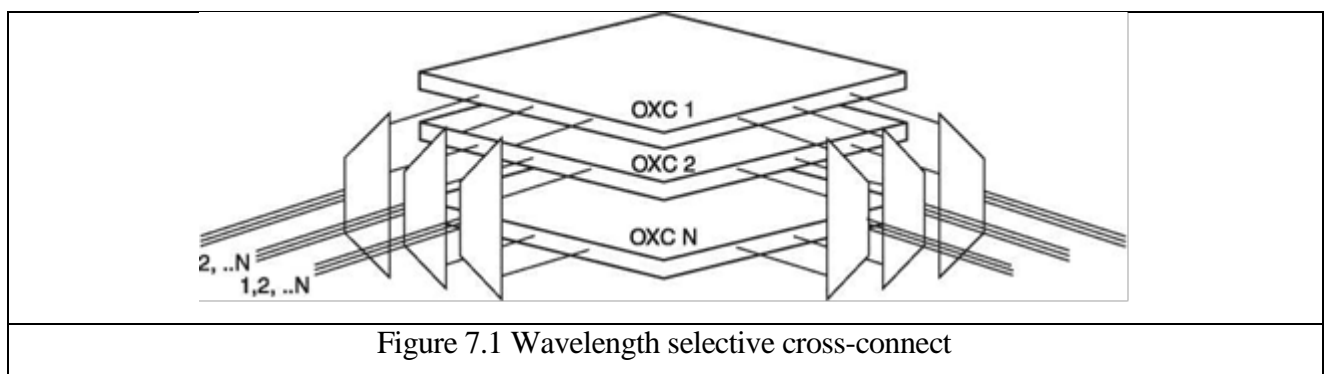
Optical add/drop multiplexer

Besides re-routing signals, another application for OXC is in an optical add/drop multiplexer (OADM). Figure 7.2(a) shows the basic principle of an OADM. After the incoming signal is de-multiplexed, each channel, which corresponds to a different wavelength, can be added or dropped by means of a 2 x 2 optical switch. The same can be achieved using an OXC. As illustrated in Fig. 7.2(b), signals of different wavelengths come in from the left, while the channels are added from the bottom. If a switch, for example in row y and column x, is turned on, the signal coming in along column x from the bottom will be added to the output while the original signal from row y will be dropped to column x at the top. OADM serves as an important component of an optical network.

7.2 Optical switch architectures

Numerous architectures of MEMS switches for OXCs have been proposed over the past decade. The basic mechanism for routing optical signal has mostly been reflection off micromirrors. Nevertheless, the configurations of the micromirrors and how they are actuated have a significant impact on the switching performance. In this section, different architectures such as two-dimensional (2D) and three-dimensional (3D) optical switches will be discussed in detail. Moreover, the pros and cons of these architectures will also be examined.

7.2.1 Two-dimensional optical switches

These switches are called 2D because the optical signals inside the switch only travel on a 2D plane, even though the switch is actually a 3D structure as the micromirrors move in and out of the signal plane.

Toshiyoshi and Fujita (1996) reported a 2 x 2 architecture that consists of four micromirrors. Each mirror can reflect light when it is in the ON position or let the light through when it is in the OFF position. When the mirror is in its OFF position, it lies horizontally above the signal path. To turn the mirror ON, one side of the mirror drops down such that it is now in a vertical position. The mirror then forms a 45° inclination to both the input and output signal paths. In order to keep the light on a 2D plane, the angle of the mirror must be well controlled. This can be easily achieved by a stopper that is implemented on the substrate below. The stopper prevents the mirror from turning more than 90°. The switch architecture is illustrated in Fig. 7.3. This architecture can be extended to TV x TV, with TV^2 micromirrors. The binary nature of the mirror position simplifies the control circuitry. However, this architecture does have its disadvantages. For instance, the micromirror count can go up very quickly as the port count increases. A 16 x 16 switch will require 256 mirrors, which take up a lot of space and hence making the size of the switch bigger. In addition, the crossbar switch also suffers from insertion loss variation due to the fact that the optical path length inside the switch is different along different

input/output paths.

Polygon switch

A polygon OXC was proposed by Lin et al. (1998). This architecture takes advantage of the connection symmetry property of the switch. In other words, if we have Input X going to Output Y and Input Y going to Output X in the TV x TV crossbar switch, the mirror arrangement will always be symmetric across the diagonal. The mirror count could be reduced if the two mirrors on each side of the axis of the symmetry can be combined into one. The polygon switch, as shown in Fig. 5.4, uses a double-sided mirror for each set of input-output pairs. For example, when the miiror at the intersection between the B2 and C2 path is in the ON state, optical signal coming in from port B2 will reflect off the mirror to output C2, while the signal from input C2 will use the other side of the same mirror to get to output B2.

Figure 7.3 2D 2 x 2 crossbar switch
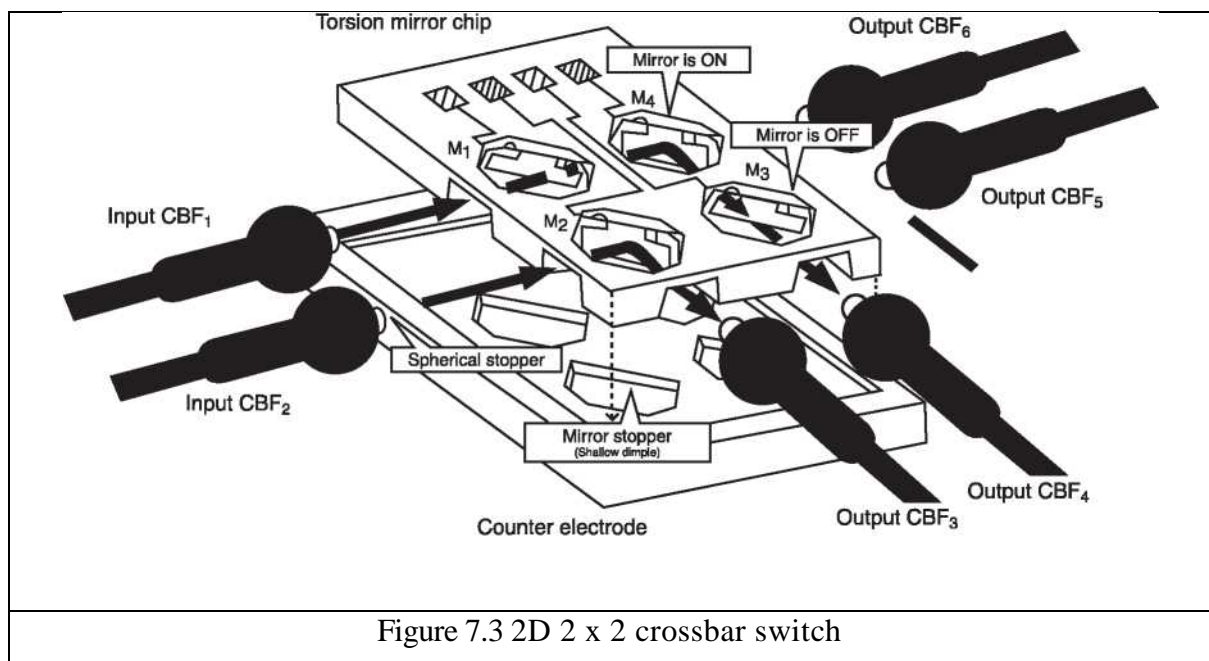
The control circuitry is further simplified with a reduced number of mirrors, in exchange for a more complicated fabrication process to produce double-sided micromiirors. However, this architecture requires a specific traffic connection, one that is symmetric, in order to function. For example, if input port B2 is routed to output port Cl, it would be impossible for the signal from input port C2 to get to output port B2.
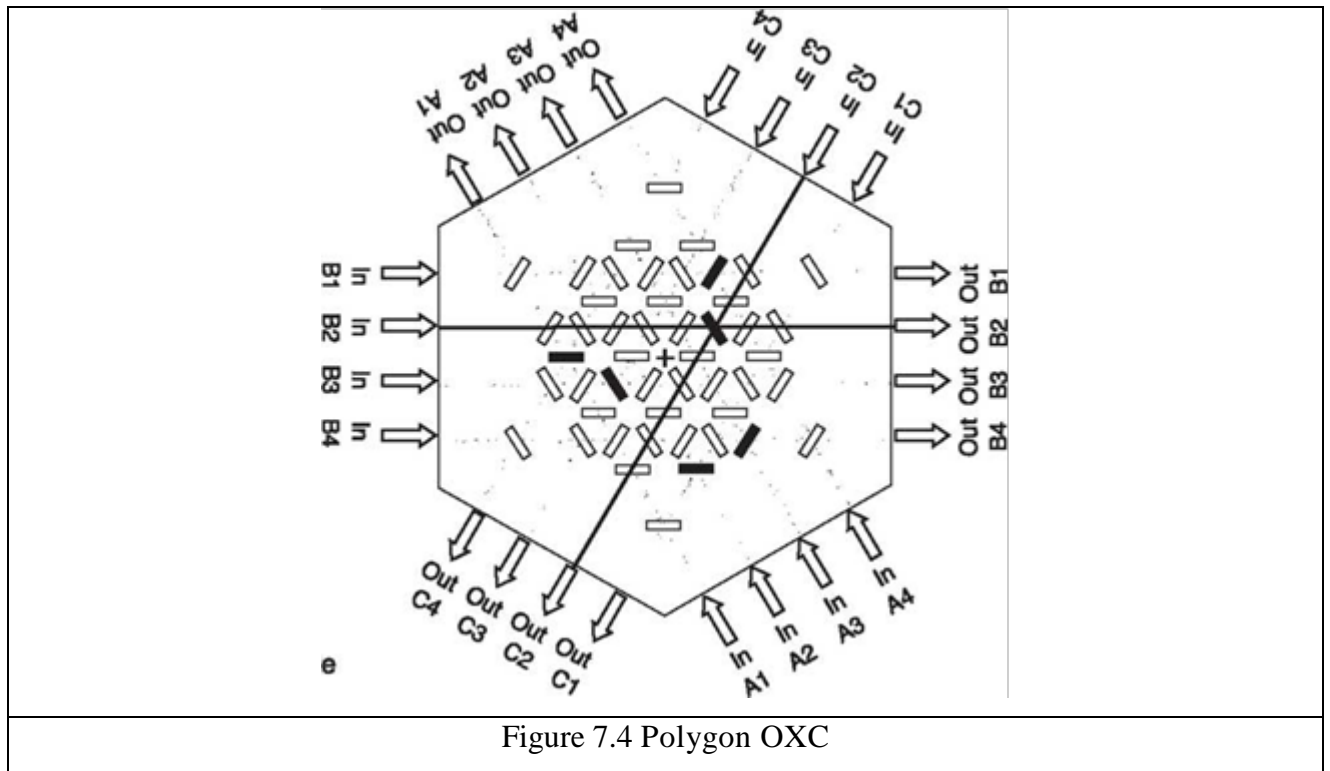
Figure 7.4 Polygon OXC

Re-arrangeable non-blocking switch

Non-blocking means that a new path can be set up even though an old path is already established (without affecting the existing connections). Re-arrangeable non-blocking, on the other hand, suggests that a new path can be formed but the old path might need to be rearranged. In comparison, a strictly non-blocking switch allows a new path to be created without any change in the existing paths. Shen et al. (2002b) introduced a re-arrangeable non-blocking architecture that reduces the mirror count from N^2 to $N(N+l)/2$. Similar to the polygon switch, this architecture also utilizes double-sided micromiirors. An optical signal might need to be reflected more than once, up to seven for a 4 x 4 switch in the worst case, before it can reach the output port. As a result, the variation in propagation delay might be a concern.

L-switching matrix

Yeow et al. (2003) came up with the L-switching matrix architecture that reduces the micromirror count of a conventional crossbar switch by one quarter. For an $N \times N$ ports switch, it requires $N^2/4$ double-sided and $2A^/4$ single-sided mirrors. Figure 5.5 presents a 4 x 4 architecture of the L-switching matrix. The top left quadrant is the input quadrant. It is where the double-sided mirrors are located. The single-sided mirrors are situated in the other two quadrants, top right and bottom left, with the output ports on the bottom right comer of the structure. To illustrate how the L-switching matrix works, imagine mirrors (1, 1), (3, 1) and (1, 4) are turned on. Optical signal from IN 1 will reflect off mirrors (1,1) and (1,4) to reach OUT1. Similarly, signal from IN3 will utilize the other side of double-sided mirror (1, 1) and mirror (3, 1) to reach OUT3. If mirror (1, 1) is switched off, only one reflection will occur for each

signal, as a result signal from IN 1 and IN3 will go to OUT3 and OUT1, respectively. One major advantage of the L-switching matrix is the difference between most distance and least distance path, as it is important to minimize the free-space propagation of optical beam. For a 32 x 32 port switch, the maximum path difference is only 30 pitches for the L-switching matrix while the number is over 60 pitches for the conventional crossbar architecture. A single pitch is defined as the distance between two adjacent mirrors. The optical signal will be reflected at most two times inside the switch. Care must be taken in the control scheme when establishing a path as multiple paths can be chosen in certain instances, as the L-switching matrix is re-arrangeably non-blocking but not strictly non-blocking. Yeow and Abdallah (2005) proposed a staircase-switching algorithm to minimize the occurrence of internal blocking conditions.

Figure 7.5 L-switching matrix
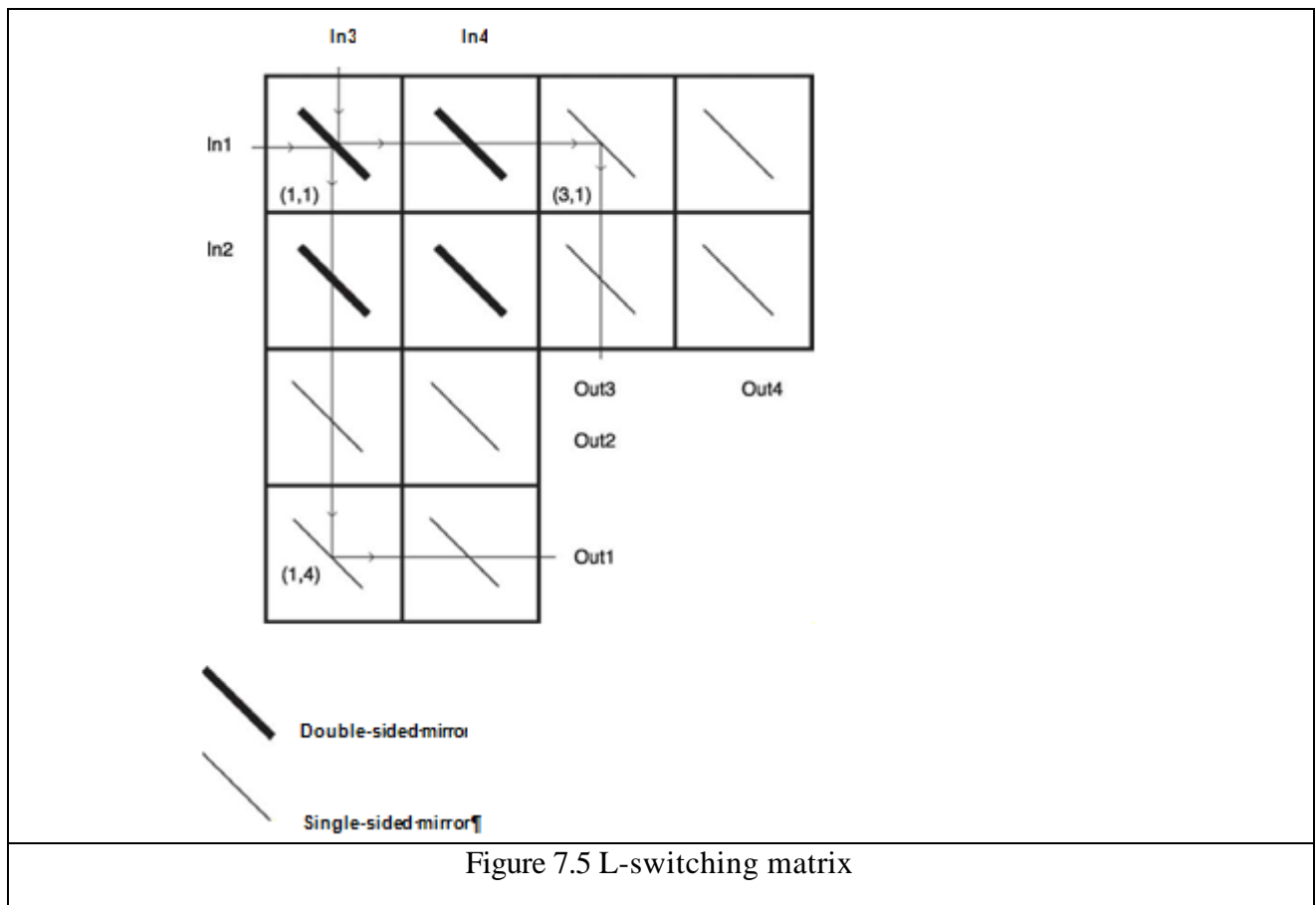
Multistage 2D switch

A typical commercially available OXC switch has a large number of input and output ports, for example, 256 x 256 in the Lucent LambdaRouter™ (Bishop et al., 2002). A switch of this size is generally not feasible using a 2D architecture as the number of mirrors, as well as the cost and the potential of having a faulty device, goes up very quickly. One solution to this problem is to construct a large matrix using multiple stages of smaller ones. Configurations such as the three- stage Clos network, the Spanke-Benes network and the Benes network have been proposed (Shen et al., 2002a; Li et al., 2003). A Clos network implementation is illustrated in Fig. 5.6. This special case of Clos network consists of only *mxm* switches, but the network is connected in a unique way so that it can be used to

route m^2 inputs to m^2 outputs. The switches are divided into three columns in a Clos network. The left column is called the ingress stage. The m^2 input signals are

divided into groups of m, with each group going into a different switch of the ingress stage. Each input switch then routes its m outputs to the switches in the middle column, called the middle stage. Similar signal routing happens between the middle stage and the output stage, or egress stage. The key point here is that every switch in the first two stages has a physical connection with all the switches in the subsequent stage. There are in fact m paths that an input signal can take to reach each output. This is necessary as all m signals from the same input switch could be destined for the same output switch in the worst case. The Clos network ensures that the input signals can be routed to the outputs in any combination.
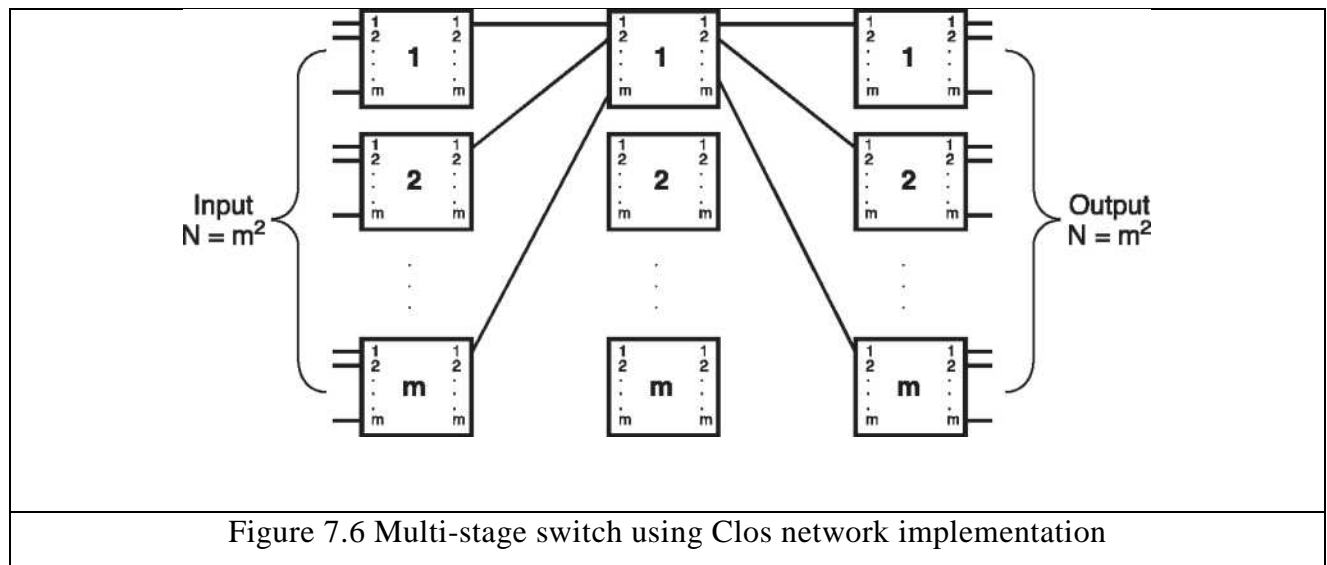
Figure 7.6 Multi-stage switch using Clos network implementation

7.2.2 Three-dimensional optical switches

To allow the optical signal to travel in 3D, the mirrors must be able to rotate about two axes. Both the fabrication process and the control scheme are more complicated for the 3D switches, yet the switch can be made smaller compared to a 2D switch with the same number of ports because the number of micromirrors required is greatly reduced. For an N x N port switch, a typical 3D implementation would require N or $2N$ switches.

7.2.2.1 Lucent Microstar

In 2000, Lucent Technologies announced a 3D OXC, the Microstar™ micromirror array, which consists of micromirrors that can rotate about two axes (Aksyuk et al., 2000). A picture of the micromirror can be seen in Fig. 5.7. It achieves two-axis rotation by using a gimbal ring, which is able to rotate about one axis, and have the mirror turn about another axis with respect to the gimbal ring. The angular range of each axis is greater than ±6°. The product, LambdaRouter, uses the $2N$ configuration for 3D switch and contains 2 x 256 mirrors in a single OXC switch. There is one micromirror associated with each input port and also one for each output port. Optical signal from the input port reflects off the input mirror to the output mirror and then to the output port. As a result, a 256 x 256

strictly non-blocking switch can be achieved. In addition, a 1024 x 1024 OXC using a similar approach was also demonstrated in a later publication (Aksyuk et al., 2003). Since the introduction of the Lucent LambdaRouter, several 3D MEMS OXCs have been reported that use similar architecture but different micromirror structures or actuation methods (Bernstein et al., 2004; Fernandez et al., 2004).

Microlens scanner

Instead of using mirrors, Takahashi et al. (2007) proposed a microlens structure that could be potentially used in an OXC switch. Two sets of microlens array are used, with one each on the input and the output sides. Each lens in the input array can be moved to direct the incoming light beam to any of the output optical fiber through the corresponding lens in the output array. A 9 x 9 OXC with coupling loss of 13.8 dB was reported
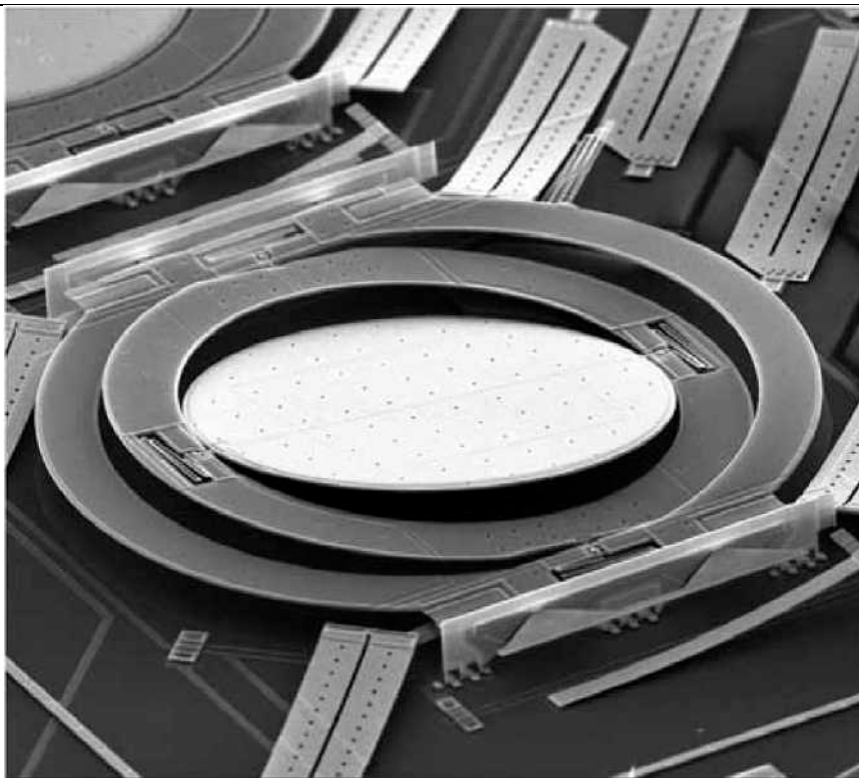
Figure 5.7 Image of 3D micromirror used in the Lucent LambdaRouter

7.3 Actuating principles of MEMS-based optical switches

The movement of MEMS micromirrors needs to be controlled accurately and reliably in order to be used in an OXC. Moreover, other properties such as small size, easy to fabricate and low power consumption are also desirable. As a result, the actuating principles of the MEMS micromirrors have attracted a great deal of attention in research. This section will present several actuating principles that have been applied to micromirror switching.

7.3.1 Electrostatic

One of the earliest implementations of MEMS optical switches, proposed by Toshiyoshi and Fujita

(1996), uses electrostatic force to drive the micromirror movement. The basic principle of electrostatic relies on the attractive forces between opposite charges. If there is an electric potential difference between two objects, an electrostatic field will be developed which pulls the two objects towards each other. In the Toshiyoshi design, the mirrors are normally held at an OFF position. When the underlying electrode is charged to a voltage different from that of the mirror, the mirror will be attracted to the electrode and be tilted to the ON position. Electrostatic actuation is very reliable as it is highly repeatable. It has also been well researched and modeled. However, one major disadvantage of using electrostatic force for micromirror actuation is the requirement of very high driving voltages. Toshiyoshi and Fujita reported that they required a voltage in the range of 100 to 150 V to turn on the mirror.

Given the high driving voltage requirement of electrostatic micromirror actuation, several researchers have come up with ways to relax this constraint. Yoon et al. (2002) reported that they were able to lower the driving voltage to about 50 V by introducing an extra vertical electrode. Kuo et al. (2004) proposed a structure using stress-induced bending micromirrors, as illustrated in Fig. 5.8. Electrostatic force was used to attract a bending beam downward, moving the mirror into its ON position. Driving voltages of 25 and 18 V, depending on the side of the beam that was attracted, were reported.

It should be noted that the 3D optical switch developed by Lucent Technologies also employed electrostatic actuation (Aksyuk et al., 2003). Four electrodes, two for each axis, were used. The voltage required was less than that of a 2D switch, so as to avoid the 'snap-down' point of the mirror.

7.3.2 Electromagnetic

Using electric current to generate magnetic field is another efficient way to actuate MEMS micromirrors. Electromagnetic actuation requires much lower driving voltage than that of the electrostatic mechanism. However, switches employing electromagnetic actuation suffer from crosstalk as the magnetic field can potentially affect adjacent micromirrors if shielding is not done properly. Miller et al. (1997) from the California Institute of Technology demonstrated an electromagnetic 2x2 fiber optic bypass switch. The double-sided mirror in the switch lies below the optical fibers in its OFF state to let the signal through. When the switch is activated, the mirror moves up and reflects the signal into neighboring optical fibers (see Fig. 7.9). It was reported that only 30 mA of current is needed to achieve a mirror displacement of 200-300 pm that is necessary to get the mirror completely out of the signal path in the OFF state.

A team from LG Electronics Institute of Technology presented another electromagnetic MEMS optical switch (Ji et al., 2004). In their paper, they also proposed the use of an electro-permanent magnet to make a latchable optical switch. The nonzero remnant flux of the magnet helps maintain the mirror in a certain state even when the power is off. On the other hand, the magnetic field can be removed by applying an opposite current to the coil. Power consumption can be greatly reduced in this scheme as current is needed only during state changing of switches.
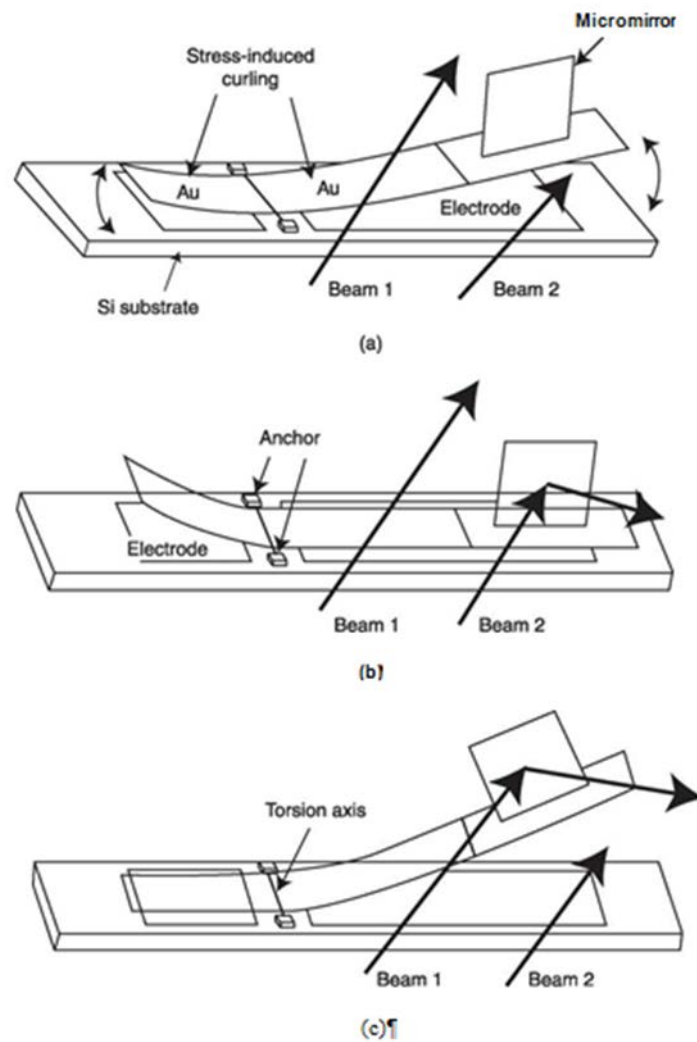
Figure 7.8 Stress-induced bending micromirror: (a) OFF state, (b) ON state on the right side and (c) ON state on the left side

Electromagnetic actuation can be used in 3D optical switches as demonstrated by a number of researchers. The challenge lies in the fact that the angle of the micromirror has to be controlled precisely over two axes. Bernstein et al. (2004) demonstrated a mirror that can be tilted by more than 10° per mA in each axis,

Figure 5.9 Electromagnetic fiber optic bypass switch
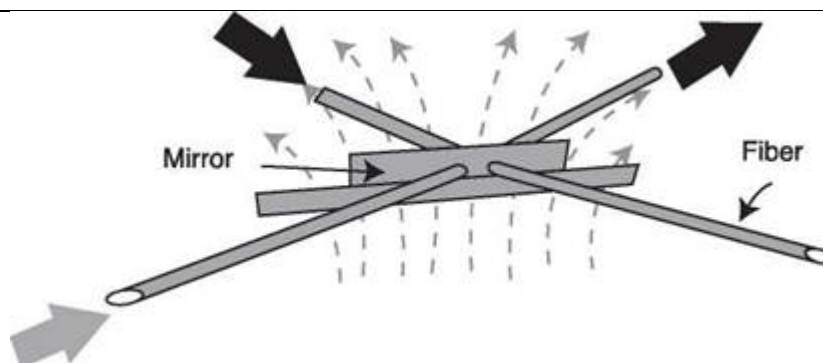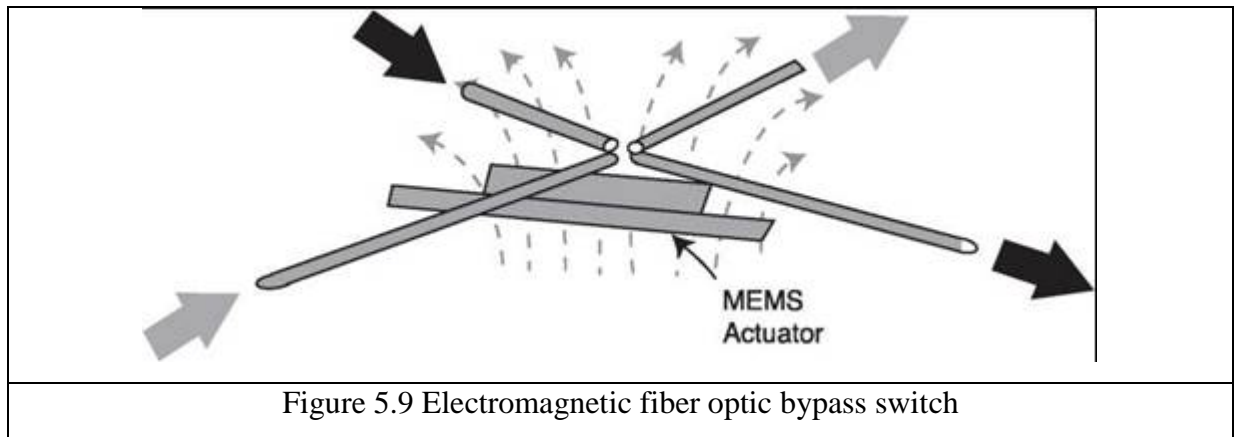
7.3.2 Other actuating mechanisms

Electrostatic and electromagnetic mechanisms are the two most widely used actuating methods in MEMS optical switches today. However, there are several other actuating mechanisms that are gaining popularity and are worth mentioning.

Comb-drive

The comb-drive could be grouped into electrostatic actuation but it does not use any electrodes to attract the micromirror directly. Instead, a voltage is applied between two combs to generate an electrostatic field and move the mirror (Jung et al., 2006). This more complicated structure provides a larger force density as the surface area between the two combs is bigger. Also, since the two combs will never be in contact with each other, the mirror's range of motion can be made larger as it does not suffer from the typical snap-down or pull-in effect.

Scratch drive actuator

A team from AT&T research labs has utilized scratch drive actuator (SDA) to rotate the mirror (Lin et al., 1999). The mirror is connected to a translation plate through pushrods and micro-hinges. The SDA mechanism causes the translation plate to move, which in turn gets the mirror to rotate up if the translation plate is moving away from the mirror or vice versa. The basic idea is presented in Fig. 5.10. It has been shown that the tilt angle of the mirror can be controlled precisely using SDA. But the fact that the mirror can only rotate about one axis limits its application to 2D switches only.

Thermal

The micromirror movement can also be controlled thermally. Werber and Zappe (2006) proposed a thermo-pneumatically actuated micromirror that can be tilted up to 13°. This is caused by the expansion of the gas in a sealed cavity. Michael et al. (2005) fabricated a beam that can be expanded when heated. The beam, which can be buckled upward or downward, is part of a bi-stable microbridge that puts the micromirror either in and out of position. Since the heat can be conducted to nearby cells, crosstalk is a potential problem for thermally actuated mirror array.

Figure 7.10 Switching of micromirror using scratch drive actuator
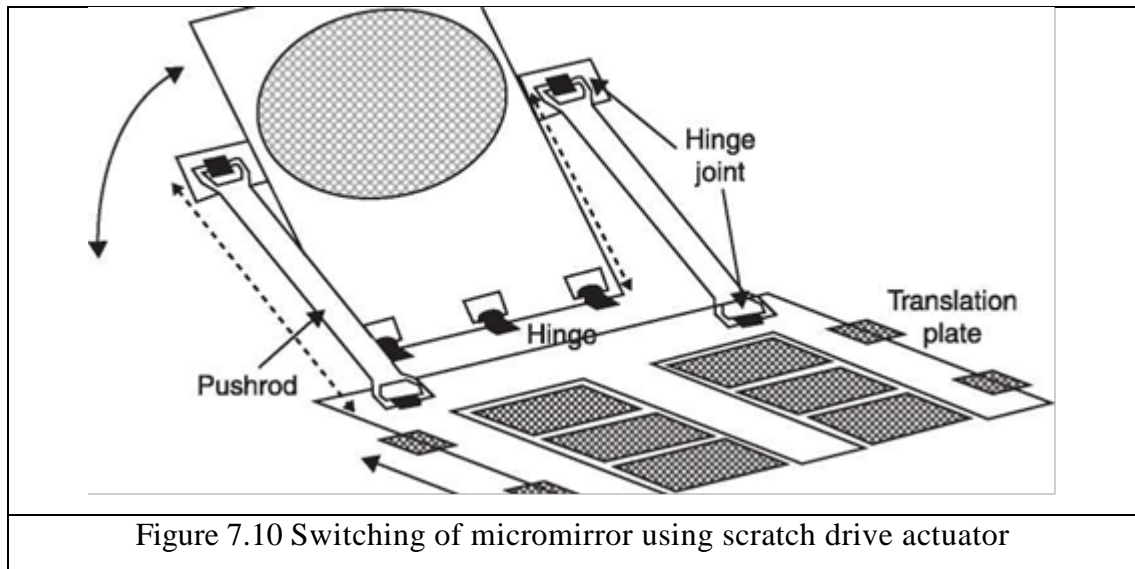
Micro-motor

Finally, a micro-motor can be used to rotate the mirror as Oohira et al. (2004) demonstrated. As in a normal motor, an AC signal is needed to power it. It requires low driving voltage as a 5 V peak-to-peak sinusoidal signal at 80 kHz is required. However, the angular resolution of the motor was limited to 0.5°.

Closed-loop versus open-loop control

Pointing accuracy is an important performance parameter for micromirrors in 3D switches. Unlike their 2D counterparts, which operate only in binary states, mirrors in 3D switches must have their tilt angle controlled precisely. The easiest control scheme is by means of an open-loop control system. As the driving voltage goes up, the mirror is moved by a larger angle. Each driving voltage corresponds to a different tilt angle. If the relationship between the driving voltage and the mirror angle is consistent, the desired tilt angle can be achieved using a look-up table. On the other hand, factors such as process variation, temperature and aging could change the actuation relationship. If that is the case, a closed-loop system is preferable as the driving voltage is adjusted continuously until the desired mirror tilt angle is reached. However, some form of sensing mechanisms must be present, which makes the switch bigger and more complicated. Also, the servo control algorithm must be carefully designed to ensure good stability.

Challenges surrounding MEMS-based optical switches

Among all the various technologies that can be used to implement optical switches, MEMS seems to show the most promise. However, there are still important issues that need to be addressed before the MEMS approach can become the technology of choice for the next-generation optical switches.

Reliability

MEMS-based optical switches must be able to function in adverse conditions, as well as over an extended period of time. The fact that there are mechanical moving parts inside the switch makes reliability a bigger concern in MEMS than in traditional solutions such as integrated circuits. As

mentioned in section 5.4, therelationship between mirror displacement and driving voltage could change due to factors such as temperature and aging. The closed-loop approach helps combat that problem. The change, if any, in the switching performance after millions of switching cycles also needs to be studied. In order to gain wide acceptance in optical networks, the reliability and repeatability of MEMS optical switches must be well researched and understood.

Scalability

The demand of network bandwidth is always increasing. An optical switch installed today, even with port redundancy built in, could potentially run out of port count in the next few years. Instead of replacing the switch with a bigger one, using a switch with the ability to expand port count is a more cost-effective approach. A software-based switch might be able to increase its port count by a firmware upgrade. But MEMS-based optical switches are physically limited by the number of mirrors. A combination of clever algorithm and novel architecture is needed to overcome this disadvantage.

Manufacturability

Performance of MEMS-based devices could change from wafer to wafer. A large process variation means a low yield, which will ultimately drive the cost high. A simpler process is highly desirable because fewer fabrication steps means less variation. Process control for MEMS is difficult because unlike normal integrated circuit process, in which there are a fixed number of layers with pre-defined dimensions, the MEMS fabrication steps are different among designs. Research in novel fabrication processes and materials will advance the MEMS manufacturing process. Moreover, standardization of processing steps will allow a better process control.

Packaging

Because of the mechanical nature of MEMS-based optical switches, packaging has a huge impact on their performance. MEMS packaging is more demanding than integrated circuit packaging since MEMS devices have a closer interaction with the outside world. In addition, MEMS external signals are not just electrical in nature. Standardization of packaging, as in the case of the integrated circuit industry, would reduce MEMS design cost.

Competing technologies

MEMS-based optical switch has an advantage over traditional OEO switch, mainly because it is future-proof. However, it is not the only all-optical switch technology MEMS-based optical switches must be able to outperform other switch technologies in performance parameters such as noise figure and switching speed.

8 MEMS thermal actuator MEMS actuation created by thermal expansion

8.1 Introduction

Thermal actuation makes use of the thermal expansion of solids as they are heated, or the differences in the rates of thermal expansion between different materials as in a bimorph actuator. Relative to electrostatic actuation, thermal actuation requires lower voltages, on the order of 10–25 V, and can generate larger forces, on the order of mN. Thermal actuators typically use Joule heating, where a current is passed through a resistor, requiring relatively high currents (mA) and the dissipation of continuous power (mW) to maintain their positions. Electrostatic actuators require relatively high voltages, on the order of 10–100 V, but they do not require high continuous current levels to maintain their positions and only dissipate power when the electrostatic actuator is charged or discharged. They generate smaller forces relative to thermal actuators, on the order of mN. They are used where lower power and lower forces are required.

In a resistive thermal actuator, Joule heating causes an actuator material to heat up by an amount ΔT. The heating leads to thermal expansion, which can be used to create a deflection for an actuator. If we consider the thermal expansion of a bar of length L, as shown in Figure 5.3, as the bar is heated by an amount ΔT it will expand by an amount ΔL, giving rise to a strain ε, where

$$\varepsilon_t = \frac{\Delta L}{L} = \beta \Delta T \qquad\qquad 8.1$$

where β is the coefficient of thermal expansion (CTE) of the material. If the bar is constrained, the thermal expansion will give rise to a stress σ, where

$$\sigma = E\varepsilon \qquad\qquad 8.2$$

where E is the Young's modulus of the material. The stress σ will give rise to a force F, where

$$F = A\sigma \qquad\qquad 8.3$$

where the bar has a cross-sectional area A.

Types of thermal actuators:

- Asymmetric (bimorph)
- Symmetric (bent beam, chevron)

Main parameters of MEMS thermal actuators are maximum displacement, achieved maximum force and maximum frequency. These parameters are interrelated and interdependent. Correlation between maximum force and maximum displacement for different thermal actuator types is depicted on Fig. 8.1.

8.1 Thermal actuators: force and distance
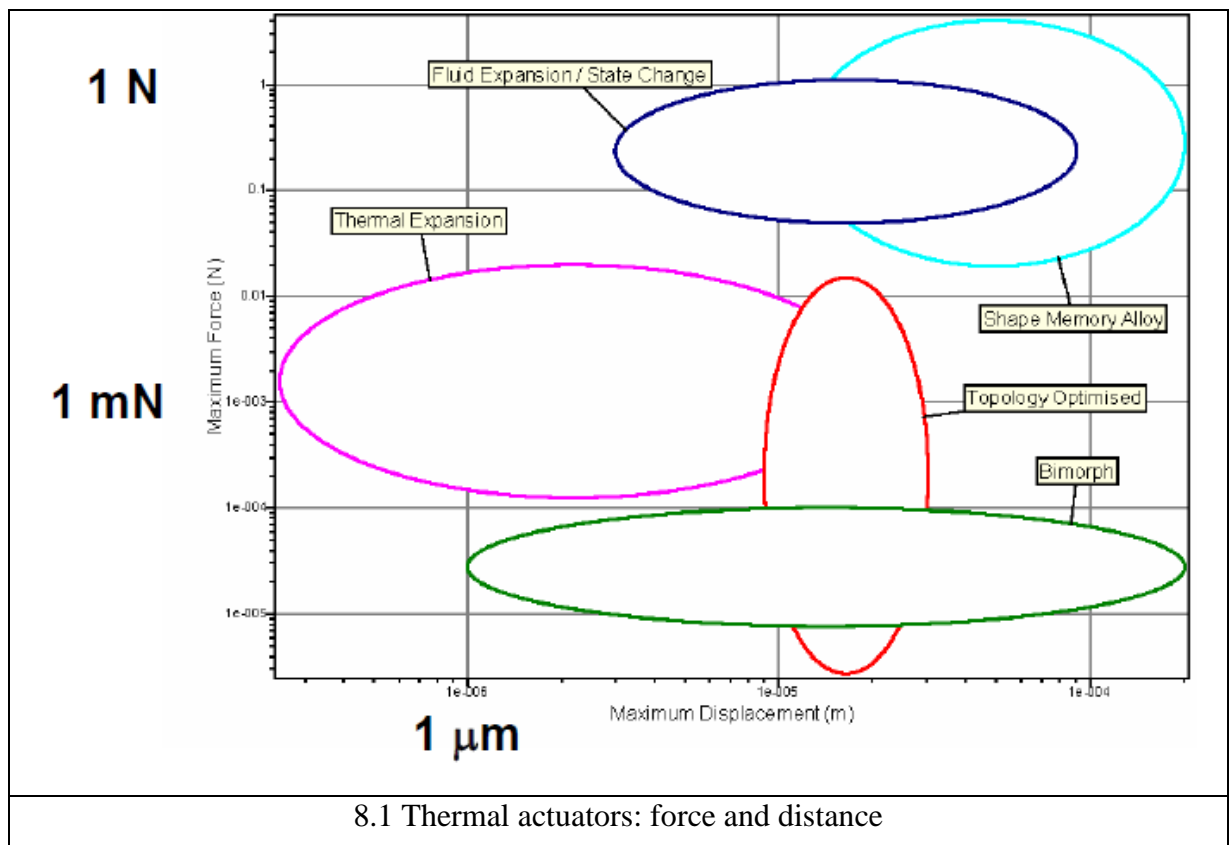
. Correlation between maximum force and maximum frequency for different thermal actuator types is depicted on Fig. 8.2.

MEMS thermal actuator requirements, wish list of consumers and developers and up-to-date implementation level is shown in Table 8.1.

Table 8.1 Thermal actuator wish list

Low power operation	✘
Fast switching	✔
High force	✔
Large analog-controllable actuator travel	✔
Simple fabrication(is possible, at least): • Few masks • Standard, CMOS-compatible processes and materials • Ability to be fabricated on the same chip as circuits	✔
Low voltage	✔
Robust operation, not prone to failures☐	✔
Ability to sense as well as actuate	✘

8.2 Thermal actuators: frequency and force
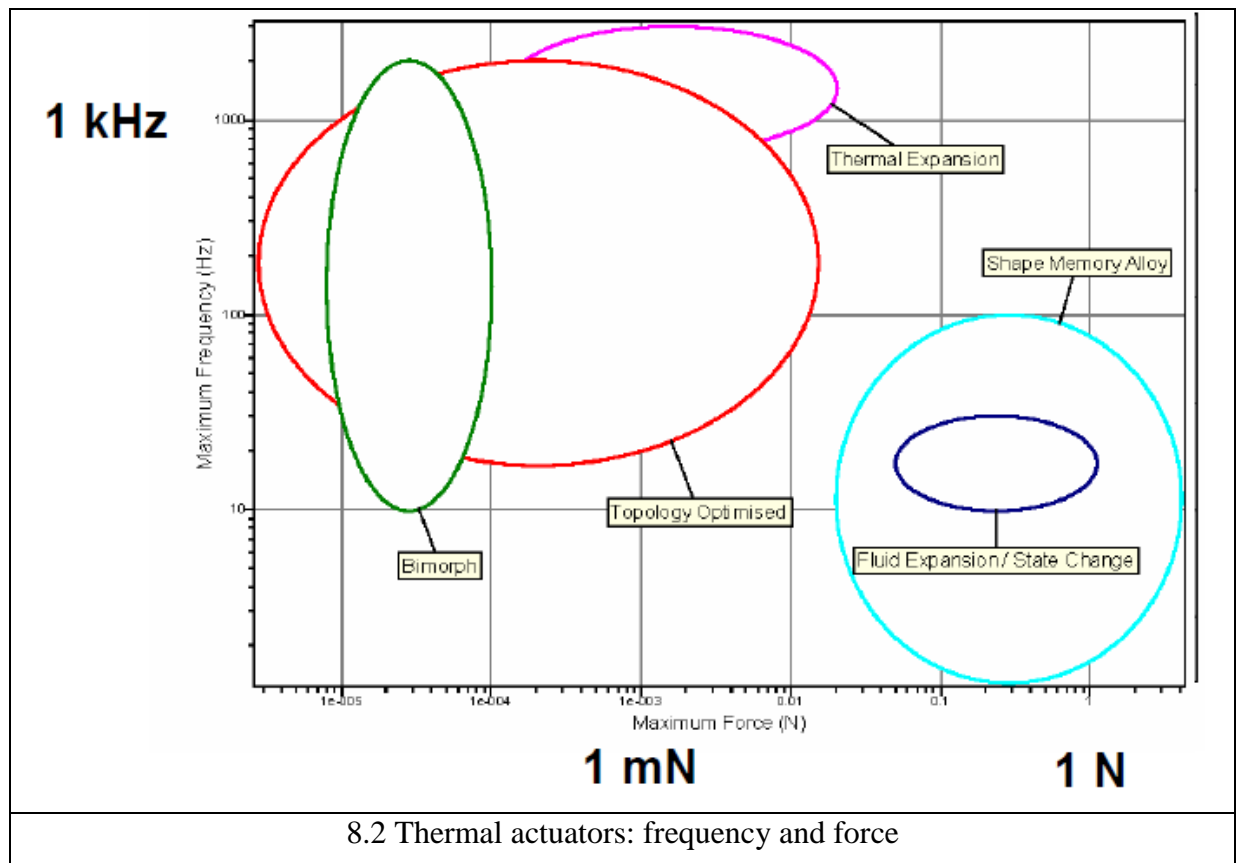
8.2   Basic of thermal actuation

8.2.1 Thermal strain.

Temperature increasing leads to thermal expansion of bar: $\Delta T\uparrow \rightarrow \Delta L\uparrow$.

Fig. 8.3 One side fixed bar
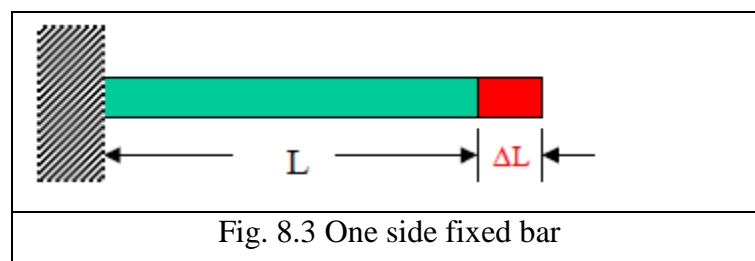
Example: L=1 mm, Area=2mm×2mm, ΔT =100ºC, β_{Si}=2.3·10^{-6} /ºC, E_{Si}=10^{11} N/m^2

$$F = AE\beta\Delta T = (2\times10^{-6})^2 \cdot (10^{11}) \cdot (2.3\times10^{-6}) \cdot (100) = 9.2\times10^{-5}N = 92mN$$

Comparison	Thermal actuation force F=92 *mN*
	Electrostatic force =10^{-8}~ 10^{-9} N
Conclusion	Thermal actuation force is larger comparatively to electrostatic force

Displacement $\Delta L = \beta\Delta TL = (2.3\times10^{-6}) \cdot (100) \cdot (10^{-3}) = 0.23\mu m$

Conclusion: displacement is rather small.

8.2.2 Thermal actuation in a pivoting design
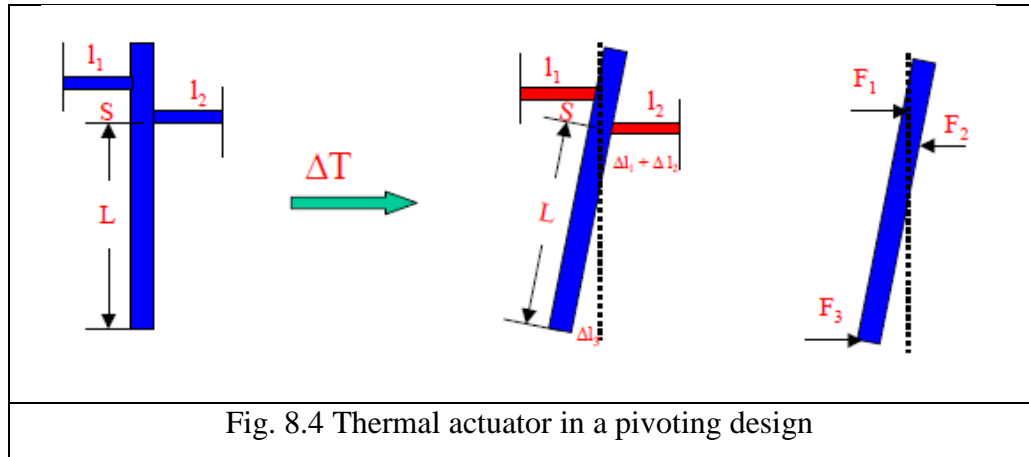
Fig. 8.4 Thermal actuator in a pivoting design

$$\Delta l_3 = (\Delta l_1 + \Delta l_2)\left(\frac{S+L}{S}\right)\uparrow \qquad\qquad 8.4$$

$$F_3 = F_2\left(\frac{S+L}{S}\right) = F_1\left(\frac{S}{L}\right)\downarrow \qquad\qquad 8.5$$

Deflection is increased, but force becomes smaller

8.2.3 Thermal actuation in buckling design.

Small displacement of thermal expansion $\Delta L = \beta \Delta T L$. Displacement can be increased if beam buckling occurs at some critical force $F_{critical}$.
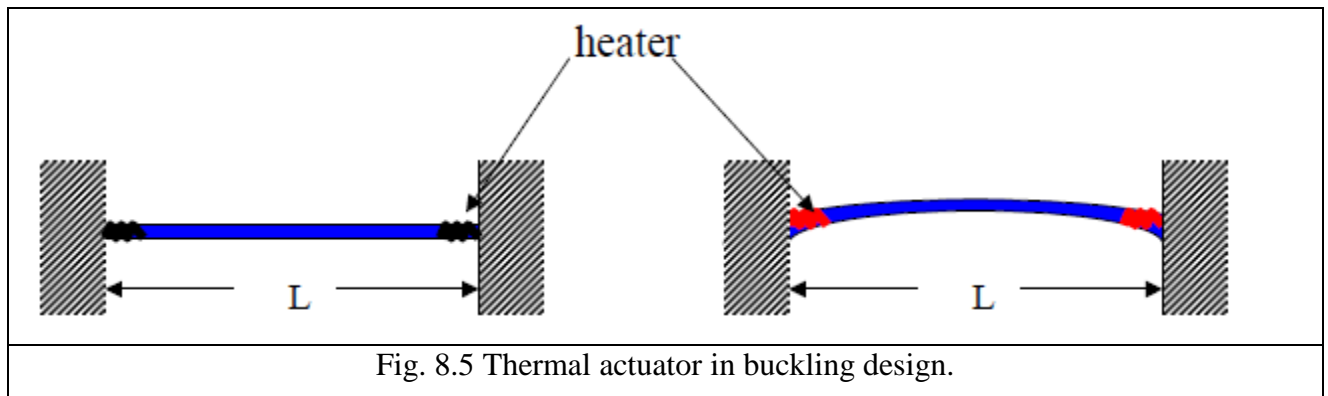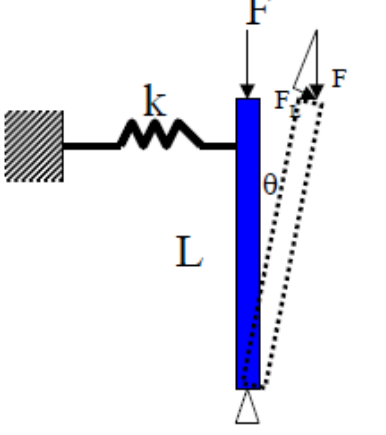
Fig. 8.5 Thermal actuator in buckling design.

	If F is small: => no buckling because kx spring keeps system stable. If F is large: => spring force: $F_s = -kx = -kL\theta$ $F_L = F\sin\theta$
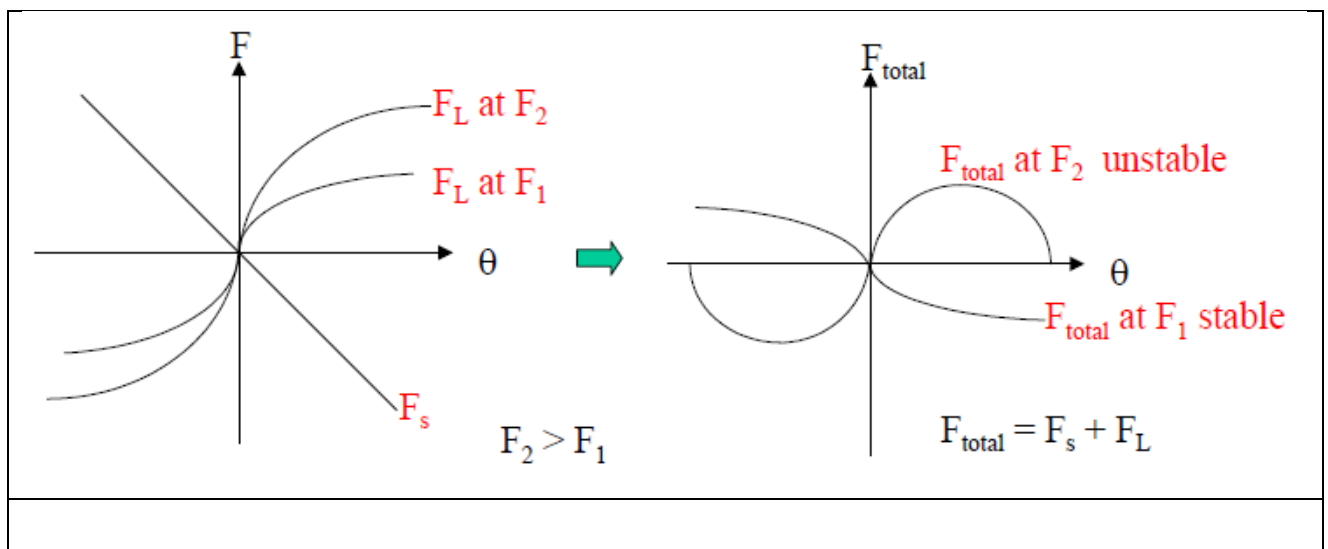
$$\frac{\partial F_{total}}{\partial \theta} = \frac{\partial (F_s + F_L)}{\partial \theta} = \frac{\partial (-KL\theta + F\sin\theta)}{\partial \theta} = -KL + F\cos\theta$$

$$\left.\frac{\partial F_{total}}{\partial \theta}\right|_{\theta=0} = -KL + F = 0$$

$$F_{critical} = KL$$

$$F_{critical} \approx \frac{\pi^2 EI}{L^2} = \frac{\pi^2 E a^3 b}{12 L^2}$$

For a beam with 2μm×2μm cross section

$$EI = \frac{10^{11} \cdot (2 \times 10^{-6})^3 \cdot (2 \times 10^{-6})}{12} \approx 10^{-13}$$

$$F_{critical}^1(a, b, L) = F_{critical}(2\mu m, 2\mu m, 1\mu m) \approx \frac{\pi^2 EI}{L^2} \approx \frac{\pi^2 \cdot 10^{-13}}{(10^{-6})^2} = 1N$$

$$F^2_{critical}(a, b, L) = F_{critical}(2\mu m, 2\mu m, 1mm) \approx \frac{\pi^2 EI}{L^2} \approx \frac{\pi^2 \cdot 10^{-13}}{(10^{-3})^2} = 1\mu N$$

Example: L=1 mm, Area=2mm×2mm, ΔT =100°C

$$F_s = AE\beta\Delta T = (2\times10^{-6})^2 \cdot (10^{11}) \cdot (2.3\times10^{-6}) \cdot (100) = 9.2\times10^{-5}N = 92mN$$

$$F^1_{critical} = 1N > F_s = 92mN \Rightarrow \text{no buckling}$$

$$F^2_{critical} = 1\mu N < F_s = 92mN \Rightarrow \text{buckling} => \text{Large displacement}$$

8.2.3 Thermal Bimorph design

Example		Al thickness t_f=10⁻⁷ m
		E_{Al}=73 GPa
		β_{Si}=22.5·10⁻⁶ °C
		Si thickness t_{Si}=2·10⁻⁶ m
		E_{Al}=10¹¹ Pa
		β_{Si}=2.3·10⁻⁶ °C
		ΔT=100 °C
		Neglect possion's ratio: υ=0
		Sample width: 2 μm
		Sample length: 10 μm

$$F = \left(\frac{E}{1-\vartheta}\right)_s \frac{wt_s^3}{12\rho \cdot l}$$

$$Z = \frac{l^2}{2\rho}$$

$$\rho = \frac{\frac{t_s}{2} + \left(\frac{1-\upsilon}{E}\right)_f \left(\frac{E}{1-\upsilon}\right)_s \frac{t_s^2}{6t_f}}{(\beta_f - \beta_s)\Delta T}$$

$$\rho = \frac{\frac{2 \times 10^{-6}}{2} + \left(\frac{10^{11}}{73 \times 10^9}\right)\frac{(2 \times 10^{-6})^2}{6 \times 10^{-7}}}{(22.5 \times 10^{-6} - 2.3 \times 10^{-6}) \cdot 100} = 0.005m$$

$$Z = \frac{22.5 \times 10^{-6}}{2 \cdot 0.005} = 10nm$$

$$F \approx 10^{11} \cdot \frac{(2 \times 10^{-6})(2 \times 10^{-6})^3}{12 \cdot 0.005 \cdot 10 \times 10^{-6}} = 2.67 \times 10^{-6}N$$

Thermal actuator speed

Because thermal actuation is based on a temperature change, the speed of these actuators is limited to the heating and cooling of materials. The time constant for the thermal expansion of a slim, uniform beam of length l_0 has been approximated as:

$$\tau = \frac{l_0^2 C_p \rho}{\pi^2 \lambda}$$

where C_p is the specific heat capacity, ρ the density and λ the thermal conductivity of the material.

Two common expander materials are (poly)silicon and polyimide. Table shows the work densities and response times for 1 mm long beams of these two materials. Although polyimide requires a lower temperature to reach the same work density as polysilicon, and thus is more energy efficient, its low thermal conductivity makes it too slow for valve applications.

Table

Title	E_y [N·m^{-2}]	α [K^{-1}]	C_p [J·kg^{-1}·K^{-1}]	ρ [kg·m^{-3}]	λ [W·m^{-1}·K^{-1}]	ΔT_{avg} [K]	ξ_{max} [J·m^{-3}]	τ [s]
Silicon	1.6×10^{11}	2.6×10^{-6}	700	2,400	157	600	4.9×10^4	1.1×10^{-3}
Polyimide	2.5×10^9	5.5×10^{-5}	2,000	1,420	0.16	250	5.9×10^4	1.8

Thermal actuator designs

Surface-micromachined thermal actuators utilize constrained thermal expansion to achieve amplified motion. The thermal expansion is most commonly caused through Joule heating by passing a current through thin actuator beams. There are two different thermal actuator designs that have been demonstrated and commonly used in the literature, the pseudobimorph or "U" shaped actuator, and the bent-beam or "V" shaped actuator. Both designs amplify the small input displacement created by thermal expansion, at the expense of a reduction in the available output force.

The U shaped actuator operation, illustrated in Figure , relies on creating a temperature difference between a hot-arm and cold-arm segment. The temperature difference is due to the reduction in Joule heating in the cold-arm because of its decrease in electrical resistance resulting from the increase in cross-sectional area. This results in a thermal expansion difference between the two segments. Because both segments are constrained at their base the actuator end experiences a

rotary motion. Multiple actuators can be connected together in parallel to increase the output force and to create a linear output motion if desired.

U shaped thermal actuator.
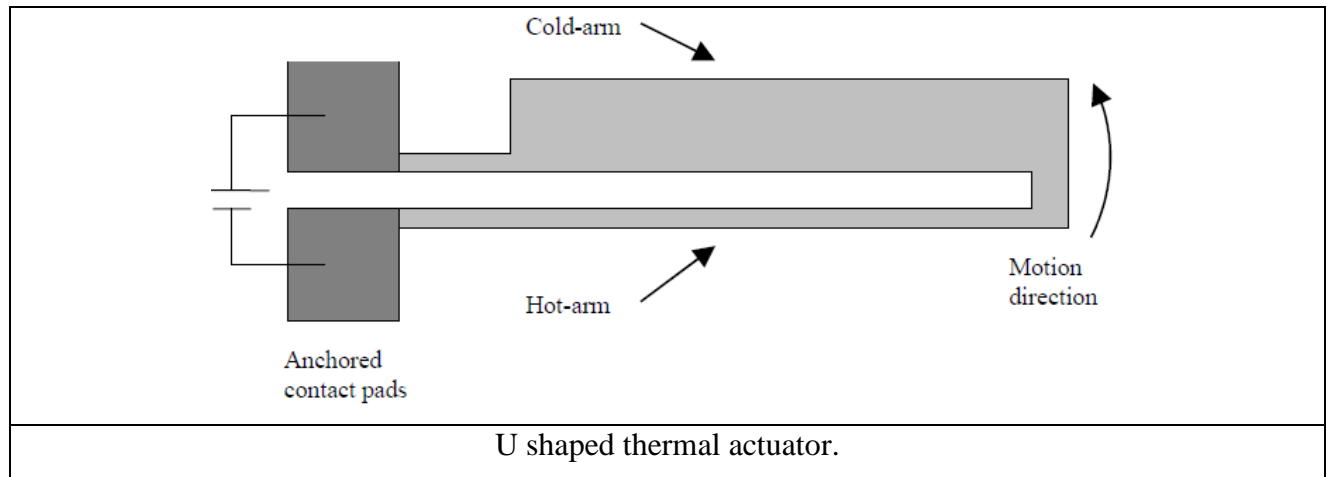
The V shaped, or chevron style actuator is illustrated in Figure. This design is characterized by one or more V shaped beams, also commonly called legs, arranged in parallel. As current is passed through the beams they heat and expand, and because of the shallow angle of the beams, the center shuttle experiences an amplified displacement in the direction of the offset.

Illustration of V shaped actuator.
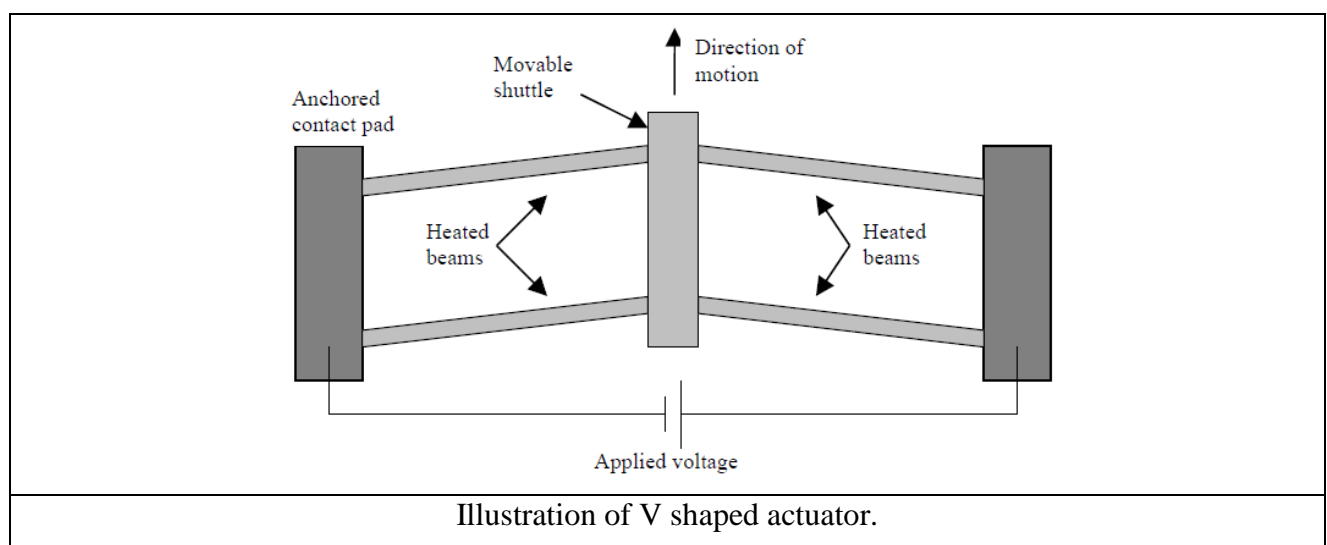
Application of MEMS thermal actuator

MEMS Proportional Control Valve

Precise fluid flow control is a key requirement in many reaction, production or analytical processing systems. Over the past decade, miniaturization of complex fluidic systems has led to many new and potential applications, including microreactor assemblies , microscale chemical analysis systems and medical systems such as rapid DNA sequencers and drug micro-dosers. However, the need for a small-volume, high-precision mass flow controller for minute fluid flows has not yet been fulfilled. A typical mass flow controller can be modeled as shown in Figure . It

consists of a mass flow sensor, a proportional control valve and control electronics connected in a closed control loop, set to adjust flow to a specific set point.
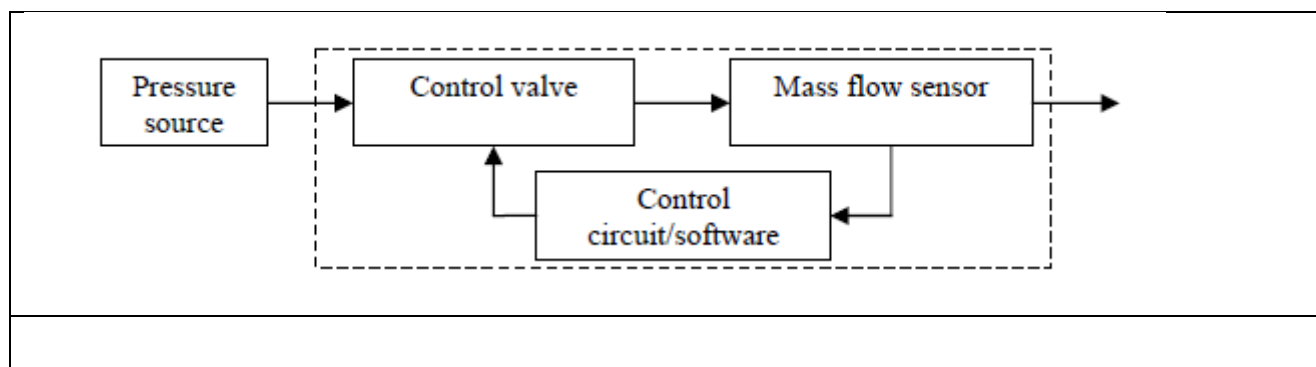
The potential applications of such a miniature mass flow controller include the following:

- Evaporation systems for the production of solar cells, requiring a very stable flow rate without overshoot;
- Microreactors for flow chemistry, demanding chemical compatibility and low leakage for safety reasons;
- Gas and liquid chromatographs, requiring a very stable flow rate and a large dynamic control range to facilitate both high-pressure and low-pressure flows;
- Dosing systems for the food production and pharmaceutical industries, demanding high control precision and easily cleanable or replaceable components;
- Flow control for medical purposes, e.g., in blood pressure measurement systems, requiring compatibility with the human body.
- Other common applications of dosing systems include lab-on-a-chip, point-of-care test systems or in-situ drug micro-dosers.

Given a certain differential pressure, controlling fluid flow requires control over the flow resistance of at least one element in the flow circuit. In most cases, this resistance change is achieved by changing the mechanical geometry. Some non-mechanical alternatives have been reported, such as electro-capillary or diffuser microvalves, but since there is no actual closure these systems cannot guarantee low leakage. They are therefore not considered in this review.

A basic design of a mechanical control valve is shown in Figure , consisting of an orifice in a fixed plate (the valve seat), which is covered by a vertically translating plunger.
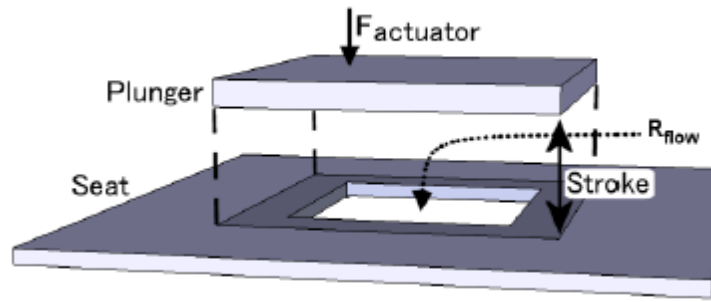
Typical structure and operating principle of a microvalve. The moving plunger covers an orifice in the valve seat, thus changing the flow resistance

The maximum flow supported by the valve is defined by the structural dimensions—specifically, the diameter of the orifice and the separation between the two plates. The maximum separation that can be achieved is called the stroke. The leakage performance of a valve is determined by the closing surface area, the relative surface roughness and flatness of seat and plunger and the force by which the plunger is pressed against the seat.

Mechanical microvalves are commonly divided into active and passive systems, *i.e.* with and without a powered actuator. Passive valve designs are not evaluated in this study, because they are either one-way flow rectifiers or designed to control a single, specific flow. In order to achieve precise control over a range of flows, a powered actuator is required. The maximum actuator force determines the maximum differential pressure across the valve, as the actuator needs to be able to counteract the fluid force acting on the plates. The maximum actuator stroke limits the valve stroke, unless a mechanical amplification scheme is applied. The actuator also defines the power dissipation of a control valve, as well as its open-loop response time.

Valve Specifications. Many properties determine the performance of (micro) valves, including flow range, control precision, leakage flow, power usage, speed, physical dimensions and dead volume. The specific set of demands for any microvalve design is listed in Table.

Table

	Specification	Influenced by
Max flow range	1 g/h (liquids), 100 sccm (gases)	Orifice diameter/Stroke
Flow control resolution	0.1 mg/h (liq.), 0.01 sccm (gas)	Resolution of plate separation
Leakage	$<10^{-6}$ mbar L/s He	Seal surface area/Relative roughness & flatness
Max. differential pressure	>10 bar	Closing force
Static operating power	<100 mW	Actuator/Dimensions
Settling time	<1 ms	Actuator/Dimensions
Dimensions	$<10 \times 10 \times 10$ mm^3	

A typical value for the closed-loop settling time of an entire mass flow controller is 100 ms, which is generally dominated by the response time of the flow sensor and the stabilization of the

fluid flow. To ensure that the valve does not become a limiting factor, an open-loop response time of 1 ms is demanded for the microvalve. Finally, the total valve volume should be kept as low as possible, so that the flow control system as a whole remains compact and easy to integrate.

One of the "translating plate" concepts is to fix one side of the plunger to the surrounding bulk. This turns a translating motion into a tilting motion. With an appropriately chosen pivot point, this geometry can supply an amplification of actuator stroke (or force). An early tilting plate design used a flexible beam fixed on one side to the bulk as shown in Figure. An electrode embedded in the beam enabled direct electrostatic attraction to the substrate, but due to the electrostatic pull-in effect the valve suffered from very poor control precision.

Electrostatic tilting plate microvalve
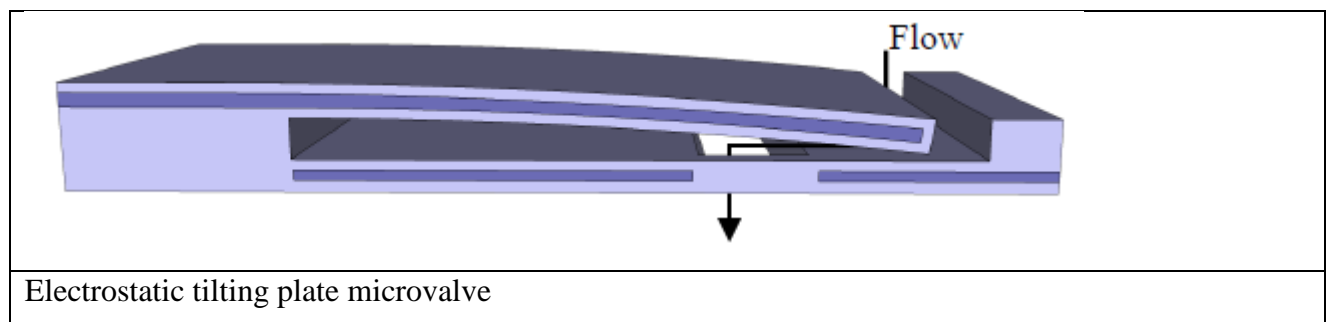
A design better suited to flow control applications is used in the commercially available NC-1500 Fluistor, shown schematically in Figure 6. This valve uses thermal actuation, where the formation of evaporation bubbles creates a small expansion of a movable membrane. This small initial stroke is then transferred so that the actuator and plunger tilt away from the valve seat over a larger distance.

Design and operating principle of the NC-1500 Fluistor, a commercially available
tilting plate microvalve
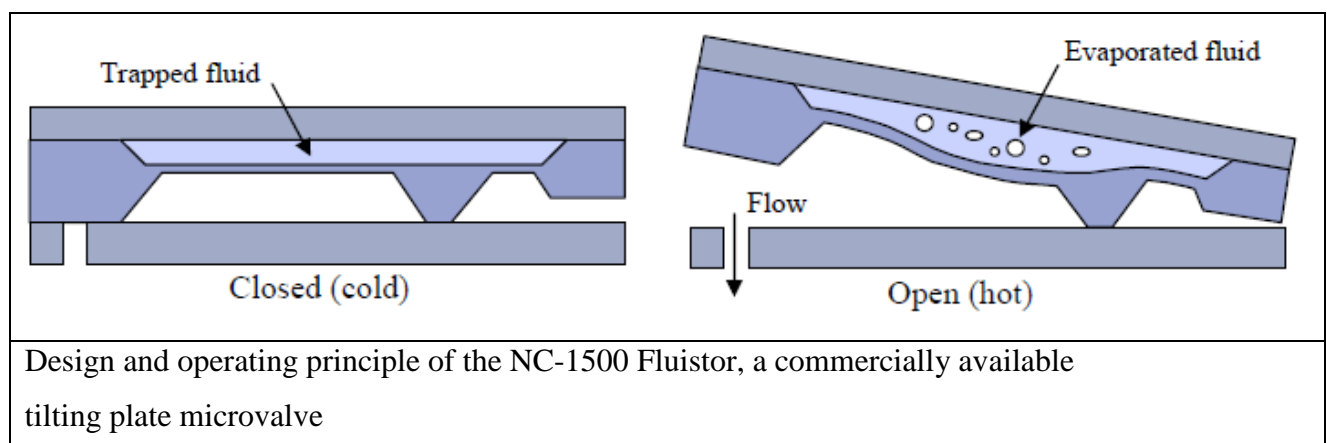
A more elaborate variation on the tilting plate concept utilizes specific materials, or combinations of materials, to fabricate a plunger that detaches itself from the valve seat. Material properties, such as thermal expansion, piezoelectric strain or phase-change shape memory, can be applied to obtain plates that bend or curl away from the valve orifice, as shown in Figure.

Schematic drawings of bending platemicrovalve design concepts
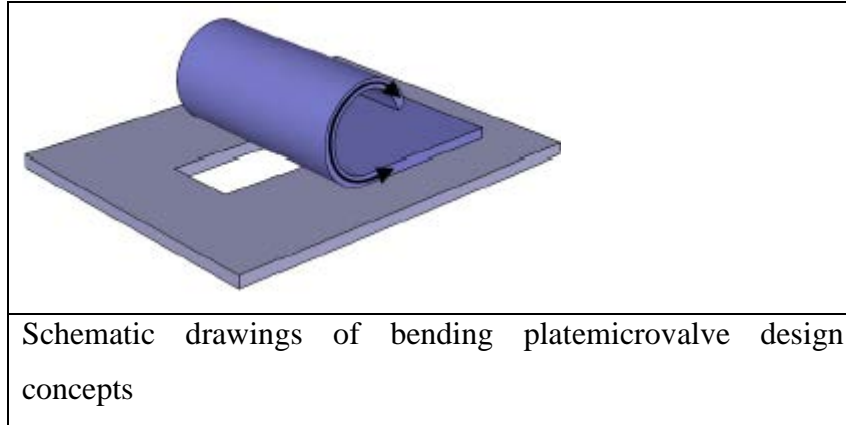
The gradual "peeling" of the plate can significantly reduce the force required to overcome plate stiction. It can also allow for precise flow control, provided the actuation has sufficient resolution. A drawback is the large bending moment created by the fluid pressure on the necessarily compliant plate, which leads to low leakage performance at large differential pressures.

A variation of the bending plate scheme that could circumvent this problem is presented in Figure. It uses an annular piezoelectric disc that contracts radially in the in-plane direction, eliminating any moving parts in the third dimension and so reducing the total volume of the device. Because the actuation direction is perpendicular to the fluid pressure, the stiffness in the direction of the fluid pressure can be increased without penalty.

Annular disc actuation scheme
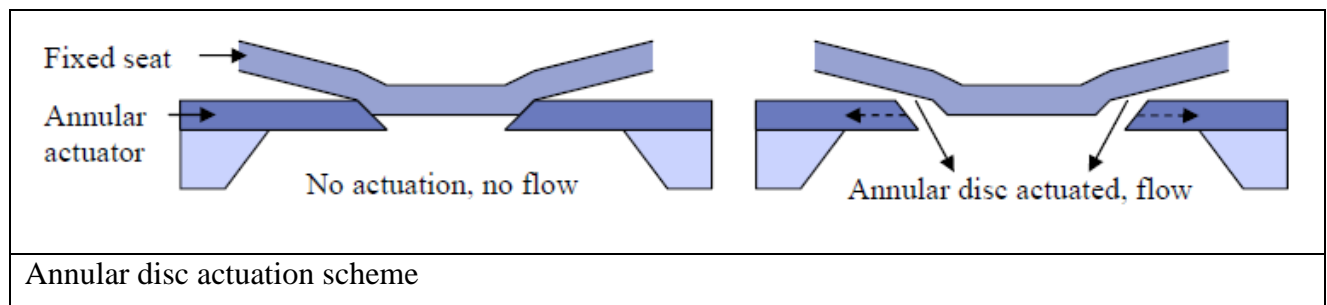
A concept intuitively well suited to flow control applications is the use of a horizontally translating or sliding plate to vary the overlap between two orifices, shown in Figure.

Schematic drawings of horizontally translating plate platemicrovalve design concepts
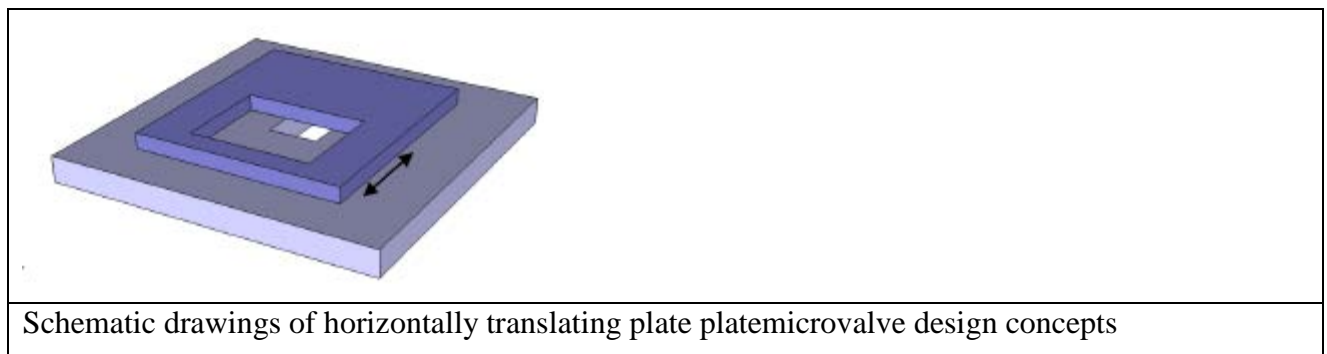
Apart from straightforward control, a big advantage of such a design is that it can be made to require zero power in the steady state. The sliding motion can be executed linearly, in rotation or using a pivoting structure. The a pivoting structure design is used in a valve shown in Figure

Thermal expansion of silicon rods pushes a perforated plate across the valve seat, changing the overlap of the orifices.

Horizontally translating plate microvalve using thermal actuation in a pivoting design. The dashed line marks the perforated plate, which overlaps with the inlet and outlet orifices.
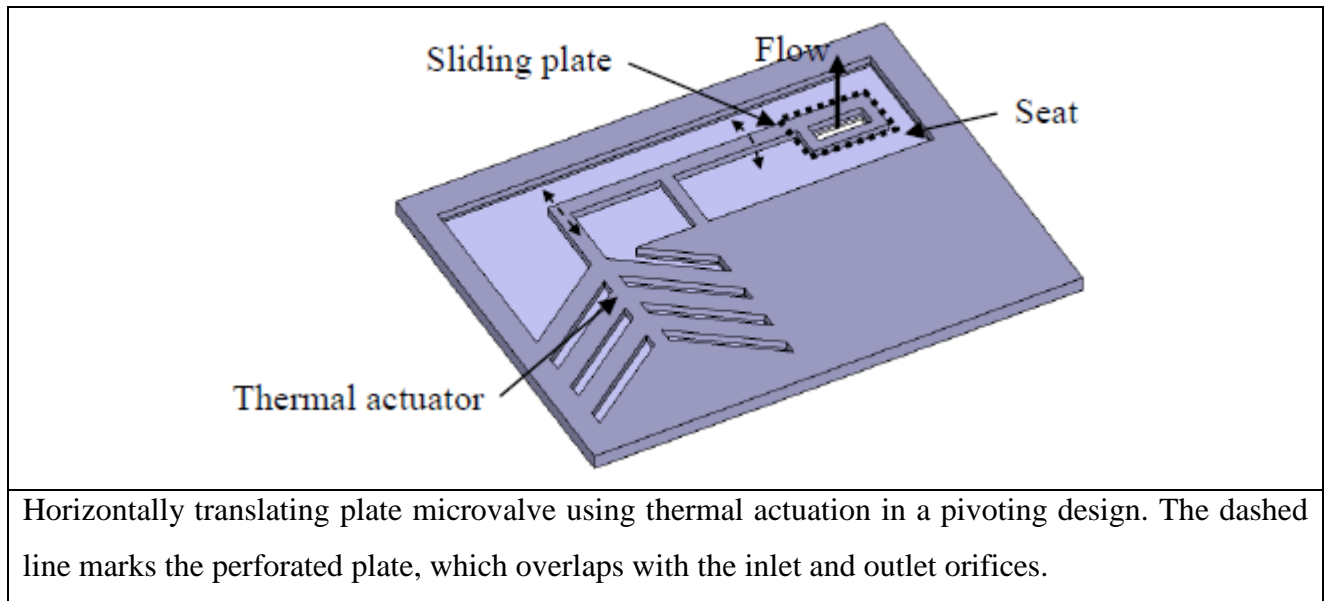
In contrast to its good control properties, the sliding or rotating plate concept suffers from very poor leakage performance. This is because the sliding movement takes place in the same plane as that in which the channel closure needs to be achieved. For low-force, high-precision actuation it is desirable to have low friction between the two plates, whereas good closure requires pretension between the plates, which leads to friction. Thus there is an inherent tradeoff between friction and closure.

Microgrippers

Nanorobotic pick-and-place is one of the most encouraging approaches towards the prototyping of nanostructure-based devices, which is a real challenge. Due to the availability of achieving high forces at a relatively small size, electrothermally actuated microgrippers are promising tools for overcoming this challenge.

Micro-grippers find applications in microrobotics, microsurgery, micro-fluidics, microrelays, assembling and miniature medical instrumentation. Actuation principle involved may be electrothermal, electrostatic, piezoelectric, shape memory and electromagnetic. It has been found that thermal actuation provides greater displacement at low voltages when compared to other mechanisms. Microgripper is comprised of two microactuators (hot-and-coldarm actuator) which operates on the basis of Joule heating and thermal expansion. The hot-and-cold arm actuator consists of one narrow (hot) arm, one wide (cold) arm and the flexure. Flexure joins the wider arm with the anchor as shown in figure
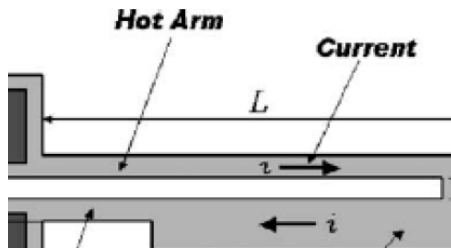
When voltage is applied in series to this structure, the current will flow through these arms with same heat distribution. Thus the narrower arm gets more heated due to small resistance according to the relations; $R=\rho L/A$ and $H=I^2R$, where R is the resistance, L is the length and A is the area of cross section of the arm, H is the Joule heat produced in the arm and I is the current flowing through the arm. As a result of more heated narrower arm, it will deflect more than wider arm. Thus, the narrower arm creates mechanical force and pushes the structure in direction narrower to wider arm. Also, when the voltage is applied in parallel, then the structure bends in opposite direction.

The narrow arm gets more heated than the wider arm as shown in figure

Surface temperature, K

Total displacement, μm
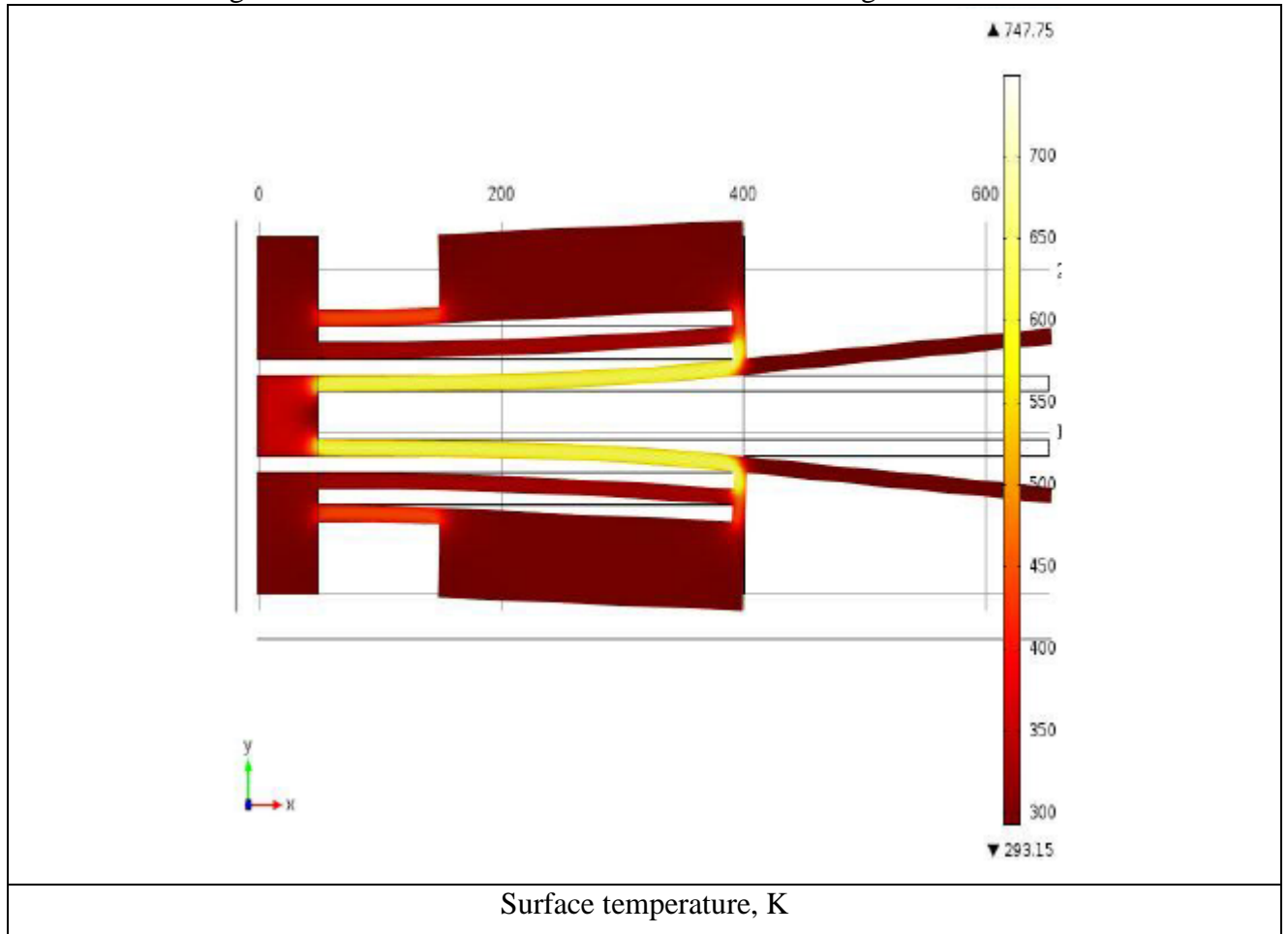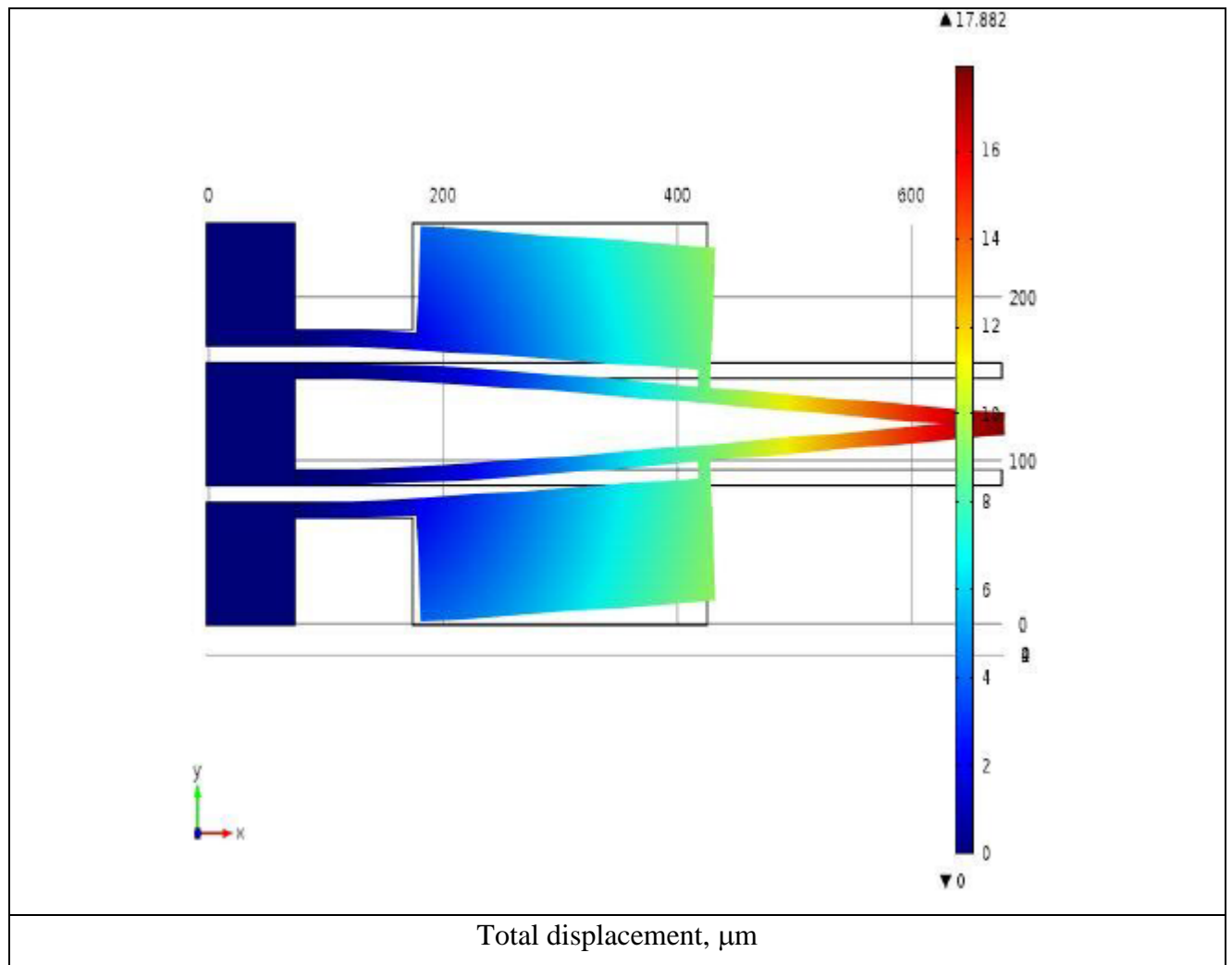
The maximum temperature lies in the middle of the hot arm i.e., 747.75K and displacement at the tip is 17.882μm at 3V as shown in figure.

Three microgripper generations
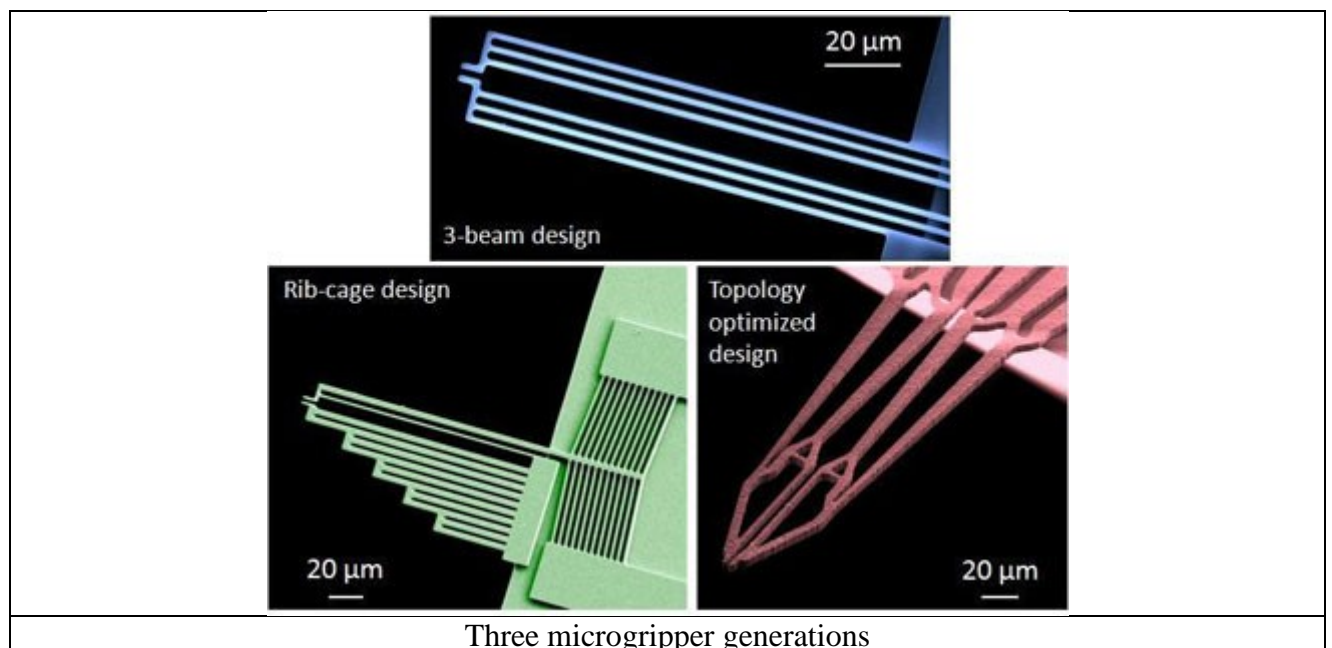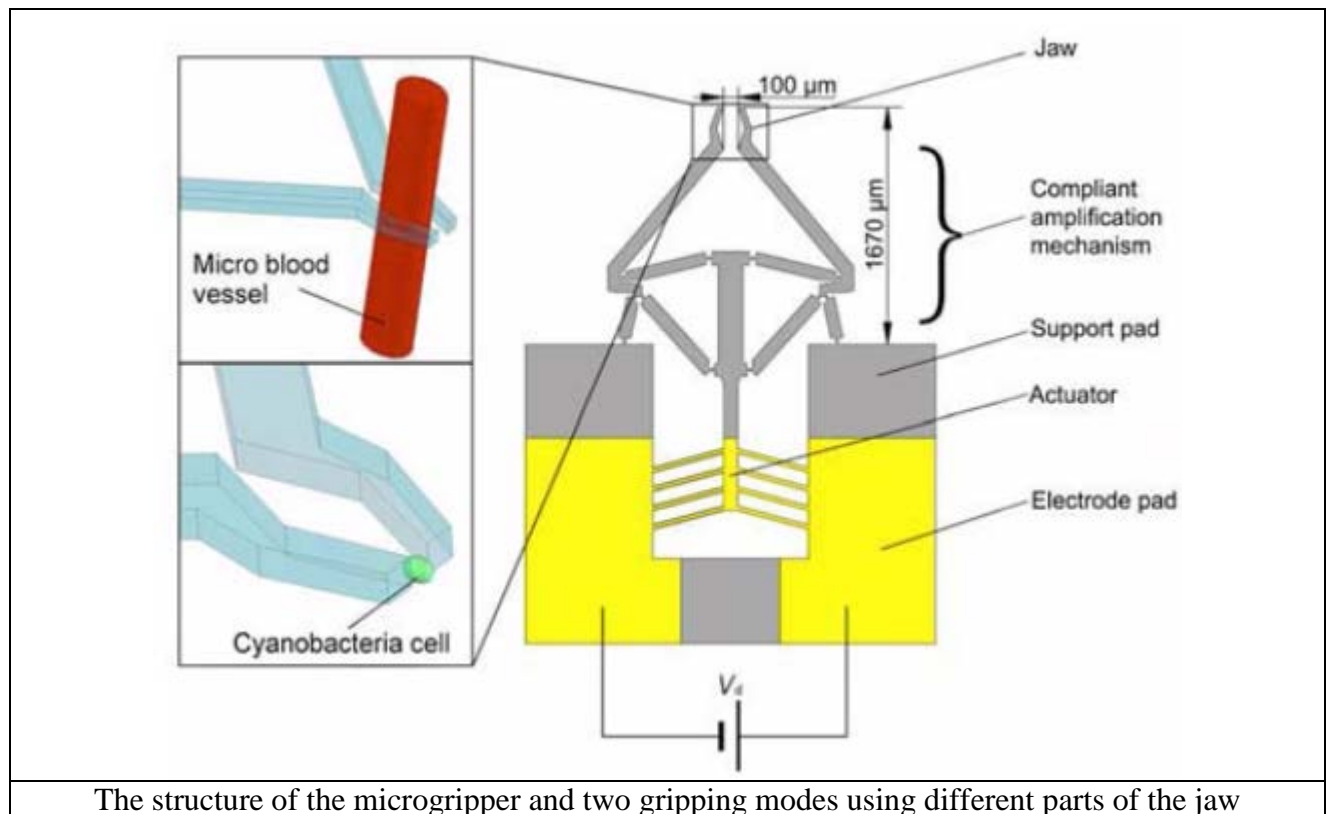
The structure of the microgripper and two gripping modes using different parts of the jaw

MEMS thermal actuator for micro relays.

There exists expanding demands for very small relays(micro relays) which will be used for small components or systems in many fields. Micro relays fabricated by micromachining technologies are expected to have many advantages such as very small size, low power consumption, low cost and integration capability with other devices on the same silicon die. There are several researches on micro relays fabricated by micromachining processes. Some use electrostatic and the other use electromagnetic force. Electrostatic suffers from trade off between travel and excitation voltage. And electromagnetic inherently has scale-merit disadvantage and coil fabrication complexity.

Micromachined relay using a thermal buckling actuator
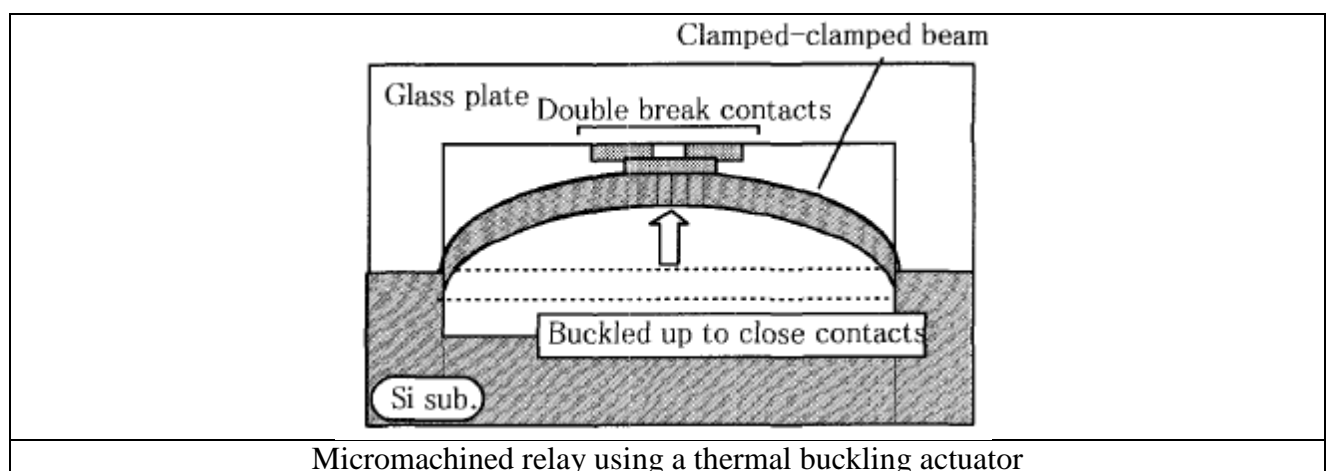
On the contrary, thermal actuators have the advantage of operation time reduction as those are fabricated in smaller dimension. Also since structures are fairly simple, low cost devices can be expected. From those backgrounds, the clamped-clamped beam buckling actuator which is driven by thermal expansion can be used for micro relay applications. The buckled clamped-clamped beam structure has larger derivative of strain energy at stable equilibrium than that of cantilever structure

with the same size, material and deflection in general. Therefore, the actuator can generate larger force without sacrificing deflection. Also in this way small deformation of thermal expansion can be converted to larger deflection without using atmosphere temperature-sensitive bimetallic approach. As an example, the actuator can be applied to a micro relay by bonding it to a glass plate which has fixed contacts (Fig.).

Fig. illustrates the cross section of the actuator. It has a thin silicon clamped-clamped beam which is sandwiched with thin silicon dioxide layers. A poly-Si heater line is formed on the silicon dioxide.

Thermal buckling Clamped-clamped beam actuator
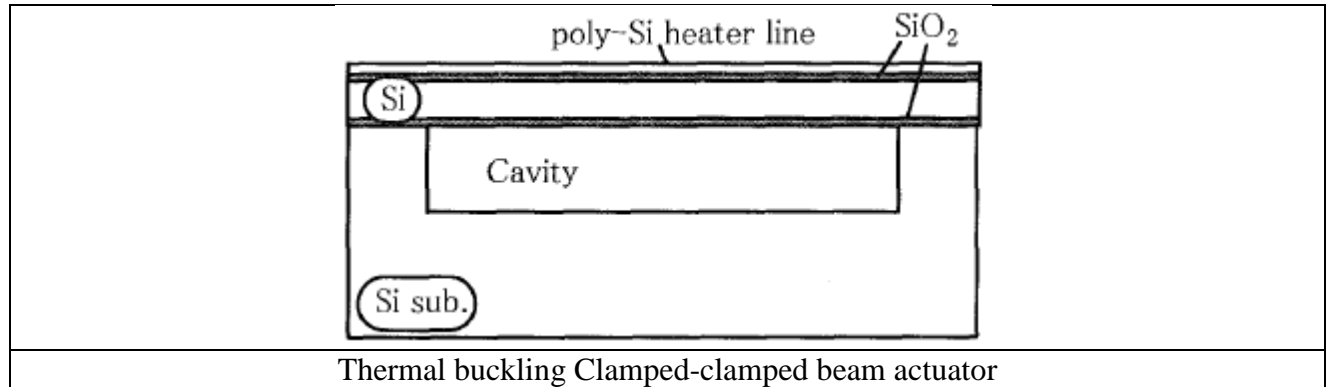
MEMS thermal actuator in medicine.

One of the most commonly used implants for the brain is the microelectrode that is used to either record or stimulate single neurons or neuronal networks in the brain. Long-term functionality of the microelectrode implants in the brain is critical for the success of emerging applications in cortical neural prostheses and other long-term neurobiology studies on the progression of disease states, learning and memory, development of neuronal networks etc. However, current microelectrode technologies are inconsistent in their performance and unreliable in long-term experiments. Multi-unit recordings obtained from current fixed microelectrode technologies typically show deterioration in signal quality with time and almost all of them fail over periods ranging from few weeks to few years. Further, the same microelectrode technology can often yield successful multi-unit recordings for widely varying durations of time before they fail. MEMS based technologies can move microelectrodes in rodent brains in long-term experiments resulting in improvements in signal quality. Further improvements in packaging and surgical techniques will potentially enable movable microelectrodes to record cortical neuronal activity in chronic experiments.

Movable microelectrodes have the potential capability of moving past the glial sheath surrounding the microelectrodes to restore the electrical interface with functional neurons in long-term experiments. Further, in short-term experiments, the microelectrodes can be moved to counter the variabilities in signal amplitudes due to brain motion caused either by animal behavior, or other endogenous factors. They also have the potential ability to track neurons or groups of neurons, and can be useful in locating specific targets in the brain. In addition, in the event of a failure in electrical recordings due to any number of reasons indicated earlier, one could potentially move the

microelectrodes to a new location and re-establish electrical connectivity with another neuron located in the same vertical axis as the neuron before.

The first generation of movable microelectrodes used manually movable microwire electrodes. However, the manual movement of these microelectrodes requires constant supervision making it cumbersome. Subsequently, the first motorized movable microelectrodes used DC motors to move the microelectrodes. More recently, there have been several attempts at using piezoelectric motors (Cham et al., 2005 ; Park et al., 2008), servomotors (Yamamoto and Wilson, 2008), and hydraulic motors (Sato et al., 2007) to move microelectrodes in the brain. However, the size and weight of these devices are often large. There is therefore a need to reduce the size and weight of movable microelectrodes while automating their ability to move within the brain after implantation in order to produce a potential chronic neural interface. A MEMS based movable microelectrode device can allow for batch fabrication, integration with advanced signal conditioning circuitry such as signal processors and wireless transmission and ensure the integrity of the interconnects.

MEMS based movable microelectrode arrays use electro-thermal actuators to allow bi-directional movement post-implant by applying low-voltage square pulses. Using the SUMMiT V™ (Sandia Ultra-planar Multi-level MEMS Technology 5, with 5-layers of polysilicon) process, three polysilicon microelectrodes, whose movements are controlled by electro-thermal actuators were surface micromachined on each chip. The overall dimensions of each of the microelectrodes are 50 μm × 4 μm × 5 mm. The microelectrodes are spaced by ~800 μm within the chip. The overall dimensions of the entire chip are 3 mm × 600 μm × 6 mm. Micrographs and SEM images of the MEMS movable microelectrode devices are shown in Figure 1.

Figure . (a) Micrograph of the movable microelectrode chip wire bonded on the chip carrier showing the three microelectrodes extended out at the bottom of the chip. The picture on the right shows a closer view of the bottom of the microchip with the microelectrodes and microactuators. (b) Schematic of a single polysilicon microelectrode and the associated thermal-microactuators/electro-thermal strips (chip is 3-mm wide and 6-mm long). (c) SEM of the middle microelectrode (50-μm wide) and the associated electro-thermal strips, and ratchets that move the microelectrodes. The device is externally controlled by application of voltage pulses via the labeled bond pads.

The principle of the electro-thermal actuators is illustrated in Figure 2 . Each of the microelectrodes has six bond pads connected to an external OmneticsTM connector, by which the microelectrode can be controllably moved and used to record multi-unit activity. Each microelectrode is controlled by four microactuators (Figure 2 A) – one each to deactivate the release-up lock and release-down lock, and one each to move the microelectrode up and down. Briefly, the V-beam microactuators work by using electro-thermal strips, where the applied low-

voltage pulse (7–10V) causes an increase in temperature and subsequent thermal expansion of the strips. Before the microelectrodes are moved up or down, the corresponding locks that prevent the microelectrode from moving are activated to release the microelectrode for motion (Figure 2 B). The microactuators are coupled to a ratcheting system that drives the center shuttle either down or up (Figure 2 C), depending on whether the move-up or the move-down microactuator is being activated respectively. As the applied voltage pulse is reduced to zero, the electro-thermal strips rapidly cool down and return to their normal position. During the return to their normal deactivated position (Figure 2 D), the center shuttle grabs the next adjacent teeth on the microelectrode and pulls it up or down depending on whether the move-up or the move-down microactuator is activated. Electro-thermal actuation mechanism allows for bi-directional movement with approximately 9 μm resolution in movement (equivalent to the spacing of the teeth on the side of the microelectrode). The aim of this study is to validate the array of movable microelectrodes in long-term rodent experiments and test if movable microelectrodes can improve the quality and reliability of neuronal recordings from the cortical brain region of rats.
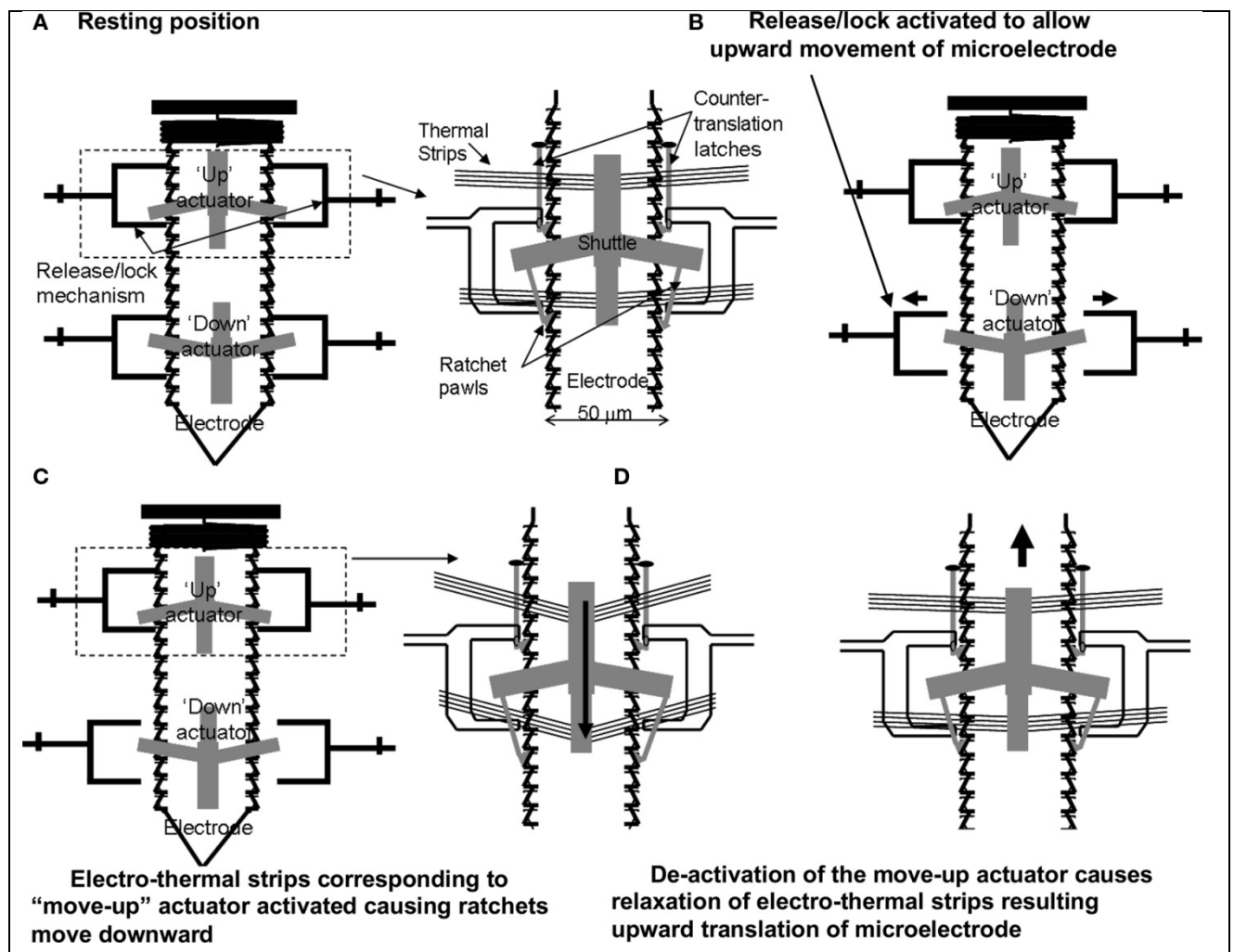
Figure 2. A schematic showing the principle of the electro-thermal actuators and how they move a microelectrode bi-directionally. (A) Resting position. (B) Release/lock activated to allow upward movement of microelectrode. (C) Electro-thermal strips corresponding to "move-up" actuator activated causing ratchets to move downward. (D) De-activation of the move-up actuator causes

relaxation of electro-thermal strips resulting upward translation of microelectrode.

Reference list

Design Considerations for a Micromachined Proportional Control Valve

APPLICATION OF THERMAL ACTUATOR

THERMAL BUCKLING ACTUATOR FOR MICRO RELAYS

Long-term neural recordings using MEMS based movable microelectrodes in the brain

9.   MEMS pressure sensors

9.1 Introduction

9.2 Pressure Sensor Types and Classification

      9.2.1   Absolute Pressure Sensors

      9.2.2   Gauge Pressure sensors

      9.2.3   Differential Pressure Sensors

9.3 Construction of MEMS presssure sensors

      9.3.1   Strain gauges and Piezoresistive pressure sensors

      9.3.2   Capacitive Pressure Sensors

      9.3.3   Piezoelectric pressure sensors

9.4 Performance Characteristics

9.5 Application

      9.5.1   Air flow monitors

      9.5.2   Aircraft and avionic systems

      9.5.3   Altimeters and barometers

      9.5.4   Anesthesia

      9.5.5   Appliances and consumer electronics

      9.5.6   Automobiles and trucks

      9.5.7   Automotive tire pressure

      9.5.8   Biomedical instruments

      9.5.9   Catheter pressure

      9.5.10 Disposable blood pressure

      9.5.11 Hydraulic systems and valves

      9.5.12 Industrial automation

      9.5.13 Medical equipment

      9.5.14 Pneumatic controls

      9.5.15 Portable gauges and manometers

      9.5.16 Pressure switches and controllers

      9.5.17 Process control

      9.5.18 Process control systems

      9.5.19 Refrigeration and HVAC controls

      9.5.20 Respirator monitoring

      9.5.21 Respiratory applications

      9.5.22 Ship and marine systems

      9.5.23 Sleep apnea

      9.5.24 Underground cable leak detection

9.5.25 Ventilation

9.1 Introduction

Pressure sensors in their primitive form existed as strain gauges for over several decades. The miniaturization of pressure sensors and other mechanical sensors gained considerable attention soon after the invention of piezoresistivity in silicon and germanium [Smith, 1954]. This activity gained further momentum with the recognition of the excellent mechanical properties of silicon [Peterson,1982]

The advent of the micromachining of silicon to carve out mechanical microstructures in silicon and the already existing expertise in manufacturing microelectronic devices and integrated circuits in silicon, opened the doors of the highly interdisciplinary area of MEMS and microsystems. During the past two decades, several industries and academic institutions all over the globe have been involved in the development and commercialization of micro sensors for industrial, automobile, defense, space and biomedical applications. Among the various devices, pressure sensors using MEMS technology have received great attention because the pressure sensors find applications in everyday life involving sensing, monitoring and controlling pressure, and they therefore constitute 60 to 70 percent of the market amongst the various MEMS devices.

As the requirements widen, new challenges emerge because the pressures range from a few Pascal (Pa) to several Mega Pascal (MPa) depending on the application and the environment, which vary from being very sensitive in biomedical applications to being very harsh in industrial and automobile applications. Hence, they need to be biocompatible in some applications while they need to be rugged and capable of performing reliably in temperatures well in excess of 80°C to 100°C.They also need to survive in corrosive fluids like ocean water in applications such as Oceanography. In several applications, such as mapping the pressure on the aero foil of an aircraft, the package needs to be flat and the device height needs to be restricted to below a millimeter. Similarly, in biomedical applications such as an intra cranial pressure (ICP) sensor, where the sensor is inserted into the ventricle, the packaged size should not exceed a diameter of 1mm.

As a result of these several constraints, the packaging of pressure sensors is not a universal technique, and has to be tailor-made depending upon the application. These restrictions on the size of the finished device also impose tremendous constraints on the pressure sensor chip design and fabrication. The demand for pressure sensors over a wide range of pressures varying from a few Pa to several MPa, operation capability at temperatures from -25 0C to +125C for aerospace applications and the growing demand for pressure measurements in high-temperature environments (>500 °C) have spurred the development of robust, reliable MEMS-based pressure sensor technologies involving silicon, Silicon on Insulator (SOI), Silicon on Sapphire (SOS), Silicon

Carbide (SiC) and Carbon Nanotubes (CNT). As a result, the constraints and the capability requirements on the ruggedness of the microsystem packaging technology have been tremendous.

This paper aims at providing an overall scenario of pressure sensor technology, beginning from basic principles. This is followed by design criteria for different ranges of pressures and the related technology, based on silicon as the diaphragm material and piezoresistor, SOI and polysilicon piezoresistors for higher temperatures and high pressures followed by more exotic materials like SiC and CNT for capacitive and piezoresistive pressure sensors. The paper elucidates the use of bossed structures, which result in sculptured diaphragms for low-pressure applications with greater linearity. The paper also gives a glimpse of some of our work which has been supported by the NPMASS program at the Centre for Nano Science and Engineering (CeNSE) at IISc and includes a section on the challenges of reliable pressure sensor packaging technology for operation in harsh corrosive environments.

9.2 Pressure Sensor Types and Classification

Pressure sensors are categorized as absolute, gauge and differential pressure sensors based on the reference pressure with respect to which the measurement is carried out. Particular applications are as follows.

9.2.1 Absolute Pressure Sensors

2Absolute pressure sensors measure the pressure relative to a reference vacuum encapsulated within the sensor as shown in Figure 9.1

Such devices are used for atmospheric pressure measurement and as manifold absolute pressure (MAP) sensors for automobile ignition and airflow control systems. Pressure sensors used for cabin pressure control, launch vehicles, and satellites also belong to this category
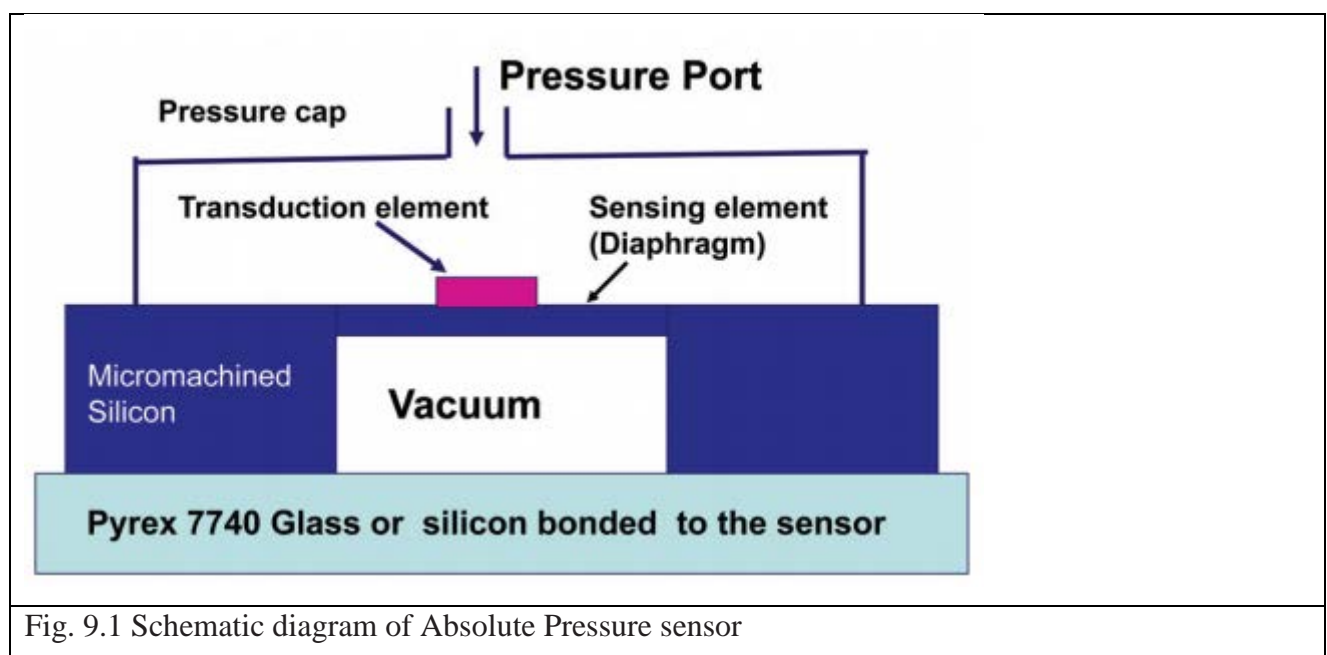
Fig. 9.1 Schematic diagram of Absolute Pressure sensor

9.2.2 Gauge Pressure sensors

Gauge pressure sensors measure pressure relative to atmospheric pressure. One side of the diaphragm is vented to atmospheric pressure as shown in Figure 2. Blood pressure (BP), intra-cranial pressure (ICP), gas cylinder pressure and most of ground-based pressure measurements are gauge pressure sensors. Vacuum sensors are gauge sensors designed to operate in the negative pressure region

Figure 9.2 Schematic diagram of Gauge Pressure sensor
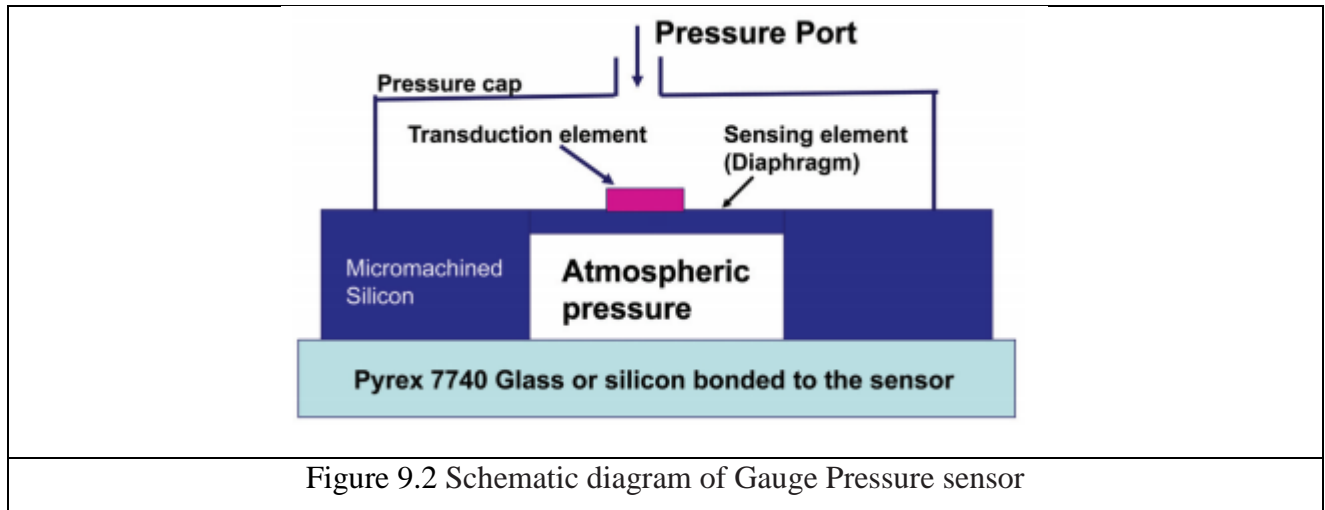
9.2.3 Differential Pressure Sensors

Differential Pressure Sensors measure accurately the difference ΔP between two pressures P1 and P2 across the diaphragm (with AP << P1 or P2), and hence need two pressure ports as shown in Figure 3. They find applications in airplanes used in warfare. They are also used in high pressure oxidation systems where it is required to maintain an oxygen pressure ranging from 1 to 10 atmospheres inside a quartz tube during the oxidation of silicon. In this system, the outside of the quartz tube is maintained at a slightly higher gas pressure of nitrogen, and the pressure difference is monitored using a differential pressure sensor which ensures that the quartz tube does not experience a differential pressure greater than its rupture stress of 1 atmosphere (10 Pascal). The differential pressure sensor is also used in some applications where it is desirable to detect small differential pressures superimposed on large static pressures.

In almost all types of pressure sensors, the basic sensing element is the diaphragm, which deflects in response to the pressure. As the deflections in diaphragm-based sensors are small they cannot be directly measured. This mechanical deflection or the resulting strain in the diaphragm is converted ultimately into electrical signals using suitable transduction mechanisms, namely, capacitive, piezoresistive or piezo-electric techniques, which are usually employed as adjectives for the pressure sensors as described below: directly on the silicon diaphragm by implanting or diffusing boron in the selected regions of maximum stress as shown in Figure 4(a). These resistors are connected in the form of a Whetstone Bridge which gives an output when the resistors are strained under the action of the pressure sensed by the diaphragm. It will be seen in subsequent sections that piezoresistive pressure sensors enable linear operation over a wide range of pressures.

They are also simple to fabricate. As a result, they have captured the major market of pressure sensors encompassing the automobile industry, defense, space as well as biomedical applications.

Figure 9.3 Schematic diagram of Differential Pressure sensors
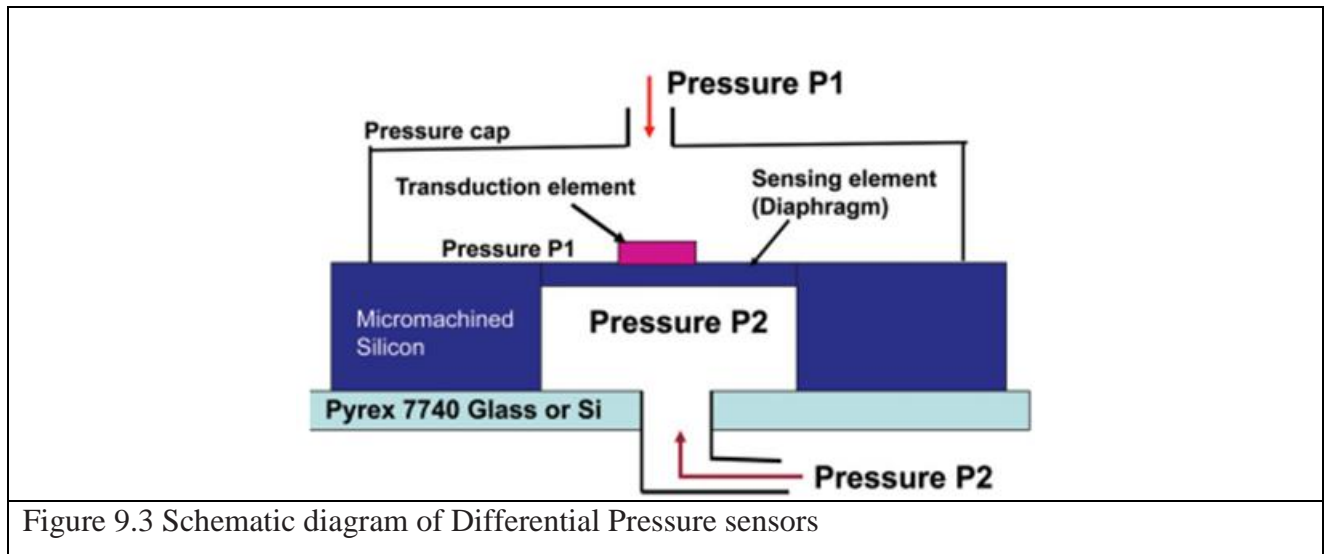
Other transduction techniques are capacitance and piezoelectric approaches and they are identified as capacitive pressure sensor and piezoelectric pressure sensors respectively.

Capacitive Pressure Sensor: A schematic diagram of a silicon micro machined sensor of this type is shown in Figure 4(b).This approach uses the diaphragm as one electrode of a parallel plate capacitor structure and diaphragm displacement causes a change in capacitance with respect to a fixed electrode. The merits of capacitive pressure sensors are their high sensitivity, which is practically invariant with temperature. However, in this case, an electronic circuit is required to convert the achieved because the piezo-resistors are embedded capacitance change into an electrical output. An additional disadvantage of this approach is the nonlinear relationship between the capacitance and displacement and hence a force-balancing and linearizing electronic circuit is essential to capture a wide range of pressures.

Piezoelectric pressure sensors: Silicon does not show a piezoelectric effect. Therefore, a piezoelectric sensing element, such as Lead Zirconate Titanate (PZT) or Zinc Oxide (ZnO) are placed/deposited on to the silicon diaphragm as shown in Figure 4(c). The deflection of the diaphragm induces strain in the piezoelectric material and hence a charge is generated. These sensors are only suitable for measuring dynamic pressures and are not suitable for static pressure sensing because piezoelectric materials only respond to changing strains. The major advantage of this approach is that an external power supply is not required.

In another approach, in which pressure sensors are identified as resonant sensors, the vibration frequency of a mechanical beam or a membrane, which depends on the extent to which it is stretched, is used This is similar to the vibration frequency of a violin string. The output signal from a resonant sensor is a frequency, which can easily be transferred into a digital signal and interfaced with computer systems, without having to use an Analog to Digital converter. In most of

the cases, a quartz resonating beam is used because these sensors are noted for their high stability and high resolution as a frequency signal is much more robust than an amplitude (e.g. a voltage). The stability is determined only by the mechanical properties of the resonator material, which is generally very stable. On the flip side, resonant silicon sensors are not easy to fabricate and hence become expensive. The high costs may be compensated by innovative simpler mechanical structures and by simpler electronics.

Figure 9.4 Schematic representation of micrmachined silicon pressure sensors (a) Piezoresistive (b) capacitive (c) Piezoelectric
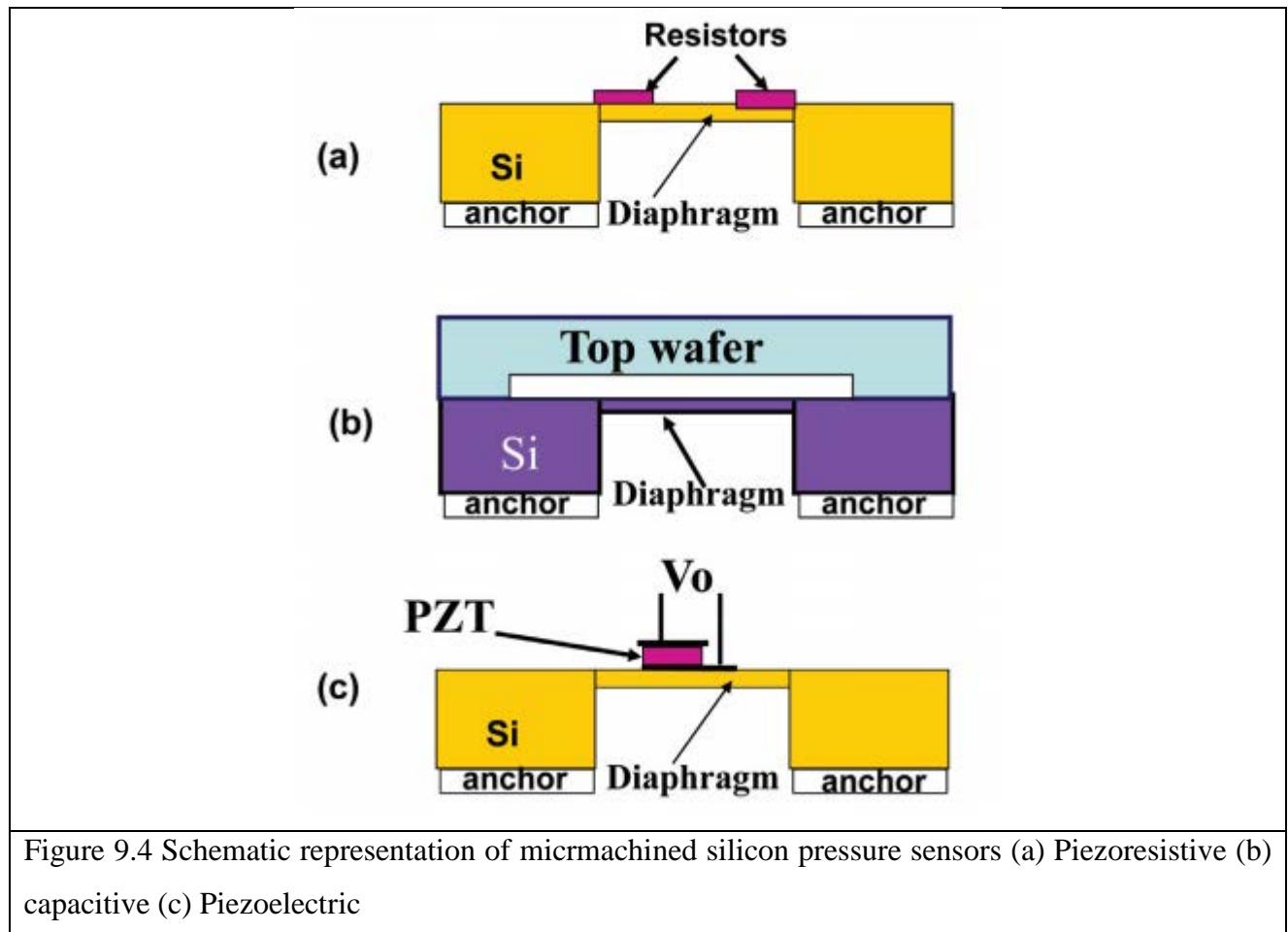
In all the types of pressure sensors, the diaphragm is invariably, if not always, chosen as the sensing element. The static and dynamic performance characteristics of the diaphragm are of great importance and provide suitable design guidelines for designing pressure sensors. These aspects of micromachined diaphragms are discussed in the following section.

Silicon Carbide Pressure Sensors for Harsh Environments

Silicon Carbide Piezoresistive Pressure Sensors

The applications of conventional piezoresistive silicon pressure sensors are limited to below 150°C due to the degradation of the PN junction isolation between the resistors on the diaphragm. Silicon On Insulator, along with oxide isolated piezoresistive elements, has enabled their use up to about 300°C. However, the need for pressure measurements in high-temperature (>600°C) and harsh environments (automotive and aerospace applications: combustion processes or gas turbine control; oil industry; industrial process control; nuclear power) has spurred the development of

robust, reliable pressure sensors. This is mainly because of the poor thermomechanical properties of silicon, *viz* the degradation of the elastic modulus above 600°C, and its inability to withstand a corrosive environment. As a result, during the past decade, considerable effort has been directed towards taking advantage of the superior thermo mechanical properties of silicon carbide (SiC) to develop micro-pressure sensors that would extend the sensing capability to 600°C and beyond. The wide band gap of 3.3eV 6H-SiC enables high temperature operation of the junctions. The fairly inert nature of the SiC makes it suitable for use in corrosive environments, and the electronic circuits on SiC are radiation-insensitive. As already presented in Table-1, the gauge factor of single crystal SiC is in the range 10 to 30 and in polycrystalline SiC, it is in the range of 3 to 5. The properties of Single crystal SiC are presented in Table-3, and are compared with Silicon and diamond to show that the yield strength , Young's modulus and hardness of SiC are considerably higher than those of Silicon and are closer to those of diamond.

Table- Some important properties of single crystal Si, SiC and diamond [Peterson, 1982]

Property	Yield strength (GPa)	Hardness (Kg/mm^2)	Young's modulus (GPa)	Melting point (°C)	Gauge factor
Si	7	850	190	1410	100-200
6H-SiC	21	2480	700	2830 (sublimes)	10-30
Diamond	53	7000	1035	4000 (Phase change)	—

In spite of its several merits, the development of SiC pressure sensors has been slow because of the difficulties involved in its micromachining. Initially, SiC piezoresistors were fabricated on oxide grown on a silicon micromachined diaphragm. In this approach, silicon becomes the limiting factor. The benefits of SiC material can be fully utilized by fabricating both the piezoresistive sensing elements and the sensing diaphragm on SiC. This has, indeed, been made possible, as reported in [Ned et al., 2004], by micromachining N-type SiC in the DRIE system in a gas plasma mixture containing SF_6 and Oxygen to realize the SiC diaphragm. SiC piezoresistors were realized on the SiC diaphragm by chemical vapor deposition of SiC using a gas mixture of SiH_4 and CH_4 and patterning it by RIE. Ohmic contacts on the SiC resistors were obtained with a Ti/Al bi-layer having 125nm titanium thickness and 1200nm of Aluminum. Both low and high pressure SiC piezoresistive pressure sensors operational up to 600°C have been demonstrated [Ned et al., 2004], using relatively thick (60 pm) diaphragms for the 1000 psi sensors, while the diaphragm for the 25 psi pressure sensor was micromachined to be significantly thinner with optimized sculptured sensing diaphragms of the type discussed in section-8. Although the performance characteristics such as sensitivity, linearity and hysteresis were reasonably good, a couple of issues still needed to

be sorted out (1) A large thermal set (10mv) was observed in all SiC sensors after exposure to 600°C. (2) The bridge resistance versus temperature showed a non-monotonic resistance change with temperature, giving rise to a monotonic reduction in resistance up to 350 °C followed by an increase in resistance. This continues to pose a big challenge for the temperature compensation of such SiC sensors.

An exploratory effort [Vandelli, 2008] has shown that capacitive SiC MEMS devices fabricated and characterized showed a repeatable hysteresis less than 2.5% at 300°C. It was also demonstrated that the combined effect of temperature and pressure on non-linearity and hysteresis was never much higher than 1% FSO over an operating temperature range of 0 to 200°C.

According to a SiC road map [Fraga et al., 2010], the development of SiC sensors is based on progress in (1) the improved electrical and mechanical properties of SiC films produced by the optimization of the SiC deposition process, (2) SiC film processing by the optimization of the etching process and metallization appropriate for high temperature applications, (3) micro-fabrication technology to fabricate miniaturized sensors and (4) sensor packaging for harsh environments. The immediate goal is to improve the performance of SiC pressure sensors and strain gauges by optimizing all the four technologies above.

SiC -Capacitive Pressure Sensors for High Temperatures > 500°C

Capacitive pressure sensors are attractive for high-temperature applications because their performance is almost independent of temperature, and high sensitivity can be achieved with a low turnon temperature drift. These devices are also less sensitive to side stress and environment variations. Capacitive sensors are best suited for low pressure and high temperature sensing applications ranging from low temperatures up to 500 °C and beyond with SiC diaphragms. A schematic structure of one such device [Du et al., 2003] fabricated with a 0.5 im single crystal 3C-SiC diaphragm on a single crystal <100> silicon is shown in Figure 23. A thin layer of phosphorous silicate glass (PSG) on the surface of the DRIE-etched silicon substrate provides electrical isolation between the SiC diaphragm and the DRIE-etched silicon substrate. When the diaphragm is subjected to pressure, P as shown in Figure XX(a), the capacitance increases nonlinearly due to its downward deflection till it touches the substrate at a designed touch pressure point (TPP). At this pressure, the diaphragm touches only at its middle.

SiC capacitive Pressure Sensor (a) the schematic structure (b) typical Capacitance, C, versus Pressure, P, characteristics showing the Touch Point Pressure (TPP)
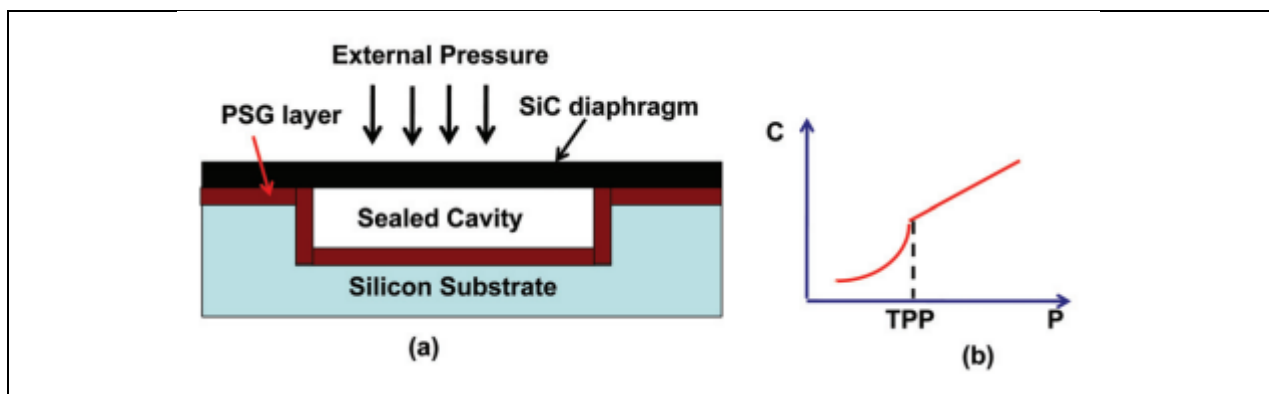
Beyond the TPP, the touching area increases and the capacitance increases linearly as shown in Figure XX(b). With the capacitance structure having a circular SiC diaphragm of 400pm, thickness 0.5 pm suspended over a 2 pm cavity in the silicon substrate, high sensing capability up to 400 °C has been reported in the literature [Du et al., 2003] with excellent linear characteristics in the pressure range between 1100 Torr and 1760 Torr with a sensitivity of 7.7fF/Torr.

These results have indeed demonstrated the benefits of using SiC diaphragms for high temperature operations and a capacitive sensing approach for low-pressure operations and this has opened up new avenues for future work for sensors using SiC. The development of high temperature

SiC electronics with SiC pressure sensors and SiC temperature sensors for intelligent engine systems is being pursued in academic institutions and industries. Semiconductor integrated circuit chips capable of giving over 100-fold improvement at 500 °C operational durability have been reported using Silicon Carbide transistors and logic circuits [Mohamed et al., 2010]

Carbon Nano Tube (CNT) Based Sensors

In recent years CNTs-based pressure sensors have drawn considerable attention due to their high sensitivity, small size, low power consumption, and strong mechanical stability, in addition to thermal stability. The CNT-based pressure sensor can be operated up to 250 °C as against the 125°C of conventional PN junction-isolated silicon pressure sensors. The high gauge factors of about 1000 and the small size(diameter 1 to 100nm) of CNTs make them ideal for miniaturized pressure sensors for biomedical applications as well. The piezoresistive effect in the aligned multi-walled carbon nanotube (CNT) array arises due to the buckling of individual carbon nanotubes when compressed, leading to a decrease in the electrical resistance of the CNT array. Recently, it has been experimentally proved by mechanical loading of the array that the decrease in resistance is almost fully recoverable once the load is removed [Mohamed et al., 2010]. In the same paper, it has been shown that the change in the resistance increases linearly when 100 g and 500 g load are applied at a temperature of 20°C to 180°C, thus demonstrating that the multi-walled carbon nanotube (MWCNT) array is an excellent pressure and strain sensing element capable of operation at elevated temperatures.

A novel polymer-based MEMS pressure sensor has been reported [Fung et al., 2005] using bulk MWNT as the piezoresistive sensing elements. The development of the pressure sensor includes the fabrication of 300 thick Polymethylmethacrylate (PMMA) diaphragms using an SU8 molding/hot- embossing technique and AC electrophoretic manipulation of MWNT bundles on the

diaphragms. The advantage of MWNT bundles is the ease with which they can be formed across electrodes on top of pressure diaphragms to serve as sensing elements. A schematic structure of the pressure sensor reported in [Wong et al., 2000] is shown in Figure XXX. It has been shown that the diaphragm deflects due to an applied pressure and the MWNT piezoresistor gets strained, as a result, causing a change of resistance of 4 kW at a pressure of 70kPa from an initial value of 153kW at zero pressure. The CNT, as a sensing element on diaphragms using various materials is of great interest, and considerable effort has been expended over the past decade with varied degrees of success.

Schematic diagram of MWCNT based pressure sensor reported in [Mohamed et al., 2010]
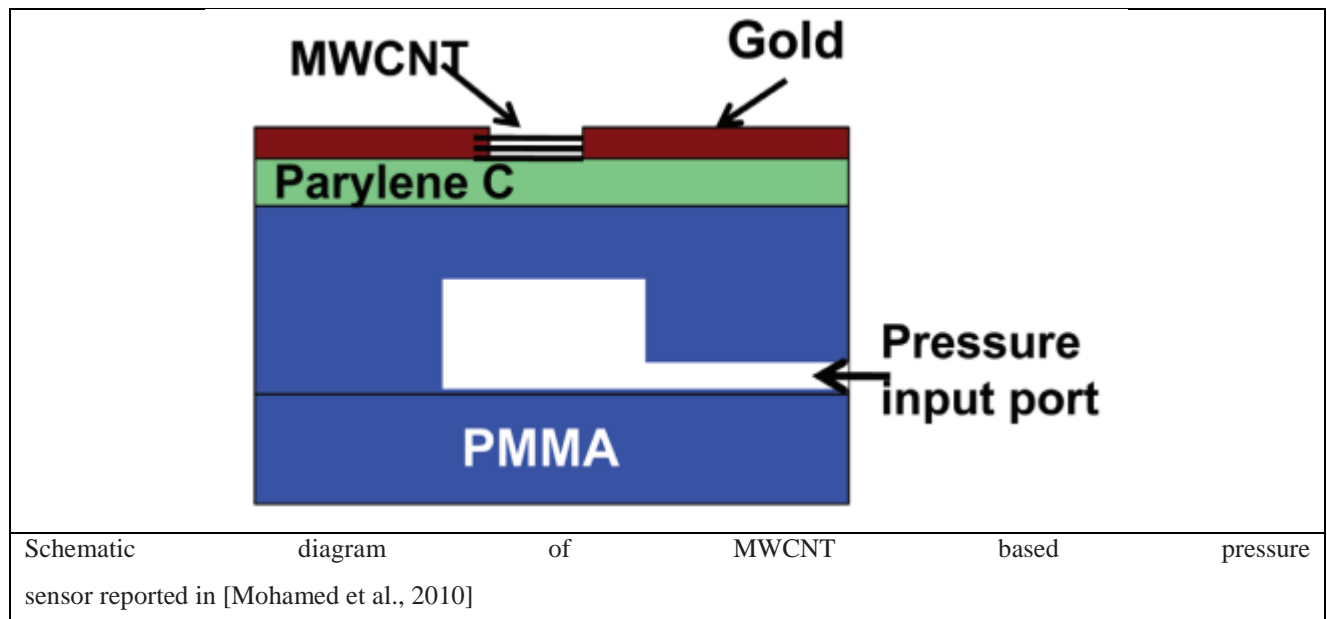
Packaging Technology

The requirements and constraints imposed on MEMS pressure sensor/transducer packages are briefly presented here. The associated challenges and techniques used for realizing microsystem packages are presented with examples drawn from various types (Absolute, Gauge or Relative and Differential) of pressure sensors.

Micromachined (MEMS) pressure sensors play an important role in present day microsystems. These microsystems are also required to accommodate electronics to harness raw signals into acceptable levels and also to make the signal insensitive to the effects of signals which are outside the domain of the system. Thus for instance, if a microsystem is realized for sensing and monitoring pressure, it should be able to track the effects of temperature and correct the output to compensate for the effect of temperature. The micromachined components of these microsystems are fragile structures and they are invariably exposed to harsh aggressive environments such as the corrosive water of the ocean which comes into contact with the pressure sensing microsystem used in oceanography. As a result, the packaging technology for microsystems plays a very important role in protecting these micro devices from the shock, the chemical environment and temperature extremities. In some biomedical applications such as intracranial pressure (ICP) and blood pressure

monitoring sensors, the package should invariably be biocompatible. In aerospace applications which map the pressure in the aero foil, the package needs to be flat and should not exceed 1mm in height. The packaging technology for microsystems is thus more involved and varied than the packaging technology for integrated circuit chips. Here, we discuss the techniques used for packaging microsystems considering specific case of micromachined silicon piezoresistive pressure sensors.

The Importance and Challenges

MEMS pressure sensor packaging includes three major tasks, namely, (i) Assembly, (ii) Packaging and (iii) Testing, abbreviated as AP&T. The AP&T of microsystems contributes to a significant portion of the overall cost of production and could vary from 20% of the overall production cost with plastic passivation designed for a friendly environment in mass production, to as high as 95% of total cost for special pressure sensors for high temperature application with toxic-pressurizing media. Thus the AP&T cost for microsystems and especially for MEMS pressure sensors/ transducers varies from one product to another. Currently, this cost represents, on an average, 80% of the total production cost. The source of the failure of microsystems can be traced, in most cases, to inadequate packaging. As a result, microsystem packaging technology has become a key factor in MEMS product design and development, and it has attracted great interest and the attention of the microsystem community all over the world.

Packaging the micromachined component of the MEMS pressure transducer should be designed to meet the demands and requirements of the end user, whether those requirements be generic or specific. It consists of a complex matrix of solutions and is a multi disciplinary field, irrespective of whether it is an application driven or performance driven requirement. The reliable packaging of these devices and systems is a major challenge to the industry because microsystem of packaging has not reached the maturity level of its counterpart, the microelectronics packaging, has reached.

The packaging of the MEMS pressure transducer is required to protect the sensing or actuating elements when the sensing device is in contact with the medium, which, in itself, is the source of action, as in a gas sensor. Many media are hostile to these elements. The pressure sensor/ transducer packaging is highly challenging because the core devices of microsystems, such as microsensors usually involve delicate, complex threedimensional geometry made of layers of dissimilar materials [Gardener, 1994; Malshe et al., 1999].

Functional Requirements

The mechanical package and its associated manufacturing technologies must fulfill functions like mechanical protection for day-to-day use in the service intended, and media protection when operated in harsh environments such as humidity, salt water, body fluids, fuels and gases etc. It should incorporate a convenient low cost means to interface and to isolate a very sensitive silicon die from undesirable mechanical stresses. The sensor package must be carefully designed to isolate

the stress and reject extraneous mechanical straining inputs from any other variables such as acceleration, vibration etc. Simultaneously the package should allow the efficient transfer of the mechanical variable of desired inputs i.e., pressure in this case. Silicon, which is used for fabricating micro sensors and actuators, is a relatively low-expansion material while most packaging materials such as metals and ceramic exhibit considerably higher expansion. Hence, the package must allow for an appropriate means of reducing the undesirable effects of thermally-created stresses due to expansion or contraction.

An electrical interface is to be achieved by depositing and patterning aluminum or a noble metal thin film layer on the die. Modern silicon sensors often incorporate calibration and compensation either by using them on the silicon die itself or by employing hybrid components / signal conditioning ICs / resistors. The package must provide an appropriate means of incorporating and protecting these additional components. Hence the package provided must be suitable for housing and protecting the associated signal conditioning electronics.

In addition to the functional requirements that the package must accomplish, a suitable packaging scheme must have low cost, reliability, compatibility with subsequent assembly techniques and high volume production capability.

7.1 Microsystem Packaging Techniques

The packaging techniques used in silicon sensors originate from two distinct roots. One root is its adaptation from the well-established IC industry. The second major root of sensor packaging technology is derived from the conventional mechanical technology of aerospace and other process control industries. This excellent packaging technology involves the more traditional, reliable and well-established mechanical arts like machining (Precision CNC turning, milling, grinding, drilling, honing, polishing, Lapping, electro discharge machining etc.,), welding (Electron Beam Welding, Laser welding, precision TIG/MIG welding, Electric welding), Brazing (Vacuum Brazing and micro brazing), Engraving, metal etching, electro plating, anodizing, Glass to metal sealing, casting of alloys etc. The modern silicon sensor package is a blend or amalgamation of these two arts. Hence, to produce a package which is rugged and reliable, the silicon sensor-packaging engineer seeks to employ the best from each art.

The Future and the Vision

The future is tending towards MEMS to Nano technology. Nano technology provides the ability to work at the molecular level, atom by atom, to create large structures with a fundamentally new molecular organization. The emerging approach for both chemical and biological MEMS sensors is based on carbon nanotube (CNT) and graphene-based technology. It is essentially concerned with materials, devices and systems to achieve structures and components which exhibit novel and significantly improved physical, chemical and biological properties, phenomena and processes due

to their nano-scale size. These sensors have better sensitivity, selectivity and stability than commercially available sensors.

Focused R & D, with more attention and effort in design and packaging is needed in India to deliver MEMS-based devices to commercial as well as strategic programs. Effective collaborations between successful organizations may dramatically shorten the time frame between the concept to the systems. The following areas require careful attention and efforts:

♦ Polymer-clay, Nano-micro composite material for load bearing structures.

♦ Composite materials with embedded MEMS sensors.

♦ Nano and micro sensors and devices for space and terrestrial applications.

♦ MEMS based vibration sensing and analysis system for structural health monitoring

♦ Sensors for spacecraft structural health monitoring.

♦ Silicon on sapphire (SOS) based miniaturized sensors for high temperature operation in upto 30 °C for operation in harsh environment

♦ Silicon carbide (SiC) sensors for operation at temperatures in excess of 500°C and for operation in automotive applications as well as for corrosive fluids.

Chapter 23

Atomic force microscopy (AFM)

Atomic force microscopy (AFM) is a technique for analyzing the surface of a rigid material all the way down to the level of the atom. AFM uses a mechanical probe to magnify surface features up to 100,000,000 times, and it produces 3-D images of the surface.

The technique is derived from a related technology, called scanning tunneling microscopy (STM). The difference is that AFM does not require the sample to conduct electricity, whereas STM does. AFM also works in regular room temperatures, while STM requires special temperature and other conditions.

AFM is being used to understand materials problems in many areas, including data storage, telecommunications, biomedicine, chemistry, and aerospace. In data storage, it is helping researchers to "force" a disk to have a higher capacity. Today's magnetic storage devices typically have a capacity limit of between 20 and 50 gigabits (billions of bits) per square inch of storage medium. Researchers are looking into AFM to help raise read and write densities to between 40 gigabits and 300 gigabits per square inch. No one has yet commercialized AFM technology for this purpose, but IBM and others are actively pursuing it.

Nano, from the Greek word for 'dwarf', corresponds to a prefix denoting a factor of 10^{-9}. Thus, a nanometer is one billionth of a meter, which is the length scale at which intermolecular force and quantum effect take hold. To put the nanoscale in a more understandable perspective, consider that the size of an atom relative to an apple is similar to the size of an apple relative to the planet Earth! Atomic Force Microscopes (AFMs) give us a window into this nanoscale world.
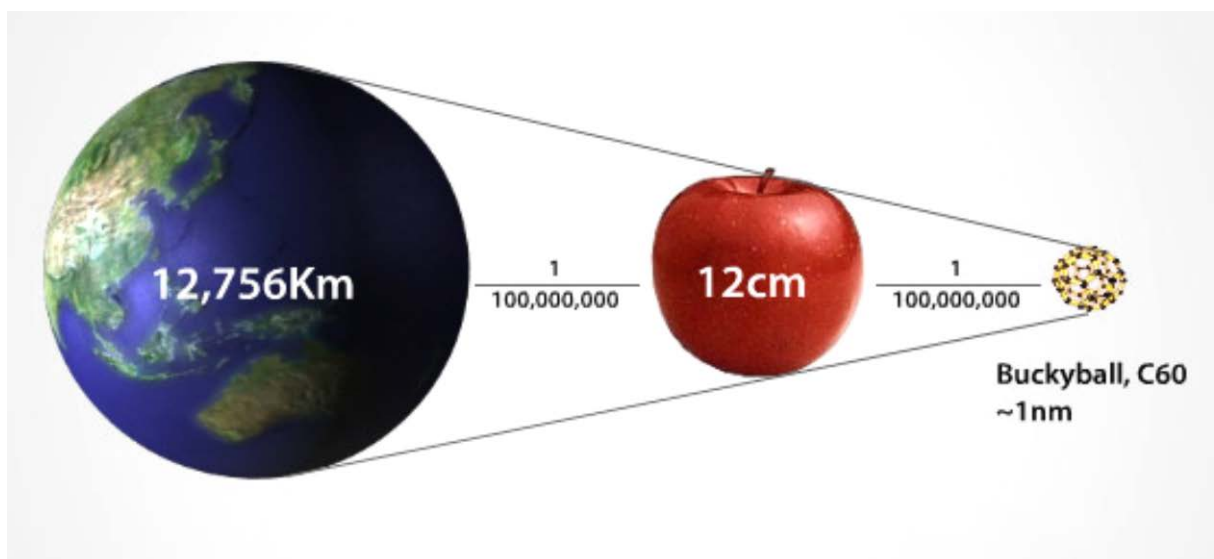
Fig. 1. Understanding of nano- scale.

An AFM uses a cantilever with a very sharp tip to scan over a sample surface. As the tip approaches the surface, the close-range, attractive force between the surface and the tip cause the cantilever to deflect towards the surface. However, as the cantilever is brought even closer to the surface, such that the tip makes contact with it, increasingly repulsive force takes over and causes the cantilever to deflect away from the surface.

At the right side of the curve the atoms are separated by a large distance. As the atoms are gradually brought together, they first weakly attract each other. This attraction increases until the atoms are so close together that their electron clouds begin to repel each other electrostatically. This electrostatic repulsion progressively weakens the attractive force as the inter-atomic separation continues to decrease. The force goes to zero when the distance between the atoms reaches a couple of angstroms, about the length of a chemical bond. When the total van der Waals force becomes positive (repulsive), the atoms are in contact.
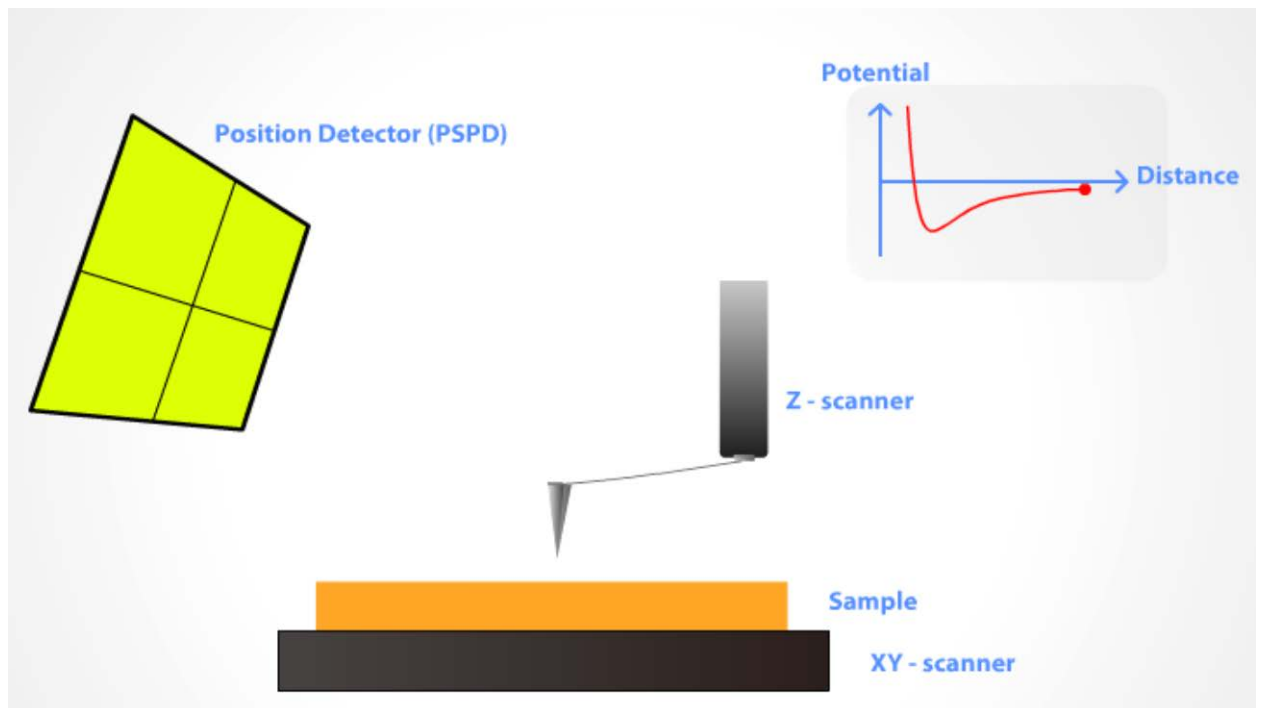
Fig. 2. Scheme of AFM.

The tip is attached to the end of a cantilever with a low spring constant, lower than the effective spring constant holding the atoms of the sample together. As the scanner gently traces the tip across the sample (or the sample under the tip), the contact force causes the cantilever to bend to accommodate changes in topography. A laser beam is used to detect cantilever deflections towards or away from the surface. By reflecting an incident beam off the flat top of the cantilever, any cantilever deflection will cause slight changes in the direction of the reflected

2

beam. A position-sensitive photo diode (PSPD) can be used to track these changes. Thus, if an AFM tip passes over a raised surface feature, the resulting cantilever deflection (and the subsequent change in direction of reflected beam) is recorded by the PSPD.
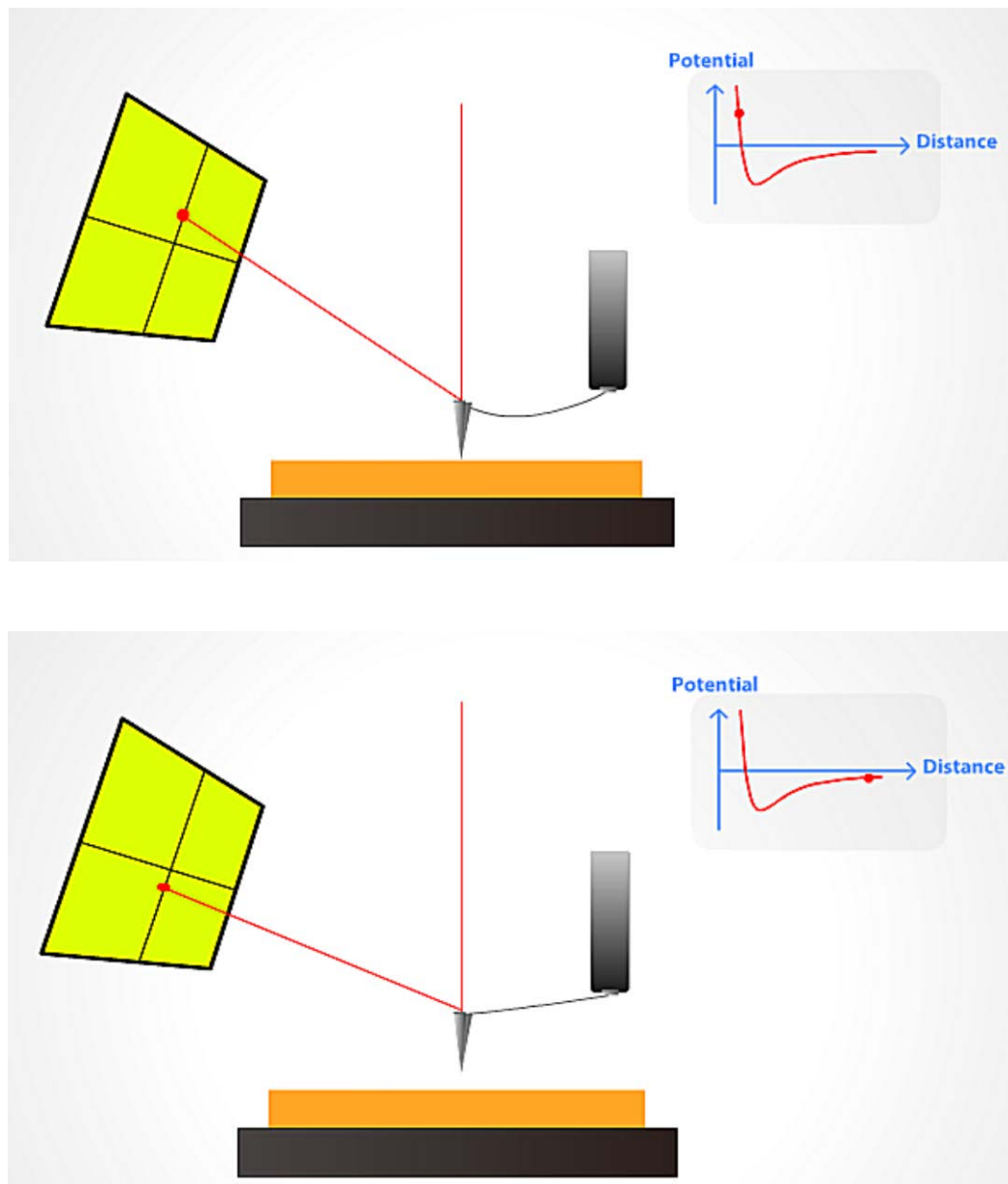
Fig. 3. AFM principles.

An AFM images the topography of a sample surface by scanning the cantilever over a region of interest. The raised and lowered features on the sample surface influence the deflection of the cantilever, which is monitored by the PSPD. By using a feedback loop to control the height of the tip above the surface—thus maintaining constant laser position—the AFM can generate an accurate topographic map of the surface features.
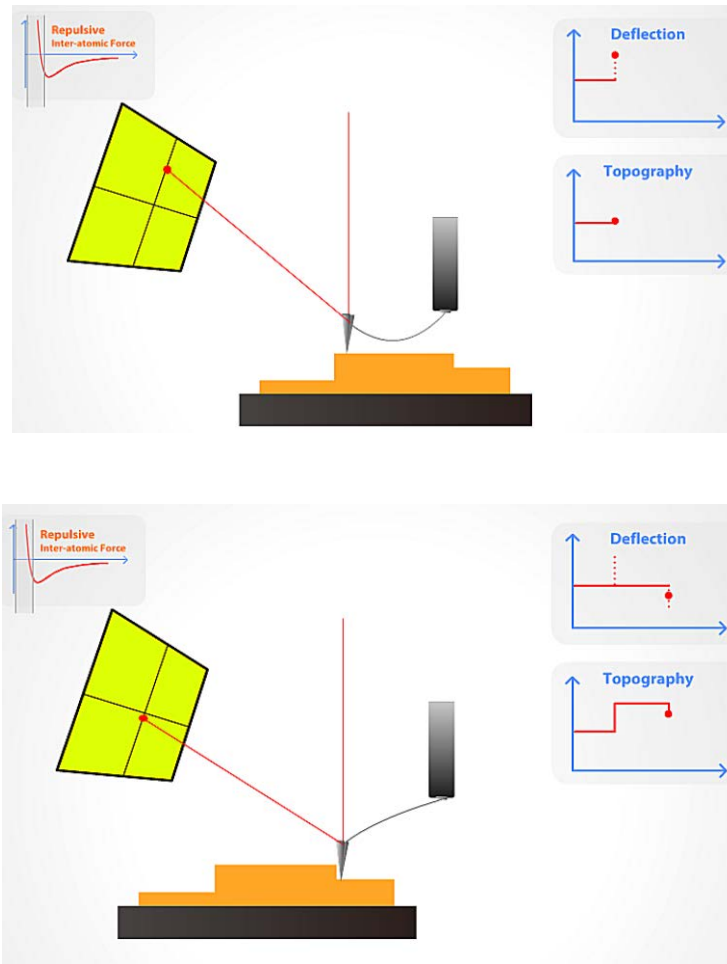
Fig. 4. Contact mode.

EFM-DC utilizes a lock-in amplifier to study the electrical properties and topography of a sample surface in a single scan, unlike the two scans required by standard EFM. Here, the cantilever is biased with an AC current different than the resonance of the cantilever. The oscillation component of the PSPD signal is extracted by the lock-in amplifier, resulting in the EFM signal.

Once the contact AFM of the XE-series has detected the cantilever deflection, it can generate the topographic data set by operating in one of two modes - constant-force or constant-height mode. In constant-force mode, the deflection of the cantilever can be used as input to a feedback circuit that moves the scanner up and down in Z, responding to the topography by keeping the cantilever deflection constant. In this case, the image is generated from the scanner's motion. With the cantilever deflection held constant, the total force applied to the sample is constant. In constant force mode, the speed of scanning is limited by the response time of the overall feedback circuit, but the total force exerted on the sample by the tip is well controlled. Constant-force mode is a default contact mode in the XE-series and preferred for most of applications.
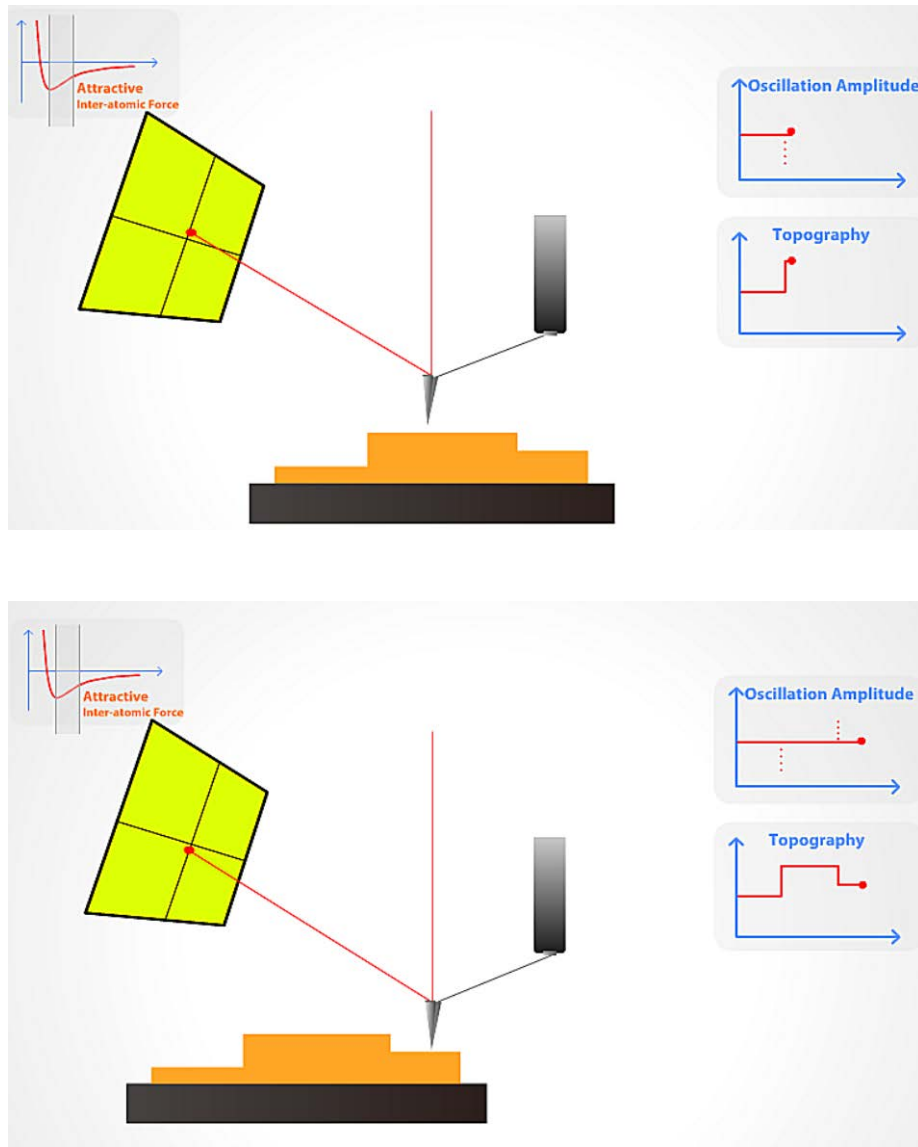
Fig. 5. Non-contact mode.

In this alternative technique to non-contact mode, the cantilever again oscillates just above the surface, but at a much higher amplitude of oscillation. The bigger oscillation makes the deflection signal large enough for the control circuit, and hence an easier control for topography feedback. It produces modest AFM results but blunts the tip's sharpness at a higher rate, ultimately speeding up the loss of its imaging resolution.

Constant-height mode is often used for taking atomic-scale images of atomically flat surfaces. Here, the Z-servo is turned off and a sample is scanned without Z-feedback. Instead, the cantilever deflection, error signal, is used directly to generate the topographic data set because the height of the scanner is fixed as it scans. The cantilever deflections and thus variations in applied force are small.

Constant-height mode is essential for recording real-time images of changing surfaces, where high scan speed is required.
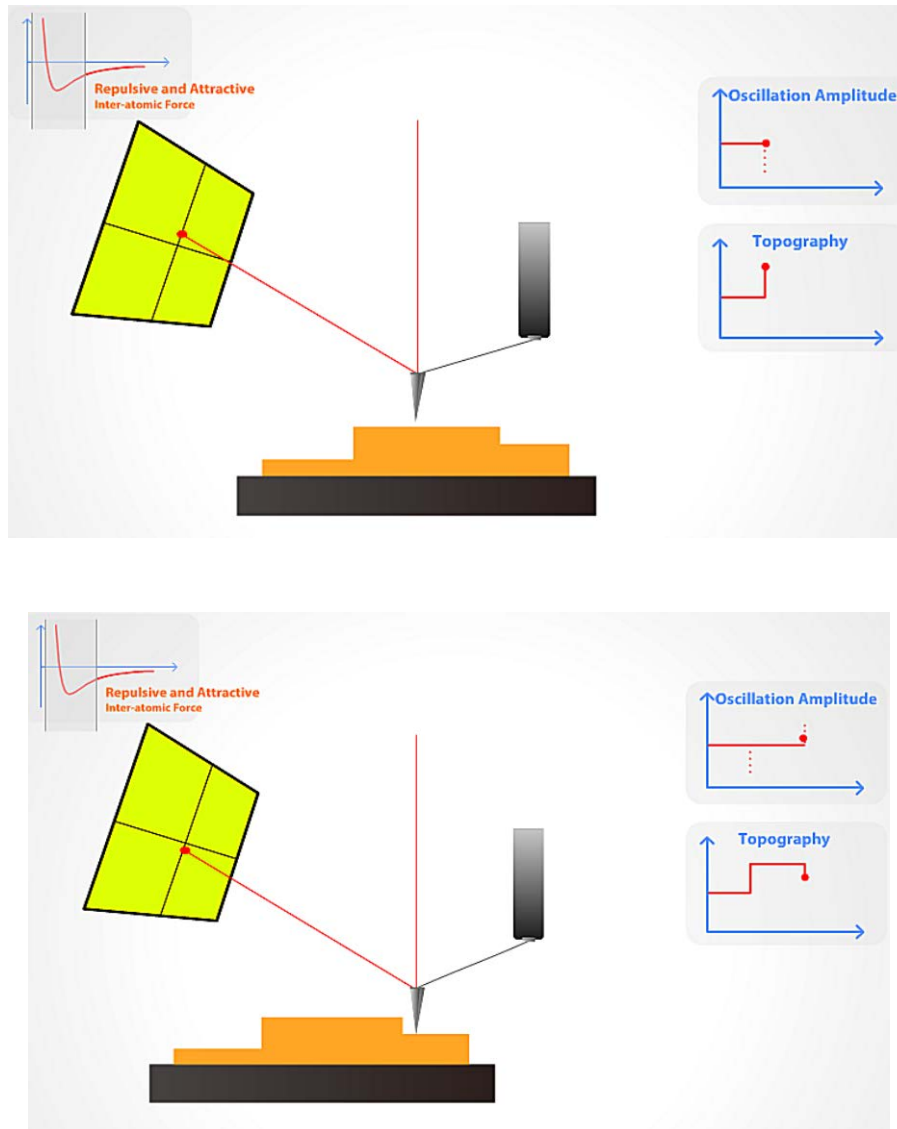
Pic. 6. Tapping mode.

In this technique, the cantilever oscillates just above the surface as it scans. A precise, high-speed feedback loop prevents the cantilever tip from crashing into the surface, keeping the tip sharp and leaving the surface untouched. As the tip approaches the sample surface, the oscillation amplitude of the cantilever decreases. By using the feedback loop to correct for these amplitude deviations, one can generate an image of the surface topography.

Dynamic Force Microscopy (DFM) of the XE-series is very similar to True Non-Contact mode of the XE-series in many ways such as the applied force and the measurement principle. DFM is a hybrid of the two most fundamental measurement methods, represented by contact mode and true non-contact mode.

In DFM of the XE-series, the cantilever vibrates in free-space in the vicinity of the resonant frequency like in non-contact mode, and at the same time, as the vibrating cantilever gets very close to the sample surface, it taps the surface repeatedly. The tip "contacts" the sample surface as in contact mode.
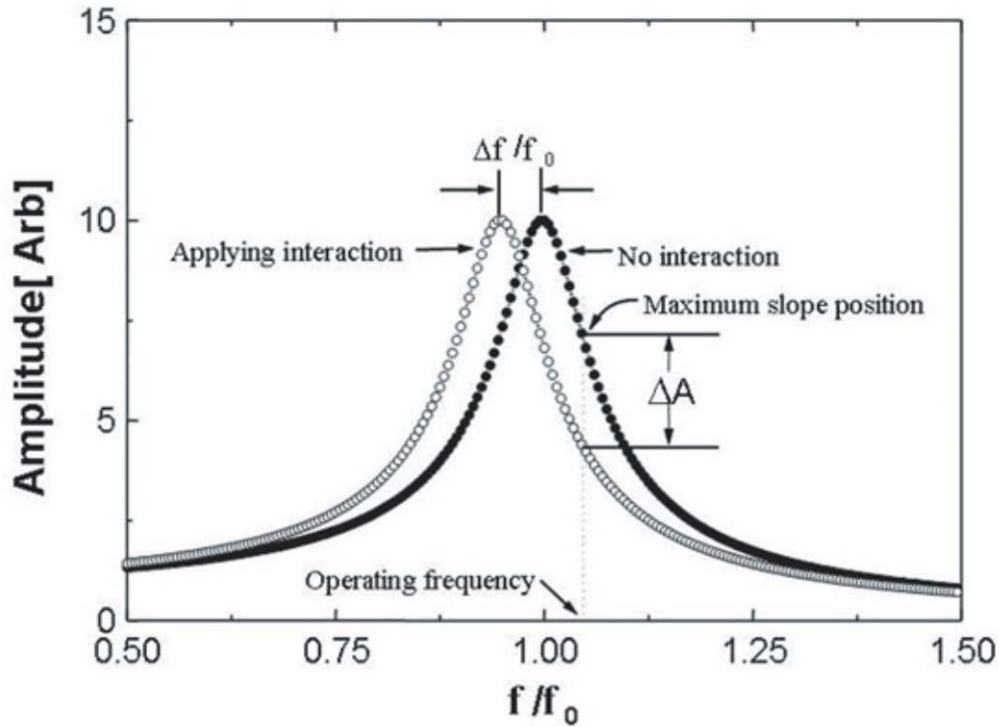
Fig. 7. Dynamic Force Microscopy (DFM).

If you measure the amplitude of vibration of the cantilever used in DFM of the XE-series while changing the frequency, there appears a special frequency where the amplitude resonates and amplifies greatly. This is called the intrinsic frequency (f_0). DFM of the XE-series uses the non-contact mode feedback circuit with keeping the vibrating frequency (f_1) a little bit lower than the resonant frequency while oscillating in free-space. Note that non-contact AFM uses the vibrating frequency higher than the resonant frequency. As the tip is lowered, the real spring constant reduces due to the attractive van der Waals force which becomes larger as the tip comes closer to the sample surface. Therefore the resonant frequency changes to effective frequency (f_{eff}) in non-contact regime and the amplitude at the frequency f_1 increases by ΔA. Since the amplitude increases by ΔA, the non-contact mode feedback circuit decreases the distance between the tip and the sample surface by Δd, indicated in the graph of vibration amplitude vs. tip-sample distance and z-feedback. Therefore, the vibrating cantilever, which is oscillating above the sample, approaches the sample almost in contact or in collision with the surface. This method, keeping intermittent contact between the sample surface and the vibrating cantilever is called Dynamic Force Microscopy (DFM).
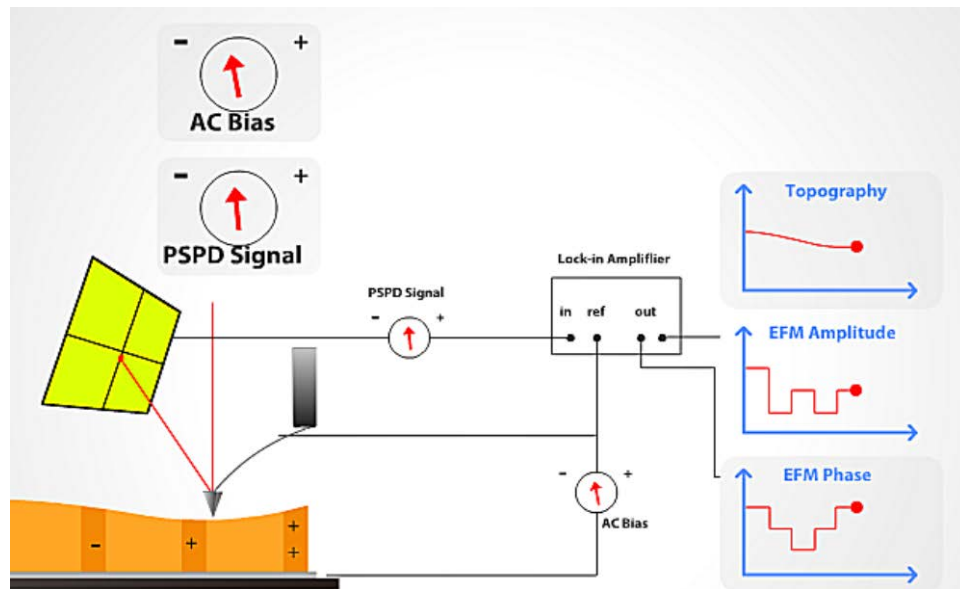
Fig. 7. Dynamic Force Microscopy (DFM).

Electrostatic Force Microscopy (EFM) probes ferroelectric regions of a sample surface with a conductive cantilever. An EFM image is the result of two separate scans: one scan probes the topography, while in the other the cantilever is raised away from the surface to the region where long-range, electrostatic force begins to dominate. In this electrostatic domain, the attractive and repulsive deflections of the cantilever correspond to regions of positive and negative charge on a sample surface. EFM gives users an image that couples topography with the electrical properties of a nanoscale region.
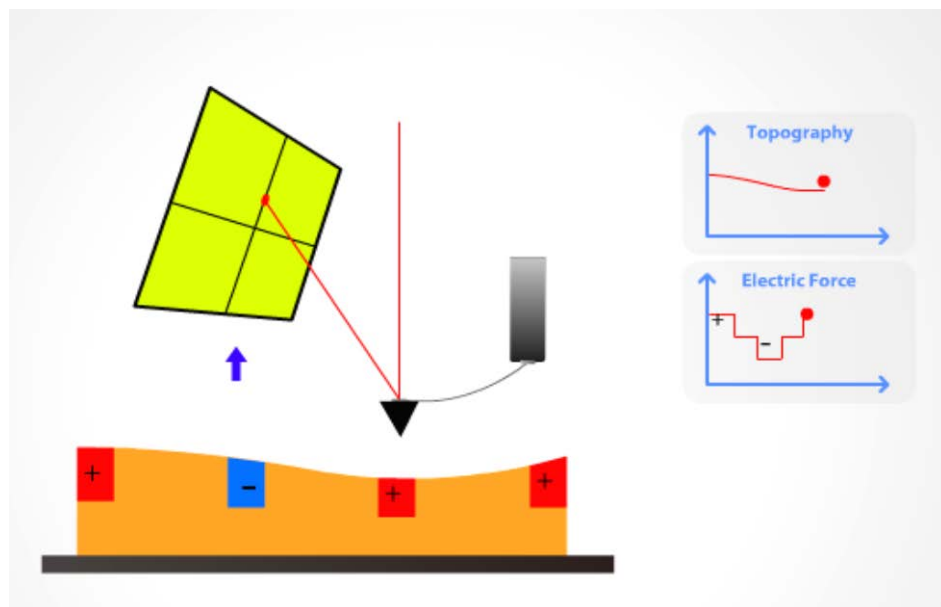
Fig. 8. Electrostatic Force Microscopy (EFM).

Electrostatic Force Microscopy (EFM) of the XE-series maps electric properties on a sample surface by measuring the electrostatic force between the surface and a biased AFM cantilever. EFM applies a voltage between the tip and the sample while the cantilever hovers above the surface, not touching it. EFM images contain information about electric properties such as the surface potential and charge distribution of a sample surface. EFM maps locally charged domains on the sample surface, similar to how MFM plots the magnetic domains of the sample surface. The magnitude of the deflection, proportional to the charge density, can be measured with the standard beam-bounce system. Thus, EFM can be used to study the spatial variation of surface charge carrier. For instance, EFM can map the electrostatic fields of a electronic circuit as the device is turned on and off. This technique is known as "voltage probing" and is a valuable tool for testing live microprocessor chips at the sub-micron scale.

Four different EFM modes, distinguished by the method which the surface electrical information is obtained, are provided by XE-series AFM. These are Standard EFM, enhanced EFM(ext), Park Systems's own patented Dynamic-Contact EFM (EFM-DC), and Scanning Kelvin Probe Microscope (SKPM). The enhanced EFM option of the XE-series contains enhanced EFM(ext), EFM-DC, and SKPM.

For EFM, the sample surface properties would be electrical properties and the interaction force will be the electrostatic force between the biased tip and sample. However, in addition to the electrostatic force, the van der Waals forces between the tip and the sample surface are always present. The magnitude of these van der Waals forces change according to the tip-sample distance, and are therefore used to measure the surface topography.
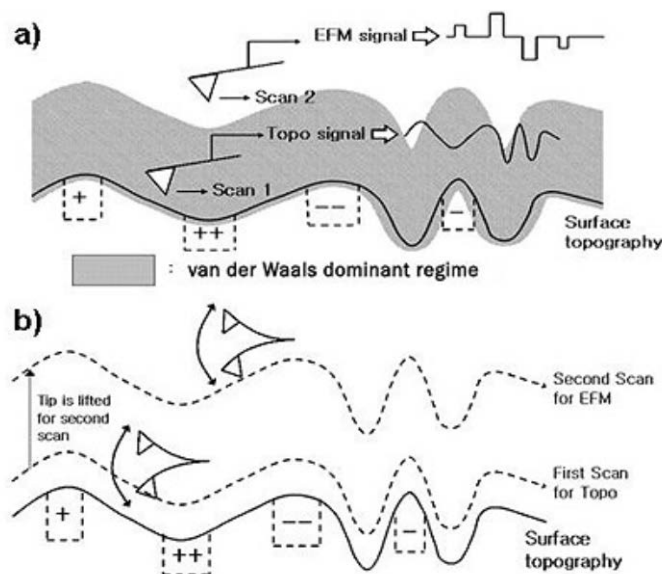
Fig. 9. The schematics of (a) Force Range technique and (b) two pass technique.

The standard EFM of the XE-series is based on the two facts. One fact is that van der Waals forces and electrostatic forces have different dominant regimes. van der Waals forces are proportional to $1/r^6$, while electrostatic forces are proportional to $1/r^2$. Thus, when the tip is close to the sample, van der Waals forces are dominant. As the tip is moved away from the sample, the van der Waals forces rapidly decrease and the electrostatic forces become dominant. The other fact is that the topography line is the line of the constant tip sample distance, which equals the line of constant van der Waals force.

In the Force Range technique, the first scan is performed by scanning the tip in the region where the van der Waals force is dominant for topography image. Then, the tip-sample distance is varied to place the tip in the region where the electrostatic force is dominant and scanned for EFM image as shown in Figure 9 (a).

In the Two Pass technique, the first scan is performed to obtain the topography by scanning the tip near the surface as it is done in NC-AFM, in the region where the van der Waals forces are dominant. In the second scan, system lifts the tip and increases the tip-sample distance in order to place the tip in the region where electrostatic forces are dominant. The tip is then biased and scanned without feedback, parallel to the topography line obtained from the first scan as shown in Figure 9 (b), therefore maintaining constant tip-sample distance.
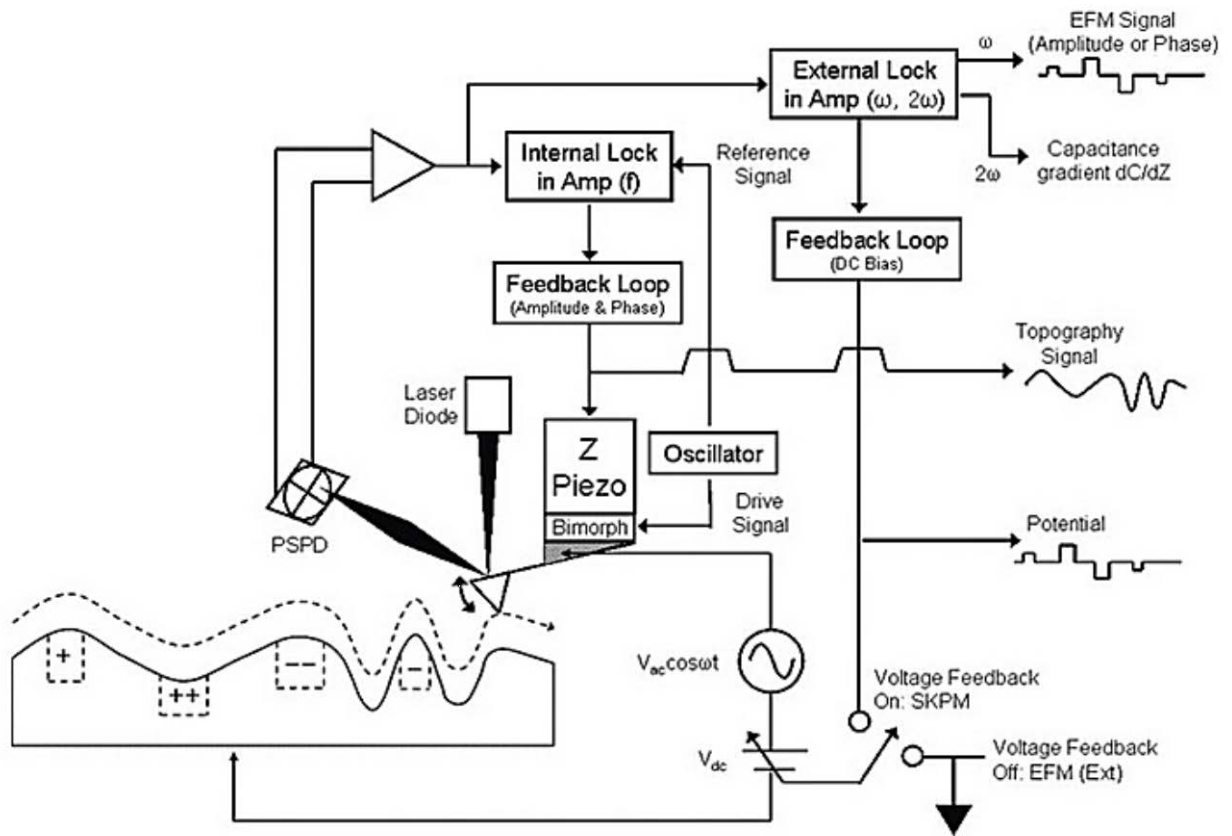
Fig. 10. Schematic diagram of enhanced EFM of the XE-series.

Since the topography line is the line of constant van der Waals force, the van der Waals forces applied to the tip during the second scan are constant. Thus, the only source of the signal change will be the change of the electrostatic force. So, from the second scan, a topography free EFM signal can be obtained.

Three extra EFM modes are supported by the enhanced EFM option of the XE-series. They are EFM (ext), EFM-DC (EFM-DC is patented by Park Systems US Patent 6,185,991), and Scanning Kelvin Probe Microscopy (SKPM), also known as Scanning Electrical Potential Microscopy.

In the enhanced EFM of the XE-series whose schematic diagram is shown in Figure 10, an external Lock-in Amplifier is connected to the XE-series AFM for two purposes. One purpose is to apply AC bias of frequency ω, in addition to the DC bas applied by the XE controller, to the tip. The other purpose is to separate the frequency ω component from the output signal. This unique capability offered by the XE-series enhanced EFM is what excels in performance when compared to the Standard EFM.

In the enhanced EFM, the voltage between the tip and the sample can be expressed by the following equation:

$$V(t) =_S +V_{AC} \, Sin\left(\omega t\right) \tag{1}$$

$$F = q \times E = q \times \frac{V}{d} = C \times \frac{V^2}{d} \tag{2}$$

$$F(t) = \left(\frac{C}{d}\right) \times V(t)^2 = \tag{a}$$

$$\left(\frac{C}{d}\right) \times \left[\left(V_{DC} - V\right)^2 + \frac{1}{2} V_{AC}^2\right] + \tag{b}$$

$$2 \times \left(\frac{C}{d}\right) \times \left(V_{DC} - V\right) \times V_{AC} \, Sin\left(\omega t\right) \tag{c}$$

Where V_{dc} is the DC offset potential, Vs is the surface potential on the sample and Vac and ω is the amplitude and frequency of the applied AC voltage signal, respectively. Equation (1) is appropriate if the geometry of the tip and sample can be approximate using two parallel plates. Other geometries can be assumed as well. Equation (2) can be used to derive an expression for the electrostatic force between the tip and the sample: (Again, parallel-plate geometry is assumed.)

Here, F is electrostatic force applied to the tip, q is charge, E is electric field, V is electric potential, C is capacitance, and d is tip to sample spacing. Note that since both AC and DC bias are applied between the tip and the sample, three terms arise in the expression for the force between the tip and the sample. These terms can be referred to as the DC term (a), the ω term (b), and the 2ω term (c), respectively.

The total cantilever deflection signal, which represents the force between the tip and sample, can be analyzed in terms of its separate parts: DC part, which comes from the van der waals interaction between the tip and the sample, AC part with a frequency of ω, which reflects the electrostatic properties of the sample, and AC part with a frequency of 2ω. The DC part comes from van der waals interaction between the tip and the sample, and the AC part with a frequency of ω and 2ω reflects the electrostatic properties and the capacitive properties of the sample, respectively.

The DC cantilever deflection signal can be read directly from the signal channels accessible using XEP Data Acquisition software. The AC parts of the cantilever deflection signal can be read by sending the signal to a Lock-in Amplifier, which can read either the part of the signal with a frequency of ω, or the part of the signal with a frequency of 2ω. Together, the three signals can be used to gain information about the electrical properties of the sample. For example, the capacitance appears in the equation as the ratio of capacitance to tip-to-sample spacing, C/d. If the tip-to-sample spacing is kept constant by the z feedback loop, then C/d is proportional to the capacitance. The ω signal, which is the coefficient of the term labeled (b) in Equation (2) above, contains contributions from both C/d and the surface potential, VS. Assuming VDC and VAC are known, you still cannot separate the contributions of the capacitance and the surface potential to the measured ω signal.

However, the 2ω signal, which is the coefficient of the term labeled (c) above, only includes a contribution from the capacitance.

Thus, the 2ω signal can be used to normalize the ω signal, isolating the contribution of the surface potential. Images can be generated from any of the above-mentioned signals. Analysis of an image involves understanding the contributions to the signal used to generate the image.

EFM(Ext) mode is the enhanced EFM that operates in Non-Contact mode. In the EFM (ext), tip scans over the surface while oscillating in frequency f to obtain the non contact AFM topography image. At the same time, AC bias of frequency ω is applied to the tip via Lock in Amplifier and DC bias is applied from the control electronics. This results the force between the AC biased tip and charged surface. Using the external lock in amplifier, signal resulting from the tip's motion by the force can be decomposed and analyzed into DC part, frequency ω part and frequency 2ω part. ω part of the signal contains information of surface charge, and

2ω part of the signal contains information of the gradient of surface capacitance between the tip and the sample (dC/dz). The frequency ω is chosen to be smaller enough than the cantilever oscillation frequency f, so that the two signals do not interfere each other.

The key advantages of Enhanced EFM(ext) are as follows:
• Nonevasive True Non-Contact AFM
• Small electrical load due to a small capacitance
• High spatial resolution
• Ambient operation
• High voltage resolution (50 mV ~ 1 mV)
• High measurement bandwidth(DC ~ 100 GHz)

(a)   **(b)**
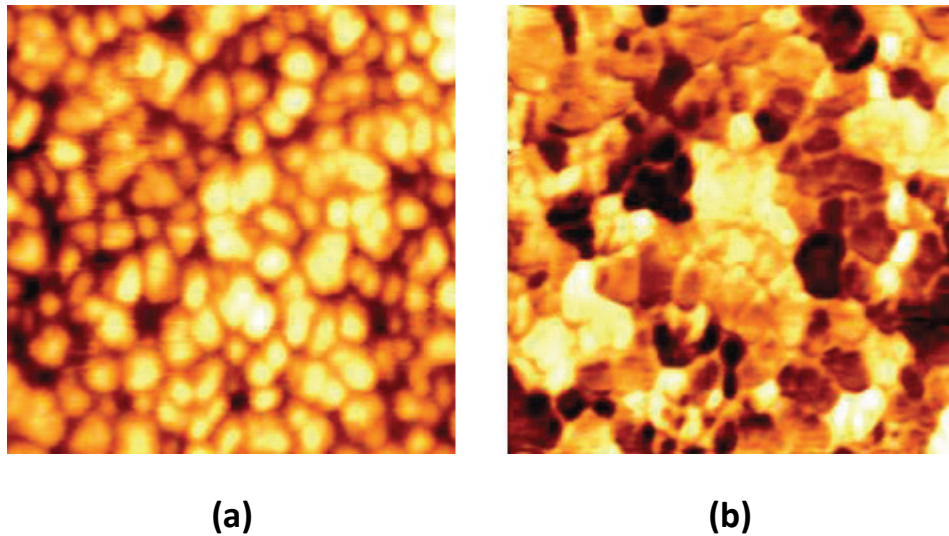
Figure 11. (a) Topography and (b) EFM phase image of the PZT film by Enhanced EFM(ext).

MFM measurements in the XE-series AFM follow the same procedure. For MFM, the surface properties would be magnetic properties and the interaction force will be the magnetic force between the magnetized tip and magnetic sample. However, in addition to the magnetic forces, Van der Waals forces always exist between the tip and the sample. These Van der Waals forces vary according to the tip-sample distance and therefore are used to measure the surface topography.

MFM images the spatial variation of magnetic forces on a sample surface with the MFM tip coated with a ferromagnetic thin film. In the XEseries AFM, surface topography is obtained in True Non-Contact mode. At the same time, the MFM image is generated by measuring either the amplitude or the phase change of the cantilever oscillation from the magnetic force between the surface and the magnetized MFM cantilever. MFM images contain information on magnetic

domain distributions on the sample surface. MFM can be used to image naturally occurring and deliberately written domain structures in magnetic materials.

During the MFM measurements, there are two forces acting on the tip; magnetic and Van der Waals forces. Hence, in MFM, the signal contains both information of surface topography (called 'Topo signal') and surface magnetic property (called 'MFM signal') generated by Vander Waals and magnetic forces, respectively. The dominating force depends on the distance between the tip and the sample surface, because the inter-atomic magnetic force persists for greater tip-to-sample separations than the van der Waals force. These van der Waals forces vary according to the tip-sample distance and therefore are used to measure the surface topography. If the tip is close to the surface, in the region where standard Non-Contact AFM is operated, the image will be predominantly topographic. As you increase the separation between the tip and the sample, magnetic effects become apparent. Collecting a series of images at different tip heights is one way to separate magnetic from topographic effects.
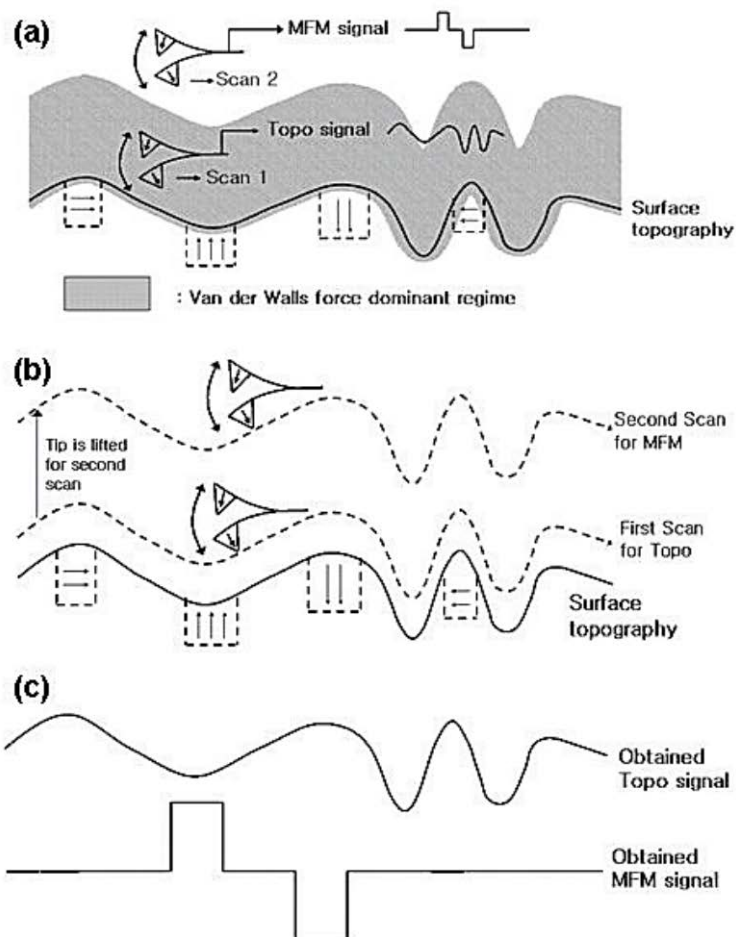
Fig. 12. The schematics of (a) Force Range technique and (b) two pass technique and (c) corresponding.

14

The key to successful MFM imaging lies in separating the magnetic signal from the entire signal. In MFM, this is done by 'Force Range' technique or 'Two Pass' technique.

Van der Waals force is a short range force while the magnetic force is a long range force. In the first scan, the tip is scanned in the region where the van der Waals force is dominant for topography image. Then, the set-point is varied to place the tip in the region where the magnetic force is dominant and scanned for MFM image, as shown in Figure 12 (a).

In this MFM mode of the XE-series AFM, sample is scanned twice to separate the signal as shown in Figure 12(b). In the first scan, the tip scans the surface as in True Non-Contact AFM to obtain the topography of the sample. In the second scan, the tip-sample distance is increased and the biased tip is scanned along the topography line obtained from the first scan as shown in Figure 12(c). The tip is only affected by the magnetic force and MFM image is obtained as a result.

The topographical line maintains constant tip sample distance, which equals the line of the constant van der Waals force. So, when the tip follows the topography line in the second scan of 'MFM mode', the van der Waals forces acting on the tip are kept constant. Thus, the only change in force affecting the signal is the change of the magnetic force. So, from the second scan, the MFM image, free of topography signal, can be obtained.
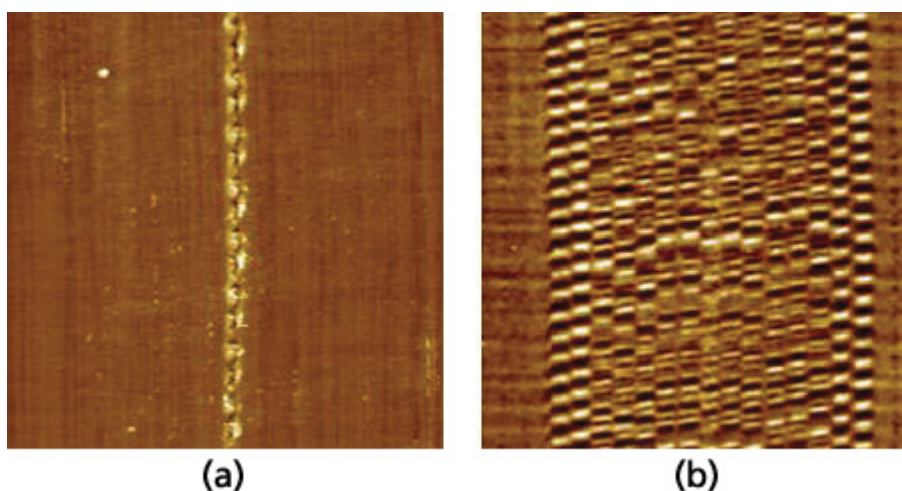
Figure 13.Topography (a) and MFM phase (b).

Force Modulation Microscopy (FMM) of the XE-series is an extension of AFM imaging that operates in contact atomic force microscopy (C-AFM) mode and is used to detect variations in the mechanical properties of the sample surface such as surface elasticity, adhesion, and friction.
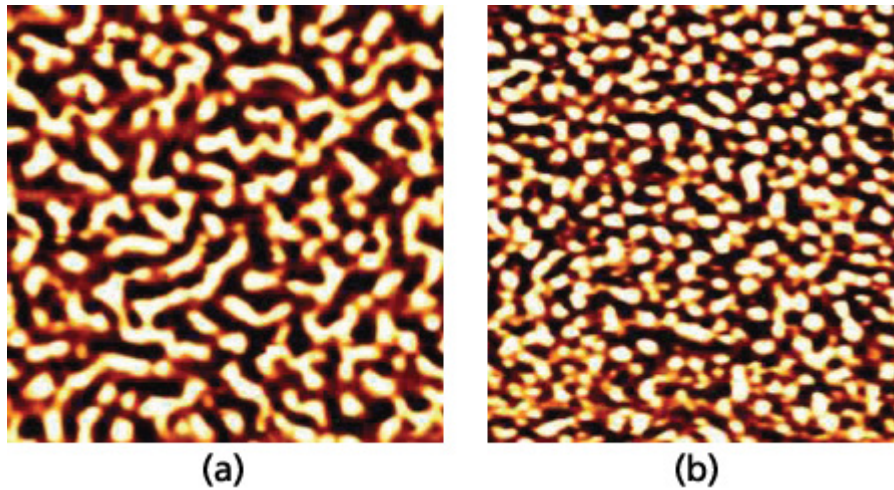
Fig. 14. MFM phase image of (Co84Cr16)100-xPtx X=13 (a) and X=28 (b).

In FMM mode, the AFM tip is scanned in contact with the sample surface, and the Z feedback loop maintains a constant cantilever deflection as in constant-force mode AFM. In addition, a periodic signal known as the 'driving signal' is applied to the bimorph piezo and vibrates either the tip or the sample. The resulting tip motion is converted to an electrical signal. This electrical signal is separated into AC and DC components for analysis.

The DC signal represents tip deflection as in contact AFM. The Z feedback loop uses this signal to maintain a constant force between the tip and the sample to generate a topographic image. The AC signal contains the tip response due to oscillation. The amplitude of the AC signal (called 'FMM Amplitude') is sensitive to the elastic properties of the sample surface. A hard surface will deflect the oscillation, resulting in a large amplitude response. On the other hand, a soft surface will absorb the oscillation, resulting in a small amplitude response. The FMM image, which is a measure of the sample's elastic properties, is generated from variations in the FMM amplitude. An example of the Amplitude Response using FMM is shown in Figure 15.

The phase shift of the AC signal (called 'FMM Phase') also differs according to the elastic properties of the sample surface. Hence, FMM Phase can be used to generate an FMM image. This technique is called Phase Detection Microscopy (PDM). Often, FMM Phase is more sensitive to the elastic properties of the surface than FMM Amplitude. PDM provides an additional contrast mechanism within a region of homogeneous hardness. Figure 15 shows a schematic diagram of the PDM principle.
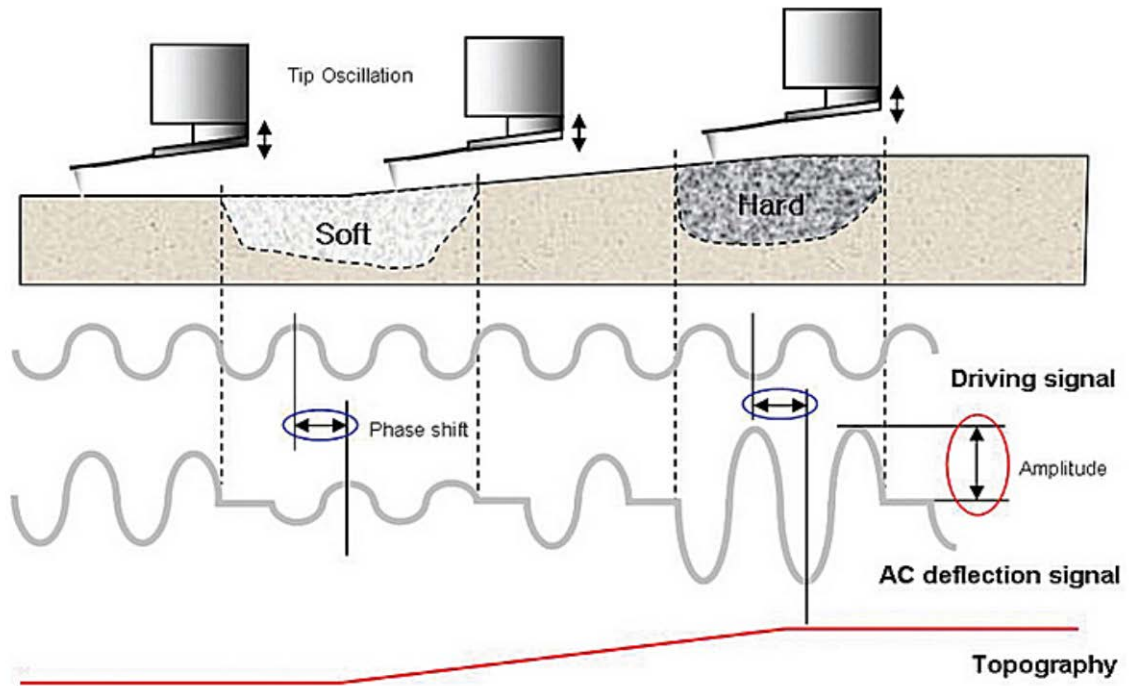
Fig.15. The amplitude of cantilever deflection varies according to the mechanical properties of the sample surface.

The FMM generates a force modulation image, which is a map of the sample's elastic properties, from the changes in the amplitude of cantilever modulation. The frequency of the applied signal is in the order of hundreds of kilohertz, which is faster than the Z feedback loop is set up to track. Thus, topographic information can be separated from local variations in the sample's elastic properties, and the two types of images can be collected simultaneously as shown in Figure 15.

In the XE-series, the key aspects of both Contact and True Non-Contact AFM are combined to generate FMM imaging. The feedback parameters of the Contact AFM of the XE-series are used to modulate the contact force between the tip and the sample while the frequency sweep, drive frequency, and drive gain of the True Non-Contact AFM in the XE-series are used to control the tip or sample vibration, subsequently affecting the contrast of an FMM image.

Figure 16 shows the topography and FMM image of glass fiber-PP (poly-propylene) composite. One can observe the difference of stiffness between the glass and the PP regions. In Figure 17 the topography and FMM image of bland polymer (SIBS 1027 60%+SMA 14 40%) are shown. In topography alone, the different components (SIBS and SMA) of the sample cannot be distinguished. However, the regions of SIBS and SMA having different stiffness property are clearly manifested in the FMM image.
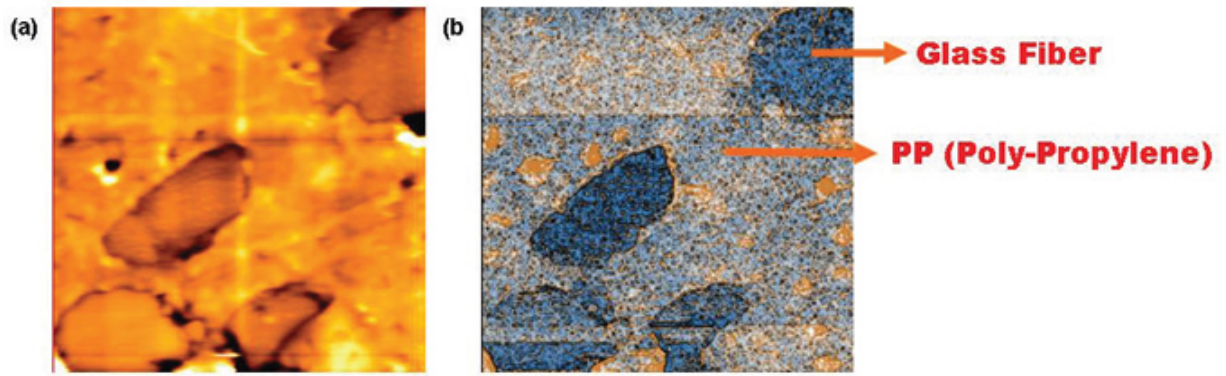
Figure 16. (a) Topography and (b) FMM image of glass fiber-PP composite (30 µm scan size).
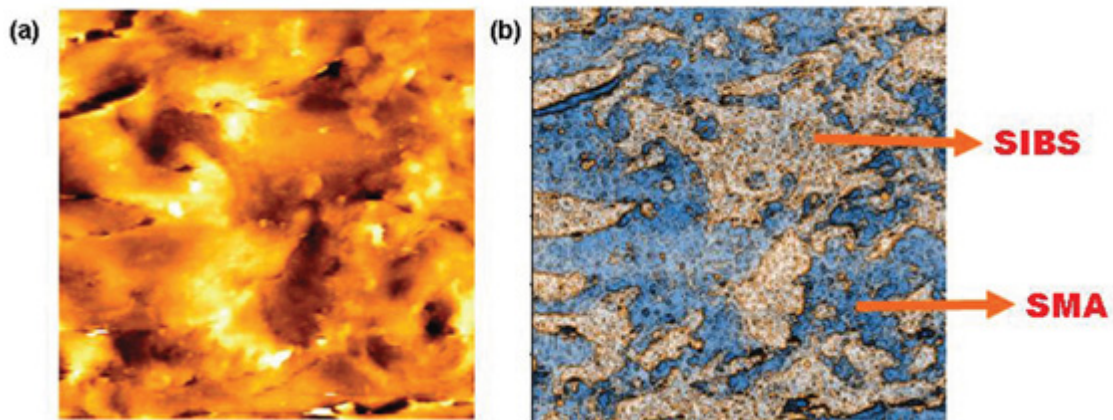
Figure 17. (a) Topography and (b) FMM image of blend polymer : SIBS 1027 60% +SMA 14 40% (30 µm scan size).

Hardness measures the capability of a material to with stand external force while maintaining its structure and form. NanoIndentation is one of the XE Modes that determines the hardness of a local region. Its application includes hard coating (titanium nitride coating), microelectronics (reliability of metal bond), thin film technology (adhesion failure), and polymers. By pressing down on a sample with a sharp tip as shown in Figure 18, an indent is made. The depth and area of the indent are correlated with the hardness. Other properties such as elasticity, adhesion, creep, and tribology can also be calculated from the indentation data.

The indentation itself is very simple. The difficulty lies in the analysis. The important factors in this technique are the force calibration and the depth, or so-called displacement. The force constant of a cantilever, the shape of the tip, and other mechanical properties such as tip glue determine the total force exerted onto a sample. Therefore, the accurate imaging of the shape of the indenter as shown in Figure 19, becomes a huge advantage when it comes to a reliable force calibration.

18
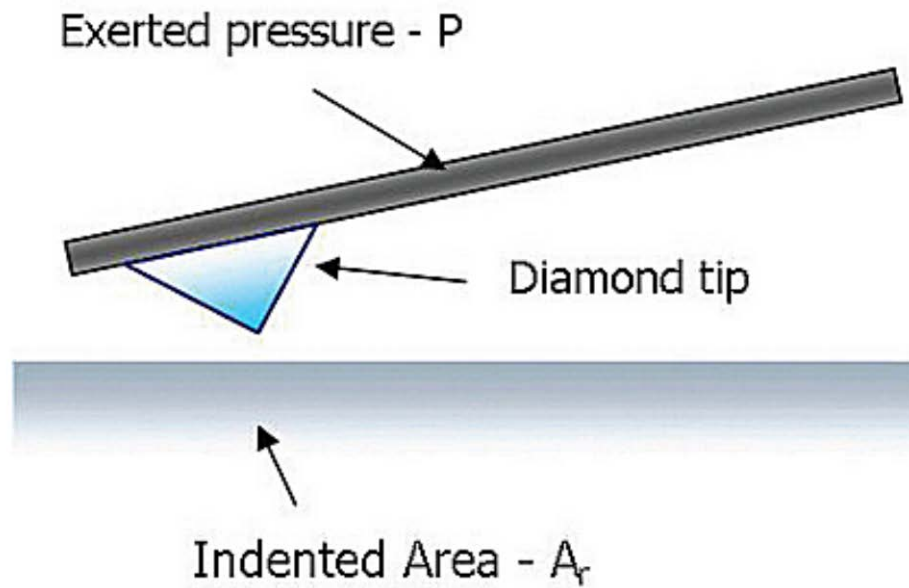
Fig. 18. NanoIndentation measures the hardness of a local region by pressing the indenter tip Into a sample.
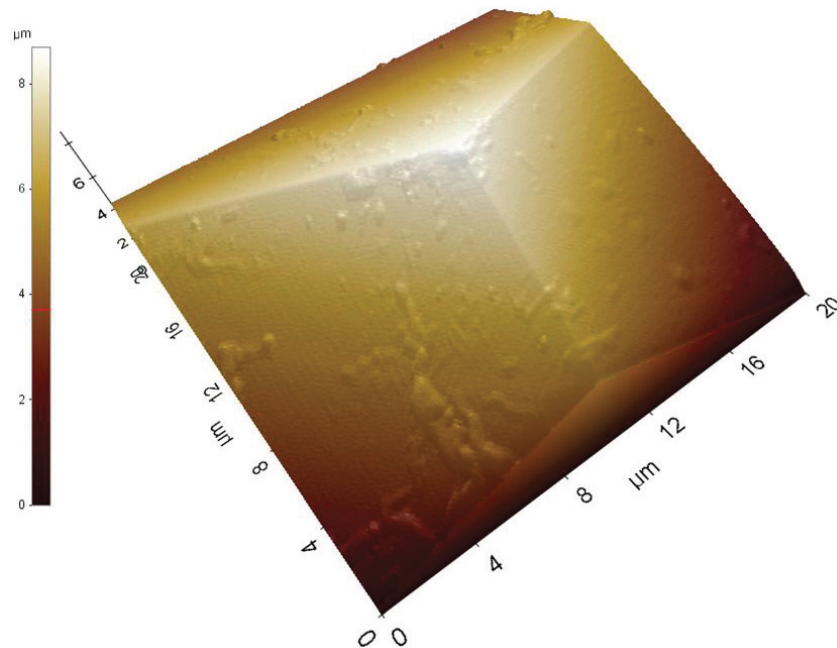
Fig. 19. The image of Berkovich indenter. The image was taken using True Non-Contact mode by the XE-series AFM. This kind of image can be only taken by True Non-Contact mode, and impossible to image by either tapping or contact AFM.

Only the unique capability of True Non-Contact AFM by the XE-series can accomplish such imaging of the Berkovich indenter. Coupled with the spring constant calibration by a reference cantilever array (see Spring Constant Calibration by Reference Cantilever Array), the force calibration in the XE-series AFM is the unparalleled accuracy and reliability. As to the depth sensing, accurate detection of the bending of the cantilever and mechanical simulation of the system are important.

The SEM image of our MEMS fabricated cantilever, made by the ECM technique, is shown in Figure 20. Its merit is that the cantilever has a well defined actuation part, so that one can accurately predict the mechanical characteristics. A common indenter is a three-sided pyramidal single-crystal natural diamond tip, a Berkovich indenter. A sharper indenter may be more effective in high resolution & sensitivity indentation, but it is very difficult to simulate and get quantitative results from such sharp tip features. The advantage of Berkovich indenter is that we can simulate the result and make a quantitative force measurement without damaging the indenter tip.
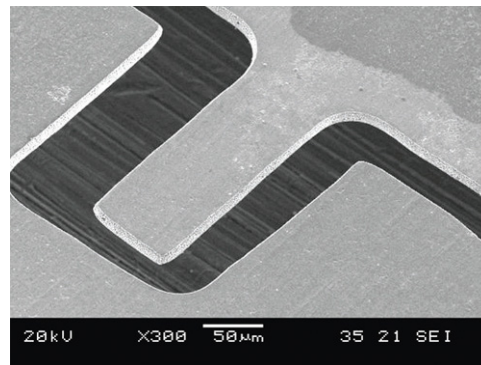
Fig. 20. SEM image of Park Systems's cantilever beam.
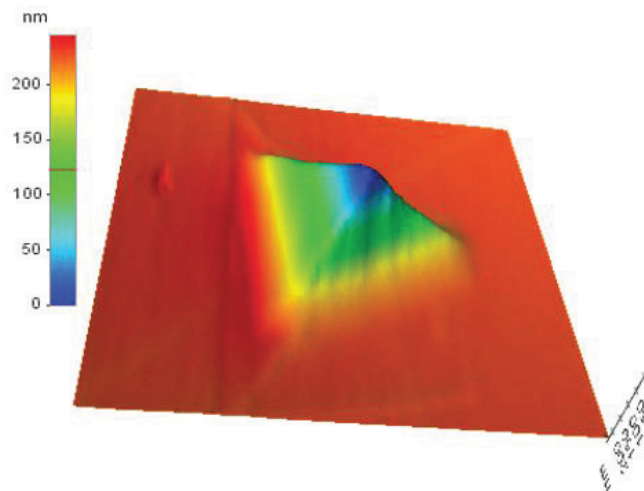
Fig. 21. An upside down image of a NanoIndentation pit, taken by True Non-Contact mode of the XE-series AFM.
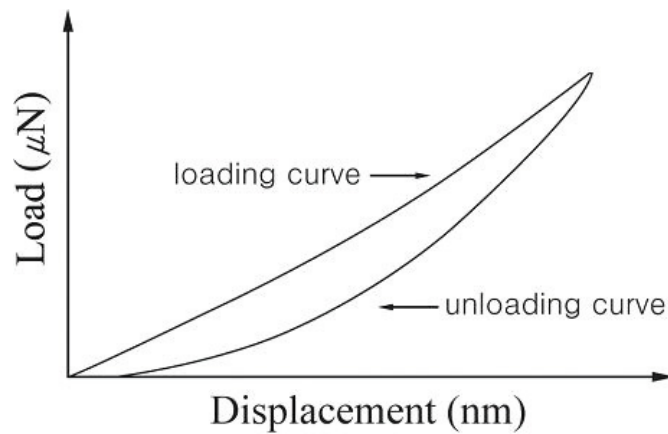
Fig. 22. A loading and unloading displacement curves during nanoindentation.

The indenter is mounted on the beam such that one sharp edge is parallel to the long axis of the beam. Indenter stresses the surface of a sample at a specified force for about 1-2 seconds. The indentation pit has a distinct geometry, which can be characterized using a high precision AFM as shown in Figure 4. Moreover, AFM is the only method capable of exact geometry determination of such tiny areas. It goes without saying that True Non-Contact AFM of the XE-series enables a very accurate and reliable imaging of the indentation tip.

Hardness is calculated by dividing the loading force by the projected residual area of the indentation. Besides hardness, the Young's modulus of elasticity can be obtained from the slope of the unloading curve. A typical loading and unloading displacement curves during a NanoIndentation cycle are depicted in Figure 22.

The hysteresis indicates that the deformation is not fully elastic and partially inelastic. For compliant materials the loading and unloading curves would be closely identical since the deformation is mainly elastic. Here, NanoIndentation is primarily used for determination of elasticity and viscoelasticity.

Normally AFM is used to image a surface without damaging it. However, AFM can be used to modify the surface deliberately, by applying either excessive force, or by applying high-field pulses. Not only scientific literature, but also newspapers and magazines have shown examples of surfaces that have been modified atom by atom. This technique is known as Nanolithography.

Nanolithography patterns are created on the sample surface by two methods as shown in Figure 23. First method is to scratch the sample surface with hard tips, mechanically deforming the sample surface. Second method is to apply bias between the tip and the surface, inducing the change of the chemical properties of the surface. The lithographic patterns can be generated in diverse forms: point or grid. Raster Nanolithography is common where the entire image is scanned.
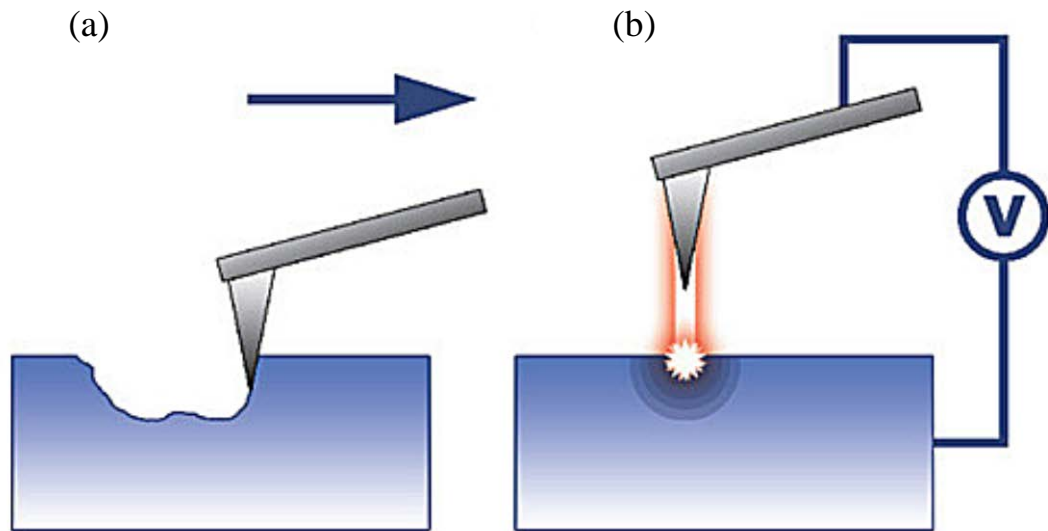
Fig. 23. Pattern created on the surface by plowing the surface with the tip (a) and by changing the surface with applied bias (b).

When it comes to Nanolithography, the value of advanced closed-loop system cannot be emphasized enough. It is no exaggeration that the reliability and efficiency of Nanolithography depend on it. With the state-of-the-art closed-loop system offered by the XE-series AFM, the vector Nanolithography can be realized.

The XE-series AFM use the Position Sensitive Photo Detector (PSPD) in the XY-scanner as the closed loop sensor, and PSPD monitors the movement of XY-scanner and controls it in real time. This enables the vector movement of the cantilever in XY direction. In other words, rasterizing, a common method used in Nanolithography of other AFM systems, is not necessary for XE-series AFM even though it is provided by the XE-series AFM. In a way, the vector Nanolithography is the ultimate measure of the performance in one's closed-loop system.

From Figure 24, the advanced performance of vector nanolithography is demonstrated. The Nanolithography image in Figure 24 (b) was written with two circles and two lines, individually in vector as shown in Figure 24 (a). The width of lines depends on several factors such as the applied voltage and/or the atmosphere humidity during the scan. Note that there is no distortion of the image even though the lines were drawn in vector movement, not rasterizing. This shows the superior performance of the closed-loop system of the XE-series AFM.
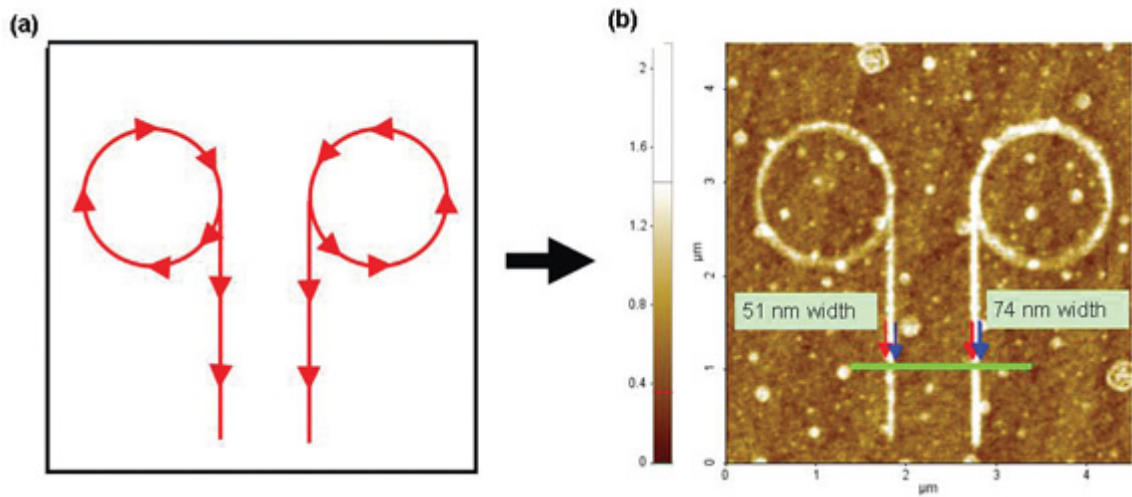
Fig. 24. Vector Nanolithography. The image is generated in vector by applying negative voltage between -5 and -10 V. Scanning rate was varied between 1 and 0.1 μm/s. The height of deposited oxide is about 2-4 nm.

Nanolithography process of the XE-series AFM is controlled by XEL lithography control software. XEL has a convenient user interface which makes the lithography process as easy as drawing a picture with simple graphic software. Just by mouse manipulation, objects can be easily drawn, resized and moved. Bitmap images can be imported for rasterizing and vector Nanolithography respectively.

The XEL software has the following advantages:

- A variety of lithography modes
- Separates Background and Object properties.
- Supports both Vector mode and Raster mode
- Provides a convenient Graphic Editor
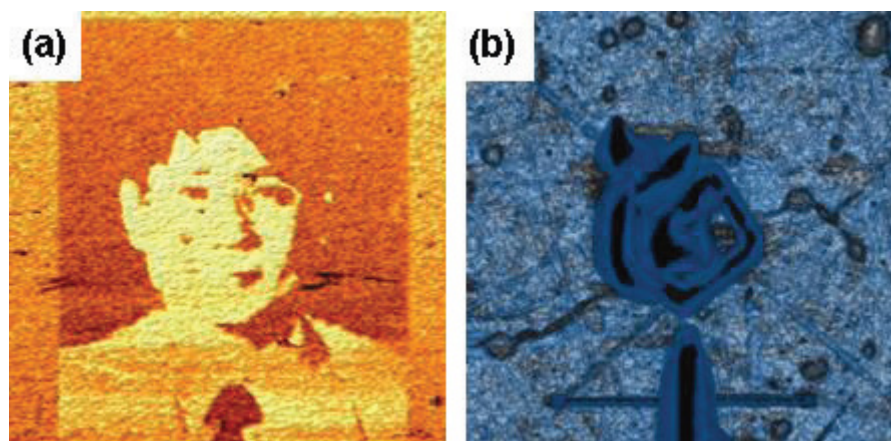
Fig. 25. Bitmap image pattern generation using (a)domain switching on PZT film, and (b) nanoscratch on polycarbonate film, 30 μm scan size.

Figure 25 is a bitmap image pattern generation using voltage induced Si oxidation. Electrical domain of PZT film can be switched by applying different bias between the tip and sample at different position. Figure 26 shows the before as shown in Figure 26 (a) and after as shown in Figure 4 (b) the domain switching of EFM images of the PZT film. Light colored regions are biased by 10 V and dark regions are biased by -10 V.
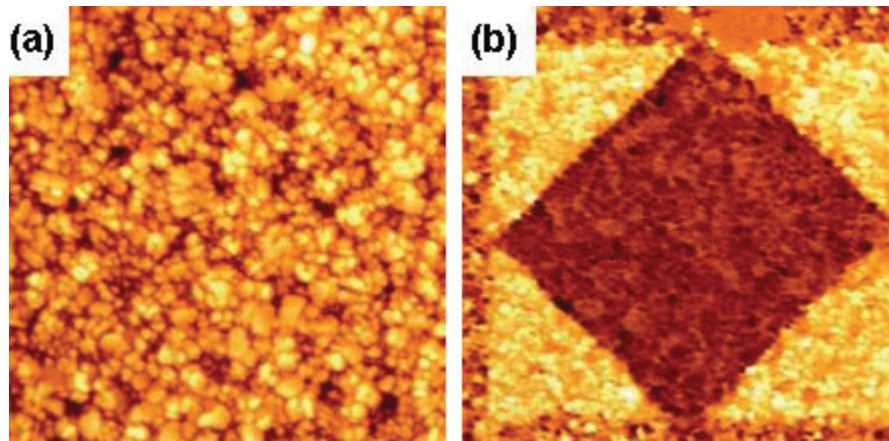
Fig. 26. EFM image of lithography performed on PZT film (a) before and (b) after the domain switching, 5 µm scan size.

Nanolithography in the XE-series is a technique used to pattern nano-scale shapes on the sample surface. For ordinary Nanolithography mode using the XE-series AFM, the available range of the applied voltage is -10 V to +10 V. For special situations requiring a higher voltage range, High Voltage Nanolithography mode has been developed. With the External High Voltage toolkit, XE AFM can be connected to external voltage amplifier enabling the experiments or measurements with tip or sample bias exceeding 10 V. The available voltage range for High Voltage Lithography depends on the external voltage amplifier.

Applications:

- Develop NanoLithographic media and techniques
- Study minimum feature size
- Minimum line spacing
- Resist exposure speed
- Exposure threshold

Tip-Enhanced Raman Spectroscopy is used to be when a sharp tip coated with gold approaches an illuminated sample surface. Using an AFM cantilever as the enhancer, nano- scale chemical properties are measured by an integrated Raman spectrometer while the topographic data is acquired simultaneously by AFM.
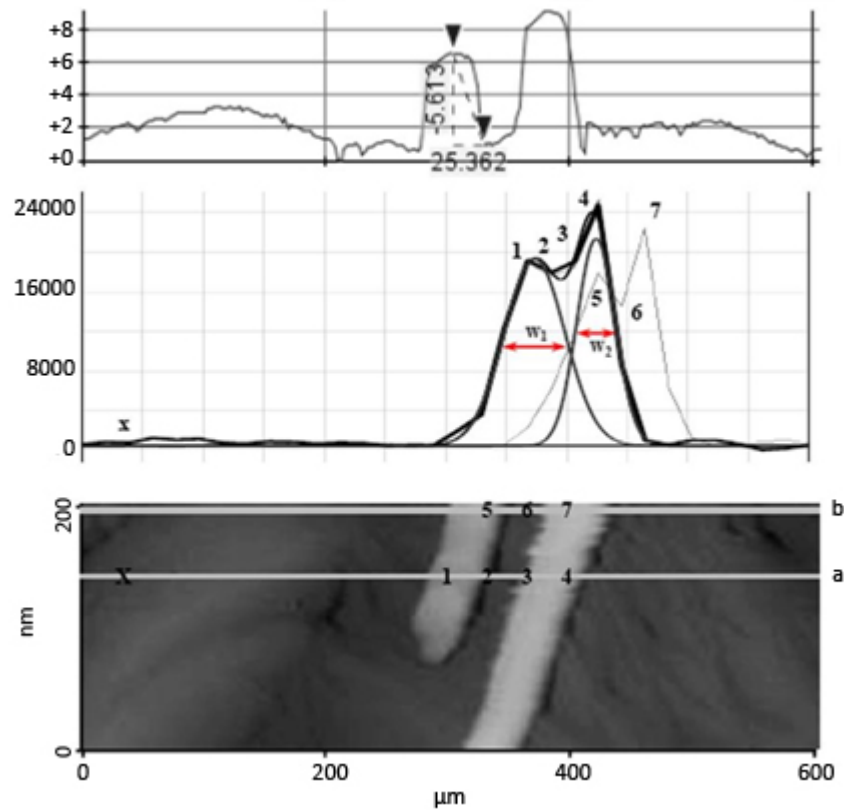
Fig. 27. Raman spectra and images obtained by TERS AFM.

There has been growing interest in the heat dispersion of nanostructured materials. **The XE-series Scanning Thermal Microscopy** (SThM) mode was developed to probe thermal properties at the nanoscale level. The XE-series SThM uses nanofabricated thermal probes to achieve unprecedented high spatial and thermal resolution and sensitivity with a unique signal detection scheme. The SThM technique of the XE-series maps the thermal properties of the sample surface by using a nanofabricated thermal probe with a resistive element. The XEseries SThM is available in two modes, Temperature Contrast Microscopy (TCM) and Conductivity Contrast Microscopy (CCM). TCM allows the user to measure the temperature variations on a sample surface. CCM allows the user to measure variations of thermal conductivity on a sample surface.
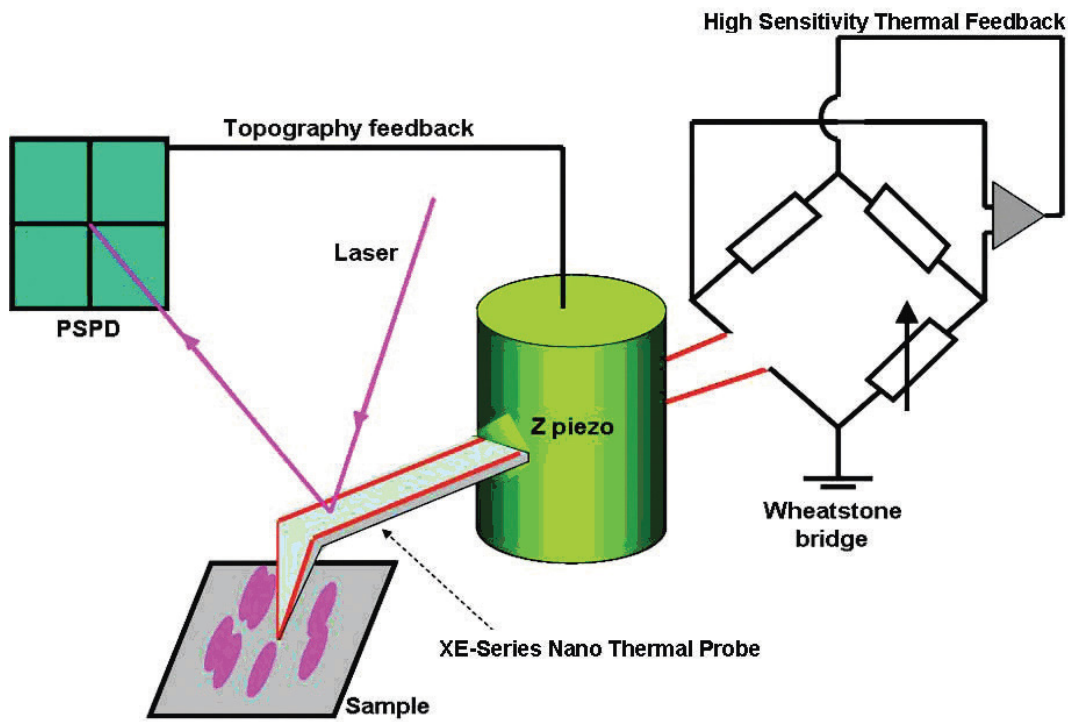
Fig. 28. Schematic diagram of the series SThM system

Figure 28 shows the schematic diagram of the XE-series SThM system. A "V" shaped resistive element is mounted at the end of a cantilever. While the distance between the probe tip and sample surface is controlled by usual AFM scheme, the thermal probe forms one leg of a Wheatstone bridge. It is this Wheatstone bridge which feedbacks, adjusts, and balances the bridge voltage in order to measure the probe's temperature (TCM) or maintain a constant probe temperature (CCM). A topographical AFM image can be generated from changes in the cantilever's amplitude deflection. Thus, topographic information can be separated from local variations in the sample's thermal properties, and the two types of images can be collected simultaneously.

The key part of the SThM is the SThM tip, which serves as a resistance thermometer (or a heater in CCM mode) at the same time as an AFM tip. The resitivity of the SThM tip is changed according to the teperature of the surface in contact, and the resistivity is read from the wheatstone bridge. Previous SThM designs, which was made by etching a wire, could not provide sufficient spatial and thermal resolution, critically limited by the geometry of a wire-based thermal probe, i.e. Wollastone wire. The XE-series SThM uses a nanofabricated thermal probe where a resistive element is lithographically patterned on the AFM tip.
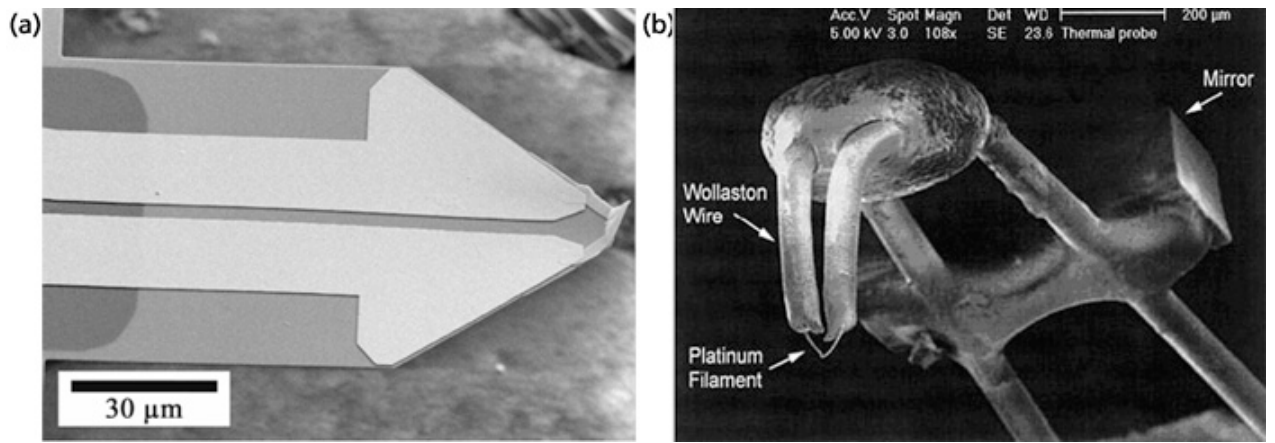
Fig. 29. The SEM images of (a) a XE-series Nano Thermal Probe and (b) a Wollaston wire.

Figure 29 (a) and 29 (b) shows scanning electron microscopy (SEM) images of a Wollaston wire thermal probe and the nanofabricated thermal probe used in the XE-series SThM. The tip radius of the nanofabricated probe is about 100 nm enabling high resolution thermal image scan while that of a Wollaston wire probe is larger than several hundred nm.
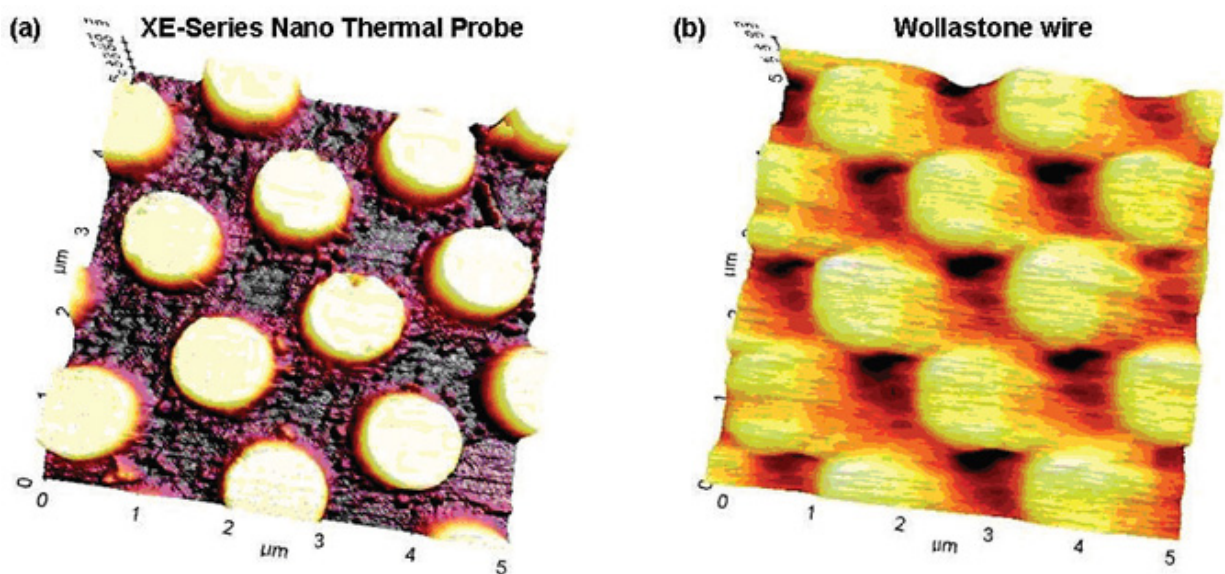
Fig. 30. Topography image comparison of HSQ posts of 1 µm diameter patterned on a silicon substrate (5 µm scan size) using (a) XE-series Nano Thermal Probe and (b) Wollastone wire.
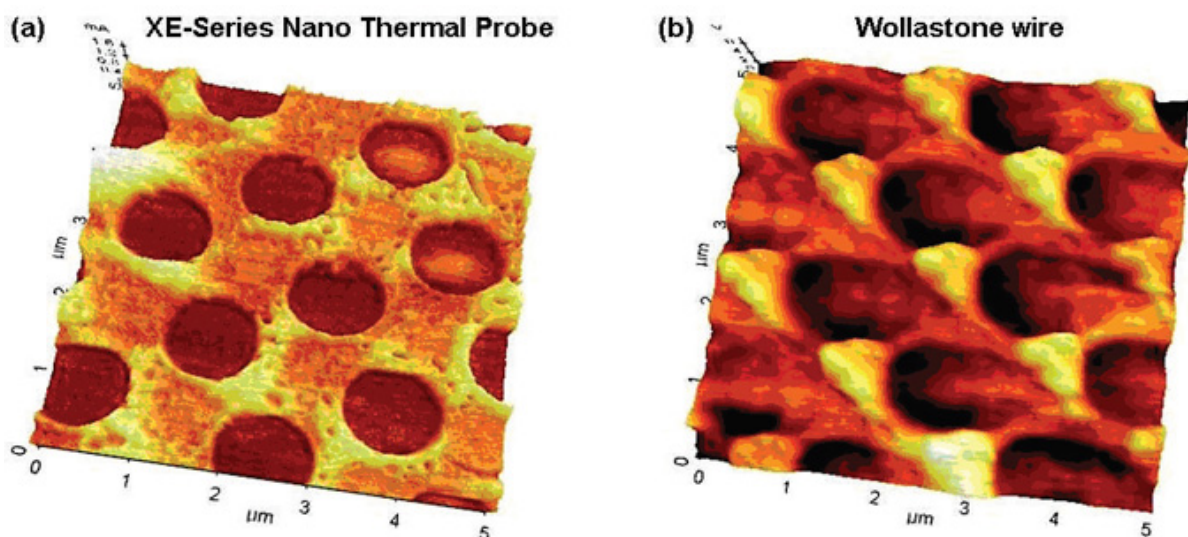
Fig. 31. Thermal conductivity image comparison of HSQ posts of 1 μm diameter patternedon a silicon substrate (5 μm scan size) using (a) XE-series Nano Thermal Probe and (b) Wollastone wire.

In Figure 30 and 31, a comparison is made between the XE-series Nano Thermal Probe and a Wollastone wire probe. The imaged sample is hydrogen silsesquioxane (HSQ) posts with 1 μm diameter on a silicon substrate. The detailed differences in topographic and thermal conductivity resolution are clearly demonstrated with the XE-series Nano Thermal Probe which has superior spatial and thermal resolution. Please note that such dramatic enhancements in resolution and sensitivity are realized only by combining the advantages of the nanofabricated thermal probe and the SThM mode sensitivity offered by the XE-series.

In TCM mode, the resistive element of the XE-series Nano Thermal Probe is used as a resistance thermometer. The temperature of the thermal probe changes as the tip scans the surface according to the surface temperature. Change of the wire temperature leads to change of its resistance. The temperature of a very small region can be measured by running a constant current, referred to as the 'Probe Current,' through the probe and measuring the resistance as shown in Figure 32. First, the tip is put into thermal equilibrium with the sample surface and thus its resistance is constant. At this time, the variable resistor in the bridge is adjusted so that the potential difference between the point 1 and 2 becomes zero. Then, the temperature of the probe changes as the probe scans over the surface. The corresponding change in probe resistance will alter the voltage balance of the bridge, changing the voltage difference between the points 1 and 2. This is referred to as 'SThM error'. This SThM error is used to generate the SThM image in TCM mode.

The current passed through the probe in TCM is set to be small enough that no self-heating of the probe occurs. (Resistance change due to the self heating would cause errors in temperature measurement.) Also in TCM mode, the scanning speed is limited by the time it takes for the tip to reach thermal equilibrium with the sample surface.

In Conductivity Contrast Mode, (CCM) the resistive element of the XE-series Nano Thermal Probe is used as a resistive heater. Sufficient energy is applied to the probe tip to keep it at a set temperature via a feedback loop. The energy required to maintain the set temperature represents the local thermal conductivity. Schematic diagram of the CCM is shown in Figure 33.

When the heated probe, preset at a value much higher than a sample temperature, makes contact, heat flows from the probe to the sample, resulting in the cooling of the probe. The feedback senses this shift, balances the bridge voltage, and restores the probe's resistance (or temperature) to its preset value. The raw data from the SThM of the XE-series reflects the feedback voltage, V_{out}, applied to the bridge. However, the thermal conductivity of the specimen is proportional to the heat flow ($\sim V_{out}^2$), when the tip is in contact with a sample. A simple calibration method can be implemented for absolute thermal conductivity measurement.

The heat flow between tip and specimen under investigation is controlled by the following three factors:

  • Thermal conductivity of the sample
  • Contact area of the probe
  • Temperature difference of the probe and the sample
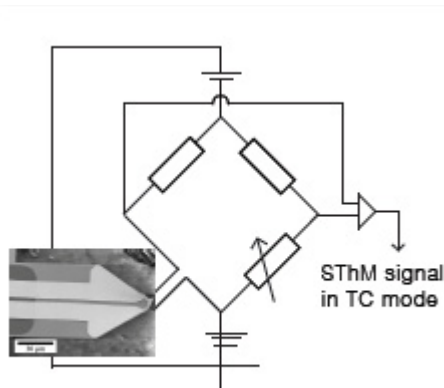
Fig. 32. Schematic diagram of the TCM mode.
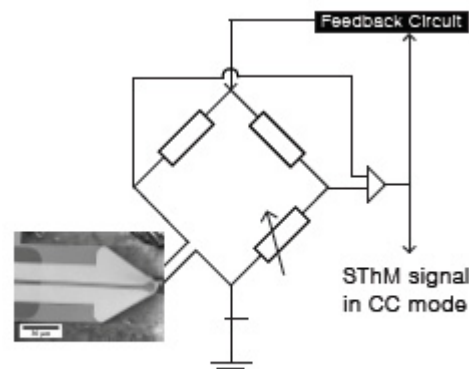
Fig. 33. Schematic diagram of the CCM mode.

For most of the samples the contact area changes of the probe-sample are negligible and, due to its large thermal mass, the sample remains at a constant temperature (the temperature difference between the probe tip and the sample also stays constant since the temperature of the probe is controlled by the feedback loop). As a result, the changes in heat flow will be only caused by changes in thermal conductivity of the sample.

As the thermal conductivity of the sample varies during the scan, the probe's temperature tends to change, however, the Wheatstone bridge uses the SThM error and feedback loop to balance the voltage applied to the tip in order to maintain its temperature constant, at the preset value.

Figure 34 shows the high resolution topography and thermal conductivity image of a 4.3 µm diameter HSQ post on a silicon substrate by the XE-series SThM with Nano Thermal Probe. Inhomogeneity in the thermal conductivity, due to impurities in HSQ composition, is observed in contrast to a flat topography.

In Figure 35 the high resolution topography and thermal conductivity of smaller HSQ posts with 0.2µm diameter on a silicon substrate are imaged, again, using the XE-series SThM with Nano Thermal Probe. In the thermal conductivity image, one can also observe the impurities, which is not apparent in topography.

It is evidently demonstrated that the XE-series SThM has a superior spatial and thermal resolution. It opens up great possibilities in the nanoscale investigation of thermal properties in various nanostructured materials.
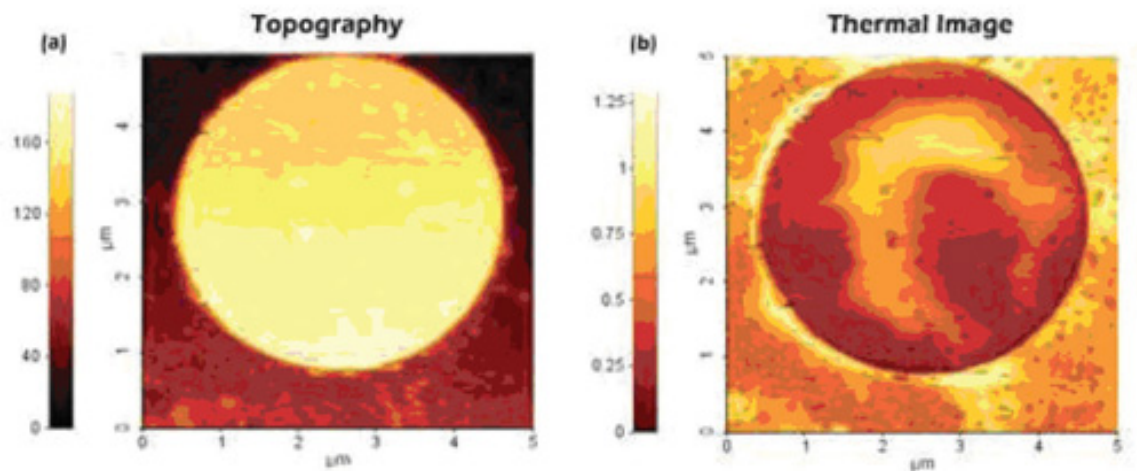
Fig. 34. (a) High resolution SThM topography and (b) thermal conductivity image of a HSQ post with 4.3 µm diameter on a silicon substrate (5 µm scan size) by the XE-series SThM with Nano Thermal Probe.
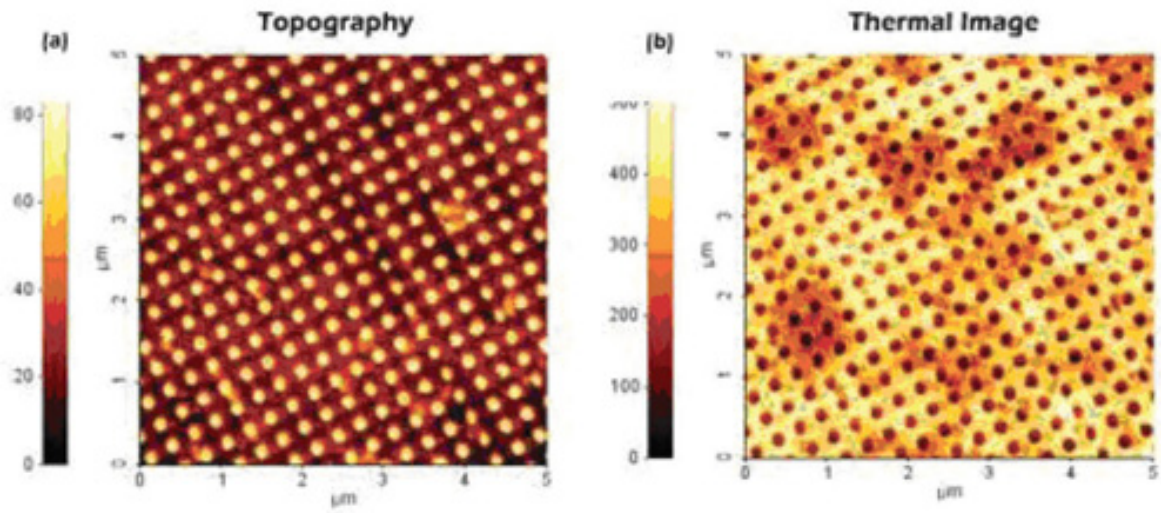
Fig. 35. (a) High resolution SThM topography and (b) thermal conductivity image of HSQ posts with 0.2 μm diameter on a silicon substrate 5 μm scan size) by the XE-series SThM with Nano Thermal Probe.

Chapter 11 Transdusers for nanobiosensors

4.1 Introduction

Biosensors and related bioarray techniques represent the end product of a rapidly growing field, which combines fundamental biological, chemical, and physical sciences with engineering and computer science to satisfy needs in a broad range of application areas. A biosensor is "an analytical device, which converts the concentration of the target substance, the analyte, into an electrical or optical signal through a combination of a biological or biologically derived recognition system either integrated within or intimately associated with a suitable physico-chemical transducer"(fig.4.1)
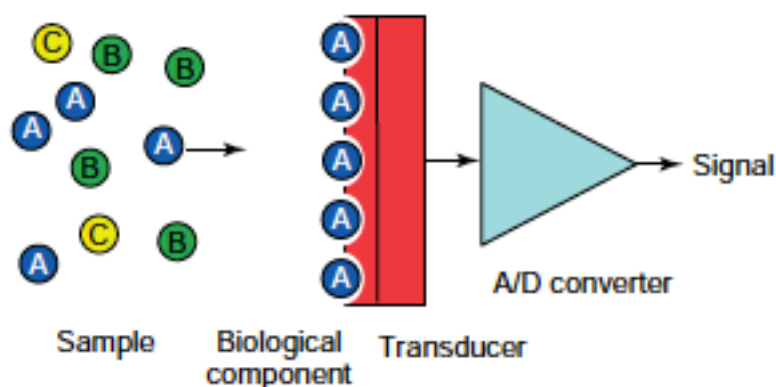
Fig.4.1 Biosensors recognition system

The biological or biologically derived element is capable of recognizing the presence, activity, or concentration of a specific target analyte in a complex mixture of other components. The recognition element may comprise one of three different types: affinity biosensors are based on ligand-receptor interactions such as those involving antibodies, nucleic acids, aptamers, peptides, protein, or cell receptors, while the second principal class involves binding and catalysis, and involves enzymes, microorganisms, organelles, plant or animal cells or tissue slices, and the third involves biomimetic receptors based on various synthetic binding and/or catalytic systems. The interaction or subsequent reaction of the recognition element with the analyte in the sample matrix results in a measurable change in a solution property, such as depletion of a reactant or formation of a product, immediately proximal to the transducer. The latter converts the change in solution property into a quantifiable and processible electrical signal. The transducer is a device, usually electronic, electroacoustic, electro-optical, electromagnetic, electrothermal, or electromechanical, that converts one type of energy (electricity, sound, light, magnetism, heat, or mechanical) into another (usually electrical) for various purposes including measurement or information transfer. Biosensors that use single or multiple enzyme sequences as the recognition and response elements are the most extensively investigated area where the target analyte is amenable to enzymatic modification and its concentration is typically in the range 10^{-1}–10^{-7} M. The high specificity and high

turnover rates of enzymatic reactions suggests ways to create tailor-made sensitive and specific enzyme-based biosensors for their respective substrates. The enzyme catalyst is integrated either on, adjacent to, or into the materials comprising the transducer in order to ensure that the biocatalytic transformation is selectively transduced into the electrical signal.

The oxidation of glucose catalyzed by GO_x is an example of the archetypal enzyme biosensor:

$$\text{Glucose} + O_2 \rightarrow \text{Gluconic acid} + H_2O_2$$

$$H_2O_2 \rightarrow 2H^+ + O_2 + 2e^-$$
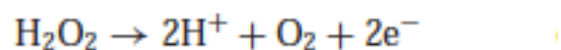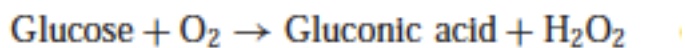
The father of the biosensor concept can be identified clearly as the late Professor Leland C. Clark Jr. In 1956, Clark published his definitive paper on the oxygen electrode, a schematic of which is shown in Figure 4.2.
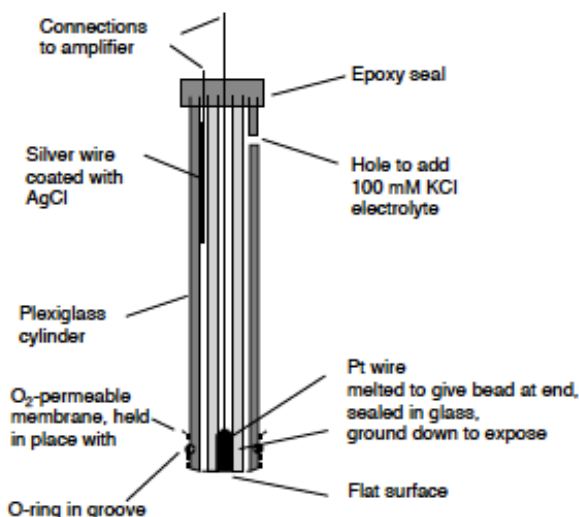
Figure 4.2. Schematic of Clark oxygen electrode.

Clark-type devices rely on the application of a potential of between −0.6 and −0.8V between the Ag/AgCl reference and the platinum working electrode. The oxygen dissolved in the solution between the electrodes is reduced (Figure 4.3). The resultant current is proportional to the local oxygen concentration.

From the reaction scheme in Figure 4.3, it is apparent that it is also possible to measure glucose via the oxidation of the hydrogen peroxide produced by the enzymatic reaction, since this is also directly proportional to the concentration of glucose. This can be measured amperometrically at a potential of approximately +0.7V versus Ag/AgCl, when a platinum working electrode is used.

The Clark-style sensor is not restricted to measuring glucose. Particularly in the early years of biosensor development, numerous variations on the basic design and many other (oxidase) enzymes were immobilized by various workers. Other electrode configurations, featuring carbon and other working electrodes can be constructed and operated via amperometry in a similar manner.
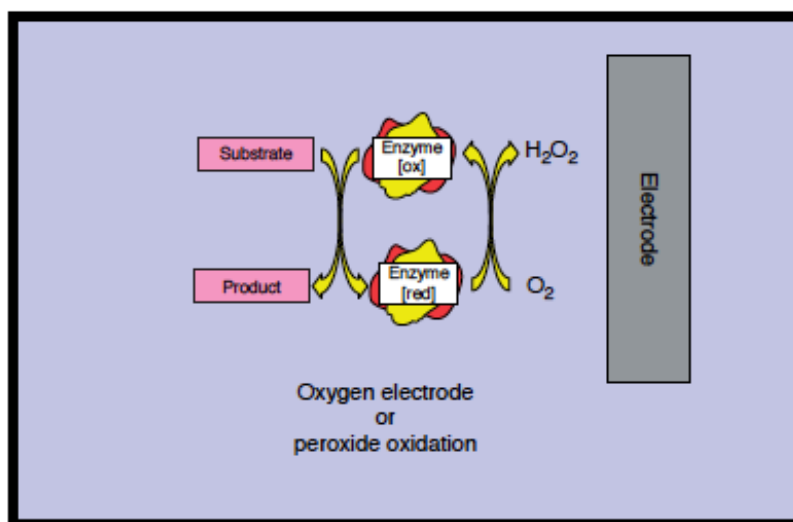
Figure 4.3. Clark-type biosensor schematic.

In principle, many transducers can be used in a biosensor and early researchers were quick to exploit these opportunities. In 1969, Guilbault and Montalvo were the first to detail a potentiometric enzyme electrode. They described a urea sensor based on urease immobilized at an ammonium-selective liquid membrane electrode (see reaction scheme below). Others have used pH electrodes and pH ion-selective field-effect transistors (ISFETs) as transducers.

$$\text{Urea} + \text{H}_2\text{O} \xrightarrow{\text{Urease}} \text{CO}_2 + 2\text{NH}_3$$
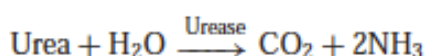
Antibodies and their various truncated antigenbinding elements are the most commonly employed biorecognition systems for use in biosensors for analytes at concentrations in the 10^{-6}–10^{-10} M range. Whole polyclonal and monoclonal antibodies, recombinant antibodies, phagedisplay systems, Fab, single chain antibodies (scfv), and minibodies have all been exploited in biosensors. The sensitivity and specificity of an immunosensor are governed by specificity and affinity of the antigen–antibody interaction and the signal-noise ratio of the transducer. However, in practice, adequate sensitivity can only be achieved.

Molecularly imprinted polymers are artificial recognition systems, which mimic natural molecular recognition and are of two types, covalent and noncovalent. These systems have been exploited in sensors for a variety of analytes, including pesticides, biotin derivatives, metalloporphyrins, sterols, polymers, anti-idiotypes and chloramphenicol.

4.2 Type of transducers

The transducer is the device that converts a wide range of physical, chemical, or biological effects into an electrical signal with high sensitivity and minimum disturbance to the measurand. Transducers are often described in terms of their sensitivity to input signals or responsivity, which is simply the ratio of the output to input signals. Different types of transducers act on heat, light, sound,

magnetism, electricity, radiation, strain, vibrations, pressure, and acceleration, and most types have been used in biosensors at some point in time.

The mid to late 1970s witnessed an explosion of biosensor concepts. It was quickly realized that there were an enormous number of combinations of transducers (Table 4.1) and bioreceptors. During this time, many of these combinations were tested, initially with enzymes and then with other biological components.

Table 4.1. Examples of transducers used in biosensor construction

Transducer	Examples
Electrochemical	Clark electrode, mediated electrodes, ISEs, FET-based devices, LAPS
Optical	Photodiodes, waveguide systems, integrated optical devices
Piezoelectric	Quartz crystals, SAW devices
Calorimetric	Thermistor, thermopile
Magnetic	Hall effect, magnetoresistive
Micromechanical	Viscosity, pressure, force, and oscillating cantilevers

ISEs, ion-selective electrodes; FET, field-effect transistor; LAPS, light-addressable potentiometric sensors; SAW, surface acoustic wave.

4.3 Chemical and biological field-effect sensors for liquids

Semiconductor-type field-effect devices (FED) based on the electrolyte–insulator–semiconductor (EIS) concept, that is, ISFET (ion-sensitive field effect transistor), capacitive EIS sensor, and LAPS (light-addressable potentiometric sensor) are currently one of the basic structural elements of chemical and biological microsensors with new functional and application possibilities. They provide a lot of potential advantages such as small size and weight, robustness, fast response time, high reliability, batch processing capability, and so on. On the other hand, the possible field of application of these three kinds of FEDs reaches from medicine, biotechnology, and environmental monitoring over food and drug industries up to defense and security purposes including antibioterrorism and biological warfare agents field. The miniaturization of FEDs and their compatibility with advanced microfabrication technology also make them very attractive for the integration into microfluidic platforms in order to build up miniaturized analytical systems such as micro total analysis system (μTAS), "lab on chip", and electronic tongue devices.

Figure 4.4 shows a typical structure of a capacitive EIS sensor (a), a LAPS (b), and an ISFET (c). These sensor structures are obtained by replacing the metallic gate of the metal–insulator– semiconductor (MIS) capacitance and the insulated-gate field-effect transistor (IGFET), respectively, by an electrolyte solution and a reference electrode. Since (bio-)chemical FEDs are very sensitive for any kind of electrical interaction at or near the gate insulator/electrolyte interface, in general, nearly each (bio-)chemical reaction leading to chemical or electrical changes at this interface can be measured. Therefore, the ISFET, capacitive EIS sensor or LAPS must be coupled with the respective chemical or

biological recognition element.
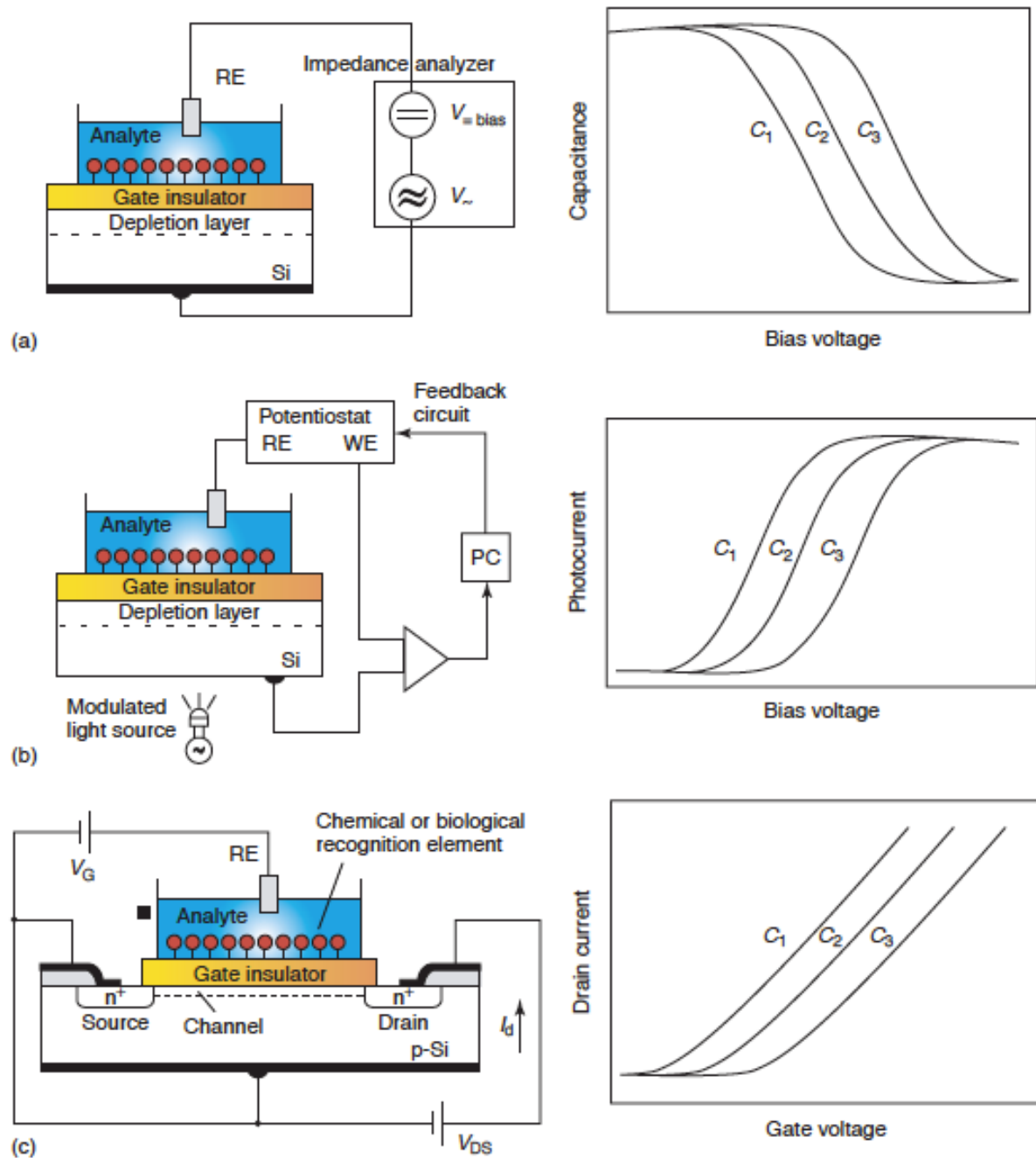
Figure 4.4. Setup and typical sensor response of a capacitive EIS sensor (a), a LAPS (b), and an ISFET (c); RE: reference electrode; WE: working electrode; PC: computer; c1,2,3: analyte concentration. For operation, the gate or bias voltage is applied via a reference electrode (e.g., Ag/AgCl liquid-junction electrode), to set the working point of the sensor as well as to provide a stable potential in the solution. In the case of the EIS sensor setup, an additional small alternating voltage $V\sim(\sim10–50\text{mV})$ is applied to the system in order to measure the capacitance of the EIS sensor. To detect the variation of the capacitance of the depletion layer, the LAPSis illuminated with a modulated light beam, which induces an ac photocurrent to be measured as the sensor signal.

Changes in the chemical composition will induce changes in the electrical surface charge of the gate insulator and in the potential drop at the electrolyte/insulator interface that consequently modulates the current in the ISFET's channel, the capacitance of the EIS sensor or the photocurrent of the LAPS (see Figure 4.4, right). Generally, the following, basic mechanisms of potential generation can be considered: a pH or ion-concentration change, enzymatic reactions, adsorption of charged macromolecules (e.g., polyelectrolytes, deoxyribonucleic acid (DNA)), affinity binding of molecules (e.g., antigen–antibody affinity reaction, DNA hybridization), and potential changes that are coming from living biological systems as a result of more sophisticated (bio-) chemical processes (e.g., action potentials of nerve cells).

4.3.1 pH ISFET

Different oxides and nitrides, like SiO_2, Si_3N_4, SiO_xN_y (oxynitride), Al_2O_3, Ta_2O_5, ZrO_2, SnO_2, TiO_2, WO_3, $PbTiO_3$, AlN, TiN, GaN, Al_2O_3–ZrO_2, and Al_2O_3–Ta_2O_5 double oxides, hydrogenated diamond, hydrogenated amorphous Si (a-Si:H), and so on, have been proven as pH sensitive material for FEDs. However, sometimes these results have been "rediscovered" from results that have already been obtained more than 10–30 years ago. At present, Si_3N_4, Al_2O_3, and Ta_2O_5 serve as pH-sensitive gate insulator materials in commercial ISFETs. Other more exotic materials such as AlN, TiN, $PbTiO_3$, WO_3, and so on, sometimes show nearly Nernstian sensitivity, but have been only rarely studied. Ta_2O_5 is considered as the best pH-sensitive material for field-effect sensors, combining a practically ideal Nernstian pH sensitivity, minimal drift, and hysteresis. In addition, it seems that Ta_2O_5 is also the best corrosion-resistant pH-sensitive material.

4.3.2 Enzyme-modified FETs (EnFET)

Table 4.2 Recently developed EnFETs with enzyme system used and analyte to be detected

Analyte	Enzyme system	
	Glucose oxidase	
Glucose	Glucose oxidase/MnO_2 powder	
	Glucose oxidase/MnO_2 nanoparticles	
	Glucose oxidase/SiO_2 nanoparticles	
Urea	Urease	
Penicillin	Penicillinase	12,
Creatinine	Creatinine deiminase	
Organophosphate compounds	Organophosphate hydrolase	
Fluorine-containing organophosphates	Organophophorus acid anhydrolase	
Phenolic compounds	Tyrosinase	
Glycoalkaloids	Butyryl cholinesterase	

Most reported enzyme-modified field-effect transistors (EnFETs) are built-up of pH-sensitive ISFETs, where hydrogen ions are produced or consumed by the enzymatic reaction. A multitude of EnFETs differing in their sensor design or gate material, enzyme–membrane composition, or immobilization method have been reported for the detection of glucose, urea, penicillin, organophosphorus pesticides, creatinine, phenolic compounds, glycoalkaloids, and so on. Some recently developed EnFETs are summarized in Table 4.2

In general, the main disadvantage of field-effect transistor (FET)-based urea and glucose biosensors for biomedical applications is a rather narrow dynamic range (a few millimolars) implying a necessary dilution of the biological sample before the measurement. To extend the dynamic range of a glucose-sensitive EnFET, a glucose-oxidase membrane has been doped with MnO_2 powder that catalyzes hydrogen peroxidase (as by-product of the glucose-oxidation reaction). As a result, the additional product oxygen is produced, which can be recycled for the glucose-oxidation reaction. In this way, the upper detection limit of the biosensor was extended up to 20mM. Hence, the developed EnFET can be applied for the glucose detection in undiluted blood samples. A glucose-sensitive EnFET with an extended dynamic range, a good reproducibility, and stability has been realized by the co-immobilization of glucose oxidase and MnO_2 nanoparticles, where the MnO_2 nanoparticles act as an oxidant rather than a catalyst.

With the same background, a urea-sensitive EnFET with an extended dynamic range of up to 80mM was developed taking advantage of a recombinant urease with a genetically modified enzymatic active site. The concept of in-situ electrochemical generation of OH− ions in the enzyme membrane was used to improve the recovery time of a glucose-sensitive EnFET. A highly sensitive, low detection limit, and long lifetime penicillin-sensitive EnFET was developed.

4.3.3 DNA-modified FEDs

The possibility of a label-free detection of DNA hybridization utilizing FEDs offers a next generation of DNA chips with direct electrical readout for a fast, simple, and inexpensive real-time analysis of nucleic acid samples. Therefore, in recent years, a considerable research effort has been devoted to the label-free electronic detection of biomolecules (DNA, proteins) by their intrinsic molecular charge using FEDs. In most cases, the experimentally observed sensor response is interpreted in a way that during the binding event (hybridization of immobilized single-strand DNA (ssDNA) with its complementary target molecule (cDNA); see Figure 4.5) the charge associated with the target molecule effectively changes the charge applied to the gate of the FED. As a result, the operating characteristics of the FED, that is, the flat-band voltage and capacitance of the EIS sensor or the threshold voltage and drain current of the FET device, will also change.
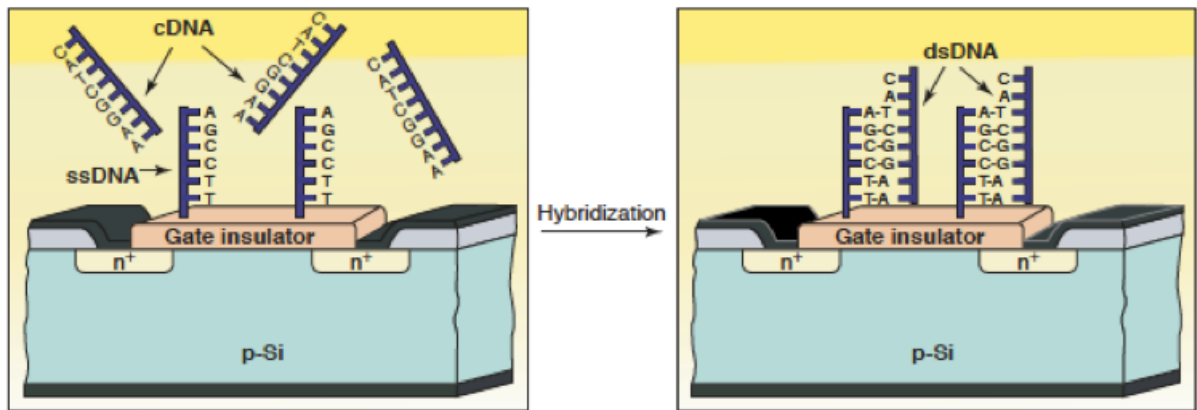
Figure 4.5. Schematic structure of a DNA-modified FET and principle of DNA-hybridization detection (ssDNA: single-stranded DNA; dsDNA: double-stranded DNA; cDNA: complementary DNA)

4.3.5 (Bio-)chemical Sensor as a Physical Sensor
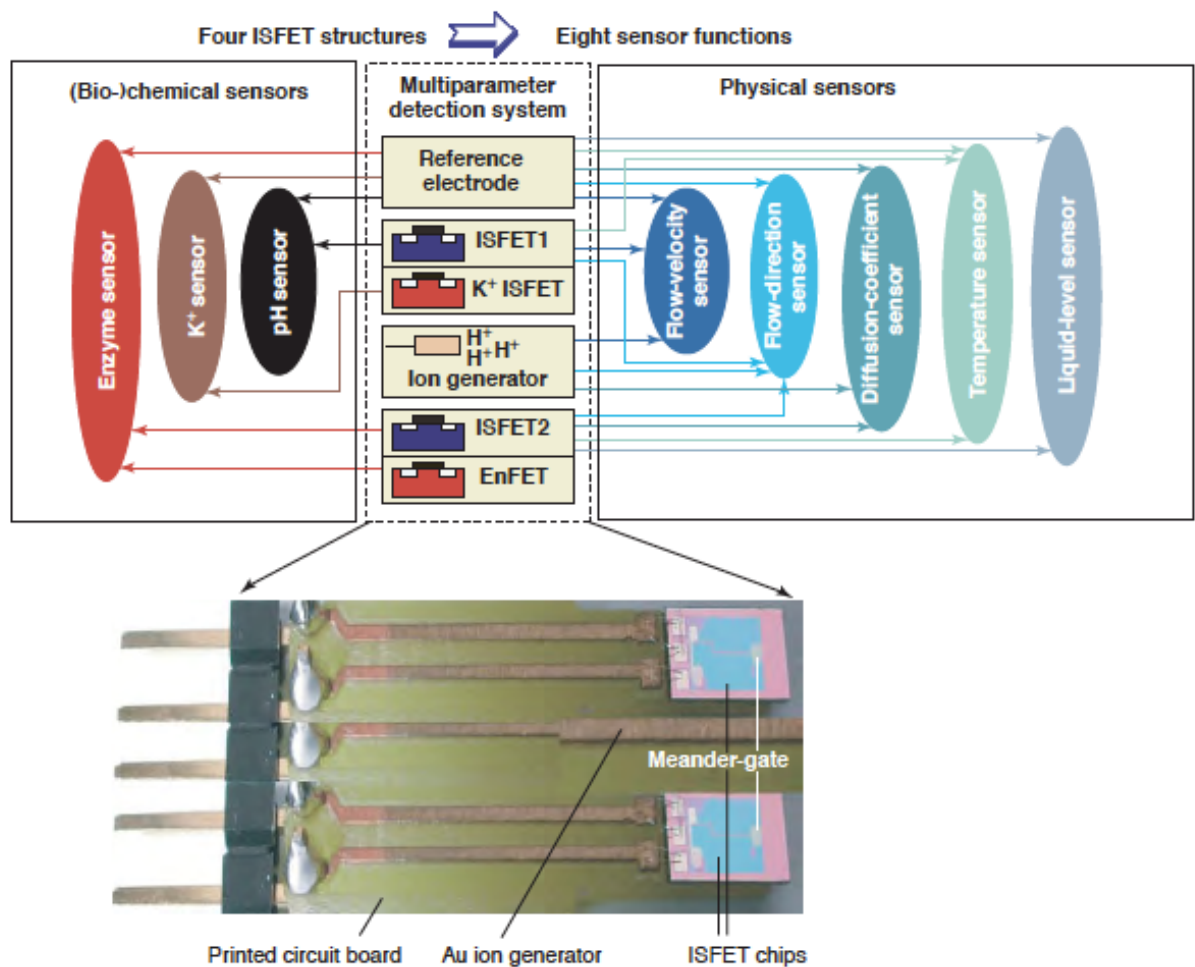
Figure 4.6. Schematic of an ISFET-based multiparameter system for detecting three (bio-)chemical (pH, K+, penicillin) and five physical quantities (temperature,

flow velocity, flow direction, diffusion coefficient of ions, liquid level) by using only four ISFET transducers. In this approach, the same ISFET, which is well known as a (bio-)chemical sensor, also serves as a physical sensor; the multifunctionality is achieved by means of different sensor configurations and/or operation modes.

In multiparameter analysis systems, besides (bio-)chemical parameters it is often necessary to also measure physical parameters of the liquid, like temperature, flow rate, flow direction, and so on. A simple and mostly used way to construct an ISFET-based multiparameter detection system is the integration of an array of ISFETbased (bio-)chemical sensors with different well established single-function physical sensors . In this approach, the same ISFET, which is well known as a (bio-)chemical sensor, also serves as a physical sensor and thus, the amount of obtained (bio-)chemical and physical information can be significantly higher than the number of sensors present in the system ("high-order" system). On the basis of this concept, an ISFET-based multiparameter system for the detection of three (bio-)chemical (pH, K+, and penicillin concentration) and five physical quantities (temperature, flow velocity, flow direction, diffusion coefficient of ions, and liquid level) has been realized using only four ISFET transducers (see Figure 4.6).

4.4. Capacitive EIS sensors

In general, owing to the same functional mechanism, applied materials, and deposition techniques, the sensitive characteristics obtained with EIS and ISFET sensors are quite similar. Therefore, in the following part, we focus on some innovative developments for the fabrication of EIS sensors, like the pulsed laser deposition (PLD) technique, CIP-suitable pH-sensitive materials, porous Si, and the integration of miniaturized EIS sensors with micromachined flow-through microcells.

Capacitive EIS sensors based on porous Si exhibit the advantages of a protected embedment of chemical or biological receptor molecules inside the pores against a fast leaching out, and the enlargement of the effective sensor area due to the porous structure (see Figure 4.7). Porous pH sensors with an n-$Si/SiO_2/Si_3N_4$ structure have been reported: the pH sensitivity (54mV/pH) is in good agreement with results obtained for nonporous (planar) sensors with the same Si_3N_4 layer. Here, the effective sensor surface and thus, the capacitance of the porous pH sensor is larger by a factor of 30 with respect to a comparable nonporous EIS sensor. The possibility to use oxidized porous Si as a transducer material for ion-sensor applications has been demonstrated. Surprisingly, an unusual super-Nernstian sensitivity toward Na^+ and Cu^{2+} ions has been found, that was interpreted by using a quantum effect model. Besides the possibility of miniaturization, porous EIS sensors should offer the advantage of fixing biomolecules inside the pores just by means of a physical adsorption process. For example, two porous Si EIS biosensors for the determination of penicillin and triglycerides utilizing the adsorptively immobilized enzymes penicillinase and

lipase, respectively, have been investigated.
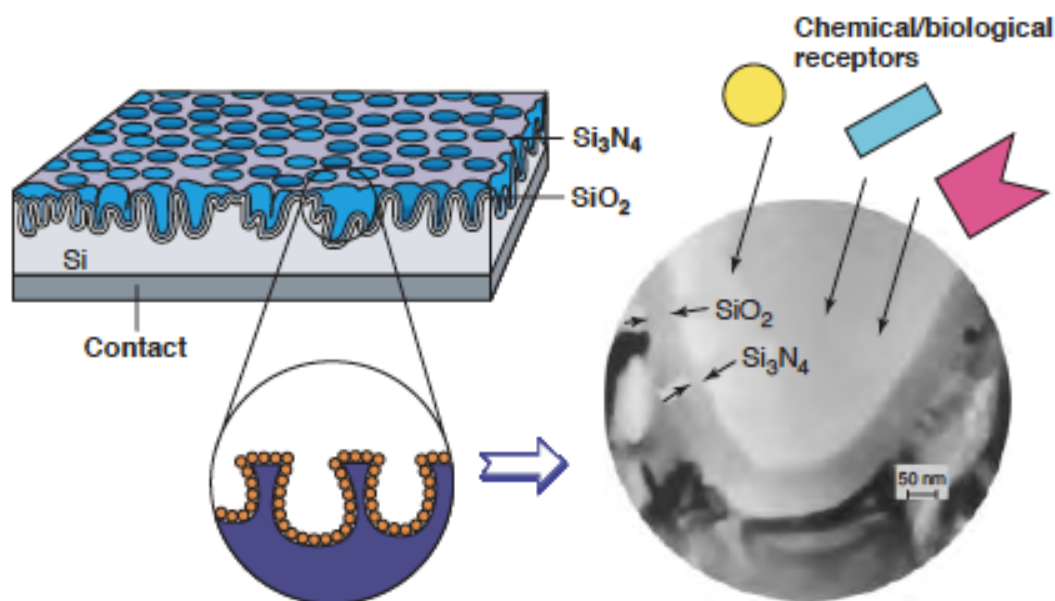
Figure 4.7 Schematic and photo of a macroporous EIS sensor; the mean pore diameter varies between 0.5 and 3 μm.

4.5 Micromachined flow-through cell with wafer-level-integrated capacitive EIS sensor

In literature, two platforms have been applied to integrate EIS sensors into a flow-through cell. In a first and most often used approach, the EIS sensor represents a separate component in a homemade flow-through cell. In this way, for example, pH sensitive and penicillinase-modified Ta_2O_5-gate EIS sensors have been integrated into a flow through cell with a variable internal volume from 12 to 48 μl that is combined with a commercial flow-injection analysis (FIA) system. The second platform favors a monolithic wafer level integration of the flow channel together with the EIS sensor structure. In this case, the sensor represents an integral part of the whole flow-through microcell. Here, a micromachined flow-through microcell with integrated EIS sensor was realized by combining Si and SU-8 technologies; the flow-through micro-channel has been formed in a thick SU-8 layer directly onto an already prepared p-Si-SiO$_2$–Ta$_2$O$_5$ EIS structure (see Figure 4.8). In order to extend the functional possibilities of this microcell, two thin film Pt microelectrodes have been deposited onto the same chip for additional amperometric and flow-velocity measurements. The EIS structures have been integrated in a pH- and penicillin sensitive configuration in flow-through and FIA mode, respectively, yielding a comparable sensor behavior as for the current single pH- and penicillin-sensitive EIS structures.
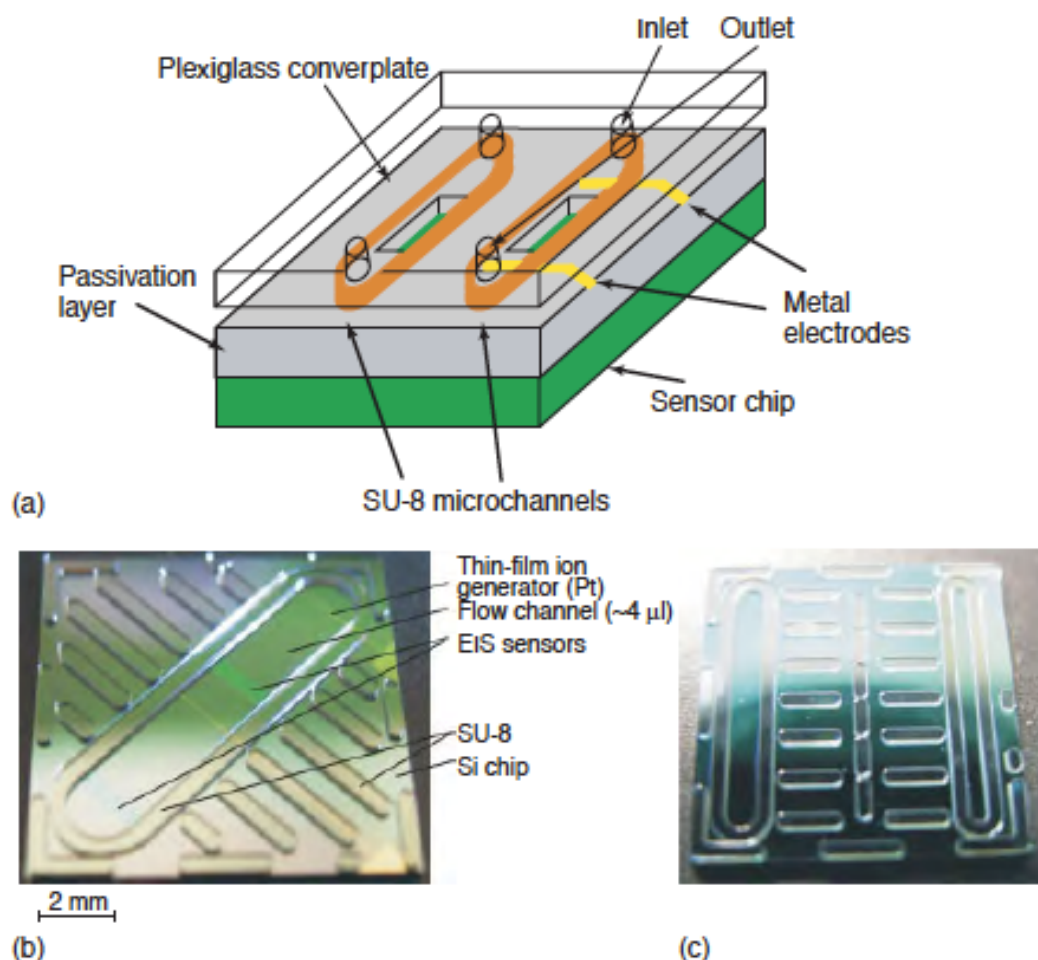
Figure 4.8. Schematic flow-through FIA cell setup of the capacitive EIS sensor (a); micromachined flow-through microcells fabricated by combining Si and SU-8 technologies, where the microchannels have been formed as thick SU-8 layers directly onto the already prepared Si–SiO$_2$–Ta$_2$O$_5$ EIS structures: microcell with EIS sensor and Pt ion-generator (b), and microcell with two separate microchannels and EIS sensor for multisensor and/or differential setup applications (c).

4.6 LAPS

In contrast to the capacitive EIS sensor, where the measured value of the analyte concentration is an average value over the whole sensing surface in contact with the analyte, the LAPS measurement has the advantage of being spatially resolved. The measured area on the sensing surface is defined by the area of illumination, where the ac photocurrent to be measured has been generated. LAPS devices became popular in many chemical and biological applications such as the detection of bacterial growth, the measurement of cell metabolism, the study of mechanisms of drug action on cell physiology, and so on.

Thanks to the structural similarity between LAPS, capacitive EIS sensor and ISFET, many of the sensitive materials, membrane-deposition techniques, and enzyme-immobilization strategies already developed for ISFETs and EIS sensors are also applicable to LAPS devices. Traditionally, the LAPS is employed for pH

recording. Si$_3$N$_4$ is the most frequently used pH-sensitive material in LAPS devices, although several alternative materials such as PLD-deposited Ta$_2$O$_5$ and Al2O3 have also been proved as alternative pH-sensitive materials.
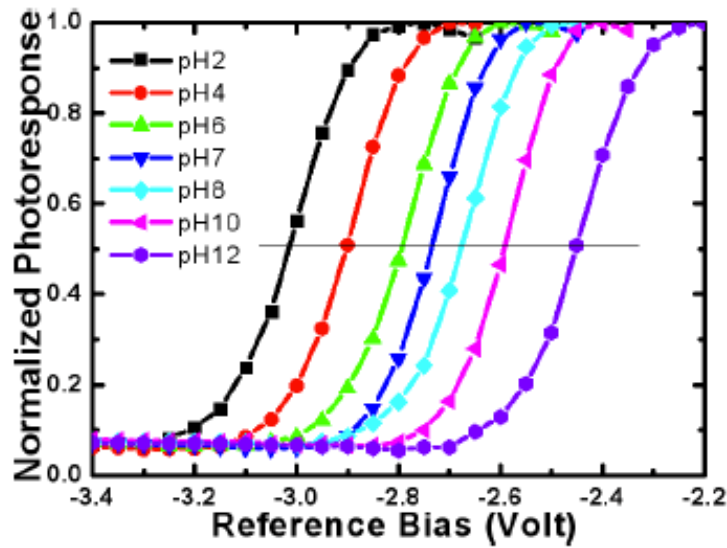
(a)
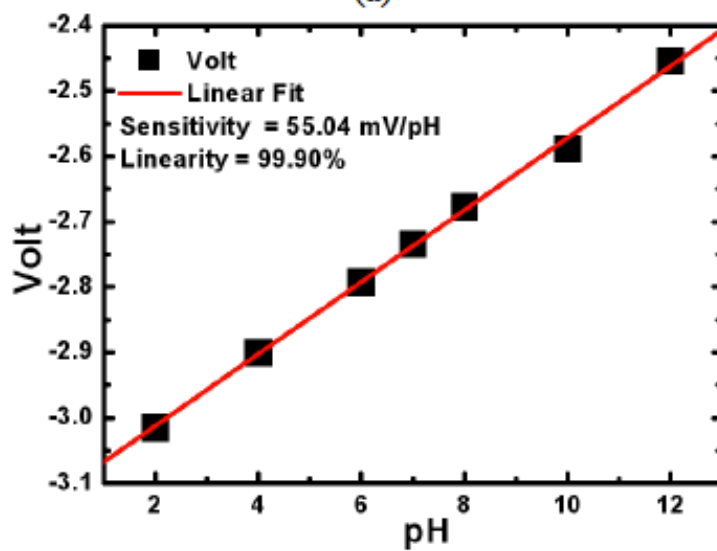
(b)
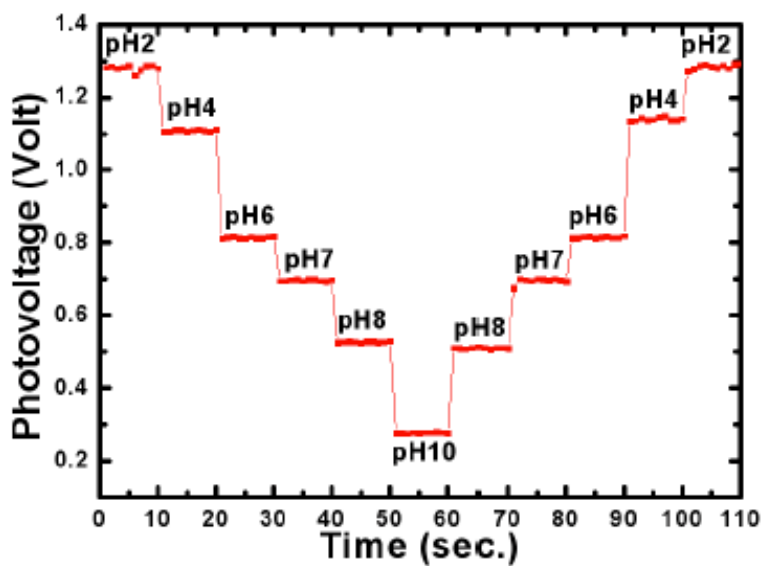
(c)

Figure 4.9 (a) Normalized photoresponse vs. reference bias characteristic of the LAPS chip for different pH buffers with a light spot size of 50X50 μm. (b) Sensitivity and linearity of the LAPS chip for measurement of pH values ranging from 2 to 12. (c) Constant bias measurement of the LAPS chip for different pH values.

      To test the pH sensing performance of LAPS chip with the present set-up, a square shaped white spot of size 50x50 μm was used to illuminate the LAPS chip. The amplitude of the photoresponse was measured as a function of the bias voltage for buffer solutions with different pH values. The corresponding measurement plot is depicted in fig. 4.9(a), which shows the shift of the photoresponse vs bias curve along the voltage axis as the pH value changes. According to the site-binding model, a potential difference appears at the interface of the sensing membrane (Si_3N_4) and the solution that depends on the concentration of H+ in the electrolyte. The larger concentration of H^+ in solution leads to a larger potential difference, resulting in a shift of the photoresponse vs bias curve towards the negative side of the bias axis as the pH value decreases. To calculate the sensitivity and the linearity, the curve shift was plotted against the pH value as shown in fig. 4.9(b). The sensitivity of the measurement can be estimated by using following equation:

$$Sensitivity = \Delta V/pH$$

where ΔV is the change in the reference bias due to a change in pH at a constant photovoltage. The calculated sensitivity is approximately 55 mV/pH and has a linearity of 99.9%, which is in good agreement with the ideal Nernstian value. For the further application of constructing a chemical image, the result of the constant bias measurement mode becomes highly relevant. This mode is the fastest and suitable for chemical imaging because chemical imaging requires measuring the photosignal for a large numbers of pixels. Fig. 4.9c) shows the measurement result of the constant bias mode. The amplitude change of the photosignal is plotted when the pH value of the buffer solution is varied from 2 to 10.

4.7 Optical Biosensing Techniques

      The development of optical biosensing technology is an extremely important scientific and technological issue for (i) the diagnosis and monitoring of diseases, (ii) drug discovery, (iii) proteomics, and (iv) the environmental detection of pollutants and/or biological agents. Traditionally, a biosensor is derived from the coupling of a ligand–receptor binding reaction to a transduction mechanism.

      In an optical biosensor, either the reaction product effects a significant change in the response of a transducer to incoming light or the reaction produces an optical signal that is sensed by the transducer, possibly after amplification and/or conversion to some other form (Figure 4.10).
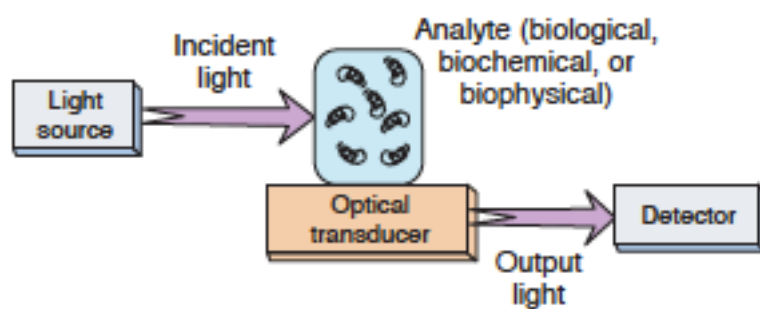
Figure 4.10. Schematic showing the main parts of an optical biosensor.

Table.4.3 Summary of the optical techniques and measured parameters used in biosensing

Technique	Measured parameters
Ellipsometry and polarimetry	Polarization state
Absorption spectroscopy	Intensity, spectrum, polarization dependence
Elastic scattering	Intensity; angular, wavelength, and polarization dependences; correlation length
SPR	Intensity, phase, peak position, polarization
Phosphorescence	Intensity, wavelength, polarization state, lifetime
Raman scattering	Intensity, peak position, polarization state
Guided-wave resonance	Intensity, resonance frequency
Evanescent wave	Mode profile, intensity, spectrum
Interference	Intensity, phase

Several light signals at different frequencies can be sent over the same optical beam because they do not interfere with one another. By measuring differences in wavelengths, arrival times, or polarization states, optical signals can be readily multiplexed and demultiplexed. Some optical techniques, such as fluorescence, have intrinsic amplification in which a single label can lead to a million photons. Surface enhanced techniques such as surface-enhanced Raman scattering (SERS) from molecules located near metallic nanoparticles allow sensing of small concentrations and the ability to recognize specific analytes in the sample. In addition, some optical techniques, such as surface plasmon resonance (SPR) and null ellipsometry, are zero- or blackbackground techniques: the only source of the signal is due to the presence of the analyte species, thereby enabling

high-sensitivity measurements. Finally, optical signals travel in an open path; no wires or other transmitting conduits are needed, although the use of optical fibers may be sometimes necessary for nonoptical reasons and may have certain advantages. This attractive feature enables remote measurements to be made. Several optical-sensing mechanisms exist. These mechanisms are identified in Table 1, along with the variables measured. Several variations are possible on each of these mechanisms as well ascombination of more than one optical measurement in the same setup.

4.8 Total Internal Reflection (TIR) and Evanescent Waves

Several sensing configurations based on TIR exist. They rely on the existence of the evanescent field which is modified when the refractive index in the evanescence region is varied. The phenomenon of TIR is well known in optics. If incident light strikes a specularly smooth and planar interface between two different nonabsorbing and uniform materials (with refractive indices n_i and n_r) at an angle θ_i, as shown in Figure 4.11(a), a fraction of light misreflected and the rest is refracted at an angle θ_r. If $n_i > n_r$, as θ_i is increased there is a value of θ_i for which the refracted wave propagates parallel to the interface (and is called a surface wave). As shown in Figure 4.11(b), then $\theta_r = 90°$ and the incidence angle is called the critical angle:

$$\theta_c = \sin^{-1}\left(\frac{n_r}{n_i}\right)$$

When the angle of incidence exceeds the critical angle, no refraction occurs; instead, the light is reflected back into the medium of incidence. The TIR phenomenon is illustrated in Figure 4.11(c).

(a)　　　　　(b)　　　　　(c)
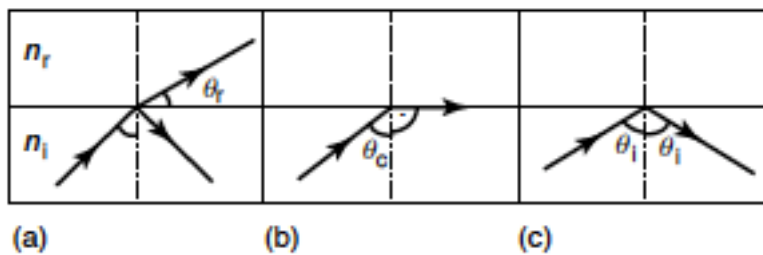
Figure 4.11. Illustration of total internal reflection (TIR) at a specularly smooth and planar interface of two different nonabsorbing materials.

No net flow of energy occurs across the interface under the TIR conditions. In order to satisfy the boundary conditions, the electromagnetic field must penetrate the optically rarer medium (Figure 4.12) and an evanescent wave is then generated with exponentially decaying amplitude:

$$E = E_0 \exp\left(\frac{-z}{d_p}\right)$$

where z > 0 is the distance from the interface, and dp is the penetration depth:

$$d_p = \frac{\lambda}{2\pi n_1 \sqrt{\sin^2 \theta_1 - (n_r/n_1)^2}}$$

The penetration depth is of the order of few hundred nanometers or less, and can be controlled by an appropriate choice of the ratio of refractive indices of the two materials, the angle of incidence, and the wavelength of the incident light.
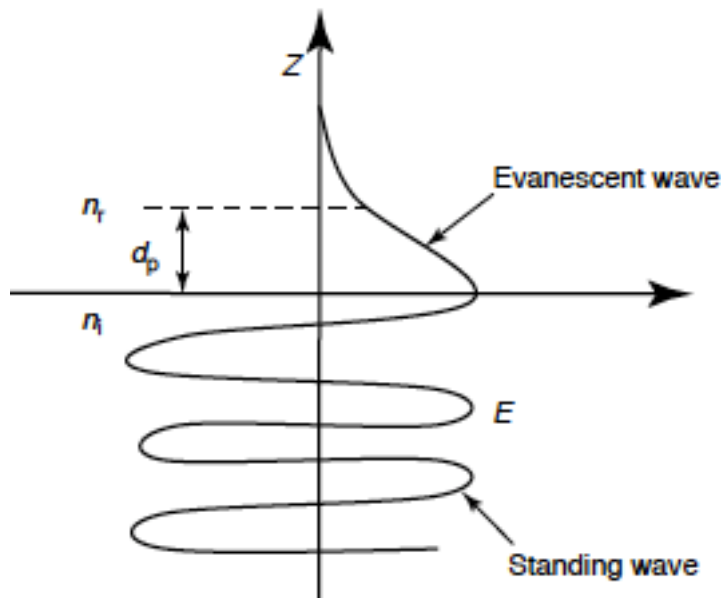
Figure 4.12. Schematic of the field structure when TIR occurs at a smooth planar interface.

4.8.1 Planar Waveguides (PW)

Waveguides are structures that confine and guide electromagnetic radiation (Figure 4.13). A PW, at its simplest, consists of a thin, transparent, dielectric film deposited on a substrate, and a cover may be put on top of the thin film. In order to achieve a true-guided mode in the thin film, the following condition must be satisfied: $n_f > \max\{n_s, n_c\}$. If $n_s = n_c$ then the PW is referred to as being asymmetric; otherwise, it is symmetric. A further requirement is that the incidence angle θ_i must exceed the critical angle. Thus, the incidence angle must satisfy the following criteria: $\theta_i > \sin^{-1}(n_s/n_f)$ and $\theta_i > \sin^{-1}(n_c/n_f)$. Figure 4.14 illustrates the distinctions between radiation modes (no confinement), substrate modes and guided modes (full confinement).
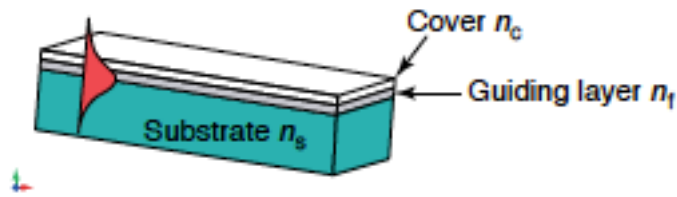
Figure 4.13. Schematic showing a standard PW, along with the transverse intensity profile of a guided mode.
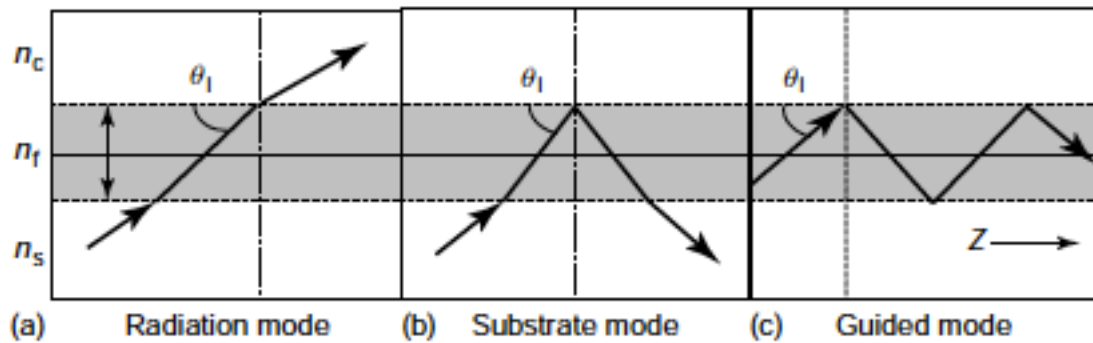
Figure 4.14. Ray patterns of (a) a radiation mode, (b) a substrate mode, and (c) a guided mode in a PW, where $n_f > n_s \geq n_c$.

4.8.2 Planar-waveguide Sensing Configurations

The fact that the (evanescent) fields of the guidedmode extend into the cover and substrate of the PW makes the device useful for sensing purposes. The cover can be coated with a reagent, such as an indicator or recognition element. A chemical reaction within the region of the evanescent field is sensed due to the changes induced by the analyte on the sensor surface, such as scattering, absorption, fluorescence, or changes in the refractive index.

PW sensors are attractive for several reasons. Their small size and flexibility facilitates the miniaturization of analytical instrumentation for in situ measurements. Their availability encouraged the development of remote and real-time continuous sensing in gaseous and liquids without sampling. They are immune to electromagnetic interference and resist hostile and hazardous environments. PW sensors can be fabricated from a variety of materials and can be relatively easily adapted for specific biomedical applications that require biocompatibility, sterilization, disposability, and the capability for in vivo measurements if needed. Light signals in PW sensors can be modulated in several ways to eliminate interference and facilitate the use of internal references. Their adaptability to act as electrodes allows integration with other techniques such as electric-field and magnetic-field techniques to provide more information, reduce nonspecific binding, shorten the analysis time, and improve the sensitivity.

Asymmetric PWs are useful as direct biochemical sensors, as the TIR

boundary can be between the chemically selective waveguiding layer and the sample. The basic sensor is a thin waveguiding layer (on which the recognition element is immobilized) that is either deposited or engraved on either a polymer or a glass substrate. Air serves as the cover medium in the simplest case (Figure 4.15a).

Typical PW sensors use the high-index waveguide formed by depositing silicon nitride (Si_3N_4), indium tin oxide (ITO), or tantalum pentoxide (Ta_2PO_5) on a glass or polymer substrate. The efficacy of such sensors have been demonstrated for label-free and fluorescence detection. The planar waveguide can be a multilayer and theintegrated waveguide SPR biosensor.
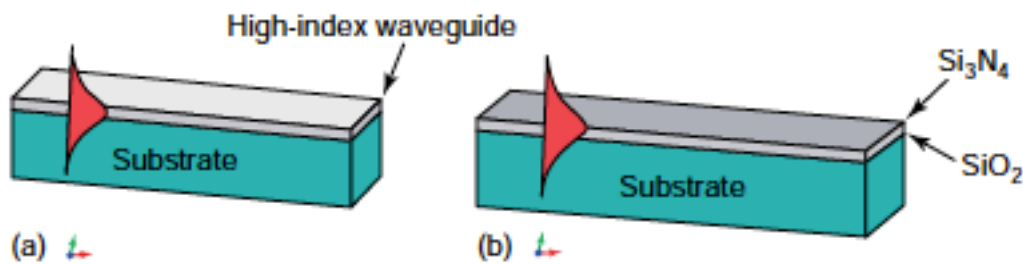
Figure 4.15. Schematics of (a) a single-layer PW (b) and a multilayer PW that is an RM.

Evanescent-field biosensors have proven to be a highly sensitive tool for interactions in the close vicinity of the sensor surface. But, as they generally have penetration depths d_p on the order of 100–150 nm, they are mostly suitable for detection of interactions with small targets such as viruses (10–100 nm), proteins (1–10 nm), and DNA. Detection of large targets such as bacterial cells (0.5–5 µm) and eukaryotic cells (5–50µm) is problematic (Figure 4.16a). Several researchers have attempted to extend d_p up to 1µm (Figure 11b). It s have been used for detecting refractive-index changes, scattering, and fluorescence from bacterial cells captured on an immobilized antibody. Their detection limit is a 1000-fold better than of the conventional evanescent-field sensors based on SPR and RM.
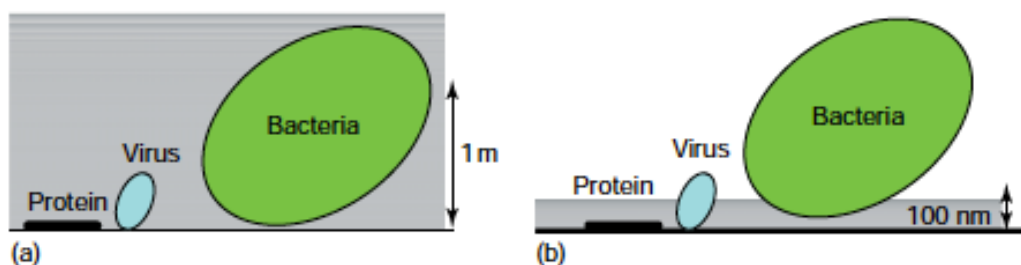
Figure 4.16. Schematics of (a) the evanescent field of a conventional evanescent-field sensor of 100-nm penetration depth and (b) a long-range evanescent field, in comparison to biologically significant targets to be sensed.

4.8.3 Cylindrical Waveguides

The concept of combining optical detection with capillary device is unique as it achieves dual use of the capillary as a light-guiding and a fluidic channel. There are two possible configurations: (i) liquid-core waveguide, in which, the glass or silica capillary can be illuminated by either direct or partial reflections at the interior (Figure 4.18a), and (ii) guiding-wall capillary, wherein an indirect evanescent field arises from the guided mode in the capillary wall (Figure 4.18b). The latter is favored as it guides light strongly and leads to the excitation of an evanescent field uniformly along the entire length of the capillary. In contrast, the liquid-core waveguide has nonuniform illumination along the capillary due to the weak guiding in the core (n = 1.33 for water), which is lower in refractive index than the cladding (capillary). Light guiding in the liquid core is improved by coating the capillary internally or externally with materials with refractive index lower than water, for example, fluoropolymers, Teflon, and nanoporous dielectric materials.
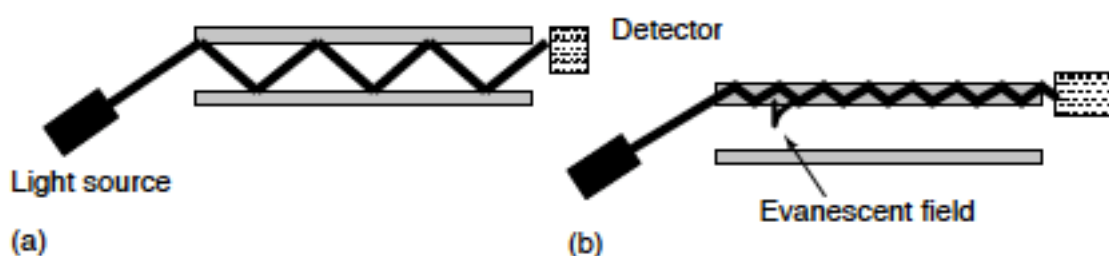
Figure 4.18. Schematics of (a) a liquid-core waveguide and (b) guiding-wall capillary.
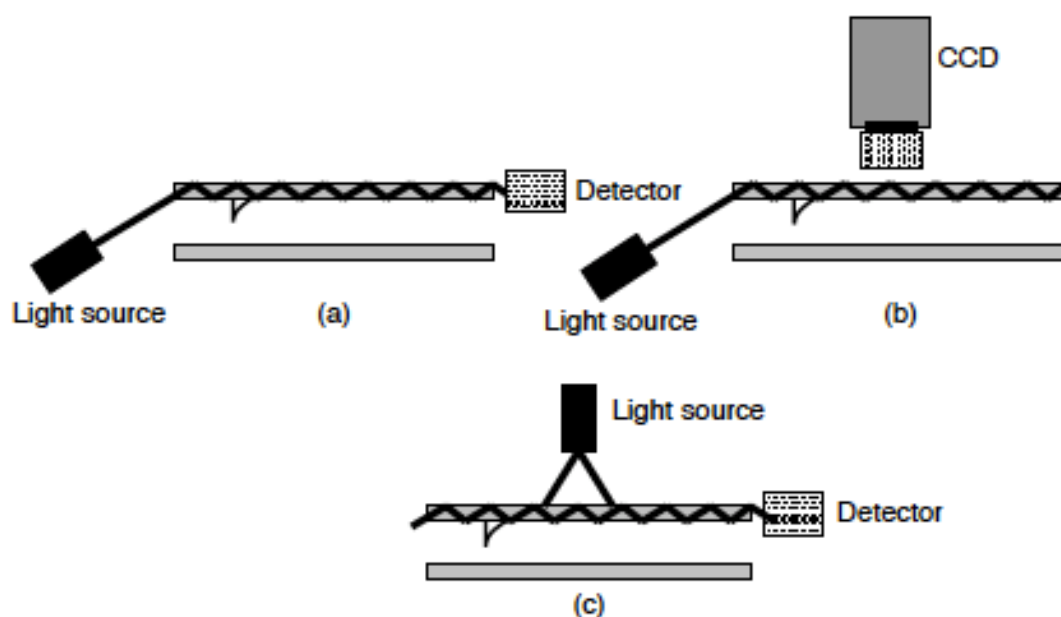
Figure 4.19. Schematics of three setups for excitation and collection of fluorescence emission.

Guiding-wall capillaries have been used mainly for fluorescence sandwich

immunoassays with different excitation and emission probing configurations. In one setup, light is introduced at the end of the capillary and propagates on the inner surface as an evanescent wave. To probe fluorescence for example, a portion of the resulting fluorescence tunnels into the guided mode in the capillary wall and is captured at either or both ends of the capillary (Figure 4.19a). In another setup, which is analogous to a PW, the excitation is evanescent, but the emitted light is collected normal to the capillary axis using a CCD camera (Figure 4.19b).

4.9 Surface Plasmon Resonance (SPR)

SPR is a quantum optical-electrical phenomenon arising from the interaction of light with the free electrons at the metal surface. Under certain conditions the energy carried by photons of light is transferred to collective excitation of electrons, called plasmons, at the interface between a metal and a dielectric. Energy transfer occurs only at a specific resonance wavelength of light when the momentum of the photons and the plasmons are matched. These plasmons are strongly localized electromagnetic waves that propagate along the interface between the metal and the ambient medium, and decay exponentially with penetration distance into an emergent dielectric medium. At the resonance condition, SPR is responsible for a dip in reflectance, which results from the absorption of optical energy in the metal layer.

This wave is extremely sensitive to changes in the refractive index near the metal surface within the range of the plasmon field. This change may result in a shift in the resonant wavelength of the incident light, change in the intensity of the reflected light, or change in the resonant angle of the incident light (4.20). The magnitude of this shift is quantitatively related to the magnitude of the refractive-index change of the dielectric medium.
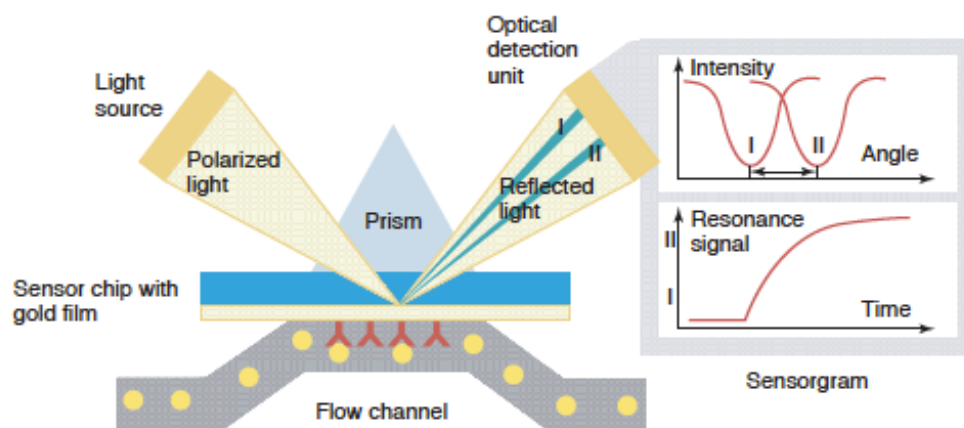
Figure 4.20. Surface plasmon resonance.

4.9.1 SPR configurations

In order to excite SPR by a light wave at a metal-dielectric interface, the incident light's wavevector must match the SPR wavevector, and the polarization

state must be transverse magnetic (TM). To achieve the first condition easily, the incident light's wave vector can be increased in magnitude either by passing the light through a medium with a refractive index higher than that of the dielectric medium at the boundary at which the surface plasma wave (SPW) is to be excited, or by using diffraction effects. Therefore, couplers are needed in SPR sensors.

Prism Couplers

The prism-based excitation of SPRs was proposed by Kretschmann. The Kretschmann configuration (Figure 4.21a) is the most common geometrical setup as it is the most efficient technique for generating the plasmon. In this configuration, a metal film is deposited directly on top of the prism surface.The metal film is illuminated through the prism at an angle of incidence greater than the critical angle for TIR. The light wave undergoes TIR at the interface between the prism coupler and the metal film, and excites the plasmon at the outer boundary of the metal film by evanescent tunneling.

In the Otto configuration (Figure 4.22b), the prism is placed close to a metal surface, so that the photon tunneling occurs through the air gap between the prism and the metal surface. This configuration is useful in the study of SPR with solid media. The Otto configuration is less useful for applications with solutions, since the gap between the metal and the prism is filled with a dielectric, thereby reducing the SPR efficiency.
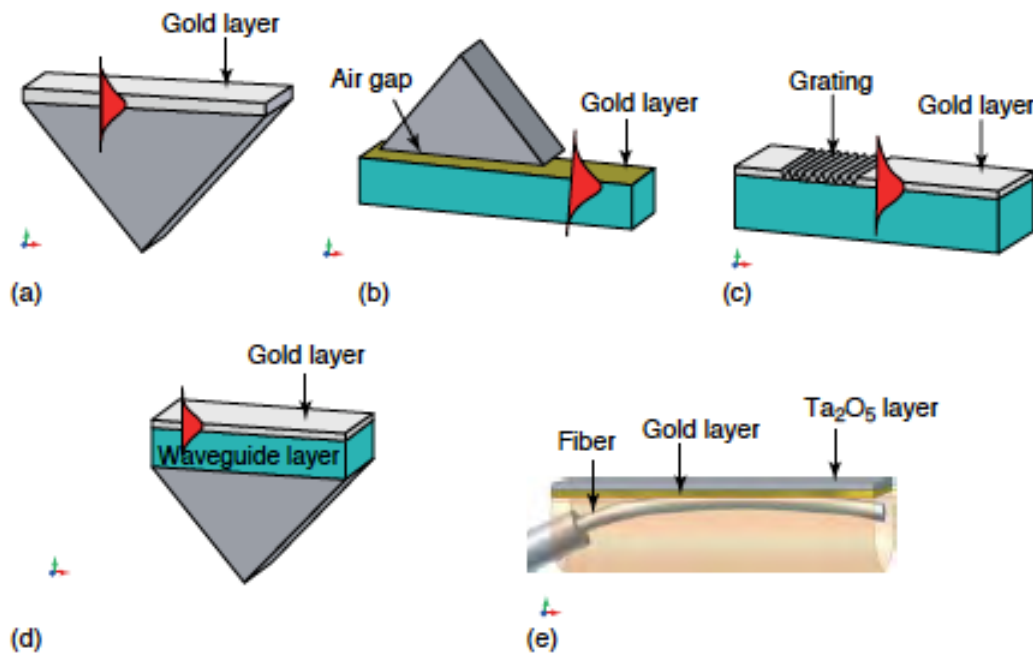
Figure 4.22. Schematics of (a) the Kretschmann configuration of a prism coupler, (b) the Otto configuration of a prism coupler, (c) grating coupler, (d) waveguide coupler, and (e) fiber-optic coupler.

Grating coupling

Another technique to overcome the momentum mismatch is to use a periodically corrugated metal/dielectric interface. The diffracted orders from the

corrugation have wave vectors larger in magnitude than the incident light. A light beam is directed towards a medium whose surface has a spatial periodicity comparable to the wavelength of the incident light, for example, a reflection diffraction grating (Figure 4.22c).

Waveguide Couplers

The process of exciting an SPR using the waveguide is similar to that in the Kretschmann configuration. The light is guided by either a single or multilayer (slab or channel) waveguide to a region with a thin metal overlayer. In that region, the light penetrates evanescently through the metal layer (Figure 4.22d).

Fiber-optic Couplers

SPR sensors with fiber-optic couplers operate using either wavelength- or intensity-interrogation on a formed SPR active sensing area that is located either at the end of the fiber or in the middle of the fiber where the cladding of optical fiber core is partially removed (Figure 4.22e).

4.9.2 Sensing with porous materials and photonic crystals

Porous materials provide several modalities for optical biosensing, provided the porosity is on the scale of a wavelength or smaller. A simple modality emerges by way of homogenization theory. If the length scale of porosity is less than a tenth of the wavelength of incident light, a porous material may be thought of in terms of a nonporous homogeneous material with an equivalent refractive index. The equivalent refractive index depends on the refractive index of the skeleton material and the porosity or void fraction. When this porous material is submerged in a liquid or a vapor that does not react chemically with the skeleton material, the equivalent refractive index must change. Thus, the transmission/reflection spectra of a slab of the porous material will be altered by any infiltrant fluid as well as by any solute molecules in that fluid. If the infiltrant molecules chemically react with the skeleton material, the spectra would also be altered. It is possible the optically sense small organic molecules, 16-nucleotide DNA oligomers, and proteins at pico- and femtomolar analyte concentrations in porous silicon. Insertion of the porous silicon between two dielectric mirrors increases the sensitivity.

If a slab of a porous material possesses a periodic variation of optical constitutive properties, different types of Bragg phenomena can be exhibited by it. The Bragg phenomena could depend on the polarization state and the direction of propagation of the incident light in relation to the illuminated surface of the slab. Any infiltration of the voids could shift the center-wavelength and the bandwidth of a Bragg range. Photonic crystals and chiral sculptured thin films are two candidate materials. (fig.4.23).
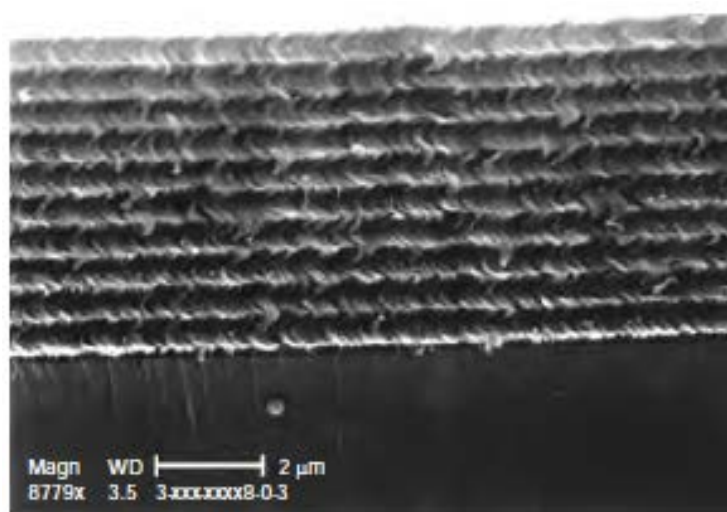
Figure 4.23. Scanning electron microscope image of a typical sculptured thin film prepared by tilted evaporation. This thin film is structurally chiral

4.10 Localized surface plasmon resonance (LSPR) spectroscopy in biosensing

Interaction of electromagnetic waves with the interface between a metal and a dielectric medium can result in excitation of electron density oscillations called surface plasmon (SP) polaritons. In the case of a planar interface the SP propagates along the surface. At interfaces with a sharp roughness on the nanometer scale, as well as in dispersed metal structures, these oscillations are confined to high-curvature areas. These two types of electronic excitations are known as propagating and localized SPs, respectively, and can be monitored using an appropriate experimental setup. Propagating SPs are usually excited in the total internal reflection mode (Kretschmann configuration) using a complicated experimental setup. Observation of localized SPs is considerably simpler and can be done with standard spectrophotometric equipment in the transmission (or reflection) configuration. The conditions for excitation of SPs and light absorption are sensitive to the dielectric properties of the medium adjacent to the metal–dielectric interface.

A major advantage of SP spectroscopy with respect to other detection methods is its inherent label-free nature: the measured quantity is the optical response of the metal nanostructures, which is sensitive to changes in the dielectric constant of the medium.

4.10.1 Solution-based and Immobilized Metal Nanoparticles (NPs)

LSPR biosensing using metal NPs in solution is based on formation of NP–biomolecule conjugates without aggregation; systems based on bioinduced NP aggregation are beyond the scope of this text. Although both Au and Ag NPs have been widely used as markers in biological systems, Au NPs have been the system-of-choice in the preparation of LSPR biosensors. Water-soluble citrate-stabilized

Au NPs have been almost exclusively used in solution-based LSPR biosensing, as they allow adjustment to the particular system requirements as well as tuning of the average size between 5 and 50 nm. Preparation procedures for citrate stabilized Au NPs are well established and available in the literature, providing NPs with a narrow size distribution (size dispersion of ca. 10%). Immobilization of receptor molecules on the NP surfaces is performed by exploiting either electrostatic interactions or exchange of the citrate stabilizer with sulfur-containing receptor molecules the sensitivity of LSPR to the immediate dielectric environment is seen in the spectra of Au NPs in solution and in air (Figure 4.24a, spectra 1 and 2). The increase in the dielectric constant from air ($\varepsilon 1 = 1$) to aqueous solution ($\varepsilon 1 \approx 1.332$) causes a red shift of the maximum of the LSPR band from 514 to 520 nm, in accordance with the Mie theory, although the influence of the solid substrate on the resonance conditions should be taken into account for Au NPs immobilized on glass.

(a)
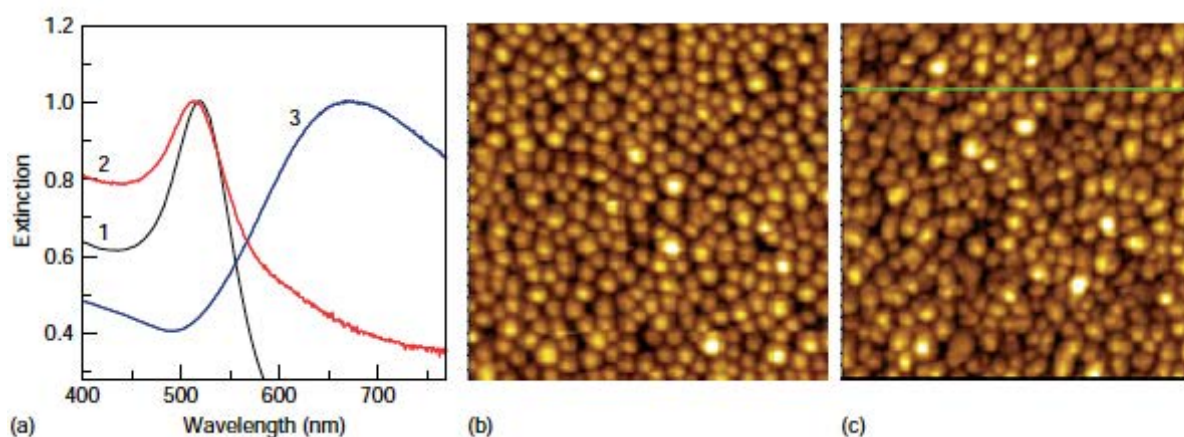
Figure 4.24. Transmission spectra (a) and AFM images (b,c) (600 °ø 600 nm2) of Au nanoparticulate systems. Part (a) shows (1) citrate-stabilized Au NPs (mean diameter, 2 nm) in water; (2) a monolayer of citrate-stabilized Au NPs immobilized on mercaptosilane-modified glass; an image of the monolayer is shown in (b); (3) Au island film (nominal thickness, 5.0 nm) evaporated on aminosilane-modified glass; an image of the island film is shown in (c). Spectra 2 and 3 were taken in air; all spectra are normalized to maximum extinctio

The general scheme for the preparation of a metal island film using evaporation through a mask of densely-packed microspheres is shown in Figure 4.25. Variation of the sphere size and the nominal thickness of the evaporated metal allows control of the metal NP shape and size, as depicted in Figure 4.25. The LSPR band position can be tuned in a rather large wavelength range. The triangle footprint and large interisland separation are thought to be advantageous in sensing applications. As in the case described above of metal evaporation on unpatterned surfaces, thermal annealing is a simple way to affect NP shape.

In the case of metal deposition through a microsphere mask the NPs are hundreds of nanometers in size and annealing effects island reshaping with no substantial island displacement. Upon annealing the well-defined crystallographic habit seen in Figure 4.25(a) disappears and the NPs become roundshaped, while the SP band maximum is shifted to shorter wavelengths by as much as 200 nm.
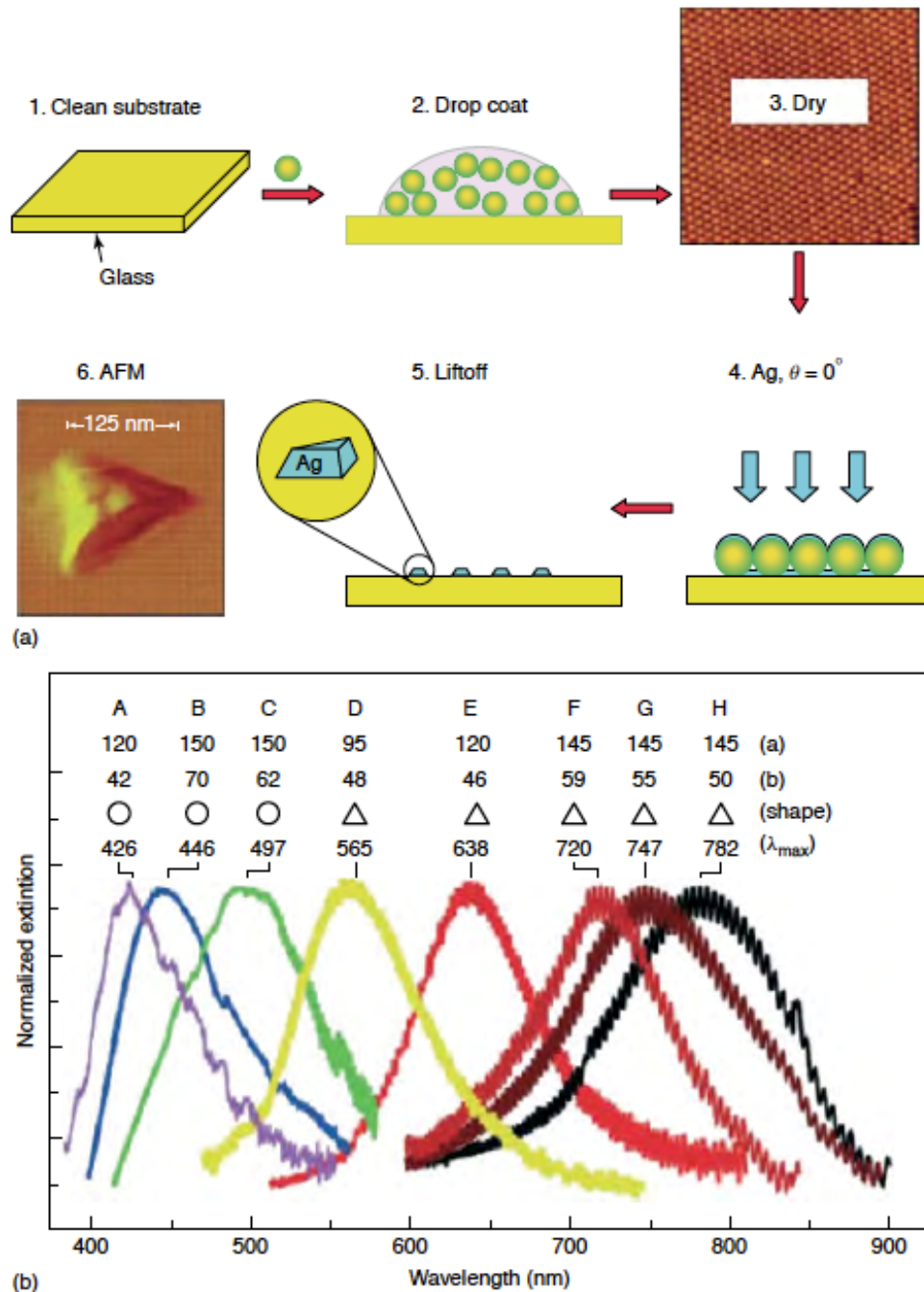
Figure 4.25 (a) Schematic representation of the nanosphere lithography (NSL) fabrication process. Glass and mica were used as substrates. The AFM image in step 3 is 5 - 5µm². (b) Size- and shape-tunable LSPR spectra of various Ag nanoparticles on mica (labeled A–H). The extinction oscillation is caused by light interference in the thin mica substrates. The wavelength of maximum extinction, λmax, changes by varying the in-plane width (a) and out-of-plane height (b) of the nanoparticles.

These rather dramatic changes in LSPR spectra, as well as those described above for the random evaporated metal island films, illustrate the sensitivity of the LSPR extinction spectra to change in metal island shape, presenting a potential difficulty in sensing applications.

Several sandwich-type structures combining a flat, continuous metal film and an island or a rough metal film, as well as a combination of a metal island film and bound NPs, were used as LSPR biosensors. An example of the preparation of a sandwich-type LSPR transducer based on encapsulation of silica beads in Au is shown in Figure 4.26.
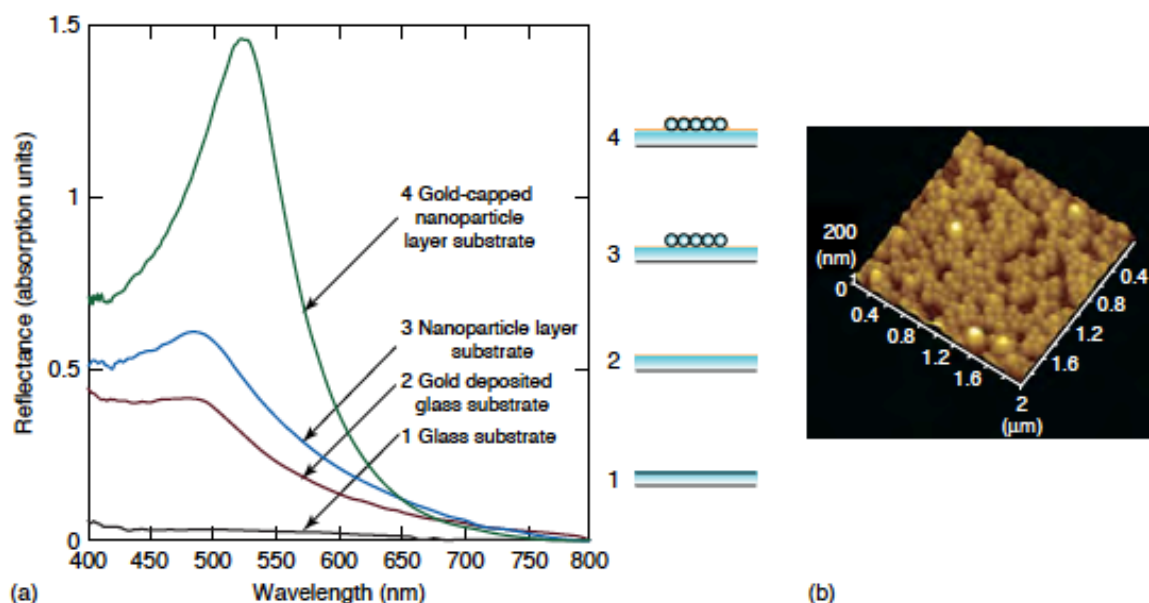
Figure 4.26. (a) Reflectance spectra for the different steps in the preparation of a gold-capped silica nanoparticle layer on a gold substrate. (b) AFM image of the gold-capped nanoparticle layer.

Figure 4.27 shows the optical response of an LSPR transducer to layer-by-layer (LbL) construction of a self-assembled multilayer on Au island films using metal–organic coordination. Evolution of the transmission spectra during incremental increase in the thickness of a thin uniform layer on Au islands (Figure 4.27a,c) is qualitatively similar to that observed for change of the bulk refractive index . The differential extinction spectra (Figure 4.27b,d) show a maximum sensitivity at nearly constant wavelengths red shifted with respect to the maximum of the SP band, as also seen in Figure 4.27. The decrease in sensitivity due to distance dependence of the evanescent field for the 2.5-nm film is evident as saturation after 7–8 layers, corresponding to an overlayer thickness of ca. 11 nm. The decrease in sensitivity for the 5.0-nm films is much less pronounced, extending the response to > 15 nm, with a near-linear region resembling the response to SAMs with variable thickness. Since biorecognition interfaces usually comprise several layers, generally including biomolecules, thicker metal films with better-defined islands—providing higher distance sensitivity—should be preferred

as platforms for LSPR biosensors.
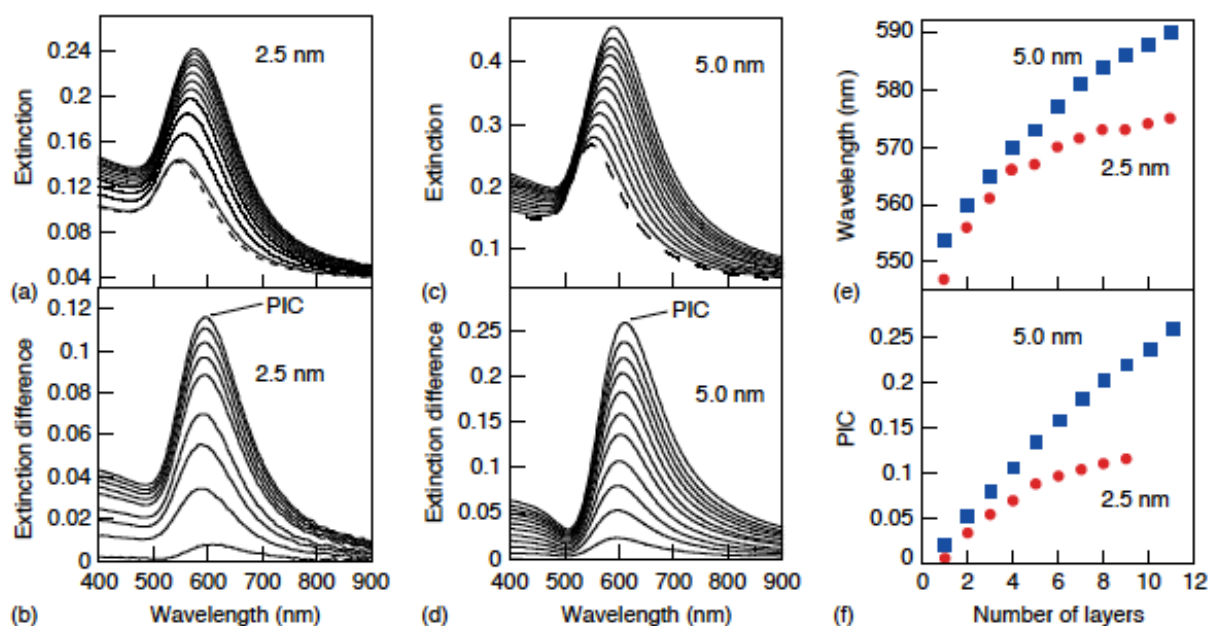
Figure 4.27. Transmission UV–vis spectra of Au island films (annealed at 200 ∘C; nominal thickness indicated) recorded during films. The thickness of the layer deposited on the Au islands increases by ca. 1.4 nm/layer. All spectra were taken in air. Transmission spectra (a,c) and corresponding difference spectra (b,d) are shown. (e,f)Wavelength of maximum extinction and the PIC obtained during film preparation
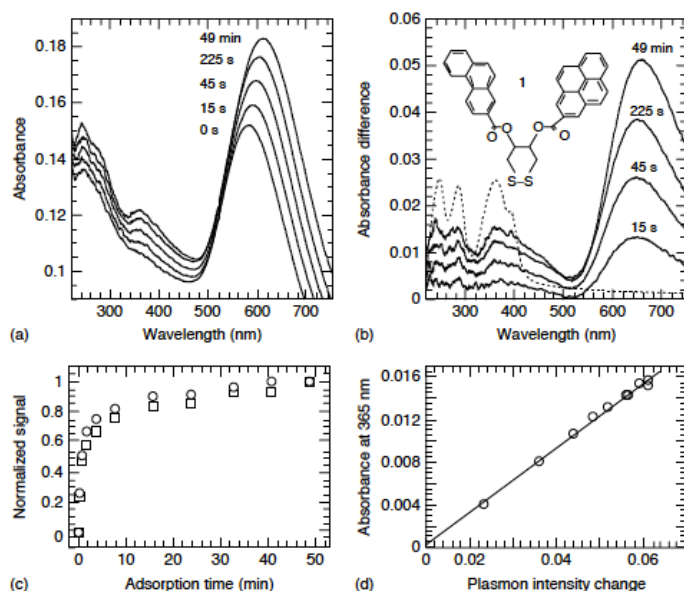
Figure 4.28. Transmission UV–vis spectroscopy for the formation of a monolayer of 1 (inset in b) on an Au island substrate (2.5 nm nominal thickness, evaporated on quartz, unannealed), adsorbed from 2mM solution of 1 in chloroform. (a) Absolute spectra. (b) Difference spectra (obtained by subtraction of the 0-s

spectrum from the other spectra in a); dashed line corresponds to the spectrum of a thick layer of 1, obtained by evaporation of a drop on quartz (original spectrum divided by 6). (c) Normalized quantities of the plasmon absorbance versus adsorption time: squares—shift of the wavelength of maximum absorbance (from a); circles—intensity of the maximum difference peak (plasmon intensity change, PIC) (from b). (d) Correlation between the PIC and the maximum absorbance of 1 at 365 nm (both from b). All UV–vis spectra were taken ex situ, after rinsing the sample with chloroform and ethanol and drying under a nitrogen stream.

A basic question in LSPR spectroscopy concerns the quantitative relationship between surface coverage by the analyte (in the submonolayer regime) and the LSPR response. We addressed this issue by studying the sensitivity of the SP band to change in the fractional coverage of a self-assembled chromophoric monolayer on an Au island film up to a full monolayer coverage. Au island films similar to the one presented in Figure 4.24(c,d) were used as LSPR transducers, and the disulfide chromophoric molecule 1 (Figure 4.28b)15 or Co-tetraphenylporphyrin (Co- TPP)14 (not shown) served as a target analyte. The spectral separation between the molecular absorption band and the Au SP band allowed independent determination of the fractional surface coverage and the SP change during formation of a SAM. The results in Figure 4.28(a,b) show a clear correlation between the two absorption bands, increasing simultaneously during formation of a SAM of 1. As in the previous examples, the maximum change in SP extinction (the PIC) is red-shifted with respect to the maximum SP absorbance. The kinetics of the shift of both the wavelength of the SP band and the PIC are similar (Figure 4.28c), showing saturation upon completion of monolayer adsorption. The PIC and the chromophore absorption, the latter directly proportional to the fractional coverage by the SAM, are linearly correlated over the entire range of surface coverage (Figure4.28). The same linearity applies to the shift of the wavelength of the SP band (not shown).

Immobilization of biological receptors on oxide surfaces can be carried out by converting an amine-terminated siloxane SAM to a carboxylateterminated surface followed by coupling of a biomolecular receptor through an amide bond. This scheme and corresponding transmission spectra are shown in Figure 4.29 for the preparation of an immunoglobulin (IgG) protein recognition interface on silica-stabilized Au islands on glass. Specific recognition was demonstrated in experiments where bound mouse-IgG and rabbit- IgG antigens were exposed to the corresponding specific and nonspecific IgG antibodies, showing highly specific protein recognition (Figure 4.30).
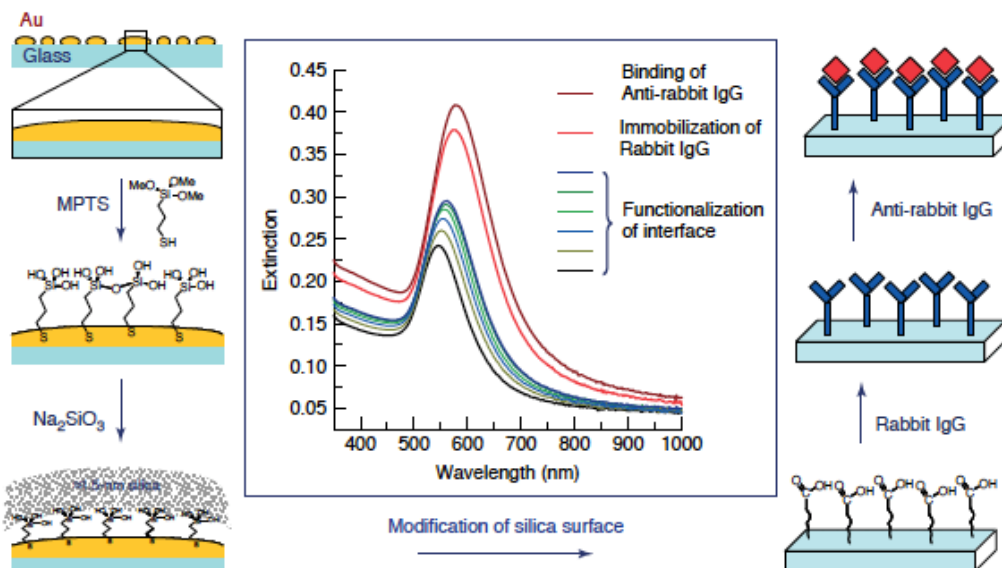
Figure 4.29. Schematic chart showing the preparation of a protein recognition interface using a T-LSPR transducer based on a silica-encapsulated, 5.0-nm Au island film evaporated on silanized glass and annealed, followed by specific protein recognition. Inset: actual transmission UV–vis measurements, carried out ex situ
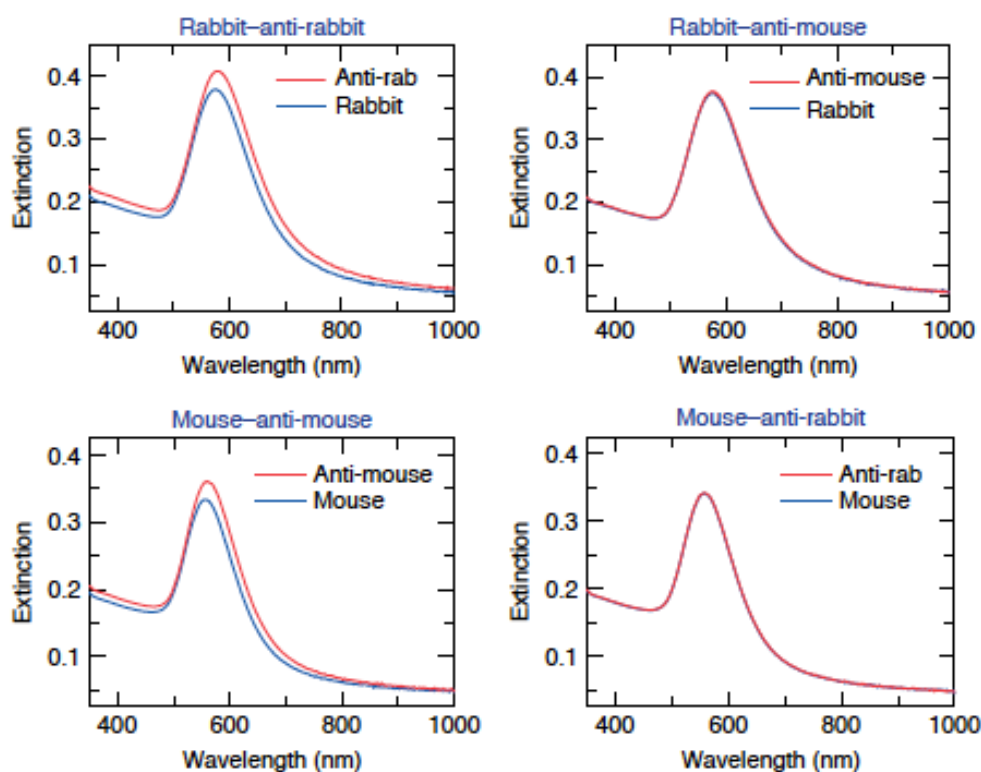
Figure 4.30. Specific recognition of IgG antibodies using an array configuration of T-LSPR sensors with bound IgG antigens. Preparation of the biorecognition interfaces and experimental conditions as in Figure 4.29.
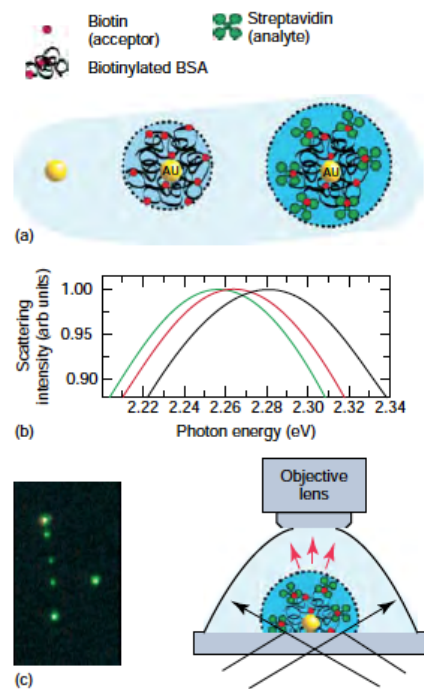
Figure 15. Schematic representation of a biosensor based on light scattering from single Au NPs. (a) Au NPs are functionalized with biotinylated BSA which subsequently binds streptavidin. (b) Mie-type calculations for the three situations shown in (a). (c) Left: true-color photograph of a sample comprising functionalized Au NPs, obtained using dark-field illumination. Right: experimental setup enabling dark-field microscopy of single Au NPs immersed in a liquid.

An example of a biorecognition experiment using single NPs as LSPR transducers is shown in Figure 4.31 Au NPs (40 nm diameter) were functionalized with biotinylated bovine serum albumin (BSA) and the sample was exposed to a solution of streptavidin.

Flow acceleration MEMS

12.1 Introduction

An electrostatic fluid accelerator (EFA) is a device which pumps a fluid such as air without any moving parts. Instead of using rotating blades, as in a conventional fan, an EFA uses an electric field to propel electrically charged air molecules. Because air molecules are normally neutrally charged, the EFA has to create some charged molecules, or ions, first. Thus there are three basic steps in the fluid acceleration process: ionize air molecules, use those ions to push many more neutral molecules in a desired direction, and then recapture and neutralize the ions to eliminate any net charge.

The basic principle has been understood for some time but only in recent years have seen developments in the design and manufacture of EFA devices that may allow them to find practical and economical applications, such as in micro-cooling of electronics components.

Improvements in the cooling capacity of conventional rotary fans have stagnated due to the increasing fluid viscosity as the channels between heat sink fins become narrower. Dynamic velocity profiles, comparable air velocity, and a decreased boundary layer make corona pumps a viable alternative to conventional methods of microelectronic cooling. Liquid cooling is a prevalent alternative process that has been researched for improved convection in heat pipes and micro-channels. Recently, this process has found commercial applications, for example, in IBM's Think Pad laptop line. The advantage offered by liquid coolants is their high specific heat, which enables

an intensive thermal transfer out of a system with a relatively small increase in coolant temperatures. This advantage must be weighed against the need for exotic high dielectric fluids for insulation and structures for fluid containment.

The corona air pump increases the potential thermal transfer of the device, while maintaining air as the coolant In the present study, corona discharge is utilized as the driving mechanism for a high air velocity electrostatic pump. A corona pump prototype was fabricated for analysis and optimization of electrical characteristics and airflow profiles. Figure 1 illustrates the operation of this device. Gas molecules surrounding the corona electrode are ionized by the high intensity electric field at the tip, forming an ion stream between the corona and collector electrodes, thus creating airflow. Ultimately, corona pumps allow for airflow generation with low noise, static filtration, air sterilization, and no moving parts. These ideal characteristics of corona air pumps give rise to a variety of applications as air handling systems. The principle of ionic air propulsion with coronagenerated charged particles has been known nearly as long as electricity. As the demand for enhanced cooling grows, corona driven electrostatic air pumps continue to be investigated for improved thermal transfer.

Rapidly shrinking electronics and MEMS devices generate heat from electrical resistance, mechanical friction, and even combustion processes. Existing thermal management solutions have inadequate heat removal and energy efficiency characteristics to meet industry needs. Conventionally, heat dissipates away from a hot substrate through a thermally conductive heatsink. Fluid flow then removes this heat to the atmosphere. The traditional rotary fan fails to offer the performance and flexibility demanded by advanced thermal management applications. High speed moving parts produce unwanted vibration and noise and are difficult to scale down. Resulting cross sectional air velocity profiles are static and experience a dip in the center opposite the motor assembly. Electrostatic fluid acceleration which ionizes and accelerates an atmospheric fluid in a desired flow pattern is a promising replacement for the removal, or fluid flow, stage in micro cooling. EFA technology has been around for some time, and has been investigated for use in air filtration, propulsion, humidification, propulsion, and audio speakers. For the cooling of microelectronic and MEMS devices, EFA offers several important advantages over the rotary fan: near laminar air profile with controllable velocities, simple solid-state electrode geometry which offers excellent design flexibility and ease of manufacture, low power consumption, large convective heat transfer, and the possibility to decrease the effective boundary layer at the solid fluid interface within an enclosed structure like a channel. The mechanisms of EFA action by corona discharge based ionization have been previously investigated and modeled with respect to electrostatic, fluid dynamic, and space charge effects.

12.2 Physical principles of work

The mechanism of EFA action is illustrated in Figure 12.1. Voltage is applied between a high tip curvature corona electrode and a low tip curvature collector electrode. High electric field intensity in the vicinity of the corona electrode ionizes air molecules. These ions are accelerated towards the collector, transferring kinetic energy to surrounding neutral molecules. The EFA has three regions of operation. At low levels of applied voltage, the EFA is cut-off and has zero current. At some onset voltage the EFA current begins to increase exponentially with applied voltage. The EFA operates between the onset voltage and the breakdown voltage, where dielectric breakdown occurs between the corona and collector electrodes, causing sparking between the electrodes and poor net fluid flux. A wide range of different air velocities and voltage characteristics have been reported for different devices. The variables of import are the corona surface area, air gap size, and corona electrode curvature. Experiments at the macro and meso scale have investigated the variation of these parameters and what follows is the application of understood optimizations to micro devices.
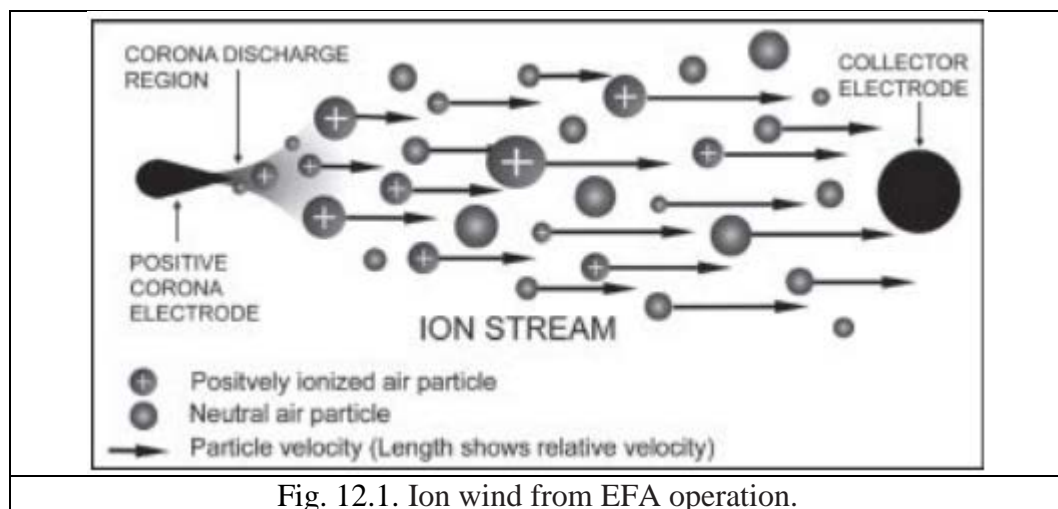
Fig. 12.1. Ion wind from EFA operation.

Ionization

Giving a charge to air molecules is a process known as ionization. Ions are air molecules that have a net electrical charge. Air under normal circumstances has no net charge. Whenever a charge imbalance does occur, the natural attraction of positive and negatively charged ions tends to eventually cancel this charge out, as they attract and combine. Static shocks are an example of this, as is lightning.

It is possible to ionize air artificially, and there are many methods for doing so, as is done for example in arc welding and light bulbs. However many of the methods known to science do not operate in conditions which are conducive to everyday uses; for example, very high temperatures or very low pressures might be required. Or as in light bulbs, specialized materials and gases may be used and extraneous light and heat might be produced.

Because of these restrictions most applications of EFA have relied on a process known as the corona discharge, which has a number of attractive characteristics. It requires no exotic materials, temperatures, or pressures. It works using air at normal levels of humidity and at normal temperatures. It doesn't produce significant negative side effects such as heat or light. It also requires only fairly simple electrical principles in order to function, and uses only low electrical currents, making it relatively safe.

Production of thrust

Ionization is only the first step in electrostatic fluid acceleration. Once the ions are created they have to be used to generate a thrust. This process relies on the same electrodes and electric field as the corona process.

The ions that have been created have a net charge. Because of this charge they will be repelled from the electrode having the same charge and simultaneously attracted to the other electrode with the opposite charge. However, in between these electrodes are enormous numbers of neutral air molecules that have been unaffected by the ionization process. To reach the attracting electrode the ions must literally push their way through the mass of neutral molecules. In doing so, they tend to push those neutrals along with them; this is the process which results in a thrust.

It is important to note that because the ions are continuously propelled by the electric field they can continue to bump into and accelerate neutral ions the entire distance between the electrodes. This is possible because they are given potential energy by the electric field. The electric potential energy of the ions is converted to kinetic energy of the neutrals in each collision. This is the mechanism whereby electrical energy is used to do work by accelerating the neutral air. Some energy is also wasted of course, by slightly raising the temperature of the air and the electrodes, and increasing motion of the molecules in unwanted directions.

Once the ions reach the attracting electrode most of them will lose their charge, i.e., by gaining an electron from the electrode. The fraction of the ions that do not collide with the attracting electrode will tend to be drawn back (up stream) to the attracting electrode. This causes the EFA device to be, more or less, being "driven with a foot on the gas and a foot on the brake." An alternating driving voltage of the right frequency can, in principle, minimize this effect. The neutralized molecules may bounce off the attracting electrode in any random direction. The neutral molecules are not influenced by either electrode and thus their net flow is unaffected as they exit the EFA device.

Physicists and engineers have developed mathematical models for some aspects of corona ionization and fluid acceleration; but in general, due to its complexity there is no general-purpose model which can predict exactly what will happen under any given set of circumstances. For example, air temperature, humidity, electrode shape, and airflow all can affect the exact amount of

energy required, the number of ions generated, etc. Because of these difficulties, developments in EFA have relied on experimentation more than modeling to fine-tune and refine ionization designs.

12.3 Parameters

The velocity of the air at the outlet of EFA is limited by three main factors:

1. Device size
2. energy density released in the volume of the device
3. efficiency of conversion of input electrical energy into the kinetic energy of the moving air.

Each of these factors suggests a possible class of improvements to the canonical needle-to-plane and wire-to-rod geometries of ionic pumps.

Device size

The size of the EFA device limits the amount of airflow that can be generated. A possible improvement involves layering or stacking a series of EFA devices to increase the airflow. Existing limitations and focus areas in this approach are that larger device sizes are inconvenient, especially for smaller applications such as netbooks. There can be interference between the stages, where the attractor electrode of the first stage and the ionizing electrode of the next stage produce an unwanted corona effect, which can actually result in reverse airflow. Alternatively, sparking can also occur if the layers are placed too close together. It is believed that careful design of the geometry of the devices can improve on all these issues.

One way to increase the air velocity at the outlet of the EFA is to stack sev eral corona-collector pairs in the direction parallel to the air stream. A straight-forward stacking of several devices leads to several undesired side effects, and, ultimately, inability to achieve theoretical optimal performance. One of these side effects is excessive size requirements for multi-stage EFA devices, since several stages of EFAs, placed in succession, require a substan tial length along an air duct (i.e., along air flow direction). This lengthy duct further presents greater resistance to air flow. When stages are placed close to each other, the reduced spacing between the stages may produce a strong electric field between an attractor (or collecting) electrode of one stage and a corona discharge electrode of an adjacent next stage. This electric field may induce corona that will result in a reversed air flow. Moreover, due to the electrical capacitance between the neighboring stages, there is a parasitic current flow between them. This current is caused by non-synchronous high voltage ripples or high voltage pulses between neighboring stages. Yet another problem with multiple stage designs is that each separate (or groups of) stage(s) is equipped with its own high voltage power supply (HVPS). In this case, the high voltage required to create the corona discharge may lead to an unacceptable level of sparks generated between the electrodes. When a spark is generated, the HVPS must completely shut down for some period of time required for deionization and spark quenching prior to resuming its operation. As the number of electrodes increases, sparks

are generated more frequently than with one set of electrodes. If one HVPS feeds several sets of electrodes (i.e., several stages), then it is necessary to shut down more frequently to extinguish the increased number of sparks generated. That leads to an undesirable increase in power interruption for the system as a whole. To address this problem, it may be beneficial to feed each stage from its own dedicated HVPS. However, using separate HVPS requires that consecutive stages be more widely spaced to avoid undesirable electrical interactions caused by stray capacitance between the electrodes of neighboring stages and to avoid production of a back corona. The EFA designs used for experiments discussed later in this paper employ innovative solutions to increase airflow by closely spacing EFA stages while minimizing or avoiding the introduction of undesired effects. The improved performance is achieved through a combination of electrode geometry, placement of the stages with respect to each other, and voltage pattern applied to the electrodes. The corona electrodes and collecting electrodes are positioned parallel to each other extending between respective planes perpendicular to an airflow direction. All electrodes of neighboring stages are parallel to each other, with all the electrodes of the same kind (i.e., corona discharge electrodes or collecting electrodes) placed in the same parallel planes that are orthogonal to the planes where electrodes of the same kind or electrodes edges are located. The stages are spaced to avoid or minimize any corona discharge activity between the electrodes of neighboring stages. The experimental studies were conducted to determine the minimum possible distance between the neighboring stages. The distance between the corona discharge wire of one stage and the closest part of the neighboring stage should be a factor of 1.2 to 2.0 larger than the characteristic distance of the single stage a_n. Assuming voltage V_1 applied to 4 the corona electrode and voltage V_2 to the closest accelerating electrode of the same stage, the characteristic distance is defined here as $a_n = |(V_1 - V_2)/E_n|$, where E_n is the onset electric field intensity in the vicinity of the corona electrode for a given geometry of the accelerating stage.

Finally, voltages applied to neighboring stages should be synchronized and syn-phased. That is, a.c. components of the voltages applied to the electrodes of neighboring stages should rise and fall simultaneously and have substantially the same waveform and magnitude and/or amplitude. Operating in such a synchronous manner between stages, electrical potential differences between neighboring electrodes of adjacent EFA components remains constant and any resultant stray current from one electrode to another is minimized or completely avoided. Synchronization may be implemented by different means, but is accomplished most easily by powering neighboring EFA components with respective synchronous and syn-phased voltages from corresponding power supplies, or with power supplies synchronized to provide similar amplitude a.c. components of the respective applied voltages. This may be achieved with the same power supply connected to neighboring EFA components or with different, preferably matched power supplies that produce synchronous and syn-phased a.c. component of the applied voltage.

Energy density

Higher velocity airflow can be generated if more ions are generated by the ionization process; a greater density of ions would mean that a greater number of neutral molecules are pushed along. However the maximum voltage which can be applied to the electrodes is limited by the breakdown strength of the air – too high a voltage would result in a spark, which is a sudden discharge of electrical energy (much like lightning), instead of a corona. To avoid this a larger number of electrodes are required. Again, finding the optimal geometry will be the key determinant of success, as too many electrodes too close together tend to diminish the corona effect.

The electric field energy is converted into a kinetic energy of moving air through collisions of ions with the air molecules. Naturally, higher velocities of air at the device outlet can be achieved by increasing the ion density. The larger number of ions can be generated by placing a multiple corona electrodes close to one another. The energy could also be increased by increasing the electric field in the space between the electrodes. However, the power available from a single corona electrode is limited by air breakdown voltage. The range between the corona inception and the air breakdown is quite narrow, for any geometry. The overall conclusion is that in order to increase the velocity of air, the electrode density should be increased. A straightforward way to achieve that is to stack the electrodes above each other, in a direction perpendicular to the air flow. However, doing so would lead to reduced electric field intensity near the corona wire. In the limit, an array of neighboring corona wires acts just like a plane electrode, not suitable for creating an inhomogeneous electric field. The solution to this problem lies in "screening" or shielding the corona electrodes from each other . The corona electrodes can be placed close to each other and yet maintain the corona discharge if they are electrically shielded from each other. The shield ing effect can be achieved by designing geometry and excitation pattern for shielding electrodes in such a way that they still produce non-homogeneous electric field needed to generate a corona discharge. These electrodes must be placed between adjacent corona electrodes and, thus, across the intended di rection of flow for the fluid molecules. This design feature leads to the third element of discussion, distribution of electric fields and design of electrode shapes most suitable for producing the air flow in the desired direction.

Energy efficiency

Optimization of the energy efficiency requires optimization of the overall fluid flow, so that a maximal amount of electrical energy is used to produce kinetic energy of the flow in the right direction. One approach to "tune" the fluid flow is to use additional accelerating electrodes, strategically placed to draw the charged fluid in the desired direction. This concept also requires substantial development and testing.

Future research is likely to consider improvements in some or all of these areas as compared to the simple models. Finally, although as-yet unproven, the potential for MEMS-level integration may further reduce costs, increase electrical and thermal efficiency, and also increase reliability. Other work remains in testing prototype devices to ascertain if they are ready for widespread commercial use. For example, a determination of the operating life of an EFA device would need to be made before they could be widely adopted. Long-term running effects might include collection of particles on the electrodes, with as-yet-unknown effects. Likewise, air humidity effects need to be fully considered.

In order to cause ions to create fluid flow, either the exciting electrode must be asymmetrically located between the adjacent corona electrodes or there must be an accelerating electrode. Such an accelerating electrode is an attracting electrode placed downstream from the corona electrodes in order to cause the ions to move in the intended direction. The electric polarity of the attracting electrode is opposite to that of the corona electrode. The effect of electric field distribution on the efficiency of energy conversion has been previously studied in, however, at that time, the approach was focused strictly on electric fields and did not take into account the effects of fluid dynamics. Although the concept of electrode placement and energization is relatively easy to conceive, the practical solution is not obvious, since the design space is very large. The reasoning for initial design is described above. Theoretical modeling and numerical simulations become especially important for the iterative design process of individual EFA stages. The next section presents our approach to multi-physics modeling of EFAs.

12.4 Application

12.4.1 Cooling

As integrated circuits scale down, alternative technologies are required to cool local regions of high heat flux i.e., "hot-spots". Air cooling remains an attractive option for thermal management because of its cost advantages and relative ease of implementation. An increasing challenge in the future is that many components are expected to reach heat flux levels at which heat transfer augmentation will be required. Because it is impractical to add spreaders and heat pipes to these higher power components, there is a need for technologies where the heat transfer can be augmented without adding extra heat transfer area to the component. Microscale ionic winds can potentially enhance the bulk cooling of forced convection at the location of a hot-spot for more effective and efficient cooling.
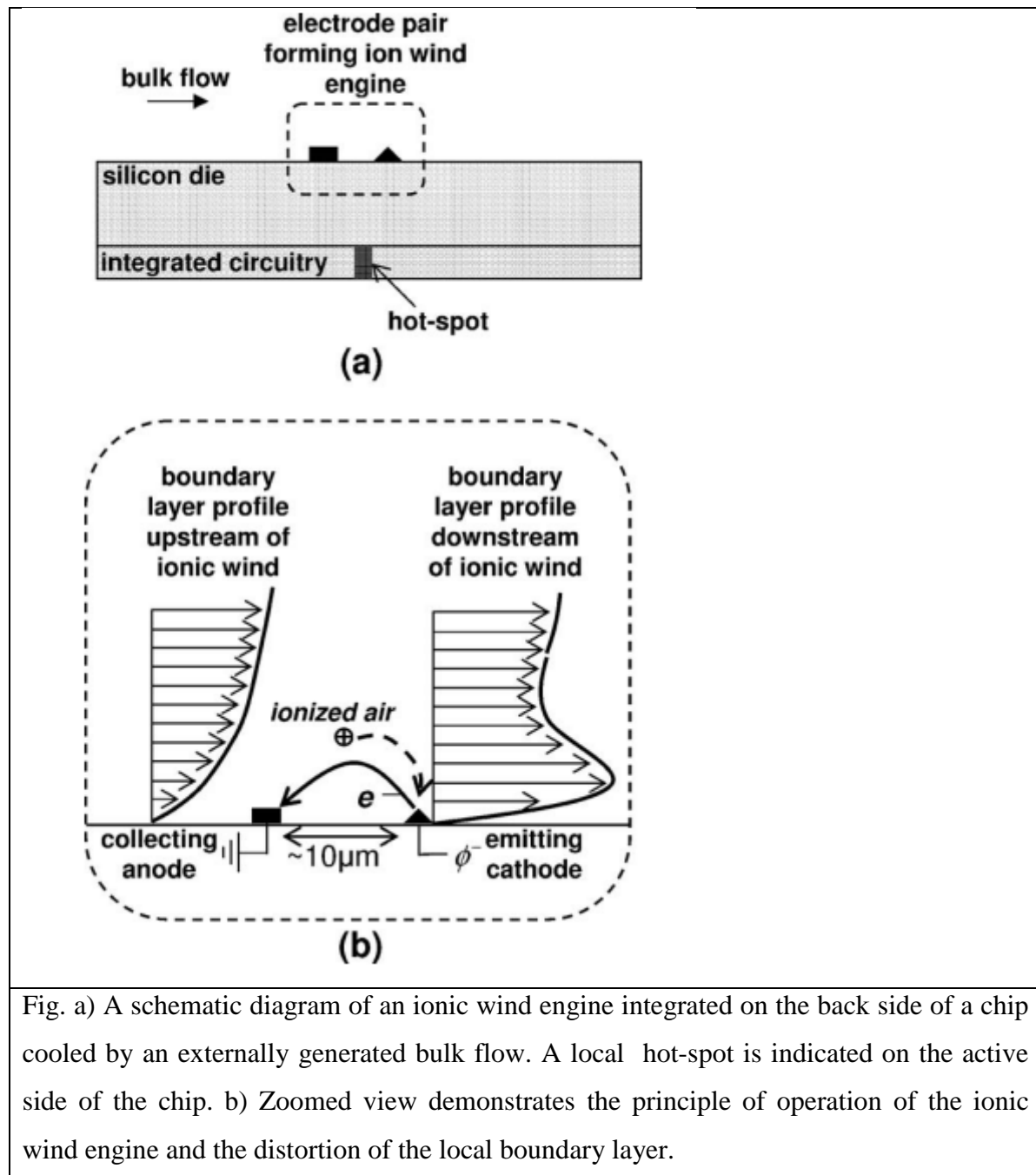
Fig. a) A schematic diagram of an ionic wind engine integrated on the back side of a chip cooled by an externally generated bulk flow. A local hot-spot is indicated on the active side of the chip. b) Zoomed view demonstrates the principle of operation of the ionic wind engine and the distortion of the local boundary layer.

Because microscale ionic wind devices i.e., "ionic wind engines" can be fabricated on an electronic chip, on a spreader, or on the skin of a notebook, they offer an attractive method for hot-spot thermal management. Microscale ionic winds have been suggested previously as an air pumping technique for electronics cooling but without detailed experimental evidence. Furthermore, because the air pumping device must also draw air through the entire electronic system, relying on ionic winds alone may be insufficient because they are often unable to generate the necessary pressure rise.

The use of ionic winds as a method of enhancing a preexisting bulk flow rather than as a method of pumping static air was modeled. In this configuration, a fan would still be used for cooling, and the ionic wind engines would be selectively placed in order to enhance heat transfer where required. Computational simulations at the continuum scale were conducted to explore the

feasibility of this approach and indicated that a microscale ionic wind can enhance the local heat transfer coefficient of a 1 m/s bulk flow by approximately 50%.

For a device cooled by an external i.e., fan-drivenbulk flow on its back side, an integrated ionic wind engine would be situated over a known hot-spot location and would consist of an electrode pair anode and cathode under an applied potential difference of the order of hundreds to thousands of volts, depending on the geometry.

The configuration and spacing between the electrodes determine whether ions are generated by corona discharge or electron field emission. A corona is typically formed by a sufficiently high potential that is applied between a sharp electrode a pin or wire and a blunt, collecting electrode with an electrode gap separation of the order of millimeters. The geometrically enhanced electric field near the sharp electrode accelerates naturally occurring free electrons in the atmosphere so that they collide with neutral molecules. Recent work has suggested that as the gap decreases to ~10 μm or less, the impacting electrons are not naturally occurring, but tunnel from the surface of the cathode into the atmosphere. In both cases, at sufficient kinetic energies, collisions between accelerated electrons and neutral air molecules strip an electron from the neutral molecule to form an ion. The ions are also accelerated through the atmosphere by the electric field, colliding with neutral molecules, exchanging momentum, and generating an ionic wind. The ionic wind acts as a secondary wind in the existing bulk flow, causing local fluid acceleration, and disrupts the velocity and thermal boundary layers to enhance the local cooling effect. Figure 1 depicts the concept of a chipintegrated ionic wind engine and demonstrates the principle of operation and the expected distortion of the local boundary layer due to the ionic wind.

Published seminal works investigated the use of ion drag as pumps for both air and liquids. Since then, ionic winds have been investigated for applications ranging from blowers to air filtration and heat transfer enhancement. From a thermal perspective, a variety of corona configurations have been investigated including point-to-plane impingement, pumping air in ductlike structures and enhancing internal flows. The use of ionic winds in the presence of a bulk flow to modulate an external boundary layer has been an area of growing interest in the aerospace community. The ability of corona discharges to reduce drag using electrodes perpendicular to the flow and either flush or in contact with a flat plate was investigated and demonstrated. Little work exists on the heat transfer effects of ionic wind-enhanced external flow. At the microscale, recent experimental and numerical work has also been completed by the authors' group. Electron emission was demonstrated in air for turn-on potentials on the order of tens of volts for 10 μm electrode gaps using surfaces enhanced by nanofabricated, polycrystalline diamond.

12.4.2 Fluid Accelerator

Devices were fabricated from bulk silicon using industry standard photolithographic techniques. Lowering the necessary onset voltage for EFA operation was the principle design goal. Figure 12.2 shows the geometry of the devices alongside an SEM image of an actual device.
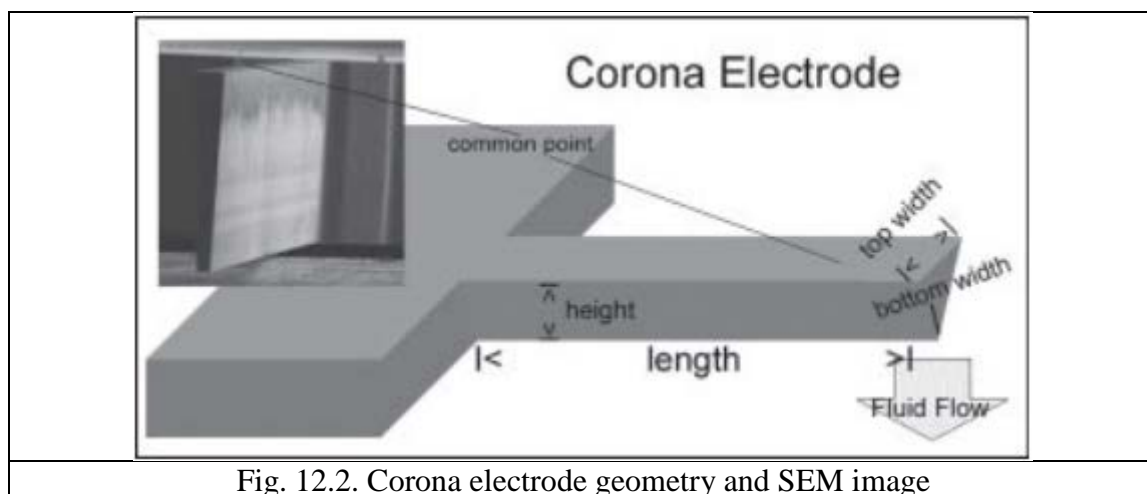
Fig. 12.2. Corona electrode geometry and SEM image

The wafers were etched using deep reactive ion etching (DRIE). High curvature was desired at the very tip of the long cantilever on the underside of the wafer. The overall tip length serves to isolate the tips' electric field from that of the low curvature base. The negative sidewall tapering (top width > bottom width) was achieved by patterning large spacing around the cantilever for masking during the DRIE. Devices were fabricated with varying tip length, height, top width, and bottom width. Etch times were continually modified to establish reliable expectations of tip dimensions and curvature. Figure 12.3 shows a device with sharpening by RIE for high tip curvature.
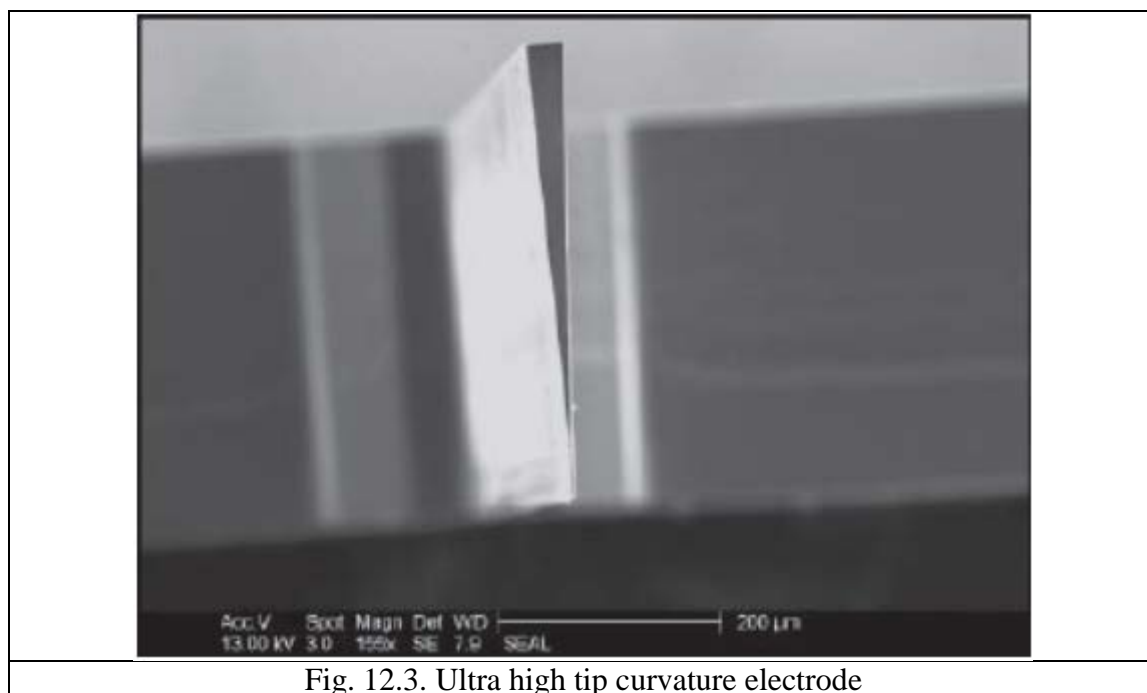
Fig. 12.3. Ultra high tip curvature electrode

The devices were tested to obtain current-voltage relationships for micro EFA device operation at known corona electrode to collector electrode orientations and separation distances. The devices were held at a distance above a semiconductive foam collector electrode by a micropositioning x-z

stage. SEM imaging was used before and after tests to ensure devices conformed to desired dimensions and did not change during operation or from breakdown effects.
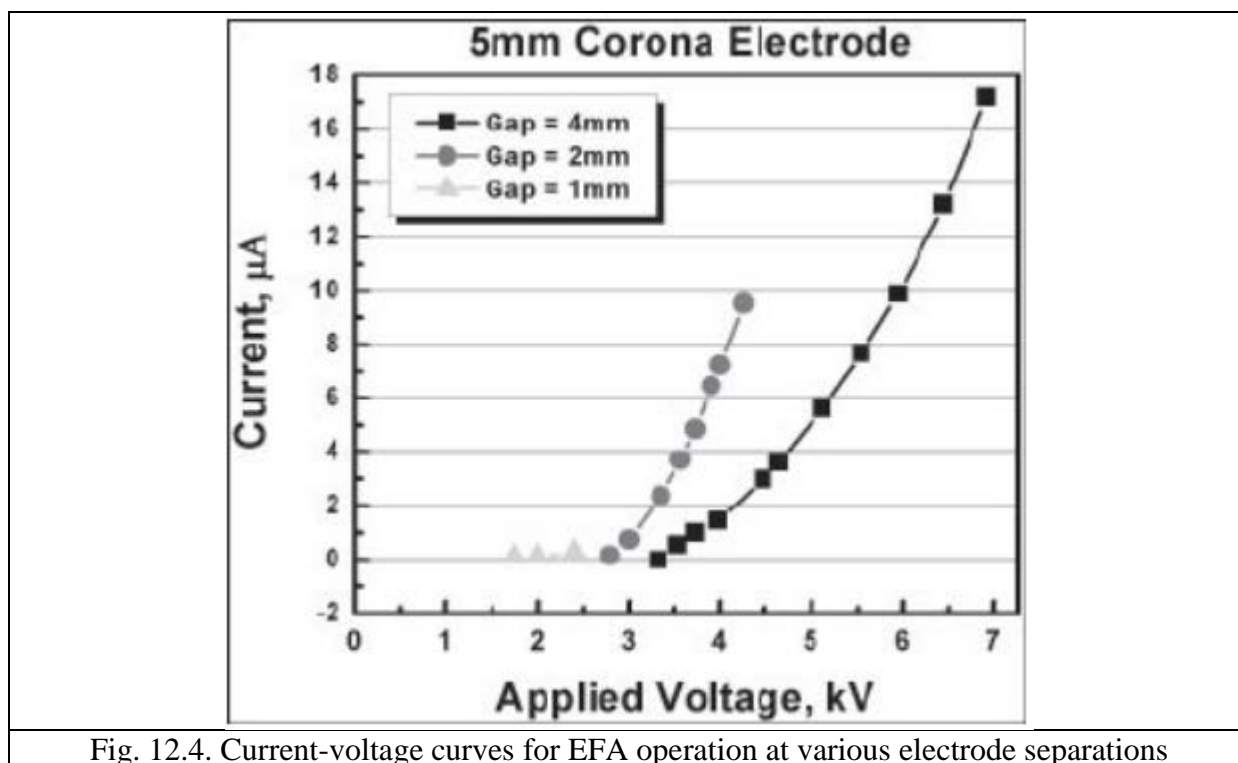
Fig. 12.4. Current-voltage curves for EFA operation at various electrode separations

Micro devices were tested to determine the dependency of onset and breakdown voltages on tip length as well as air gap size between electrodes. Positive voltage was applied to the Corona electrode and the collector was held at a lower voltage. Current data points were taken for a particular device geometry and air gap size throughout the cut-off and operational voltage range up until the expected point of breakdown or appearance of sparks. Figure 12.4 shows the effects of variations in air gap size on the operational voltage range of a 5 mm long tip.

Operational voltage range increased along with the air gap size. Experimental data, not shown here, indicates that onset voltage is lowered and operational range extended for devices with longer tips. This was observed via extensive testing of 3, 5, and 8 mm tip lengths. Preliminary tests indicated air velocities for some devices in excess of 2 m/s. Future work includes further miniaturization and characterization of micro EFA devices moving towards a fully optimized single wafer on-chip solution.

13. Micro-scale energy harvesting MEMS.

13.1   Introduction

13.2   Mechanical to Electrical Conversion

    13.2.1 Piezoelectric Energy Harvesting

        13.2.1.1   Basic Principles

        13.2.1.2   Piezoelectric Transducers

        13.2.1.3   Piezoelectric Microgenerators

        13.2.1.4   Energy Harvesting Circuits

        13.2.1.5   Strategies for Enhancing the Performance of Energy Harvester

            13.2.1.5.1 Multi-modal Energy Harvesting

            13.2.1.5.2 Magnetoelectric Composites

            13.2.1.5.3 Self-Tuning

            13.2.1.5.4 Frequency Pumping

    13.2.2 Electromagnetic Energy Harvesting

        13.2.2.1   Basic Principles

        13.2.2.2   Device constructions

        13.2.2.3   Microscale Implementations

        13.2.2.4   Macro-Scale Implementations

    13.2.3 Electrostatic Generators

    13.2.4 Reverse Electro-Wetting Generators

13.3   Thermal to Electrical Energy Converters

    13.3.1 Basic Principles

    13.3.2 Heat Engines

        13.3.2.1   MEMS Stirling Engines

        13.3.2.2   MEMS Brayton Engines

        13.3.2.3   MEMS Vapor Cycle Engines

        13.3.2.4   MEMS Ericsson Engines

        13.3.2.5   Thermoacoustic Engines

        13.3.2.6   Shape Memory Alloy Heat Engines

        13.3.2.7   Thermomagnetic Generators

    13.3.3 Thermal to Electrical Energy Converters

        13.3.3.1   Thermionic Generators

        13.3.3.2   Pyroelectric Generators

        13.3.3.3   Seebeck Thermoelectric Generators

        13.3.3.4   Alkali Metal Thermal-to-Electric Converter

13.1 Introduction

Energy harvesting materials and systems have emerged as a prominent research area and continues to grow at rapid pace. A wide range of applications are targeted for the harvesters, including distributed wireless sensor nodes for structural health monitoring, embedded and implanted sensor nodes for medical applications, recharging the batteries of large systems, monitoring tire pressure in automobiles, powering unmanned vehicles, and running security systems in household conditions. Recent development includes the components and devices at micro–macro scales covering materials, electronics, and integration. The growing demand for energy harvesters has motivated the publication of this book to present the current state of knowledge in this field.

In most sources of low-temperature waste heat, the heat is dissipated to the atmosphere due to a lack of cost-effective solutions that can efficiently convert the heat into useable electrical energy. The ability to capture this thermal energy could increase the efficiency of existing processes and machinery, supply isolated sensors, allow for extended portable electronic power supply, and much more. The aim of this Springer Brief is to summarize a very broad range of thermal energy harvesting methods, and describe the potential of applying these methods to lowcost, batch process manufactured, micro electromechanical systems (MEMS). The brief will focus on the functionality of the device, rather than the MEMS production methods. An additional motivation for the production of this Springer Brief is the rapid growth of the MEMS market, which grew to $11 billion in 2012 and is expected to double by 2018.

13.2   Mechanical to Electrical Conversion

13.2.1 Piezoelectric Energy Harvesting

In last decade, the field of energy harvesting has increasingly become important as evident from the rising number of publications and product prototypes. At the same time, several applications have been projected for the energy harvesters covering wide range of civilian and defense components. Out of these different applications, the prominent use of harvester is to power the wireless sensor node. A major challenge in the implementation of multi-hop sensor networks is supplying power to the nodes. Powering of the densely populated nodes in a network is a critical problem due to the high cost of wiring or replacing batteries. In many cases, these operations may be prohibited by the

infrastructure. Outdoor solar energy has the capability of providing power density of 15, 000 W/cm3 which is about two orders of magnitudes higher than other sources. However, solar energy is not an attractive source of energy for indoor environments as the power density drops down to as low as 10–20 W/cm3. Mechanical vibrations (300 W/cm^3) and air flow (360 W/cm^3) are the other most attractive alternatives. In addition to mechanical vibrations, stray magnetic fields that are generated by AC devices and propagate through earth, concrete, and most metals, including lead, can be the source of electric energy. The actual AC magnetic field strengths encountered within a given commercial building typically range from under 0.2 mG in open areas to several hundred near electrical equipment such as cardiac pace makers, CRT displays, oscilloscopes, motor vehicles (approximately up to 5 G max); computers, magnetic storage media, credit card readers, watches (approximately up to 10 G max); magnetic power supply, liquid helium monitor (approximately up to 50 G max); magnetic wrenches, magnetic hardware, and other machinery (approximately up to 500 G max). AC magnetic fields decrease naturally in intensity as a function of distance (*d*) from the source. The rate of decrease, however, can vary dramatically depending on the source. For example, magnetic fields from motors, transformers, and so on, decrease very quickly ($1/d^3$), while circuits in a typical multi-conductor circuit decay more slowly ($1/d^2$). Magnetic fields from "stray" current on water pipes, building steel, and so on, tend to decay much more slowly ($1/d$). The other important sources of energy around us are radio frequency waves and acoustic waves.

13.2.1.1   Basic Principles

Vibration of a rigid body can be caused by several factors such as unbalanced mass in a system, tear and wear of materials and can occur in almost all dynamical systems. The characteristic behavior is unique to each system and can be simply described by two parameters: damping constant and natural frequency. Most commonly, a single degree of freedom lumped spring mass system is utilized to study the dynamic characteristics of a vibrating body associated with energy harvesting. The single degree of freedom helps to study unidirectional response of the system. Figure 13.1 shows a diagram of a cantilever beam with piezoelectric plates bonded on a substrate and a proof mass at the end; multilayer piezoelectric plates and equivalent lumped spring mass with external excitation. Cantilever structure with tip mass is the most widely used configuration for piezoelectric energy harvesting device. The source of vibration is shown with an arrow at the base of the contact point. The stiffness of the structure depends on the loading condition, material, and cross-sectional area perpendicular to the direction of vibration. The governing equation of motion for the system shown in Fig. 13.1c can be obtained from energy balance equation. This configuration applies to both the energy harvesting mechanisms shown in Fig. 13.1a and 13.1b.
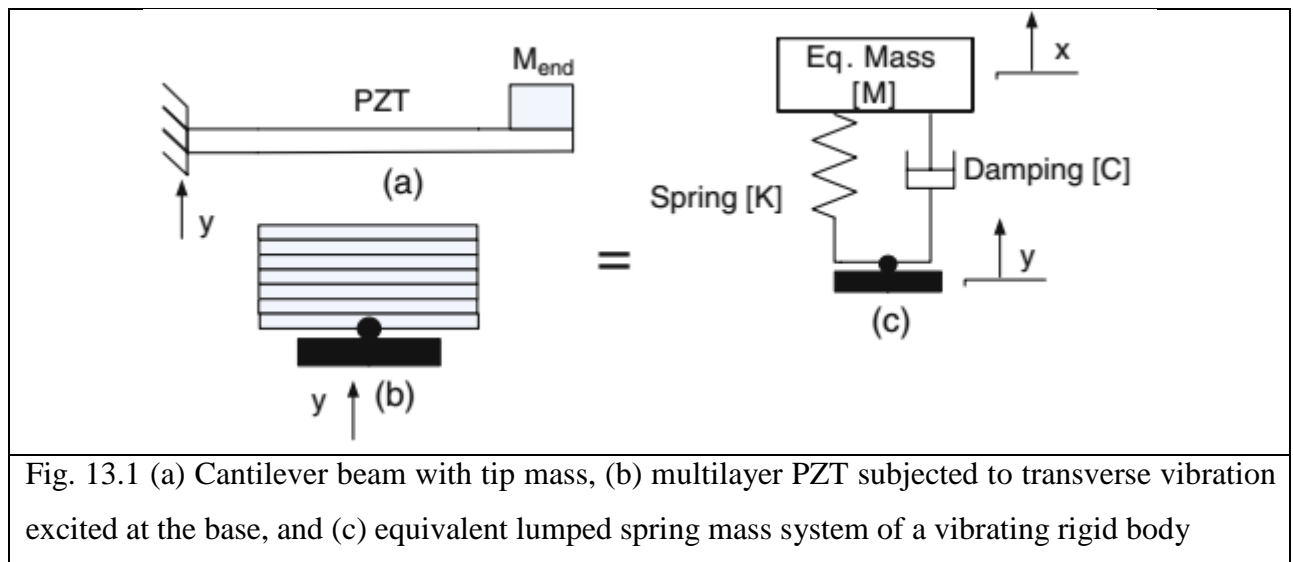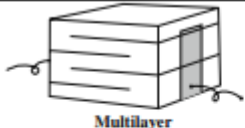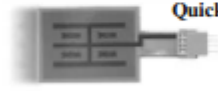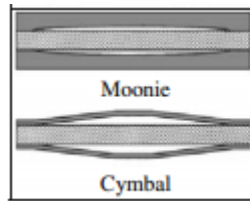
Fig. 13.1 (a) Cantilever beam with tip mass, (b) multilayer PZT subjected to transverse vibration excited at the base, and (c) equivalent lumped spring mass system of a vibrating rigid body

13.2.1.2   Piezoelectric Transducers

There has been significant progress in the design and fabrication of piezoelectric transducer structures. 13.1 lists the designs which can be easily obtained from commercial sources and are promising for energy harvesting application.

Table 13.1

Transducer Products	Company/Characteristics
Multilayer	Supplier example: Morgan electroceramics, APC International, Tokin, PI. Characteristics: low frequency (~10 Hz), suitable under large uniaxial stress condition, easy mounting.
Macro Fiber Composite (MFC)	Supplier example: Smart Material. Characteristics: flexible, both d_{33} and d_{31} mode possible, low strain high frequency application, large area coverage, can be used as a bimorph element.
Thunder	Supplier example: Face International Characteristics: Various curvatures and heights possible providing wide range of stress amplification, suitable for very low frequencies (~1 Hz).
Bimorphs	Supplier example: APC International. Characteristics: resonance frequency can be tuned in the range of 5–100 Hz, used in various configuration such as cantilever, end–end clamped, etc.
Amplified Piezoelectric Actuator	Supplier example: Cedrat Characteristics: higher efficiency under large stress, resonance frequency can be tuned to lower ranges (~100 Hz).
QuickPack	Supplier example: Mide Characteristics: similar to bimorphs but easier mounting, wide bandwidth, widely used in cantilever configuration
Rainbow	Characteristics: curved surface resulting in higher charge under a given stress level, can be stacked to amplify charge.

	Supplier: Micromechatronics
Moonie	Characteristics: metal caps protect ceramic allowing application under high stress levels, higher charge due to stress amplification, resonance frequency can be tuned by changing cap dimensions and material.
Cymbal	

13.2.1.3 Piezoelectric Microgenerators

There are five components of energy harvesting device based on MEMS process, namely thin film composition and deposition technique, device design, fabrication process, including etching procedure, electrical connections, and packaging. As mentioned earlier, PZT composition close to MPB are commonly selected and deposited using chemical solution deposition technique. Chemical deposition process for PZT films was proposed. In this process, lead acetate trihydrate is added to 2-methoxyethanol (2-MOE) under argon atmosphere in the flask maintained in an oil bath at 70°C on a hot plate and the temperature of the bath is gradually increased to 120°C.
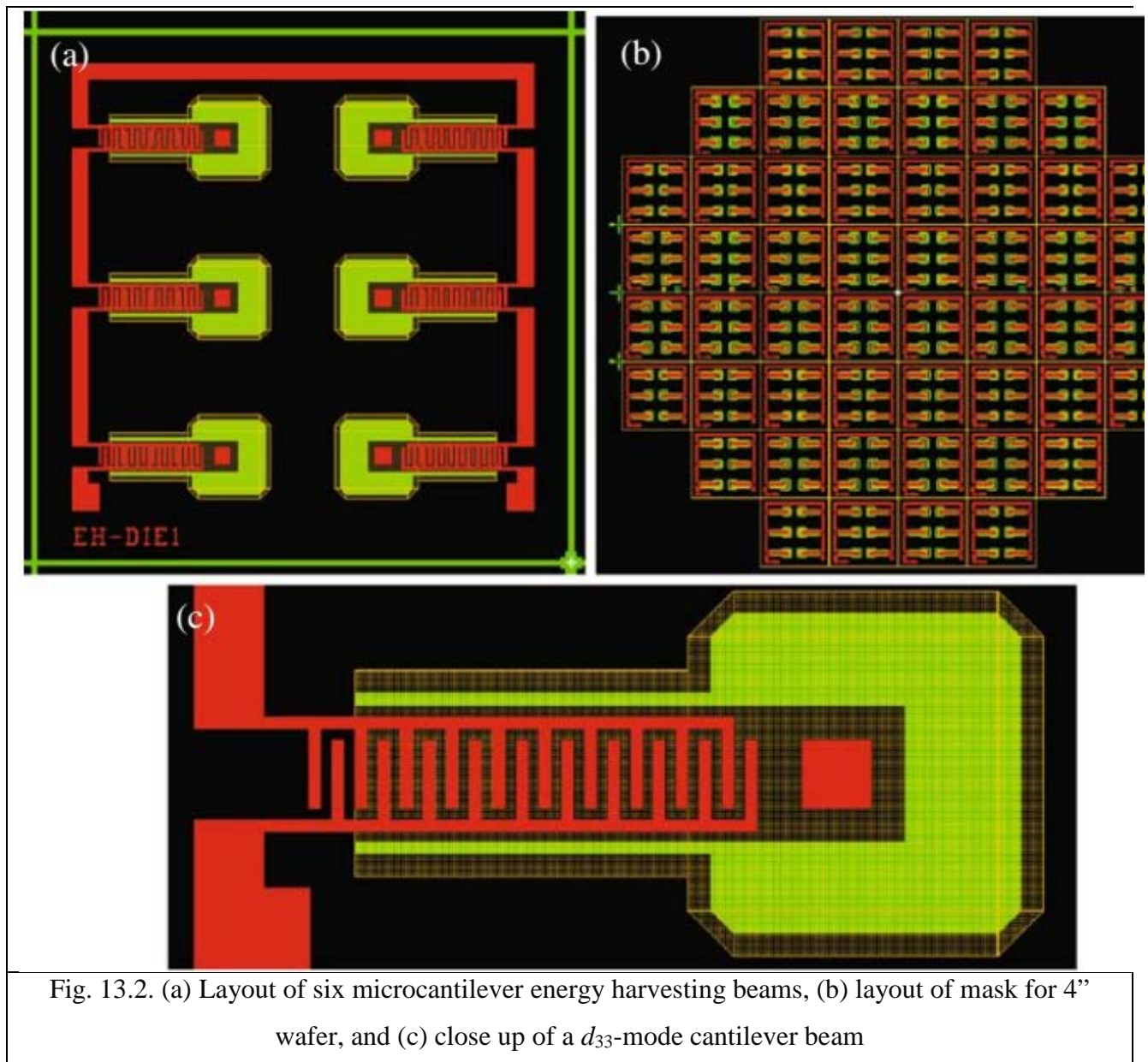
Fig. 13.2. (a) Layout of six microcantilever energy harvesting beams, (b) layout of mask for 4" wafer, and (c) close up of a d_{33}-mode cantilever beam

The solution is then dehydrated under vacuum until a semi-dry powder remains. A mixture of zirconium *n*-propoxide and titanium *iso*-propoxide in 2-MOE at room temperature is added and the entire solution is refluxed for 2 h under Ar at 120°C. After refluxing, the solution is vacuum distilled and 2-MOE is added until the desired solution molarity is achieved. Acetylacetonate and acetic acid are used to adjust the molarity of the solution. Desired modifiers can be added before the refluxing stage.

Piezoelectric mirco-cantilevers-based energy harvester utilizing lead-free barium titanate thin film was demonstrated. This device has interdigital electrode pattern to access the d_{33} response. The microcantilevers were connected in series to increase the output voltage. Figure 13.2 shows the design of mircogenerator and cantilever beam layouts. The fabrication of microcantilever beams requires three masks: electrode mask, cantilever beam mask, and backside mask. The electrode mask is used to pattern metal electrode (red color), the cantilever beam mask is used.

The SEM picture of the fabricated MEMS microcantilever beam is shown in Fig. 13.3.
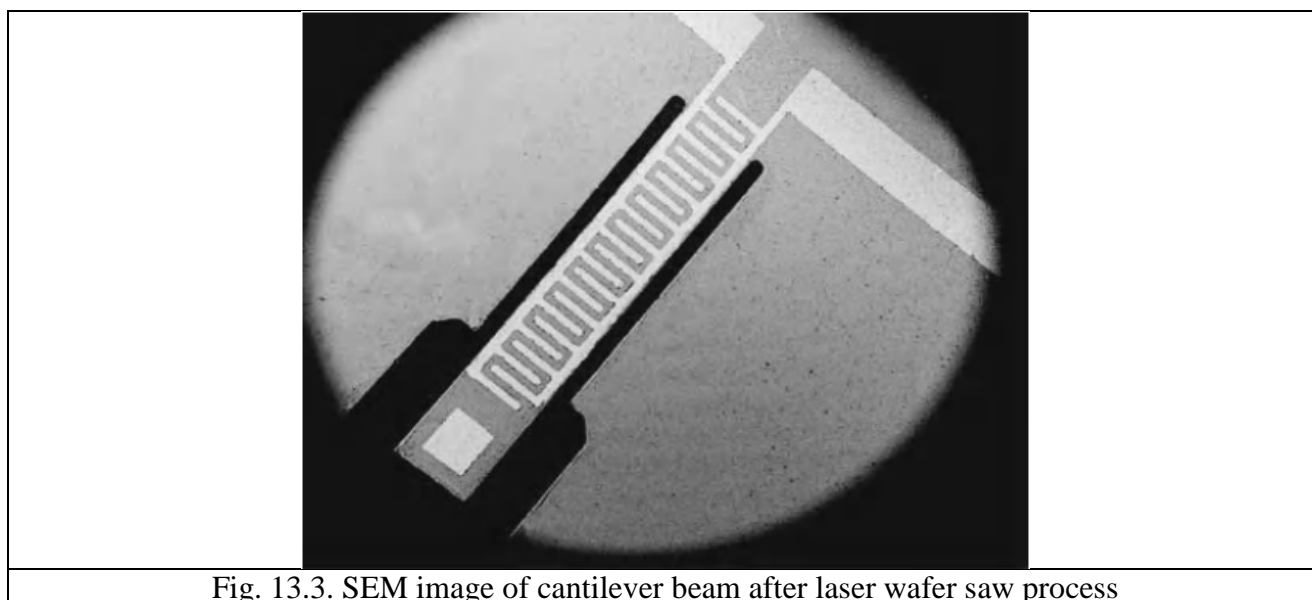
Fig. 13.3. SEM image of cantilever beam after laser wafer saw process

13.2.1.4 Energy Harvesting Circuits

A simple energy harvesting circuit consists of a diode rectifier (AC/DC) and a DC– DC converter, as shown in Fig. 13.4. The addition of DC–DC converter has been shown to improve energy harvesting by a factor of 7. The efficiency of the step-down converter was between 74 and 88%. Through the exploitation of discontinuous conduction mode operation of the DC–DC converter, a stand-alone energy harvesting system with significantly simplified control circuitry has been proposed. A nonlinear processing technique "Synchronized Switch Harvesting on Inductor" (SSHI) was presented for harvesting energy, which consists of a switching device in parallel with piezoelectric element. The device is composed of a switch and an inductor connected in series. The switch is in open state except when the maximum displacement occurs in the transducer. At that instant, the switch is closed and the capacitance of the piezoelectric element and inductance

together constitute an oscillator. The switch is kept closed until the voltage on the piezoelectric element has been reversed. In case of nonlinear AC device, a resistive load is directly related on the piezoelectric element in parallel with the switching device. This nonlinear technique has been shown to significantly enhance the performance of the energy harvesting circuit and will be well suited for the resonating structures.
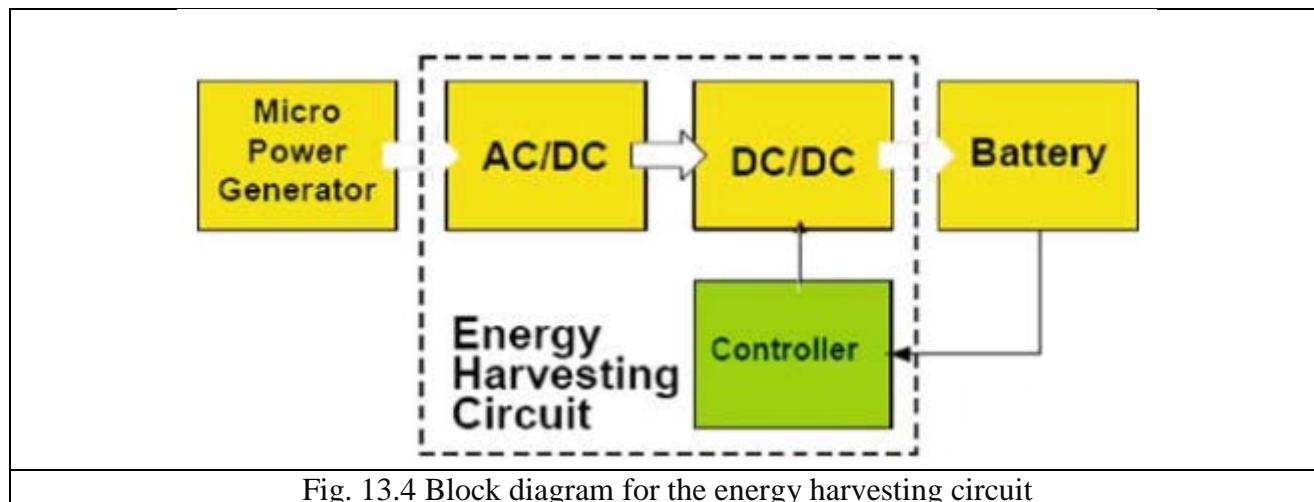
Fig. 13.4 Block diagram for the energy harvesting circuit

Piezoelectric microgenerators require development of CMOS-based components for efficient energy harvesting and storage. Typically, piezoelectric generator's internal impedance is high in the range of few tens of k. This high-internal impedance restricts the amount of output current that can be driven by the source to microampere range. Therefore, it is important to have low dissipation of the quiescent current for both rectifier and controller of the SC-regulated doubler in order to enhance the efficiency of the power conversion system. Recently, Guo and Lee have proposed a high-efficiency rectifier for transcutaneous power transmission in biomedical application as shown in Fig. 13.5a. The rectifier was implemented in a standard 0.35 m CMOS and was found to achieve peak conversion ratio of 95% with output current handling capability of up to 20 mA. The rectifier adopts two comparator-controlled switches to allow unidirectional current flow with only $2V_{ds}$ drop in the conducting path, where V_{ds} is the drain–source voltage across a transistor in linear region. With the use of comparator-controlled switches, this design was not only functional under the lowest input amplitude, but also has good power efficiency. In the subsequent work, Lee and Mok have developed a switched-capacitor (SC) voltage doubler with pseudo-continuous control as shown in Fig. 13.5b. The SC-regulated voltage doubler provides a constant DC voltage through a step-up DC–DC conversion. These developments are important for full implementation of microgenerator.

In the case of wireless sensor node, an energy storage mechanism with significant energy density is required between the energy harvesting circuitry and the wireless transmission unit. Possible solutions are lithium ion rechargeable batteries and double-layer ultracapacitors. The

appropriate choice is based on the required energy density and power density of the application, with lithium ion batteries being used if high "energy" density is required, and ultracapacitors if high "power" density is required.
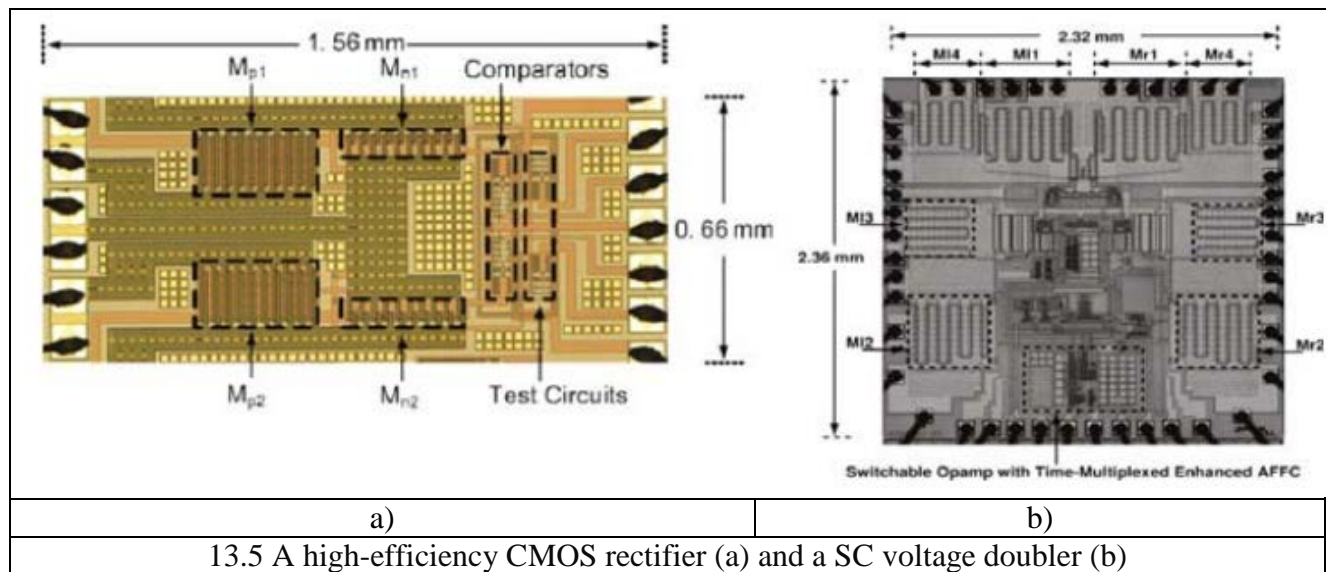
| a) | b) |

13.5 A high-efficiency CMOS rectifier (a) and a SC voltage doubler (b)

13.2.1.5   Strategies for Enhancing the Performance of Energy Harvester

13.2.1.5.1   Multi-modal Energy Harvesting

In order to effectively harvest energy from the available environmental resources, the concept of multi-modal energy harvesting is being pursued. This concept improves upon the existing systems by integrating the following ideas: (i) in a given scenario, two or more energy sources may be available for harvesting energy such as solar, wind, and vibration and (ii) two different energy harvesting schemes can be combined in one system such that one assists the other.

Fig. 13.6. Schematic structure of the multi-modal energy harvester, side-view, and top-view
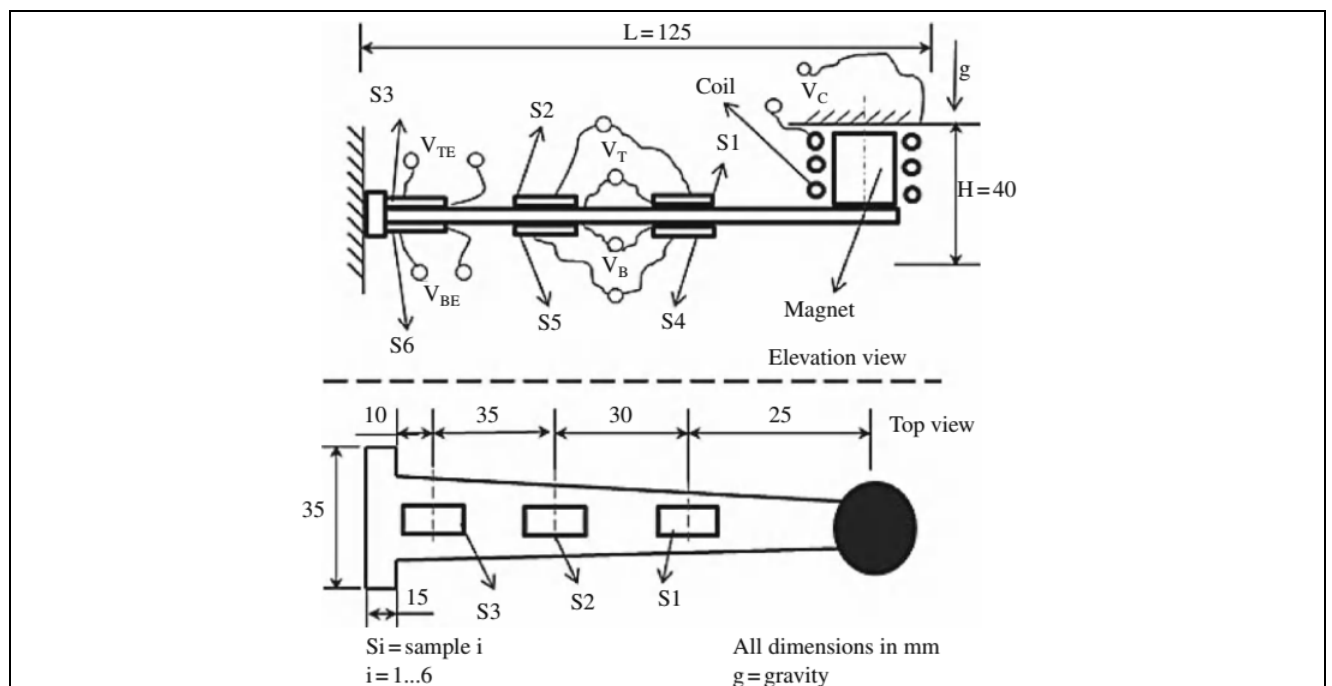
A suitable example of multi-modal energy harvesting device is a system that combines electromagnetic and piezoelectric energy harvesting mechanisms. Such a device consists of

piezoelectric transducer and electromagnetic system mutually contributing towards power generation. The design consists of cantilever beam with piezoelectric plates bonded at maximum stress locations as shown in Fig. 13.6a and 13.6b.

At the tip of cantilever beam, a permanent magnet is attached which oscillates within a stationary coil fixed to the top of package. The permanent magnet serves two functions: acts as a proof mass for the cantilever beam and acts as a core which oscillates between the inductive coils resulting in electric current generation through Faraday's effect. Figure 13.7 shows the diagram of the prototype device. The piezoelectric single crystal plates were bonded on both top and bottom of the beam. The plates were isolated from beam using an insulating epoxy. The output voltages of top and bottom row of piezoelectric were connected in parallel.
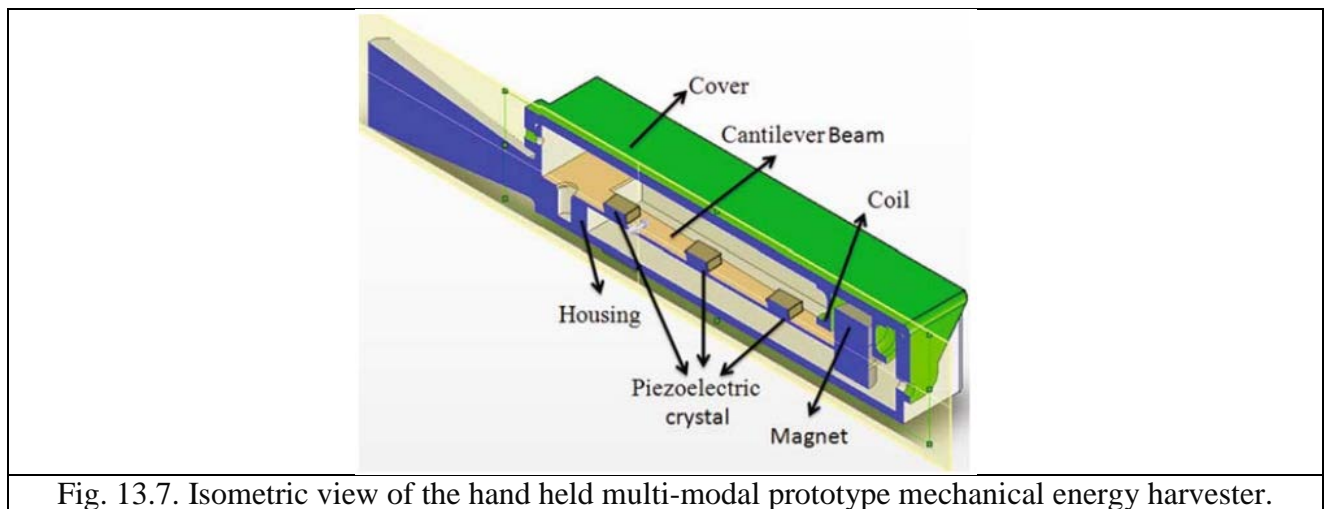
Fig. 13.7. Isometric view of the hand held multi-modal prototype mechanical energy harvester.

The beam has tapered geometry with linear variation in dimension along the length such that its moment of inertia along the axis perpendicular to the direction of vibration also varies linearly. This allows the beam to exhibit higher sensitivity. When a cantilevered beam is subjected to external excitation, various modes occur at respective resonance frequencies depending on the geometry, modulus of elasticity, density, and boundary conditions. The first two vibration modes were considered for experimentation as they cover the desired low-operating range of 20–400 Hz for the dimension of cantilever ($25 \times 30 \times 125$ mm^3). The advantage of this design is that at constant acceleration, the output power from electromagnetic is much higher at lower frequencies (first transversal resonance mode), whereas the output from piezoelectric is higher at higher frequencies (second transversal resonance mode). Thus, this combination allows obtaining significant magnitude of output power from the same device over a wide operating frequency range. Since the frequency content of a random vibration is wide, such a device can utilize multiple spectrum. The bonded piezoelectric plates were located at maximum stress position for the cantilever beam operating in the second resonance mode.

Figure 13.8 shows the photograph of EVA 100/105 wireless transceiver and the fabricated multi-modal harvester. By connecting the on board RS232 DB9 connector through a serial PC port, the received data can be saved on a PC through the application software. The push button transmitter module, PTM 200, works based on electro-dynamic energy transducer actuated by a bow that can be pushed from outside the module on the left or right. The transmission frequency and power were 868.3 MHz and 10 mW max, respectively. The electrical power generated from he fabricated prototype is able to power up the processor and HF transmitter on the board.
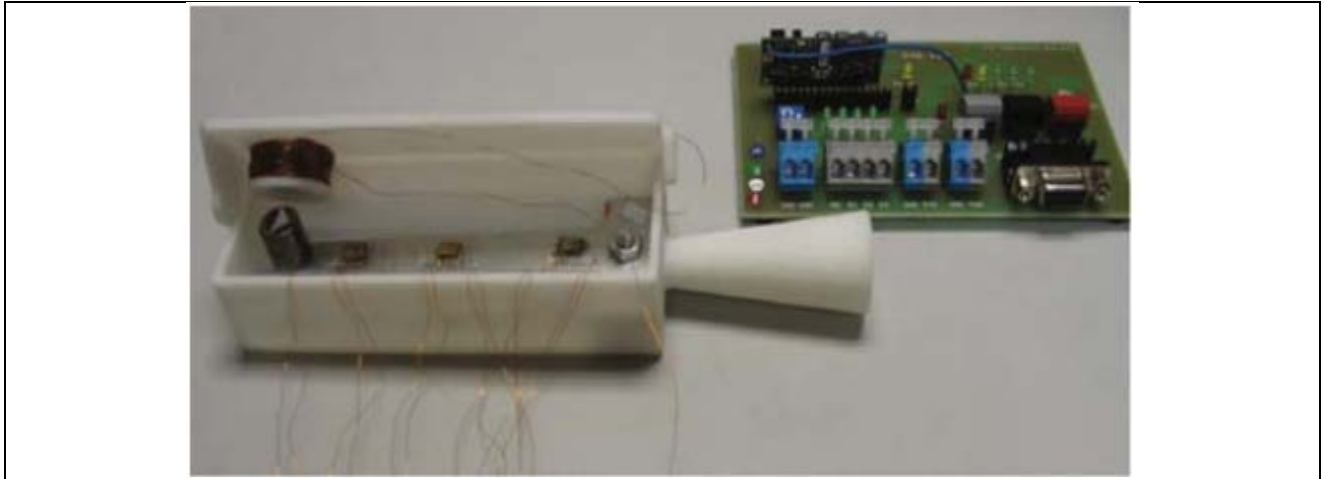
Fig. 13.8. EVA 100/105 evaluation kits and the multi-modal harvester.

13.2.1.5.2   Magnetoelectric Composites

Magnetoelectric composites combine piezoelectric and magnetostrictive phases in one material thus providing the opportunity to convert magnetic field into stress. Furthermore, it can improve the efficiency of vibration energy harvesting system. The voltage generated across the piezoelectric device under an applied stress X can be simply expressed by the relation, $V = gXt$, where t is the thickness. The voltage can be significantly enhanced using the magnetic stress, through the expression, $V + V^{\text{magnetic}} = g.(X + X^{\text{magnetic}})t$, generated through magnetostriction. Exact calculations show the voltage generated through the magnetoelectric composite (piezoelectric – magnetostrictive (fiber – matrix)) structure is given as:

$$V = \frac{N d_{33,\text{m}} g_{33,\text{p}}}{N s_{33}^{\text{E}}(1 - k_{33}{}^2) + (1 - N)s_{33}^{\text{H}}} H.t \qquad (13.1)$$

where s_{33}^E and s_{33}^H are the elastic compliances for the piezoelectric and magnetostrictive layers, respectively, $k33$ is the electromechanical coupling coefficient of the piezoelectric layer, $d_{33,m}$ and $g_{33,p}$ are the longitudinal piezomagnetic and piezoelectric voltage coefficients, respectively, and n is the thickness fraction of magnetostrictive layers. The magnetoelectric coefficients as high as 20 V/cm e at 1 Hz have been reported as predicted by the above expression (after using the correction factor). This result indicates that by utilizing magnetoelectric laminate composites, a higher power

density can be obtained from the vibration energy harvester where vibration is used to create magnetic field by oscillating the magnets. Figure 13.8 shows some examples of the structures that can be designed using the two phases in thin film and bulk forms. It can be seen from this figure that the possibilities are numerous and there can be several phases (amorphous, ceramic, metal, polymer, and so on), shapes (disk, cylinder, plate, toroid, sphere, and so on), and sizes (number of layers, layer thicknesses, length and width can be varied differently) for obtaining the magnetoelectric properties. All of these combinations will exhibit ME response with varying degrees of magnitude. For n phases, the number of connectivity patterns is given as $(n + 3)!/3!n!$, which for two phase composites comes out to be 10, for three phases as 20, and 35 for four phase patterns. Furthermore, in each of these shapes, there is a possibility of orienting the polarization along different axes and applying the electric (E) and magnetic (H) fields along different axes. In addition, there are several choices for materials depending on magnetostriction constant, resistivity, permeability, permittivity, piezoelectric strain and voltage constant, sintering temperature, and chemical reac tivity (magnetostrictive: $MnFe_2O_4$, $CoFe_2O_4$, $NiFe_2O_4$, $ZnFe_2O_4$, YFe_5O_{12}, $Sm Fe_5O_{12}$, YIG, Terfenol-D, Metglas 2605SC, Ni, Co, and so on; piezoelectric: PZT, $BaTiO_3$, PMN – PT, PVDF, $SrBi_4Ti_4O_{15}$, $(Na_{0.5}K_{0.5})NbO_3$), and so on). Several reports have been published on the magnetoelectric composites with different materials and geometry and these have contributed in understanding the effect of some of the variables associated with design. Recently, the focus of much research has been on the laminated magnetoelectric composites made by using piezoelectric and magnetostrictive materials

Fig. 13.8. Schematics of the magnetoelectric composites illustrating the multiple possibilities of (a) self-assembled nanopillar structures, (b) layer-by-layer deposition, (c) deposition accompanied by masking and patterning, (d) three phase depositions for gradient materials, (e) particles dispersed in a polymer matrix, (f) lamination, (g) checkerboard arrangement, and (h) sintered particulates.

13.2.1.5.3 Self-Tuning

One element of efficiency is the magnitude of power harvested when the mechanical input frequency is different from the resonance of piezoelectric harvester. It is highly desired to have an

energy harvester that is able to self-tune to optimize its power output in virtually any vibration environment. One common approach is using the shunt.

This method has drawback that it requires a complex electronic with programmable microcontroller to connect or disconnect a series of capacitive loads. Another method which can be conceived to alter the frequency of the harvester is by cascading electrical connections, i.e., changing the number of transducers connected in series or parallel by an active switch. The switch can regulate the connections by comparing the power levels over a frequency band. This method has drawback that it requires several transducers to be available and arranged in a systematic fashion for implementing the cascading. The device is bulky and has low power density. Another approach can be developed based on the nonlinearity of piezoelectric material. It is known that the applied electrical voltage has an effect on the resonance which is attributed to the changes in stiffness. As the applied voltage increases the resonance frequency of the piezoelectric material shifts to lower frequencies.

13.2.1.5.4   Frequency Pumping

A method to increase the power harvested by piezoelectric transducers is to achieve more deflections in the same time period ($P \propto f$). In other words, enhance the ratio, $f_{piezo}/f_{source} > 1$. Typical vibration frequency of the machines, aircraft, ship hulls, and deck is in the range of 5–100 Hz, which is much lower than the operating frequency of the microgenerator. If the source vibration frequency can be amplified before being applied to the piezoelectric transducer than an improved efficiency can be obtained. One of the simple ways to implement this is using mechanical gears or springs. We have used this technique before to design piezoelectric windmills as shown in Fig. 13.9, where each bimorph goes through 5 Hz oscillation for source frequency of 1 Hz.
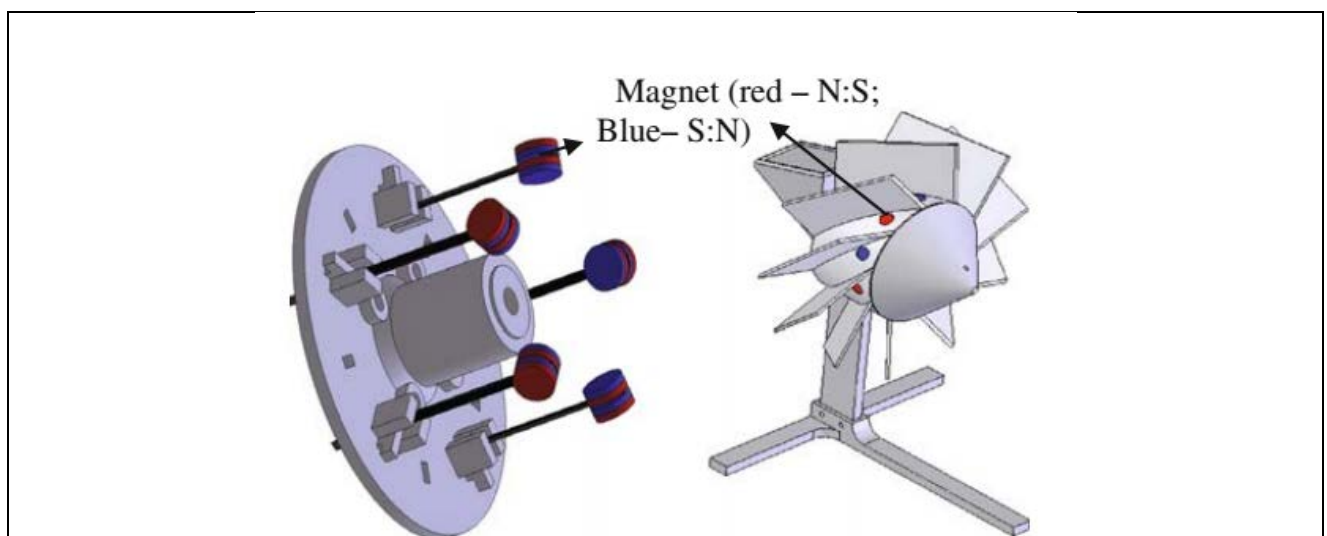
Fig. 13.8. Two magnets sandwich the tip of the bimorphs. These interact with 10 magnets mounted on the vanes in alternating polarities which causes the bimorphs to go through five cycles for every rotation of the vanes

13.2.2 Electromagnetic Energy Harvesting

Electro-magnetism has been used to generate electricity since the early 1930s, not long after Faraday's fundamental breakthrough in electromagnetic induction. The majority of generators used today are based on rotation and are used in numerous applications from the large-scale generation of power to smaller scale applications in cars to recharge the battery. Electromagnetic generators can also be used to harvest micro- to milli-Watt levels of power using both rotational and linear devices. Provided a generator is correctly designed and not constrained in size, they can be extremely efficient converters of kinetic energy into electrical. Attempts to miniaturise the technique, however, using micro-engineering technology to fabricate a generator, invariably reduce efficiency levels considerably. This chapter introduces the fundamental principles of electromagnetic induction before exploring the scaling effects that work against successful miniaturisation. Conventional discrete magnets and coils are compared with their micro-machined equivalent and the technical challenges of associated with micro-coils and deposited magnetic materials are highlighted. The chapter concludes with a comprehensive and up to date review and comparison of energy harvesters realised to date. The generators presented demonstrate many of the issues previously discussed.

Microscale Implementations

The earliest microscale device was reported by researchers from the University of Sheffield, UK. Figure 13.9 shows the electromagnetic approach. It consists of two parts: an upper gallium arsenide (GaAs) wafer containing the spring mass arrangement and a lower wafer that incorporates the planar integrated coil. The seismic mass consists of a vertically polarised 1 mm × 1 mm × 0.3 mm samarium–cobalt magnet of mass 2.4×10^{-3} kg which is attached to a 7 mm thick cured polyimide circular membrane that forms the spring element.

The planar coil is formed on the bottom wafer from a 2.5 mm thick gold (Au) thin film layer is patterned using a lift off process which yielded 13 turns of 20mm line width and 5 mm spacing. The two wafers were bonded together using silver epoxy to form the assembled generator. The overall size of the electromagnetic transducer described by is around 5 mm × 5 mm × 1 mm and a schematic is shown in Fig. 13.10. When the generator is subject to external vibrations, the mass moves vertically out of phase with the generator housing producing a net movement between the magnet and the coil. Early design work presented the familiar equation predicting power output from such a device. This analysis predicted that $+/- 50$ mm inertial mass displacement would produce 1m W at 70 Hz and 100 mW at 330 Hz. In these cases, the driving amplitude, Y, was fixed at 30 mm and, since acceleration $A = \omega^2 Y$, the corresponding acceleration levels are 5.8 and 129 m/s^2 respectively. The device was tested and generated 0.3 mW at excitation frequency of 4.4 kHz with external vibrations, Y, equalling 0.5 mm equating to an acceleration of 382 m/s^2.
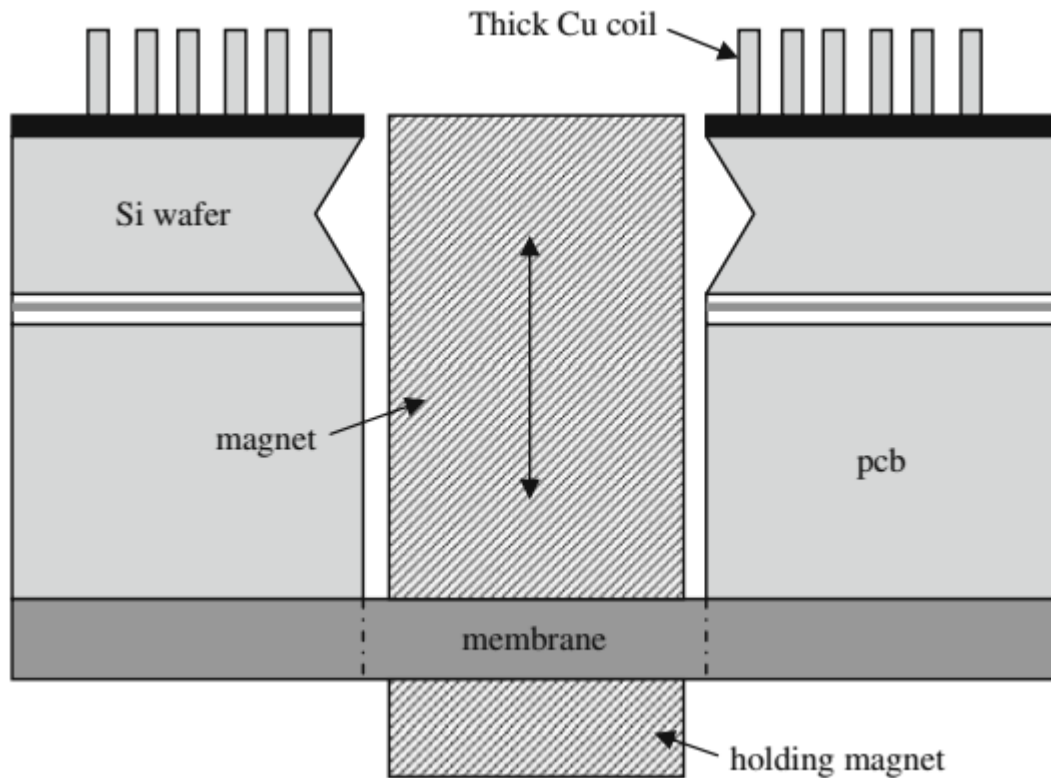
Fig. 13.9. Improved EM generator with electroplated copper coil.

The measured electrical power output was considerably lower than the predicted value and this was thought to be due to the non-linear effects arising from the membrane spring. This exhibited the hard-spring effect where the resonant frequency increases with increasing excitation amplitude.
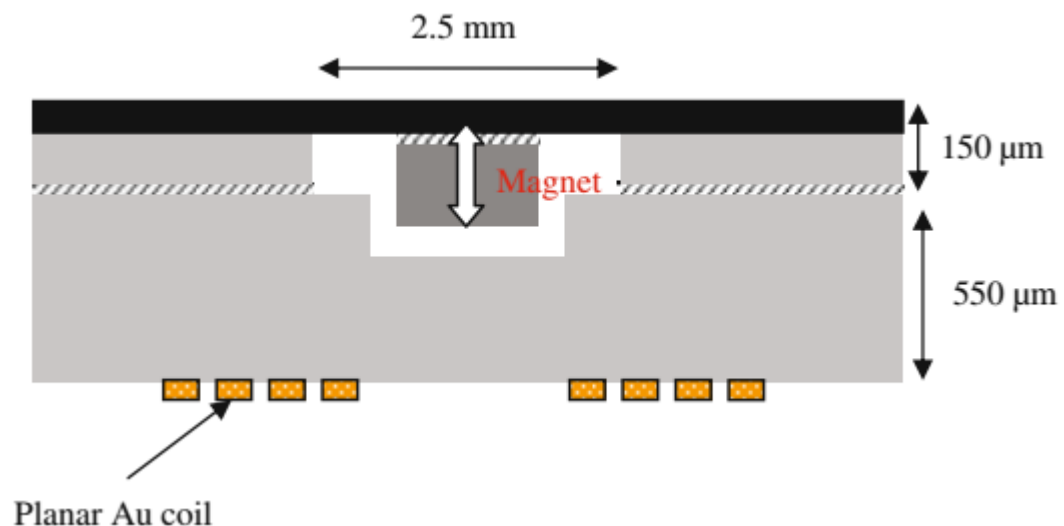
Fig. 13.10. Cross section of the electromagnetic generator

More significantly, however, is the low electromagnetic damping which would reduce the practical power output achieved. The damping could be improved by bringing the coil closer to the magnets. However, the low number of coil turns would always limit the voltage achievable and increasing the number of turns would only result in higher coil resistance and decreased EM damping. This highlights the key fabrication challenge facing microscale electromagnetic devices. Another

consideration highlighted by this design is the high-resonant frequency achieved in practice. This arises from the small device size and is typical for micro-machined generators. Similar designs to the Sheffield generator have been demonstrated by other researchers. A two wafer device that uses an electroplated nickel–iron suspension spring element upon which the magnet is mounted was presented. The second wafer hosts the integrated coils which were fabricated by electroplating copper into a 100m m thick photoresist mould. The coil has an internal resistance of 2Ω although the number of turns is not stated. The device is targeted at generating power from human motion and is reported to generate 0.16 mW from a 'finger tap'. The device has a resonant frequency of 100 Hz and when operated at resonance with an excitation amplitude of 50 mm (19.7 m/s^2) it produces 1.4 mW.

Researchers from the University of Barcelona have described a similar structure. They have bonded a neodymium iron boron (NdFeB) magnet to a polyimide membrane to form the spring–mass system with the polyimide membranes being attached to a PCB square frame around which the coil has been formed. Initial results were from a planar coil made from a 1.5 mm thick aluminium layer. The power output is reported to be 1.44 mW for an excitation displacement of 10 mm and a resonant frequency of 400 Hz, which equates to 63 m/s^2. It was observed that the parasitic damping in this device was far greater than the electromagnetic damping factor and therefore maximum power output was not achieved. This is due to the poor method of assembly introducing large damping effects and the poor electromagnetic coupling factor due to the geometry of the device. A later paper by the group reported on some improvements and the revised design is shown in Fig. 13.9. The device has a top 1 mm thick silicon (Si) (100) wafer which contains the 15mm thick, 20 mm wide electroplated copper tracks which form a coil with 52 turns and a resistance of 100Ω. The spring– mass element is formed by bonding a Kapton membrane between two PCB frames with the NeFeB magnet held at the centre of the membrane by a second smaller holding magnet. The improved version of the device generates about 55 mW at 380 Hz from a 5 mm excitation amplitude which equates to 29 m/s^2. Unfortunately, the device still demonstrates stress stiffening and therefore hysteresis. The use of polymer membranes, such as Kapton, as the spring element in these devices also suffers from the potential for creep and fatigue failure.

A a micro-machined electromagnetic generator that combines an electroplated copper planar spring with an integrated coil was presented. The electroplated spring is formed on a silicon substrate with the silicon being etched away in KOH to leave the freestanding copper structure. A NeFeB permanent magnet is manually glued to the centre of the spring and the overall structure has resonant frequencies of around 55, 121 and 122 Hz. A two layer coil was fabricated on a glass wafer using a combination of spin-coated resists, copper electroplating and polyimide coating to insulate the two layers. The coil and magnet spring layers were simply glued together to assemble the device. Maximum output at modes 2 and 3 where around 60 mV generated from 14.7 m/s^2.

A team of researchers from The Chinese University of Hong Kong have developed a magnet spring arrangement which comprises of a NdFeB magnet mounted on laser-micro-machined spiral copper spring. Meandering and spiral spring designs were simulated and the spiral design was found to have twice the deflection. The springs were laser machined from 110 mm thick copper had diameters ranging from 4 to 10 mm with width/gap dimensions ranging from 40/40m m to 100/100 mm. The spring–mass structure has a resonant frequency of 64 Hz. The wire-wound coil is fixed in position on the housing of the structure. The device has an overall volume of around 1 cm^3 and produces 2 V at its resonant frequency. This is a power output of 10 mW for an excitation amplitude of 100 mm (16.2 m/s^2). A later publication describes a similar generator fabricated on a printed circuit board. An improved spiral spring resulted in a peak-to-peak output voltage of up to 4.4 V and a maximum power of 830 mW when driven by a 200 mm displacement in its third mode with a frequency of 110 Hz (equates to 95.5 m/s^2). Mizuno and Chetwynd (2003) proposed an array of micro-machined cantileverbased electromagnetic generators to boost total power output. Each cantilever element comprises an inertial mass, an integrated coil and is aligned adjacent to a common fixed external magnet. The dimensions of the proposed micro-cantilever beam were 500 mm × 100 mm × 20m m. The theoretical analysis predicted a power output of only 6 nW and a voltage output of 1.4 mV for a typical single cantilever element at a resonant frequency of 58 kHz. For evaluation, the authors fabricated a larger scale version of their proposed device. The size of the cantilever beam was increased to 25 mm × 10 mm × 1 mm which was placed next to a fixed NdFeB magnet whose size was 30 mm × 10 mm × 6 mm. The cantilever contained four coil turns fabricated from 50 m thick gold with a track width of 1 mm. The resonant frequency was 700 Hz. For an input vibration of 0.64 mm (corresponding to 12.4 m/s^2 at 700 Hz), the output power was found to be 320 V and 0.4 nW across a 128 resistive load (which matched the four turn coil resistance). Suggested methods to improve the power output include implementing the array of cantilevers as described initially and reducing the coil magnet gap. The author's conclusion is pessimistic regarding the opportunities for microscale electromagnetic devices, especially considering the low-output voltage. The device presented does indeed lead to this conclusion due, as with the Sheffield generator described above, to the very low-electromagnetic damping achieved with their design. Another two part micro-machined silicon-based generator has been described by Kulah and Najafi (2004). The device consists of two micro machined resonant systems that are bought together in order to achieve mechanical up-frequency conversion. This approach is designed to overcome the high inherent frequency exhibited by the majority of microscale devices which is not compatible with the majority of ambient vibration frequencies which are often less than 100 Hz. The generator is fabricated as two separate chips. The upper chip consists of an NdFeB magnet bonded to a low stiffness parylene diaphragm giving low resonant frequencies in the range of 1–100 Hz. The NdFeB magnet is used to excite a lower structure into resonance through magnetic

attraction. The lower chip consists of an array of resonating cantilevers again fabricated from parylene. These cantilevers also include metal tracks which form the coil. These metal tracks are formed using sputtered nickel the ferromagnetic properties of which also enable the magnetic excitation. The microscale designed cantilevers have a resonant frequency of 11.4 kHz. The analysis of the microscale design predicts a theoretical maximum power of 2.5 mW, however, a larger millimetre scale mock up only delivered 4 nW in practice (the level of input mechanical excitation was not quoted). This mock up validated the frequency conversion principle and little attention was paid to the design of the electromagnetic energy harvesting circuit. One of the main limitations of microscale electromagnetic generators is the limited number of coil turns possible with integrated circuit technology.

Macro-Scale Implementations

El-Hami et al. at the University of Southampton, presented the simulation, modelling, fabrication and characterisation of a cantilever springbased electromagnetic generator. The cantilever beam is clamped at one end and has a pair of NdFeB magnets attached to a c-shaped core located at the free end. The coil is made up of 27 turns of 0.2 mm diameter enamelled copper wire and is fixed in position between the poles of the magnets. The device volume is 240 mm3 and is shown in Fig. 13.11. The device was found to produce 0.53 mW at a vibration frequency of 322 Hz and an excitation amplitude of 25 m ($102 \ m/s^2$).
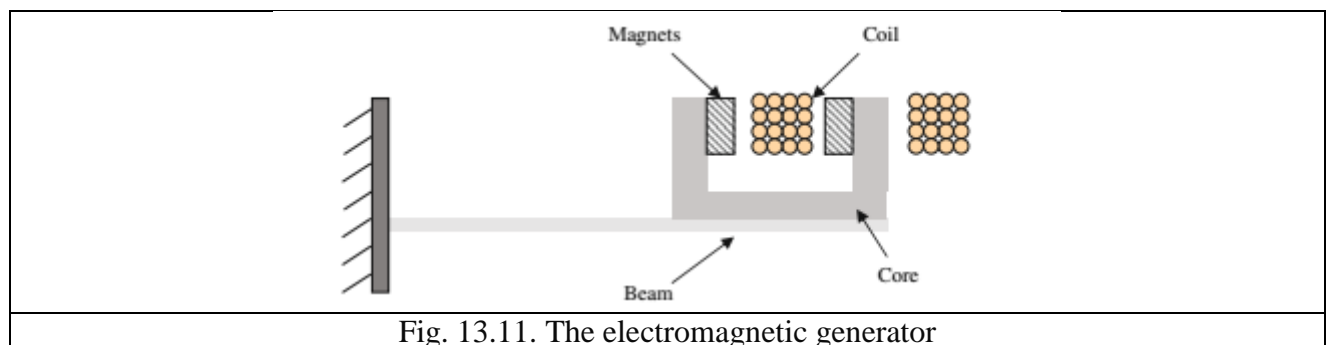
Fig. 13.11. The electromagnetic generator

The team at Southampton further developed the device concentrating on improving the magnetic circuit and therefore the power generated. The team describes a cantilever-based prototype with a four magnet and fixed coil arrangement. The overall generator volume is 3.15 cm^3 and is shown in Fig. 13.12.

At its resonant frequency of 106 Hz and an acceleration level of 2.6 m/s^2, the improved generator produced an output voltage of 1 V, compared with $\sim$ 150 mV for the first arrangement. The generator was mounted on the engine block of a car and its effectiveness naturally depended on the frequency of vibrations which in turn depended on the engine speed. The maximum peak power produced was 4 mW, while the average power was found to be 157 μW over a journey of 1.24 km.
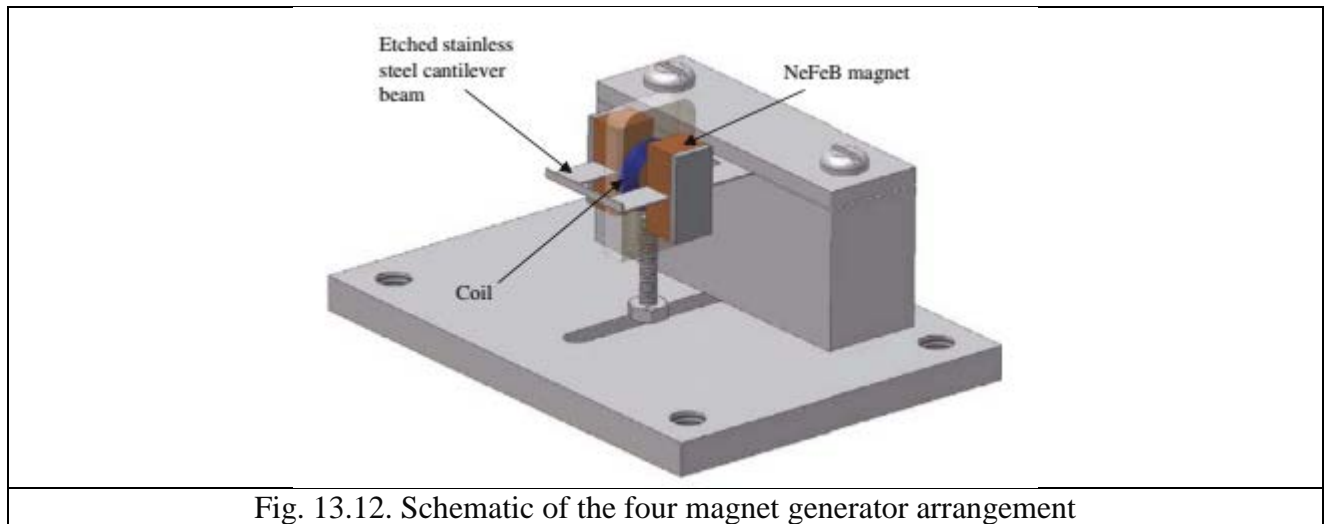
Fig. 13.12. Schematic of the four magnet generator arrangement

This cantilever-based four magnet arrangement has more recently been realised in miniaturised form during an EU funded Framework 6 project 'Vibration Energy Scavenging' (VIBES). The miniature device utilises discrete magnets, conventionally wound coils and machined components. The use of wire erosion and a CNC micro-mill to fabricate these parts has, however, enabled components with sub-millimetre features to be realised. The use of bulk magnets maintains high-flux densities and a coil wound from enamelled copper wire as thin as 12 μm in diameter has resulted in useful levels of energy and power from a device volume of 150 mm^3. The generator produced 46 μW and 428 mV at its resonant frequency of 52 Hz and an acceleration level of 0.59 m/s^2 (excitation amplitude 5.5 μm). Further improvements enabling a degree of manual tunability, increased magnet size and rearrangement of the centre of gravity increased the power output to 58 μW from the same excitation levels. This device is shown in Fig. 13.13.
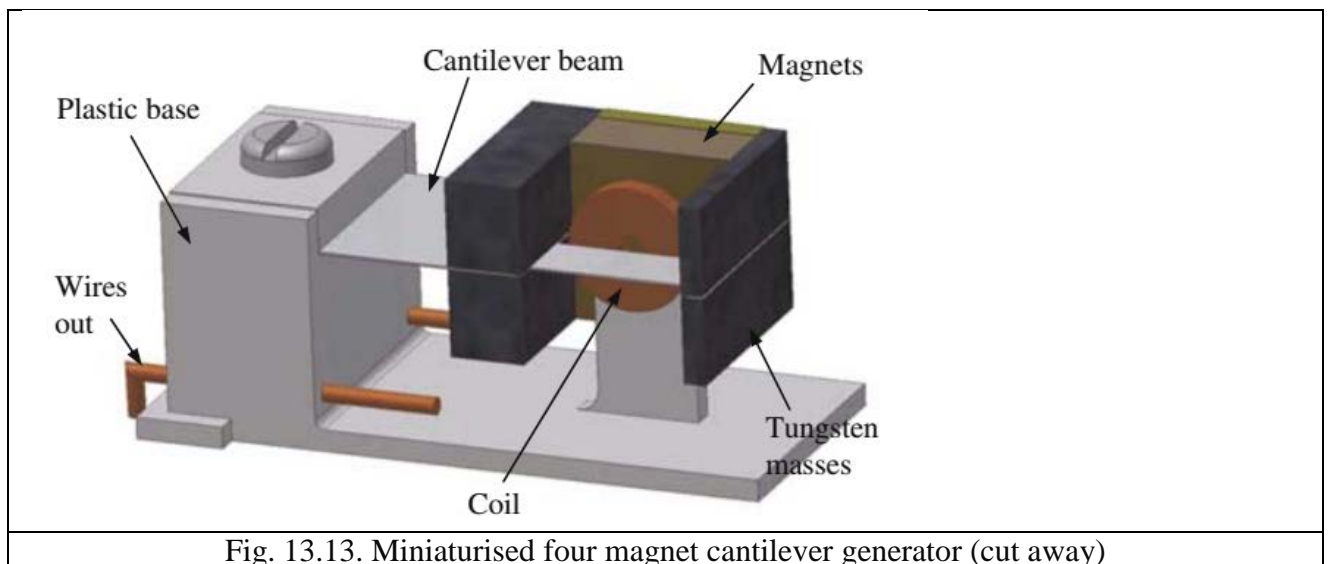
Fig. 13.13. Miniaturised four magnet cantilever generator (cut away)

The design consists of a tubular translator which contains a number of cylindrical magnets separated by spacers. The translator moves vertically within a series of stator coils and the spacing and dimensions of the components was optimised using finite element analysis. The vertical motion of the translator is controlled by a parallel spring stage consisting of two parallel beams. The stator

and translator designs were realised with six magnets, five coils and a volume of 0.5 cm^3, while the total device volume is 30.4 cm^3. The fabricated prototype produced an average power output of 35 mW when located just below a subject's knee.

13.2.3 Electrostatic Generators

Electrostatic generators can utilize a micro-electro-mechanical structure (MEMS) to modulate the capacitance between two surface areas so as to amplify a reference voltage across another fixed value capacitance. The input voltage to the structure is amplified by the mechanical work done against the electrostatic forces between the two plates. It is important to note that electrostatic generators require an initial voltage source to create the initial charge separation for amplification. These converters are not capable of converting as much power per unit volume. However, they are still attractive because they offer more potential for integration with microsystems. There are several types of devices that can transform vibrations into capacitance changes. UC Berkeley developed three types of MEMS structures and evaluated two types referred to as in-plane overlap and in-plane gap. The study concluded that the former required large spring deflections and high Q-factors to produce maximum power, while the latter has an optimum spring deflection and can accommodate lower Q-factors. The Q-factor is of significant interest. A high Q-factor will be more vibration–frequency selective, while a lower Q-factor can accommodate a wider range of vibration frequencies. Therefore, it might become necessary to invent a resonant adaptive structure. A significant portion of any initial design involves carefully characterizing the acceleration profile of vibration sources for the intended application. This determines the design parameters and optimization criteria for a vibration transducer that can deliver maximum energy

Electrostatic transducers can convert mechanical energy into electrical energy by using the input mechanical energy to change capacitance of the generator between a high and low value. The capacitance is most commonly reduced by increasing the distance between the capacitor plates, or by reducing the overlapping area of the plates. Electrostatic transducers operate in one of two modes. The first is constrained charge mode, in which the charge and electric field remain constant on the capacitor plates, causing an increase in voltage as the capacitance reduces. The second is constrained voltage mode, which causes charge to move from the capacitor to a storage device as the capacitance decreases. Figure 13.14 and Table 13.2 compare the differences between the two modes.

Fig. 13.14. An electrostatic transducer shown in the (a) low energy (b) high energy states (C capacitance; E electric field; I current; Q charge; V voltage)
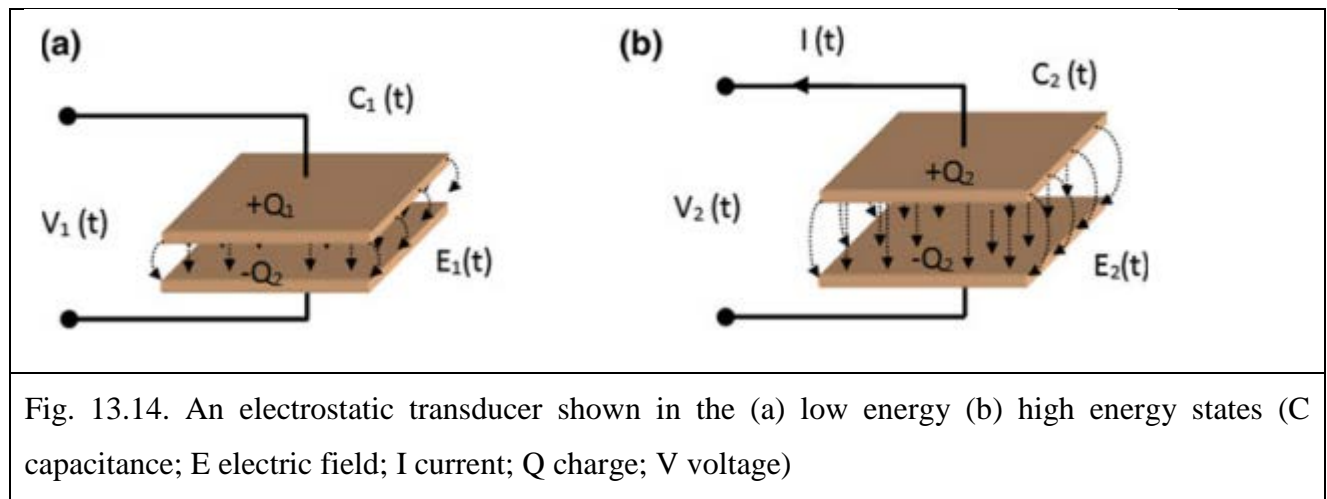
	Constrained Q mode	Constrained V mode
Capacitance	$C1(t) < C2(t)$	$C1(t) < C2(t)$
Charge	$Q1(t) = Q2(t)$,	$Q1(t) > Q2(t)$
Electric field	$E1(t) = E2(t)$	$E1(t) > E2$
Current	$I(t) = 0$	$|I(t)| > 0$
Voltage	$V1(t) < V2(t)$	$V1(t) = V2(t)$

Both methods convert mechanical kinetic energy to electrical energy from work done against the electrostatic forces between the two plates. A control circuit is required to bias and extract the energy from the transducer; the control and bias configuration is dependent on the mode of operation. The use of biasing can be avoided by using electrets, which is a permanent charge buried in a dielectric layer. Electrostatic converters have the significant advantage of easy manufacture from silicon-based MEMS technology with minimal complexity. As mentioned previously, commercial MEMS production companies do not provide the option to include materials such as PZT and permanent magnets in their production processes, although they can easily produce an electrostatic transducer. This means that electrostatic transducers can be externally produced, which allows institutions without production facilities to conduct research and prototyping. The associated disadvantages are the extra complexity and volume required for the bias and energy extracting circuitry, and the high voltages generated in constant charge mode, which can be in the order of thousands of volts to obtain high power generation. Since silicon can withstand high temperatures—single crystal Si melts at 1414 °C—electrostatic transducers can withstand a large range of temperature differentials for heat engine applications. However, the strength of silicon degrades as the temperature increases. Most applications of electrostatic transducers harvest energy from ambient vibration, implementing an inertial mass spring damper system with a designed resonance to target a particular vibration frequency. Three device configurations was presented that can be used for this application: in-plane overlap, where capacitance varies dependent on the overlapping area of fingers; in-plane gap closing, where capacitance varies with the gap between fingers; and out-ofplane gap closing, where capacitance varies by changing gap

distance between two capacitor plates. Roundy et al. discuss the challenges of each configuration, including aspects such as fluid damping. Their dynamic simulations reveal that an output power density of 116 mW/cm3 is possible from a vibration source of 2.25 m/s^2 at 120 Hz.

13.2.4 Reverse Electro-Wetting Generators

Reverse electro-wetting is a method of converting mechanical energy into electrical energy. The principle of its operation was first published in Nature Communications in 2011, but the idea is based on the concept of electro-wetting, which has been recognized for more than a century. In electro-wetting, microdroplets of liquid are dispersed over an insulating surface. When a voltage is applied between a droplet and a conducting surface beneath the insulating layer, the capacitance between the two materials causes the droplet to lower the interfacial energy and flatten out. Reverse electro-wetting is the reverse of the above phenomena, and functions as a liquid dielectric electrostatic transducer. Again, micro-droplets are dispersed across an insulating surface or a micro-channel with a conducting electrode beneath it; this is shown in Fig. 13.15a–c.
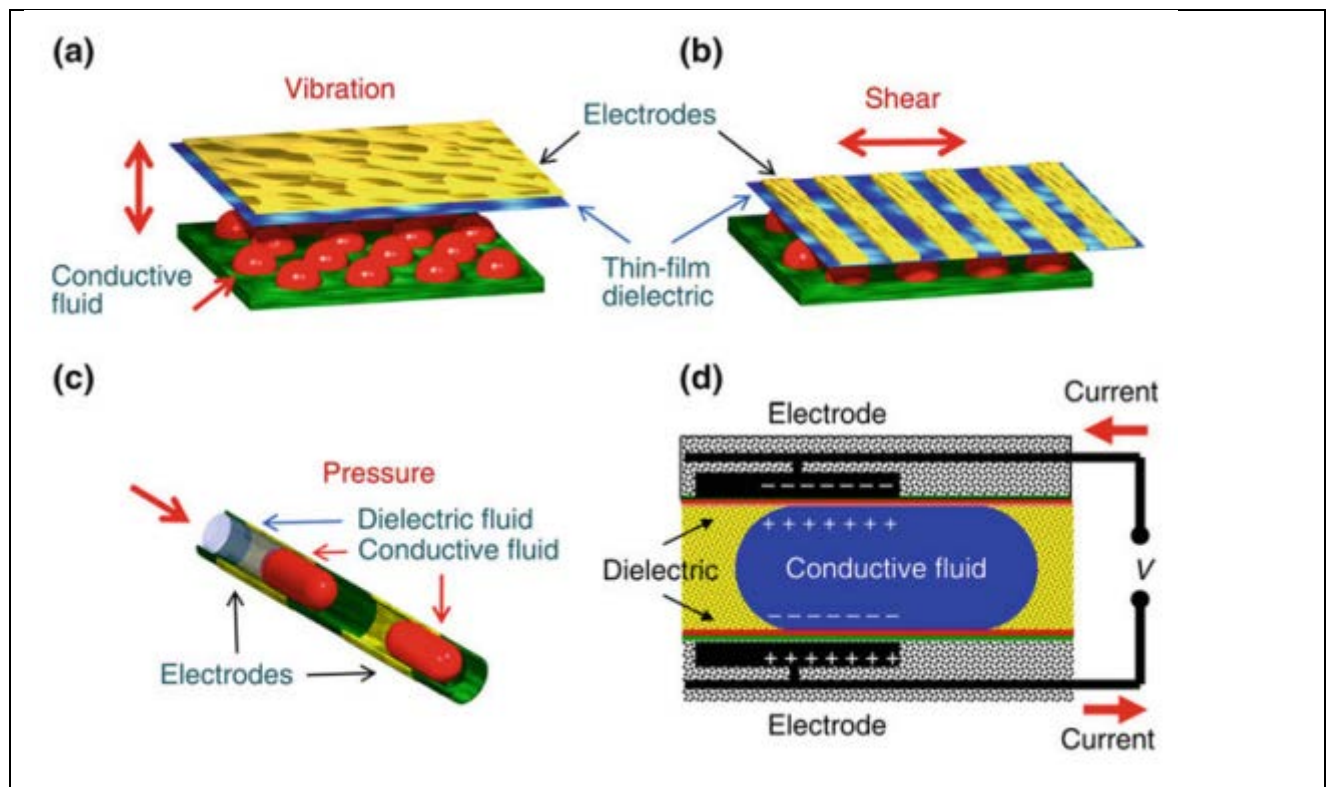
Fig. 13.15. Schematics of three major droplet actuation mechanisms. These include (a) droplets between oscillating plates (b) droplets between sliding plates, and (c) droplets in a micro-channel. (d) shows in greater detail schematics of reverse-electro-wetting-based energy generation process in a micro-channel geometry.

A bias electric field is applied between the droplet and the electrode, causing charges to move in the droplet, which then accumulates at the border with the dielectric layer. External mechanical actuation can then be used to move the droplet such that the area of overlap with the electrode is

decreased. This reduces the capacitance of the field and causes electric current to flow, in the same method as that used in an electrostatic transducer . While no MEMS heat engine using this technique has been demonstrated to date, it may be possible to select the operating temperature of the liquid to allow it to operate as a transducer method over a wide temperature range. Reverse electro-wetting at the time of writing is still in the very early stages of development. As such, it is yet to be successfully applied to the field of thermal energy harvesting. Typical energies as high as 0.4 mJ/mm^2 per oscillation was demonstrated. A design for development of the energy harvesting device, with a projected application being human movement energy harvesting was presented. Energy generation from this device is approximately 10 W using a bias voltage of 35 V. Possible thermal harvesting applications could include harvesting energy from fluid convection flow, or coupled with oscillatory systems.

Chapter 10. Fuel Cells

10.1. INTRODUCTION

A fuel cell is a device that converts the chemical energy from a fuel into electricity through a chemical reaction of positively charged hydrogen ions with oxygen or another oxidizing agent [1]. Fuel cells are different from batteries in that they require a continuous source of fuel and oxygen or air to sustain the chemical reaction, whereas in a battery the chemicals present in the battery react with each other to generate an electromotive force (emf). Fuel cells can produce electricity continuously for as long as these inputs are supplied.

The first fuel cells were invented in 1838. The first commercial use of fuel cells came more than a century later in NASA space programs to generate power for satellites and space capsules. Since then, fuel cells have been used in many other applications. Fuel cells are used for primary and backup power for commercial, industrial and residential buildings and in remote or inaccessible areas. They are also used to power fuel cell vehicles, including forklifts, automobiles, buses, boats, motorcycles and submarines.

There are many types of fuel cells, but they all consist of an anode, a cathode, and an electrolyte that allows positively charged hydrogen ions (or protons) to move between the two sides of the fuel cell. The anode and cathode contain catalysts that cause the fuel to undergo oxidation reactions that generate positively charged hydrogen ions and electrons. The hydrogen ions are drawn through the electrolyte after the reaction. At the same time, electrons are drawn from the anode to the cathode through an external circuit, producing direct current electricity. At the cathode, hydrogen ions, electrons, and oxygen react to form water. As the main difference among fuel cell types is the electrolyte, fuel cells are classified by the type of electrolyte they use and by the difference in startup time ranging from 1 second for proton exchange membrane fuel cells (PEM fuel cells, or PEMFC) to 10 minutes for solid oxide fuel cells (SOFC). Individual fuel cells produce relatively small electrical potentials, about 0.7 volts, so cells are "stacked", or placed in series, to create sufficient voltage to meet an application's requirements [2]. In addition to electricity, fuel cells produce water, heat and, depending on the fuel source, very small amounts of nitrogen dioxide and other emissions. The energy efficiency of a fuel cell is generally between 40–60%, or up to 85% efficient in cogeneration if waste heat is captured for use.

The fuel cell market is growing, and Pike Research has estimated that the stationary fuel cell market will reach 50 GW by 2020 [3].

10.2. TYPES OF FUEL CELLS

Fuel cells come in many varieties; however, they all work in the same general manner (Fig. 10.1). They are made up of three adjacent segments: the anode, the electrolyte, and the cathode (Fig. 10.2). Two chemical reactions occur at the interfaces of the three different segments. The net result of the two reactions is that fuel is consumed, water or carbon dioxide is created, and an electric current is created, which can be used to power electrical devices, normally referred to as the load.
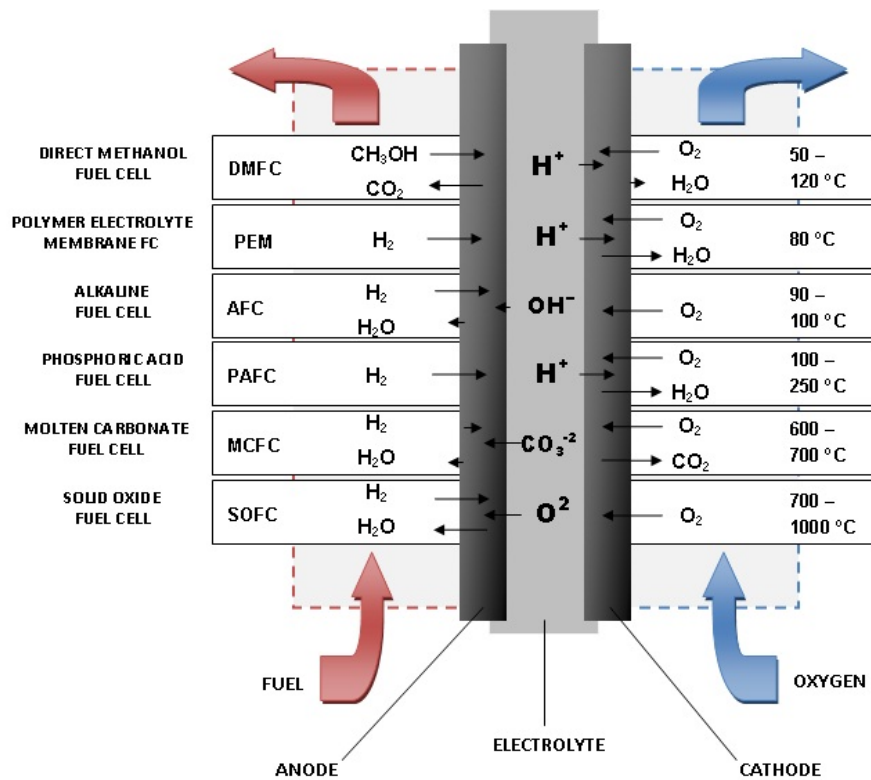
Fig. 10.1. Types of fuel cells.
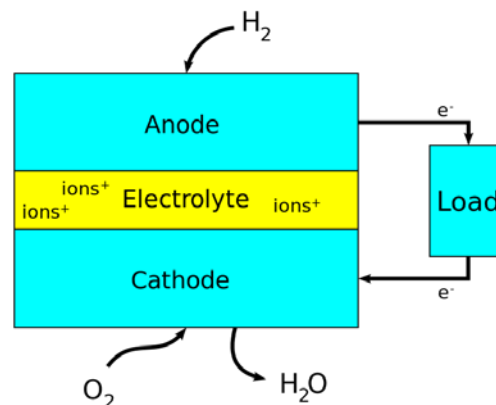
Fig. 10.2. A block diagram of a fuel cell.

At the anode a catalyst oxidizes the fuel, usually hydrogen, turning the fuel into a positively charged ion and a negatively charged electron. The electrolyte is a substance specifically designed so ions can pass through it, but the electrons cannot. The freed electrons travel through a wire creating the electric current. The ions travel through the electrolyte to the cathode. Once reaching

the cathode, the ions are reunited with the electrons and the two react with a third chemical, usually oxygen, to create water or carbon dioxide.

The most important design features in a fuel cell are:

- The electrolyte substance. The electrolyte substance usually defines the *type* of fuel cell.

- The fuel that is used. The most common fuel is hydrogen.

- The anode catalyst breaks down the fuel into electrons and ions. The anode catalyst is usually made up of very fine platinum powder.

- The cathode catalyst turns the ions into the waste chemicals like water or carbon dioxide. The cathode catalyst is often made up of nickel but it can also be a nanomaterial-based catalyst.

A typical fuel cell produces a voltage from 0.6 V to 0.7 V at full rated load. Voltage decreases as current increases, due to several factors:

- Activation loss

- Ohmic loss (voltage drop due to resistance of the cell components and interconnections)

- Mass transport loss (depletion of reactants at catalyst sites under high loads, causing rapid loss of voltage) [4].

To deliver the desired amount of energy, the fuel cells can be combined in series to yield higher voltage, and in parallel to allow a higher current to be supplied. Such a design is called a *fuel cell stack*. The cell surface area can also be increased, to allow higher current from each cell. Within the stack, reactant gases must be distributed uniformly over each of the cells to maximize the power output [5-7].

10.2.1. Proton exchange membrane fuel cells (PEMFCs)

Proton exchange membrane fuel cells, also known as polymer electrolyte membrane (PEM) fuel cells (PEMFC), are a type of fuel cell being developed for transport applications as well as for stationary fuel cell applications and portable fuel cell applications (Fig. 10.3). Their distinguishing features include lower temperature/pressure ranges (50 to 100 C) and a special polymer electrolyte membrane. PEMFCs operate on a similar principle to their younger sister technology PEM electrolysis. They are a leading candidate to replace the aging alkaline fuel cell technology, which was used in the Space Shuttle [8].

Proton exchange membrane fuel cell

1 Hydrogen fuel is channeled through field flow plates to the anode on one side of the fuel cell, while oxidant (oxygen or air) is channeled to the cathode on the other side of the cell.

2 At the anode, a platinum catalyst causes the hydrogen to split into positive hydrogen ions (protons) and negatively charged electrons.

3 The polymer electrolyte membrane (PEM) allows only the positively charged ions to pass through it to the cathode. The negatively charged electrons must travel along an external circuit to the cathode, creating an electrical current.

4 At the cathode, the electrons and positively charged hydrogen ions combine with oxygen to form water, which flows out of the cell.

Backing layers
Hydrogen gas
Oxidant
Oxidant flow field
Hydrogen flow field
Unused fuel
Water
Anode (negative)
Cathode (positive)
Polymer electrolyte membrane

Fig. 10.3. Construction of a high-temperature PEMFC: Bipolar plate as electrode with in-milled gas channel structure, fabricated from conductive composites (enhanced with graphite, carbon black, carbon fiber, and/or carbon nanotubes for more conductivity); Porous carbon papers; reactive layer, usually on the polymer membrane applied; polymer membrane.

PEMFCs are built out of membrane electrode assemblies (MEA) which include the electrodes, electrolyte, catalyst, and gas diffusion layers. An ink of catalyst, carbon, and electrode are sprayed or painted onto the solid electrolyte and carbon paper is hot pressed on either side to protect the inside of the cell and also act as electrodes. The pivotal part of the cell is the triple phase boundary (TPB) where the electrolyte, catalyst, and reactants mix and thus where the cell reactions actually occur [9]. Importantly, the membrane must not be electrically conductive so the half reactions do not mix. Operating temperatures above 100 °C are desired so the water byproduct becomes steam and water management becomes less critical in cell design.

Reactions. A proton exchange membrane fuel cell transforms the chemical energy liberated during the electrochemical reaction of hydrogen and oxygen to electrical energy, as opposed to the direct combustion of hydrogen and oxygen gases to produce thermal energy. A stream of hydrogen is delivered to the anode side of the membrane electrode assembly (MEA). At the anode side it is catalytically split into protons and electrons. This oxidation half-cell reaction or hydrogen oxidation reaction (HOR) is represented by:

At the anode:

$$\mathrm{H_2 \rightarrow 2H^+ + 2e^-} \qquad E^\circ = 0\,\mathrm{V} \quad \frac{\mathrm{d}E^\circ}{\mathrm{d}T} = 0\,\mathrm{mV\,K^{-1}} \qquad (10.1)$$
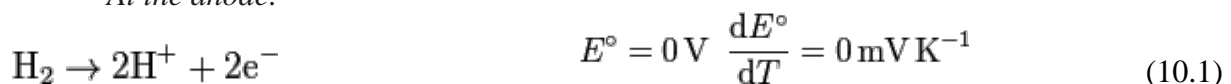
The newly formed protons permeate through the polymer electrolyte membrane to the cathode side. The electrons travel along an external load circuit to the cathode side of the MEA, thus creating the current output of the fuel cell. Meanwhile, a stream of oxygen is delivered to the cathode side of the MEA. At the cathode side oxygen molecules react with the protons permeating through the polymer electrolyte membrane and the electrons arriving through the external circuit to form water molecules. This reduction half-cell reaction or oxygen reduction reaction (ORR) is represented by:

At the cathode:

$$\frac{1}{2}O_2 + 2H^+ + 2e^- \rightarrow H_2O \qquad E^\circ = 1.2291\,V \quad \frac{dE^\circ}{dT} = -0.8456\,mV\,K^{-1} \qquad (10.2)$$

Overall reaction:

$$H_2 + \frac{1}{2}O_2 \rightarrow H_2O \qquad E^\circ = 1.2291\,V \quad \frac{dE^\circ}{dT} = -0.8456\,mV\,K^{-1} \qquad (10.3)$$

The reversible reaction is expressed in the equation and shows the reincorporation of the hydrogen protons and electrons together with the oxygen molecule and the formation of one water molecule. The potentials in each case are given with respect to the standard hydrogen electrode.

Polymer electrolyte membrane. To function, the membrane must conduct hydrogen ions (protons) but not electrons as this would in effect "short circuit" the fuel cell. The membrane must also not allow either gas to pass to the other side of the cell, a problem known as gas crossover [10, 11]. Finally, the membrane must be resistant to the reducing environment at the cathode as well as the harsh oxidative environment at the anode. Splitting of the hydrogen molecule is relatively easy by using a platinum catalyst. Unfortunately however, splitting the oxygen molecule is more difficult, and this causes significant electric losses. An appropriate catalyst material for this process has not been discovered, and platinum is the best option.

A cheaper alternative to platinum is Cerium(IV) oxide catalyst [12-14]. The PEMFC is a prime candidate for vehicle and other mobile applications of all sizes down to mobile phones, because of its compactness. However, the water management is crucial to performance: too much water will flood the membrane, too little will dry it; in both cases, power output will drop. Water management is a very difficult subject in PEM systems, primarily because water in the membrane is attracted toward the cathode of the cell through polarization. A wide variety of solutions for managing the water exist including integration of electroosmotic pumps. Furthermore, the platinum catalyst on the membrane is easily poisoned by carbon monoxide (no more than one part per million is usually acceptable) and the membrane is sensitive to things like metal ions, which can be introduced by corrosion of metallic bipolar plates, metallic components in the fuel cell system or from contaminants in the fuel/oxidant.

PEM systems that use reformed methanol were proposed, as in Daimler Chrysler Necar 5; reforming methanol, i.e. making it react to obtain hydrogen, is however a very complicated process, that requires also purification from the carbon monoxide the reaction produces. A platinum-ruthenium catalyst is necessary as some carbon monoxide will unavoidably reach the membrane. The level should not exceed 10 parts per million. Furthermore, the start-up times of such a reformer reactor are of about half an hour. Alternatively, methanol, and some other biofuels can be fed to a PEM fuel cell directly without being reformed, thus making a direct methanol fuel cell (DMFC). These devices operate with limited success.

The most commonly used membrane is Nafion by DuPont, which relies on liquid water humidification of the membrane to transport protons. This implies that it is not feasible to use temperatures above 80 to 90 °C, since the membrane would dry. Other, more recent membrane types, based on polybenzimidazole (PBI) or phosphoric acid, can reach up to 220 °C without using any water management: higher temperature allow for better efficiencies, power densities, ease of cooling (because of larger allowable temperature differences), reduced sensitivity to carbon monoxide poisoning and better controllability (because of absence of water management issues in the membrane); however, these recent types are not as common [15]. PBI can be doped with phosphoric or sulfuric acid and the conductivity scales with amount of doping and temperature [16]. At high temperatures, it is difficult to keep Nafion hydrated, but this acid doped material does not use water as a medium for proton conduction. It also exhibits better mechanical properties, higher strength, than Nafion and is cheaper. However, acid leaching is a considerable issue and processing, mixing with catalyst to form ink, has proved tricky. Aromatic polymers, such as PEEK, are far cheaper than Teflon (PTFE and backbone of Nafion) and their polar character leads to hydration that is less temperature dependent than Nafion. However, PEEK is far less ionically conductive than

Nafion and thus is a less favorable electrolyte choice [17]. Recently, protic ionic liquids and protic organic ionic plastic crystals have been shown as promising alternative electrolyte materials for high temperature (100–200°C) PEMFCs [18-19].

Electrodes. In order to enable the electrochemical reactions at the electrodes, protons, electrons and the reactant gases (hydrogen or oxygen) must gain access to the surface of the catalyst in the electrodes, while the product water must be able to permeate from the catalyst to the gas outlet. These properties are typically realized by porous composits of polymer electrolyte binder and catalyst nanoparticles supported on carbon particles [20]. Typically platinum is used as the catalyst for the electrochemical reactions at the anode and cathode, while nanoparticles realize high surface to weight ratios (as further described below) reducing the amount of the costly platinum. The polymer electrolyte binder provides the ionic conductivity, while the carbon support of the catalyst improves the electric conductivity and enables low platinum metal loading [21]. The electric conductivity in the composite electrodes is typically more than 40 times higher as the proton conductivity [22].

Gas diffusion layer. The GDL electrically connects the catalyst and current collector. It must be porous, electrically conductive, and thin. The reactants must be able to reach the catalyst, but conductivity and porosity can act as opposing forces. Optimally, the GDL should be composed of about one third Nafion or 15% PTFE. The carbon particles used in the GDL can be larger than those employed in the catalyst because surface area is not the most important variable in this layer. GDL should be around 15-35 µm thick to balance needed porosity with mechanical strength [23]. Often, an intermediate porous layer is added between the GDL and catalyst layer to ease the transitions between the large pores in the GDL and small porosity in the catalyst layer. Since a primary function of the GDL is to help remove water, a product, flooding can occur when water effectively blocks the GDL. This limits the reactants ability to access the catalyst and significantly decreases performance. Teflon can be coated onto the GDL to limit the possibility of flooding [17]. Several microscopic variables are analyzed in the GDLS such as: porosity, tortuosity and permeability. These variables have incidence over the behavior of the fuel cells [24].

Efficiency. The maximal theoretical efficiency applying the Gibbs free energy equation ΔG=-237.13 kJ/mol and using the heating value of Hydrogen (ΔH=-285.84 kJ/mol) is 83% at 298 K.

$$\eta = \frac{\Delta G}{\Delta H} = 1 - \frac{T\Delta S}{\Delta H}$$

The practical efficiency of a PEM's is in the range of 40–60% . Main factors that create losses are:
- Activation losses
- Ohmic losses
- Mass transport losses

Metal-Organic Frameworks. Metal-organic frameworks (MOFs) are a relatively new class of porous, highly crystalline materials that consist of metal nodes connected by organic linkers. Due to the simplicity of manipulating or substituting the metal centers and ligands, there are a virtually limitless number of possible combinations, which is attractive from a design standpoint. MOFs exhibit many unique properties due to their tunable pore sizes, thermal stability, high volume capacities, large surface areas, and desirable electrochemical characteristics. Among their many diverse uses, MOFs are promising candidates for clean energy applications such as hydrogen storage, gas separations, supercapacitors, Li-ion batteries, solar cells, and fuel cells. Within the field of fuel cell research, MOFs are being studied as potential electrolyte materials and electrode catalysts that could someday replace traditional polymer membranes and Pt catalysts, respectively.

As electrolyte materials, the inclusion of MOFs seems at first counter-intuitive. Fuel cell membranes generally have low porosity to prevent fuel crossover and loss of voltage between the anode and cathode. Additionally, membranes tend to have low crystallinity because the transport of ions is more favorable in disordered materials. On the other hand, pores can be filled with additional ion carriers that ultimately enhance the ionic conductivity of the system and high crystallinity makes the design process less complex.

The general requirements of a good electrolyte for PEMFCs are: high proton conductivity ($>10^{-2}$ S/cm for practical applications) to enable proton transport between electrodes, good chemical and thermal stability under fuel cell operating conditions (environmental humidity, variable temperatures, resistance to poisonous species, etc.), low cost, ability to be processed into thin-films, and overall compatibility with other cell components [25]. While polymeric materials are currently the preferred choice of proton-conducting membrane, they require humidification for adequate performance and can sometimes physically degrade due to hydrations effects, thereby causing losses of efficiency. As mentioned, Nafion is also limited by a dehydration temperature of $< 100\ °C$, which can lead to slower reaction kinetics, poor cost efficiency, and CO poisoning of Pt electrode catalysts. Conversely, MOFs have shown encouraging proton conductivities in both low and high temperature regimes as well as over a wide range of humidity conditions. Below 100 °C and under hydration, the presence of hydrogen bonding and solvent water molecules aid in proton transport, whereas anhydrous conditions are suitable for temperatures above 100 °C. MOFs also have the distinct advantage of exhibiting proton conductivity by the framework itself in addition to the inclusion of charge carries (i.e., water, acids, etc.) into their pores.

A low temperature example is work by Kitagawa, et. al. who used a two-dimensional oxalate-bridged anionic layer framework as the host and introduced ammonium cations and adipic acid molecules into the pores to increase proton concentration [26]. The result was one of the first instances of a MOF showing "superprotonic" conductivity (8 x 10^{-3} S/cm) at 25 °C and 98% relative humidity (RH). They later found that increasing the hydrophilic nature of the cations introduced into the pores could enhance proton conductivity even more. In this low temperature regime that is dependent on degree of hydration, it has also been shown that proton conductivity is heavily dependent on humidity levels.

A high temperature anhydrous example is PCMOF2, which consists of sodium ions coordinated to a trisulfonated benzene derivative [27]. To improve performance and allow for higher operating temperatures, water can be replaced as the proton carrier by less volatile imidazole or triazole molecules within the pores. The maximum temperature achieved was 150 °C with an optimum conductivity of 5 x 10^{-4} S/cm, which is lower than other current electrolyte membranes. However, this model holds promise for its temperature regime, anhydrous conditions, and ability to control the quantity of guest molecules within the pores, all of which allowed for the tunability of proton conductivity. Additionally, the triazole-loaded PCMOF2 was incorporated into a H_2/air membrane-electrode assembly and achieved an open circuit voltage of 1.18 V at 100 °C that was stable for 72 hours and managed to remain gas tight throughout testing. This was the first instance that proved MOFs could actually be implemented into functioning fuel cells, and the moderate potential difference showed that fuel crossover due to porosity was not an issue.

To date, the highest proton conductivity achieved for a MOF electrolyte is 4.2 x 10^{-2} S/cm at 25 °C under humid conditions (98% RH), which is competitive with Nafion [25]. Some recent experiments have even successfully produced thin-film MOF membranes instead of the traditional bulk samples or single crystals, which is crucial for their industrial applicability. Once MOFs are able to consistently achieve sufficient conductivity levels, mechanical strength, water stability, and simple processing, they have the potential to play an important role in PEMFCs in the near future. MOFs have also been targeted as potential replacements of platinum group metal (PGM) materials for electrode catalysts, although this research is still in the early stages of development. In PEMFCs, the oxygen reduction reaction (ORR) at the Pt cathode is significantly slower than the fuel oxidation reaction at the anode, and thus non-PGM and metal-free catalysts are being investigated as alternatives. The high volumetric density, large pore surface areas, and openness of metal-ion sites in MOFs make them ideal candidates for catalyst precursors [28]. Despite promising catalytic

abilities, the durability of these proposed MOF-based catalysts is currently less than desirable and the ORR mechanism in this context is still not completely understood.

Catalyst research. Much of the current research on catalysts for PEM fuel cells can be classified as having one of following main objectives:

- to obtain higher catalytic activity than the standard carbon-supported platinum particle catalysts used in current PEM fuel cells
- to reduce the poisoning of PEM fuel cell catalysts by impurity gases
- to reduce the cost of the fuel due to use of platinum-based catalysts

Increasing catalytic activity. As mentioned above, platinum is by far the most effective element used for PEM fuel cell catalysts, and nearly all current PEM fuel cells use platinum particles on porous carbon supports to catalyze both hydrogen oxidation and oxygen reduction. However, due to their high cost, current Pt/C catalysts are not feasible for commercialization. The U.S. Department of Energy estimates that platinum-based catalysts will need to use roughly four times less platinum than is used in current PEM fuel cell designs in order to represent a realistic alternative to internal combustion engines [29]. Consequently, one main goal of catalyst design for PEM fuel cells is to increase the catalytic activity of platinum by a factor of four so that only one-fourth as much of the precious metal is necessary to achieve similar performance.

One method of increasing the performance of platinum catalysts is to optimize the size and shape of the platinum particles. Decreasing the particles' size alone increases the total surface area of catalyst available to participate in reactions per volume of platinum used, but recent studies have demonstrated additional ways to make further improvements to catalytic performance. For example, one study reports that high-index facets of platinum nanoparticles (that is Miller indexes with large integers, such as Pt (730)) provide a greater density of reactive sites for oxygen reduction than typical platinum nanoparticles [30].

Since the most common and effective catalyst, platinum, is extremely expensive, alternative processing is necessary to maximize surface area and minimize loading. Deposition of nanosized Pt particles onto carbon powder (Pt/C) provides a large Pt surface area while the carbon allows for electrical connection between the catalyst and the rest of the cell. Platinum is so effective because it has high activity and bonds to the hydrogen just strongly enough to facilitate electron transfer but not inhibit the hydrogen from continuing to move around the cell. However, platinum is less active in the cathode oxygen reduction reaction. This necessitates the use of more platinum, increasing the cell's expense and thus feasibility. Many potential catalyst choices are ruled out because of the extreme acidity of the cell [17].

The most effective ways of achieving the nanoscale Pt on carbon powder, which is currently the best option, are through vacuum deposition, sputtering, and electrodeposition. The platinum particles are deposited onto carbon paper that is permeated with PTFE. However, there is an optimal thinness to this catalyst layer, which limits the lower cost limit. Below 4 nm, Pt will form islands on the paper, limiting its activity. Above this thickness, the Pt will coat the carbon and be an effective catalyst. To further complicate things, Nafion cannot be infiltrated beyond 10 µm, so using more Pt than this is an unnecessary expense. Thus the amount and shape of the catalyst is limited by the constraints of other materials [23].

A second method of increasing the catalytic activity of platinum is to alloy it with other metals. For example, it was recently shown that the Pt₃Ni(111) surface has a higher oxygen reduction activity than pure Pt(111) by a factor of ten [31]. The authors attribute this dramatic performance increase to modifications to the electronic structure of the surface, reducing its tendency to bond to oxygen-containing ionic species present in PEM fuel cells and hence increasing the number of available sites for oxygen adsorption and reduction.

Further efficiencies can be realized using an Ultrasonic nozzle to apply the platinum catalyst to the electrolyte layer or to carbon paper under atmospheric conditions resulting in high efficiency spray [32]. Studies have shown that due to the uniform size of the droplets created by this type of spray, due to the high transfer efficiency of the technology, due to the non-clogging nature of the

nozzle and finally due to the fact that the ultrasonic energy de-agglomerates the suspension just before atomization, fuel cells MEA's manufactured this way have a greater homogeneity in the final MEA, and the gas flow through the cell is more uniform, maximizing the efficiency of the platinum in the MEA [33]. Recent studies using inkjet printing to deposit the catalyst over the membrane have also shown high catalyst utilization due to the reduced thickness of the deposited catalyst layers [34].

Very recently, a new class of ORR electrocatalysts have been introduced in the case of Pt-M (M-Fe [35] and Co) systems with an ordered intermetallic core encapsulated within a Pt-rich shell [36]. These *intermetallic core-shell (IMCS) nanocatalysts* were found to exhibit an enhanced activity and most importantly, an extended durability compared to many previous designs. While the observed enhancement in the activities is ascribed to a strained lattice, the authors report that their findings on the degradation kinetics establish that the extended catalytic durability is attributable to a sustained atomic order.

Reducing poisoning. The other popular approach to improving catalyst performance is to reduce its sensitivity to impurities in the fuel source, especially carbon monoxide (CO). Presently, pure hydrogen gas is not economical to mass-produce by electrolysis or any other means. Instead, hydrogen gas is produced by steam reforming light hydrocarbons, a process which produces a mixture of gases that also contains CO (1–3%), CO_2 (19–25%), and N_2 (25%) [37]. Even tens of parts per million of CO can poison a pure platinum catalyst, so increasing platinum's resistance to CO is an active area of research.

For example, one study reported that cube-shaped platinum nanoparticles with (100) facets displayed a fourfold increase in oxygen reduction activity compared to randomly faceted platinum nanoparticles of similar size [38]. The authors concluded that the (111) facets of the randomly shaped nanoparticles bonded more strongly to sulfate ions than the (100) facets, reducing the number of catalytic sites open to oxygen molecules. The nanocubes they synthesized, in contrast, had almost exclusively (100) facets, which are known to interact with sulfate more weakly. As a result, a greater fraction of the surface area of those particles was available for the reduction of oxygen, boosting the catalyst's oxygen reduction activity.

In addition, researchers have been investigating ways of reducing the CO content of hydrogen fuel before it enters a fuel cell as a possible way to avoid poisoning the catalysts. One recent study revealed that ruthenium-platinum core–shell nanoparticles are particularly effective at oxidizing CO to form CO_2, a much less harmful fuel contaminant [39]. The mechanism that produces this effect is conceptually similar to that described for Pt_3Ni above: the ruthenium core of the particle alters the electronic structure of the platinum surface, rendering it better able to catalyze the oxidation of CO.

Lowering cost. The challenge for the viability of PEM fuel cells today still remains in their cost and stability. The high cost can in large part be attributed to the use of the precious metal of platinum in the catalyst layer of PEM cells. The electrocatalyst currently accounts for nearly half of the fuel cell stack cost [40]. Although the Pt loading of PEM fuel cells has been reduced by two orders of magnitude over the past decade [41], further reduction is necessary to make the technology economically viable for commercialization. Whereas some research efforts aim to address this issue by improving the electrocatalytic activity of Pt-based catalysts, an alternative is to eliminate the use of Pt altogether by developing a non-platinum-group-metal (non-PGM) cathode catalyst whose performance rivals that of Pt-based technologies. The U.S. Department of Energy has been setting milestones for the development of fuel cells, targeting a durability of 5000 hours and a non-PGM catalyst ORR volumetric activity of 300 A cm^{-3} [42].

Promising alternatives to Pt-based catalysts are Metal/Nitrogen/ Carbon-catalysts (M/N/C-catalysts). To achieve high power density, or output of power over surface area of the cell, a volumetric activity of at least 1/10 that of Pt-based catalysts must be met [40], along with good mass transport properties. While M/N/C-catalysts still demonstrate poorer volumetric activities than Pt-based catalysts, the reduced costs of such catalysts allows for greater loading to compensate. However, increasing the loading of M/N/C-catalysts also renders the catalytic layer thicker, impairing its mass transport properties. In other words, H_2, O_2, protons, and electrons have greater

difficulty in migrating through the catalytic layer, decreasing the voltage output of the cell. While high microporosity of the M/N/C catalytic network results in high volumetric activity, improved mass transport properties are instead associated to macroporosity of the network. These M/N/C materials are synthesized using high temperature pyrolysis and other high temperature treatments of precursors containing the metal, nitrogen, and carbon.

Recently, researchers have developed a Fe/N/C catalyst derived from iron (II) acetate (FeAc), phenanthroline (Phen), and a metal-organic-framework (MOF) host. The MOF is a Zn(II) zeolitic imidazolate framework (ZIF) called ZIF-8 , which demonstrates a high microporous surface area and high nitrogen content conducive to ORR activity [40]. The power density of the FeAc/Phen/ZIF-8-catalyst was found to be 0.75 W cm^{-2} at 0.6 V. This value is a significant improvement over the maximal 0.37 W cm^{-2} power density of previous M/N/C-catalysts and is much closer to matching the typical value of 1-1.2 W cm^{-2} for Pt-based catalysts with a Pt loading of 0.3 mg cm^{-2}. The catalyst also demonstrated a volumetric activity of 230 A cm^{-3}, the highest value for non-PGM catalysts to date [40] approaching the U.S. Department of Energy milestone.

While the power density achieved by the novel FeAc/Phen/ZIF-8-catalyst is promising , its durability remains inadequate for commercial application. It is reported that the best durability exhibited by this catalyst still had a 15% drop in current density over 100 hours in H$_2$/air [40]. Hence while the Fe-based non-PGM catalysts rival Pt-based catalysts in their electrocatalytic activity, there is still much work to be done in improving their durability.

10.2.2. Phosphoric acid fuel cell (PAFC)

Phosphoric acid fuel cells (PAFC) (Fig. 10.4) were first designed and introduced in 1961 by G. V. Elmore and H. A. Tanner. In these cells phosphoric acid is used as a non-conductive electrolyte to pass positive hydrogen ions from the anode to the cathode. These cells commonly work in temperatures of 150 to 200 degrees Celsius. This high temperature will cause heat and energy loss if the heat is not removed and used properly. This heat can be used to produce steam for air conditioning systems or any other thermal energy consuming system [43]. Using this heat in cogeneration can enhance the efficiency of phosphoric acid fuel cells from 40–50% to about 80% [44]. Phosphoric acid, the electrolyte used in PAFCs, is a non-conductive liquid acid which forces electrons to travel from anode to cathode through an external electrical circuit. Since the hydrogen ion production rate on the anode is small, platinum is used as catalyst to increase this ionization rate. A key disadvantage of these cells is the use of an acidic electrolyte. This increases the corrosion or oxidation of components exposed to phosphoric acid [44].
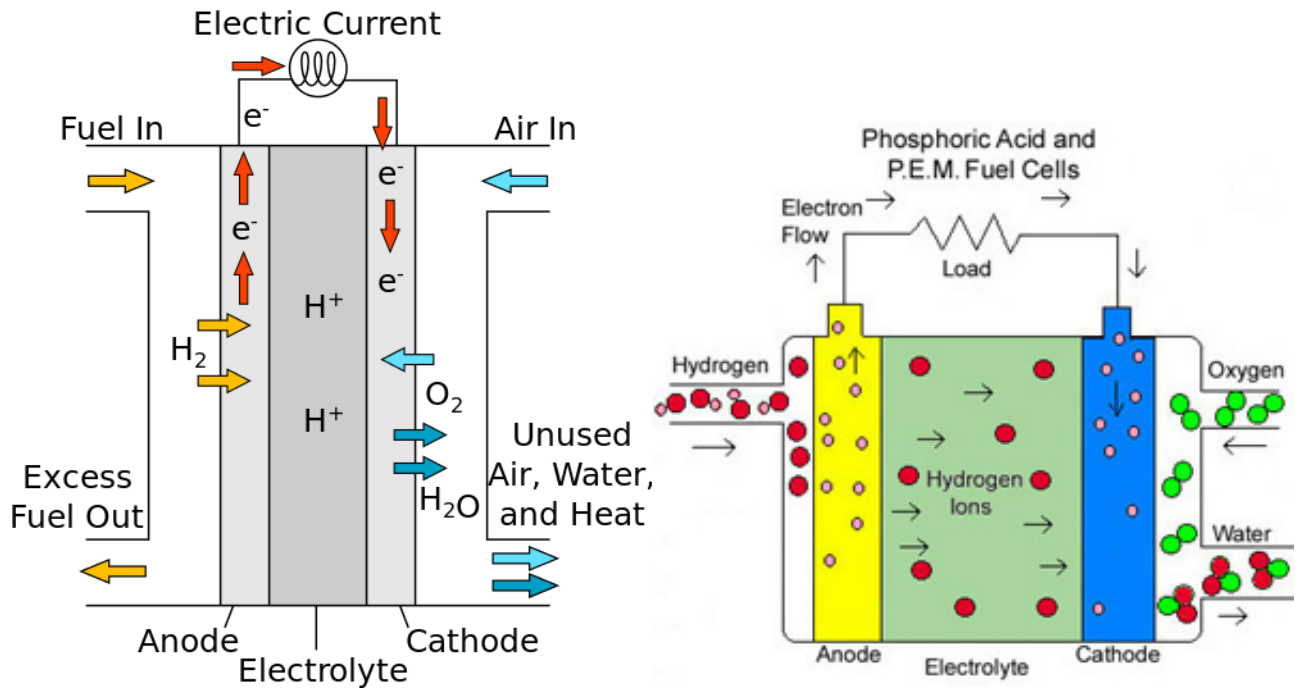
Fig. 10.4. Schemes of a phosphoric-acid fuel cell and its operations.

Electrolyte is highly concentrated or pure liquid phosphoric acid (H_3PO_4) saturated in a silicon carbide matrix (SiC). Operating range is about 150 to 210 °C. The electrodes are made of carbon paper coated with a finely dispersed platinum catalyst.

Electrode reactions:

Anode reaction: $2H_2 \rightarrow 4H^+ + 4e^-$

Cathode reaction: $O_2(g) + 4H^+ + 4e^- \rightarrow 2H_2O$

Overall cell reaction: $2 H_2 + O_2 \rightarrow 2H_2O$

Advantages and disadvantages.

At an operating range of 150 to 200 °C, the expelled water can be converted to steam for air and water heating (combined heat and power). This potentially allows efficiency increases of up to 70% [2]. PAFCs are CO_2-tolerant and even can tolerate a CO concentration of about 1.5 percent, which broadens the choice of fuels they can use. If gasoline is used, the sulfur must be removed [3]. At lower temperatures phosphoric acid is a poor ionic conductor, and CO poisoning of the platinum electro-catalyst in the anode becomes severe [45]. However, they are much less sensitive to CO than PEFCs and AFCs.

Disadvantages include rather low power density and aggressive electrolyte.

10.2.3. High-temperature PEM (HT-PEM)

High-temperature PEM (HT-PEM) fuel cells are similar to PEM fuel cells, but operate at higher temperatures, between 250°F and 390°F. HT-PEMs are often integrated with fuel reformers, permitting operation using wider variety of input fuels. HT-PEMs can be used to power vehicles as range extenders for batteries, and small scale commercial buildings and homes.

10.2.4. Solid Oxide Fuel Cell (SOFC)

A solid oxide fuel cell (SOFC) is an electrochemical conversion device that produces electricity directly from oxidizing a fuel. Fuel cells are characterized by their electrolyte material; the SOFC has a solid oxide or ceramic electrolyte (Fig. 10.5, Fig. 10.6). Advantages of this class of

fuel cells include high efficiency, long-term stability, fuel flexibility, low emissions, and relatively low cost. The largest disadvantage is the high operating temperature which results in longer start-up times and mechanical and chemical compatibility issues [46].
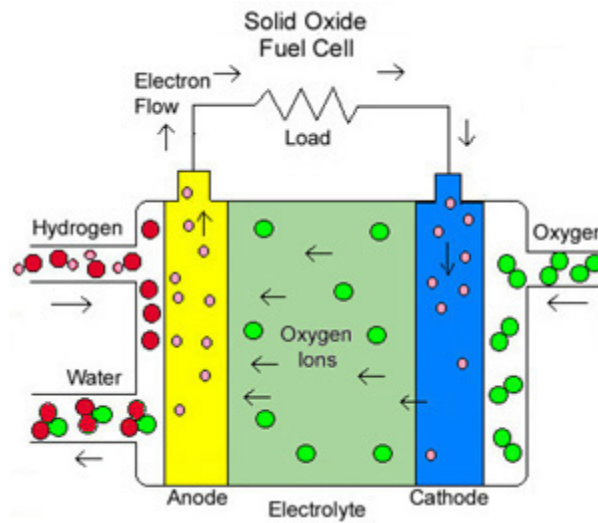
Fig. 10.5. Drawing of a solid oxide cell.
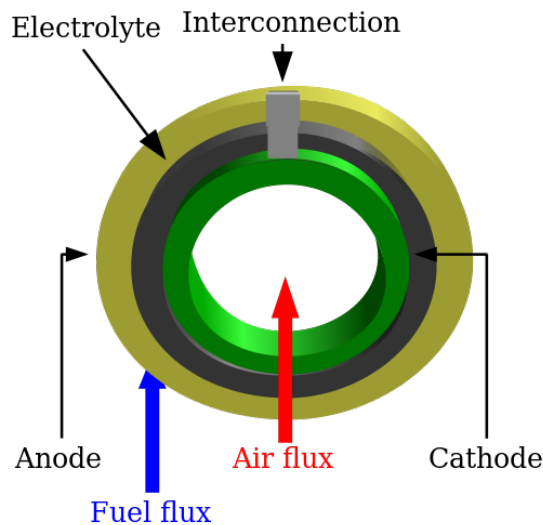
Fig. 10.6. Cross section of three ceramic layers of a tubular SOFC. From left to right: porous cathode, dense electrolyte, porous anode.

Solid oxide fuel cells are a class of fuel cells characterized by the use of a solid oxide material as the electrolyte. SOFCs use a solid oxide electrolyte to conduct negative oxygen ions from the cathode to the anode. The electrochemical oxidation of the oxygen ions with hydrogen or carbon monoxide thus occurs on the anode side. More recently, proton-conducting SOFCs (PC-SOFC) are being developed which transport protons instead of oxygen ions through the electrolyte with the advantage of being able to be run at lower temperatures than traditional SOFCs.

They operate at very high temperatures, typically between 500 and 1,000°C. At these temperatures, SOFCs do not require expensive platinum catalyst material, as is currently necessary for lower-temperature fuel cells such as PEMFCs, and are not vulnerable to carbon monoxide catalyst poisoning. However, vulnerability to sulfur poisoning has been widely observed and the sulfur must be removed before entering the cell through the use of adsorbent beds or other means.

Solid oxide fuel cells have a wide variety of applications from use as auxiliary power units in vehicles to stationary power generation with outputs from 100 W to 2 MW. In 2009, Australian company, Ceramic Fuel Cells successfully achieved an efficiency of a SOFC device up to the previously theoretical mark of 60% [47, 48]. The higher operating temperature make SOFCs suitable candidates for application with heat engine energy recovery devices or combined heat and power, which further increases overall fuel efficiency.

Because of these high temperatures, light hydrocarbon fuels, such as methane, propane and butane can be internally reformed within the anode. SOFCs can also be fueled by externally reforming heavier hydrocarbons, such as gasoline, diesel, jet fuel (JP-8) or biofuels. Such reformates are mixtures of hydrogen, carbon monoxide, carbon dioxide, steam and methane, formed by reacting the hydrocarbon fuels with air or steam in a device upstream of the SOFC anode. SOFC power systems can increase efficiency by using the heat given off by the exothermic electrochemical oxidation within the fuel cell for endothermic steam reforming process. Additionally, solid fuels such as coal and biomass may be gasified to form syngas which is suitable for fueling SOFCs in integrated gasification fuel cell power cycles.

Thermal expansion demands a uniform and well-regulated heating process at startup. SOFC stacks with planar geometry require in the order of an hour to be heated to light-off temperature. Micro-tubular fuel cell design geometries promise much faster start up times, typically in the order of minutes.

Unlike most other types of fuel cells, SOFCs can have multiple geometries. The planar fuel cell design geometry is the typical sandwich type geometry employed by most types of fuel cells, where the electrolyte is sandwiched in between the electrodes. SOFCs can also be made in tubular geometries where either air or fuel is passed through the inside of the tube and the other gas is passed along the outside of the tube. The tubular design is advantageous because it is much easier to seal air from the fuel. The performance of the planar design is currently better than the performance of the tubular design however, because the planar design has a lower resistance comparatively. Other geometries of SOFCs include modified planar fuel cell designs (MPC or MPSOFC), where a wave-like structure replaces the traditional flat configuration of the planar cell. Such designs are highly promising, because they share the advantages of both planar cells (low resistance) and tubular cells.

Operation. A solid oxide fuel cell is made up of four layers, three of which are ceramics (hence the name). A single cell consisting of these four layers stacked together is typically only a few millimeters thick. Hundreds of these cells are then connected in series to form what most people refer to as an "SOFC stack". The ceramics used in SOFCs do not become electrically and ionically active until they reach very high temperature and as a consequence the stacks have to run at temperatures ranging from 500 to 1,000 °C. Reduction of oxygen into oxygen ions occurs at the cathode. These ions can then diffuse through the solid oxide electrolyte to the anode where they can electrochemically oxidize the fuel. In this reaction, a water byproduct is given off as well as two electrons. These electrons then flow through an external circuit where they can do work. The cycle then repeats as those electrons enter the cathode material again.

Balance of plant. Most of the downtime of a SOFC stems from the mechanical balance of plant, the air preheater, prereformer, after burner, water heat exchanger, anode tail gas oxidizer, and electrical balance of plant, power electronics,hydrogen sulfide sensor and fans. Internal reforming leads to a large decrease in the balance of plant costs in designing a full system [48].

*Anode.*The ceramic anode layer must be very porous to allow the fuel to flow towards the electrolyte. Consequently, granular matter is often selected for anode fabrication procedures [49]. Like the cathode, it must conduct electrons, with ionic conductivity a definite asset. The most common material used is a cermet made up ofnickel mixed with the ceramic material that is used for the electrolyte in that particular cell, typically YSZ (yttria stabilized zirconia) nanomaterial-based catalysts, this YSZ part helps stop the grain growth of nickel. The anode is commonly the thickest and strongest layer in each individual cell, because it has the smallest polarization losses, and is often the layer that provides the mechanical support. Electrochemically speaking, the anode's

job is to use the oxygen ions that diffuse through the electrolyte to oxidize the hydrogen fuel. Theoxidation reaction between the oxygen ions and the hydrogen produces heat as well as water and electricity. If the fuel is a light hydrocarbon, for example methane, another function of the anode is to act as a catalyst for steam reforming the fuel into hydrogen. This provides another operational benefit to the fuel cell stack because the reforming reaction is endothermic, which cools the stack internally.

Electrolyte. The electrolyte is a dense layer of ceramic that conducts oxygen ions. Its electronic conductivity must be kept as low as possible to prevent losses from leakage currents. The high operating temperatures of SOFCs allow the kinetics of oxygen ion transport to be sufficient for good performance. However, as the operating temperature approaches the lower limit for SOFCs at around 600 °C, the electrolyte begins to have large ionic transport resistances and affect the performance. Popular electrolyte materials include yttria-stabilized zirconia (YSZ) (often the 8% form Y8SZ), scandia stabilized zirconia (ScSZ) (usually 9 mol% Sc_2O_3 – 9ScSZ) and gadolinium doped ceria (GDC) [50]. The electrolyte material has crucial influence on the cell performances [51]. Detrimental reactions between YSZ electrolytes and modern cathodes such as lanthanum strontium cobalt ferrite (LSCF) have been found, and can be prevented by thin (<100 nm) ceria diffusion barriers [52].

If the conductivity for oxygen ions in SOFC can remain high even at lower temperature (current target in research ~500 °C), material choice for SOFC will broaden and many existing problems can potentially be solved. Certain processing technique such as thin film deposition [53] can help solve this problem with existing material by:

– reducing the traveling distance of oxygen ions and electrolyte resistance as resistance is inversely proportional to conductor length;

– producing grain structures that are less resistive such as columnar grain structure;

– controlling the micro-structural nano-crystalline fine grains to achieve "fine-tuning" of electrical properties;

– building composite with large interfacial areas as interfaces have shown to have extraordinary electrical properties.

Lowering the temperature can lead to a 6 percent increase in the efficiency of the solid oxide fuel cell. The cathode, or air electrode, is a thin porous layer on the electrolyte where oxygen reduction takes place. The overall reaction is written in Kröger-Vink Notation as follows:

$$\frac{1}{2}O_2(g) + 2e' + V_o^{\bullet\bullet} \longrightarrow O_o^{\times}$$
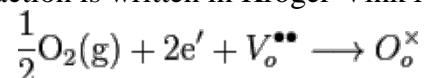
Cathode materials must be, at minimum, electronically conductive. Currently, lanthanum strontium manganite (LSM) is the cathode material of choice for commercial use because of its compatibility with doped zirconia electrolytes. Mechanically, it has similar coefficient of thermal expansion to YSZ and thus limits stresses buildup because of CTE mismatch. Also, LSM has low levels of chemical reactivity with YSZ which extends the lifetime of the material. Unfortunately, LSM is a poor ionic conductor, and so the electrochemically active reaction is limited to the triple phase boundary (TPB) where the electrolyte, air and electrode meet. LSM works well as a cathode at high temperatures, but its performance quickly falls as the operating temperature is lowered below 800 °C. In order to increase the reaction zone beyond the TPB, a potential cathode material must be able to conduct both electrons and oxygen ions. Composite cathodes consisting of LSM YSZ have been used to increase this triple phase boundary length. Mixed ionic/electronic conducting (MIEC) ceramics, such as the perovskite LSCF, are also being researched for use in intermediate temperature SOFCs as they are more active and can makeup for the increase in the activation energy of reaction.

*Interconnect.*The interconnect can be either a metallic or ceramic layer that sits between each individual cell. Its purpose is to connect each cell in series, so that the electricity each cell generates can be combined. Because the interconnect is exposed to both the oxidizing and reducing side of the cell at high temperatures, it must be extremely stable. For this reason, ceramics have been more successful in the long term than metals as interconnect materials. However, these ceramic interconnect materials are very expensive as compared to metals. Nickel- and steel-based

alloys are becoming more promising as lower temperature (600–800 °C) SOFCs are developed. The material of choice for an interconnect in contact with Y8SZ is a metallic 95Cr-5Fe alloy. Ceramic-metal composites called 'cermet' are also under consideration, as they have demonstrated thermal stability at high temperatures and excellent electrical conductivity.

Polarizations. Polarizations, or overpotentials, are losses in voltage due to imperfections in materials, microstructure, and design of the fuel cell. Polarizations result from ohmic resistance of oxygen ions conducting through the electrolyte (iRΩ), electrochemical activation barriers at the anode and cathode, and finally concentration polarizations due to inability of gases to diffuse at high rates through the porous anode and cathode (shown as ηA for the anode and ηC for cathode). The cell voltage can be calculated using the following equation:

$$V = E_0 - i R_\omega - \eta_{cathode} - \eta_{anode}$$

where: E_0 = Nernst potential of the reactants, R = Thévenin equivalent resistance value of the electrically conducting portions of the cell, $\eta_{cathode}$ = polarization losses in the cathode, η_{anode} = polarization losses in the anode

In SOFCs, it is often important to focus on the ohmic and concentration polarizations since high operating temperatures experience little activation polarization. However, as the lower limit of SOFC operating temperature is approached (~600 °C), these polarizations do become important [54].

Above mentioned equation is used for determining the SOFC voltage (in fact for fuel cell voltage in general). This approach results in good agreement with particular experimental data (for which adequate factors were obtained) and poor agreement for other than original experimental working parameters. Moreover, most of the equations used require the addition of numerous factors which are difficult or impossible to determine. It makes very difficult any optimizing process of the SOFC working parameters as well as design architecture configuration selection. Because of those circumstances a few other equations were proposed [55]:

$$E_{SOFC} = \frac{E_{max} - i_{max} \cdot \eta_f \cdot r_1}{\frac{r_1}{r_2} \cdot (1 - \eta_f) + 1}$$

where: E_{SOFC} = cell voltage, E_{max} = maximum voltage given by the Nernst equation, i_{max} = maximum current density (for given fuel flow), η_f = fuel utilization factor [55, 56], r_1 = ionic specific resistance of the electrolyte, r_2 = electric specific resistance of the electrolyte.

This method was validated and found to be suitable for optimization and sensitivity studies in plant-level modelling of various systems with solid oxide fuel cells [57]. With this mathematical description it is possible to account for different properties of the SOFC. There are many parameters which impact cell working conditions, e.g. electrolyte material, electrolyte thickness, cell temperature, inlet and outlet gas compositions at anode and cathode, and electrode porosity, just to name some. The flow in these systems is often calculated using the Navier-stokes equation.

Ohmic polarization. Ohmic losses in an SOFC result from ionic conductivity through the electrolyte. This is inherently a materials property of the crystal structure and atoms involved. However, to maximize the ionic conductivity, several methods can be done. Firstly, operating at higher temperatures can significantly decrease these ohmic losses. Substitutional doping methods to further refine the crystal structure and control defect concentrations can also play a significant role in increasing the conductivity. Another way to decrease ohmic resistance is to decrease the thickness of the electrolyte layer.

Ionic conductivity. An ionic specific resistance of the electrolyte as a function of temperature can be described by the following relationship [55]:

$$r_1 = \frac{\delta}{\sigma}$$

where: δ – electrolyte thickness, and σ – ionic conductivity.

The ionic conductivity of the solid oxide is defined as follows [55]:

$$\sigma = \sigma_0 \cdot e^{\frac{-E}{R \cdot T}}$$

where: σ_0 and E – factors depended on electrolyte materials, T – electrolyte temperature, and R – ideal gas constant.

Concentration polarization. The concentration polarization is the result of practical limitations on mass transport within the cell, and represents the voltage loss due to spatial variations in reactant concentration at the chemically active sites. This situation can be caused when the reactants are consumed by the electrochemical reaction faster than they can diffuse into the porous electrode, and can also be caused by variation in bulk flow composition. The latter is due to the fact that the consumption of reacting species in the reactant flows causes a drop in reactant concentration as it travels along the cell, which causes a drop in the local potential near the tail end of the cell.

The concentration polarization occurs in both the anode and cathode. The anode can be particularly problematic, as the oxidation of the hydrogen produces steam, which further dilutes the fuel stream as it travels along the length of the cell. This polarization can be mitigated by reducing the reactant utilization fraction or increasing the electrode porosity, but these approaches each have significant design trade-offs.

Activation polarization. The activation polarization is the result of the kinetics involved with the electrochemical reactions. Each reaction has a certain activation barrier that must be overcome in order to proceed and this barrier leads to the polarization. The activation barrier is the result of many complex electrochemical reaction steps where typically the rate limiting step is responsible for the polarization. The polarization equation shown below is found by solving the Butler–Volmer equation in the high current density regime (where the cell typically operates), and can be used to estimate the activation polarization:

$$\eta_{act} = \frac{RT}{\beta z F} \times ln\left(\frac{i}{i_0}\right)$$

where: R = gas constant, T_0 = operating temperature, β = electron transfer coefficient, z = electrons associated with the electrochemical reaction, F = Faraday's constant, i = operating current, i_0 = exchange current density.

The polarization can be modified by microstructural optimization. The Triple Phase Boundary (TPB) length, which is the length where porous, ionic and electronically conducting pathways all meet, directly relates to the electrochemically active length in the cell. The larger the length, the more reactions can occur and thus the less the activation polarization. Optimization of TPB length can be done by processing conditions to affect microstructure or by materials selection to use a mixed ionic/electronic conductor to further increase TPB length.

Variaty of SOFC.

SOEC. A solid oxide electrolyser cell (SOEC) is a solid oxide fuel cell set in regenerative mode for the electrolysis of water with a solid oxide, or ceramic, electrolyte to produce oxygen and hydrogen gas [58].

ITSOFC. SOFCs that operate in an intermediate temperature (IT) range, meaning between 600 and 800 °C, are named ITSOFCs. Because of the high degradation rates and materials costs incurred at temperatures in excess of 900 °C, it is economically more favorable to operate SOFCs at lower temperatures. The push for high performance ITSOFCs is currently the topic of much research and development. One area of focus is the cathode material. It is thought that the oxygen reduction reaction is responsible for much of the loss in performance so the catalytic activity of the cathode is being studied and enhanced through various techniques, including catalyst impregnation.

LT-SOFC. Low temperature solid oxide fuel cells (LT-SOFCs), operating lower than 650 °C, are of great interest for future research because the high operating temperature is currently what restricts the development and deployment of SOFCs. A low temperature SOFC is more reliable due to smaller thermal mismatch and easier sealing. Additionally, a lower temperature requires less insulation, and therefore has lower cost. Cost is further lowered due to wider material choices for interconnects and compressive nonglass/ceramic seals. Perhaps most importantly, at a lower temperature, SOFCs can be started more rapidly and with less energy, which lends itself for uses in

in portable and transport applications. Interestingly, as temperature decreases, the maximum theoretical fuel cell efficiency increases, in contrast to the Carnot cycle. For example, the maximum theoretical efficiency of an SOFC using CO as a fuel increases from 63% at 900 °C to 81% at 350 °C [59]. This is a materials issue, particularly for the electrolyte in the SOFC. YSZ is the most commonly used electrolyte because of its superior stability, despite not having the highest conductivity. Currently, the thickness of YSZ electrolytes is a minimum of ~10 μm due to deposition methods, and this requires a temperature above 700 °C. Therefore, low-temperature SOFCs are only possible with higher conductivity electrolytes. Various alternatives that could be successful at low temperature include gallium-doped ceria (GDC) and erbia-cation-stabilized bismuth (ERB) (Fig. 10.7). They have superior ionic conductivity at lower temperature, but this comes at the expense of lower thermodynamic stability. CeO_2 electrolytes become electronically conductive and Bi_2O_3 electrolytes decompose to metallic Bi under the reducing fuel environment [60]. To combat this, researchers created a functionally graded ceria/bismuth-oxide bilayered electrolyte where the GDC layer on the anode side protects the ESB layer from decomposing while the ESB on the cathode side blocks the leakage current through the GDC layer. This leads to near-theoretical open-circuit potential (OPC) with two highly conductive electrolytes, that by themselves would not have been sufficiently stable for the application. This bilayer proved to be stable for 1400 hours of testing at 500 °C and showed no indication of interfacial phase formation or thermal mismatch. While this makes strides towards lowering the operating temperature of SOFCs, it also opens doors for future research to try and understand this mechanism [61].
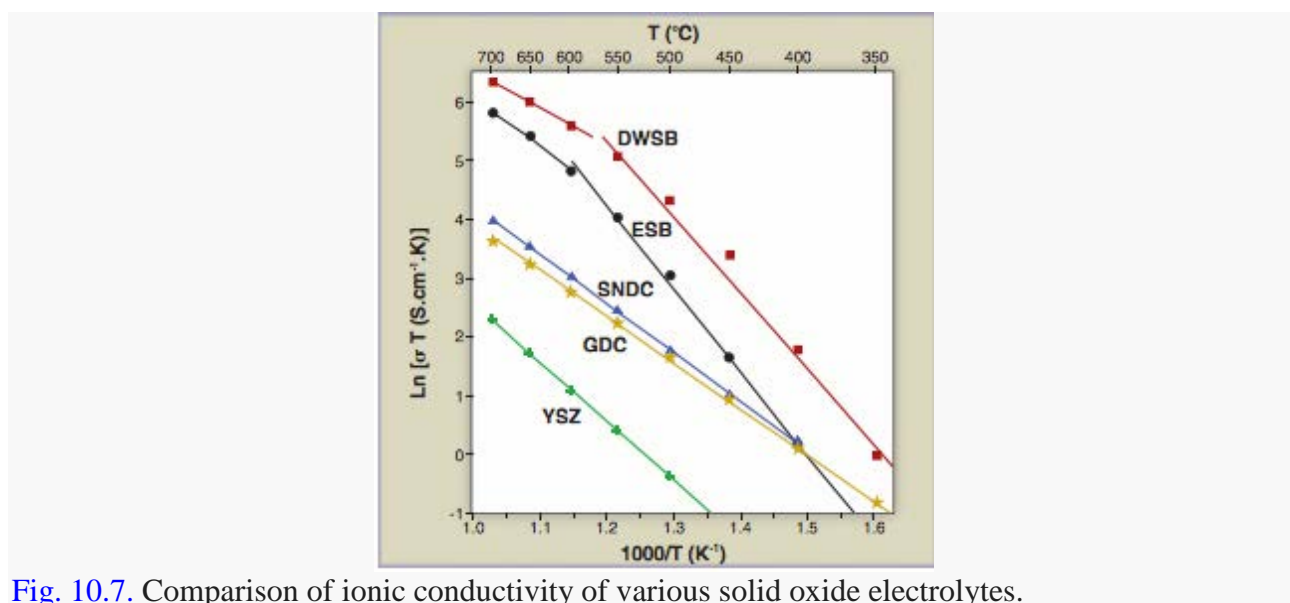
Fig. 10.7. Comparison of ionic conductivity of various solid oxide electrolytes.

Researchers at the Georgia Institute of Technology dealt with the instability of $BaCeO_3$ differently. They replaced a desired fraction of Ce in $BaCeO_3$ with Zr to form a solid solution that exhibits proton conductivity, but also chemical and thermal stability over the range of conditions relevant to fuel cell operation. A new specific composition, $Ba(Zr_{0.1}Ce_{0.7}Y_{0.2})O_3$-δ (BZCY7) that displays the highest ionic conductivity of all known electrolyte materials for SOFC applications. This electrolyte was fabricated by dry-pressing powders, which allowed for the production of crack free films thinner than 15 μm. The implementation of this simple and cost-effective fabrication method may enable significant cost reductions in SOFC fabrication [62]. However, this electrolyte operates at higher temperatures than the bilayered electrolyte model, closer to 600 °C rather than 500 °C.

Currently, given the state of the field for LT-SOFCs, progress in the electrolyte would reap the most benefits, but research into potential anode and cathode materials would also lead to useful results, and has started to be discussed more frequently in literature.

SOFC-GT. An SOFC-GT system is one which comprises a solid oxide fuel cell combined with a gas turbine. Such systems have been evaluated by Siemens Westinghouse and Rolls-Royce as a means to achieve higher operating efficiencies by running the SOFC under pressure. SOFC-GT

systems typically include anodic and/or cathodic atmosphere recirculation, thus increasing efficiency.

Theoretically, the combination of the SOFC and gas turbine can give result in high overall (electrical and thermal) efficiency [63]. Further combination of the SOFC-GT in a combined cooling, heat and power (or trigeneration) configuration (via HVAC) also has the potential to yield even higher thermal efficiencies in some cases [64].

DCFC. For the direct use of solid coal fuel without additional gasification and reforming processes, a direct carbon fuel cell (DCFC) has been developed as a promising novel concept of a high-temperature energy conversion system. The underlying progress in the development of a coal-based DCFC has been categorized mainly according to the electrolyte materials used, such as solid oxide, molten carbonate, and molten hydroxide, as well as hybrid systems consisting of solid oxide and molten carbonate binary electrolyte or of liquid anode (Fe, Ag, In, Sn, Sb, Pb, Bi, and its alloying and its metal/metal oxide) solid oxide electrolyte [65 -69].

SOFC operated on Landfill gas. Every household produces waste/garbage on a daily basis. In 2009, Americans produced about 243 million tons of municipal solid waste, which is 4.3 pounds of waste per person per day. All that waste is sent to Landfill sites. Landfill gas which is produced from the decomposition of waste that gets accumulated at the landfills has the potential to be a valuable source of energy since methane is a major constituent. Currently the majority of the landfills either burn away their gas in flares or combust it in mechanical engines to produce electricity. The issue with mechanical engines is that incomplete combustion of gases can lead to pollution of atmosphere and are also highly inefficient.

The issue with using landfill gas to fuel an SOFC system is that landfill gas contains hydrogen sulfide. Any landfill accepting biological waste will contain about 50-60 ppm of hydrogen sulfide and around 1-2 ppm mercaptans. However, construction materials containing reducible sulfur species, principally sulfates found in gypsum-based wallboard, can cause considerably higher levels of sulfides in the 100's of ppm. At operating temperatures of 750 °C hydrogen sulfide concentrations of around 0.05 ppm begin to affect the performance of the SOFCs.

$$Ni + H_2S \rightarrow NiS + H_2$$
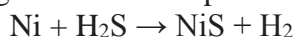
The above reaction controls the effect of sulfur on the anode.

This can be prevented by having background hydrogen which is calculated below.

At 453 K the equilibrium constant is 7.39×10^{-5}.

ΔG calculated at 453 K was 35.833 KJ/mol

Using the standard heat of formation and entropy ΔG at room temperature (298 K) came out to be 45.904 KJ/mol

On extrapolation to 1023 K, ΔG is -1.229 KJ/mol

On substitution, K_{eq} at 1023 K is 1.44×10^{-4}. Hence theoretically we need 3.4% hydrogen to prevent the formation of NiS at 5 ppm H_2S [70].

10.2.5. Hydrogen-oxygen fuel cell

The hydrogen-oxygen fuel cell or alkaline fuel cell (Fig. 10. 8) was designed and first demonstrated publicly by Francis Thomas Bacon in 1959. It was used as a primary source of electrical energy in the Apollo space program [42]. The cell consists of two porous carbon electrodes impregnated with a suitable catalyst such as Pt, Ag, CoO, etc. The space between the two electrodes is filled with a concentrated solution of KOH or NaOH which serves as an electrolyte. $2H_2$ gas and O_2 gas are bubbled into the electrolyte through the porous carbon electrodes. Thus the overall reaction involves the combination of hydrogen gas and oxygen gas to form water. The cell runs continuously until the reactant's supply is exhausted. This type of cell operates efficiently in the temperature range 343 K to 413 K and provides a potential of about 0.9 V [43].
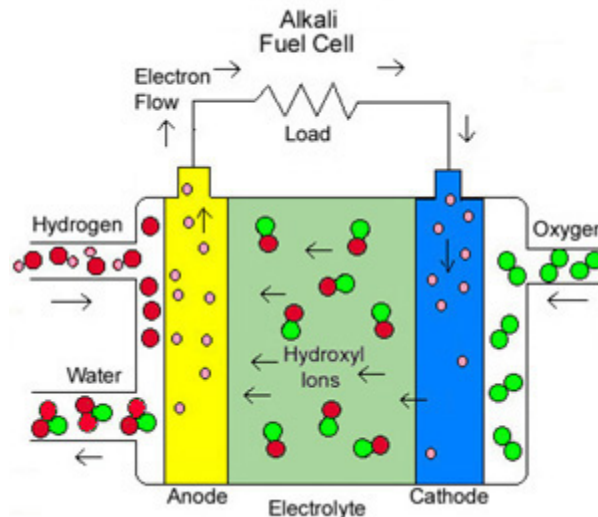
Fig. 10. 8. Drawing of an alkali cell.

10.2.6. Molten Carbonate Fuel Cells (MCFC)

Molten Carbonate fuel cells (MCFC) (Fig. 10. 9) use high-temperature compounds of salt (like sodium or magnesium) carbonates (chemically, CO_3) as the electrolyte. Efficiency ranges from 60 to 80 percent, and operating temperature is about 650 degrees C (1,200 degrees F). Units with output up to 2 megawatts (MW) have been constructed, and designs exist for units up to 100 MW. The high temperature limits damage from carbon monoxide "poisoning" of the cell and waste heat can be recycled to make additional electricity. Their nickel electrode-catalysts are inexpensive compared to the platinum used in other cells. But the high temperature also limits the materials and safe uses of MCFCs–they would probably be too hot for home use. Also, carbonate ions from the electrolyte are used up in the reactions, making it necessary to inject carbon dioxide to compensate. Molten-carbonate fuel cells (MCFCs) are high-temperature fuel cells that operate at temperatures of 600 °C and above.

Fig. 10.9. Schemes of a molten-carbonate fuel cell.

Molten carbonate fuel cells (MCFCs) are currently being developed for natural gas, biogas (produced as a result of anaerobic digestion or biomass gasification), and coal-based power plants for electrical utility, industrial, and military applications. MCFCs are high-temperature fuel cells that use anelectrolyte composed of a molten carbonate salt mixture suspended in a porous,

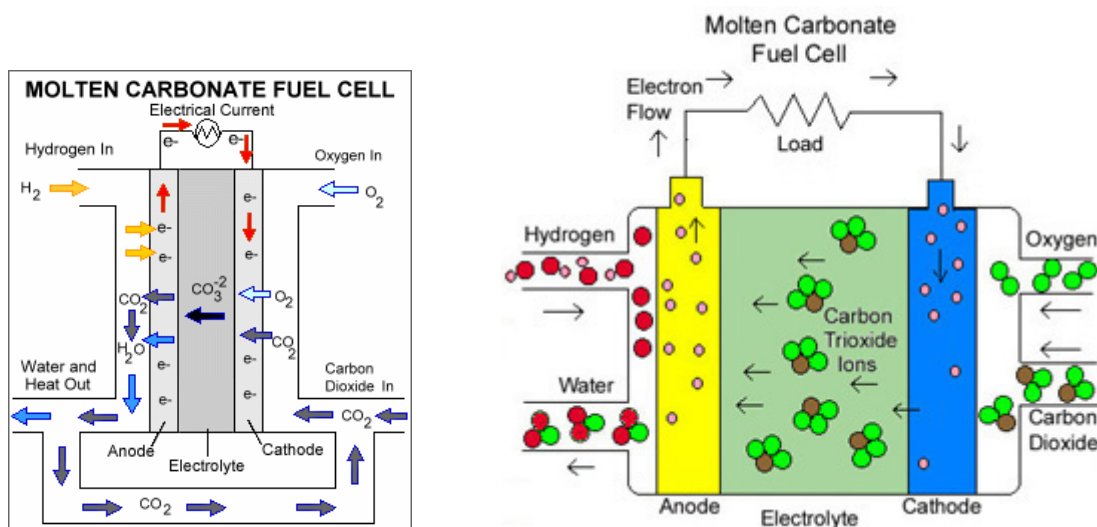chemically inert ceramic matrix of beta-alumina solid electrolyte (BASE). Since they operate at extremely high temperatures of 650 °C (roughly 1,200 °F) and above, non-precious metals can be used as catalysts at the anode and cathode, reducing costs.

Improved efficiency is another reason MCFCs offer significant cost reductions over phosphoric acid fuel cells (PAFCs). Molten carbonate fuel cells can reach efficiencies approaching 60%, considerably higher than the 37–42% efficiencies of a phosphoric acid fuel cell plant. When the waste heat is captured and used, overall fuel efficiencies can be as high as 85%.

Unlike alkaline, phosphoric acid, and polymer electrolyte membrane fuel cells, MCFCs don't require an external reformer to convert more energy-dense fuels to hydrogen. Due to the high temperatures at which MCFCs operate, these fuels are converted to hydrogen within the fuel cell itself by a process called internal reforming, which also reduces cost.
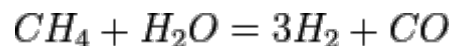
Molten carbonate fuel cells are not prone to poisoning by carbon monoxide or carbon dioxide — they can even use carbon oxides as fuel — making them more attractive for fueling with gases made from coal. Because they are more resistant to impurities than other fuel cell types, scientists believe that they could even be capable of internal reforming of coal, assuming they can be made resistant to impurities such as sulfur and particulates that result from converting coal, a dirtier fossil fuel source than many others, into hydrogen. Alternatively, because MCFCs require CO_2 be delivered to the cathode along with the oxidizer, they can be used to electrochemically separate carbon dioxide from the flue gas of other fossil fuel power plants for sequestration.

The primary disadvantage of current MCFC technology is durability. The high temperatures at which these cells operate and the corrosive electrolyte used accelerate component breakdown and corrosion, decreasing cell life. Scientists are currently exploring corrosion-resistant materials for components as well as fuel cell designs that increase cell life without decreasing performance.

Operation. Molten carbonate FC's are a recently developed type of fuel cell that targets small and large energy distribution/generation systems since their power production is in the 0.3-300 MW range [73]. The operating pressure is between 1-8 atm while the temperatures are between 600-700 ºC [74]. Due to the production of CO_2 during reforming of the fossil fuel (methane, natural gas), MCFC's are not a completely green technology, but are promising due to their reliability and efficiency (sufficient heat for co-generation with electricity). Current MCFC efficiencies range from 60-70% [75].

Reactions [76]
Internal Reformer:
$$CH_4 + H_2O = 3H_2 + CO$$
Anode:
$$H_2 + CO_3^= = H_2O + CO_2 + 2e^-$$
Cathode:
$$\frac{1}{2}O_2 + CO_2 + 2e^- = CO_3^=$$
Cell:
$$H_2 + \frac{1}{2}O_2 + CO_2 = H_2O + CO_2$$
Nernst Equation:
$$E = E^o + \frac{RT}{2F}log\frac{P_{H_2}P_{O_2}^{\frac{1}{2}}}{P_{H_2O}} + \frac{RT}{2F}log\frac{P_{CO_2,cathode}}{P_{CO_2,anode}}$$
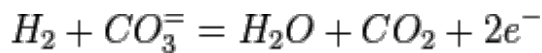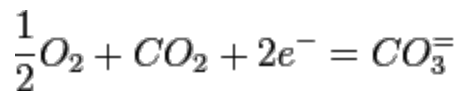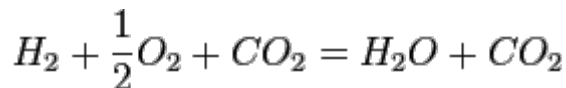
Materials. Due to the high operating temperatures of MCFC's, the materials need to be very carefully selected to survive the conditions present within the cell. The following sections cover the various materials present in the fuel cell and recent developments in research.

Anode. The anode material typically consists of a porous (3-6 μm, 45-70% material porosity) Ni based alloy. Ni is alloyed with either Chromium or Aluminum in the 2-10% range. These alloying elements allow for formation of $LiCrO_2/LiAlO_2$ at the grain boundaries, which increases the materials' creep resistance and prevents sintering of the anode at the high operating temperatures of the fuel cell [77]. Recent research has looked at using nano Ni and other Ni alloys to increase the performance and decrease the operating temperature of the fuel cell [78]. A reduction in operating temperature would extend the lifetime of the fuel cell (i.e. decrease corrosion rate) and allow for use of cheaper component materials. At the same time, a decrease in temperature would decrease ionic conductivity of the electrolyte and thus, the anode materials need to compensate for this performance decline (e.g. by increasing power density). Other researchers have looked into enhancing creep resistance by using a Ni_3Al alloy anode to reduce mass transport of Ni in the anode when in operation [79].

Cathode. On the other side of the cell, the cathode material is composed of a porous Ni that is converted to a lithiated nickel oxide (lithium is intercalated within the NiO crystal structure). The pore size within the cathode is in the range of 7-15 μm with 60-70% of the material being porous [80]. The primary issue with the cathode material is dissolution of NiO since it reacts with CO_2 when the cathode is in contact with the carbonate electrolyte. This dissolution leads to precipitation of Ni metal in the electrolyte and since it is electrically conductive, the fuel cell can get short circuited. Therefore, current studies have looked into the addition of MgO to the NiO cathode to limit this dissolution [81]. Magnesium oxide serves to reduce the solubility of Ni^{2+} in the cathode and decreases precipitation in the electrolyte. Alternatively, replacement of the conventional cathode material with a $LiFeO_2$-$LiCoO_2$-NiO alloy has shown promising performance results and almost completely avoids the problem of Ni dissolution of the cathode [81].

Electrolyte. MCFC's use a liquid electrolyte (molten carbonate) which consists of a sodium (Na) and potassium (K) carbonate. This electrolyte is supported by a ceramic ($LiAlO_2$) matrix to contain the liquid between the electrodes. The high temperatures of the fuel cell is required to produce sufficient ionic conductivity of oxygen through this electrolyte [74]. Common MCFC electrolytes contain 62% Li_2CO_3 and 38% K_2CO_3 [82]. A greater fraction of Li carbonate is used due to its higher ionic conductivity but is limited to 62% due to its lower gas solubility and ionic diffusivity of oxygen. In addition, Li_2CO_3 is a very corrosive electrolyte and this ratio of carbonates provides the lowest corrosion rate. Due to these issues, recent studies have delved into replacing the potassium carbonate with a sodium carbonate [83]. A Li/Na electrolyte has shown to have better performance (higher conductivity) and improves the stability of the cathode when compared to a Li/K electrolyte (Li/K is morebasic). In addition, scientists have also looked into modifying the matrix of the electrolyte to prevent issues such as phase changes (γ-$LiAlO_2$ to α-$LiAlO_2$) in the material during cell operation. The phase change accompanies a volume decrease in the electrolyte which leads to lower ionic conductivity. Through various studies, it has been found that an alumina doped α-$LiAlO_2$ matrix would improve the phase stability while maintaining the fuel cell's performance [83].

10.3. COMPARISON OF FUEL CELL TYPES

Fuel cell name	Electrolyte	Qualified power (W)	Working temperature (°C)	Efficiency(cell)	Efficiency (system)	Status	Cost (USD/W)

Fuel cell name	Electrolyte	Qualified power (W)	Working temperature (°C)	Efficiency(cell)	Efficiency (system)	Status	Cost (USD/W)
Zinc-air battery	Aqueous alkaline solution		< 40			Mass production	
Upflow microbial fuel cell (UMFC)			< 40			Research	
Tubular solid oxide fuel cell(TSOFC)	O^{2-}-conducting ceramic oxide	< 100 MW	850–1100	60–65%	55–60%	Commercial / Research	
Solid acid fuel cell	H^+-conducting oxyanion salt (solid acid)	10 W - 1 kW	200-300	55-60%	40-45%	Commercial / Research	
RFC – Redox	Liquid electrolytes with redox shuttle and polymer membrane (Ionomer)	1 kW – 10 MW				Research	
Regenerative fuel cell	Polymer membrane (ionomer)		< 50			Commercial / Research	
Reformed methanol fuel cell	Polymer membrane (ionomer)	5 W – 100 kW	250–300 (Reformer) 125–200 (PBI)	50–60%	25–40%	Commercial / Research	
Protonic ceramic fuel cell	H^+-conducting ceramic oxide		700			Research	

Fuel cell name	Electrolyte	Qualified power (W)	Working temperature (°C)	Efficiency(cell)	Efficiency (system)	Status	Cost (USD/W)
Proton exchange membrane fuel cell	Polymer membrane (ionomer)	1 W – 500 kW	50–100 (Nafion)[48] 120–200 (PBI) [49]	50–70%	30–50%[47]	Commercial / Research	50–100
Planar Solid oxide fuel cell	O^{2-}-conducting ceramic oxide	< 100 MW	500–1100	60–65%	55–60%[47]	Commercial / Research	
Phosphoric acid fuel cell	Molten phosphoric acid (H_3PO_4)	< 10 MW	150-200	55%	40%[47] Co-Gen: 90%	Commercial / Research	4–4.50
Molten carbonate fuel cell	Molten alkaline carbonate	100 MW	600–650	55%	45-55%[47]	Commercial / Research	
Microbial fuel cell	Polymer membrane or humic acid		< 40			Research	
Metal hydride fuel cell	Aqueous alkaline solution		> -20 (50% P_{peak} @ 0 °C)			Commercial / Research	
Magnesium-Air Fuel Cell	Salt water		−20 to 55	90%		Commercial / Research	
Enzymatic Biofuel Cells	Any that will not denature the enzyme		< 40			Research	

Fuel cell name	Electrolyte	Qualified power (W)	Working temperature (°C)	Efficiency(cell)	Efficiency (system)	Status	Cost (USD/W)
Electro-galvanic fuel cell	Aqueous alkaline solution		< 40			Commercial / Research	
Direct-ethanol fuel cell	Polymer membrane (ionomer)	< 140 mW/cm²	> 25 ? 90–120			Research	
Direct methanol fuel cell	Polymer membrane (ionomer)	100 mW – 1 kW	90–120	20–30%	10–25%[47]	Commercial / Research	125
Direct formic acid fuel cell(DFAFC)	Polymer membrane (ionomer)	< 50 W	< 40			Commercial / Research	
Direct carbon fuel cell	Several different		700–850	80%	70%	Commercial / Research	
Direct borohydride fuel cell	Aqueous alkaline solution		70			Commercial	
Alkaline fuel cell	Aqueous alkaline solution	10 – 100 kW	< 80	60–70%	62%	Commercial / Research	

10.4. EFFICIENCY OF FUEL CELL TYPES

Theoretical maximum efficiency. The energy efficiency of a system or device that converts energy is measured by the ratio of the amount of useful energy put out by the system ("output

energy") to the total amount of energy that is put in ("input energy") or by useful output energy as a percentage of the total input energy. In the case of fuel cells, useful output energy is measured in electrical energy produced by the system. Input energy is the energy stored in the fuel. According to the U.S. Department of Energy, fuel cells are generally between 40–60% energy efficient [84]. This is higher than some other systems for energy generation. For example, the typical internal combustion engine of a car is about 25% energy efficient [85]. In combined heat and power (CHP) systems, the heat produced by the fuel cell is captured and put to use, increasing the efficiency of the system to up to 85–90%.

The theoretical maximum efficiency of any type of power generation system is never reached in practice, and it does not consider other steps in power generation, such as production, transportation and storage of fuel and conversion of the electricity into mechanical power. However, this calculation allows the comparison of different types of power generation. The maximum theoretical energy efficiency of a fuel cell is 83%, operating at low power density and using pure hydrogen and oxygen as reactants (assuming no heat recapture) [86]. According to the World Energy Council, this compares with a maximum theoretical efficiency of 58% for internal combustion engines [86]. While these efficiencies are not approached in most real world applications, high-temperature fuel cells (solid oxide fuel cells or molten carbonate fuel cells) can theoretically be combined with gas turbines to allow stationary fuel cells to come closer to the theoretical limit. A gas turbine would capture heat from the fuel cell and turn it into mechanical energy to increase the fuel cell's operational efficiency. This solution has been predicted to increase total efficiency to as much as 80% [87].

In practice. In a fuel-cell vehicle the tank-to-wheel efficiency is greater than 45% at low loads [88] and shows average values of about 36% when a driving cycle like the NEDC (New European Driving Cycle) is used as test procedure [89]. The comparable NEDC value for a Diesel vehicle is 22%. In 2008 Honda released a demonstration fuel cell electric vehicle (the Honda FCX Clarity) with fuel stack claiming a 60% tank-to-wheel efficiency [90].

It is also important to take losses due to fuel production, transportation, and storage into account. Fuel cell vehicles running on compressed hydrogen may have a power-plant-to-wheel efficiency of 22% if the hydrogen is stored as high-pressure gas, and 17% if it is stored as liquid hydrogen [91]. Fuel cells cannot store energy like a battery, [92] except as hydrogen, but in some applications, such as stand-alone power plants based on discontinuous sources such as solar or wind power, they are combined with electrolyzers and storage systems to form an energy storage system. Most hydrogen, however, is produced by steam methane reforming, and so most hydrogen production emits carbon dioxide [93]. The overall efficiency (electricity to hydrogen and back to electricity) of such plants (known as *round-trip efficiency*), using pure hydrogen and pure oxygen can be "from 35 up to 50 percent", depending on gas density and other conditions [94]. While a much cheaperlead–acid battery might return about 90%, the electrolyzer/fuel cell system can store indefinite quantities of hydrogen, and is therefore better suited for long-term storage.

Solid-oxide fuel cells produce exothermic heat from the recombination of the oxygen and hydrogen. The ceramic can run as hot as 800 degrees Celsius. This heat can be captured and used to heat water in a micro combined heat and power (m-CHP) application. When the heat is captured, total efficiency can reach 80–90% at the unit, but does not consider production and distribution losses. CHP units are being developed today for the European home market.

10.5. CONCLUSION

Fuel cell technology offers clean, efficient, reliable power generation to almost any device requiring electrical power. Fuel cells are used in a wide range of portable, stationary and transport applications, from battery chargers to home heating and power to cars. Arguably, fuel cells represent the most versatile energy solution ever invented.

A fuel cell is like a battery in that it generates electricity from an electrochemical reaction. Fuel cells however use an external supply of chemical energy meaning they can run indefinitely, as long as they are supplied with a source of hydrogen and a source of oxygen (usually air).

Fuel cells are used in a large number of different applications: portable power generation, stationary power generation, power for transportation and many others.

References to Chapter 10.

1. Khurmi, R. S. Material Science.

2. Nice, Karim and Strickland, Jonathan. "How Fuel Cells Work: Polymer Exchange Membrane Fuel Cells". How Stuff Works, accessed 4 August 2011

3. Prabhu, Rahul R. (13 January 2013). "Stationary Fuel Cells Market size to reach 350,000 Shipments by 2022". Renew India Campaign. Retrieved 2013-01-14.

4 Larminie, James (1 May 2003). Fuel Cell Systems Explained, Second Edition. SAE International.ISBN 0-7680-1259-7.

5. Wang, J.Y. (2008). "Pressure drop and flow distribution in parallel-channel of configurations of fuel cell stacks: U-type arrangement". Int. J. of Hydrogen Energy **33** (21): 6339–6350.doi:10.1016/j.ijhydene.2008.08.020.

6. Wang, J.Y.; Wang, H.L. (2012). "Flow field designs of bipolar plates in PEM fuel cells: theory and applications, Fuel Cells," **12** (6): 989–1003. doi:10.1002/fuce.201200074.

7. Wang, J.Y.; Wang, H.L. (2012). "Discrete approach for flow-field designs of parallel channel configurations in fuel cells". Int. J. of Hydrogen Energy **37** (14): 10881–10897.doi:10.1016/j.ijhydene.2012.04.034.

8. Loyselle, Patricia; Prokopius, Kevin. "Teledyne Energy Systems, Inc., Proton Exchange Member (PEM) Fuel Cell Engineering Model Powerplant. Test Report: Initial Benchmark Tests in the Original Orientation". NASA. Glenn Research Center. Retrieved 15 September 2011.

9. Millington, Ben, Shangfeng Du, and Bruno G. Pollet. "The Effect of Materials on Proton Exchange Membrane Fuel Cell Electrode Performance." Journal of Power Sources 196.21 (2011): 9013-017. Web.

10. Schalenbach, Maximilian; Hoefner, Tobias; Paciok, Paul; Carmo, Marcelo; Lueke, Wiebke; Stolten, Detlef (2015-10-28). "Gas Permeation through Nafion. Part 1: Measurements". The Journal of Physical Chemistry C. doi:10.1021/acs.jpcc.5b04155.

11. Schalenbach, Maximilian; Hoeh, Michael A.; Gostick, Jeff T.; Lueke, Wiebke; Stolten, Detlef (2015-10-14). "Gas Permeation through Nafion. Part 2: Resistor Network Model". The Journal of Physical Chemistry C. doi:10.1021/acs.jpcc.5b04157.

12. http://www.isvav.cz/projectDetail.do;jsessionid=CE66F4A7484F5D40CBD6FDEC51358F2C?rowId=GAP204%2F10%2F1169

13. http://onlinelibrary.wiley.com/doi/10.1002/fuce.200900036/abstract

14. http://aktualne.centrum.cz/ekonomika/doprava/clanek.phtml?id=801592

15. Lee, J. S.; et al. (2006). "Polymer electrolyte membranes for fuel cells" (PDF). Journal of Industrial and Engineering Chemistry **12**: 175–183.

16. [Wainright, J. S. "Acid-Doped Polybenzimidazoles: A New Polymer Electrolyte." Journal of The Electrochemical Society 142.7 (1995): L121. Web.].

17. O'Hayre, Ryan P. Fuel Cell Fundamentals. Hoboken, NJ: John Wiley & Sons, 2006. Print.

18. Jiangshui Luo, Jin Hu, Wolfgang Saak, Rüdiger Beckhaus, Gunther Wittstock, Ivo F. J. Vankelecom, Carsten Agert and Olaf Conrad (2011). "Protic ionic liquid and ionic melts prepared from methanesulfonic acid and 1H-1,2,4-triazole as high temperature PEMFC electrolytes". Journal of Materials Chemistry **21**: 10426–10436. doi:10.1039/C0JM04306K.

19. Jiangshui Luo, Annemette H. Jensen, Neil R. Brooks, Jeroen Sniekers, Martin Knipper, David Aili, Qingfeng Li, Bram Vanroy, Michael Wübbenhorst, Feng Yan, Luc Van Meervelt, Zhigang Shao, Jianhua Fang, Zheng-Hong Luo, Dirk E. De Vos, Koen Binnemans and Jan Fransaer (2015). "1,2,4-Triazolium perfluorobutanesulfonate as an archetypal pure protic organic ionic plastic crystal electrolyte for all-solid-state fuel cells". Energy & Environmental Science **8**.doi:10.1039/C4EE02280G.

20. Litster, S.; McLean, G. (2004-05-03). "PEM fuel cell electrodes". Journal of Power Sources**130** (1–2): 61–76. doi:10.1016/j.jpowsour.2003.12.055.

21. Gasteiger, H. A.; Panels, J. E.; Yan, S. G. (2004-03-10). "Dependence of PEM fuel cell performance on catalyst loading". Journal of Power Sources. Eighth Ulmer Electrochemische Tage **127** (1–2): 162–171. doi:10.1016/j.jpowsour.2003.09.013.

22. Schalenbach, Maximilian; Zillgitt, Marcel; Maier, Wiebke; Stolten, Detlef (2015-07-29). "Parasitic Currents Caused by Different Ionic and Electronic Conductivities in Fuel Cell Anodes". ACS Applied Materials & Interfaces **7** (29): 15746–15751. doi:10.1021/acsami.5b02182. ISSN 1944-8244.

23. [Litster, S., and G. Mclean. "PEM Fuel Cell Electrodes." Journal of Power Sources 130.1-2 (2004): 61-76. Web.].

24. Link text, Espinoza, M., Andersson, M., Yuan, J., and Sundén, B. (2015) Compress effects on porosity, gas-phase tortuosity, and gas permeability in a simulated PEM gas diffusion layer. Int. J. Energy Res., 39: 1528–1536. doi: 10.1002/er.3348

25. Ramaswamy, Padmini; Wong, Norman E.; Shimizu, George K. H. "MOFs as proton conductors – challenges and opportunities". Chem. Soc. Rev. **43** (16): 5913–5932.doi:10.1039/c4cs00093e.

26. Li, Shun-Li; Xu, Qiang. "Metal–organic frameworks as platforms for clean energy". Energy & Environmental Science **6** (6). doi:10.1039/c3ee40507a.

27. Kitagawa, Hiroshi. "Metal–organic frameworks: Transported into fuel cells". Nature Chemistry **1**(9): 689–690. doi:10.1038/nchem.454.

28. Lux, Lacey; Williams, Kia; Ma, Shengqian. "Heat-treatment of metal–organic frameworks for green energy applications". CrystEngComm **17** (1): 10–22. doi:10.1039/c4ce01499e.

29. Hydrogen, Fuel Cells & Infrastructure Technologies Program Multi-Year Research, Development and Demonstration Plan, U.S. Department of Energy, October 2007.

30. N. Tian, Z.-Y. Zhou, S.-G. Sun, Y. Ding, Z. L. Wang (2007). "Synthesis of tetrahexahedral platinum nanocrystals with high-index facets and high electro-oxidation activity". Science **316**(5825): 732–735. doi:10.1126/science.1140484. PMID 17478717.

31. V. R. Stamenkovic, B. Fowler, B. S. Mun, G. Wang, P. N. Ross, C. A. Lucas, N. M. Marković. Activity on Pt3Ni(111) via increased surface site availability (2007). "Improved Oxygen Reduction Activity on Pt3Ni(111) via Increased Surface Site Availability". Science **315** (5811): 493–497.doi:10.1126/science.1135941. PMID 17218494.

32. Koraishy, Babar (2009). "Manufacturing of membrane electrode assemblies for fuel cells"(PDF). 6.2.1: Singapore University of Technology and Design. p. 9.

33. Engle, Robb (2011-08-08). "Maximizing the use of platinum catalyst by ultrasonic spray application" (pdf). Proceedings of asme 2011 5th international conference on energy sustainability & 9th fuel cell science, engineering and technology conference. Esfuelcell 2011-54369.

34. Shukla, S (2015). "Analysis of Low Platinum Loading Thin Polymer Electrolyte Fuel Cell Electrodes Prepared by Inkjet Printing".

35. Sagar Prabhudev, Matthieu Bugnet, Christina Bock and Gianluigi Botton (2013). "Strained Lattice with Persistent Atomic Order in Pt3Fe2 Intermetallic Core–Shell Nanocatalysts". ACS Nano **7** (7): 6103–6110. doi:10.1021/nn4019009. PMID 23773037.

36. Minna Cao, Dongshuang Wu and Rong Cao (2014). "Recent Advances in the Stabilization of Platinum Electrocatalysts for Fuel-Cell Reactions". ChemCatChem **1** (1): 26–45.doi:10.1002/cctc.201300647.

37. G. Hoogers (2003). Fuel Cell Technology Handbook. Boca Raton, FL: CRC Press. pp. 6–3.ISBN 0-8493-0877-1.

38. C. Wang, H. Daimon, T. Onodera, T. Koda, S. Sun. A general approach to the size- and shape-controlled synthesis of platinum nanoparticles and their catalytic reduction of oxygen (2008). "A General Approach to the Size- and Shape-Controlled Synthesis of Platinum Nanoparticles and Their Catalytic Reduction of Oxygen". Angewandte Chemie International Edition **47** (19): 3588–3591. doi:10.1002/anie.200800073. PMID 18399516.

39. S. Alayoglu, A. U. Nilekar, M. Mavrikakis, B. Eichhorn. Ru–Pt core–shell nanoparticles for preferential oxidation of carbon monoxide in hydrogen (2008). "Ru–Pt core–shell nanoparticles

for preferential oxidation of carbon monoxide in hydrogen". Nature Materials **7** (4): 333–338.doi:10.1038/nmat2156. PMID 18345004.

40. E. Proietti, F. Jaouen, M. Lefevre, N. Larouche, J. Tian, J. Herranz, and J.-P. Dodelet. Iron-based cathode catalyst with enhanced power density in polymer electrolyte membrane fuel cells. Nature Communications, 2(1), 2011

41. S. Litster and G. McLean. PEM fuel cell electrodes. Journal of Power Sources, 130(1-2):61–76, 2004.

42. Yun Wang, Ken S Chen, Mishler Jeffrey, Sung Chan Cho, and Cordobes Adroher Xavier. A review of polymer electrolyte membrane fuel cells: Technology, applications, and needs on fundamental research. Applied Energy, 88(4):981 – 1007, 2011.

43. "Collecting the History of Phosphoric Acid Fuel Cells".

44. "Phosphoric Acid Fuel Cells". scopeWe - a Virtual Engineer.

45. "Mr. W. R. Grove on a new Voltaic Combination". The London and Edinburgh Philosophical Magazine and Journal of Science. 1838. Retrieved October 2, 2013.

46. Badwal, SPS. "Review of Progress in High Temperature Solid Oxide Fuel Cells". Journal of the Australian Ceramics society 50 (1).

47. Ceramic fuel cells achieves world-best 60% efficiency for its electricity generator units. Ceramic Fuel Cells Limited. 19 February 2009

48 Electricity from wood through the combination of gasification and solid oxide fuel cells, Ph.D. Thesis by Florian Nagel, Swiss Federal Institute of Technology Zurich, 2008

49. Ott, J; Gan, Y; McMeeking, R; Kamlah, M (2013). "A micromechanical model for effective conductivity in granular electrode structures" (PDF). Acta Mechanica Sinica. 29 (5): 682–698.

50. Nigel Sammes, Alevtina Smirnova, Oleksandr Vasylyev (2005). "Fuel Cell Technologies: State and Perspectives". NATO Science Series, Mathematics, Physics and Chemistry 202: 19–34.doi:10.1007/1-4020-3498-9_3.

51. Steele, B.C.H., Heinzel, A. (2001). "Materials for fuel-cell technologies". Nature 414 (Nov 15): 345–352. doi:10.1038/35104620. PMID 11713541.

52. Mohan Menon, Kent Kammer; et al. (2007). "Processing of Ce1-xGdxO2-δ (GDC) thin films from precursors for application in solid oxide fuel cells". Advanced Materials Engineering. 15–17: 293–298. doi:10.4028/www.scientific.net/AMR.15-17.293.

53. Charpentier, P (2000). "Preparation of thin film SOFCs working at reduced temperature". Solid State Ionics 135 (1-4): 373–380. doi:10.1016/S0167-2738(00)00472-0. ISSN 0167-2738.

54. Hai-Bo Huo, Xin-Jian Zhu, Guang-Yi Cao (2006). "Nonlinear modeling of a SOFC stack based on a least squares support vector machine". Journal of Power Sources 162 (2): 1220–1225.doi:10.1016/j.jpowsour.2006.07.031.

55 Milewski J, Miller A. (2006). "Influences of The Type and Thickness of Electrolyte on Solid Oxide Fuel Cell Hybrid System Performance". Journal of Fuel Cell Science and Technology 3 (4): 396–402. doi:10.1115/1.2349519.

56. M. Santarelli, P. Leone, M. Calì, G. Orsello (2007). "Experimental evaluation of the sensitivity to fuel utilization and air management on a 100 kW SOFC system". Journal of Power Sources 171(2): 155–168. doi:10.1016/j.jpowsour.2006.12.032.

57. Kupecki J., Milewski J., Jewulski J. (2013). "Investigation of SOFC material properties for plant-level modeling". Central European Journal of Chemistry 11 (5): 664–671. doi:10.2478/s11532-013-0211-x.

58. Anne Hauch, Søren Højgaard Jensen, Sune Dalgaard Ebbesen, Mogens Mogensen (2009)."Durability of solid oxide electrolysis cells for hydrogen production" (PDF). Risoe Reports 1608: 327–338.

59. Choi, S.; Yoo, S.; Park, S.; Jun, A.; Sengodan, S.; Kim, J.; Shin, J. Highly efficient and robust cathode materials for low-temperature solid oxide fuel cells: $PrBa0.5Sr0.5Co(2-x)Fe(x)O(5+\delta)$. Sci Rep. 2013, 3, 2426-2428.

60. Hibini, T.; Hashimoto, A.; Inoue, T.; Tokuno, J.; Yoshida, S.; Sano, M. A Low-Operating-Temperature Solid Oxide Fuel Cell in Hydrocarbon-Air Mixtuers. Science. 2000. 288, 2031-2033.

61. Wachsman, E.; Lee, Kang T. Lowering the Temperature of Solid Oxide Fuel Cells. Science. 2011, 334, 935-939.

62. Zuo, C.; Zha, S.; Liu, M.; Hatano, M.; Uchiyama, M. Ba(Zr0.1Ce0.7Y0.2)O3-δ as an Electrolyte for Low-Temperature Solid-Oxide Fuel Cells. Advanced Materials. 2006, 18, 3318-3320

63. S.H. Chan, H.K. Ho, Y. Tian (2003). "Multi-level modeling of SOFC-gas turbine hybrid system".International Journal of Hydrogen Energy 28 (8): 889–900. doi:10.1016/S0360-3199(02)00160-X.

64. L. K. C. Tse, S. Wilkins, N. McGlashan, B. Urban, R. Martinez-Botas (2011). "Solid oxide fuel cell/gas turbine trigeneration system for marine applications". Journal of Power Sources 196 (6): 3149–3162. doi:10.1016/j.jpowsour.2010.11.099.

65. Giddey, S; Badwal SPS; Kulkarni A; Munnings C (2012). "A comprehensive review of direct carbon fuel cell technology". Progress in energy and combustion science 38 (3): 360–399.doi:10.1016/j.pecs.2012.01.003.

66. HyungKuk Ju, Jiyoung Eom, Jae Kwang Lee, Hokyung Choi, Tak-Hyoung Lim, Rak-Hyun Song, and Jaeyoung Lee, Durable power performance of a direct ash-free coal fuel cell, Electrochimica Acta 115 (2014) 511. doi:10.1016/j.electacta.2013.10.124

67. HyungKuk Ju, Sunghyun Uhm, Jin Won Kim, Rak-Hyun Song, Hokyung Choi, Si-Hyun Lee, Jaeyoung Lee, Enhanced anode interface for electrochemical oxidation of solid fuel in direct carbon fuel cells: The role of liquid Sn in mixed state, Journal of Power Sources 198 (2012) 36.doi:10.1016/j.jpowsour.2011.09.082

68. HyungKuk Ju and Jaeyoung Lee, High-temperature liquid Sn-air energy storage cell, Journal of Energy Chemistry, 24 (2015) 614. doi:10.1016/j.jechem.2015.08.006

69. Hansaem Janga, et al, Ameliorated performance in a direct carbon fuel cell using Snmediator on Ni–YSZ anode surface, Journal of Power Sources, 260 (2016) 158.doi:10.1016/j.cattod.2015.06.013

70. Khan, Feroze (2012-01-01). "Effect of Hydrogen Sulfide in Landfill Gas on Anode Poisoning of Solid Oxide Fuel Cells". Youngstown State University.

71. Williams, K.R. (1 February 1994). "Francis Thomas Bacon. 21 December 1904-24 May 1992"(PDF). Biographical Memoirs of Fellows of the Royal Society 39: 2–9.doi:10.1098/rsbm.1994.0001. Retrieved January 5, 2015.

72. Srivastava, H. C. Nootan ISC Chemistry (12th) Edition 18, pp. 458–459, Nageen Prakashan (2014) ISBN 9789382319399

73. "Types of Fuel Cells - Fuel Cell Energy". www.fuelcellenergy.com. Retrieved 2015-11-02.

74. "NFCRC Tutorial: Molten Carbonate Fuel Cell (MCFC)". www.nfcrc.uci.edu. Retrieved 2015-11-02.

75. "Types of Fuel Cells | Department of Energy". energy.gov. Retrieved 2015-11-02.

76. "High Temperature Fuel Cells" (PDF). University of Babylon. Retrieved 1 November 2015.

77. Boden, Andreas (2007). "The anode and the electrolyte in the MCFC" (PDF). Diva Portal. Retrieved 1 November 2015.

78. Nguyen, Hoang Viet Phuc; Othman, Mohd Roslee; Seo, Dongho; Yoon, Sung Pil; Ham, Hyung Chul; Nam, Suk Woo; Han, Jonghee; Kim, Jinsoo (2014-08-04). "Nano Ni layered anode for enhanced MCFC performance at reduced operating temperature". International Journal of Hydrogen Energy 39 (23): 12285–12290. doi:10.1016/j.ijhydene.2014.03.253.

79. Kim, Yun-Sung; Lim, Jun-Heok; Chun, Hai-Soo (2006-01-01). "Creep mechanism of porous MCFC Ni anodes strengthened by Ni3Al". AIChE Journal 52 (1): 359–365. doi:10.1002/aic.10630. ISSN 1547-5905.

80. Wijayasinghe, Athula (2004). "Development and Characterisation of Cathode Materials for the Molten Carbonate Fuel Cell" (PDF). Retrieved 2 November 2015.

81. Antolini, Ermete. "The stability of molten carbonate fuel cell electrodes: A review of recent improvements". Applied Energy 88 (12): 4274–4293. doi:10.1016/j.apenergy.2011.07.009.

82. *Fang, Baizeng; Liu, Xinyu; Wang, Xindong; Duan, Shuzhen (1998-01-15).* "The mechanism of surface modification of a MCFC anode". *Journal of Electroanalytical Chemistry 441 (1–2): 65–68.*doi*:*10.1016/S0022-0728(97)00202-7.

83. *Kulkarni, A.; Giddey, S. (2012-06-08).* "Materials issues and recent developments in molten carbonate fuel cells".*Journal of Solid State Electrochemistry 16 (10): 3123–3146.* doi*:*10.1007/s10008-012-1771-y. ISSN 1432-8488.

84. "Comparison of Fuel Cell Technologies". U.S. Department of Energy, Energy Efficiency and Fuel Cell Technologies Program, February 2011, accessed 4 August 2011

85. "Fuel Economy: Where The Energy Goes". U.S. Department of Energy, Energy Effciency and Renewable Energy, accessed 3 August 2011

86. "Fuel Cell Efficiency". World Energy Council, 17 July 2007, accessed 4 August 2011Archived February 9, 2014, at the Wayback Machine.

87. Milewski, J., A. Miller and K. Badyda. "The Control Strategy for High Temperature Fuel Cell Hybrid Systems". The Online Journal on Electronics and Electrical Engineering, Vol. 2, No. 4, p. 331, 2009, accessed 4 August 2011

88. Eberle, Ulrich and Rittmar von Helmolt. "Sustainable transportation based on electric vehicle concepts: a brief overview". Energy & Environmental Science, Royal Society of Chemistry, 14 May 2010, accessed 2 August 2011

89. *Von Helmolt, R.; Eberle, U (20 March 2007). "Fuel Cell Vehicles:Status 2007". Journal of Power Sources 165 (2): 833.* doi*:*10.1016/j.jpowsour.2006.12.073.

90. "Honda FCX Clarity – Fuel cell comparison". *Honda.* Retrieved 2009-01-02.

91. "Efficiency of Hydrogen PEFC, Diesel-SOFC-Hybrid and Battery Electric Vehicles" *(PDF). 15 July 2003. Archived from* the original *(PDF) on 21 October 2006.* Retrieved 2007-05-23.

92. "Batteries, Supercapacitors, and Fuel Cells: Scope". *Science Reference Services. 20 August 2007.* Retrieved 11 February 2009.

93. Nice, Karim. "How Fuel Processors Work". HowStuffWorks, accessed 3 August 2011

94. *Garcia, Christopher P.; et al. (January 2006).* "Round Trip Energy Efficiency of NASA Glenn Regenerative Fuel Cell System". *Preprint: 5.* Retrieved 4 August 2011.

15.   MEMS motors

15.1 Introduction.

MEMS motors and generators are rapidly developing in the world of microdevices. They are also discussed aspects of the larger electromechanical energy conversion systems into which they fit. When coupled to microturbomachinery, typical applications of these machines might include motor/compressors, motor/pumps, and turbine/generators, all perhaps part of much larger systems such as a microengine or a microrocket.

PowerMEMS motors and generators takes place in the context of the system shown in Fig. 15.1. This system connects an electrical source or load to a mechanical load or source through an electric or magnetic machine and its attendant power electronics. If the machine acts as a motor, then power flows from left to right. If the machine acts as a generator, then power flows from right to left. The role of the machine in this context is the conversion of electrical to mechanical energy, or vice versa. The role of the power electronics is to commutate the machine; it is assumed here that the machine is an AC machine, while the electrical load or source is a DC device. The operation of the system takes place under the supervision of a local controller, which executes the desired regulation functions. In every-day language, the terms "electric motor" and "electric generator" almost always refer to magnetic-field machines, rather than electric-field machines. This is understandable because electric-field machines are very rare in every-day life, while magnetic-field machines are ubiquitous. The language of this chapter, however, must be more precise because both types of machines will be discussed. Thus, from this point forward the adjectives "electric" and "magnetic" will be used literally so as to avoid confusion. A non-specific machine will be refered to as "electromagnetic."
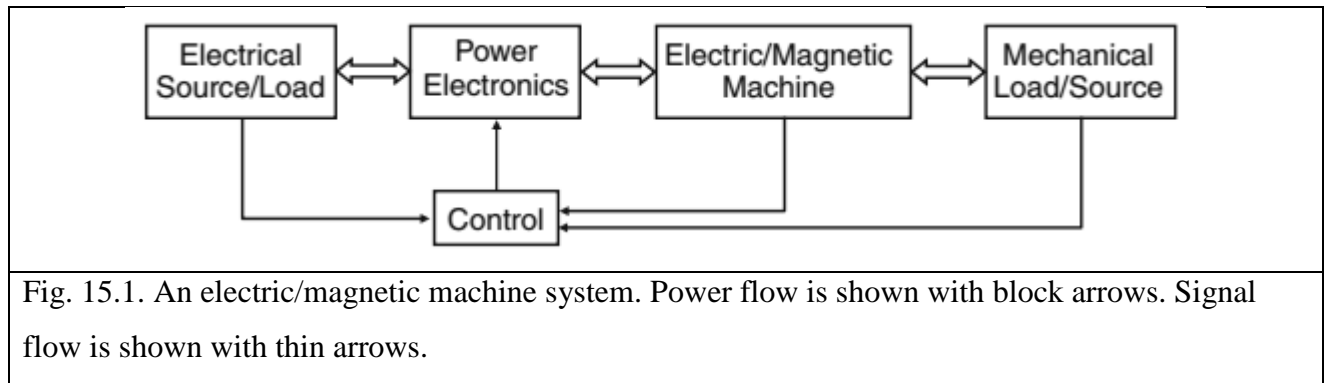
Fig. 15.1. An electric/magnetic machine system. Power flow is shown with block arrows. Signal flow is shown with thin arrows.

Maxwell's Equations exhibit symmetry in the equations which govern the behavior of electric and magnetic fields. There is also a strong formulaic similarity between the equations which govern the behavior of electric and magnetic materials. Because of this, there is a duality in the basic types of electric and magnetic machines. For example, Table 15.1 catalogs the dual pairs of some basic AC electric and magnetic machines.

Table 15.1

Electric machine	Magnetic machine
Variable capacitance	Variable reluctance
Permanent-electret	Permanent-magnet
Excited-rotor	Wound-field
Induction	Induction

 Other dual machine pairs can be created by combining the features of those listed in the table. In principle, any such machine could be implemented as a PowerMEMS machine. In that case, which one is the best choice for a given PowerMEMS motoring or generating application? While there is no universal answer to this question, there are good guidelines for a selection. To begin, those machines with externally excited rotors, specifically the electric excited-rotor and the magnetic wound-field synchronous machines, require brushes and slip rings to bring the excitation to their rotating rotor phases. At the macroscale, this poses few challenges, and is common practice. However, at the microscale there is no proven brush and slip ring technology. Further, even if such components were developed, they would likely cause a significant mechanical loss. Thus, these machines are eliminated here as impractical at the microscale. Similar concerns eliminate the use of brushes and commutator bars as a means of commutating AC machines, which would result in a traditional DC machine as viewed from the outermost electrical terminals. Consider next the electric variable-capacitance and magnetic variable-reluctance synchronous machines, which have both been implemented as MEMS machines. These machines usually use slots on both the stator and rotor to define salient poles that are excited by electrodes or windings on the stator. The poles then attract to produce a torque. In general, to be good torque producers, these machines require deep

slots and very narrow air gaps in order to create a large difference between the maximum and minimum phase capacitance or inductance. Further, they require relatively accurate rotor position information to support the commutation required for good performance, both as a motor and as a generator, which is challenging to measure at the microscale. While these machines exhibit excellent rotor integrity, they are nonetheless eliminated here due to their fabrication challenges, their control challenges, and the viscous loss they incur with a small air gap. After elimination of the others, only the permanent-electret and permanentmagnet synchronous machines, and the two asynchronous induction machines, remain. Each is a viable candidate for the Microengine Project, and the selection among them is more complex. Consider first a comparison of the two electric machines, and a separate comparison of the two magnetic machines. In an MEMS electric induction machine exhibiting a typical air-gap electric field magnitude E of 108 V/m, the induced surface charge on the rotor is approximately $\varepsilon_0 E$, or 0.9×10^{-3} C/m^2. This is very close to the $0.5 - 1.2 \times 10^{-3}$ C/m^2 surface charge reported for MEMS permanent electrets fabricated from thin films of Teflon, similar materials and silicon dioxide. Thus, the two machines appear electromechanically comparable at the microscale because it is the rotor surface charge which interacts with the stator excitation to produce a torque. However, the permanent electrets are not yet stable against leakage and decay over time, particularly at temperatures approaching 80°C and above. Further, being a synchronous machine, the permanent-electret machine requires position information to operate at its best, which is again difficult to provide at the microscale. Therefore, the electric induction machine is chosen here over the permanent-electret synchronous machine. The choice between the permanent-magnet synchronous machine and the magnetic induction machine is much less clear than in the case of the dual electric machines. Samarium-cobalt magnets are capable of high-performance operation even beyond 300◦C. For this reason, both machines have been studied in detail as PowerMEMS machines. The primary conclusion from this work is that, when permitted by thermal considerations, the permanent-magnet synchronous machine appears to be the better choice, and so it is chosen here. The permanent-magnet synchronous machine is generally more power dense and efficient than the induction machine at the microscale due largely to the challenge of inducing substantial rotor currents at that scale. Permanent magnets, on the other hand, maintain their bulk performance at the micro scale, easily offering a good rotor field excitation.

15.2 Electric Induction Motors

15.2.1 Design and Performance

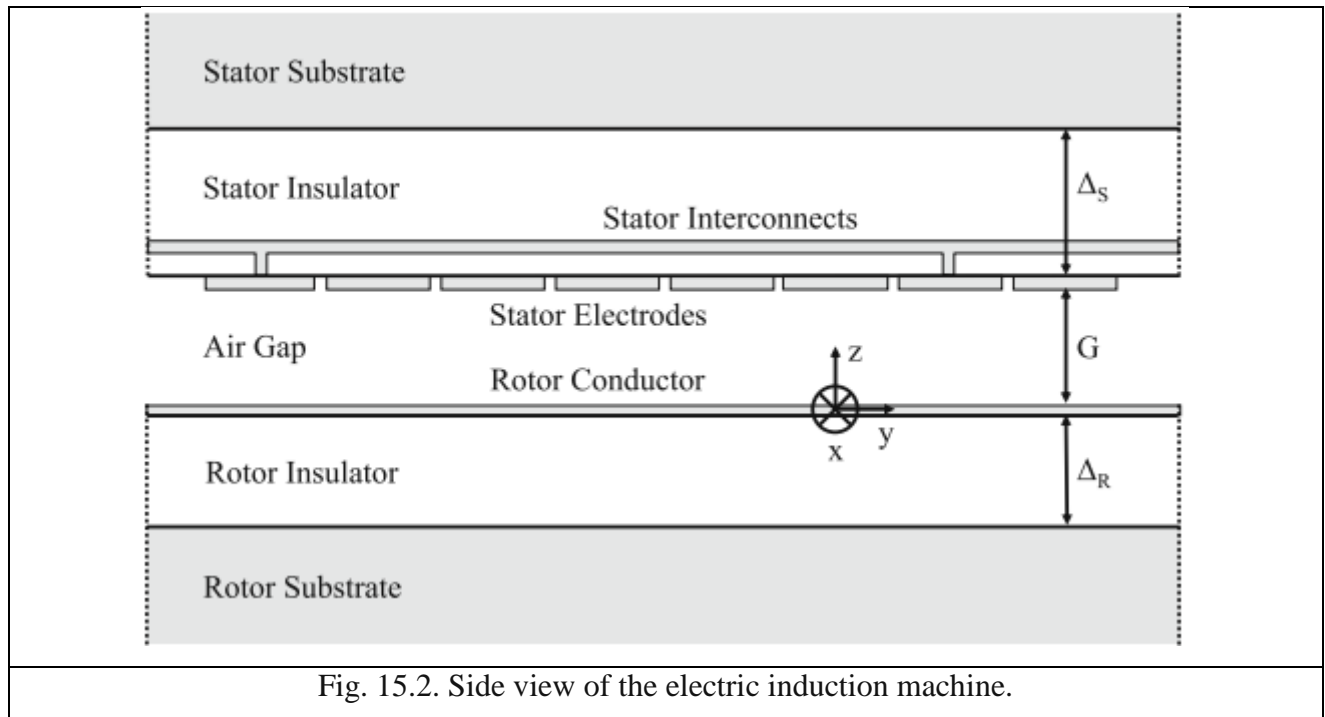A cross-section side view of the electric induction machine studied here is shown in Fig. 15.2

Fig. 15.2. Side view of the electric induction machine.

Photographs of the machine fabricated as a PowerMEMS machine are shown in Fig. 15.3. The stator comprises radially directed electrodes, like spokes, supported by an insulator on a conducting substrate. The stator electrodes are wired together through back-side interconnects to form N phases; every Nth electrode is connected to form a single phase. In both figures, N is 6. The interconnected electrode pattern is repeated P times around the stator. Thus, the total number of stator electrodes is NP.
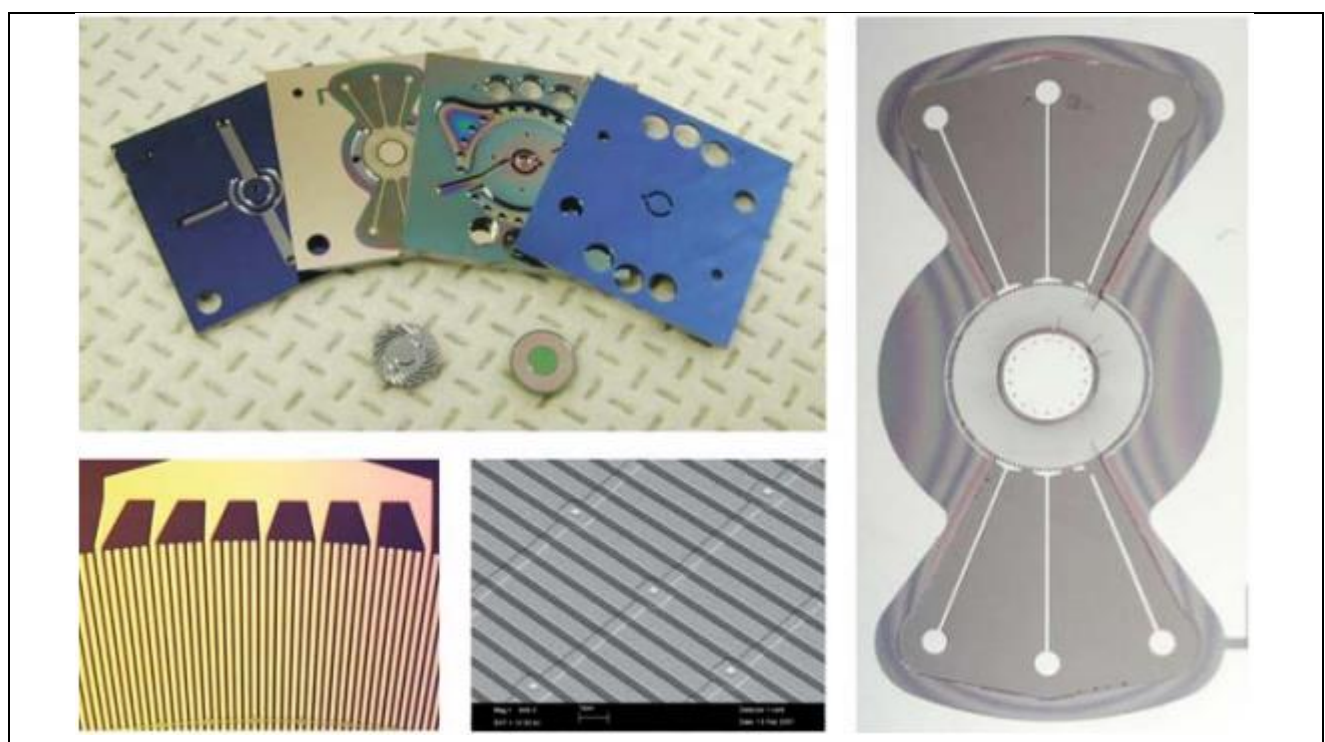
Fig. 15.3. Photographs of an MEMS electric induction machine from. The *top-left* photograph shows a five-wafer set which, when fully bonded, forms a complete motor/compressor or turbine/generator depending on the design of the blades. The middle-right wafer is already a bonded wafer pair that contains the rotor. The middle-left wafer is the stator wafer. The two outer wafers are foundation wafers that contain many of the air passages connecting to the compressor or turbine, and the air bearings. Below the wafers are two separated rotors. The rotor on the left shows its turbine blades, while the rotor on the right shows its annular rotor conductor surrounding a thrust bearing pad. The *right*-most photograph shows a close-up view of the stator fabricated on top of buried silicon dioxide in the shape of a bow tie. The six phase leads that terminate on contact pads are also visible. The stator electrodes in the central annular region are not distinguishable, but the air nozzles around the periphery of the center thrust bearing are visible. The *lower-left* photograph shows a close-up view of the stator electrodes near their connection to one of the phase leads. One interconnect ring is visible at the very bottom of the photograph. The *lower-middle* photograph shows another close-up view of the stator electrodes which shows three interconnect rings.

In Fig. 15.3, *P* is 131 for a total of 768 electrodes. The rotor consists of a thin uniform conductor also supported by an insulator on a conducting substrate. In the applications considered here, the bottom side of the rotor could be etched to form turbine or compressor blades. The blades are not shown in Fig. 15.2, but they are visible in Fig. 15.3. In the machines studied here, the substrates are silicon, the insulators are silicon dioxide, the stator electrodes and interconnects are either metal or heavily doped polysilicon, and the rotor conductor is moderately doped polysilicon. The stator phases of an electric induction machine are generally excited with balanced sinusoidal voltages, having a temporal phase shift of $2\pi/N$ from one phase to the next, and hence one electrode to the next. In this way, the phases support a traveling wave of potential that circulates around the stator. The stator potential and the corresponding charges on the stator electrodes induce image charges on the rotor conductor, as shown in Fig. 15.4.
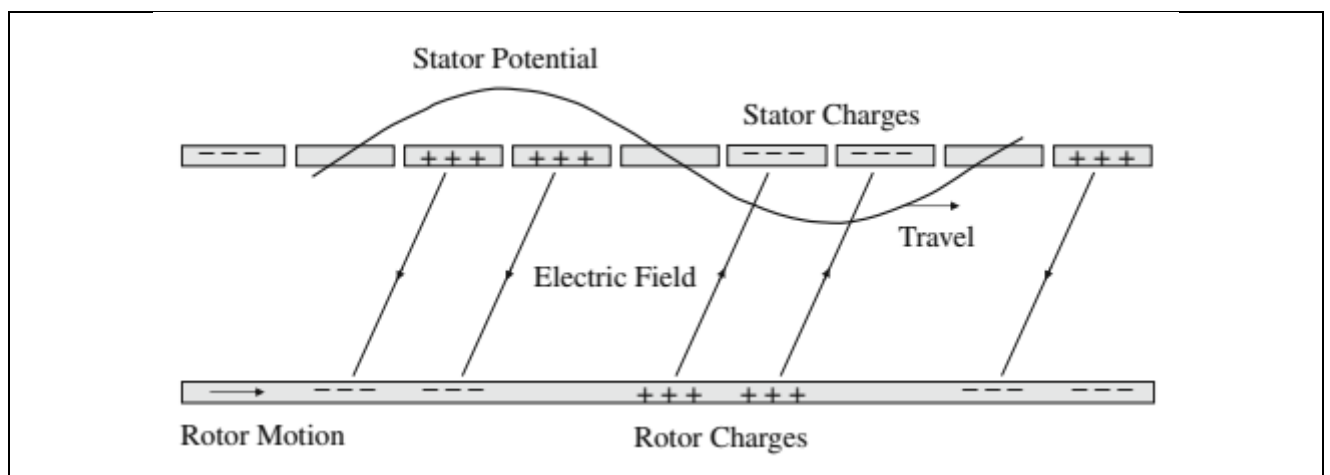
In steady state, the image charges on the rotor must travel in synchronism with the stator excitation. Part of this travel is provided by convection arising from the motion of the rotor, and part is provided by conduction through the rotor. During motoring operation, as shown in Fig. 15.4, the stator potential travels at a speed that exceeds that of the rotor motion, and so the rotor charges lag behind. To stay in synchronism with the stator potential, the rotor charges conduct forward through the rotor conductor, and this conduction is driven by the horizontal component of the air-gap electric field. This same electric field component produces the motoring torque acting on the rotor. During generating operation, the speed of the rotor motion exceeds that of the stator potential wave, and the rotor charges lead their images on the stator electrodes. In this case, the charges conduct backward through the rotor to maintain synchronism, and the horizontal component of the electric field is reversed.

15.2.2 Power Electronics and Control

To achieve its best performance as a motor or a generator, the electric induction machine must be operated at both high voltage and high frequency. Voltage amplitudes near 300 V and frequencies near 1 MHz are typical for highest power throughput. In addition, the applied stator voltage waveforms should be sinusoidal. Non-sinusoidal waveforms contain time harmonics that can degrade both motoring and generating performance, and they will unnecessarily increase the peak voltage between neighboring electrodes making electrical breakdown a more limiting phenomenon. Thus, hard-switching power electronics such as those shown in Fig. 15.5 are inappropriate for the electric induction machine, and will most likely be too lossy, particularly for generating applications.
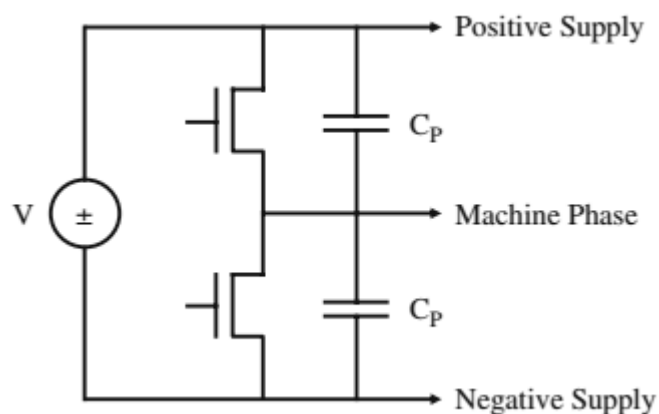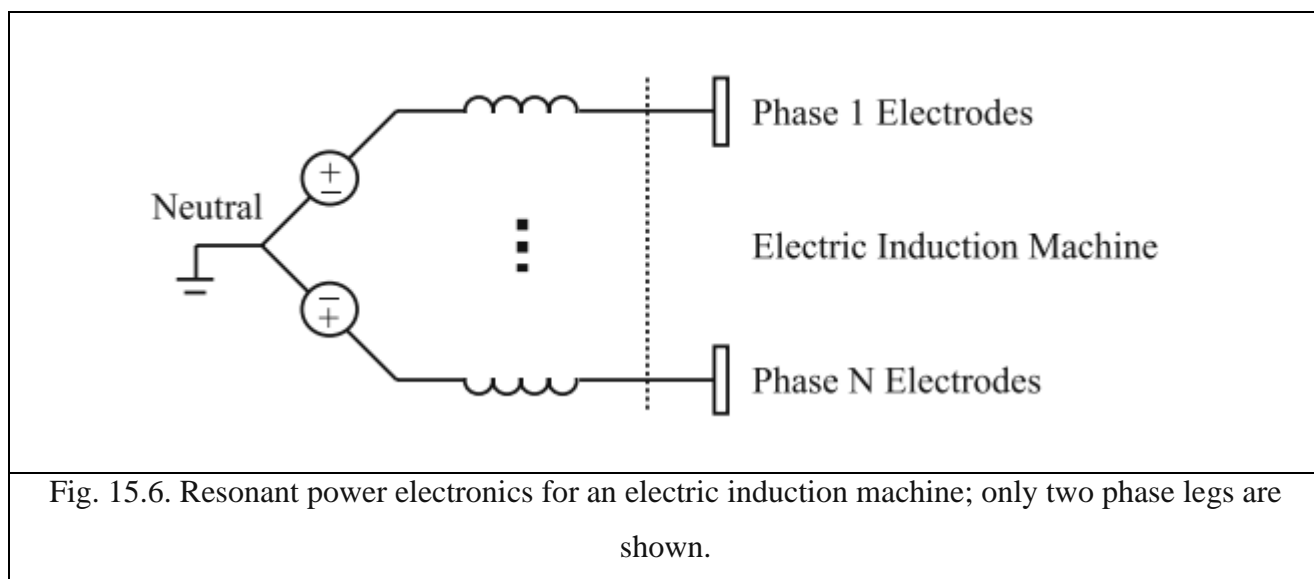

Fig. 15.5. Power electronics for an electric machine (only one phase branch is shown)

An alternative is required. One alternative is to use resonant power electronics such as those shown in Fig. 15.6.

Fig. 15.6. Resonant power electronics for an electric induction machine; only two phase legs are shown.

The electric induction machine is essentially a capacitive machine. The effective capacitance at its electrical terminals, which includes any parasitic capacitance, can be used to form a resonator with an external inductor. The function of the resonator is to step the voltage from the transistors in the power electronics up to the phases of the machine. The low-voltage sources are themselves sinusoidal to minimize the presence of time harmonics and the associated loss. This is important during generation. However, it is possible, particularly in motoring applications, that the sources can be simple switching square-wave sources as shown in Fig. 15.5 because each resonator acts as a band-pass filter.]. Whether or not this approach is acceptable will depend on the application. The advantage of resonant power electronics is that the transistors in the power electronics operate at a lower voltage, and so their switching losses can be greatly reduced. Voltage step-up ratios near 20 have been achieved, which lowers the transistor switching losses by a factor of 400. Resonant power electronics also have important disadvantages. One disadvantage is the need to tune each phase so as to maintain a balanced excitation at the electrodes of the machine. Without a balanced excitation, the amplitudes of the space harmonics may be large enough to significantly reduce the performance of the machine, particularly when used as a generator.

One source of imbalance is unbalanced parasitic capacitance in the machine itself. For example, the interconnect rings seen in Fig. 15.3 are concentric, with some rings inside others. The parasitic capacitance between these rings and the other phases will be different depending on the radius of the ring at least because the electrodes become wider as radius increases. Another source of imbalance is the inductors, no two of which will be the same. For this reason, it may be necessary to place tuning capacitors from each phase to neutral, and adjust them, and the voltage and temporal phase of the low-voltage sources, to achieve balanced excitation at the electrodes. Note that the presence of the tuning capacitors increases reactive currents and the associated losses, but the net

result can be an improvement in performance. A second disadvantage is the presence of the resonating inductors, and the losses that they bring at high voltage and high frequency. Low-loss inductors having the appropriate inductance that operate at high voltage and high frequency can be physically large, and almost certainly larger than the electric machine itself. A final disadvantage is that the effective capacitance of the electric induction machine will change with operating conditions. For this reason, it may be necessary to change the excitation frequency and temporal phase shifts of the low-voltage sources so as to keep the resonator tuned at resonance. A second alternative is to again use resonant electronics, but with a crystal as opposed to an inductor. Crystals exhibit inductive behavior, although they do so only over a limited frequency range. The advantage to using a crystal over an inductor is the potential for much smaller size, and much lower loss. Preliminary experiments with crystals indicate that they can tolerate acceptably high-voltage operation. A third alternative for providing a high-voltage and high-frequency excitation is to make use of multi-level power electronics.

Multiphase inverter can be implemented with a collection of identical single-phase inverters, one inverter per machine phase. Each single-phase inverter is in turn constructed from several identical full-bridge inverters whose outputs are transformer coupled and connected in series to produce a high net output voltage. By timing the switching of the individual inverters that constitute a single phase, the net output can be made nearly sinusoidal over a wide frequency range so that electrical breakdown between neighboring electrodes is not a limiting phenomenon. The advantages of such an inverter are again the low-voltage operation of the individual inverters that constitute a complete phase inverter. Additionally, the multi-level inverter is non-resonant, and so need not be tuned. Its primary disadvantages are the presence and size of the output transformers, and increased hardware complexity. Finally, a fourth alternative for providing a high-voltage and high-frequency excitation is the use of piezoelectric transformers. These transformers can also be used in resonant circuits to provided very-high-voltage sinusoidal outputs at veryhigh frequency. To date, however, they have not been pursued extensively in the context of inverters for MEMS electromagnetic machines. The control of an electric induction machine is identical in spirit to that of a magnetic induction machine since the two are electromagnetic duals.

As a motor, the electric induction machine may be successfully run open loop with a fixed electrical excitation, although care must be taken to anticipate the peak load so that the motor is capable of producing the necessary torque over the range of $0 \leq \gamma_1 \leq 2$. This is the manner in which most magnetic induction machines are operated today. Since it is an asynchronous machine, the electric induction machine will produce starting torque at zero speed, for which $S_m = 1$, even if it is excited at a high electrical frequency. Alternatively, a more complex control can be implemented with a controller that measures speed and adjusts the electrical excitation frequency ω so that the output torque follows the positive peak of Fig. 15.7.
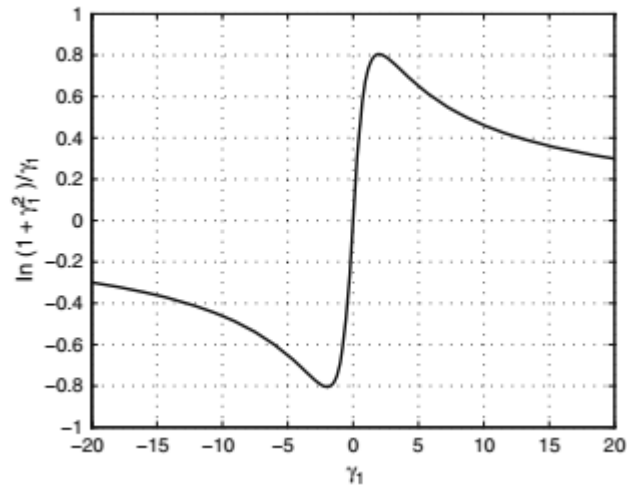
Fig. 15.7. A graph of $\ln(1 + \gamma_1 2)/\gamma_1$ as a function of γ_1

Yet more complex torque and speed controllers are discussed in the context of magnetic machines. As a generator, the electric induction machine may also be run successfully open loop with an electrical frequency that yields negative slip, namely such that $P > \omega$. Again, more complex control is possible, such as following the negative peak in Fig. 15.7. Interestingly, the electric induction generator with resonant power electronics is capable of self excitation.

15.3 Permanent-Magnet Motors

Unlike the electric induction machines, the specific permanent-magnet synchronous machines described in this section have not been fully integrated into a silicon housing with self-contained air bearings. In order to achieve this integration, several challenges must be met. These challenges include:

1. the integrated fabrication of deep laminations in a silicon substrate
2. the fabrication of multi-level three-phase windings on the stator surface
3. the design and operation of bearings that can support a heavy magnetic rotor
4. the design and fabrication of a rotor that can hold together at speeds approaching 300 krpm to 400 krpm.

The permanent-magnet synchronous machine (PMSM) consists of two-key components, a rotor with a multi-poled permanent magnet, and a stator with appropriately designed windings. During operation the rotating multi-poled permanentmagnet rotor creates a time-varying magnetic flux in the rotor–stator air gap. This flux induces AC voltages at the terminals of the stator windings, which forms the basis for power generation. The PMSMs considered in this section use an annular permanent magnet mounted on a ferromagnetic core. Interior PMSMs are not considered here due to the difficulty of embedding magnets into an electroplated ferromagnetic core. With the use of suitably thick magnets (500 μm) and highperformance magnetic materials in the rotor and stator core, the air gap can be made very large (300–500 μm) without appreciable loss of performance.

This allows the stator windings to occupy space in the air gap, fabricated on the surface of a flat stator core as shown in Fig. 15.8. PMSM fabrication is thus greatly simplified. A fabricated stator is shown in Fig. 15.9.
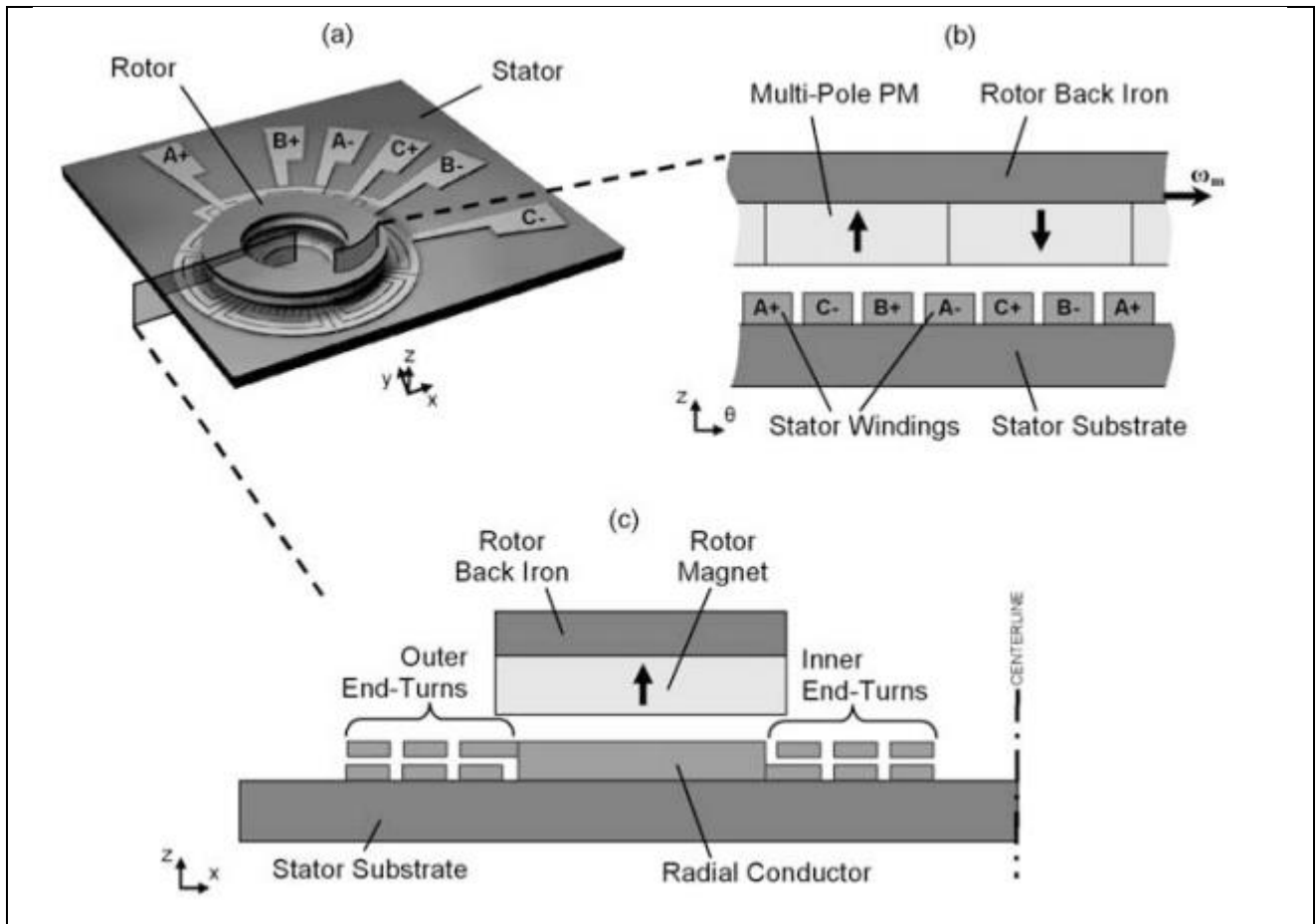

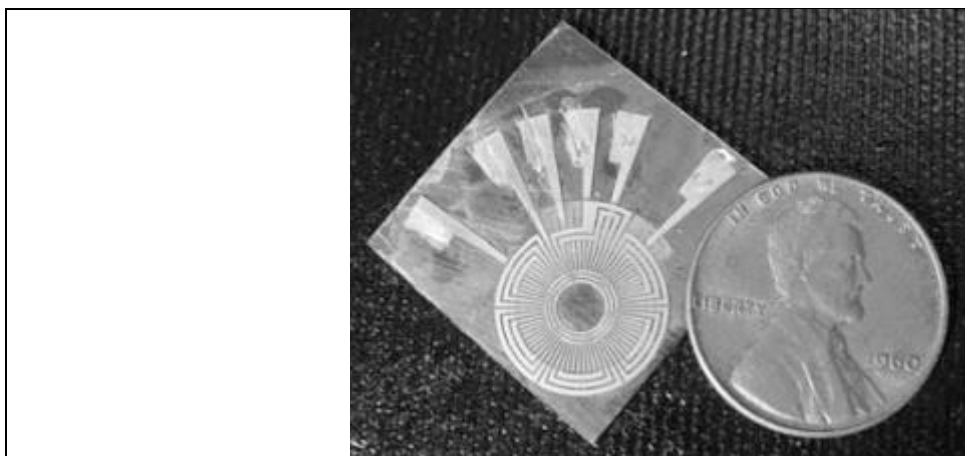Fig. 15.8. Drawing of the PMSM: (a) perspective view, (b) angular cross section, and (c) radial cross section.

Fig. 15.9. Fabricated 3-turn/pole concentric wound stator is shown along side a penny for size comparison

Microscale magnetic machines can match the linear tip speeds of their macroscale counterparts by spinning much faster, reaching speeds of 100–1000 krpm. This, combined with the higher current

densities allowed in the windings, enables microscale machines with very high-power densities. Microscale magnetic machines operate at electrical frequencies one to two orders of magnitude higher than typical macroscale machines, typically in the kHz to tens of kHz range, due to their higher rotational speeds. These high-operating electrical frequencies lead to smaller passive filtering components in their attendant switch-mode power electronics. The small size of microscale machines also leads to low stator inductances. This permits both generator and motor drives with fast transient responses, but also requires very high switching frequencies (hundreds of kHz to low MHz) in their power electronics.

Prior work has focused on planar MEMS PMSMs. Several axial-air-gap motors have been demonstrated, and speeds of up to 200 krpm with torques of up to 7.5 µNm have been achieved. Planar three-phase motors using magnetic thrust bearings have also been demonstrated. Generators too have been demonstrated. One was a turbine/generator using NdFeB permanent magnets that produced 1.1 mW at 30 krpm with an air flow of 35 L/min. The composite turbine/generator rotor was supported by conventional ball-race bearings, and measured 1 mm in thickness and 6 mm in diameter. Another demonstrated highspeed rotation (260 krpm) by using a combination of air and magnetic bearings. Using an 8-mm diameter $SmCo_5$ rotor, 14.6 mW of power was delivered to a wyeconnected resistive load at 58 krpm. Using an air-powered drill and 8-mm-diameter NdFeB permanent magnets, the same generator delivered 5 W into a wye-connected resistive load at 380 krpm. The previous work on microscale PMSMs for power generation focused solely on the generator.

15.4.1 Design

The surface-wound PMSM is chosen as the best candidate for efficient high power generation in the microengine. This section focuses on the design of a microscale PMSMs capable of generating and deliveringWatt-level power at speeds of approximately 300 krpm. The emphasis will be on DC power generation via power electronics that rectify the AC voltages delivered by the generator phases. Since integration issues are not dealt with at this stage, silicon wafers and electroplated core materials will not be used. Instead, a ferromagnetic wafer will be used as the stator core. Since this wafer will not be laminated, eddy current losses will degrade efficiency but will not affect output power provided that the turbine that spins the rotor provides sufficient mechanical power to overcome the added losses.

The PMSMs considered here are three-phase axial-air-gap machines. Each machine consists of a rotor with an annular multi-pole permanent magnet and ferromagnetic core, and a stator with multi-turn windings on the surface of a ferromagnetic substrate which serves as the stator back iron as shown in Fig. 15.10. During operation, the rotating multi-poled permanent-magnet rotor creates a timevarying magnetic flux in the rotor–stator air gap, which induces AC voltages at the terminals of

the stator windings. The magnetic flux generated by the permanent magnets can be concentrated in the rotor–stator air gap by using cores with very large permeabilities. By using rare earth permanent magnets, which have large remanences (~1 T), strong magnetic fluxes can be generated in the air gap. In fact, by using high-performance magnetic materials (permanent magnet, and rotor and stator cores) and suitably thick magnets (~500 µm), the actual rotor–stator air gap can be made relatively large (300–500 µm) to reduce windage loss while maintaining high magnetic flux in the air gap. Thus, the stator windings can be hundreds of microns thick and occupy space in the air gap on the surface of the stator substrate.
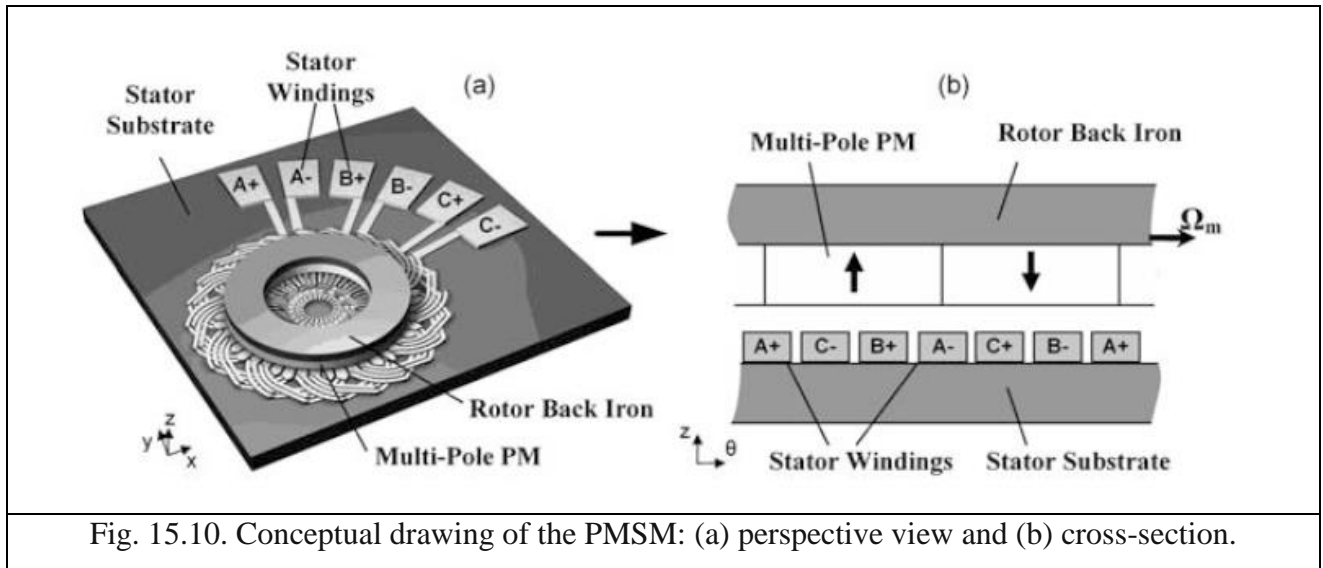
Fig. 15.10. Conceptual drawing of the PMSM: (a) perspective view and (b) cross-section.

15.4.1.1 Stator Design.

The stator uses interleaved, multi-layer electroplated copper windings that are dielectrically isolated from a 1-mm thick $Ni_{80}Fe_{15}Mo_5$ (Moly Permalloy) substrate by a 3-µm spin-on-glass layer and/or a 5-µm polyimide layer. The active area of the stator has an outer diameter of 9.5 mm, and an inner diameter of 5.5 mm. NiFeMo is selected as the stator core (substrate) for its combination of high permeability ($\mu r > 104$), low coercivity ($Hc \sim 0.16$ A/m) to minimize hysteresis losses, and commercial availability in sheets of suitable thickness.

An optimized 2nd-generation 8-pole 3-turn-per-pole stator is shown in Fig. 15.11. The winding pattern is designed for minimal resistance to ensure optimal generator performance. These stators use a single layer for the radial conductors that are 200 µm thick.
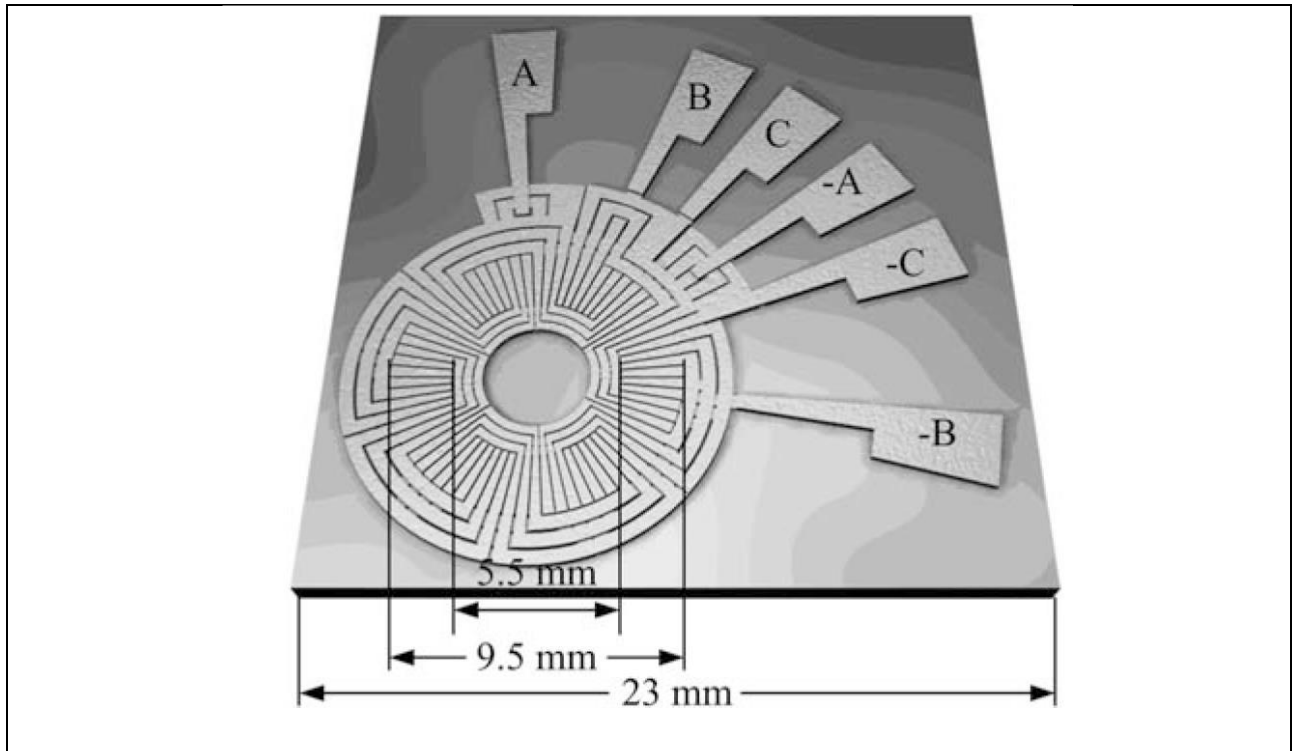
Fig. 15.11. 8-pole, 3-turn/pole concentric wound stator

The end turns occupy two layers and are each 80 μm thick. There is a 40 μm thick SU8 insulation layer between the top and bottom end turns. The stators use concentric windings, shown in Fig. 15.12, in which a complete end turn occupies a single layer. This eliminates the need for cross-over conductors like the ones used in the 1st-generation stators.
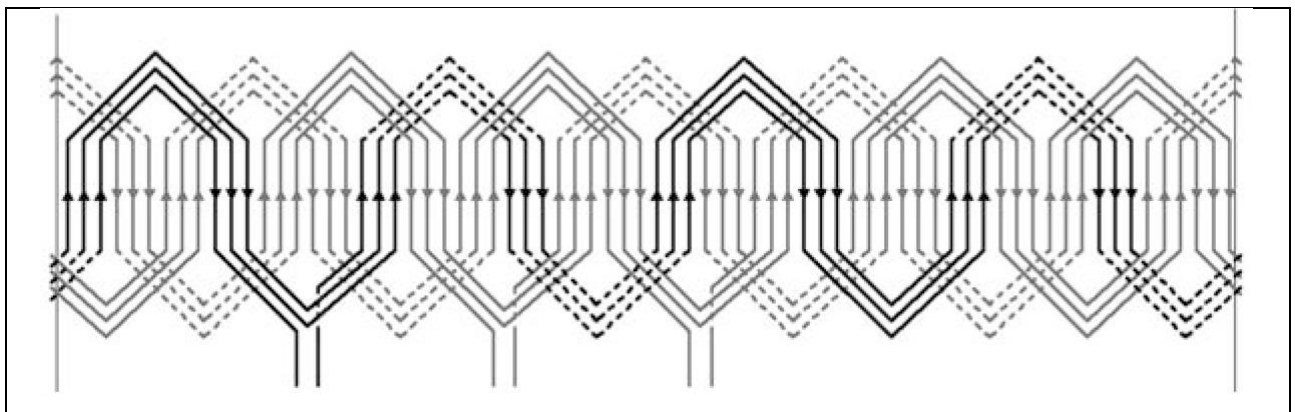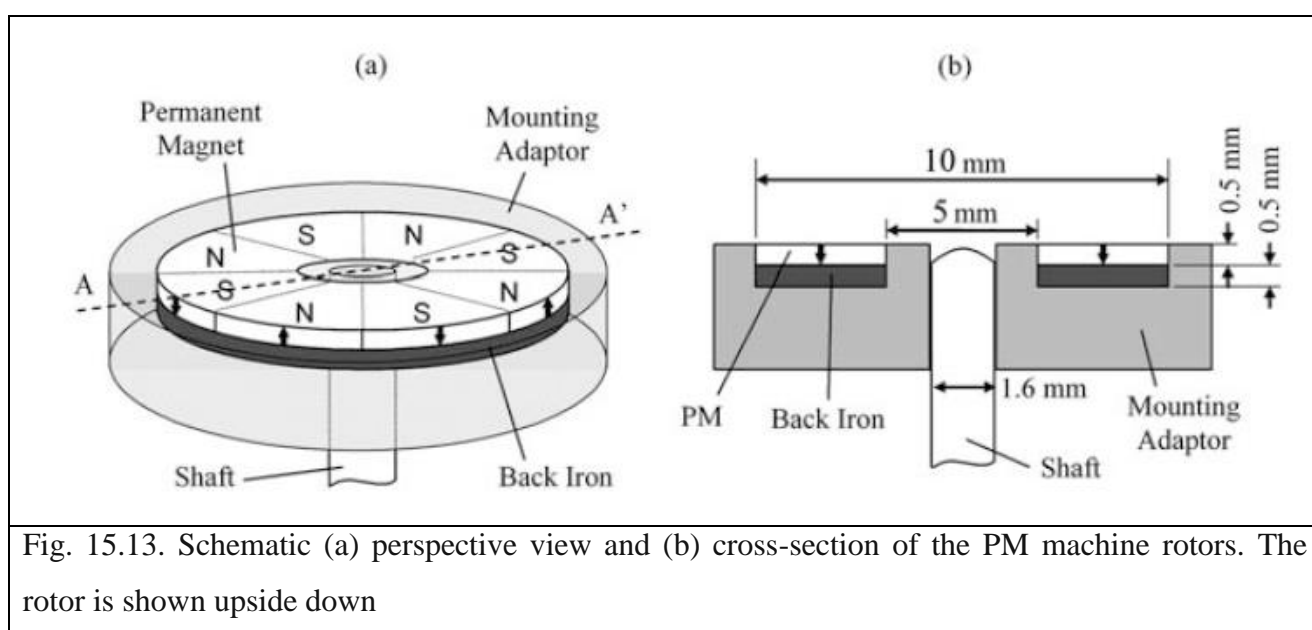

Fig. 15.12. Winding diagram for the 3-turn-per-pole concentric-wound stator. Solid lines represent end turns on the top layer while dashed lines represent end turns on the bottom layer. The radial conductors, marked with arrows, occupy both coil layers. Phase A is darkened for reference.

In fact, the only cross-over conductors needed are for connections to the pads. With fewer crossovers the number of vias is greatly reduced, from 96 down to 3. In addition, by using concentric windings, the end turns can be made to be circular arcs. This makes much more efficient use of the surface area of the inner and outer regions of the stator than the triangular end turns used

in the 1st-generation devices. Furthermore, variable-width coils are used with the gaps between adjacent radial conductors and end turns equal to 50 μm rather than 130 μm. The combination of the reduced number of vias, better end turn design, and more aggressive fabrication leads to a predicted stator resistance of 100 mΩ. This resistance is about the same as that of a 1st-generation 2-turn-per-pole machine. In fact, the predicted output power capability is increased by a factor of 2.25 over the 2-turn-per-pole machines due to the improved stator resistance.

15.4.1.2 Rotor Design

The rotor contains an 8-pole, annular, pressure-formed (sintered) Sm2Co17 permanent magnet and a ferromagnetic $Fe_{49}Co_{49}V_2$ (Hiperco 50) core, as shown in Fig. 15.13. The rotor components are mounted in a titanium adaptor for assembly onto a 1.6-mm diameter shaft for testing.

Fig. 15.13. Schematic (a) perspective view and (b) cross-section of the PM machine rotors. The rotor is shown upside down

The permanent magnet and rotor core are 500 μm thick to be compatible with standard wafer thicknesses, and have an outer diameter of 10 mm and inner diameter of 5 mm. This outer diameter was chosen because it was the closest match to the 10-mm outer diameter of the microengine generator. SmCo magnets were selected for their combination of high-energy product (BH_{max} ~240 kJ/m^3) for high-energy conversion, and high-operating temperatures (T_{max} ~300 °C). Although NdFeB has a higher energy product (BH_{max} ~400 kJ/m^3), it does not provide the necessary operating temperatures (T_{max} ~150 °C) for integration with the combustion-driven microengine. However, it would be suitable for a low-temperature turbine generator. Hiperco 50 was selected for the rotor core for its combination of a reasonably high permeability ($\mu_r > 3000$), and its high saturation flux density (B_s ~ 2.4 T) to minimize the thickness of the cores while avoiding magnetic saturation. Note

that the fields in the rotor core are nearly time-invariant, so the coercivity of the rotor back iron is not a major concern.

15.4.1.3 Power Electronics

For the microengine system to replace batteries, it must provide a constant DC output voltage. The PMSMpresented in the previous section produces a three-phase AC voltage. A PMSM can work with both passive and active power electronics to create a DC output voltage. If the microengine is to provide a 12-V battery replacement, for example, then the power electronics must rectify and boost the AC voltages from the generator, while operating under closed-loop control to provide a regulated output voltage. One way to passively step up and rectify the AC voltages is to use a Δ/wyeconnected transformer and diode bridge rectifier as shown in Fig. 15.14.

15.1
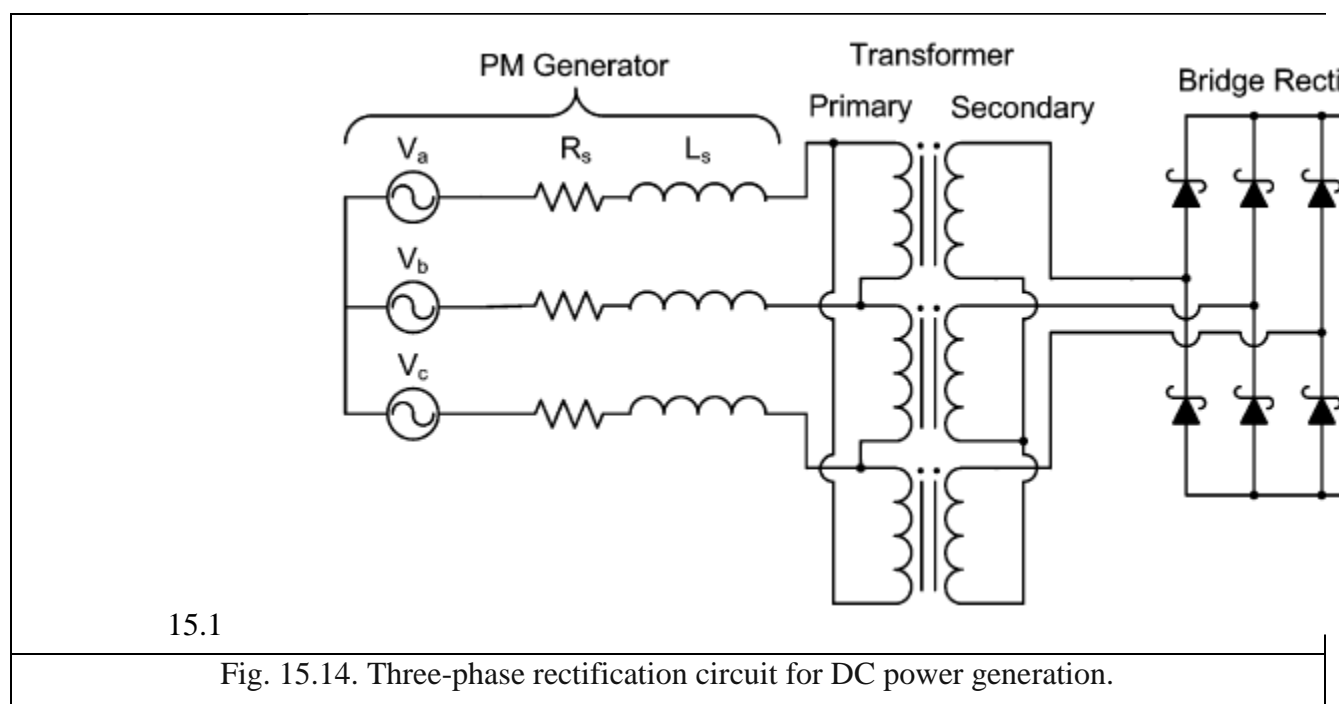
Fig. 15.14. Three-phase rectification circuit for DC power generation.

The generator windings are connected in a wye configuration and tied to the Δ-connected primary side of the transformer. This gives an additional factor-of-three increase from the line-to-neutral voltage of the PMSM to the line-to-line voltage on the transformer secondary side. A linear regulator at the output of the rectifier can then be used to achieve output voltage regulation. Passive electronics, however, tend to be large and bulky. Transformers designed for use in the stator electrical frequency range of 10–20 kHz can be very large compared to the PMSM. The three-phase transformer is eight times the volume of the PMSM stator. In addition, the transformer/diode bridge combination cannot provide power factor correction. Power factor correction is important at high speeds where the machine and transformer inductances will start to create a phase lag in the generator current relative to the generator voltage, reducing output power capability. At only 100 krpm the power factor for the transformer and diode bridge is 93.4%. At 300 krpm, the power factor drops to 90% while at 500 krpm the power factor is only 82.6%. Clearly, as the speed rises, passive

power electronics become a poor choice for power conversion. Active power electronics can eliminate the transformer, diode bridge, and linear regulator, and provide all four functions (power factor correction, rectification, voltage step-up, and regulation) in a much more compact and efficient way. Active power electronics for a macroscale PMSMoften requires speed and/or position sensing to determine the position of the magnetic poles on the rotor. This information is used to keep the stator currents in phase with the voltage to achieve maximum energy/power conversion. Hall effect sensors are sometimes used to directly measure the rotor position but are difficult to integrate into a microscale PMSM. "Sensorless" techniques to estimate the rotor position from current and terminal voltage measurements are also used in macroscale PMSMs but are computationally intensive, often requiring the use of high-performance microprocessors or digital signal processors. Real time processing of current and voltage information to determine the proper turn-on and turn-off times of the power transistors would be difficult achieve for the high speeds in the microscale PMSMs. The active power electronics designed for microscale a PMSM uses a three-phase boost semi-bridge topology, shown in Fig. 15.15, operating in discontinuous conduction mode (DCM).
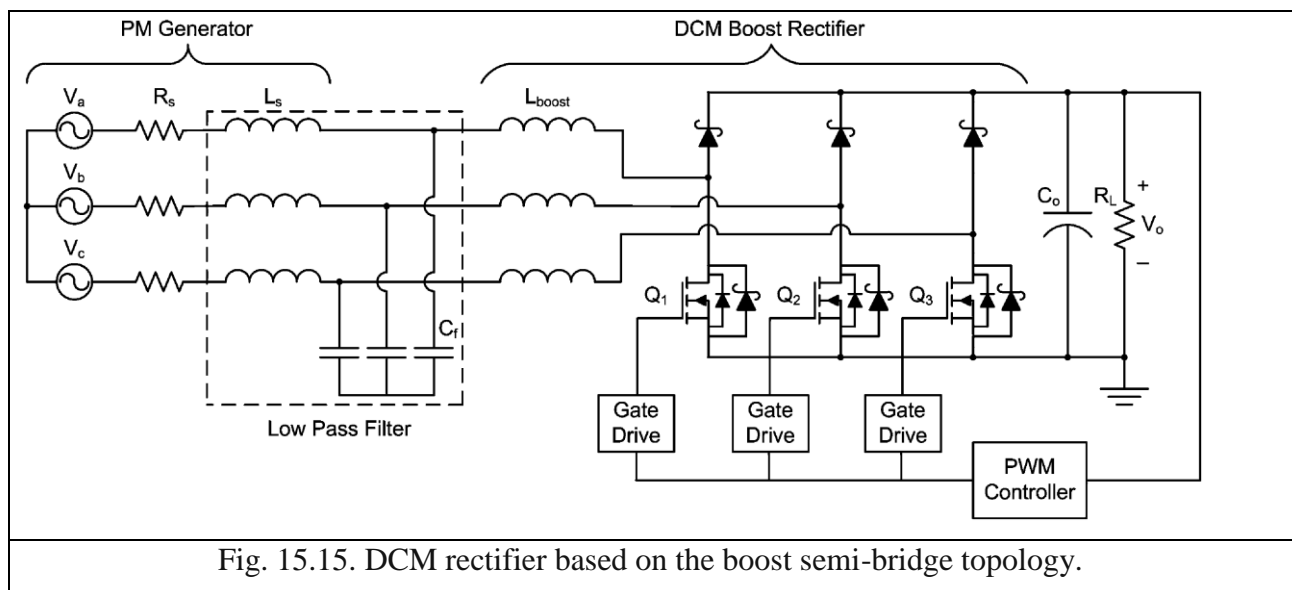

Fig. 15.15. DCM rectifier based on the boost semi-bridge topology.

Independent control of all three phases, as well as direct position, speed, current or voltage sensing is not needed to align the stator currents with the line-to-neutral voltages due to DCM nature of the circuit. The three power MOSFETs are driven from the same control signal and are turned on/off at a constant switching frequency much higher than the stator electrical frequency. The converter turns the PMSM into a current source that can be controlled with a voltage control loop implemented using an off-the-shelf pulse-width-modulated (PWM) control IC. This topology provides near unity power factor and low-harmonic distortion without the need for large low-frequency passive magnetic components. The boost inductors are separate external inductors. The machine inductance and resistance, along with an external capacitor, act as a damped low-pass LC filter to attenuate the ripple current from the boost inductors so that the line voltage sees mainly a

sinusoidal current at the line frequency. By using separate inductors, the high-frequency high-ripple currents are localized on the printed circuit board to minimize electromagnetic interference (EMI). In addition, the RMS stator currents are primarily due to the fundamental harmonic. This leads to low-harmonic distortion and reduced conduction loss in the stator windings. Even though the RMS currents are larger in the boost inductor, the resistance of the boost inductor is much smaller than the stator resistance (by design) so the conduction loss in the inductor is a small fraction of the overall conduction loss. At the beginning of each switching period the three MOSFETs are turned on and the currents through the boost inductors rise to a value proportional to their respective open-circuit line-to-neutral voltages. The duty ratio determines the fraction of the switching period that the switches remain on. When the MOSFETs are turned off, the currents will flow through the diode bridge. If the currents through the boost inductor are positive, then current will flow through the high side (or boost) diode. If the currents are negative, they will flow through the free wheeling diode in parallel with the MOSFET. The peak boost inductor current is proportional to the average value of the lineto- neutral voltage during the on-time of the MOSFET. This average value varies sinusoidally. Since the current in the boost inductor returns to zero each period, the variation in the average value of the current over a switching period will also be approximately sinusoidal, and will be in phase with the open-circuit voltage. For a DC output voltage much larger than the open-circuit voltage, such as is the case for microscale PMSMs, the time it takes for the boost inductor current to return to zero is much shorter than the on-time of the switch. Therefore, the average current over a switching period is almost independent of the output voltage leading to lowharmonic distortion of the current waveform.

16.   Millipede memory

16.1 Introduction

Millipede memory is a non-volatile computer memory stored on nanoscopic pits burned into the surface of a thin polymer layer, read and written by a MEMS-based probe. It promised a data density of more than 1 terabit per square inch (1 gigabit per square millimeter), which is about the limit of the perpendicular recording hard drives. Millipede storage technology was pursued as a potential replacement for magnetic recording in hard drives, at the same time reducing the form-factor to that of flash media. IBM demonstrated a prototype millipede storage device at CeBIT 2005, and was trying to make the technology commercially available by the end of 2007. However, because of concurrent advances in competing storage technologies, no commercial product has been made available since then.

16.2 Millipede concept

The main memory of modern computers is constructed from one of a number of DRAM-related devices. DRAM basically consists of a series of capacitors, which store data as the presence or absence of electrical charge. Each capacitor and its associated control circuitry, referred to as a cell, holds one bit, and bits can be read or written in large blocks at the same time. In contrast, hard drives store data on a disk that is covered with a magnetic material; data is represented as local magnetisation of this material. Reading and writing are accomplished by a single head, which waits for the requested memory location to pass under the head while the disk spins. As a result, the drive's performance is limited by the mechanical speed of the motor, and is generally hundreds of thousands of times slower than DRAM. However, since the "cells" in a hard drive are much smaller, the storage density is much higher than DRAM.

Millipede storage attempts to combine features of both. Like a hard drive, millipede stores data in a substrate or medium and accesses the data by moving the medium under the head as well. However, millipede uses many nanoscopic heads that can read and write in parallel, thereby

increasing the throughput. Additionally, millipede's physical medium stores a bit in a small area, leading to high storage densities. Mechanically, millipede uses numerous atomic force probes each of which is responsible for reading and writing a large number of bits associated with it. Bits are stored as a pit, or the absence of one, in the surface of a thermo-active polymer deposited as a thin film on a carrier known as the sled. Any one probe can only read or write a fairly small area of the sled available to it, a storage field. Normally the sled is moved to position the selected bits under the probe using electromechanical actuators similar to those that position the read/write head in a typical hard drive, although the actual distance moved is tiny. The sled is moved in a scanning pattern to bring the requested bits under the probe, a process known as x/y scan.

The amount of memory serviced by any one field/probe pair is fairly small, but so is its physical size. Many such field/probe pairs are used to make up a memory device. Data reads and writes can be spread across many fields in parallel, increasing the throughput and improving the access times. For instance, a single 32-bit value would normally be written as a set of single bits sent to 32 different fields. In the initial experimental devices, the probes were mounted in a 32x32 grid for a total of 1,024 probes. As the layout looked like the legs on a millipede (animal), the name stuck. The design of the cantilever array involves making numerous mechanical cantilevers, on which a probe has to be mounted. All the cantilevers are made entirely out of silicon, using surface micromachining at the wafer surface.

Non-crosslinked polymers retain a low glass temperature, around 120 °C for PMMA and if the tip is heated to above the glass temperature, it leaves a small indentation. Indentations have been made at 3 nm lateral resolution. By heating the probe immediately next to an indentation, the polymer will re-melt and fill in the indentation, erasing it (see also: thermo-mechanical Scanning Probe Lithography). After writing, the probe tip can be used to read the indentations. If each indentation is treated as one bit then a storage density of 0.9 Tb/in^2 could theoretically be achieved.

16.3 Reading and writing data

Each probe in the cantilever array stores and reads data thermo-mechanically, handling one bit at a time.To accomplish a read, the probe tip is heated to around 300 °C and moved in proximity to the data sled. If the probe is located over a pit the cantilever will push it into the hole, increasing the surface area in contact with the sled, and in turn increasing the cooling as heat leaks into the sled from the probe. In the case where there is no pit at that location, only the very tip of the probe remains in contact with the sled, and the heat leaks away more slowly. The electrical resistance of the probe is a function of its temperature, rising with increasing temperature. Thus when the probe drops into a pit and cools, this registers as a drop in resistance. A low resistance will be translated to a "1" bit, or a "0" bit otherwise. While reading an entire storage field, the tip is dragged over the entire surface and the resistance changes are constantly monitored. To write a bit, the tip of the probe is heated to a temperature above the glass transition temperature of the polymer used to

manufacture the data sled, which is generally acrylic glass. In this case the transition temperature is around 400 °C. To write a "1", the polymer in proximity to the tip is softened, and then the tip is gently touched to it, causing a dent. To erase the bit and return it to the zero state, the tip is instead pulled up from the surface, allowing surface tension to pull the surface flat again. Older experimental systems used a variety of erasure techniques that were generally more time consuming and less successful. These older systems offered around 100,000 erases, but the available references do not contain enough information to say if this has been improved with the newer techniques. As one might expect, the need to heat the probes requires a fairly large amount of power for general operation. However, the exact amount is dependent on the speed that data is being accessed; at slower rates the cooling during read is smaller, as is the number of times the probe has to be heated to a higher temperature to write. When operated at data rates of a few megabits per second, Millipede is expected to consume about 100 milliwatts, which is in the range of flash memory technology and considerably below hard drives. However, one of the main advantages of the Millipede design is that it is highly parallel, allowing it to run at much higher speeds into the GB/s. At these sorts of speeds one might expect power requirements more closely matching current hard drives. Data transfer speed is limited to the kilobits-per-second range for an individual probe, which amounts to a few megabits for an entire array. Experiments done at IBM's Almaden Research Center showed that individual tips could support data rates as high as 1 - 2 megabits per second, potentially offering aggregate speeds in the GB/s range.

Writing operation

As explained earlier, information is stored in the form of 1's and 0's by making indentations and no indentations in the storage medium. This is achieved by a process called thermo mechanical writing which involves both mechanical and heating operations. In this process, the probe tip is heated using local heaters which soften the storage surface making it ready for writing. The heating makes it easier to penetrate the surface. Initially, the heating from the tip to the storage medium is through air which inhibits the heat transfer, but this improves as the contact area increases when the tip moves into the surface. This means that the tip must be heated to a relatively high temperature of about 400ºC to initiate the softening. Simultaneously, an electric pulse is sent to the cantilever which bends due to the atomic attraction between atoms on the medium surface and on the tip.

This bending presses the tip into the polymer, forming an indentation, which is denoted as 1.Once softening has initiated, wherever a binary 1 has to be written on the storage medium there the tip is pressed into the polymer, hence forming an indentation and on the other hand for a binary 0 to be written no indentation is made which clearly distinguishes 1's and 0's.Also, to form a stable indentation the local surface must be rapidly cooled after the write operation.

Thermo mechanical writing of data in this manner can make the storage medium to occupy bits of about 40nm size without being merged, implying a potential storage density on the order of 400 Gbit/in^2.

For example, a (32×32) cantilever array with 1024 storage fields, each having an area of 100×100 m2, has a capacity of about 6 G bit on a chip area on the order of 3×3 mm2. More recently single-cantilever areal densities up to 1 Tbit/in2 have been demonstrated, although currently at a somewhat degraded write/read quality.

Fig. Millpede writting operation
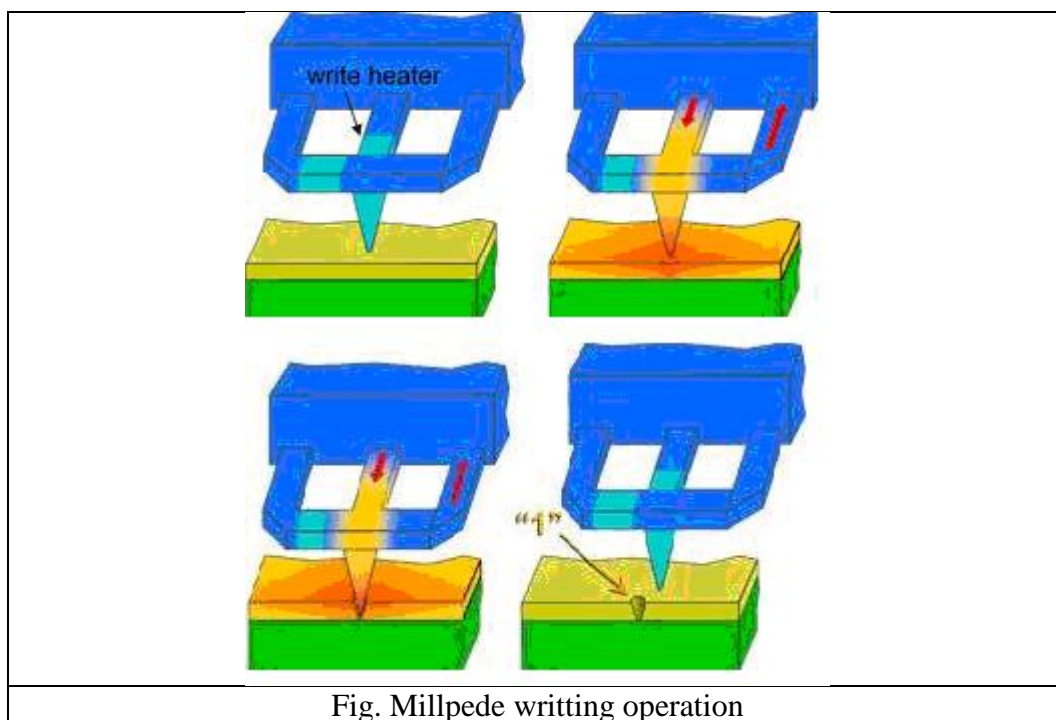
Reading operation

Information stored in the form of indentations and non-indentations on the surface can be read only when there is a conspicuous way of distinguishing them. The reading operation should involve a convenient process identifying them and designate 0's and 1's accordingly. To achieve this, the same probe tip attached cantilever is used. Also, the temperature dependent resistance of the probe tip is the main principle behind the read operation.

In general, the resistance increases nonlinearly with heating power/temperature from room temperature to a peak value of 500-700 ℃. The peak temperature is determined by the doping concentration of the heater platform, which ranges from $1×10^{17}$ to $2×10^{18}$ cm^3. First the tip is heated to a temperature of about 350 ℃ so that the storage surface does not melt. Then the cantilever is bent by applying a read electrical pulse which in turn produces a mechanical force. As the tip moves into the indentation the heat transfer rate improves and so the cantilever cools. This temperature decrease produces a change in resistance of about 10^{-4} to $5×10^{-4}$. This change in resistance is converted by a thermo mechanical sensor into an electrical signal. This sensor can be approximated to a variable resistance and the voltage across this resistance is detected by the

detector, which produces a binary 1. The change in electrical resistance of the cantilever during this process is of the same order. Thus, a signal of high resolution is required to distinguish between these two changes in resistance and to read the information correctly.
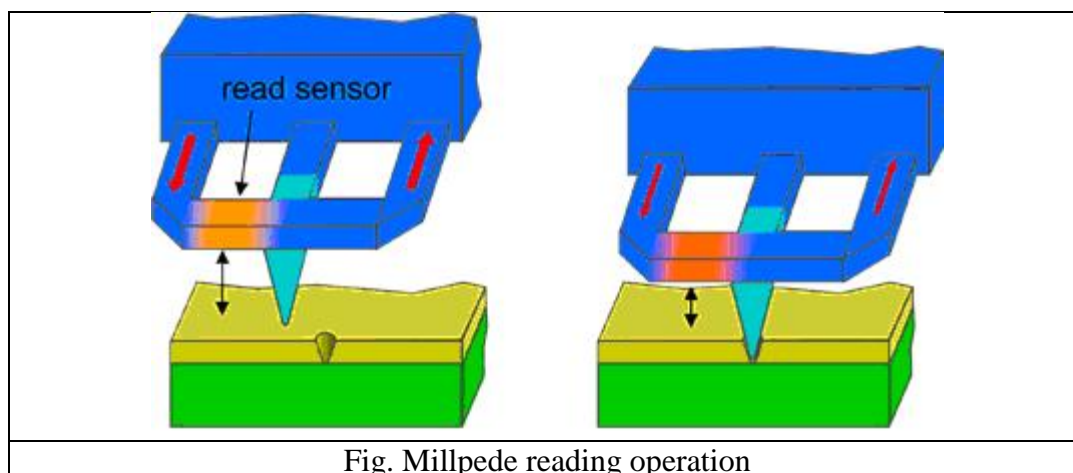

Fig. Millpede reading operation

The maximum temperature attained by the cantilever while moving over an indentation is always less than the temperature attained when there is no indentation present. As the value of the variable resistance depends on the temperature of the cantilever, the maximum value achieved by the resistance will be smaller as the cantilever moves over an indentation. Therefore, during the read process, binary 1's and 0's can be easily distinguished and transmitted.

Erase operation

In the thermal-mechanical writing process described above, indentations are created by applying a force on the cantilever. Then the indent surface is rapidly cooled, resulting in the creation of a "meta-stable" indentation. If the storage medium is reheated, it softens, thereby allowing relaxation of the surface and hence erasing the indentation. This can be done at the individual bit level by using the tip as a localized heat source.

Also, another way of erasing the old indentation is to write the new indentation very close to the old one thus erasing it. Hence, a previously written data track can be erased simply by overwriting it with a series of closely spaced indentations. In other words, erasing is essentially similar to writing at a narrower pitch between indentations.
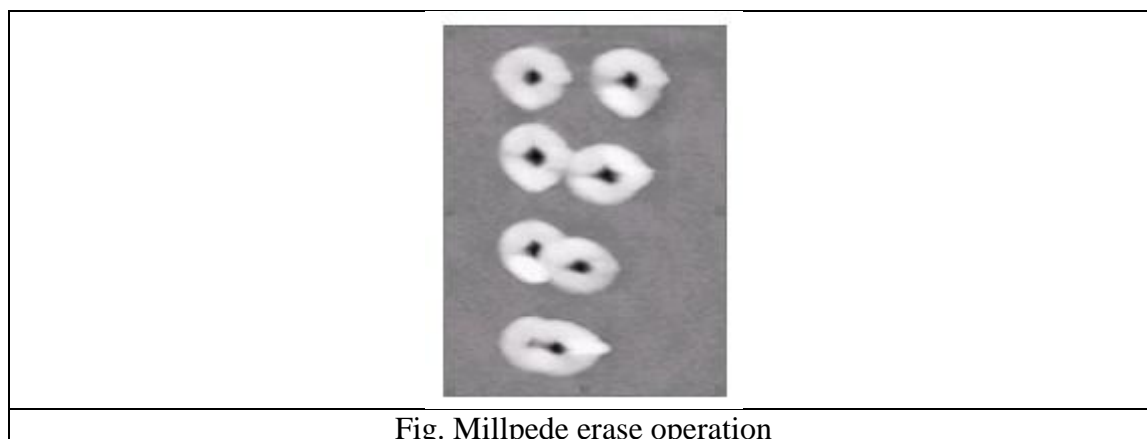
Fig. Millpede erase operation

16.4 Stored bits

Write/read operations depend on a mechanical parallel x/y scanning of either the entire cantilever array chip or the storage medium. The tip-medium contact is maintained and controlled as a whole, i.e., not on an individual cantilever basis, by using a common control for the entire chip, which greatly simplifies the system.

Experimental results indicate that overall chip tip-apex height control to within 500 nm is feasible. As the Millipede tracks the entire array without individual cantilever positioning, thermal expansion of the array chip has to be small. For 3×3 mm2 silicon array area and 10-nm tip-position accuracy, the chip temperature has to be controlled to about 1ºC. This is ensured by four temperature sensors in the corner of the array and heater elements on each side of the array. Thermal-expansion is the reason why two-dimensional arrays are preferred over one dimensional ones which would utilise space on the storage medium more efficiently.

Fig. Comparison between (a) the read back signal obtained experimentally along a data track and (b) the read back signal obtained by the synthetic model.
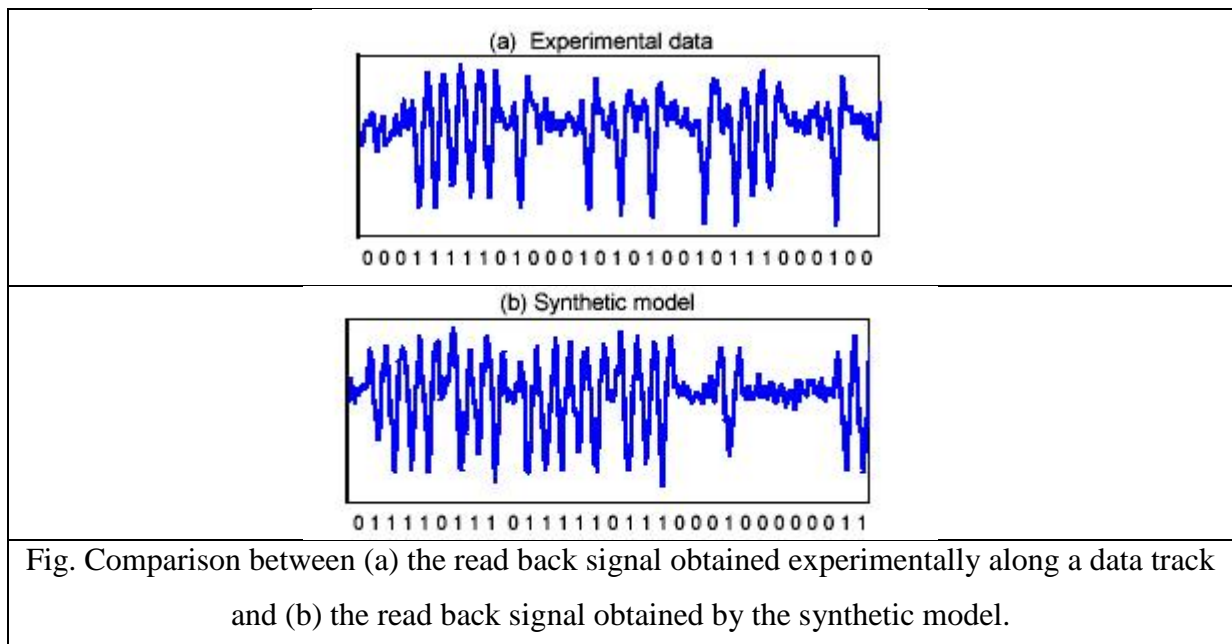
16.5 Cantilever Structure

Schematic representation of the cantilever–probe and tip–sample system is shown on Figure. The sample is a standard probe-storage polymer film of around 100 nm in thickness, spin-coated on a silicon substrate. The heated probe was used to form indents having a depth between 1 and 4 nm in the polymer by applying a negative, pulsed capacitive voltage to the substrate. The arrow and 'V' indicate the direction of motion of the sample during writing.
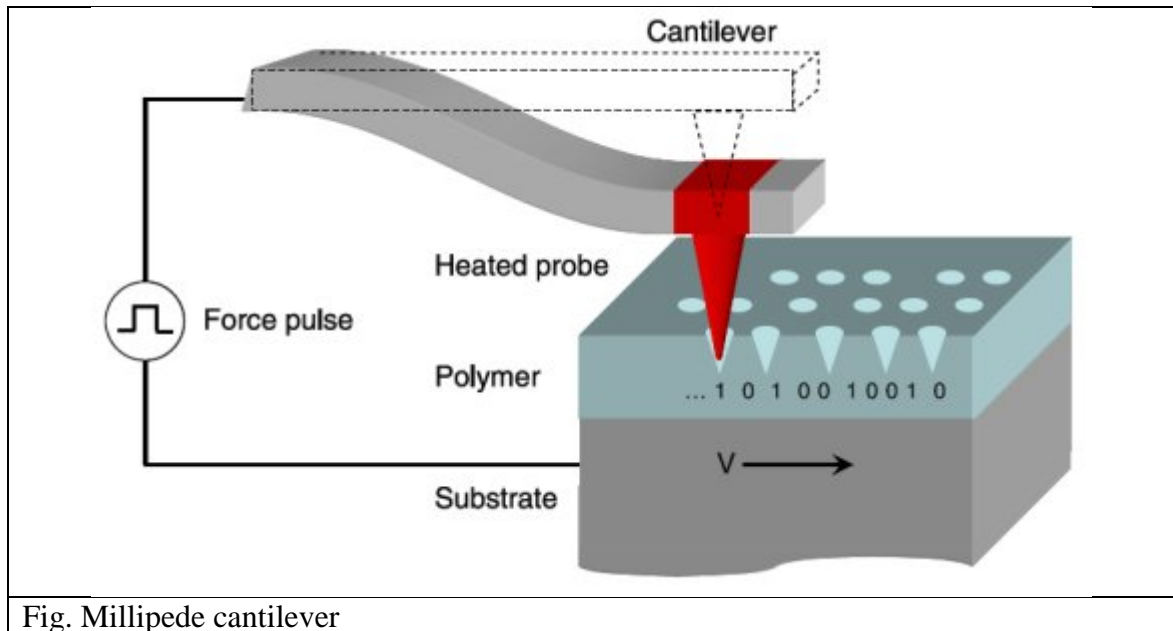
Fig. Millipede cantilever

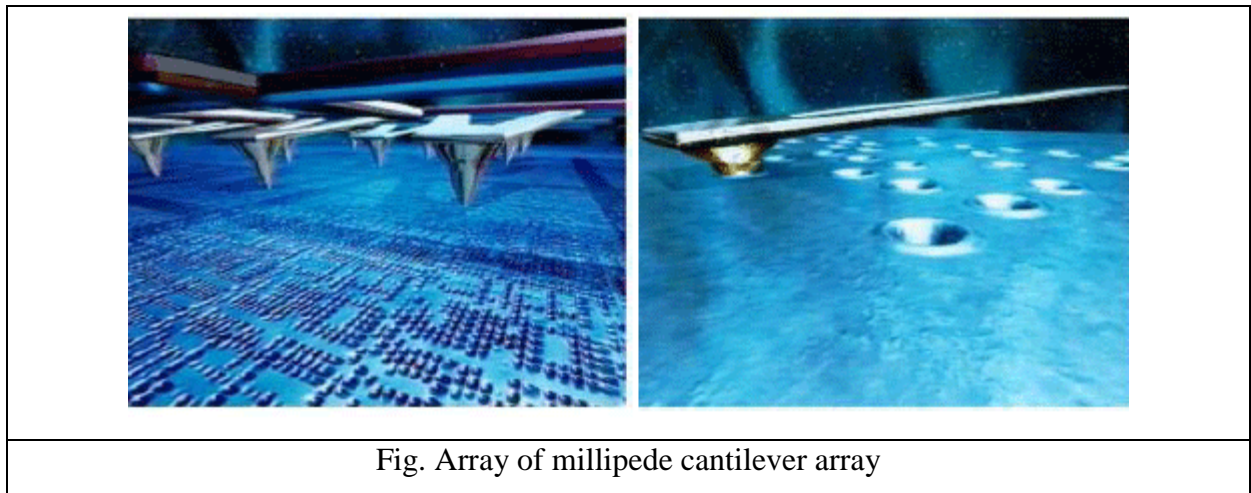Array of millipede cantilever array is shown on Fig.

Fig. Array of millipede cantilever array

16.6 Parameters

First generation millipede: as a first step, a (5 x5) array chip was designed and fabricated to test the basic Millipede concept. All 25 cantilevers had integrated tip heating for thermo mechanical writing and piezoresistive deflection sensing for read-back. Each cantilever was individually addressable for both thermo mechanical writing and piezoresistive deflection sensing. A complete resistive bridge for integrated detection has also been incorporated for each cantilever. Unfortunately, the chip was not able to demonstrate parallel writing because of electro migration problems due to temperature and current density in the Al wiring of the heater.

Once global leveling was proven as a reliable concept and (Al) wiring was avoided, the millipede was a realizable device. Using the results of the (5×5) cantilever array, a (32×32) array chip was made with some important changes in the chip functionality and fabrication processes.

The major differences are:

1. Surface micromachining to form cantilevers at the wafer surface,

2. All-silicon cantilevers,

3. Thermal instead of piezoresistive sensing,

4. First- and second-level wiring with an insulating layer for a multiplexed row/column addressing scheme.

Since the heater platform was both a write/read element and no individual cantilever actuation was required, the basic array cantilever cell became a simple device.

Technical specifications:

The present day specifications for a (32×32) Millipede device are.

1. The cell area and x/y cantilever pitch is 92 µm x 92 µm, which results in a total array size of less than 3 mm x 3 mm for the 1024 cantilevers.

2. The cantilever is fabricated entirely of silicon for good thermal and mechanical stability. It consists of the heater platform with the tip on top, the legs acting as a soft mechanical spring, and an electrical connection to the heater. They are highly doped to minimize interconnection resistance and replace the metal wiring on the cantilever to eliminate electro migration.

3. The resistive ratio between the heater and the silicon interconnection sections should be as high as possible; currently the highly doped interconnections are 400 Ω and the heater platform is 11 k Ω (at 4 V reading bias).

4. The cantilever must be soft (low mass) to avoid high loading forces and to achieve high resonant frequencies. This leads to less wear and tear and higher speed of operation.

5. Keeping the above considerations in mind, the array cantilever is designed to be 50-µm-long, 10-µm-wide, 0.5-µm-thick at the legs, and 5-µm-wide, 10-µmlong, 0.5-µm-thick at the platform platform. This cantilever has a stiffness of 1 N/m and a resonant frequency of 200 kHz.

6. The heater time constant is a few microseconds, which should allow a multiplexing rate of 100 kHz.

7. A large gap is required between the tip and the medium, but the tips must be sort to improve sensitivity, we purposely bent the cantilevers a few micrometers out of the chip plane by depositing a silicon nitride layer at the base of the cantilever.

8. The cantilevers within the array are electrically isolated from one another by integrated Schottky diodes. Because every parasitic path in the array to the addressed cantilever of interest contains a reverse-biased diode, the crosstalk current is drastically reduced.

9.   The tip-apex height uniformity within an array is to be maintained because it determines the force of each cantilever while in contact with the medium and hence influences write/read performance as well as medium   and tip wear.

Strengths:

1.   Very high storage density of about 500Gbit/in^2 is possible.
2.   High data rates are achieved through massive parallelism.
3.   Control loop structure for optimized 2D motion has been established.
4.   Track following with nm precision is achieved.
5.   It consumes very low power compared to other devices.
6.   It has reliable operation with small overhead.

Challenges:

1.   The medium must endure long term usage.
2.   It must endure long term tip & media wear.
3.   Tip uniformity in lever array chip must be maintained.\
4.   It must be shock resistant and immune to vibrations.
5.   Cantilever must be designed for low power reading & writing/
6.   The area density limits must be exploited to the maximum permissible safe limit.
7.   The signals must be processed and detected accurately for proper operation.

16.7 Application

The earliest generation millipede devices used probes 10 nanometers in diameter and 70 nanometers in length, producing pits about 40 nm in diameter on fields 92 µm x 92 µm. Arranged in a 32 x 32 grid, the resulting 3 mm x 3 mm chip stores 500 megabits of data or 62.5 MB, resulting in an areal density, the number of bits per square inch, on the order of 200 Gbit/in². IBM initially demonstrated this device in 2003, planning to introduce it commercially in 2005. By that point hard drives were approaching 150 Gbit/in², and have since surpassed it.

Devices demonstrated at CeBIT in 2005 have improved on the basic design, using a 64 x 64 cantilever chips with a 7 mm x 7 mm data sled, boosting the data storage capacity to 800 Gbit/in² using smaller pits. It appears the pit size can scale to about 10 nm, resulting in a theoretical areal density just over 1Tbit/in². IBM planned to introduce devices based on this sort of density in 2007. For comparison, as of late 2011, laptop hard drives are shipping with a density of 636 Gbit/in² and it is expected that heat-assisted magnetic recording and patterned media together could support densities of 10 Tbit/in².   Flash reached almost 250 Gbit/in² in early 2010.

There are many applications for the Millipede chip in storage devices. The concept and design vary according to the required application. Two cases in particular are discussed below.

16.7.1 Terabit drive

The potential for very high storage density makes the Millipede very attractive for high-end terabit storage applications also. Terabit capacity can be achieved in three methods:

1. Very large arrays,
2. Many smaller arrays operating in parallel and
3. Displacement of small/medium-sized arrays over large disk.

Though realization of considerably larger arrays (10^5 to 10^6 cantilevers) appears to be possible, control of the thermal linear expansion will be a delimiting factor as the array chip becomes significantly larger and control problems will definitely arise.

The second approach is appealing because the storage system can be upgraded to fulfill our requirements in a modular fashion by operating many smaller Millipede units in parallel.

The operation of the third approach is described with the example of a modified hard disk. This approach has the advantage of integration of new technology with our existing systems. A number of small arrays on the disk which can be accessed in a way similar to that of the present hard drive enhance the storage capacity greatly while keeping the basic principle unchanged. A storage capacity of several terabits appears to be achievable on 2.5- and 3.5-in. disks (Fig)
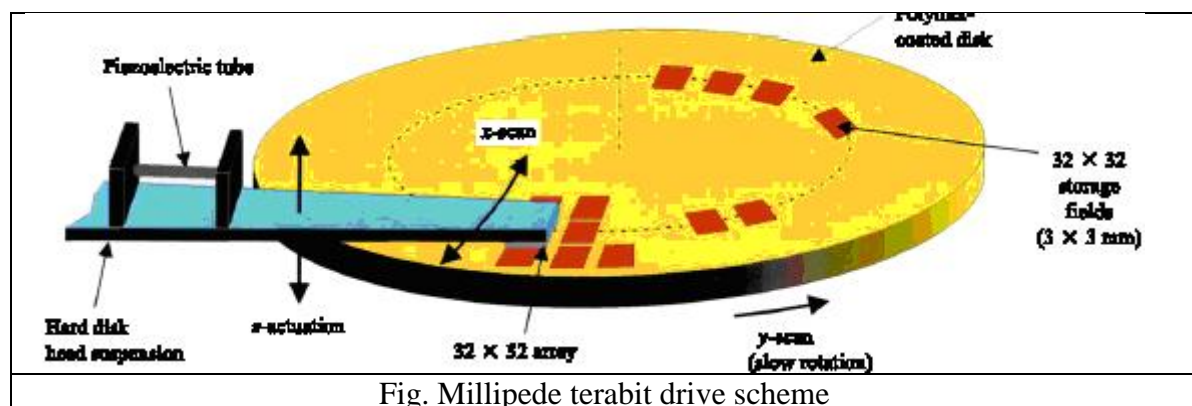
Fig. Millipede terabit drive scheme

16.7.2 Nanodrive

Using the concept of Millipede, storage devices with gigabyte capacity having a very small form factor in the range of centimeters and even millimeters can be made possible. Hence this concept opens up possibilities to integrate such "Nanodrives" into low power consuming devices such as watches, cellular telephones, laptops, etc.

Millipede based nanodrives have interesting prospects for audio and video consumer applications. All-silicon, batch fabrication, low-cost polymer media, and low power consumption make Millipede very attractive as a centimeter- or even millimeter-sized gigabyte storage system.

HYDROGEN GENERATION AND STORAGE USING NANOCOMPOSITES

1. Introduction

One of the most promising concepts of modern alternative energy is the use of hydrogen as an intermediate energy carrier (fig. 1). The oxidation of H_2 is accompanied by release of considerable energy (120.6 MJ/kg or 3 kW·h/m^3) and the formation of water, which determines the efficiency and cleanliness of this approach. Hydrogen reservoirs are the promising way of energy storage, along with mechanical, chemical and other accumulators.
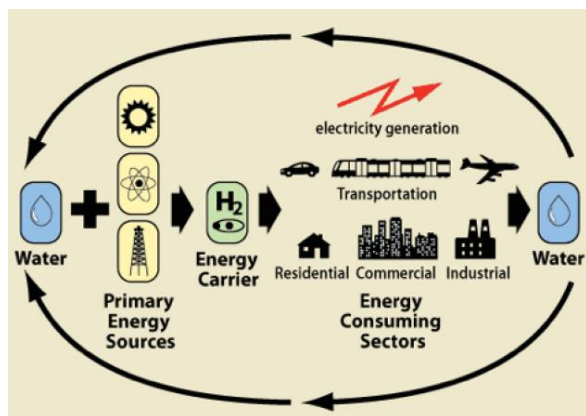

Fig. 1.

Hydrogen energy concept

The essential task of this field is investigation of safe solid-state hydrogen reservoirs. It is known that silicon could be used for hydrogen storage by way of etching in HF solutions, whereupon the porous silicon (PS) is produced. Such way its surface becomes completely hydrogenated. The preparation technology, as well as properties of the used Si wafers, appoints morphology and chemical composition of the gained PS. As was indicated, hydrogen covers dangling bonds of Si during the etching process. Thus silane SiH_x (x = 1, 2, 3) groups are formed on surface of the obtained PS.

Every type of hydrogen bonding in PS has its characteristic energies of activation and dissociation, vibrational modes, etc., and could be analyzed (Fig. 2). One of the most useful methods used to determine the hydrogen content and the type of its bonding is Fourier-transform infrared spectroscopy (FTIR).
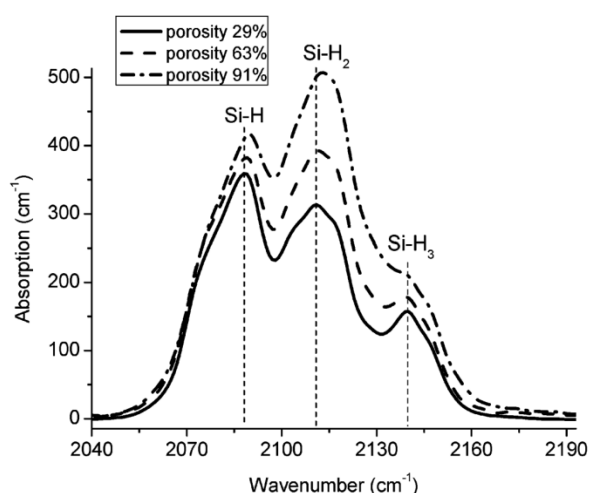
Fig. 2.
FTIR spectra for PS with different porosities. The peaks corresponding to every type of hydrogen bonding are indicated.

Theoretical maximum of the hydrogen content for the silane bonds is 3.44 mass% for SiH, 6.66 mass% for SiH_2 and 9.67 mass% for SiH_3. Maximal level of the hydrogen accumulation, which was experimentally obtained in this material, is near 60 mmol(H) per g(PS) or 6 mass% that corresponds to the H/Si ratio ~ 1.8.

The most widely used methods for producing H_2 from PS are heating, photo-desorption and chemical reaction with water. First two methods need high vacuum and are not good for practical applications. The method based on reaction with water is the easiest and the most productive, which generates up to 2.24 liters H_2 per gram of porous silicon. The overall reaction in generalized form can be presented as:

$$SiH_x + (2+y)H_2O = SiO_2 \cdot yH_2O + (2+x/2)H_2 \qquad (1)$$
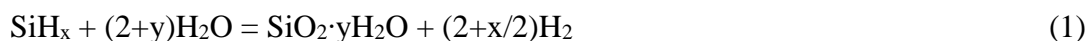
During this reaction the hydrogen is produced as from the breaked SiH_x species, so from the water interaction with Si-Si bonds. For the proper hydrogen effusion rates, the alkali catalyst should be used. The volume of evolved gas and kinetics of hydrogen production characterize hydrogen storage in PS nanocomposites. Similar methods are used to produce H_2 from water decomposition using crystalline Si and nano-Si powders, as well as from aluminum and different metals.

Conversion of hydrogen to electrical energy is made by fuel cells (FC). The power conversion is based on two reactions:

$$\text{anode: } H_2 \rightarrow 2H^+ + 2e^- \qquad (2)$$
$$\text{cathode: } 2H^+ + 1/2O_2\ 2e^- \rightarrow H_2O \qquad (3)$$
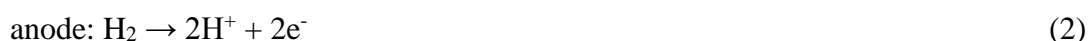
Fuel cells cover a wide range of applications for transportation and power generation to small and medium scale.

2. Experimental

2.1. FTIR.

FTIR transmission spectra are measured with a Perkin Elmer Spectrum BXII FTIR spectrometer in the spectral range of 400-7000 cm^{-1}. The samples can be studied in films as well as in pellets. The pellets are prepared through pressing of the sample powder with KBr (mass ratio 1:25, chosen for optimal spectra measuring). The device is connected to personal computer with spectra analytics software.


Fig. 3.

Perkin Elmer Spectrum BXII FTIR spectrometer

2.2. Hydrogen volume measurement.

The volume of evolved H_2 is measured by displacement of water in an inverted measuring cylinder. Systematic errors due to excessive water-column pressure, pressure of water vapors, deviations from ambient temperature and hydrogen solubility in water are neglected, resulting in measurements with a maximum error of about 10%.

Fig. 4. Hydrogen volume measurements.
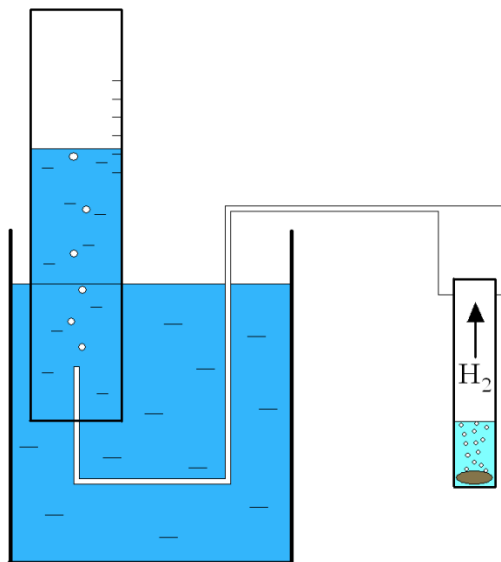
2.3. Hydrogen electricity producing device.

Application of H_2 produced from the reaction for generation of electrical energy is made with the fuel-cell based device. It consists of reservoir for mixing of the reactants, polymer electrolyte FC and output stabilizer. Molecular H_2, produced in reservoir by reaction (1), is oxidized in FC with transforming of obtained energy in electromotive force. In presence of output load, the current exists, with value corresponding to intensity of the reaction. Current and voltage measurements on the load are provided for calculation of produced power.
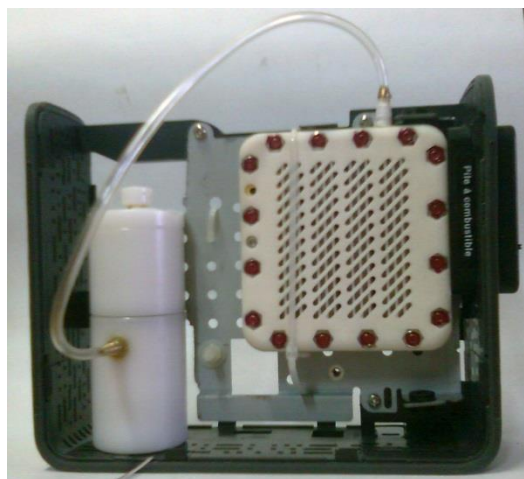
Fig. 5. Hydrogen electricity producing device.
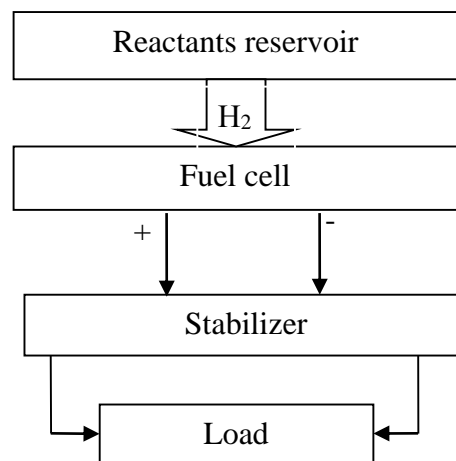
3. Work scheme

3.1. FTIR spectra measurement and analysis.

1) *Preparation of sample pellets.* The sample powder of PS is mixed with powder of KBr (1:25 by mass). The obtained composite is poured into hydraulic press. The pellet is made under one-half pressure during 1 minute.

2) *Transmission spectra measurement.* FTIR spectrometer (2.1) is switched on and performs its boot self test. Than the spectrometer software is opened on the computer. First, the transmission spectrum of the empty spectrometer is measured. The spectral range of 400-3000 cm^{-1} is used. Than the pellet is placed in the sample holder of the spectrometer. The spectrum of the sample is measured as ratio of the obtained spectrum and the empty holder spectrum. The measured data is saved in files.

3) *Spectra analysis.* The absorption peaks corresponding to SiH$_x$ bonds are found. The main spectral bands of different PS composites are listed below:

Si-H$_x$ stretching bands	2080-2140 cm^{-1}
Si-H$_x$ bending bands	627-667 cm^{-1}
SiH$_2$ scissors	906-909 cm^{-1}
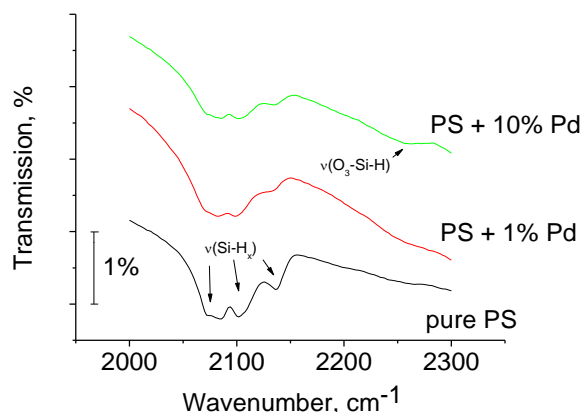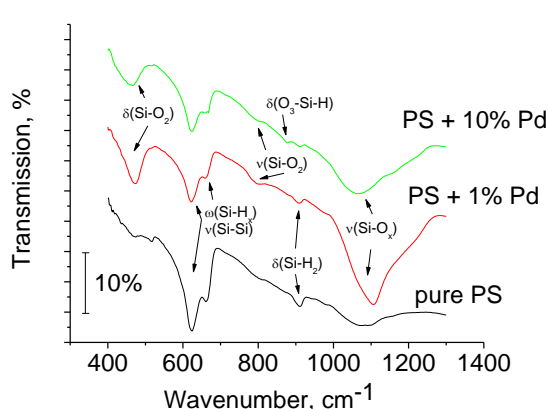
3.2. Hydrogen generation during the reaction of porous silicon with water.

1) *Preparation of reactants*. The first reactant is the sample PS powder. There could be different types of powders, as well as nanocomposites. The second reactant is the mixture of H$_2$O, C$_2$H$_5$OH and NH$_3$(10%), in 5:1:1 volume ratio. Typically, the volume of solution used is 0.2 ml per 1 mg of the PS. Here NH$_3$ is used as catalyst; ethanol allows the solution to wet the hydrophobic PS sample.

2) *Hydrogen generation*. The reactants are mixed together in the vessel, which is attached to the water displacement system (2.2). The reaction (1) is initiated. The volume of produced H$_2$ is measured as a function of time.

3) *Hydrogen volume analysis*. The general volume of the produced hydrogen and the time dependence of the reaction intensity characterize the hydrogen content and the potential power of the sample powder.

3.3. Use of the hydrogen electricity generating device.

1) *Preparation of reactants.* The same reactants as in 3.2 are used. The reactants are placed in the reservoir of the hydrogen electricity generating device. The powder is poured in the bottom part of the reservoir, while the water solution is placed in the top cell.

2) *Initiation of hydrogen generation.* The reaction (1) is started, when the top and the bottom cell of the reactant reservoir are connected by opening the valve. The produced H_2 is supplied to fuel cell that results in generation of electrical power.

3) *Electrical power generation.* Before initiation of the hydrogen generation the electrical load (resistor) should be connected to the output of the device, as well as multimeters for current and voltage measurements. The time dependences of output voltage and current make possible to obtain electrical power characteristics of the used hydrogen generation scheme.

4. Safety

The used materials are safe for human and environment. However 2 safety issues should be kept in mind.

1) *Hydrogen leakage.* Molecular hydrogen is volatile and flammable gas. The amount of H_2 generated during the measurements is too small to be dangerous. However, it should be no sources of open fire near the hydrogen generation equipment.

2) *Alkaline catalyst.* The used alkaline water solution is not corrosive and toxic. However, avoid contacts of this solution with eyes and respiratory system.

5. Sample task

The porous silicon powders of different types are given. The tasks are (a) to analyze hydrogen bonding in the samples; (b) to measure the volume of hydrogen produced by reaction (1); (c) to examine the electrical power generated using the hydrogen electricity generating device.

6. Literature

1) L.M. Gandía, G. Arzamendi, P.M. Diéguez, *Renewable Hydrogen Technologies: Production, Purification, Storage, Applications and Safety* (Elsevier, 2013).

2) U. Eberle, M. Felderhoff, F. Schuth, Chemical and Physical Solutions for Hydrogen Storage, Angew. Chem. Int. Ed. 2009, 48, 6608 – 6630.

3) A.I. Manilov, V.A. Skryshevsky, Hydrogen in porous silicon — A review, Materials Science and Engineering B 178 (2013) 942–955.

4) Lysenko V., Bidault F., Alekseev S., Zaitsev V., Barbier D., Turpin Ch., Geobaldo F., Rivolo P., Garrone E. Study of Porous Silicon Nanostructures as Hydrogen Reservoirs // J. Phys. Chem. B. – 2005. – Vol. 109. – P. 19711-19718.

5) V.P. Tolstoy, I.V. Chernyshova, V.A. Skryshevsky, *Handbook of Infrared Spectroscopy of Ultrathin Films*, John Wiley & Sons, Inc., N.J. (2003).

Lab work 4-2
LASER SCANNING METHOD

Laser scanning facility is a complex of modules and devices designed to illuminate the investigated objects by laser scanning spot, to observe the corresponding images, and to save them in digital form. It is possible to obtain both conventional optical images in reflection or transmittance mode and photovoltaic images due to induced photocurrent (Light Beam Induced Current, LBIC). It could be a useful tool for investigation of photovoltaic and photosensitive devices such as solar cells, chemical sensors with photovoltaic transducer.

View of Laser scanning facility is shown in Figure 1.

Fig. 1 Photo of laser scanning facility

The facility consists of Laser scanning module, Control module, Remote control and interface for user, Holder with investigated PV device, External circuit for the investigated PV device, Selective nanovoltmeter to receive and amplify the obtained (induced) current, additional optic elements to adjust laser raster within the investigated object.

General scheme of the facility with interconnections is shown in Figure 2.
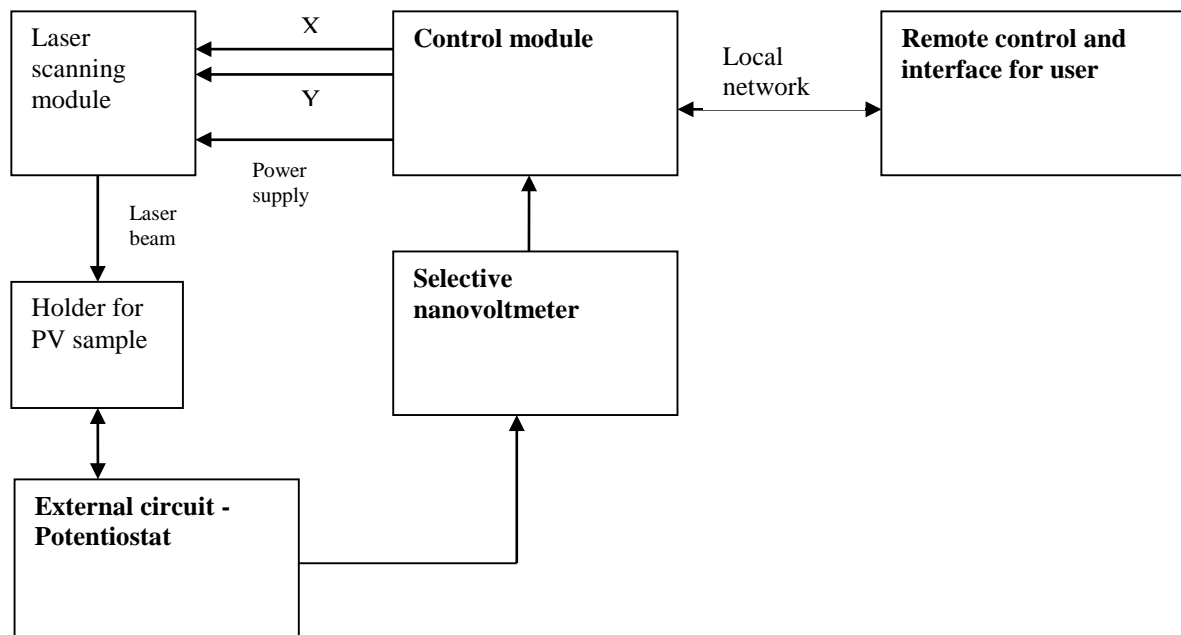
Fig. 2. General scheme of the facility with interconnections

Laser scanning module

Here we describe in few words the laser-scanning module with 2D acousto-optic deflector. It allows simultaneously scanning and modulating the laser beam to realize a technique using lock-in amplifier. Acousto-optic deflector is supplied with the two-channel computer controlled frequency synthesizer, which use two Analog Devices double deflection systems. The special software to control DDS system and to process LBIC signals is designed.

The optical scheme of the device is shown in Fig.3. It is rather simple due to the special shape of the acousto-optic cells, and because of this, the diffracted beam on the central frequency was collinear to the incident one. The front focus of the lens coincides to the virtual center of the scanning; therefore, the focused laser beam is normal to the surface of the specimen.

The control device uses 32-bit word for precision control of the laser beam position.
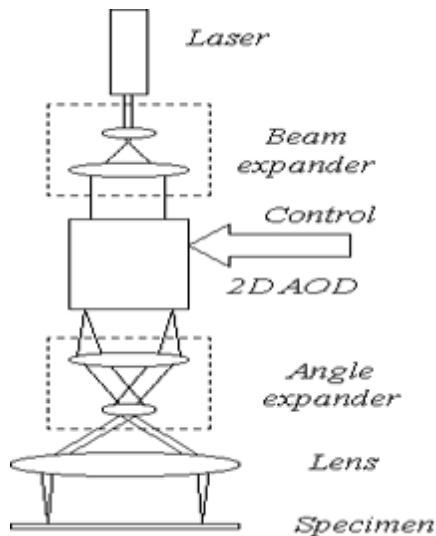
Fig. 3 Optical scheme of laser scanning module

The construction of the module and the service software allows controlling the various conditions of the experiment. There are the possibilities to use the module as a common optical microscope with additional properties allowing to storage and operate images, and in the mode of the laser beam induced current fruitful in the experiments with photosensitive devices. It is possible by means of the software to modulate the intensity of laser beam, to vary the scanning velocity, save the data in the image or ASCII standards. The size of the scanning area is controlled both by software and hardware possibilities and it changes in the wide range from 0.5*0.5 mm to several meters. The absence of mechanically moved parts makes it more reliable and multipurpose.

Wavelength of laser light is 532 nm, laser power is 10 mW, but it significantly reduced to 1-2 mW at the investigated surface after passing deflectors and optical elements.

Manual

At the beginning, it is necessary to check all the interconnections. Then, after switching-on the power supply, it is time to launch the Micro Terminal software. Interface of this application is shown on Fig.4.

Fig. 4. User interface of Micro Terminal software

When the local network connection is automatically established (Status "Connected" displayed in the left column), it is possible to set the scanning area on the investigated surface, and adjust the focus by optical elements. To do this, a button "FF" (left column) and "Start" ► (right top corner) should be clicked. A figure of square appears on the investigated surface, this figure determines the scanning area.

It is possible to choose the resolution of image among 512*512, 256*256, 128*128 and 64*64 pixels by clicking corresponding button.

Scanning velocity depends on the chosen resolution, time delay (this scale is in left column, HW parameters area) and a number of averaging (left column). Scanning velocity should be a compromise regarding the image quality and time need to obtain the image. For example, at 256*256 resolution, 1300 Hz frequency, reasonable duration of one frame scanning is about 3-4 minutes.

In HW parameters area, there is a scale allowing to tune the laser power from 0 to 100%. While tuning and operation process, protect your eyes from laser irradiation!

To illuminate with modulated light, switch on modulation and choose the frequency (left column, bottom).

To start scanning, switch off "FF" button and click "Start" ► button.

Some other buttons mean the next. "SR" shows the profile (graph) of current line. "FR" button sets a mode of only chosen line scanning. "FP" results in fixed illumination of selected point.

The obtained data can be saved in proposed formats of images or/and tables suitable for further treatment.

Additionally, a serial image mode can be launched by setting a number of scans and interval. In this mode, each filename of the series automatically gets precise time of measurement.

Task

1) Obtain 2D images of light beam induced current at scanning of Si photodetector

2) Study the influence of scanning parameters (resolution, frequency, duration of one frame scanning) on 2D images

3) Explain the operation principle of the method of light beam induced current

Lab work 4-3
The measurement of spectral response of semiconductor structures

1. Introduction

There is standard set up for photovoltaic measurement. Two possibilities can be regarded. First possibility is measurement of short circuit current $i(\lambda)$ (or open circuit voltage) dependences on wave length λ in barrier structures with inserted inner potential barrier (n-p junction, heterojunction etc). In such case, signal from investigated sample is applied to the nano voltmeter directly. Second possibility is measurement of photoconductivity. In such case electrical circuit includes illuminated sample, loading resistance, source of voltage and the signal from loading resistance is used for further processing. Our standard equipment includes five principal parts: light source, spectrometer МДР 206 with variables diffraction gratings, selective nano voltmeter, optical modulator and the power supply of the light source (Fig.1).
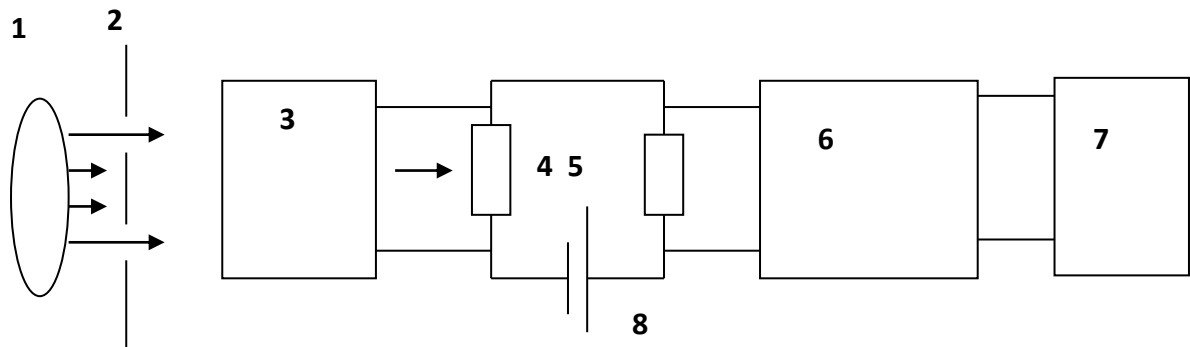
1-light source, 2- modulator,3- spectrometer МДР 206

4-investigated sample,

5-loading resistance (for photoconductivity measurement),

6- selective nano voltmeter,

7- analog to digital converter (ADC),

Fig.1 Simplified structural scheme of equipment.

2. Operation principle

The grating spectrometer МДР 206 covers spectral range 200-1000 nm (that is necessary for investigation of silicon structures). The spectrometer is illuminated by Xenon lamp (power consumption is 90 Wt) that is mainly used in ultraviolet and visible part of spectrum, Fig 2. Mechanical modulator is used. The data acquisition is done using selective nano voltmeter tuned to

the modulation frequency (typically 450 Hz in our experiment). This allows separating of chopped monochromatic signal of investigated sample from constant bias light influence. The analog to digital converter (ADC), connected with nano voltmeter output, allow to process digital signal with special computer software. It is necessary to keep investigated sample in short circuit condition if spectral response measurement takes place. The digital voltmeter GDM-8246 is used as ADC in our scheme. The scan of wavelength is controlled by computer software. As it well known, the spectral response gives the number of electrons produced in investigated sample by one incoming photon . So photo detector or solar cell with previously known spectral sensitivity should be used to calculate spectral response. For this aim the calibrating distribution of photon $N(\lambda)$ is also needed. The ratio $i(\lambda)/N(\lambda)$ will represents required spectral response.

Fig.2 Photo of experimental equipment.

To start the measurement of spectral response the program "MonoLomo 2" with very simple interface should be loaded from computer (see corresponding symbol on Рабочий стол). The buttons "Подключится к монохроматору", "Перемотать на ноль", "Перемотать на длину волны" should be successively pressed. Resulting data is stored in the file that can be further processed with standard Origin or Exel software. Fig.3 represents typical curves of silicon junction photocurrent as function of wavelength received with our equipment.
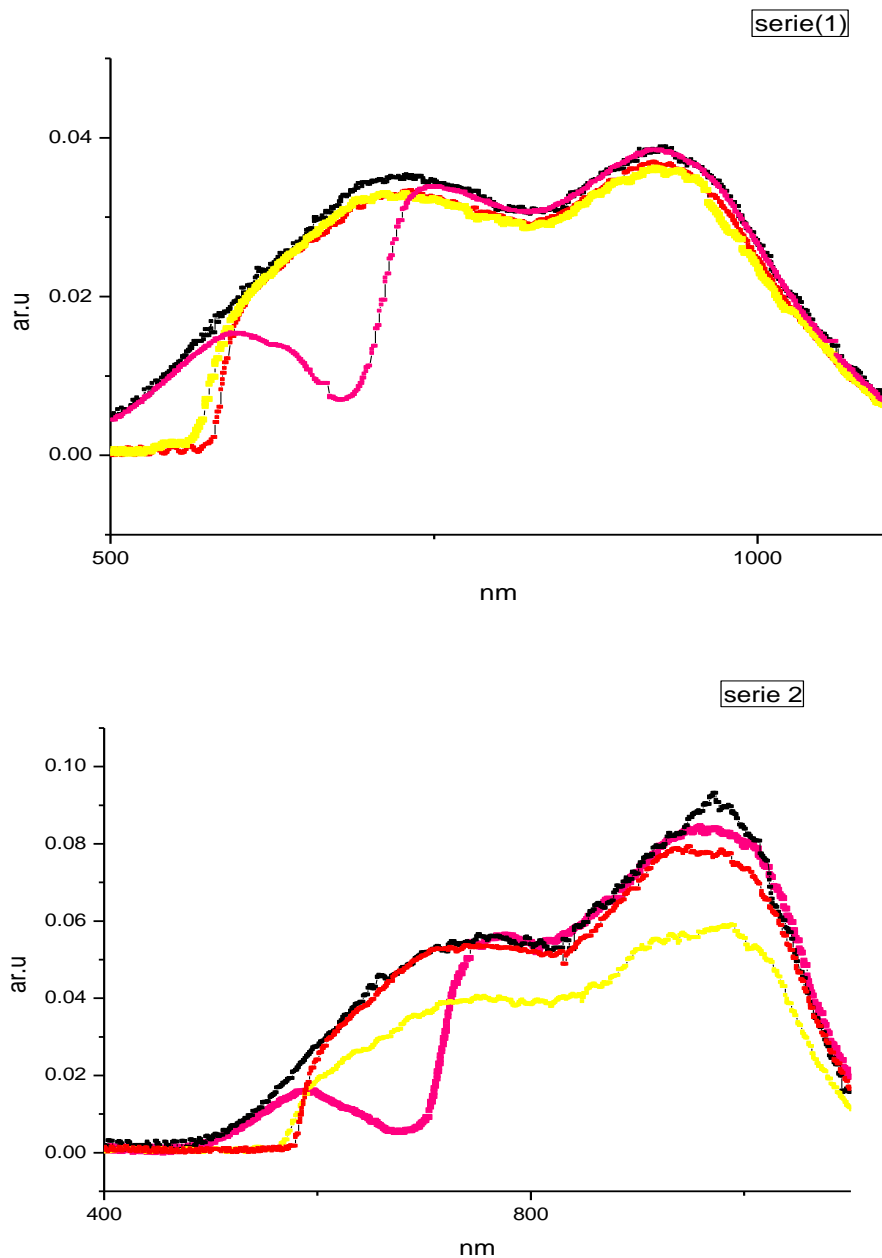
Fig.3 Measured spectral distribution of photocurrent for silicon n-p junction with thin polymeric films at the front surface

Work with source of high voltage (such as power supply ЛИПС-35 for Xenon lamp) should be performed with great caution. The operating current of this lamp is nearly 7-8 A. If transparent "ПЕРЕГРУЗКА" is lighted, the device should be turn-off immediately. After that, work can be continued.

Task

1) The measurement of short circuit current $i(\lambda)$ dependences on wave length λ for tested p-n junction.
2) Calculation of external quantum efficiency of solar cell
3) The measurement of photoconductivity dependences on wave length λ for semiconductor sample.

3. References

1) S.M. Sze, Physics of semiconductor devices, Wiley, New York (1969)
2) K. Bücher, A. Schönecker Spectral Response Measurements of Multi-Junction Solar Cells with a Grating Monochromator and a Fourier Spectrometer, In Proceeding 10 Photovoltaic conference ,Portugal 1991, pp 107-110.

Investigation of dynamic reflection spectra (kinetics) of multilayer structures (based on CCD receiver and fiber optic technology)

Purpose of work: To investigate stationary and kinetic reflection spectra response of single-layer and multilayer sensor structures on analyte deposition.

Equipment: CCD-spectrometer Thorlab **CCS200**, Fiber Optic Reflection Probe Bundle **RP20**, Diafragm, focusing lens in a frame, set of optical components, optical board, cage mounting system of optical elements, High-Intensity Fiber-Coupled Illuminator **OSL2**, pipette dispenser, set of analytes.

Software: Thorlabs OSA

Work progress:

1. Assemble the measuring reflectance spectra setup due the scheme

    1.1 Bind Vertical Mounting Plate for Cage Systems **UBP2/M** to optical board using 4 screws M6 (Fig. 2a-b).

Fig. 2a. Vertical Mounting Plate **UBP2/M.**	Fig. 2b. ER rods	Fig. 2c. Vertical Mounting Plate **UBP2/M** assembled with ER rods fixed to optical table	Fig. 2d. SM1-Threaded 30 mm Cage Plate **CP02T/M**

    1.2 Insert ER rods (Fig. 2b) into the mounting holes of 30 mm Vertical Mounting Plate **UBP2/M** (Fig. 2a). Fix each ER rod using two 4-40 setscrews (Fig. 2c). Place SM1-Threaded 30 mm Cage Plate **CP02T/M** (Fig. 2d.) on the top of Vertical Mounting Plate **UBP2/M** using ER rods as guides.

    1.3 Attach Plano-Convex Lens **LA1951-ML - Ø1"** with SM1-Threaded Mount (Fig. 3b) to Adjustable Lens Tube **SM1V10 - Ø1"** (Fig. 3a).

    1.4 Attach Graduated Ring-Actuated SM1 Iris Diaphragm **SM1D12C** (Fig. 3c) to Mount of Plano-Convex Lens **LA1951** (Fig. 3b).

Fig. 3a. FSM1V10 - Ø1" Adjustable Lens Tube, 0.81" Travel Range	Fig. 3b. N-BK7 Plano-Convex Lens, Ø1"	Fig. 3c. Graduated Ring-Actuated SM1 Iris Diaphragm SM1D12C.

1.5 Attach Adjustable Lens Tube (Fig. 3a) with attached lens and diaphragm to the bottom of the 30 mm-Cage-Compatible SM1-Threaded Kinematic Mount with Slip Plate **KC1-S/M** (Fig. 4a).

Fig. 4a. SM1-Threaded Kinematic Mount with Slip Plate KC1-S/M	Fig. 4b. SMA Fiber Adapter Plate **SM1SMA** with External SM1

1.6 Attach SMA Fiber Adapter Plate **SM1SMA** (Fig. 4b) with External SM1 to the top of Kinematic Mount **KC1-S/M** (Fig 4a)**.**

Fig. 5a. SM1 Series Alignment Disk **SM1A7**	Fig. 5b. 30 mm Cage System Alignment Plate **CPA1**

1.7 Swirl Alignment Disk **SM1A7** (Fig. 5a) into SM1-Threaded Mounting Hole of Cage Plate **CP02T/M** (Fig. 2d).

1.8 Connect Fiber Optic Reflection Probe Bundle **RP20** (Fig. 6a) to High-Intensity Fiber-Coupled Illuminator **OSL2** (Fig. 6b) according to Fiber Configurations (Fig. 7).

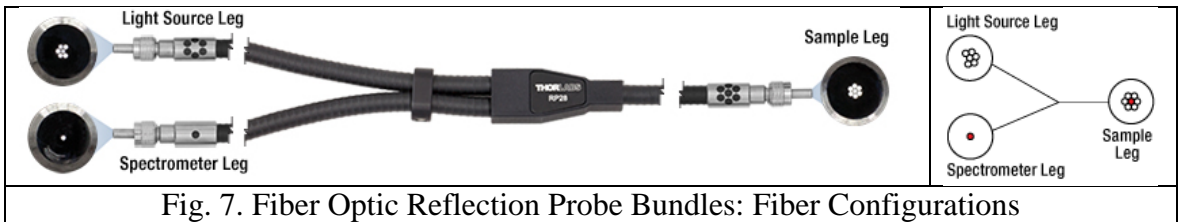|	
Fig. 6a. Fiber Optic Reflection Probe Bundle **RP20**	Fig. 6b. High-Intensity Fiber-Coupled Illuminator **OSL2**

Fig. 7. Fiber Optic Reflection Probe Bundles: Fiber Configurations

1.9 Connect Fiber Optic Reflection Probe Bundles **RP20** (Fig. 6a) to Spectrometer **CCS200/M** (Fig. 8) according to Fiber Configurations.

Fig. 8. Compact Spectrometer **CCS200/M**

1.10 Connect High-Intensity Fiber-Coupled Illuminator **OSL2 (Fig. 6b)** to the electrical mains.

1.11 Connect Spectrometer **CCS200/M** (Fig. 8) to a Personal computer using a USB cable

1.12 Compare assembled setup with the measurement setup on Fig. 9.

Fig. 9. Assembles measuring setup.

2. Calibration of the optical system.

    2.1 Place Cage System Alignment Plate **CPA1** (Fig. 5b) on Alignment Disk **SM1A7** (Fig. 5a) using as guides ER rods.

    2.2 Turn on High-Intensity Fiber-Coupled Illuminator **OSL2** (Fig. 6b). Using Variable Output Power Front Panel Knob get a clear picture of the light spot on Cage System Alignment Plate **CPA1** (Fig. 5b).

    2.3 Perform necessary Tip, Tilt, X-Y Linear Translation **KS1-M** (Fig. 4a, 9) setting to get light Spot aligned relatively to deposited concentric circles of Cage System Alignment Plate **CPA1** for different values of diameter of open Graduated Ring-Actuated Iris Diaphragm **SM1D12C** (Fig. 3c).

3. Process of measurement.

    3.1 Place the first sample on the surface of the Alignment Disk **SM1A7** (Fig. 5a).

3.2 Adjust aperture diameter of Graduated Ring-Actuated Iris Diaphragm **SM1D12C** to illuminate only the desired surface area - etched porous surface.

3.3 Start spectrometer program **Thorlabs OSA.**

    3.3.1 Click button **Scan USB** (Fig.10) on the tab **Sweep** and select **Spectrometer CCS200/M** from the list of the connected equipment to connect software **Thorlabs OSA** to Compact Spectrometer **CCS 200/M.**


Fig. 10
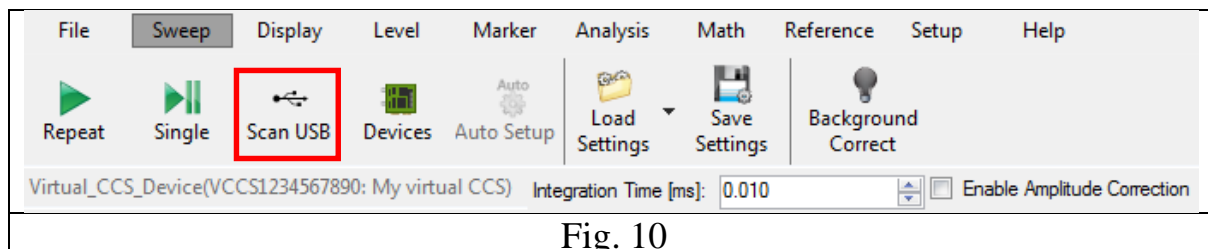
    3.3.2 Click button **Repeat** (Fig.11) on the tab **Sweep** to begin spectrum measurements.

Fig. 11
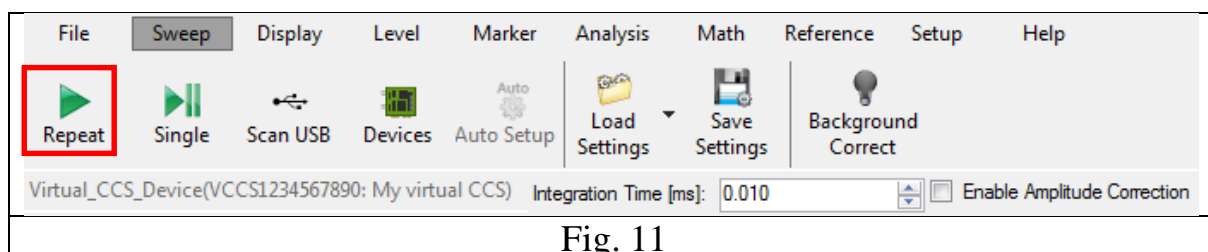
    3.3.3 Increasing Integration Time (Fig. 12) and increasing the light intensity of High-Intensity Fiber-Coupled Illuminator **OSL2** get a clear reflection spectrum of investigated sample. *Note:* with increasing light intensity avoid off-scale response of spectrometer!

Fig. 12
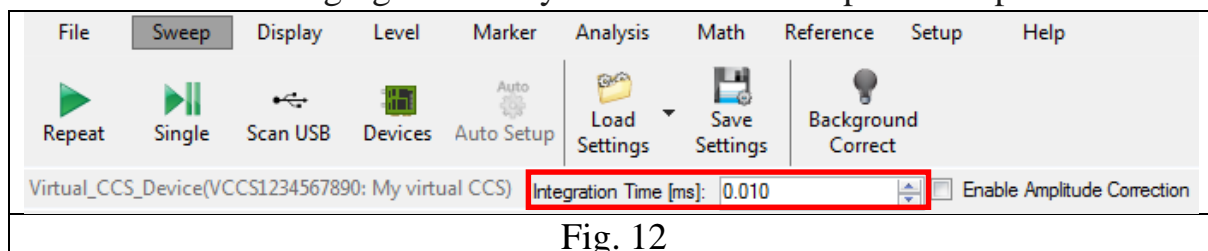
    3.3.4 Save the measured spectrum. To do this, click button tab **Stop** (Fig.13) on the tab **Sweep**.

Fig. 13
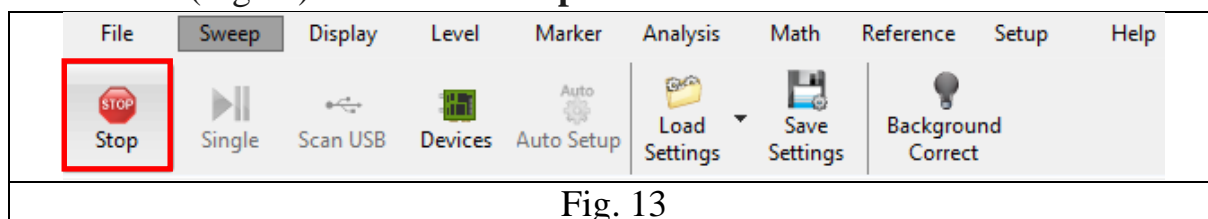
Press button **Save Trace** (Fig. 14) on the tab **File** tab and save the current measured spectrum with the desired file and file format. Saved spectrum is reflection spectrum of sample without analyte.
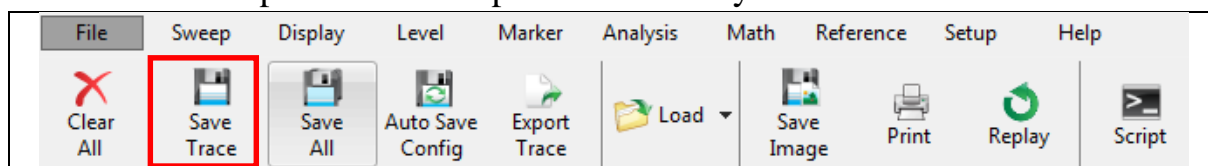

Fig. 14

3.3.5 To measure kinetics of reflection spectrum after deposition of analyte on the porous sample user should define parameters of measurement and saving.

3.3.6 Press button **Auto Save Config** (Fig.15) on the tab File.
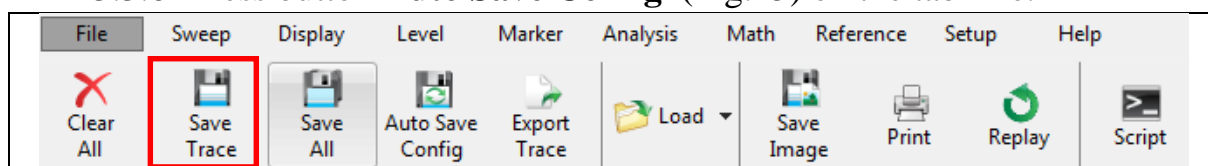

Fig. 15

3.3.7 Select **Device** and **Spectrum** at left part of open window (Fig. 16). Set the **Output Directory** on tab **File Settings**. Measured spectrums will be automatically saved in this folder.
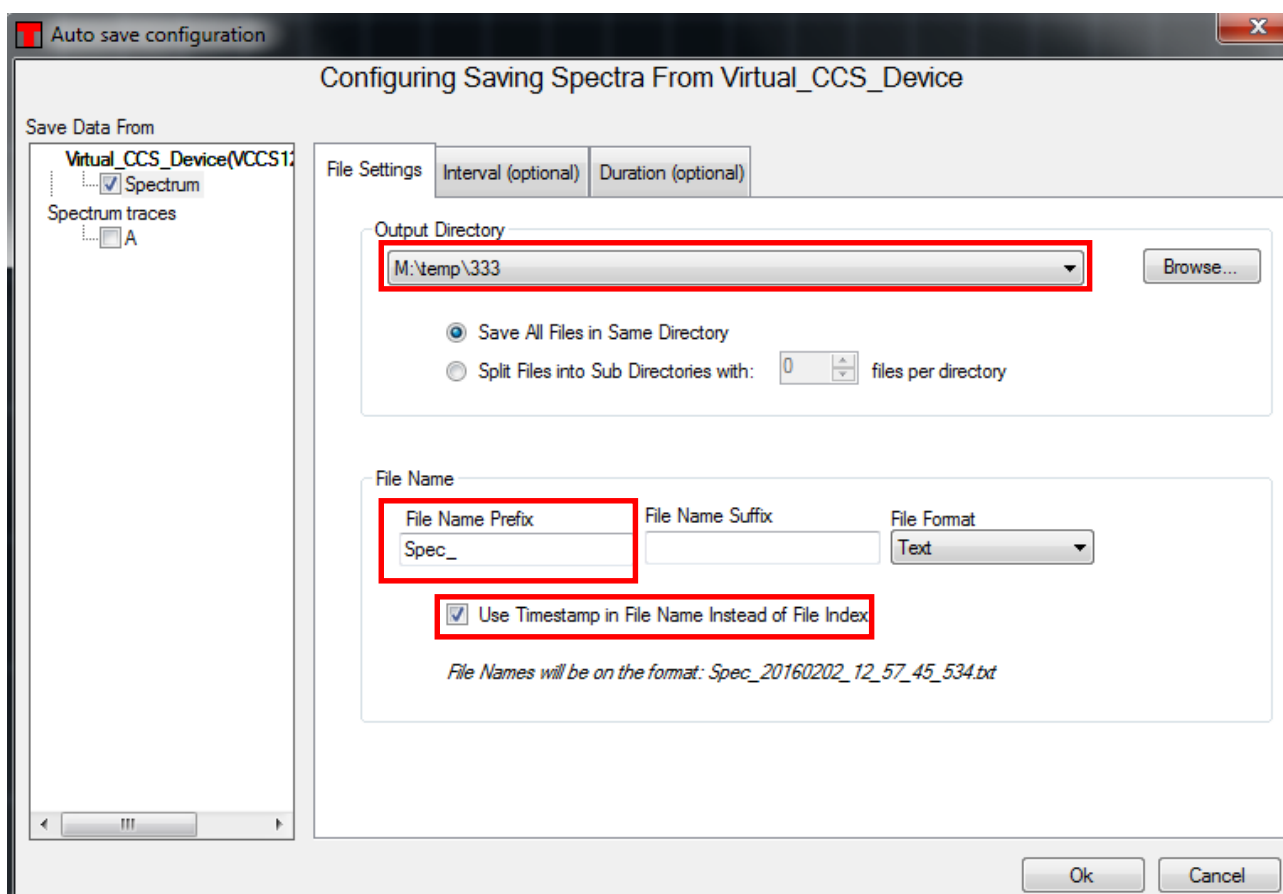

Fig. 16.

3.3.8 Write **File Name** prefix for saved files and switch on option **Use Timestamp in File Name Instead of File Index** (Fig. 16). In this case file name will have the time of spectrum measurement and kinetic spectrum can be plotted.

3.3.9 Select option **Save Every Spectrum/Interferogram** on tab **Interval** (optional) (Fig.17). In this case spectrums will be saved as fast as hardware possible.
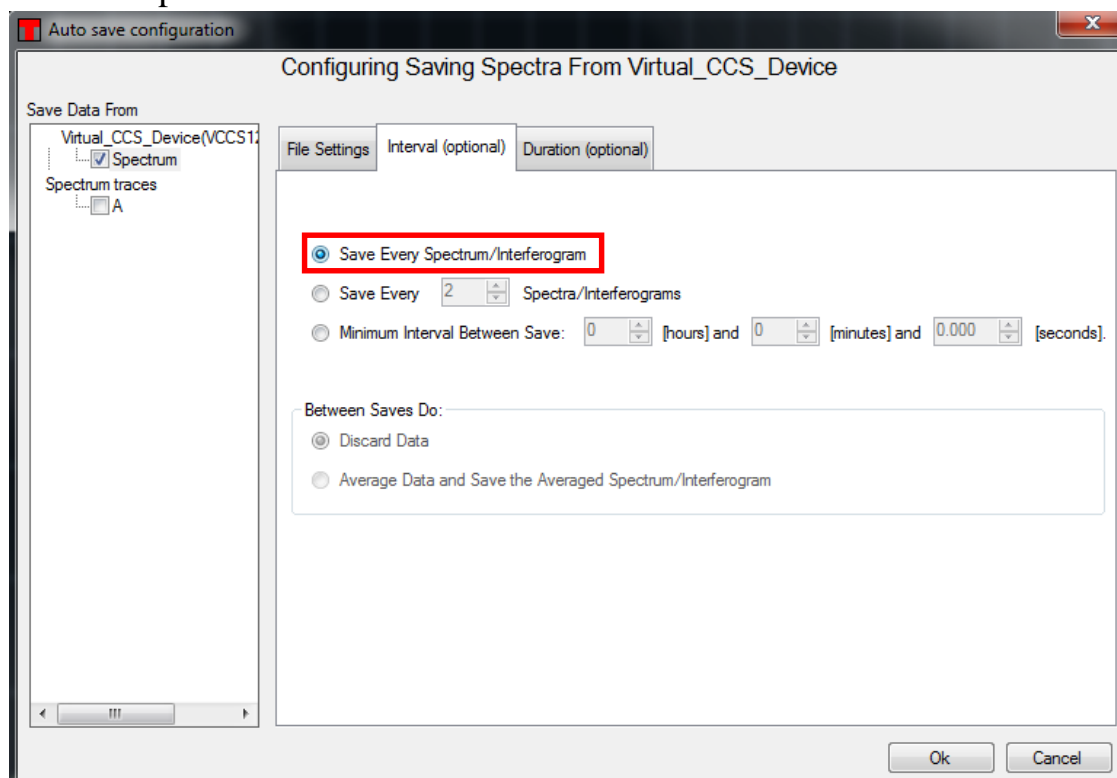

Fig.17.

3.3.10 To **Start Measurement** and **Autosaving** press button **Repeat** on tab **Sweep** (Fig. 18). Deposit a drop of analyte on the top of porous sample using pipet dispenser
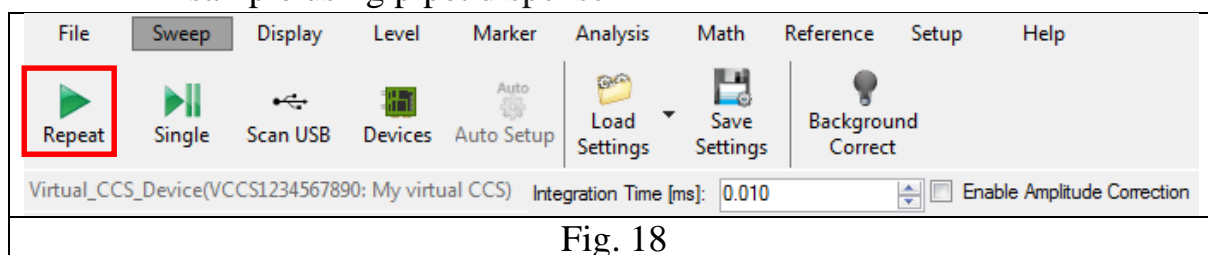

Fig. 18

3.3.1 Files with measured spectrums will be saved in chosen Folder. See example on Fig.19.

Name	Date modified	Type	Size
Spec_20160202_12_56_53_850	02.02.2016 12:56	Text Document	125 KB
Spec_20160202_12_56_53_890	02.02.2016 12:56	Text Document	125 KB
Spec_20160202_12_56_53_920	02.02.2016 12:56	Text Document	125 KB
Spec_20160202_12_56_53_940	02.02.2016 12:56	Text Document	125 KB
Spec_20160202_12_56_53_950	02.02.2016 12:56	Text Document	125 KB
Spec_20160202_12_56_53_970	02.02.2016 12:56	Text Document	125 KB
Spec_20160202_12_56_53_980	02.02.2016 12:56	Text Document	125 KB
Spec_20160202_12_56_54_000	02.02.2016 12:56	Text Document	125 KB
Spec_20160202_12_56_54_020	02.02.2016 12:56	Text Document	125 KB
Spec_20160202_12_56_54_030	02.02.2016 12:56	Text Document	125 KB
Spec_20160202_12_56_54_050	02.02.2016 12:56	Text Document	125 KB

Fig. 19.

3.3.2 Autosaving will be finished after pressing of button **Stop** on tab **Sweep**.

3.3.3 Repeat measurement for different analyte.

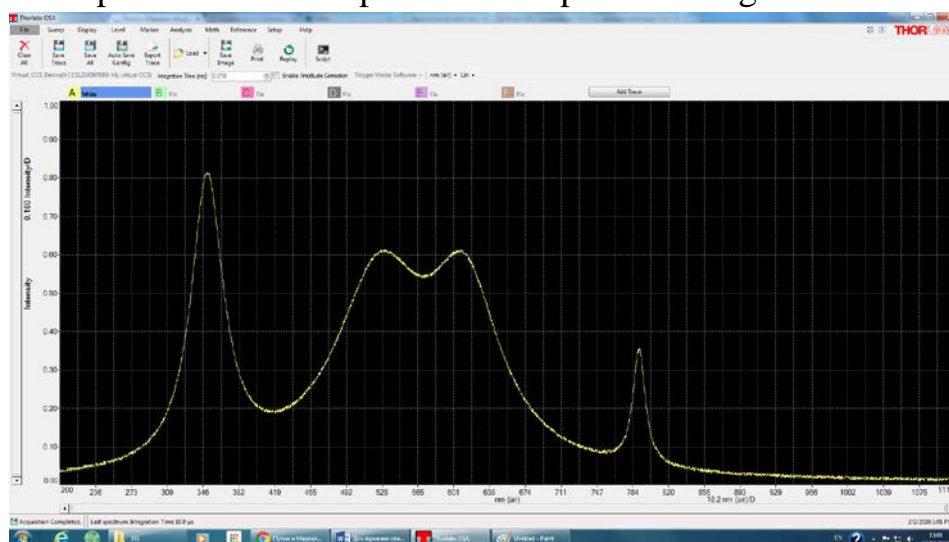3.3.4 Example of reflection spectrum is depicted on Fig. 20.

Fig. 20.

4. Result visualization

4.1 Using Origin, plot dependence of reflection coefficient on wavelength for sample with air filled pores and analyte filled pores.

4.2 Using Origin, plot dependence of reflection coefficient on wavelength for sample with air filled pores and analyte filled pores at different time moments during analyte desorbtion from pores.

4.3 Using Matlab, create 3D-plot: dependence of reflection coefficient on wavelength and time after analyte deposition on porous sample.

Reference list

1. Скришевський В. А. Фізичні основи напівпровідникових хімічних сенсорів, ВПЦ "Київський університет", 190 с., 2006.
2. Sailor M. J. Porous Silicon in Practice, Wiley-VCH, 250 p., 2011.

3. Spectrometer CCS200/M manual
   http://www.thorlabs.de/thorcat/18100/CCS200_M-Manual.pdf
4. MATLAB Programming,
   https://en.wikibooks.org/wiki/MATLAB_Programming